

APSIPA ASC 2024

Asia Pacific Signal and Information Processing Association Annual Summit and Conference 2024

Dec 3rd – 6th 2024 | MACAU, CHINA

PROGRAM BOOK

CONTENT

1 Welcome Message

2 Organizers and Sponsors

3 Organizing Committee

11 Conference Information

13 Program at a Glance

15 Speaker List

- 15 - Keynote Speakers
- 19 - Winter School Speakers
- 22 - Industrial Forum Speakers
- 26 - Tutorial Speakers
- 30 - Women's Forum Speakers

33 Full Program

- 35 - Oral Presentation
- 80 - Poster List

94 AUTHOR INDEX

130 PROCEEDING

WELCOME MESSAGE FROM PRESIDENT



On behalf of the Asia Pacific Signal and Information Processing Association (APSIPA), it is my great pleasure and honor to welcome all of you to the APSIPA Annual Summit and Conference (ASC), held in Macao from December 3 through 6, 2024. This is the 16th APSIPA ASC and the first ASC that is held in this place.

APSIPA was founded in 2009. It was targeted to cover signal and information processing, which is among the most growing research fields, in the most growing region of Asia Pacific. Since then, we have held APSIPA ASC, our flagship conference, every year on various sites in this region. In the past 15 years, we have seen steady growth in paper submissions and participants, even during the pandemic.

APSIPA now has more than one thousand members. Local chapters have been established in Japan, Taiwan, the USA, Singapore, Vietnam, Korea, and Thailand. We are now reaching out to developing countries, particularly ASEAN countries. We have organized workshops in Vietnam and Indonesia in the past years. This year, we set up a travel grant scheme to provide financial support for participants from ASEAN countries and India. More than 20 participants are awarded this grant.

The organizers of APSIPA ASC 2024 have worked very hard to make it successful. I would like to express my greatest appreciation and gratitude to the organizers led by General Co-Chairs, Jiantao Zhou and Kenneth Lam. I believe their hard work will be rewarded when the conference becomes an excellent experience for everyone. Eight technical committees (TCs) have also committed to organizing special sessions, attracting papers, and overseeing the review process. The conference has a variety of programs, including Winter School, Industrial Forum, and Women's Forum, which are organized by APSIPA VPs and BoG members to enhance diversity. In my personal experience, APSIPA provides an invaluable network through a variety of technical and regional areas, and we feel at home during the conference. I hope this tradition continues and is passed on to the future generation.

Tatsuya Kawahara

President, APSIPA (2023-2024)

ORGANIZERS AND SPONSORS

ORGANIZERS

APSIPA Signal and Information Processing

澳門大學 UNIVERSIDADE DE MACAU UNIVERSITY OF MACAU

MACETS 澳門會展旅遊業協會

科技學院 Faculdade de Ciências e Tecnologia Faculty of Science and Technology

IEEE Signal Processing Society

SKL-IOTSC 智慧城市物聯網國家重點實驗室(澳門大學)

IEEE

SPONSOR

玖治科技 Ninenovo

GALAXY MACAU 澳門銀河

ORGANIZING COMMITTEE

Honorary Co-Chairs

K. J. Ray Liu, University of Maryland
C.-C. Jay Kuo, University of Southern California
Haizhou Li, The Chinese University of Hong Kong (Shenzhen)
Wan-Chi Siu, The Hong Kong Polytechnic University
Hitoshi Kiya, Tokyo Metropolitan University
Yonghua Song, University of Macau

Advisory Committee Co-Chairs

Tatsuya Kawahara, Kyoto University
Cheng-Zhong Xu, University of Macau

General Co-Chairs

Jiantao Zhou, University of Macau
Kenneth Lam, The Hong Kong Polytechnic University
Anthony Kuh, University of Hawaii at Manoa
Woon-Seng Gan, Nanyang Technological University
Kosin Chamnongthai, King Mongkut's University of Technology Thonburi

Finance Co-Chairs

Yuanman Li, Shenzhen University
Xiaowei Wu, University of Macau

TPC Co-Chairs

Shaodan Ma, University of Macau
Bonnie Law The Hong Kong Polytechnic University
Yuanfang Guo, Beihang University
(SPS) Zhiyi Yu, Sun Yat-Sen University
(SIPTM) Zhengguo Li, Institute for Infocomm Research, A*STAR
(SLA) Jun Du, University of Science and Technology of China
(BioSiPS) Koichi Fujiwara, Nagoya University
(IVM) Chang-Su Kim, Korea University
(MSF) Minoru Kuribayashi, Tohoku University
(WCN) Po-Chiang Lin, Yuan Ze University, Taiwan
(MLDA) Jen-Tzung Chien, National Yang Ming Chiao Tung University

Plenary Co-Chairs

Yih-Fang Huang, University of Notre Dame

Antonio Ortega, University of Southern California

Zhiguo Gong, University of Macau

Special Session Co-Chairs

Sanghoon Lee, Yonsei University

Weisi Lin, Nanyang Technological University

Toshihisa Tanaka, Tokyo University of Agriculture and Technology

Li Dong, Ningbo University

Tutorial Co-Chairs

Mingyi He, Northwestern Polytechnical University

Shoji Makino, Waseda University

Publicity Co-Chairs

Waleed H. Abdulla, The University of Auckland

Yui Lam Chan, The Hong Kong Polytechnic University

Nam Ik Cho, Seoul National University

Junhui Hou, City University of Hong Kong

Yoshinobu Kajikawa, Kansai University

Publication Co-Chairs

Yi Wang, The Hong Kong Polytechnic University

Liming Zhang, University of Macau

Chi Man Pun, University of Macau

Jinyu Tian, Macau University of Science and Technology

Registration Co-Chairs

Yuan Wu, University of Macau

Li Li, University of Macau

Xianwei Zheng, Foshan University

Industrial Forum Co-Chairs

Gwo-Giun Lee, National Cheng Kung University

Lap-Pui Chau, The Hong Kong Polytechnic University

Jie Chen, Hong Kong Baptist University

Jing-Ming Guo, National Taiwan University of Science and Technology

Ning Xu, Dobly Laboratories
Seishi Takamura, NTT Corporation
Weiwei Sun, Alibaba Group

Sponsorship and Exhibition Co-Chairs
Ryan U, University of Macau
Derek Wong, University of Macau

Overview Section Co-Chairs
KokSheik Wong, Monash University
Isao Echizen, National Institute of Informatics
Thomas Fang Zheng, Tsinghua University

Local Arrangement Co-Chairs
Andrew Jiang, Macau Expo Group
Hui Kong, University of Macau



TPC Members – Area Chairs **Sorted by First Name*

Ahmed Khan
Chang-Su Kim
Chau-Wai Wong
Chul Lee
Daigo Muramatsu
Gaoyan Zhang
Haiwei Wu
Hemant Patil
Jaymar Soriano
Jiandian Zeng
Jinyu Tian
Kazunori Hayashi
Koki Yamada
KokSheik Wong
Kouei Yamaoka
Lai-Kuan Wong
Masaki Kawamura
Michiharu Niimi
Mingyi He
Naohiro Tawara
Nguyen Linh Trung
Peng Shen
Qing Wang
Ruifan Li
Shinsuke Ibi
Shogo Muramatsu
Shoji Makino
Shoko Imaizumi
Simying Ong
Tatsuya Komatsu
Tetsushi Ohki
Thi Lan Le
Thushara Abhayapala
Ven Jyn Kok
Warit Sirichotedumrong
Wei Wang
Woon Seng Gan
Xianwei Zheng
Xiaoxu Li
Xugang Lu
Yuanman Li
Yuchao Dai
Yuen Peng Loh
Yuki Takashima
Yu-Shuen Wang
Yusuke Hioka

TPC Members – Reviewers **Sorted by First Name*

Aaliya Sarfaraz
Abe Narishige
Abel Yu Hao Chai
Ahmed Khan
Amy Bastine
Andy W H Khong
Anh Quang Tran
Ankur Patil
April Pyone Maung Maung
Asad Malik
Bach Kim Ngoc
Beichen Li
Bin Deng
Binbin Song
Binze Wang
Bo Chen
Bo Xiao
C.-C. Jay Kuo
Changfu Xu
Chang-Hee Han
Cheng Zhong
Chenghua He
Chern Hong Lim
Chih-Chang Yu
Chih-Peng Fan
Ching-Ting Tu
Christian Ritz
Chuan Qin
Chul Lee
Chung-Che Wang
Chuong Vo
Daigo Muramatsu
Daisuke Niizumi
Daniel Hladek
Dilshani Mallikarachchi
Dong Wang
Poompat Saengudomlert
Duc-Tan Tran
Duc-Trong Le
Dung Nguyen
Duong Nguyen
Fangxiang Feng
Fariza Fauzi
Felix Albu
Fengpeng Li
Fengyi Huang
Ga-Young Choi
Gejian Zhao
Gene Cheung
Guangxue Zhang
Guo Li
Ha Do
Ha Le
Haihui Liu
Haiwei Wu
Hanh Tran
Hanyu Chen
Hanzhou Wu
Haodong Zou
Haoran Qin
Harumi Murata
He Min
Hemant Patil
Heqi Peng
Hideki Kawahara
Hirohisa Hioki
Hiroshi Higashi
Hiroshi Tsutsui
Hoang Nguyen
Hoang-Dieu Vu
Hong Kook Kim
Hong-Goo Kang
Hrishav Barua
Hsiao-Chi Li
Huaming Du
Hui Miao
Hung-Kuo Chu
Hung-Shin Lee
Huy Nguyen
Hyunho Kang
I-Chen Lin
Imad Barhumi
Ingon Chanpornpakdi
Isao Echizen
Jae-Young Sim
Jen-Yi Pan
Jiajia Pan
Jianing Lin
Jianwei Fei
Jiaying Liu
Jijie Wu
Jin Xuan Teh
Jing Dong
Jing Zhao
Jingdong Chen
Jintae Kim
Jinuk Kwon
Jinyu Tian
Joel Than

Jong-Ok Kim

Juan Wen

Jun Liu

Junichi Yamagishi

Kazuaki Nakamura

Kazuhiro Kono

Kazuhiro Nakadai

Kazunori Hayashi

Kazunori Komatani

Kazushi Ikeda

Keiichi Funaki

Keiichiro Inagaki

Keisuke Imoto

Keita Emura

Kemou Li

Kenta Iwai

Kenta Takahashi

Khac-Hoang Ngo

Khoa Tran

Kien Tran

Kiyoshi Nishikawa

Koi Yee Ng

Koichi Adachi

Koichi Ito

Koksheik Wong

Konstantin Markov

Kotaro Sonoda

Kouei Yamaoka

Kyong Hwan Jin

Lam Sinh Cong

Le Thanh

Lei Chen

Li Li

Li Yujiang

Liang Wang

Liangbin Xie

Lin Wang

Lit Yen Yeo

Liyun Yu

Madhu Kamble

Madoka Hasegawa

Mai Ohta

Man On Pun

Masaaki Fujiyoshi

Masafumi Nishida

Masahiro Okuda

Masakatsu Nishigaki

Masaki Kawamura

Masao Yamagishi

Masashi Unoki

Masato Akagi

Masato Nakayama

Masatsugu Ichino

Masnizah Mohd

Matthew Teow

Michiharu Niimi

Ming-Sui Lee

Minoru Kuribayashi

Miseon Shim

Mitsunori Mizumachi

Mohammad Shahid

Mohammad Kamrul Hasan

Mustafa S. Abdul Karim

Myo Thida

Nam Ik Cho

Naofumi Aoki

Naohiro Tawara

Nawara Mahmood Broti

Ngai Fong Law

Nghia Nguyen

Nguyen Linh Trung

Nguyen Thinh

Nie Shihao

Nirmesh Shah

Noboru Harada

Nobuaki Minematsu

Nobutaka Ono

Norishige Fukushima

Nurul Japar

Olarik Surinta

Osamu Takyu

Peiya Li

Pei-Yun Tsai

Peng Shen

Pengyue Lin

Phat Nguyen Huu

Po-Chyi Su

Qi Liu

Quân Nguyễn

Quynh Tran

Rajeev Kumar

Rongqin Liang

Ruifan Li

Ruijie Yang

Ruikui Wang

Ruixue Duan

Runmin Cong

Ruoyu Wang

Ryo Hayakawa

Ryo Matsuoka

Ryoichi Takashima

Ryouichi Nishimura

Sanghoon Lee

Sansanee Auephanwiriyakul

Sayaka Shiota

Se-Ho Lee

Seisuke Kyochi

Seung-Won Jung

Shang-Chih Lin

Sheng Li

Shih-Yu Chen

Shilin Wang

Shinji Watanabe

Shinnosuke Takamichi

Shinsuke Ibi

Shintami Hidayati

Shogo Muramatsu

Shogo Okada

Shoichi Koyama

Shoji Makino

Shoko Imaizumi

Shuji Komeiji

Shuning Xu

Shunsuke Ono

Simying Ong

Sinan Zhang

Siok Yee Tan

Soohwan Song

Sophia Chulif

Sue Han Lee

Sunao Hara

Taichi Nishimura

Taishi Nakashima

Tai-Shih Chi

Takafumi Koshinaka

Takahiro Shinozaki

Takanobu Nishiura

Takeshi Ikenaga

Tan-Nhu Nguyen

Tatsuya Kawahara

Tatsuya Komatsu

Tatsuya Suzuki

Tatsuya Yokota

Tay Fei Siang

Terumi Umematsu

Tetsuji Ogawa

Tetsuya Kojima

Thai-Mai Dinh

Thanh Huong Ha Thi

Thanh Pham

Thanh-Hai Tran

Thi Lan Le

Thien Huynh-The

Thushara Abhayapala

Thuy-Binh Nguyen

Tomoki Toda

Tomoko Matsui

Tonghui Li

Toru Takahashi

Toshihisa Tanaka

Toshio Irino

Tran Hiep Dinh

Tso-Bing Juang

Uchida Hidetsugu

Vannhu Le

Vantoi Nguyen

Ven Jyn Kok

Viet-Dung Nguyen

Vu-Hoang Tran

Wageesha Manamperi

Wai Lam Hoo

Wakako Nakamura

Waleed Abdulla

Wandeep Singh

Wang Jiayan

Weibin Zhu

Weina Xu

Wei-Qiang Zhang

Weiwei Lin

Wei-Wen Hsu

Wei-Yang Lin

Wen-Liang Hwang

Wenqi Zhuo

Wenting Zhu

Xiangui Kang

Xiangyu Chen

Xiangyu Zhao

Xianwei Zheng

Xiao Na Zeng

Xiao Yang
Xiaohan Zhao
Xina Liu
Xingrun Wang
Xingyan Chen
Xinqi Jiang
Xinyu Yang
Xugang Lu
Xuran Yao
Xuyang Zhao
Yan Chen
Yang Lin
Yaxin Mei
Yi Wen Kerk
Yi-Chin Huang
Yi-Chong Zeng
Yijun Liu
Yiming Chen
Ying-Ren Chien
Yingxin Yi
Yiqi Tew
Yong Xu
Yoonsik Choe
Yoshiaki Bando
Yoshihiro Maeda
Yoshikazu Washizawa
Yoshimitsu Kuroki
Youzhi Tu
Yu Zhao
Yuanchen Niu
Yuanman Li
Yue Long
Yuen Peng Loh
Yufei Hou
Yuhang Zhang
Yuhao He
Yuki Okamoto
Yukoh Wakabayashi
Yuma Kinoshita
Yuma Shirahata
Yunqi Zhang
Yunqian Wang
Yurong Guo
Yusuke Fujita
Yuzhu Liang
Zewei Chen
Zhen Yang
Zheng Li
Zheng Yao
Zheyuan Li
Zhiguang Yang
Zhiyuan Tang
Zihan Li
Zihao Peng
Zihao Zheng
Zou Yongpeng

CONFERENCE INFORMATION

CONFERENCE VENUE

Galaxy International Convention Center,

Galaxy Macau Cotai Macau, Macau

REGISTRATION

Foyer, 2F

Open Hours:

Tuesday, December 3, 2024	08:30-18:00
Wednesday, December 4, 2024	08:30-18:00
Thursday, December 5, 2024	08:30-18:00
Friday, December 6, 2024	08:30-13:30

BADGE

Please wear and clearly display your name badge to attend the conference. All participants will receive the badge upon registration. Access to the conference venue will not be granted without a proper name badge.

CONFERENCE POLICY

✓ Smoking is prohibited at all times in the conference rooms and the entire building.

✓ Please switch your mobile phone off or to vibration mode during all sessions.

✓ No photography or recording are allowed in any sessions.

LANGUAGE

The official language of the conference is English, which will be used in all presentations and printed materials.

LUNCH/COFFEE BREAK

Lunch will be served at LG1 Ball Room.

Coffee Break will be served at Foyer.

SOCIAL PROGRAMS

Welcome Reception

Date/Time: 17:30-20:00, Dec. 3

Ball Room LG1

Banquet

Date/Time: 18:00-21:00, Dec. 5

Merged Room 16+17+18

INSTRUCTIONS FOR ORAL PRESENTERS

Upload Presentation Files

All the presenters are required to use the laptops provided by the organizer for presentation. If you use a PowerPoint presentation, please save the file in USB and upload it to the conference laptop in the meeting room 15 minutes before your session starts. Our staff will assist you to save your file. We highly recommend you bring two copies of your file as a backup.

Presentation Time

15 minutes presentation + 5 minutes Q&A, total 20 minutes.

INSTRUCTIONS FOR POSTER PRESENTERS

- ✓ The poster board in portrait format is reserved for each poster.
- ✓ Authors shall prepare their posters in standard A0 size (841mm x 1189 mm) and mount them on the poster board.
- ✓ The presenting author must present the work in front of the poster during the poster session.

FLOOR PLAN


2F
: Meeting Room
: Foyer
: Elevator (To LG for Lunch and Reception)
Foyer
Foyer
SPEAKERS PREP ROOM 2
ORGANIZERS OFFICE 2
DRESSING ROOM
1
2
3
4
5
6
8
9
10
16
17
18
19
APSIPA
Signal and Information Processing
GALAXY
INTERNATIONAL CONVENTION CENTER
銀河國際會議中心

PROGRAM AT A GLANCE

3-Dec		
Time/Location	**Meeting Room 6**	**Meeting Room 9**
09:30-11:30 (Coffee Break 10:10-10:30 **Foyer**)	**Tutorial 01** EEG Signal Processing and Machine Learning	**Tutorial 02** From Statistical to Causal Inferences for Time-Series and Tabular Data
11:30-13:00	Lunch **(LG Ball Room)**	
13:00-15:00	**Tutorial 03** Human-Centric RF Sensing: Pose Estimation, ECG Monitoring and Self-Supervised Learning	**Winter School** **Part 1:** Overview of Neural Network AI **Part 2:** Hopfield Neural Network Fundamental for Machine Learning
15:00-15:30	Coffee Break **(Foyer)**	
15:30-17:30	**Tutorial 04** Emerging Topics for Speech Synthesis: Versatility and Efficiency	**Winter School** **Part 3:** Deep Learning for Image forensics **Part 4:** Generative Modeling and Learning for Conversational AI
17:30-20:00	Welcome Reception **(LG Ball Room)**	

4-Dec										
Time/Location	**Meeting Room 6**	**Meeting Room 8**	**Meeting Room 9**	**Meeting Room 1**	**Meeting Room 2**	**Meeting Room 3**	**Meeting Room 4**	**Meeting Room 5**	**Meeting Room 10**	**Foyer**
09:00-09:40	Opening Ceremony & Invited Talk for the APSIPA Sadaoki Furui Prize Paper Award (**Merged Room 16+17**)									
09:40-10:40	**Keynote Speech I: "Rate-Distortion Optimization in Video/Image Compression: From Temporal Dependency Formulation to Learning-based Modeling" (Merged Room 16+17)**									
10:40 - 11:00	Coffee Break **(Foyer)**									
11:00-12:20	Audio Processing (84, 111, 115,116)	Biomedical Signal Processing and Systems (77, 107, 130, 136)	Machine Learning and Data Analytics (17, 57, 62)	Machine Learning and Data Analytics (171, 209, 211, 56)	Machine Learning and Data Analytics (240, 323, 385, 386)	Image, Video, and Multimedia (44, 53, 58, 66)	Signal and Information Processing & Systems (51, 74, 86, 88)	Speech and Language Processing (26, 29, 37, 54)	Speech and Language Processing (55, 78, 79)	Poster (25, 32, 39, 229, 59, 63, 65, 75, 83, 189, 195)
12:20-14:00	[TC1] TC Meeting (SPS)	[T02] TC Meeting (MSF)	[TC3] TC Meeting (BioSiPS)	[TC4] TC Meeting (SLP)	[TC5] TC Meeting (AA)	[TC6] TC Meeting (WCN)	[TC7] TC Meeting (IVM)	Lunch **(LG Ball Room)**		
14:00-16:00	Audio Processing (145, 153, 203, 212, 218, 214)	Biomedical Signal Processing and Systems (140, 161, 221, 255, 302, 339)	Best Student Paper Competition (192, 460, 433, 333, 125)	Industrial Forum	Multimedia Security and Forensics (34, 48, 49, 186, 315)	Image, Video, and Multimedia (90, 122, 131, 138, 495)	Signal and Information Processing & Systems (100, 109, 124, 133, 143)	Speech and Language Processing (80, 82, 97, 110, 120)	Speech and Language Processing (126, 146, 147, 162, 163)	Poster (89, 98, 99, 102, 108, 112, 113, 117, 204, 144, 206, 150, 157,181)
16:00-16:20	Coffee Break **(Foyer)**									
16:20-18:00	Audio Processing (230, 235, 243, 246, 277)	Biomedical Signal Processing and Systems (370, 398, 466, 431, 456)	Best Paper Competition (159, 382, 85, 172, 50)	Overview Session	Multimedia Security and Forensics (359, 360, 417, 434, 505)	Image, Video, and Multimedia (193, 219, 238, 247,496)	Signal and Information Processing & Systems (170, 173, 149, 217, 225)	Speech and Language Processing (184, 190, 200, 118, 237)	Speech and Language Processing (239, 244, 249, 274, 301)	Poster (160, 164, 167, 177, 188, 194, 216, 222)
18:00-20:00	BoG Meeting **(Meeting Room 19)**									

5-Dec										
Time/Location	Meeting Room 6	Meeting Room 8	Meeting Room 9	Meeting Room 1	Meeting Room 2	Meeting Room 3	Meeting Room 4	Meeting Room 5	Meeting Room 10	Foyer
09:00-10:00	Keynote Speech II: Learning from Unreliable Sources via Crowdsourcing" (Merged Room 16+17)									
10:00-10:20	Coffee Break **(Foyer)**									
10:20-12:00	Audio Processing (284, 288, 290, 296, 309)	Overview Session	Emerging Technologies and Applications Of Image Processing And Computer Vision (45, 46, 47, 95, 166)	Advanced Topics on Sound Event and Scene Analysis (52, 158, 380, 406, 424)	Recent Advances in Multimedia Enrichment and Security (123, 198, 202, 207, 208)	Image, Video, and Multimedia (252, 253, 269, 271, 295)	Signal and Information Processing & Systems (226, 231, 260, 273, 307)	Speech and Language Processing (320, 326, 342, 350, 361)	Advanced Signal Processing for Information Collection and Data Analysis in Wireless Environmental Sensing (105, 142, 199, 279, 303)	Poster (228, 248, 245, 262, 265, 266, 272, 283, 286)
12:00-14:00	APSIPA Women Forum	Lunch **(LG Ball Room)**		TC Chair Meeting	Lunch **(LG Ball Room)**					
14:00-15:20	Audio Processing (331, 347, 362)	Audio Processing (363, 378, 383, 400)	High Performance Image and Video Processing and Applications (92, 215, 258, 280)	New Frontiers in Biometric Authentication (220, 453, 317, 409)	Recent Advances in Multimedia Enrichment and Security (210, 227, 233, 251)	Image, Video, and Multimedia (297, 300, 305, 330)	Signal and Information Processing & Systems (318, 328, 340, 376)	Speech and Language Processing (388, 402, 404, 407)	Advancements in Biosignal Decoding and Neuromodulation for Human Function Enhancement (72, 178, 256, 259, 476)	Poster (287, 298, 308, 314, 319, 321, 324, 329, 334)
15:20-15:40	Coffee Break **(Foyer)**									
15:40-16:40	Keynote Speech III: "AI and Cognitive Health" (Merged Room 16+17)									
16:40-18:00	Audio Processing (408, 412, 419, 425)	Audio Processing (446, 478, 470, 487)	High Performance Image and Video Processing and Applications (345, 349, 368, 440)	Wireless Communications and Networking (344, 381)	Recent Advances in Multimedia Enrichment and Security (257, 263, 267, 336)	Image, Video, and Multimedia (337, 341, 364, 490)	Signal and Information Processing & Systems (397, 414, 442)	Speech and Language Processing (415, 435, 438)	Advancements in Biosignal Decoding and Neuromodulation for Human Function Enhancement (270, 365, 373)	Poster (352, 353, 366, 396, 389, 391, 393, 401, 338, 413)
18:00-21:00	Banquet **(Merged Room 16+17+18)**									

6-Dec										
Time/Location	Meeting Room 6	Meeting Room 8	Meeting Room 9	Meeting Room 1	Meeting Room 2	Meeting Room 3	Meeting Room 4	Meeting Room 5	Meeting Room 10	Foyer
9:00-10:20	Signal Processing for Drone Audition & Recent Advances in Intelligent Signal Processing (128, 180, 278, 292, 294)	Converging AI and Computer Vision: Innovations and Potential (104, 175, 205, 254)	AI-Driven Innovations in Cybersecurity Advanced Applications in Signal Processing, Multimedia Security, and Privacy (106, 241, 261, 327)	Embedded and Real-Time Systems for AI and Signal Processing Applications (71, 165, 304, 311)	Selected Papers from APSIPA Workshop on Advanced Signal and Information Processing (41, 42, 43, 94)	Image, Video, and Multimedia (372, 410, 428, 451)	Acoustic Scene Analysis and Signal Enhancement Based on Advanced Signal Processing and Machine Learning (127, 155, 168, 355)	Speech and Language Processing (441, 444, 449, 452)	Overview Session	Poster (427, 432, 436, 437, 439, 443, 445, 450, 223, 481)
10:20-10:40	Coffee Break **(Foyer)**									
10:40-12:00	Signal Processing for Drone Audition & Recent Advances in Intelligent Signal Processing (411, 430, 457,387)	Converging AI and Computer Vision: Innovations and Potential (282, 354, 357, 371)	Multimedia Processing Systems in the AI Era (291, 293, 374, 405)	Embedded and Real-Time Systems for AI and Signal Processing Applications (312, 343, 375, 394)	Selected Papers from APSIPA Workshop on Advanced Signal and Information Processing (121, 132, 134, 135)	Image, Video, and Multimedia (455, 462, 489)	Advanced Topics for Automatic Speakers Recognition (67, 332, 467)	Speech and Language Processing (468, 483, 485)	Few-shot Vision, Language, and Multimedia Processing under LLMs (176, 448, 465, 492)	Poster (458, 461, 429, 469, 471, 473, 464, 493)
12:10-13:30	APSIPA General Assembly & Closing Ceremony **(Merged Room 16+17)**									

KEYNOTE SPEAKERS



Ce Zhu

Dean, Glasgow College

University of Electronic Science and Technology of China (UESTC)

IEEE/Optica/IET/AAIA Fellow

ChangJiang Distinguished Professor

Rate-Distortion Optimization in Video/Image Compression: From Temporal Dependency Formulation to Learning-based Modeling

The past decades have witnessed remarkable advancements in image/video coding techniques and their diverse applications, where rate-distortion optimization (RDO) plays a crucial role in maximizing coding efficiency. In the traditional block-based hybrid coding framework, the extensive utilization of spatial-temporal prediction makes the processing of each coding unit highly dependent and entangled, rendering the global RDO problem particularly challenging. One feasible solution is to delicately formulate the spatial-temporal dependencies accurately. I will first discuss achieving temporally dependent RDO in one pass coding, and present our RDO work on top of various video coding standards, from H.264/AVC to H.265/HEVC, and to the latest Versatile Video Coding (H.266/VVC). On the other hand, leveraging the powerful nonlinear transformation and entropy modeling capabilities of neural networks, end-to-end learning-based approaches have emerged as a highly promising alternative for achieving high efficiency image/video compression. In the second part, I will touch some work on end-to-end learning-based image compression, particularly highlighting our recent advancements in optimizing the entropy model.

Biography

Ce Zhu (IEEE/Optica Fellow) is currently a Professor at the University of Electronic Science and Technology of China, Chengdu, China. His research interests include image/video coding and communications, 3D video, visual analysis, perception and applications. He has served on the editorial boards of a few journals, including as an Associate Editor of IEEE Transactions on Image Processing (2016-2020). He has also served as a Guest Editor of a few special issues in international journals, including as a Guest Editor in the IEEE Journal of Selected Topics in Signal Processing (2020). He was an APSIPA Distinguished Lecturer (2021-2022), and also an IEEE Distinguished Lecturer of Circuits and Systems Society (2019-2020). He is serving as the Chair of IEEE ICME Steering Committee (2024-2025), and the Chair of IEEE Chengdu Section. He is a co-recipient of multiple paper awards at international conferences, including the Best Demo Award in IEEE MMSP 2022, and the Best Paper Runner Up Award in IEEE ICME 2020. For more information, please visit his homepage at http://www.avc2-lab.net/~eczhu.



Georgios Giannakis

Department of Electrical and Computer Engineering

University of Minnesota, Twin Cities

McKnight Presidential Chair in ECE

Learning from Unreliable Sources via Crowdsourcing

Crowdsourcing, as the name suggests, harnesses the information provided by crowds of human annotators to perform learning tasks, such as word tagging in natural language processing, crowdsensing, and ChatGPT, among others. Even though crowdsourcing can be efficient and relatively inexpensive, combining the noisy, scarce, and potentially adversarial responses provided by multiple annotators of unknown expertise can be challenging, especially in unsupervised setups, where no ground-truth data is available.

Focusing on the classification task, the first part of this talk will touch upon models and algorithms for label fusion along with their performance. Approaches will be also discussed for data-aware crowdsourcing, and links will be outlined with deep-, self-supervised, and meta-learning. Aiming to robustify crowdsourced classification against adversarial attacks, the last part will cover spectrum-based algorithms to flag and mitigate the effect of spammers. If time allows, means of dealing with dependent annotators will be discussed briefly.

Biography

Georgios B. Giannakis received his Diploma in Electrical Engr. (EE) from the Ntl. Tech. U. of Athens, Greece, 1981. From 1982 to 1986 he was with the U. of Southern California (USC), where he received his MSc. in EE, 1983, MSc. in Mathematics, 1986, and Ph.D. in EE, 1986. He was with the U. of Virginia from 1987 to 1998, and since 1999 he has been with the U. of Minnesota (UMN), where he held an Endowed Chair of Telecommunications, served as director of the Digital Technology Center 2008-21, and since 2016 he holds a UMN Presidential Chair in ECE.

His interests span the areas of statistical learning, communications, and networking - subjects on which he has published more than 500 journal papers, 810 conference papers, 26 book chapters, two edited books and two research monographs. Current research focuses on Data Science with applications to IoT, and power networks with renewables. He is the (co-) inventor of 36 issued patents, and the (co-)recipient of 10 best journal paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize. He received the IEEE-SPS Norbert Wiener Society Award (2019); EURASIP's A. Papoulis Society Award (2020); Technical Achievement Awards from the IEEE-SPS (2000) and from EURASIP (2005); the IEEE ComSoc Education Award (2019); and the IEEE Fourier Technical Field Award (2015). He is a member of the Academia Europaea, Greece's Academy of Athens, and Fellow of the National Academy of Inventors, the European Academy of Sciences, UK's Royal Academy of Engineering, Life Fellow of IEEE, and EURASIP. He has served the IEEE in several posts, including that of a Distinguished Lecturer for the IEEE-SPS.



Helen Meng

Systems Engineering & Engineering Management

The Chinese University of Hong Kong, Hong Kong

Patrick Huen Wing Ming Chair Professor

AI and Cognitive Health

With the global population ageing rapidly, a key health concern lies in Neurocognitive Disorders (NCD), also known as dementia -- a common form being Alzheimer's Disease (AD). NCDs are particularly prominent in older adults, which has an insidious onset followed by gradual, irreversible deterioration in cognitive domains (memory, communication, etc.). Thus the screening NCD is crucial for timely intervention to slow down disease progression. We will present our work in the use of spoken language for assessing cognitive health, including a carefully designed speech data collection protocol that has contextual relevance to the linguistic environment in Hong Kong, Macao and the Greater Bay Area, the development of AI-enabled speech and language analytics approaches, and the use of different acoustic and linguistic features for the AD detection. Experimental results demonstrate the feasibility of automating cognitive health monitoring with the use of speech and language technologies, which offers the advantages of accessibility, non-invasiveness and affordability. This work is funded by the HKSAR Government's Theme-based Research Scheme.

Biography

Helen Meng is Patrick Huen Wing Ming Professor of Systems Engineering & Engineering Management at The Chinese University of Hong Kong. She received all her degrees from MIT and joined CUHK in 1998. She is the Founding Director of the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies in 2005, which has been recognized as a Ministry of Education of China (MoE) Key Laboratory since 2008. In 2006, she founded the Tsinghua-CUHK Joint Research Centre for Media Sciences, Technologies and Systems and has served as its Director. In 2013, she helped establish the CUHK Stanley Ho Big Data Decision Analytics Research Center and serves as its Founding Director. She served as former Associate Dean (Research) of Engineering (2006-2010), and former Chairman of the Department (2012-2018).

Helen’s professional services include former Editor-in-Chief of the IEEE Transactions on Audio, Speech and Language Processing, and a member of the IEEE Board of Governors. She has served or is serving as a member of the Advisory Panel of the Hong Kong Science and Technology Park Corporation, the review panels of the Swedish Research Council European Research Infrastructure Initiative, and the National Centres of Competence in Research of the Swiss National Science

Foundation. She is a member of the HKSAR Government's Steering Committee on eHealth Record Sharing, Convenor of the Engineering Panel HKSAR Government's Competitive Research Funding Schemes for the Self-financing Degree Sector, member of the Hong Kong/Guangdong ICT Expert Committee and Coordinator of the Working Group on Big Data Research and Applications, Council membership of the Open University of Hong Kong, member of the Research Grants Council, former Council Member of the Hong Kong Productivity Council, former member of the HKSAR Government's Digital 21 Strategy Advisory Committee, and Chairlady of the Working Party of the Manpower Survey of the Information Technology Sector (undertaken by the Hong Kong Census and Statistics Department) for 2014-2017.

Helen is a recognized scholar in her field. She leads the interdisciplinary research team that received the first Theme-based Research Scheme Project in Artificial Intelligence in 2019. Her recent awards include 2019 IEEE Signal Processing Society Leo L Beranek Meritorious Service Award, 2018 CogInfoComm Best Paper Award, 2017 Outstanding Women Professional Award (one of 20 since 1999), 2016 Microsoft Research Outstanding Collaborator Award (one of 32 academics worldwide), 2016 IBM Faculty Award, 2016 IEEE ICME Best Paper Award, 2015 ISCA Distinguished Lecturer, 2015 HKCS inaugural Outstanding ICT Women Professional Award and 2012 Asia-Pacific Signal and Information Processing Association (APSIPA) inaugural Distinguished Lecturer. Prior to that, she has also received such awards as the CUHK Faculty of Engineering Exemplary Teaching Award, Young Researcher Award and Service Award; APSIPA Best Oral Paper Award, and 2009 Ministry of Education Higher Education Outstanding Scientific Research Output Award in Technological Advancements. She has delivered numerous invited and keynote talks, such as IEEE SIDAS 2016, ASTRI-HPE Conference 2016, Internet Economy Summit 2017, GMIC 2017, INTERSPEECH 2018 Plenary, SIGDIAL 2019 Keynote, etc. She is a Fellow of the Hong Kong Computer Society, Hong Kong Institution of Engineers, International Speech Communication Association and IEEE.

WINTER SCHOOL SPEAKERS

AI SIP: Hopfield Neural Network and Deep Learning

Mingyi He



Northwestern Polytechnical University

Hopfield Neural Network Fundamental for Machine Learning

Mingyi He, Professor of Northwestern Polytechnical University (NPU), visiting professor of Adelaide Univ and Sydney Univ, Australia. His research interests focus on advanced machine vision and intelligent processing, including signal and image processing, computer vision, integrated image and graphics processing, hyper-spectral remote sensing, 3D information acquisition and processing, and neural network artificial intelligence. Prof. He published over 300 journal and conference papers, as well as 3 monographs (Neural Computing, Neural Network Signal Processing Systems, Digital Image Processing) etc. He is the recipient of 2012 IEEE CVPR best paper award, 2017 APSIPA best deep/machine learning paper award. He has obtained 11 scientific prizes, 3 teaching prizes from China (ministry and province government). He is a recipient of the government lifelong subsidy from the state council of China, 2012 Scientific Chinese, 2017 Baosteel outstanding teacher award. He also received certificates from the IEEE Signal Processing Society in 2014, APSIPA in 2019, the Chinese Institute of Electronics in 2018 and 2020, China Remote Sensing Committee in 2023. He has acted as the general chair or the TPC (co)chair for a number of international conferences. He had addressed a number of keynote talks or invited plenary talks. He is/was an Associate Editor or guest editor for IEEE TGRS, IEEE Jstars, APSIPA T-SIP, RS, JIG, SP etc. Professor He is CIE Fellow, CSIG Fellow, Vice president of APSIPA (21-26).



Ngai Fong (Bonnie) Law

The Hong Kong Polytechnic University

Deep Learning for Image forensics

Bonnie is currently an associate professor at the Hong Kong Polytechnic University. Her research interests focus on signal and image processing, in particular image forensics which involves works on seam carving and source camera identification using deep learning. Throughout the years, she has engaged in various professional activities. For example, she served as the general co-chair of the 2017 International Conference on SP, Communication and Computing. She also held positions as the Treasurer of the IEEE HK Chapter of SP from 2009 to 2016 and as the secretary of APSIPA Headquarters since 2009. In 2011, she was awarded the Best paper award in the computing track at the IEEE Annual India Conference.



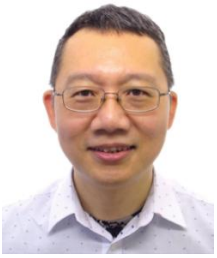
Dr Jen-Tzung Chien

Kyushu National Yang Ming Chiao Tung University

Generative Modeling and Learning for Conversational AI

Jen-Tzung Chien is currently the Lifetime Chair Professor in National Yang Ming Chiao Tung University, Hsinchu, Taiwan. He has authored more than 250 peer-reviewed articles in machine learning, deep learning, and Bayesian learning with applications on speech and natural language processing, and three books including Bayesian Speech and Language Processing, Cambridge University Press, 2015, Source Separation and Machine Learning, Academic Press, 2018, and Machine Learning for Speaker Recognition, Cambridge University Press, 2020. He was a Tutorial Speaker of AAAI, IJCAI, ACL, KDD, ICASSP, COLING and Interspeech. He received the Best Paper Award in IEEE Workshop on Automatic Speech Recognition and Understanding in 2011, and IEEE International Workshop on Machine Learning for Signal Processing in 2023.

INDUSTRIAL FORUM SPEAKER



Dr. Yusen Eason Lin

Founder & CEO, Wellgen Medical, Co., LTD. Taiwan

Research, Clinical, and Business Challenges of AI and Machine Learning Applications in Medicine - A Case Study in Metastatic Cancer and Infectious Diseases Detections by Microscopic Imaging

Integrating Artificial Intelligence and Machine Learning in medical diagnostics presents transformative opportunities, particularly in detecting metastatic cancer and infectious diseases. However, significant challenges impede widespread adoption and effective implementation. This talk examines four critical barriers: data imbalance, data overfitting, data sharing insecurity, and user-payer misalignment in healthcare AI applications. We explore how data imbalance and overfitting compromise model generalization across diverse patient populations and propose solutions through advanced validation techniques and synthetic data generation. The presentation addresses the concerns of medical data security and sharing. Finally, we analyze the economic challenges of AI implementation in healthcare settings, where end-users often differ from payment stakeholders. Practical solutions, including value-based pricing models and tiered service structures, are discussed. This comprehensive analysis provides a roadmap for healthcare institutions, researchers, startups, and industry partners to navigate the complex landscape of medical AI implementation.

Biography

Dr. Eason Lin is the founder of Wellgen Medical and a full professor at National Kaohsiung Normal University, TAIWAN. He was a research assistant professor at the Department of Medicine (2000 ~ 2001), a visiting professor at the University of Pittsburgh (2013 ~ 2016), and a Visiting Scholar at the School of Public Health, UC Berkeley (2022). His professional expertise is in AI/machine learning in medical image processing, medical device design, and commercialization. Dr. Lin was awarded as “Extinguished Researcher Award” by the National Science and Technology Council and the “Young International Affair Advocate” by Executive Yuan, Taiwan. Dr. Lin has served as an expert consultant at the Taiwan CDC, the Taiwan Accreditation Foundation, and numerous healthcare organizations. Dr. Lin can be reached at easonlin@nknu.edu.tw or www.linkedin.com/in/eason-lin-tw.



Dr. Hao Wu

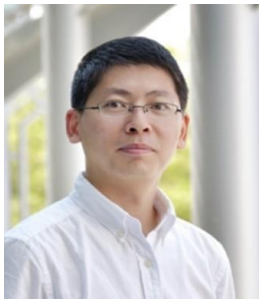
Chief-Executive-Officer (CEO) of Shenzhen Ninenovo Technology and RingConn LLC, China

Smart Rings: Pioneering Biomedical Technologies for Transformative Healthcare Applications

Over the past decade, smart wearables like smart watches and fitness bands have been developed and widely used for health monitoring, including tracking steps, heart rate, and blood oxygen levels. These devices utilize advancements in sensor technology and data analytics to deliver real-time health insights, making personal health tracking more accessible and efficient. However, smartwatches are primarily used for commercial health monitoring rather than in clinical healthcare or medical applications due to several limitations: they are uncomfortable to wear at night, lack comprehensive medical diagnostic capabilities, and offer limited healthcare services. With ongoing advancements in sensor and electronics miniaturization, new wearables like smart rings are emerging, providing more compact and specialized health monitoring solutions. RingConn, a global leader in smart wearables, has released two generations of smart rings, serving over 100,000 customers worldwide. Notably, the RingConn Gen 2 is the first smart ring to feature sleep apnea monitoring, and its cutting-edge technological specifications set a new standard in the smart ring market. In this workshop, Dr. Tony Wu, Co-Founder and CEO of RingConn, will showcase how RingConn integrates pioneering biomedical technologies into its smart rings. He will also discuss the company's long-term vision for expanding healthcare applications and services.

Biography

Hao Wu (Member, IEEE) received the B.Sc. degree from the Nanjing University of Posts and Telecommunications, China, in 2013, and the M.Sc. and Ph.D. degrees from The University of Hong Kong, Hong Kong, in 2014 and 2018, respectively. From 2018 to 2019, he was a Research Fellow with Tencent, Shenzhen, China. From 2019 to 2024, he was an Assistant Professor with the College of Electronics and Information Engineering, Shenzhen University, China. He is currently the Chief-Executive-Officer (CEO) of Shenzhen Ninenovo Technology and RingConn LLC. His research interests include computational intelligence, machine learning, signal processing, and their applications on intelligent transportation, biomedical engineering, and medial image analysis. He has published over ten articles in his research area.



Dr. Jianquan Liu, Director & Senior Principal Researcher, NEC Corporation, Japan

An Industry Perspective: Video Analytics meets Generative AI

In this talk, Dr. Jianquan Liu presents an industry perspective on the convergence of video analytics and generative AI. The talk begins with an overview of video analytics, covering advancements in action recognition, object tracking, human-object interactions, scene recognition, and behavioral pattern analysis. These technologies enable efficient extraction, retrieval, visualization, and summarization of video content. The presentation then explores the impact of generative AI, particularly large language models (LLMs), on video understanding. It discusses how LLMs enhance object recognition, semantic segmentation, action recognition, captioning, visual question answering, and storytelling. Dr. Liu provides industry case studies to illustrate these applications while also addressing limitations and challenges. The talk introduces NEC's narrative summarization framework, designed to tackle key challenges in video analytics. It concludes with a demonstration of "Video with LLM" technology, showcasing its practical application in automating traffic accident investigation reports. This presentation offers valuable insights into the current state and future potential of AI-driven video intelligence, bridging the gap between technical innovation and practical application for both industry professionals and general audiences.

Biography

Jianquan Liu is currently the Director and Head of Video Insights Discovery Research Group at the Visual Intelligence Research Laboratories of NEC Corporation, working on the topics of multimedia data processing. He is also a Visiting Professor at Nagoya University and an Adjunct Professor at Hosei University, Japan. Prior to NEC, he was a development engineer in Tencent Inc. from 2005 to 2006, and was a visiting researcher at the Chinese University of Hong Kong in 2010. His research interests include high-dimensional similarity search, multimedia databases, web data mining and information retrieval, cloud storage and computing, and social network analysis. He has published 70+ papers at major international/domestic conferences and journals, received 30+ international/domestic awards, and filed 70+ PCT patents. He also successfully transformed these technological contributions into commercial products in the industry. Currently, he is/was serving as the Industry Co-chair of IEEE ICIP 2023,

2025 and ACM MM 2023, 2024; the General Co-chair of IEEE MIPR 2021; the PC Co-chair of IEEE IRI 2022, ICME 2020, AIVR 2019, BigMM 2019, ISM 2018, ICSC 2018, ISM 2017, ICSC 2017, IRC 2017, and BigMM 2016; the Workshop Co-chair of IEEE AKIE 2018 and ICSC 2016; the Demo Co-chair of IEEE MIPR 2019 and MIPR 2018. He is a senior member of ACM and IEEE, and a member of IEICE, IPSJ, APSIPA and the Database Society of Japan (DBSJ), a member of expert committee for IEICE Mathematical Systems Science and its Applications (2017-), and IEICE Data Engineering (2015-2021), and an associate editor of IEEE TMM (2023-), ACM TOMM (2022-), EURASIP JIVP (2023-), IEEE MultiMedia Magazine (2019-2022), ITE Transaction on Media Technology and Applications (2021-), APSIPA Transactions on Signal and Information Processing (2022-), and the Journal of Information Processing (2017-2021). Dr. Liu received the M.E. and Ph.D. degrees from the University of Tsukuba, Japan.



TUTORIAL SPEAKERS

[T01] EEG Signal Processing and Machine Learning



Imperial College London, UK

Saeid (Saeed) Sanei

SAEID SANEI, a Fellow of British Computer Society (FBCS), received his PhD in biomedical signal processing from Imperial College London, UK. He is an internationally known researcher in electroencephalography (EEG) signal and machine learning. His current research focus is on application of adaptive and nonlinear signal processing, subspace analysis, and tensor factorization to EEG, BCI, speech, and medical signal and images. He has published five books (monograms), several edited books and book chapters, and over 440 papers in peer reviewed journals and conference proceedings. He has served as an Associate Editor for the IEEE Signal Processing Letters, IEEE Signal Processing Magazine, and Journal of Computational Intelligence and Neuroscience. He also served as the IEEE MLSP and SPTM Committee Member, and as the Chair and Organizer of many IEEE prestigious conferences including ICASSP 2019 in Brighton, UK. Currently, he is a Visiting Academic in digital health to Imperial College London, UK

[T02] From Statistical to Causal Inferences for Time-Series and Tabular Data



Zhejiang University, University of Illinois Urbana Champaign, China

Pavel Loskot

Pavel Loskot joined the ZJU-UIUC Institute in January 2021 as an Associate Professor after being 14 years with Swansea University in the UK as the Senior Lecturer. He received his PhD degree in Wireless Communications from the University of Alberta in Canada, and the MSc and BSc degrees in Radioelectronics and Biomedical Electronics, respectively, from the Czech Technical University of Prague in the Czech Republic. In the past 25 years, he was involved in numerous collaborative research and development projects with academia and industry, and also held a number of consultancy contracts. He is the Senior Member of the IEEE, Fellow of the Higher Education Academy in the UK, and the Recognized Research Supervisor of the UK Council for Graduate Education. His current research interests focus on mathematical and probabilistic modeling, statistical signal processing and machine learning for multi-sensor data.

[T03] Human-Centric RF Sensing: Pose Estimation, ECG Monitoring and Self-Supervised Learning



University of Science and Technology of China, China

Yan Chen

Dr. Chen is a Full Professor and the Vice Dean at the School of Cyber Science and Technology of USTC. He received B.S., M.Phil., and Ph.D. degrees from USTC, HKUST, and University of Maryland College Park, respectively. He has authored 200+ papers and 2 books. He is the associate editor for IEEE TSIPN and TNSE, and has leadership roles in multiple conferences, including WCSP, PCM, APSIPA ASC, and ACM MM. His awards include Best Paper at GLOBECOM 2013 and APSIPA ASC 2020, Best Student Papers at PCM 2017 and ICASSP 2016, and an Honor Mention at MMSP 2022.



University of Science and Technology of China, China

Dongheng Zhang

Dr. Zhang is currently a research associate professor at USTC. He received his Ph.D from UESTC. His research focuses on smart sensing technologies. He has authored 40+ papers in the field.



University of Science and Technology of China, China

Zhi Lu

Dr. Lu is a postdoctoral researcher at USTC. He received Ph.D from UESTC. His research focuses on smart sensing and multimedia.

[T04] Emerging Topics for Speech Synthesis: Versatility and Efficiency



The University of Tokyo, Japan

Yuki Saito

Yuki Saito received the Ph.D. degree in information science and technology from the Graduate School of Information Science and Technology, The University of Tokyo, Japan, in 2021. His research interests include speech synthesis, voice conversion, and machine learning. He is a member of the Acoustical Society of Japan, IEEE SPS, and the Institute of Electronics, Information and Communication Engineers. He was a recipient of eight paper awards, including the 2020 IEEE SPS Young Author Best Paper Award.



Keio University, Japan

Shinnosuke Takamichi

Shinnosuke Takamichi received the Ph.D. degree from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2016. He is currently an Associate Professor with Keio University, Japan. He was the recipient of more than ten paper/achievement awards, including the 3rd IEEE SPS Japan Young Author Best Paper Award.



The University of Tokyo, Japan

Wataru Nakata

Wataru Nakata received the B.E. degree from The University of Tokyo, Tokyo, Japan, in 2023. He is now a master's student at the University of Tokyo. His research interests include, speech synthesis, natural language processing and deep learning. He is a Student Member of the Acoustical Society of Japan. He recieved the best student presentation award of ASJ in 2022.

WOMEN'S FORUM SPEAKERS



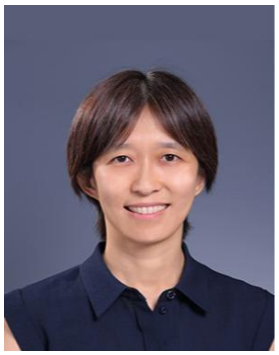
Professor Helen Meng

Helen Meng is Patrick Huen Wing Ming Chair Professor of Systems Engineering & Engineering Management at The Chinese University of Hong Kong. She joined CUHK in 1998 and later became the first female department head in the Faculty of Engineering. She has also served as Associate Dean of Research, and oversaw research across all engineering departments. Her collaborations include partnerships with Microsoft Research Asia (MSRA), MIT, Tsinghua University, and the Dr. Stanley Ho Medical Development Foundation. Her work with MSRA led to CUHK's first Ministry of Education Key Laboratory in engineering. In 2019, her interdisciplinary team was awarded the first AI project under the HKSARG RGC Theme-based Research Scheme, developing AI technologies for dementia screening. She later founded and directs CUHK's AI InnoCentre, the Centre for Perceptual and Interactive Intelligence. Helen led her department in launching Hong Kong's first undergraduate FinTech degree in 2017. She also spearheaded the development of the city's first pre-tertiary AI curriculum, with a focus on AI ethics, set for rollout in all local secondary schools in 2024 by the Education Bureau. The project earned the 2021 Hong Kong ICT Awards Smart Education and Learning Gold Award. Helen's professional service roles include Editor-in-Chief of IEEE Transactions on Audio, Speech, and Language Processing and member of the IEEE SPS Board of Governors, Awards Board, and Fellows Evaluation Committee. She also serves on the HKSAR Government's Digital Economy Development Committee, eHealth Record Sharing Committee, Research Grants Council, RGC Engineering Panel, and Research Assessment Exercise Group. Recent accolades of Helen's team include winning the US National Academy of Medicine's Healthy Longevity Global Grand Challenge 2024, Gold and Silver Medals at the 49th International Exhibition of Inventions Geneva 2024, Best Student Paper Awards at INTERSPEECH 2024 and 2023, and First Prize at the DialDoc@ACL 2022 Challenge. Helen herself has received the 2019 IEEE SPS Leo Beranek Meritorious Service Award for her leadership in and services for the Signal Processing Society, along with the HKCS 2015 Outstanding Women Professional Award, the 2017 HKWPEA Outstanding Women Professionals Award, and the 2022 JESSICA Most Successful Women Award. She is a Fellow of IEEE, ISCA, HKIE, and a Distinguished Fellow of HKCS.



Professor Sansanee Auephanwiriyakul

Sansanee Auephanwiriyakul received a B.Eng. (Hons.) degree in electrical engineering from Chiang Mai University, Thailand (1993); an M.S. degree in electrical and computer engineering; and a Ph.D. degree in computer engineering and computer science, both from the University of Missouri, Columbia, in 1996, and 2000, respectively. After receiving her Ph.D. degree, she worked as a post-doctoral fellow at the Computational Intelligence Laboratory, University of Missouri-Columbia. Dr. Auephanwiriyakul is an Associate Professor in the Department of Computer Engineering and an associate director of the Biomedical Engineering Institute, Chiang Mai University, Thailand. She is a senior member of the Institute of Electrical and Electronics Engineers (IEEE). She is an Associate Editor of the IEEE Transactions on Fuzzy Systems, IEEE Transactions on Neural Networks and Learning Systems, ECTI Transactions on Computer and Information Technology (ECTI-CIT), and was an Editorial Board of the Neural Computing and Applications, and other various major journals. She is a part of several IEEE CIS technical committees and several organizing committees of several major conferences. She has written many conference papers that have been published in several major conferences and journal papers published in several major journals.



Professor Hong (Vicky) Zhao

H. Vicky Zhao received her B.S. and M.S. degree from Tsinghua University, China, in 1997 and 1999, respectively, and her Ph. D degree from the University of Maryland, College Park, in 2004, all in electrical engineering. She was with the Department of Electrical and Computer Engineering at the University of Alberta, Canada, from 2006 to 2016. She is currently an Associate Professor in the Department of Automation at Tsinghua University, Beijing, China. Her research interests include social network analysis, information security and forensics, digital signal processing and communications. Dr. Zhao received the IEEE Signal Processing Society (SPS) 2008 Young Author Best Paper Award and the 2020 APSIPA Annual Summit and Conference Best Paper Award. She has co-authored 3 research monographs. She serves as a Member-at-Large of the IEEE Signal Processing Society Board of Governors (2022-2024), and a Member-at-Large of the APSIPA Board of Governors (2024). She is an APSIPA Distinguished Lecturer (2023-2024).

1. Overview Session 1 Meeting Room 1 04-Dec

1-1 16:20-16:40
A Decade of Progress in Sound Event Localization and Detection: Transforming Environmental Sound Analysis for Real-World Impact
Woon-Seng Gan, Nanyang Technological University

1-2 16:40-17:00
Exploring the Forward-Forward Algorithm: A Novel Learning Approach
Waleed H. Abdulla, The University of Auckland

1-3 17:00-17:20
Eye-gaze-based Human-Intention Detection
Kosin Chamnongthai, King Mongkut's University of Technology Thonburi

1-4 17:20-17:40
From GPT Evolution to Enterprise Deployment: Key Trends in Generative AI
Jing-Ming Guo, National Taiwan University of Science and Technology

1-5 17:40-18:00
An Overview of Online Distributed Kernel Methods for Supervised and Unsupervised Learning
Anthony Kuh, University of Hawaii

2. Overview Session 2 Meeting Room 8 05-Dec

2-1 10:20-10:40
An AI-based Diagnostic-aid for Epileptic Electroencephalography
Toshihisa Tanaka, Tokyo University of Agriculture and Technology

2-2 10:40-11:00
Machine Learning for Analytics Architecture: AI to Design AI Video
Chris Gwo Giun Lee, National Cheng Kung University

2-3 11:00-11:20
Compression of Large AI Models
Weisi Lin, Nanyang Technological University

2-4 11:20-11:40
Introduction to Multi-Camera Systems and 3D Quality Assessment
Sanghoon Lee, Yonsei University

2-5 11:40-12:00
Highlight of New Image Generative Models and Applications to Image Manipulations
Wan-Chi Siu, Hong Kong Polytechnic University & St. Francis University

3. Overview Session 3 Merged Room (Room 10 + 11) 06-Dec

3-1 9:00-9:20
Overview of Source Camera Identification Techniques
Bonnie N. F. Law, The Hong Kong Polytechnic University

3-2 9:20-9:40
Recent Advances in Complete Quality Preserving Data Hiding
KokSheik Wong, Monash University Malaysia

3-3 9:40-10:00
Real or Fake? Frontiers of Countering Fake Media in the Age of Infodemics

Isao Echizen, National Institute of Informatics

3-4 10:00-10:20
User Preference Modeling and Analysis in Choice Problems
H. Vicky Zhao, Tsinghua University

4. Machine Learning and Data Analytics Meeting Room 1 04-Dec

4-1 11:00-11:20
Psychological Driving Style Estimation from GPS Sensor Data Alone
Hiroto Horimoto (Japan Advanced Institute of Science and Technology); Ryusei Kimura (Japan Advanced Institute of Science and Technology); Takahiro Tanaka (Nagoya University); Shogo Okada (Japan Advanced Institute of Science and Technology)*

4-2 11:20-11:40
Adversarial Augmentation and Adaptation for Speech Recognition
Jen-Tzung Chien (National Yang Ming Chiao Tung University)*; Wei-Yu Sun (National Yang Ming Chiao Tung University)

4-3 11:40-12:00
Empathetic Response Generation via Regularized Q-Learning
Jen-Tzung Chien (National Yang Ming Chiao Tung University)*; Yi-Chien Wu (National Yang Ming Chiao Tung University)

4-4 12:00-12:20
Continual Learning with Self-Organizing Maps: A Novel Group-Based Unsupervised Sequential Training Approach
Gaurav Hirani (The University of Auckland)*; Kevin I-Kai Wang (The University of Auckland); Waleed Abdulla ("University of Auckland, New Zealand")

5. Machine Learning and Data Analytics Meeting Room 2 04-Dec

5-1 11:00-11:20
YOLO for High Resolution Images without Retraining
Daisuke Minami (Tokyo metropolitan university)*; Kiyoshi Nishikawa (Tokyo Metropolitan University)

5-2 11:20-11:40
Noise-Robust Estimation of Early-part Room Impulse Responses based on Physics-Informed Neural Network with Dynamic Pulling Method
Ken Kurata (Tokyo Denki University)*; Gen Sato (Tokyo Denki University); Izumi Tsunokuni (Tokyo Denki University); Yusuke Ikeda (Tokyo Denki University)

5-3 11:40-12:00
A Multi-Domain Camera Model Identification Feature Restoration Network to Counter AI Compression Attacks
zhang jinkai (Shandong University, Jinan, China)*

5-4 12:00-12:20
Deep Learning-based Intraoperative Video Analysis for Cataract Surgery Instrument Identification
Zhe Guo (The Hong Kong Polytechnic University)*; Yuk Hee Chan (The Hong Kong Polytechnic University); Ngai Fong Law

6. Image, Video, and Multimedia Meeting Room 3 04-Dec

6-1 11:00-11:20
GSBIQA: Green Saliency-guided Blind Image Quality Assessment Method
Zhanxuan Mei (University of Southern California)*; Yun-Cheng Wang (University of Southern California); C.-C. Jay Kuo (USC)

6-2 11:20-11:40
AFSDet: Video Small Object Detection Based on Adaptive Focused Slicing
Kangjian Huang (Southwest Jiaotong University); Yan Yang (Southwest Jiaotong University)*; Yongquan Jiang (Southwest Jiaotong University); Xiaobo Zhang (School of Computing and Artificial Intelligence, Southwest Jiaotong University); Zhuyi Angelina Li (Renmin University of China)

6-3 11:40-12:00
Dual Motion Attention and Enhanced Knowledge Distillation for Video Frame Interpolation
Deng Zhang (Changsha University of Science and Technology)*; runqi lou (changsha university of science and technology); Jiaxin Chen (School of Computer and Communication Engineering, Changsha University of Science and Technology); Xin Liao (College of Computer Science and Electronic Engineering, Hunan University); Gaobo Yang (Hunan University); xiangling ding (Hunan University of Science and Technology)

6-4 12:00-12:20
EavaNet: Enhancing Emotional Facial Expressions in 3D Avatars through Speech-Driven Animation
seyun um (yonsei university)*; YongJu Lee (Yonsei University); WooSeok Ko (Yonsei University); Yuan Zhou (01.ai); Sangyoun Lee (Yonsei University); Hong-Goo Kang (Yonsei University)

7. Signal and Information Processing & Systems Meeting Room 4 04-Dec

7-1 11:00-11:20
On the Importance of Time and Pitch Relativity for Transformer-based Symbolic Music Generation
Tatsuro Inaba (Kyoto University)*; Kazuyoshi Yoshii (Kyoto University); Eita Nakamura (Kyushu University)

7-2 11:20-11:40
Optimal Investment With Incomplete Information and Herd Effect
Huisheng Wang (Tsinghua University); Mingxiao Liu (Tsinghua University); Ji Qi (Tsinghua University); H. Vicky Zhao (Tsinghua University)*

7-3 11:40-12:00
YOLO-DC: Enhancing object detection with deformable convolutions and contextual mechanism
Deng Zhang (Changsha University of Science and Technology)*; Chuanzhen Xu (Changsha University of Science and Technology); Jiaxin Chen (School of Computer and Communication Engineering, Changsha University of Science and Technology); Xin Liao (College of Computer Science and Electronic Engineering, Hunan University)

7-4 12:00-12:20
One-step Spectral Estimation for Euclidean Distance Matrix Approximation
Yicheng Li (Sun Yat-sen University)*; Xinghua Sun (Sun Yat-sen University)

8. Speech and Language Processing Meeting Room 5 04-Dec

8-1 11:00-11:20
SDNet: Noise-Robust Bandwidth Extension under Flexible Sampling Rates
Junkang Yang (Chongqing University of Posts and Telecommunications)*; Hongqing Liu (Chongqing University of Posts and Telecommunications); Lu Gan (Brunel University London); Yi Zhou (CQUPT); Xing Li (vivo Mobile Commun co Ltd); Jie Jia (vivo Mobile Commun co Ltd); Jinzhuo Yao (Chongqing University of Posts and Telecommunications)

8-2 11:20-11:40
GLASS: Investigating Global and Local context Awareness in Speech Separation
Kuan-Hsun Ho (NTNU)*; En-Lun Yu (National Taiwan Normal University); Jeih-weih Hung (National Chi Nan University); Shih-Chieh Huang (Realtek Semiconductor Corp.); Berlin Chen (National Taiwan Normal University)

8-3 11:40-12:00
Low-resource Language Adaptation with Ensemble of PEFT Approaches
Chin Yuen Kwok (Nanyang Technological University)*; Sheng Li (National Institute of Information & Communications Technology (NICT)); Jia Qi Yip (Alibaba Group / Nanyang Technological University); Eng Siong Chng (Nanyang Technological University)

8-4 12:00-12:20
Diverse Time-Frequency Attention Neural Network for Acoustic Echo Cancellation
Jinzhuo Yao (Chongqing University of Posts and Telecommunications)*; Hongqing Liu (Chongqing University of Posts and Telecommunications); Yi Zhou (CQUPT); Lu Gan (Brunel University London); Junkang Yang (Chongqing University of Posts and Telecommunications)

9. Audio Processing Meeting Room 6 04-Dec

9-1 11:00-11:20
Scale-invariant Online Voice Activity Detection under Various Environments
Ryu Takeda (Osaka University)*; Kazunori Komatani (Osaka University)

9-2 11:20-11:40
Sound Quality Improvement in Visual Microphone by Emphasizing Focused Area Based on Focal Rate
Hayata Nakano (Ritsumeikan University)*; Yuting Geng (Ritsumeikan University); Kenta Iwai (Ritsumeikan University); Takanobu Nishiura (Ritsumeikan University)

9-3 11:40-12:00
Deep-Learning-Based Speech Enhancement with Rough-Focused Optical Laser Microphone by Reconstructing Complex Spectrum
Yuki Nakano (Ritsumeikan University)*; Yuting Geng (Ritsumeikan University); Kenta Iwai (Ritsumeikan University); Takanobu Nishiura (Ritsumeikan University)

9-4 12:00-12:20
A Study on Multimodal Fusion and Layer Adapter in Emotion Recognition
Xiaohan Shi (Nagoya University)*; Yuan Gao (Kyoto University); Jiajun He (Nagoya University); Jinyi Mi (Nagoya University); Xingfeng LI (City University of Macau); Tomoki Toda (Nagoya University)

10. Biomedical Signal Processing and Systems Meeting Room 8 04-Dec

10-1 11:00-11:20
Bluemarble: Bridging Latent Uncertainty in Articulatory-to-Speech Synthesis with a Learned Codebook
seyun um (yonsei university)*; Miseul Kim (Yonsei University); Doyeon Kim (Yonsei University); Hong-Goo Kang (Yonsei University)

10-2 11:20-11:40
Iterative Demographic Attentional Feature Fusion-based CNN and Transformer Network for Accurate Cuffless Blood Pressure Estimation
Liwen Tang (Southern University of Science and Technology); Dingchang Zheng (Coventry University); Fei Chen (Southern University of Science and Technology)*

10-3 11:40-12:00
Sampling Pattern Augmentation to Enhance Deep Learning-based Image Reconstruction of MRI
Kazuki Yamato (Utsunomiya University)*; Satoshi Ito (Utsunomiya University)

10-4 12:00-12:20
Data Augmentation and Assessment for Enhanced Ovarian Tumor Classification
Loan Pham (Hai Duong University, Hanoi University of Science and Technology)*; Gia-Minh Pham (Hanoi University of Science and Technology); Tien-Dat Nguyen (Hanoi University of Science and Technology); Hung Le (International Research Institute MICA, HUST - CNRS/UMI-2954 -GRENOBLE INP, Vietnam); Chi-Mai Pham (National Hospital of Obstetrics and Gynecology); Thi Lan Le (HUST); Duy-Hai Vu (Hanoi University of Science and Technology); Hai Vu (Hanoi University of Science and Technology); Thanh-Hai Tran (Hanoi University of Science and Technology)

11. Machine Learning and Data Analytics Meeting Room 9 04-Dec

11-1 11:00-11:20
GMA: Green Multi-Modal Alignment for Image-Text Retrieval
Tsung-Shan Yang (University of Southern California)*; Yun-Cheng Wang (University of Southern California); Chengwei Wei (University of Southern California); Suya You (University of Southern California); C.-C. Jay Kuo (University of Southern California)

11-2 11:20-11:40
Improving Semi-Supervised Object Detection by ROI-Enhanced Contrastive Learning
Teng-Kuan Huang Huang (National Taiwan Normal University); Mei-Chen Yeh (National Taiwan Normal University)*

11-3 11:40-12:00
Real-time Segmentation of Coronary Artery Calcification Using Spatial Attention and Parallel Convolution
Tetsuya Asakawa (Toyohashi University of Technology)*; Masashi Hashimoto (Toyohashi Heart center); Takeshi Miyaji (Toyohashi Heart center); kazuki shimizu (Toyohashi Heart center); Kei Nomura (Toyohashi Heart center); Masaki Aono (Toyohashi University of Technology)

12. Speech and Language Processing Meeting Room 10 04-Dec

12-1 11:00-11:20
LDMSE: Low Computational Cost Generative Diffusion Model for Speech Enhancement
Yuki Nishi (Tokyo Insutitute of Technology)*; Koji Iwano (Tokyo City University); Koichi SHINODA (Tokyo Institute of Technology)

12-2 11:20-11:40
MTFNet: Multi-Scale Transformer Framework for Robust Emotion Monitoring in Group Learning Settings
Yi Zhang (Beijing Normal University,ZhuHai)*

12-3 11:40-12:00
Target Speaker Extraction Method by Emphasizing the Active Speech with an Additional Enhancer
Xue Yang (Beijing University of Technology); Changchun Bao (Beijing University of Technology)*; Xu Zhang (Beijing University of Technology); Xianhong Chen (Beijing University of Technology)

13. Multimedia Security and Forensics Meeting Room 2 04-Dec

13-1 14:00-14:20
Compressed Deepfake Video Detection Based on 3D Spatiotemporal Trajectories
Zongmei Chen (Hunan University); Xin Liao (College of Computer Science and Electronic Engineering, Hunan University)*; Xiaoshuai Wu (College of Computer Science and Electronic Engineering, Hunan University); Yanxiang Chen (Hefei University of Technology)

13-2 14:20-14:40
A Document Presentation Attack Detection Scheme with Optical Flow under a Flashlight
Changsheng Chen (Shenzhen University)*; Wenyu Chen (Shenzhen University); Ximin Chen (Shenzhen University); Haodong Li (Shenzhen University)

13-3 14:40-15:00
Robust Image Watermarking Scheme under Halftone Distortion with Surrogate Model
Changsheng Chen (Shenzhen University)*; Xijin Li (Shenzhen University)

13-4 15:00-15:20
A Diffusion-Based Approach for Restoring Face-swapped Images
Yuanchen Niu (Shenzhen University); Yuanman Li (Shenzhen University)*; Guijia Zhang (Shenzhen university); Xia Li (Shenzhen University)

13-5 15:20-15:40
AI-generated image detectors are surprisingly easy to mislead... for now
Zihang Lyu (The Hong Kong Polytechnic University)*; Jun Xiao (The Hong Kong Polytechnic University); Cong Zhang (The Hong Kong Polytechnic University); Kin-Man Lam (The Hong Kong Polytechnic University)

14. Image, Video, and Multimedia Meeting Room 3 04-Dec

14-1 14:00-14:20
Green Video Camouflaged Object Detection
Xinyu Wang (University of Southern California)*; Hong-Shuo Chen (USC); Zhiruo Zhou (University of Southern California); Suya You (University of Southern California); Azad Madni (University of Southern California); C.-C. Jay Kuo (University of Southern California)

14-2 14:20-14:40
A Survey on Objective Quality Assessment of Omnidirectional Images
Xiangjie Sui (City University of Macau)*; Shiqi Wang (City University of Hong Kong); Yuming Fang (Jiangxi University of Finance and Economics)

14-3 14:40-15:00
Enhancing YOLOv7 with GLF-Trans for Precision in Small Object Detection
Naohito Yoshikawa (Electrical and Information Engineering (of Keio University))*; Masaaki Ikehara (Keio university)

14-4 15:00-15:20
Ablation Study to Derive a Computationally Efficient Deep Learning-Based Super-Resolution Approach
Asfa Jamil (CYENS - Center of Excellence)*; Alessandro Artusi (DeepCamera/CYENS)

14-5 15:20-15:40
Adaptive Spatial Re-sampling Method for Video Coding for Machines
Eunbin An (Yonsei Univ.); Ayoung Kim (Yonsei University); Soon Heung Jung (ETRI); Hyon-Gon Choo (Electronics and Telecommunications Research Institute (ETRI)); Kwang-Deok Seo (Yonsei Univ.)*

15. Signal and Information Processing & Systems Meeting Room 4 04-Dec

15-1 14:00-14:20
Multi-Channel Fusion Human Activity Recognition Algorithm Based on Millimeter-Wave Radar
Junda Zhu (University of Electronic Science and Technology of China)*; Shisheng Guo (University of Electronic Science and Technology of China); Longzhen Tang (University of Electronic Science and Technology of China); Cui Guolong (UESTC)

15-2 14:20-14:40
Optimizing Computational Efficiency: In-Memory Computing with Dynamic Switching
Chao-Ting Huang (Tunghai University)*; Kun-Lin Tsai (Tunghai University)

15-3 14:40-15:00
Modeling and Analysis of the Interaction between Opinions and Actions among Heterogeneous Agents
Hangjing Zhang (Tsinghua University); H. Vicky Zhao (Tsinghua University)*

15-4 15:00-15:20
Adaptive Subspace Clustering for Matrix Completion
Takuto Wada (Hosei University)*; Ryohei Sasaki (Tokyo university of technology); Katsumi Konishi (Hosei University)

15-5 15:20-15:40
A High-Isolation Sub-6 GHz In-Band Full-Duplex Communication System
chengzhe shi (University of Electronic Science and Technology of China)*; Wensheng Pan (University of Electronic Science and Technology of China); Wanzhi Ma (University of Electronic Science and Technology of China); Ying Liu (University of Electronic Science and Technology of China); Qiang Xu (University of Electronic Science and Technology of China); Zhiya Zhang (Xidian University); Shihai Shao (University of Electronic Science and Technology of China)

16. Speech and Language Processing Meeting Room 5 04-Dec

16-1 14:00-14:20
GE2E-AC: Generalized End-to-End Loss Training for Accent Classification
Chihiro Watanabe (NTT Communication Science Laboratories)*; Hirokazu Kameoka (NTT Communication Science Laboratories)

16-2 14:20-14:40
Efficient Feature Selection for Word Embedding Dimension Reduction
Jintang Xue (University of Southern California)*; Yun-Cheng Wang (University of Southern California); Chengwei Wei (University of Southern California); C.-C. Jay Kuo (University of Southern California)

16-3 14:40-15:00
Improving Speaker Consistency in Speech-to-Speech Translation Using Speaker Retention Unit-to-Mel Techniques
Rui Zhou (Graduate School of Engineering, Tohoku University)*

16-4 15:00-15:20
Speech Separation using Neural Audio Codecs with Embedding Loss
Jia Qi Yip (Alibaba Group / Nanyang Technological University)*; Chin Yuen Kwok (Nanyang Technological University); Bin Ma (Alibaba); Eng Siong Chng (Nanyang Technological University)

16-5 15:20-15:40
Speech Synthesis from IPA Sequences through EMA Data
Koki Maruyama (Tokyo University of Science)*; Shun Sawada (Tokyo University of Science); Hidefumi Ohmura

(Tokyo University of Science); Kouichi Katsurada (Tokyo University of Science)

17. Audio Processing Meeting Room 6 04-Dec

17-1 14:00-14:20
Learnable Cross-Correlation based Filter-and-Sum Networks for Multi-channel Speech Separation
Xianrui Wang (Northwestern Polytechnical University)*; Shiqi Zhang (Audio Research Group, Tampere University); Bo He (Graduate School of Information, Production and Systems, Waseda University); Shoji Makino (Graduate School of Information, Production and Systems, Waseda University); Jingdong Chen (Northwestern Polytechnical University)

17-2 14:20-14:40
Enhancing Neural Speech Embeddings for Generative Speech Models
Doyeon Kim (Yonsei University)*; Yanjue Song (Ghent University - imec); Nilesh Madhu (IDLab, Ghent University - imec); Hong-Goo Kang (Yonsei University)

17-3 14:40-15:00
Design of Spectrogram-Consistency Regularization Term Dependent on Observation in Independent Low-Rank Matrix Analysis for Blind Source Separation
Takaaki Kojima (The University of Tokyo)*; Norihiro Takamune (the University of Tokyo); Daichi Kitamura (National Institute of Technology, Kagawa College); Hiroshi Saruwatari (The University of Tokyo)

17-4 15:00-15:20
On Joint Dereverberation and Single Moving Source Separation with Online Source Steering
Yiting Zhang (Waseda University)*; Kaien Mo (Waseda University); Tetsuya Ueda (Waseda University); Yichen Yang (Waseda University); Shoji Makino (Waseda University)

17-5 15:20-15:40
New Perspectives and Insights on Distortionless Microphone Array Beamforming
Fan Zhang (CIAIC; Northwestern Polytechnical University)*; Jacob Benesty (INRS); Chao Pan (CIAIC; Northwestern Polytechnical University); Jingdong Chen (Northwestern Polytechnical University)

17-6 15:40-16:00
Block Refinement Learning for Improving Early Exit in Autoregressive ASR
Naotaka Kawata (NTT)*; Shota Orihashi (NTT Corporation); Satoshi Suzuki (NTT Computer and Data Science Laboratories / The University of Electro-Communications); Tomohiro Tanaka (NTT); Mana Ihori (NTT); Naoki Makishima (NTT); Taiga Yamane (NTT Corporation); Ryo Masumura (NTT Corporation)

18. Biomedical Signal Processing and Systems Meeting Room 8 04-Dec

18-1 14:00-14:20
Postoperative Delirium Prediction Based on Preoperative Electrocardiogram and Electroencephalogram
Shogo Mito (Kyoto University); Miho Miyajima (Tokyo Medical and Dental University); Hirofumi Tomioka (Tokyo Medical and Dental University); Hitomi Sato (Tokyo Medical and Dental University); Takashi Takeuchi (Tokyo Medical and Dental University); Hitoshi Muto (Tokyo Medical and Dental University); Yuji Kabasawa (Tokyo Medical and Dental University); Hiroyuki Harada (Tokyo Medical and Dental University); Kana Eguchi (Kyoto University); Shota Kato (Kyoto University); Manabu Kano (Kyoto University)*

18-2 14:20-14:40
A method for classification NEO–FFI answers fabricated and advantageous due to psychological bias using brainwave specific brain activity networks
YUTO ASHIKAWA (Aoyama Gakuin University)*; Takashi Ito (Aoyama Gakuin University); Syohei Ishizu (Aoyama Gakuin University); Yosuke Kurihara (Aoyama Gakuin University)

18-3 14:40-15:00
Effect of White Noise on Working Memory Using Event-Related Potentials
Seung-won Lee (Korea university); Jun-Seok LEE (Korea University); Han-Jeong Hwang (Korea University)*

18-4 15:00-15:20
Automated prediction of loudness growth curve using EEG signals
Nitya Tiwari (Indian Institute of Technology Bhubaneswar)*

18-5 15:20-15:40
Separation of Cardiopulmonary Sound Signals for Classification of Respiratory Diseases
Ruxin Zheng (Ruxin Zheng)*

18-6 15:40-16:00
Performance Improvement of Single Plane-Wave Imaging Using U-Net and Discrete Wavelet Transform
Hiromi Shidara (Tohoku University)*; Kanta Miura (Tohoku University); Takuro Ishii (Tohoku University); Koichi Ito (Tohoku University); Takafumi Aoki (Tohoku University); Yoshifumi Saijo (Graduate School of Biomedical Engineering, Tohoku University); Jun Ohmiya (KONICA MINOLTA, INC.)

19. Best Student Paper Competition Meeting Room 9 04-Dec

19-1 14:00-14:20
Rotation Invariant Spatio-Spectral Total Variation for Hyperspectral Image Denoising
Shingo Takemoto (Institute of Science Tokyo)*; Shunsuke Ono (Institute of Science Tokyo)

19-2 14:20-14:40
Peer Learning via Shared Speech Representation Prediction for Target Speech Separation
Xusheng Yang (Peking University)*; Zifeng Zhao (Peking University); Yuexian Zou (Peking University)

19-3 14:40-15:00
NecoBERT: Self-Supervised Learning Model Trained by Masked Language Modeling on Rich Acoustic Features Derived from Neural Audio Codec
Wataru Nakata (The University of Tokyo)*; Takaaki Saeki (The University of Tokyo); Yuki Saito ("The University of Tokyo, Japan"); Shinnosuke Takamichi (Keio University); Hiroshi Saruwatari (The University of Tokyo)

19-4 15:00-15:20
Multi-Task Learning Approaches for Music Similarity Representation Learning Based on Individual Instrument Sounds
Takehiro Imamura (Nagoya University)*; Yuka Hashizume (Nagoya University); Tomoki Toda (Nagoya University)

19-5 15:20-15:40
ViP-CBM: Reducing Parameters in Concept Bottleneck Models by Visual-Projected Embeddings
Ji Qi (Tsinghua University); Huisheng Wang (Tsinghua University); H. Vicky Zhao (Tsinghua University)*

20. Speech and Language Processing Meeting Room 10 04-Dec

20-1 14:00-14:20
BEES: A New Acoustic Task for Blended Emotion Estimation in Speech
Xingfeng LI (City University of Macau)*; Xiaohan Shi (Nagoya University); Yuke Si (China Mobile Research Institute); Zilong Zhang (Hainan University); Feifei Cui (Hainan University); Yongwei Li (Chinese Academy of Sciences); Yang Liu (Qingdao University of Science and Technology); Masashi Unoki (JAIST); Masato Akagi (Japan Advanced Institute of Science and Technology)

20-2 14:20-14:40
Is Corpus Truth for Human Perception?: Quality Assessment of Voice Response Timing in Conversational Corpus through Timing Replacement
Sadahiro Yoshikawa (Japan Advanced Institute of Science and Technology)*; Ryo Ishii (NTT); Shogo Okada (Japan Advanced Institute of Science and Technology)

20-3 14:40-15:00
Enhancing Branchformer with Dynamic Branch Merging Module for Code-Switching Speech Recognition
Hong-Jie Hu (NSYSU)*; Chia-Ping Chen (National Sun Yat-sen University)

20-4 15:00-15:20
Optimizing Multi-Speaker Speech Recognition with Online Decoding and Data Augmentation Strategies
Yizhou Peng (Nanyang Technological University)*; Eng Siong Chng (Nanyang Technological University)

20-5 15:20-15:40
Adapting OpenAI's Whisper for Speech Recognition on Code-Switch Mandarin-English SEAME and ASRU2019 Datasets
Yuhang Yang (College of Computer Science and Electronic Engineering, Hunan University, China); Yizhou Peng (Nanyang Technological University)*; Hao Huang (Xinjiang University); Eng Siong Chng (Nanyang Technological University); Xionghu Zhong (Hunan University)

21. Multimedia Security and Forensics Meeting Room 2 04-Dec

21-1 16:20-16:40
A Study on Variable Embedding Locations of Reversible Spectral Speech Watermarking
Xuping HUANG (Tohoku University/Shimane University)*; Akinori Ito (Tohoku University)

21-2 16:40-17:00
Normalizing Flows-Based Latent Variable Rearrangement for Generative Image Steganography
Sifan Wu (Ningbo university)*; Li Dong (Ningbo University)

21-3 17:00-17:20
Detecting Spoof Voices in Asian Non-Native Speech: An Indonesian and Thai Case Study
Aulia Adila (Japan Advanced Institute of Science and Technology)*; Candy Olivia Mawalim (Japan Advanced Institute of Science and Technology); Masashi Unoki (JAIST)

21-4 17:20-17:40
Privacy-Preserving Anomaly Detection in Bitstream Video based on Gaussian Mixture Model
Yike Chen (Sun Yat-sen University); Yuru Song (Sun Yat-sen University); Peijia Zheng (Sun Yat-sen University)*; Yusong Du (School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, P.R. China); Weiqi Luo (Sun Yat-sen University)

21-5 17:40-18:00
Source Attribution for Images Generated by Diffusion-Based Text-to-Image Models: Exploring the Forensics Approach
Xinqi Jiang (Macau University of Science and Technology); Jinyu Tian (Macau University of Science and Technology)*

22. Image, Video, and Multimedia Meeting Room 3 04-Dec

22-1 16:20-16:40
Hyperspectral Unmixing With Row-Sparsity Enhancement: A Difference-of-Convex Approach

Kazuki Naganuma (Institute of Science Tokyo)*; Shunsuke Ono (Institute of Science Tokyo)

22-2 16:40-17:00
How Accurate Can Large Vision Language Model Perform for Images with Compression Degradation?
Xiaohan Fang (City University of Hong Kong)*; PEILIN CHEN (City University of Hong Kong); Meng Wang (Lingnan University); Shiqi Wang (City University of Hong Kong)

22-3 17:00-17:20
Enhanced RefineDNet for Single Image Dehazing
Jingyu Ren (Macau University of Science and Technology)*

22-4 17:20-17:40
Tsnake: A Time-Embedded Recurrent Contour-Based Instance Segmentation Model
Chen-Jui Hsu (National Taiwan University); Jian-Jiun Ding (National Taiwan University)*; Chun-Jen Shih (National Taiwan University)

22-5 17:40-18:00
A Multi-Perceptual Learning Network for Retina OCT Image Denoising and Classification
Kin-Man Lam (The Hong Kong Polytechnic University)*

23. Signal and Information Processing & Systems Meeting Room 4 04-Dec

23-1 16:20-16:40
Affine Combination of General Adaptive Filters
Danqi Jin (Wuhan University)*; Yitong Chen (Northwestern Polytechnical University); Jie Chen (Northwestern Polytechnical University); Gongping Huang (Wuhan University)

23-2 16:40-17:00
An Annealing-Inspired Gradient-Descent Based Suboptimal Solver for Combinatorial Problems
Chang Shu Ping (National Taiwan University); Cheng-Che Lee (National Taiwan University); Hsin-Jung Lee (National Taiwan University); Chieh-Hsiung Kuan (National Taiwan University); Jason Gemsun Young (Industrial Technology Research Institute); Chia-Yu Yao (National Taiwan University of Science and Technology); Jian-Jiun Ding (National Taiwan University)*

23-3 17:00-17:20
A Solution For Anomaly Detection of Red Beans In A Product Processing Line
Duc Hai Nguyen (Phenikaa University); Hiep Do (Phenikaa University); Hoang-Linh-Phuong Nguyen (Phenikaa University); Quoc-Khanh Nguyen (Phenikaa University); Duc-Tan Tran (Phenikaa University); Tien Son Bui (Hanoi University of Industry); VanToi Nguyen (Phenikaa University)*

23-4 17:20-17:40
A Novel kind of WVD Associated with the Linear Canonical Transform
Jia-Yin Peng (Beijing Institute of Technology); Jian-Yi Chen (Beijing Institute of Technology); Bing-Zhao Li (Beijing Institute of Technology)*

23-5 17:40-18:00
A Discrete-Valued Signal Estimation by Nonconvex Enhancement of SOAV with cLiGME Model
Satoshi Shoji (Institute of Science Tokyo)*; Wataru Yata (Institute of Science Tokyo); Keita Kume (Institute of Science Tokyo); Isao Yamada (Institute of Science Tokyo)

24. Speech and Language Processing Meeting Room 5 04-Dec

24-1 16:20-16:40
Frequency & Channel Attention Network for Small Footprint Noisy Spoken Keyword Spotting
Yuanxi Lin (Bauman Moscow State Technical University)*; Yuriy Gapanyuk (Bauman Moscow State Technical University)

24-2 16:40-17:00
Long Audio File Speaker Diarization with Feasible End-to-End Models
Kai-Wei Huang (National Sun Yat-sen University)*; Chia-Ping Chen (National Sun Yat-sen University)

24-3 17:00-17:20
Analysis of Various Self-Supervised Learning Models for Automatic Pronunciation Assessment
Haeyoung Lee (Seoul National University)*; Sunhee Kim (Seoul National University); Minhwa Chung (Seoul National University)

24-4 17:20-17:40
Band-Split Inter-SubNet: Band-Split with Subband Interaction for Monaural Speech Enhancement
Yen-Chou Pan (National Yang Ming Chiao Tung University); Yih-Liang Shen (National Yang Ming Chiao Tung University)*; Yuan-Fu Liao (National Yang Ming Chiao Tung University); Tai-Shih Chi (National Yang Ming Chiao Tung University)

24-5 17:40-18:00
Run-Time Adaptation of Neural Beamforming for Robust Speech Dereverberation and Denoising
Yoto Fujita (Kyoto University)*; Aditya Arie Nugraha (RIKEN); Diego Di Carlo (RIKEN); Yoshiaki Bando (National Institute of Advanced Industrial Science and Technology); Mathieu Fontaine (Télécom Paris); Kazuyoshi Yoshii (Kyoto University)

25. Audio Processing Meeting Room 6 04-Dec

25-1 16:20-16:40
A Low-Complexity Adaptive Beamformer for Joint Reverberation and Noise Suppression
Fan Zhang (CIAIC; Northwestern Polytechnical University)*; Chao Pan (CIAIC; Northwestern Polytechnical University); Jingdong Chen (Northwestern Polytechnical University); Jacob Benesty (INRS)

25-2 16:40-17:00
Multichannel Speech Enhancement Using Complex-Valued Graph Convolutional Networks and Triple-Path Attentive Recurrent Networks
Xingyu Shen (Concordia University); Prof. Wei-Ping Zhu (Concordia University)*

25-3 17:00-17:20
Anomalous Machine Sound Detection Based on Time Domain Gammatone Spectrogram Feature and IDNN Model
Primanda Adyatma Hafiz (Bandung Institute of Technology)*; Candy Olivia Mawalim (Japan Advanced Institute of Science and Technology); Dessi Puji Lestari (Bandung Institute of Technology); Sakriani Sakti (Nara Institute of Science and Technology / Japan Advanced Institute of Science and Technology); Masashi Unoki (JAIST)

25-4 17:20-17:40
Annotation-free Fine-tuning for Unsupervised Anomalous Sound Detection
Kai Guo (Beijing Institute of Technology)*; Xiang Xie (Beijing Institute of Technology); Fengrun Zhang (Beijing Institute of Technology)

25-5 17:40-18:00
Real-Time Monophonic Dual-Pitch Extraction Model
Ngoc-Son Tran (National Yang Ming Chiao Tung University); Pei-Chin Hsieh (National Yang Ming Chiao

Tung University); Yih-Liang Shen (National Yang Ming Chiao Tung University)*; Yen-Hsun Chu (Realtek Semiconductor Corporation); Tai-Shih Chi (National Yang Ming Chiao Tung University)

26. Biomedical Signal Processing and Systems Meeting Room 8 04-Dec

26-1 16:20-16:40
Predictive Analysis of Driver Drowsiness Progression: Multi-Level Drowsiness Classification Using Physiological Signals
Natchira Dachoponchai (Mahidol University); Yodchanan Wongsawat (Mahidol University); Jetsada Arnin (Mahidol University)*

26-2 16:40-17:00
Feature Extraction for Machine Learning-based Sleep Stage Classification Using PPG-Derived Parameters and Skin Temperature
Suphachok Buaruk (Sirindhorn International Institute of Technology, Thammasat University); Sikawat Thanaviratananich (King Chulalongkorn Memorial Hospital); Peerasit Treesuthacheep (King Chulalongkorn Memorial Hospital); Somrudee Deepaisarn (Sirindhorn International Institute of Technology, Thammasat University)*

26-3 17:00-17:20
Parameterizing Hierarchical Particle Filters with Concept Drift for Time-varying Parameter Estimation
Joshua Murphy (University of Liverpool)*; Conor Rosato (University of Liverpool); Andrew Millard (University of Liverpool); Simon Maskell (University of Liverpool)

26-4 17:20-17:40
Pop Noise Detection Using Group Delay Cepstral Coefficients
Arth Shah (DA-IICT)*; Hemant Patil (DAICT, Gujrat)

26-5 17:40-18:00
Novel Estimators for the Number of Susceptible Individuals in SIR Models of Infectious Epidemics
Anton van Wyk (University of the Witwatersrand); Andre McDonald (Council for Scientific and Industrial Research)*; David Rubin (University of the Witwatersrand); FangFang Zhang (Qilu University of Technology)

27. Best Paper Competition Meeting Room 9 04-Dec

27-1 16:20-16:40
Generalized Graph Signal Sampling under Subspace Priors by Difference-of-Convex Minimization
Keitaro Yamashita (Institute of Science Tokyo)*; Kazuki Naganuma (Institute of Science Tokyo); Shunsuke Ono (Institute of Science Tokyo)

27-2 16:40-17:00
Robust Adaptive Filtering Based on Adaptive Projected Subgradient Method: Moreau Enhancement of Distance Function
Daiki Sawada (Keio University); Masahiro Yukawa (Keio University)*

27-3 17:00-17:20
Fine-Grained Quantitative Emotion Editing for Speech Generation
Sho Inoue (Chinese University of Hong Kong, Shenzhen)*; Kun Zhou (Alibaba Group); Shuai Wang (Shenzhen Research Institute of Big Data, Chinese University of Hong Kong (Shenzhen)); Haizhou Li (The Chinese University of Hong Kong, Shenzhen)

27-4 17:20-17:40
Physical Domain Adversarial Attacks Against Source Printer Image Attribution

Nischay Purnekar (University of Siena)*; Benedetta Tondi (University of Siena); Mauro Barni (University of Siena)

27-5 17:40-18:00
SRC-gAudio: Sampling-Rate-Controlled Audio Generation
Chenxing Li (Tencent AI Lab)*; Manjie Xu (Beijing Institute of Technology); Dong Yu (Tencent AI Lab)

28. Speech and Language Processing Meeting Room 10 04-Dec

28-1 16:20-16:40
Domain Adaptation by Alternating Learning of Acoustic and Linguistic Information for Japanese Deaf and Hard-of-Hearing People
Kaito Takahashi (Toyohashi University of Technology)*; Yukoh Wakabayashi (Toyohashi University of Technology); Kengo Ohta (National Institute of Technology, Anan College); Akio Kobayashi (Yamato University); Norihide Kitaoka (Toyohashi University of Technology)

28-2 16:40-17:00
Speech emotion recognition based on crossmodal transformer and attention weight correction
Ryusei Terui (University of Tsukuba)*; Takeshi Yamada (University of Tsukuba)

28-3 17:00-17:20
Unsupervised Discovery of Non-Categorical L2 Error Patterns Using Wav2Vec2.0 Code Vectors
Eunsoo Hong (Seoul National University)*; Sunhee Kim (Seoul National University); Minhwa Chung (Seoul National University)

28-4 17:20-17:40
An Effective Contextualized Automatic Speech Recognition Approach Leveraging Self-Supervised Phoneme Features
Li-Ting Pai (National Taiwan Normal University)*; Yi-Cheng Wang (National Taiwan Normal University); Bi-Cheng Yan (National Taiwan Normal University); Hsin-Wei Wang (NTNU); Jia-Liang Lu (National Taiwan Normal University); Chi-Han Lin (E.SUN Financial Holding Co., Ltd); Juan-Wei Xu (E.SUN Financial Holding Co., Ltd); Berlin Chen (National Taiwan Normal University)

28-5 17:40-18:00
COIN-AT-PVAD: A Conditional Intermediate Attention PVAD
En-Lun Yu (National Taiwan Normal University)*; Chang Ruei-Xian (National Chi Nan University); Jeih-weih Hung (National Chi Nan University); Shih-Chieh Huang (Realtek Semiconductor Corp.); Berlin Chen (National Taiwan Normal University)

29. Advanced Topics on Sound Event and Scene Analysis Meeting Room 1 05-Dec

29-1 10:20-10:40
Multi-Modal Video Summarization Based on Two-Stage Fusion of Audio, Visual, and Recognized Text Information
Zekun Yang (Nagoya University)*; Jiajun He (Nagoya University); Tomoki Toda (Nagoya University)

29-2 10:40-11:00
Prediction-error-based Adaptive SpecAugment for Fine-tuning the Masked Model on Audio Classification Tasks
Xiao Zhang (WASEDA University)*; HAORAN XING (Waseda University); Mingxue Song (Waseda University); Daiki Takeuchi (NTT Corporation); Noboru Harada (NTT); Shoji Makino (Waseda University)

29-3 11:00-11:20
Synchronization of Signals with Sampling Rate Offset and Missing Data Using Dynamic Pro-

gramming Matching
Hayato Takeuchi (Tokyo Metropolitan University)*; Nobutaka Ono (Tokyo Metropolitan University)

29-4 11:20-11:40
LEAD Dataset: How Can Labels for Sound Event Detection Vary Depending on Annotators?
Naoki Koga (Doshisha University); Yoshiaki Bando (National Institute of Advanced Industrial Science and Technology); Keisuke Imoto (Doshisha University)*

29-5 11:40-12:00
SSL-based Chewing and Swallowing Detection Using Multiple Skin-contact Microphones
Toshihiro Tsukagoshi (Shizuoka university)*; Kazuhiro Koiwai (Shizuoka University); Masafumi Nishida (Shizuoka University); Masafumi Nishimura (Shizuoka University)

30. Recent Advances in Multimedia Enrichment and Security Meeting Room 2 05-Dec

30-1 10:20-10:40
Enhancing Security Using Random Binary Weights in Privacy-Preserving Federated Learning
Hiroto Sawada (Chiba university)*; Shoko Imaizumi (Chiba University); Hitoshi Kiya (Tokyo Metropolitan University)

30-2 10:40-11:00
Estimation of rotation angle and anisotropic scaling rate using pilot signals for watermarking
Rinka Kawano (Yamaguchi University)*; Masaki Kawamura (Yamaguchi University)

30-3 11:00-11:20
On the Security of Bitstream-level JPEG Encryption with Restart Markers
Mare Hirose (Chiba University)*; Shoko Imaizumi (Chiba University); Hitoshi Kiya (Tokyo Metropolitan University)

30-4 11:20-11:40
Improved Ultimate Link without Markers for Projective Transformation
Keiji Yamadera (Kyushu Institute of Technology); Michiharu Niimi (Kyushu Institute of Technology)*

30-5 11:40-12:00
Detection of Diffusion-Generated Images Using Sparse Coding
Daishi Tanaka (Kyushu Institute of Technology); Michiharu Niimi (Kyushu Institute of Technology)*

31. Image, Video, and Multimedia Meeting Room 3 05-Dec

31-1 10:20-10:40
Improved Architecture for High-resolution Piano Transcription to Efficiently Capture Acoustic Characteristics of Music Signals
Jinyi Mi (Nagoya University)*; Sehun Kim (Nagoya University); Tomoki Toda (Nagoya University)

31-2 10:40-11:00
Ev3DGS:Event Enhanced 3D Gaussian Splatting from Blurry Images
Junwu Huang (Northwestern Polytechnical University); Zhexiong Wan (Northwestern Polytechnical University); Zhicheng Lu (Northwestern Polytechnical University); Juanjuan Zhu (Northwestern Polytechnical University); Mingyi He (Northwestern Polytechnical University); Yuchao Dai (Northwestern Polytechnical University)*

31-3 11:00-11:20
New Abnormal Behavior Detection for Patient Surveillance System
Yujin Han (Dongduk Women's University); taewan kim (Dongduk women's university)*

31-4 11:20-11:40

Utilizing Cross Layer Attentions for Semantic Segmentation of Small Objects
Chi-Hsuan Lu (Chung Yuan Christian University); Yu-Hsien Chung (Chung Yuan Christian University); Jung-Hui Cho (Chung Yuan Christian University); Chih-Chang Yu (Chung Yuan Christian University)*

31-5 11:40-12:00
Music2Fail: Transfer Music to Failed Recorder Style
Chon In Leong (National Taiwan University)*; I-Ling Chung (national taiwan university); Kin Fong Chao (National Taiwan University); Jun-You Wang (Academia Sinica); Yi-Hsuan Yang (National Taiwan University); Roger Jang

32. Signal and Information Processing & Systems Meeting Room 4 05-Dec

32-1 10:20-10:40
U-Mamba-Net: A highly efficient Mamba-based U-net style network for noisy and reverberant speech separation
Shaoxiang Dang (Nagoya University)*; Tetsuya Matsumoto (Nagoya University); Yoshinori Takeuchi (Daido University); Hiroaki Kudo (Nagoya University)

32-2 10:40-11:00
Graph Filter Transfer for Time-Varying Signal Estimation Between Two Networks
Tsutahiro Fukuhara (Osaka University)*; Junya Hara (Osaka University); Hiroshi Higashi (Osaka University); Yuichi Tanaka (Osaka University)

32-3 11:00-11:20
Few-Shot Audio Classification Model for Detecting Classroom Interactions Using LaSO Features in Prototypical Networks
Md Rashed Iqbal (University of Wollongong)*; Christian Ritz (University of Wollongong); Jie Yang (University of Wollongong)

32-4 11:20-11:40
Subset Random Sampling of Finite Time-vertex Graph Signals
Hang Sheng (Fudan University); Qinji Shu (Fudan University); HUI FENG (Fudan University)*; bo Hu (Fudan university)

32-5 11:40-12:00
Dynamic Sensor Placement on Graphs Based on Graph Signal Sampling Theory
Saki Nomura (Tokyo University of Agriculture and Technology)*; Junya Hara (Osaka University); Hiroshi Higashi (Osaka University); Yuichi Tanaka (Osaka University)

33. Speech and Language Processing Meeting Room 5 05-Dec

33-1 10:20-10:40
Can We Estimate Purchase Intention Based on Zero-shot Speech Emotion Recognition?
Ryotaro Nagase (Ritsumeikan University); Takashi Sumiyoshi (Hitachi, Ltd.); Natsuo Yamashita (Hitachi, Ltd.); Kota Dohi (Hitachi Ltd.); Yohei Kawaguchi (Hitachi, Ltd.)*

33-2 10:40-11:00
Assessment and Improvement of Customer Service Speech with Multiple Large Language Models
So Watanabe (University of yamanashi); Chee Siang Leow (University of Yamanashi)*; Junichi Hoshino (University of Tsukuba); Takehito Utsuro (University of Tsukuba); Hiromitsu Nishizaki (University of Yamanashi)

33-3 11:00-11:20
JAM: A Unified Neural Architecture for Joint Multi-granularity Pronunciation Assessment and Phone-level Mispronunciation Detection and Diagnosis Towards a Comprehensive CAPT System

Yue-Yang He (National Taiwan Normal University)*; Bi-Cheng Yan (National Taiwan Normal University, Taiwan); Tien-Hong Lo (National Taiwan Normal University); Meng-Shin Lin (National Taiwan Normal University); Yung-Chang Hsu (EZ-AI); Berlin Chen (National Taiwan Normal University)

33-4 11:20-11:40
Effective Speech Data Augmentation Method To Improve Customer Service Representative Speech Recognition System Performance
Huiyong Bak (Lotte innovate)*; Changhyeon Jeong (Lotte innovate)

33-5 11:40-12:00
Empower Typed Descriptions by Large Language Models for Speech Emotion Recognition
Haibin Wu (National Taiwan University); Huang-Cheng Chou (Department of Electrical Engineering at National Tsing Hua University (NTHU))*; Kai-Wei Chang (National Taiwan University); Lucas Goncalves (The University of Texas at Dallas); Jiawei Du (National Taiwan University); Jyh-Shing Roger Jang (National Taiwan University); Chi-Chun Lee (National Tsing Hua University); Hung-yi Lee (National Taiwan University)

34. Audio Processing Meeting Room 6 05-Dec

34-1 10:20-10:40
Wind Noise Reduction with Orthogonal Polynomial Expansion
Li Du (Northwestern Polytechnical University)*; Lijun Zhang (Northwestern Polytechnical University)

34-2 10:40-11:00
Few-Shot Open-Set Keyword Spotting with Multi-Stage Training
LoYa Li (National Taiwan Normal University)*; Tien-Hong Lo (National Taiwan Normal University); Jeih-weih Hung (National Chi Nan University); Shih-Chieh Huang (Realtek Semiconductor Corp.); Berlin Chen (National Taiwan Normal University)

34-3 11:00-11:20
Self-Supervised Augmented Diffusion Model for Anomalous Sound Detection
Jiawei Yin (Ecust); yu gao (Midea)*; Wenbin Zhang (Midea); Mingjun Zhang (Midea)

34-4 11:20-11:40
Murmur Separation and Classification from Heart Sound Using Constrained Singular Spectrum Analysis and Wavelet Transform
Yuanyang Qi (Imperial College London)*; Saeid Sanei (Imperial College London)

34-5 11:40-12:00
A Non-Intrusive Speech Quality Assessment Model using Whisper and Multi-Head Attention
Guojian Lin (Southern University of Science and Technology); Yu Tsao (Academia Sinica); Fei Chen (Southern University of Science and Technology)*

35. Emerging Technologies and Applications Of Image Processing And Computer Vision Meeting Room 9 05-Dec

35-1 10:20-10:40
Confidence-Aware Learning for Person Re-identification with Noisy Labels
Duhyun Kim (Ulsan National Institute of Science and Technology)*; Jae-Young Sim (Ulsan National Institute of Science and Technology)

35-2 10:40-11:00
Test-Time Optimization for Post-Processing of Compressed Videos
Hongil Kim (Korea university); Changwoo Han (Korea University); Donghyun Kim (ETRI); Sung-Chang Lim (ETRI); Seung-Won Jung (Korea University)*

35-3 11:00-11:20

Lifelong Person Re-Identification with Backward-Compatibility
Minyoung Oh (Ulsan National Institute of Science and Technology); Jae-Young Sim (Ulsan National Institute of Science and Technology)*

35-4 11:20-11:40
Enhancing Semiconductor X-RAY Images: A Framework Combining Denoising and Super-Resolution Modules With a Novel Dataset
Jae Hoon Shim (Seoul National University)*; Min Woo Kim (Seoul National University); Sang Hwa Lee (Seoul National University); Nam Ik Cho (Seoul National University)

35-5 11:40-12:00
Monocular Depth Estimation for Autonomous Driving Based on Instance Clustering Guidance
Dahyun Kim (Korea University)*; Dongkwon Jin (Korea university); Chang-Su Kim (Korea university)

36. Advanced Signal Processing for Information Collection and Data Analysis in Wireless Environmental Sensing Meeting Room 10 05-Dec

36-1 10:20-10:40
Data-Driven Tuning for Weighted Least Square of BLE-AoA-based Indoor Localization
Ginji Ohashi (Doshisha university); Shinsuke Ibi (Doshisha University)*; Takumi Takahashi (Osaka Univ.); Hisato Iwai (Doshisha University)

36-2 10:40-11:00
Observation of the terrestrial radio environment using the low earth orbit satellite constellation
Takatoshi Obata (Shinshu University)*; Osamu Takyu (Shinshu University); Kei Inage (Tokyo Metropolitan College of Industrial Technology); Takeo Fujii (The University of Electro-Communications); Kohei Yoshida (NEC Corporation); Masayuki Ariyoshi (NEC Corporation)

36-3 11:00-11:20
Deep Unfolding Aided Parameter Optimization for Multi-task Diffusion LMS Algorithm
Xiaoqing Tong (Kyoto University)*; Kazunori Hayashi (Kyoto University)

36-4 11:20-11:40
Reduced-dimensional MUSIC Algorithm for Frequency Diverse Array in MIMO Radar System
Beizuo Zhu (Kyoto University)*; Kazunori Hayashi (Kyoto University); Hiroki Mori (Toshiba Corporation)

36-5 11:40-12:00
Collection of Correlated Information from Superimposed Multiple Chirp Signals
Koki Aoyama (The University of Electro-Communications)*; Koichi Adachi (The University of Electro-Communications)

37. New Frontiers in Biometric Authentication Meeting Room 1 05-Dec

37-1 14:00-14:20
A Quasilinear-Time CVP Algorithm for Triangular Lattice Based Fuzzy Extractors and Fuzzy Signatures
Kenta Takahashi (Hitachi, Ltd.)*; Wataru Nakamura (Hitachi, Ltd.)

37-2 14:20-14:40
Enhancing Remote Adversarial Patch Attacks on Face Detectors with Tiling and Scaling
Masora Okano (Shizuoka University)*; Koichi Ito (Tohoku University); Masakatsu Nishigaki (Shizuoka University); Tetsushi Ohki (Shizuoka University)

37-3 14:40-15:00

Multibiometrics Using a Single Face Image
Koichi Ito (Tohoku University)*; Taito Tonosaki (Tohoku University); Takafumi Aoki (Tohoku University); Tetsushi Ohki (Shizuoka University); Masakatsu Nishigaki (Shizuoka University)

37-4 15:00-15:20
Multi-Observed Authentication: A secure and usable authentication based on multi-point observation of a single physical credential
Wataru Hatakeyama (Shizuoka University)*; Shinnosuke Nozaki (Mitsubishi Electric Corporation); Ayumi Serizawa (Shizuoka University); Mizuho Yoshirira (Shizuoka University); Masahiro Fujita (Mitsbishi Electric); Ayako Yoshimura (Mitsubishi Electric Corporation); Tetsushi Ohki (Shizuoka University); Masakatsu Nishigaki (Shizuoka University)

38. Recent Advances in Multimedia Enrichment and Security Meeting Room 2 05-Dec

38-1 14:00-14:20
Generation of Target Speech with Speaker Individuality Based on Accent Conversion for English Pronunciation Learning
Rei Hamakawa (Kyushu Institute of Technology); Michiharu Niimi (Kyushu Institute of Technology)*

38-2 14:20-14:40
Proposal of Blind Extractable Additive Video Watermarking Method
Nao Harada (Yamaguchi University)*; Rinka Kawano (Yamaguchi University); Masaki Kawamura (Yamaguchi University)

38-3 14:40-15:00
Transfer-Based Adversarial Attack Against Multimodal Models by Exploiting Perturbed Attention Region
Raffaele Disabato (Unibo)*; April Pyone Maung Maung (National Institute of Informatics); Huy Nguyen (National Institute of Informatics); Isao Echizen (National Institute of Informatics)

38-4 15:00-15:20
A Permutation-based Reversible Data Hiding Method with Zero Visual Distortion
Wendi Zhu (Monash University)*; KokSheik Wong (Monash University Malaysia); Minoru Kuribayashi (Tohoku University)

39. Image, Video, and Multimedia Meeting Room 3 05-Dec

39-1 14:00-14:20
VietSing: A High-quality Vietnamese Singing Voice Corpus
Minh Vu (Hanoi University of Science and Technology)*; Zhou Wei (Nanyang Technological University); Binit Bhattarai (Vellore Institute of Technology); Kah Kuan Teh (Institute for Infocomm Research (I2R)); Tran Huy Dat (Institute for Infocomm Research (I2R))

39-2 14:20-14:40
Inertial Strengthened CLIP model for Zero-shot Multimodal Egocentric Activity Recognition
Mingzhou He (University of Electronic Science and Technology of China); Haojie Wang (University of Electronic Science and Technology of China); Shuchang Zhou (University of Electronic Science and Technology of China); Qingbo Wu (University of Electronic Science and Technology of China)*; King Ngi Ngan (University of Electronic Science and Technology of China); Fanman Meng (University of Electronic Science and Technology of China); Hongliang Li (University of Electronic Science and Technology of China)

39-3 14:40-15:00
Optimization of the Intensity Aware Loss for Dynamic Facial Expression Recognition
Davy Tec-Hinh Lau (National Taiwan University); Jian-Jiun Ding (National Taiwan University)*; Guillaume Muller (Ecole des Mines de Saint-Etienne)

39-4 15:00-15:20
Dictionary Learning Based Two-stage Near-lossless Video Compression
Zuhai Zhang (Beijing University of Technology); Luheng Jia (Beijing University of Technology)*; Li Song (Shanghai Jiao Tong University); Shuyuan Zhu (University of Electronic Science and Technology of China); Yuanfang Guo (Beihang University); Kebin Jia (Beijing University of Technology)

40. Signal and Information Processing & Systems Meeting Room 4 05-Dec

40-1 14:00-14:20
Dictionary Learning for Directed Graph Signals via Augmented GFT
Tsubasa Naito (Niigata University)*; Ryuto Ito (Niigata University); Yuichi Tanaka (Osaka University); Shogo Muramatsu (Niigata Univ.)

40-2 14:20-14:40
Robust Quantile Regression Under Unreliable Data
Yoshifumi Shoji (Keio University)*; Masahiro Yukawa (Keio University)

40-3 14:40-15:00
Ensemble learning based head-related transfer function personalization using anthropometric features
Yih-Liang Shen (National Yang Ming Chiao Tung University)*; Tai-Shih Chi (National Yang Ming Chiao Tung University)

40-4 15:00-15:20
Blind Estimation of Room Volume from Reverberant Speech Based on the Modulation Transfer Function
Nutchanon Siripool (Sirindhorn International Institute of Technology)*; Waree kongprawechnon (Sirindhorn International Institute of Technology, Thammasat University); Masashi Unoki (JAIST)

41. Speech and Language Processing Meeting Room 5 05-Dec

41-1 14:00-14:20
Disentangling Speaker Representations from Intuitive Prosodic Features for Speaker-Adaptative and Prosody-Controllable Speech Synthesis
Cheng Pengyu (Xiaomi)*

41-2 14:20-14:40
A Pilot Study of Applying Sequence-to-Sequence Voice Conversion to Evaluate the Intelligibility of L2 Speech Using a Native Speaker's Shadowings
Haopeng Geng (The University of Tokyo)*; Daisuke Saito (The University of Tokyo); Nobuaki Minematsu (The University of Tokyo); Haopeng Geng (The University of Tokyo)

41-3 14:40-15:00
EADSum: Element-Aware Distillation for Enhancing Low-Resource Abstractive Summarization
Jia-Liang Lu (National Taiwan Normal University)*; Bi-Cheng Yan (National Taiwan Normal University, Taiwan); Yi-Cheng Wang (National Taiwan Normal University); Tien-Hong Lo (National Taiwan Normal University); Hsin-Wei Wang (NTNU); Li-Ting Pai (National Taiwan Normal University); Berlin Chen (National Taiwan Normal University)

41-4 15:00-15:20
A Tiny Whisper-SER: Unifying Automatic Speech Recognition and Multi-label Speech Emotion Recognition Tasks

Huang-Cheng Chou (Department of Electrical Engineering at National Tsing Hua University (NTHU))*

42. Audio Processing Meeting Room 6 05-Dec

42-1 14:00-14:20
EEND-EM: End-to-End Neural Speaker Diarization with EM-Network
Beom Jun Woo (Seoul National University)*; Ji Won Yoon (Chung-Ang University); Min Hyun Han (Seoul National University); Chan Yeong Moon (Seoul National University); Nam Soo Kim (Seoul National University)

42-2 14:20-14:40
Personal Voice Activity Detection With Ultra-Short Reference Speech
Longting Xu (Donghua University); Mingjun Zhang (DongHua University); Wenbin Zhang (Midea); Tianyi Wang (Midea); Jiawei Yin (Midea); yu gao (Midea)*

42-3 14:40-15:00
An Investigation on the Speech Recovery from EEG Signals Using Transformer
Tomoaki Mizuno (the University of Electro-Communications, Tokyo)*; Takuya Kishida (Aichi Shukutoku University); Natsue Yoshimura (Tokyo Institute of Technology); Toru Nakashika (The University of Electro-Communications, Tokyo)

43. Audio Processing Meeting Room 8 05-Dec

43-1 14:00-14:20
WavLM and Omni-Scale CNNs: Enhancing Boundary Detection in Partially Spoofed Audio
Menghan Li (XinJiang University)*; Zhihua Huang (Xinjiang university)

43-2 14:20-14:40
Semi-Supervised Far-Field Speaker Verification with Distance Metric Domain Adaptation
Han Wang (Donghua University)*; Mingrui He (Donghua University); Mingjun Zhang (DongHua University); Longting Xu (Donghua University)

43-3 14:40-15:00
Non-Target Conversion Based Speech Steganography for Secure Speech Communication System
Mingjun Zhang (DongHua University); Yan Feng (Donghua University); yu gao (Midea); Longting Xu (Donghua University)*

43-4 15:00-15:20
Enhancing Acoustic Scene Classification with Layer-wise Fine-Tuning on the SSAST Model
Shuting Hao (The University of Tokyo)*; Daisuke Saito (The University of Tokyo); Nobuaki Minematsu (The University of Tokyo)

44. High Performance Image and Video Processing and Applications Meeting Room 9 05-Dec

44-1 14:00-14:20
Forward Prediction-Guided Cross-Partition Targeted Pruning for VVenC
Jingyuan Tang (Beijing University of Posts and Telecommunications)*; Songlin Sun (Beijing University of Posts and Telecommunications)

44-2 14:20-14:40
Contrastive Learning Based Knowledge Distillation for Enhancing Defect Detection
Jing-Ming Guo (National Taiwan University of Science and Technology); Lun-Da Yuan (National Taiwan University of Science and Technology); CIAN HUANG (National Taiwan University of Science and Technology)*; Yi-Chong Zeng (Taiwan)

44-3 14:40-15:00
Screen Content Encoding Network Based on Deep Contextual Information
Tianyu Gong (Beijing Jiaotong University)*; Tao Zhang (Beijing Polytechnic College); Ye Zhong (Beijing Jiaotong University); Mengmeng Zhang (Beijing Union University); Huihui Bai (Beijing Jiaotong University)

44-4 15:00-15:20
A Coarse-to-Fine Change Detection Framework for Remote Sensing Sparse Cultivated Land
yuan hu (Northwestern Polytechnical University)*; Yifan Zhang (Northwestern Polytechnical University); Mingyang Ma (Northwestern Polytechnical University); Shaohui Mei (Northwestern Polytechnical University, China)

45. Advancements in Biosignal Decoding and Neuromodulation for Human Function Enhancement Meeting Room 10 05-Dec

45-1 14:00-14:20
Context-FFT: A Context Feed Forward Transformer Network for EEG-based Speech Envelope Decoding
Ximin Chen (Southern University of Science and Technology); Yuting Ding (Southern University of Science and Technology); Nan Yan (Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences); Changsheng Chen (Shenzhen University); Fei Chen (Southern University of Science and Technology)*

45-2 14:20-14:40
Effect of Dynamic Binaural Beats on Concentration Enhancement
Jun-Seok LEE (Korea University); Yun-Sung Lee (Korea University); Han-Jeong Hwang (Korea University)*

45-3 14:40-15:00
EEG-based Evaluation of Enjoyment Emotion during cognitive-motor task
Haruna Aoki (Graduate School of Science and Technology, Meiji University)*; Sinan Zhang (Organization for the Strategic Coordination of Research and Intellectual Properties, Meiji University); Yumie Ono (Meiji University)

45-4 15:00-15:20
Exploring Brain Connectivity Patterns and Cognitive Resilience in Aging: A Study with the LEMON Dataset
Kapeleshh ks (Macnica)*; Chen Wei (Macnica); Prince Aldrin Domer (Macnica); Hong Ji (Xi'an Polytechnic University)

45-5 15:20-15:40
A Study on Packet-Level Index Modulation Using Frequency Offsets within a LoRaWAN Channel
mai ohta (Fukuoka university)*; Hiroki Matsuura (The University of Electro-Communications); Takeo Fujii (The University of Electro-Communications)

46. Wireless Communications and Networking Meeting Room 1 05-Dec

46-1 16:40-17:00
Blind Self-Interference Analog Canceller with Differential Delay for Backscatter Communications
Koichi Nishikawa (Doshisha University); Shinsuke Ibi (Doshisha University)*; Takumi Takahashi (Osaka Univ.); Hisato Iwai (Doshisha University)

46-2 17:00-17:20
IoT-based Smart Attendance System using Face Recognition and Motion Detection
Umi Syamimi Saadon (Monash University)*; Chern Hong Lim (Monash University)

47. Recent Advances in Multimedia Enrichment and Security Meeting Room 2 05-Dec

47-1 16:40-17:00
Generation of Photo Slideshow with Song based on Closeness between Concept of Lyrics and That of Images
Mei Hashimoto (Kyushu Institute of Technology); Michiharu Niimi (Kyushu Institute of Technology)*

47-2 17:00-17:20
Disposable-key-based image encryption for collaborative learning of Vision Transformer
Rei Aso (Tokyo Metropolitan Univercity)*; Sayaka Shiota (Tokyo Metropolitan University); Hitoshi Kiya (Tokyo Metropolitan University)

47-3 17:20-17:40
Significance of Lower Frequency Regions for Audio Deepfake Detection
Arth Shah (DA-IICT)*; Hemant Patil (DAICT, Gujrat)

47-4 17:40-18:00
EAViT: External Attention Vision Transformer for Audio Classification
Aquib Iqbal (University of Massachusetts Amherst); Abid Hasan Zim (Aligarh Muslim University); Md Asaduzzaman Tonmoy (Aligarh Muslim University); Limengnan Zhou (University of Electronic Science and Technology of China); Asad Malik (Monash University Malaysia)*; Minoru Kuribayashi (Tohoku University)

48. Image, Video, and Multimedia Meeting Room 3 05-Dec

48-1 16:40-17:00
A Two-Stage Method for 3D Architecture Wireframe Reconstruction from Airborne LiDAR Point Cloud
Jiahao Zhang (Northwestern Polytechnical University); Qi Liu (Northwestern Polytechnical University)*; Le Hui (Northwestern Polytechnical University); Yuchao Dai (Northwestern Polytechnical University)

48-2 17:00-17:20
Secure Moving Object Detection Transformer in Compressed Video with Feature Fusion
Yuru Song (Sun Yat-sen University); Yike Chen (Sun Yat-sen University); Peijia Zheng (Sun Yat-sen University)*; Yusong Du (School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, P.R. China); Weiqi Luo (Sun Yat-sen University)

48-3 17:20-17:40
NeRF-FCM: Attention-based Feature Calibration Mechanisms for 3D Object Detection Using NeRF
Hana Lebeta Goshu (The Hong Kong Polytechnic University)*; Jun Xiao (The Hong Kong Polytechnic University); Kin-Chung Chan (The Hong Kong Polytechnic University); Cong Zhang (The Hong Kong Polytechnic University); Mulugeta Tegegn Gemeda (Jimma Institute of Technology); Kin-Man Lam (The Hong Kong Polytechnic University)

48-4 17:40-18:00
High-Quality Facial Pose Generation with Latent Space Processing
Wan-Chi Siu (The Hong Kong Polytechnic University)*; Wing-Ho Cheng (St. Francis University, Hong Kong); H Anthony Chan (St. Francis University, Hong Kong)

49. Signal and Information Processing & Systems Meeting Room 4 05-Dec

49-1 16:40-17:00
Significance of Entropy Based Features For Dysarthric Severity Level Classification
Meghana Avula (Dhirubhai Institute of Information and Communication Technology)*; Aditya Pusuluri (DAI-

ICT); Hemant Patil (DAICT, Gujrat)

49-2 17:00-17:20
Incorporating Auditory Processing into Undergraduate Signal Processing Courses to Enhance Student Learning
Kaibao Nie (University of Washington, Bothell)*

49-3 17:20-17:40
A Real-Time Platform for Portable and Scalable Active Noise Mitigation for Construction Machinery
Santi Peksi (NTU Singapore); Woon Seng Gan (NTU)*; Chung Kwan Lai (University of Southampton); Yen Theng Lee (Nanyang Technological University,); Dongyuan Shi (Nanyang Technological University); Bhan Lam (NTU)

50. Speech and Language Processing Meeting Room 5 05-Dec

50-1 16:40-17:00
A Comparative Study on the Biases of Age, Gender, Dialects, and L2 speakers of Automatic Speech Recognition for Korean Language
Jonghwan Na (Inha university); Yeseul Park (Inha University); Bowon Lee (Inha University)*

50-2 17:00-17:20
Targeted Representation with Information Disentanglement Encoding Networks in Tasks
Takumi Nagawaki (Gifu University)*; Keisuke Ikeda (Gifu University); Satoshi Tamura (Gifu University); Kohei Chike (RICOH); Hiroyuki Nagano (RICOH); Masaki Nose (RICOH)

50-3 17:20-17:40
PG-MDD: Prompt-Guided Mispronunciation Detection and Diagnosis Leveraging Articulatory Features
Meng-Shin Lin (National Taiwan Normal University)*; Bi-Cheng Yan (National Taiwan Normal University, Taiwan); Tien-Hong Lo (National Taiwan Normal University); Hsin-Wei Wang (NTNU); Yue-Yang He (National Taiwan Normal University); Wei-Cheng Chao (Telecommunication Laboratories, Chunghwa Telecom Co); Berlin Chen (National Taiwan Normal University)

51. Audio Processing Meeting Room 6 05-Dec

51-1 16:40-17:00
Experimental Evaluation of Speech Enhancement for In-Car Environment Using Blind Source Separation and DNN-based Noise Suppression
Yutsuki Takeuchi (Tokyo Metropolitan University)*; Taishi Nakashima (Tokyo Metropolitan University); Nobutaka Ono (Tokyo Metropolitan University); Takashi Takazawa (MIRISE Technologies Corporation); Shuhei Shimanoe (MIRISE Technologies Corporation); Yoshinori Tsuchiya (MIRISE Technologies Corporation)

51-2 17:00-17:20
Auxiliary-Function-Based Steering Vector Estimation Method for Spatially Regularized Independent Low-Rank Matrix Analysis
Sota Hirata (The Universiy of Tokyo)*; Norihiro Takamune (the University of Tokyo); Kouei Yamaoka (The University of Tokyo); Daichi Kitamura (National Institute of Technology, Kagawa College); Hiroshi Saruwatari (The University of Tokyo); Yu Takahashi (Yamaha Corporation); Kazunobu KONDO (Yamaha Corp.)

51-3 17:20-17:40
Two-stage Framework for Robust Speech Emotion Recognition Using Target Speaker Extraction in Human Speech Noise Conditions
Jinyi Mi (Nagoya University)*; Xiaohan Shi (Nagoya University); Ding Ma (Nagoya University); Jiajun He (Nagoya University); Takuya Fujimura (Nagoya University); Tomoki Toda (Nagoya University)

51-4 17:40-18:00
Data generation for speaker diarization by speaker transition information
Keigo Ichikawa (Nagoya Institute of Technology)*; Sei Ueno (Nagoya Institute of Technology); Akinobu Lee (Nagoya Institute of Technology)

52. Audio Processing Meeting Room 8 05-Dec

52-1 16:40-17:00
Generating Room Impulse Responses Using Neural Networks Trained with Weighted Combinations of Acoustic Parameter Loss Functions
Hualin Ren (University of Wollongong)*; Christian Ritz (University of Wollongong); Jiahong Zhao (University of Southampton); Xiguang Zheng (University of Wollongong); Daeyoung Jang (Electronics and Telecommunications Research Institute)

52-2 17:00-17:20
Audio Similarity Detection
Siddharth Malhotra (Nirma University); Sapan Mankad (Nirma University)*

52-3 17:20-17:40
Towards a B-format Ambisonic Room Impulse Response Generator Using Conditional Generative Adversarial Network
Hualin Ren (University of Wollongong)*; Christian Ritz (University of Wollongong); Jiahong Zhao (University of Southampton); Xiguang Zheng (University of Wollongong); Daeyoung Jang (Electronics and Telecommunications Research Institute)

52-4 17:40-18:00
What to Refer and How? - Exploring Handling of Auxiliary Information in Target Speaker Extraction
Tomohiro Hayashi (Waseda University)*; Riku Ogino (Waseda University); Kohei Saijo (Waseda University); Tetsuji Ogawa (Waseda University)

53. High Performance Image and Video Processing and Applications Meeting Room 9 05-Dec

53-1 16:40-17:00
Efficient Adaptation for Real-World Omnidirectional Image Super-Resolution
Cuixin Yang (The Hong Kong Polytechnic University)*; Rongkang Dong (The Hong Kong Polytechnic University); Kin-Man Lam (The Hong Kong Polytechnic University)

53-2 17:00-17:20
More Direct and stage-wise network for Face Super Resolution
Yohei Horiguchi (Keio univ)*

53-3 17:20-17:40
Camera Focal Length Prediction for Neural Novel View Synthesis from Monocular Video
Dipanita Chakraborty (King Mongkut's University of Technology Thonburi)*; Werapon Chiracharit (King Mongkut's University of Technology Thonburi); Kosin Chamnongthai (KMUTT); Minoru Okada (Nara Institute of Science and Technology)

53-4 17:40-18:00
Scene-Segmentation-Based Exposure Compensation for Tone Mapping of High Dynamic Range Scenes
Yuma Kinoshita (Tokai University)*; Hitoshi Kiya (Tokyo Metropolitan University)

54. Advancements in Biosignal Decoding and Neuromodulation for Human Function Enhancement Meeting Room 10 05-Dec

54-1 16:40-17:00
Effect of Phase-Locked Transcranial Alternating Current Stimulation on Vocal tremor
JUNTING WANG (Meiji University)*; Satoko Koganemaru (Human Brain Research Center, Graduate School of Medicine Kyoto University); Atsushi Shima (Human Brain Research Center, Graduate School of Medicine Kyoto University); Yedi Cao (Human Brain Research Center, Graduate School of Medicine Kyoto University); Kana Hirakawa (Graduate School of Medicine and Faculty of Medicine, Kyoto University); Ken Iwagana (Graduate School of Medicine and Faculty of Medicine, Kyoto University); Atsushi Suehiro (Graduate School of Medicine and Faculty of Medicine, Kyoto University); Keiko Maekawa (Department of Otolaryngology, Kobe City Medical Center General Hospital); Tatsuya Mima (The Graduate School of Core Ethics and Frontier Sciences, Ritsumeikan University); Yumie Ono (Meiji University)

54-2 17:00-17:20
Complex CNN incorporating Hilbert transform for steady-state visual evoked potential BCI
Rintaro Takata (The University of Electro-Communications)*; Yoshikazu Washizawa (The University of Electro-Communications)

54-3 17:20-17:40
Electroencephalogram-Based Effective Features for Sustained Attention Assessment in Conversation
Masaya Togashi (Tokyo University of Agriculture and Technology); Ingon Chanpornpakdi (Tokyo University of Agriculture and Technology); Toshihisa Tanaka (Tokyo University of Agriculture and Technology)*

55. Embedded and Real-Time Systems for AI and Signal Processing Applications Meeting Room 1 06-Dec

55-1 09:00-09:20
Accelerated Real-Time Local Maxima Detection in Video Streams Using FPGA Technology
Anindhita Nayazirly (Institut Teknologi Bandung); Yahwista Salomo (ITB)*; Trio Adiono (Institut Teknologi Bandung); Infall Syafalni (Bandung Institute of Technology); Nana Sutisna (Institut Teknologi Bandung); Rahmat Mulyawan (Institut Teknologi Bandung)

55-2 09:20-09:40
A Configurable OFDM Baseband Processor for RF-UOWC System-on-Chip
Trio Adiono (Institut Teknologi Bandung); Erwin Setiawan (Microelectronics Center ITB)*; Michael Jonathan (Microelectronics Center ITB); Rahmat Mulyawan (Bandung Institute of Technology); Nana Sutisna (Institut Teknologi Bandung); Infall Syafalni (Bandung Institute of Technology); Wasiu Popoola (University of Edinburgh)

55-3 09:40-10:00
Hammering Sound Inspection System Using HPSS and Gradient Boosting with a Wall-Climbing Robot
Nichika Koyama (Tokyo University of Science)*

55-4 10:00-10:20
Implementation of Real Time Oscillometric Based Algorithm for Blood Pressure Measurement in Patient Monitor
Trio Adiono (Institut Teknologi Bandung); Clarence Amadeus (Bandung Institute of Technology)*; Teuku Rafifsyah Thomi (PT. Xirka Dama Persada); Sindy Novaria Cicilya Sinaga (PT. Xirka Dama Persada)

56. Selected Papers from APSIPA Workshop on Advanced Signal and Information Processing Meeting Room 2 06-Dec

56-1 09:00-09:20
Automated Pseudo-Label Generation and Parallel Computing for Enhanced Few-Shot Medical Image Segmentation
Ha Do (VNU University of Science)*; Duc Nguyen Trong (VNU University of Science); Tien-Dung Do (Post and Telecommunication Institute of TechnologyPost)

56-2 09:20-09:40
Enhanced Sparse Convolutional Detection Model for 3D Object Detection in Autonomous Vehicles Adapted to Traffic Conditions in Vietnam
Ha Do (VNU University of Science)*; Vu Hoang Dung (Phenikaa-X Company); Kien Nguyen (Hanoi University of Science and Technology)

56-3 09:40-10:00
Enhancing Cell Segmentation using Deep Learning Models by Custom Processing Techniques
Ha Do (VNU University of Science)*; Van De Nguyen (Department of Pathological Anatomy, Central Military Hospital 108); Minh Huong Dang Hoang (VNU University Of Science); Nguyen Huy (Hanoi University of Science); Cuong Dinh Manh (Cuong2003)

56-4 10:00-10:20
Marker-Aware Ovarian Tumor Segmentation from Ultrasound Images
Hoang-Son Bui (Hanoi University of Science and Technology)*; Sy-Hoang Tran (School of Electrical and Electronic Engineering (SEEE), Hanoi University of Science and Technology (HUST)); Thuy-Binh Nguyen (University of Transport and Communications); Thanh-Hai Tran (School of Electrical and Electronic Engineering (SEEE), Hanoi University of Science and Technology (HUST)); Hai Vu (Hanoi University of Science and Technology); Le Lan (HUST)

57. Image, Video, and Multimedia Meeting Room 3 06-Dec

57-1 09:00-09:20
ACE-Flow: Auto Color Encoding for Enhanced Low-Light Image Restoration
Jiachen Qiu (The Hong Kong Polytechnic University); Yushen Zuo (The Hong Kong Polytechnic University); Kin-Man Lam (The Hong Kong Polytechnic University)*

57-2 09:20-09:40
PBJDT: Point-Based Joint Detection-and-Tracking
Zhen-Xun Lee (National Taiwan University); Jian-Jiun Ding (National Taiwan University)*

57-3 09:40-10:00
Capturing Dynamic Identity Features for Speaker-Adaptive Visual Speech Recognition
Sara Kashiwagi (Waseda University)*; Keitaro Tanaka (Waseda University); Shigeo Morishima (Waseda Research Institute for Science and Engineering)

57-4 10:00-10:20
A Byte-based GPT-2 Model for Bit-flip JPEG Bitstream Restoration
Hao Qin (The Hong Kong Polytechnic University); Haoran SUN (Hong Kong Polytechnic University); Yi Wang (The Hong Kong Polytechnic University)*

58. Acoustic Scene Analysis and Signal Enhancement Based on Advanced Signal Processing and Machine Learning Meeting Room 4 06-Dec

58-1 09:00-09:20
Successive Speaker Relative Transfer Function Estimation Through Relative Transfer Matrix in

Noisy Reverberant Environments
Wageesha Manamperi (University of Moratuwa)*; Thushara Abhayapala (The Australian National University)

58-2 09:20-09:40
Heavy-tailed Distributions-Based Online Semi-blind Source Separation for Nonlinear Echo Cancellation
Liyuan Zhang (Waseda University)*; Xianrui Wang (Northwestern Polytechnical University); Yichen Yang (Northwestern Polytechnical University); Tetsuya Ueda (Waseda University); Shoji Makino (Waseda University); Jingdong Chen (Northwestern Polytechnical University)

58-3 09:40-10:00
A Single-InputBinaural-Output Perceptual Rendering Based Speech Separation Method in Noisy Environments
tianqin zheng (Northwestern Polytechnical University)*; Hanchen Pei (Wuhan University); Ningning Pan (Southwestern University of Finance and Economics); Jilu Jin (Northwestern Polytechnical University); Gongping Huang (Wuhan University); Jingdong Chen (Northwestern Polytechnical University); Jacob Benesty (INRS)

58-4 10:00-10:20
Real-Time Noise Estimation for Lombard-Effect Speech Synthesis in Human–Avatar Dialogue Systems
Yuto Ishikawa (The University of Tokyo)*; Osamu Take (The University of Tokyo); Tomohiko Nakamura (National Institute of Advanced Industrial Science and Technology (AIST)); Norihiro Takamune (the University of Tokyo); Yuki Saito ("The University of Tokyo, Japan"); Shinnosuke Takamichi (Keio University); Hiroshi Saruwatari (The University of Tokyo)

59. Speech and Language Processing Meeting Room 5 06-Dec

59-1 09:00-09:20
EMO-Codec: An In-Depth Look at Emotion Preservation Capacity of Legacy and Neural Codec Models With Subjective and Objective Evaluations
Wenze Ren (National Taiwan University)*; Yi-Cheng Lin (National Taiwan University); Huang-Cheng Chou (Department of Electrical Engineering at National Tsing Hua University (NTHU)); Haibin Wu (National Taiwan University); Yi-Chiao Wu (-); Hung-yi Lee (National Taiwan University); Chi-Chun Lee (National Tsing Hua University); Hsin-Min Wang (Academia Sinica); Yu Tsao (Academia Sinica)

59-2 09:20-09:40
Analytic Study of Text-Free Speech Synthesis for Raw Audio using a Self-Supervised Learning Model
Joonyong Park (The University of Tokyo)*; Daisuke Saito (The University of Tokyo); Nobuaki Minematsu (The University of Tokyo)

59-3 09:40-10:00
Investigating the Language Independence of Voice Activity Projection Models through Standardization of Speech Segmentation Labels
Yuki Sato (Nagoya University)*; Yuya Chiba (NTT Communication Science Laboratories); Ryuichiro Higashinaka (Nagoya University)

59-4 10:00-10:20
A Preliminary Study on Analysing Mandarin Tone Values of Romance L2 Mandarin Learners
Wu-Hao Li (National Taipei University)*; Te-hsin Liu (National Taiwan University); Chen Yu CHIANG (National Taipei University)

60. Signal Processing for Drone Audition & Recent Advances in Intelligent Signal Processing Meeting Room 6 06-Dec

60-1 09:00-09:20
Successive Speaker Relative Transfer Function Estimation Through Relative Transfer Matrix in Noisy Reverberant Environments
Wageesha Manamperi (University of Moratuwa)*; Thushara Abhayapala (The Australian National University)

60-2 09:20-09:40
Beamforming informed independent low-rank matrix analysis for sound source enhancement in unmanned aerial vehicles
Jin Xuan Teh (University of Auckland)*; Norihiro Takamune (University of Tokyo); Hiroshi Saruwatari (University of Tokyo); Benjamin Yen (Tokyo Institute of Technology); Michael Kingan (University of Auckland); Yusuke Hioka (University of Auckland)

60-3 09:40-10:00
SMoLnet-T: An Efficient Complex-spectral Mapping Speech Enhancement Approach with Frame-wise CNN and Spectral Combination Transformer for Drone Audition
Zhi-Wei Tan (Nanyang Technological University)*; Andy W H Khong (Nanyang Technological University)

60-4 10:00-10:20
Integrating VGGSK and BEATs for Enhanced Sound Event Detection: A Semi-Supervised GRU-Based System with Weak Labels and Synthetic Soundscapes
Po-Cheng Chan (Advanced Technology Laboratory, Chunghwa Telecom Laboratories)*; Wei-Yu Chen (Advanced Technology Laboratory, Chunghwa Telecom Laboratories); Jia-Ching Wang (National Central University); Chung-li Lu (Advanced Technology Laboratory, Chunghwa Telecom Laboratories); Hsiang Feng Chuang (Advanced Technology Laboratory, Chunghwa Telecom Laboratories); yu-han cheng (Advanced Technology Laboratory, Chunghwa Telecom Laboratories)

60-5 10:20-10:40
Drone audition: implementation of an indoor multi-drone system for sound source tracking
Benjamin Yen (Tokyo Institute of Technology)*; Kazuhiro Nakadai (Tokyo Institute of Technology)

61. Converging AI and Computer Vision: Innovations and Potential Meeting Room 8 06-Dec

61-1 09:00-09:20
Hyperspectral Anomaly Detection Using Robust Principal Component Analysis with Autoencoding Adversarial Networks
Atsuya Emoto (The University of Kitakyushu); Ryo Matsuoka (The University of Kitakyushu)*

61-2 09:20-09:40
Optimising Neural Networks with Fine-Grained Forward-Forward Algorithm: A Novel Backpropagation-Free Training Algorithm
James Gong (The University of Auckland); Bruce Li (The University of Auckland); Waleed Abdulla ("University of Auckland, New Zealand")*

61-3 09:40-10:00
Two-Way Malaysian Sign Language Communication System for Inclusive Education
Veron Zhen Liang HII (Swinburne University of Technology Sarawak Campus); Aaron Ken Kiat LO (Swinburne University of Technology Sarawak Campus); Ida Pei Xin LEE (Swinburne University of Technology Sarawak Campus); ALEC VINCE GONZALES ABUAN (Swinburne University of Technology Sarawak Campus); Sue Han Lee (Swinburne University of Technology Sarawak Campus)*; Patrick HangHui Then (Swinburne University of Technology Sarawak Campus)

61-4 10:00-10:20
PRTGaussian: Efficient Relighting Using 3D Gaussians with Precomputed Radiance Transfer
Libo Zhang (Tsinghua University)*; Yuxuan Han (Tsinghua University); Wenbin Lin (Tsinghua University);

Jingwang Ling (Tsinghua University); Feng Xu (Tsinghua University)

62. AI-Driven Innovations in Cybersecurity Advanced Applications in Signal Processing, Multimedia Security, and Privacy Meeting Room 9 06-Dec

62-1 09:00-09:20
ET-SSM: Linear-Time Encrypted Traffic Classification Method Based On Structured State Space Model
Li Yanjun (University of Science and Technology of China)*; Xiangyu Zhao (University of Science and Technology of China); Zha Zhengpeng (USTC); Zhen-Hua Ling (University of Science and Technology of China)

62-2 09:20-09:40
Toward Universal Detector for Synthesized Images by Estimating Generative AI Models
Ryota Seo (Okayama University)*; Minoru Kuribayashi (Tohoku University); Akinobu Ura (Okayama University); Antoine Mallet (UTT); Rémi Cogranne (Troyes University of Technology); Wojciech Mazurczyk (Warsaw University of Technology); David Megías (Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya)

62-3 09:40-10:00
Innovative Information Hiding in H.266/VVC Using Sub-Block Transform Technique
Joan Hau (Tunku Abdul Rahman University of Management and Technology)*; Yiqi Tew (TARUC); Li Peng Tan (Tunku Abdul Rahman University of Management and Technology)

62-4 10:00-10:20
GGMDDC: An Audio Deepfake Detection Multilingual Dataset
Ravindrakumar Purohit (DA-IICT)*; Arth Shah (DA-IICT); Hemant Patil (DAICT, Gujrat)

63. Embedded and Real-Time Systems for AI and Signal Processing Applications Meeting Room 1 06-Dec

63-1 10:40-11:00
Exploration Robot Based On YOLOv8 Algorithm
Infall Syafalni (Bandung Institute of Technology)*; Angelica Winasta Sinisuka (Bandung Institute of Technology); Dwi Kalam Amal Tauhid (Bandung Institute of Technology); Farrel Ahmad (Bandung Institute of Technology); Muhammad Alif Putra Yasa (Bandung Institute of Technology); Steven Alexander Wen (Bandung Institute of Technology); Erwin Setiawan (Bandung Institute of Technology); Nana Sutisna (Institut Teknologi Bandung); Trio Adiono (Institut Teknologi Bandung)

63-2 11:00-11:20
Optimizing Deep Q-Network for Shortest Path Computation of Mobile Robot Agents
A Sumarudin (ITB)*; Nana Sutisna (Institut Teknologi Bandung); Infall Syafalni (Bandung Institute of Technology); Bambang Riyanto Trilaksono (ITB); Trio Adiono (Institut Teknologi Bandung)

63-3 11:20-11:40
Leveraging IoT and Machine Learning for Efficient Rice Stock Monitoring and Prediction
Nana Sutisna (Institut Teknologi Bandung)*; Aditya Prawira Nugroho (Institut Teknologi Bandung); Christopher Jeffrey (Institut Teknologi Bandung); Rizky Ramadhana (Institut Teknologi Bandung); Ronggur Mahendra (Institut Teknologi Bandung); Michael Jonathan (Institut Teknologi Bandung); Infall Syafalni (Institut Teknologi Bandung); Trio Adiono (Institut Teknologi Bandung)

63-4 11:40-12:00
Comparative Evaluation of Fine-Tuned Hybrid Transformer and Band-Split Recurrent Neural Networks for Music Source Separation
Ken Kalang Al Qalyubi (Bandung Institute of Technology); Nur Ahmadi (Bandung Institute of Technology)*;

Dessi Puji Lestari (Bandung Institute of Technology)

64. Selected Papers from APSIPA Workshop on Advanced Signal and Information Processing Meeting Room 2 06-Dec

64-1 10:40-11:00
Enhancing Shear Wave Propagation Analysis in Tissue with Directional Filtering of Reflected Waves
Hai Luong (Le Quy Don Technical University)*; Nghia Tran (Institute of Information Technology, Vietnam Academy of Science and Technology); Hiep Nguyen (Thai Nguyen University of Information and Communication Technology); Lam Sinh Cong (VNU University of Engineering and Technology); Duc-Tan Tran (Phenikaa University)

64-2 11:00-11:20
Structural Analysis of Asian and African Rice Panicles via Transfer Learning
Tran Hiep Dinh (VNU University of Engineering and Technology)*

64-3 11:20-11:40
New approach for Alzheimer's disease classification using topographic maps and deep learning model
Quoc Anh Le (VNU University of Engineering and Technology)*; Nguyen Thinh (UET-VNU)

64-4 11:40-12:00
M-IRRA: A Multilingual Model for Text-based Person Search
Phong Tran (Hanoi University of Science and Technology); Thi-Hoai Phan (Hanoi University of Science and Technology); Thuy-Binh Nguyen (University of Transport and Communications); Ngoc-Diep Do (Hanoi University of Science and Technology); Quân Nguyen (Viet-Hung Industrial University); Thanh-Hai Tran (Hanoi University of Science and Technology); Thanh Duong (Hanoi University of Mining and Geology); Thi Lan Le (HUST)*

65. Image, Video, and Multimedia Meeting Room 3 06-Dec

65-1 10:40-11:00
GMNER-LF: Generative Multi-modal Named Entity Recognition Based on LLM with Information Fusion
Huiyun Hu (Beijing University of Posts and Telecommunications)*; Junda Kong (Beijing University of Posts and Telecommunications); Bo Xiao (Beijing University of Posts and Telecommunications); Fei Wang (State Grid Shandong Electric Power Company Dezhou Power Supply Company); Yang Ge (State Grid Shandong Electric Power Company Dezhou Power Supply Company); Hongzhi Sun (State Grid Shandong Electric Power Company Dezhou Power Supply Company)

65-2 11:00-11:20
WildPose: HRNet-based Lightweight and Efficient Wildlife Pose Estimation
SIBUSISO BAKANA (Beihang University)*; Yongfei Zhang (Beihang University); Bhekisipho Twala (Tshwane University of Technology)

65-3 11:20-11:40
New approach on Smiling faces with Domain Transfer in Latent Space
Wan-Chi Siu (The Hong Kong Polytechnic University)*; Mingfei DUAN (Saint Francis University); Chun Chuen Hui (The Hong Kong Polytechnic University)

66. Advanced Topics for Automatic Speakers Recognition Meeting Room 4 06-Dec

66-1 10:40-11:00
JOSEPH: PHONETIC-AWARE SPEAKER EMBEDDING FOR FAR-FIELD SPEAKER VERIFICATION
Zezhong JIN (The Hong Kong Polytechnic University)*; Youzhi TU (The Hong Kong Polytechnic University); Manwai Mak (The Hong Kong Polytechnic University)

66-2 11:00-11:20
Vocal Tract Length Perturbation-based Pseudo-Speaker Augmentation Considering Speaker Variability for Speaker Verification
Hengyi Zou (Tokyo Metropolitan University)*; Sayaka Shiota (Tokyo Metropolitan University)

66-3 11:20-11:40
Differences Between Singer and Speaker Verification: Training Singer Feature Representation Extractor Utilizing Singing Voice Characteristics
Sayaka Toma (Waseda University)*; Tomoki Ariga (Waseda University); Yosuke Higuchi (Waseda University); Ichiju Hayasaka (DAIICHIKOSHO CO., LTD.); Rie Shigyo (DAIICHIKOSHO CO., LTD.); Tetsuji Ogawa (Waseda University)

67. Speech and Language Processing Meeting Room 5 06-Dec

67-1 10:40-11:00
Developing a Multilingual Spontaneous L2 Speech Corpus for Automated Proficiency Assessment
Seunghee Han (Seoul National University)*; Sunhee Kim (Seoul National University); Minhwa Chung (Seoul National University)

67-2 11:00-11:20
Prediction of Negative User Reactions Towards System Responses During Attentive Listening
Divesh Lala (Kyoto University)*; Koji Inoue (Kyoto University); Tatsuya Kawahara (Kyoto University)

67-3 11:20-11:40
Data Selection using Spoken Language Identification for Low-Resource and Zero-Resource Speech Recognition
Jianan Chen (Kyoto University)*; Chenhui Chu (Kyoto University); Sheng Li (National Institute of Information & Communications Technology (NICT)); Tatsuya Kawahara (Kyoto University)

68. Signal Processing for Drone Audition & Recent Advances in Intelligent Signal Processing Meeting Room 6 06-Dec

68-1 10:40-11:00
Drone audition: dataset and methods for ground surface material classification using drone noise in outdoor environment
Tsubasa Yano (Tokyo Institute of Technology)*; Benjamin Yen (Tokyo Institute of Technology); Kazuhiro Nakadai (Tokyo Institute of Technology)

68-2 11:00-11:20
Seismic-ionospheric Precursor Prediction Using Deep Learning
Tung Pham (National Central University)*; Pao-Chi Chang (Department of Communication Engineering, National Central University); Jia-Ching Wang (National Central University)

68-3 11:20-11:40
Swarm Active Audition System with Robots and Drones for a Search and Rescue Task
Kazuhiro Nakadai (Tokyo Institute of Technology)*; Makoto Kumon (Kumamoto University); Yoko Sasaki

(National Institute of Advanced Industrial Science and Technology); Kotaro Hoshiba (Tokyo Institute of Technology); Benjamin Yen (Tokyo Institute of Technology)

68-4 11:40-12:00
Implementation of a Robot Operation System-based network for sound source localization using multiple drones
Takumi Yamamoto (Tokyo Institute of Technology)*; Kotaro Hoshiba (Tokyo Institute of Technology); Benjamin Yen (Tokyo Institute of Technology); Kazuhiro Nakadai (Tokyo Institute of Technology)

69. Converging AI and Computer Vision: Innovations and Potential Meeting Room 8 06-Dec

69-1 10:40-11:00
RepViT Based Lightweight Architecture for Distracted Driving Detection
Muwei Jian (Shandong University of Finance and Economics)*; Yukun Ling (Shandong University of Finance and Economics)

69-2 11:00-11:20
HSIC as Information Compression for Training Deep Neural Network
Roshan Birjais Sofi (The University of Auckland)*; Kevin I-Kai Wang (The University of Auckland); Waleed Abdulla ("University of Auckland, New Zealand")

69-3 11:20-11:40
Zero-Shot Learning for Haze Removal Using Fusion of Near-Infrared and Color Images
Onhi Kato (Chuo University)*; Akira Kubota (Chuo University)

69-4 11:40-12:00
Color Enhancement for the Colorblind Using Color Correction Intensity Map and Pix2pix Image Conversion
Shu Komatsu (Chuo University)*; Akira Kubota (Chuo University)

70. Multimedia Processing Systems in the AI Era Meeting Room 9 06-Dec

70-1 10:40-11:00
Detecting Abnormal Machine Sounds Using An Ensemble Approach with Data Augmentation Techniques
Po-Cheng Chan (Advanced Technology Laboratory, Chunghwa Telecom Laboratories)*; Chung-li Lu (Advanced Technology Laboratory, Chunghwa Telecom Laboratories); Jia-Ching Wang (National Central University)

70-2 11:20-11:40
Leveraging Attention Mechanisms for Breast Cancer Diagnosis
Brahma reddy akumalla (National Central University)*; Tung Pham (National Central University); Yung-Yu Zhuang (National Central University); Bima Prihasto (Institut Teknologi Kalimantan); Pao-Chi Chang (National Central University); Jia-Ching Wang (National Central University)

70-3 11:40-12:00
Enhanced Detection of Illegally Parked Vehicles Using YOLO and Good Feature to Track Methods
Fauzan Maftuh Alwafi (Institut Teknologi Kalimantan); Boby Mugi Pratama (Institut Teknologi Kalimantan); Phuong Thi Le (Fu Jen Catholic University); Bima Prihasto (Institut Teknologi Kalimantan)*; Jia-Ching Wang (National Central University)

71. Few-shot Vision, Language, and Multimedia Processing under LLMs Meeting Room 10 06-Dec

71-1 10:40-11:00
A Noisy Context Optimization Approach for Chinese Spelling Correction
Guangwei Zhang (Beijing University Of Posts and Telecommunications); Yongping Xiong (Beijing University of Posts and Telecommunications); Ruifan Li (Beijing University of Posts and Telecommunications)*

71-2 11:00-11:20
GVDIE: A Zero-Shot Generative Information Extraction Method for Visual Documents Based on Large Language Models
Siyang Qi (Beijing University of Posts and Telecommunications)*; Fei Wang (State Grid Shandong Electric Power Company Dezhou Power Supply Company); Hongzhi Sun (State Grid Shandong Electric Power Company Dezhou Power Supply Company); Yang Ge (State Grid Shandong Electric Power Company Dezhou Power Supply Company); Bo Xiao (Beijing University of Posts and Telecommunications)

71-3 11:20-11:40
META: Text Detoxification by leveraging METAmorphic Relations and Deep Learning Methods
Alika Choo (Monash University Malaysia)*; Arghya Pal (Monash University); Sailaja Rajanala (Monash University Malaysia); Arkendu Sen (Jeffrey Cheah School of Med & Health sc, Monash University Malaysia)

71-4 11:40-12:00
Visual semantic alignment network based on pre-trained ViT for few-shot image classification
Jiaming Zhang (Lanzhou University of Technology); Jijie Wu (Lanzhou University of Technology); Xiaoxu Li (Lanzhou University of Technology)*

72. Poster Foyer 04-Dec

72-1 11:00-12:20
Speech Depression Recognition from the Self-reference Effect Using LSTM with ResNet
Xiaoyong Lu (Northwest normal university)*

72-2 11:00-12:20
Temporal-Spatial Correlation Analysis for Ship-Radiated Noise Based on Random Matrix Theory
Qing Feng (PKU-Wuhan Institute for Artificial Intelligence)*; Zhiqiang Wu (PKU-Wuhan Institute for Artificial Intelligence); Xuebin Li (Jilin Engineering Normal University); Heping Shen (Wuhan Puhui Ocean Opto-Electronic Technology Co., Ltd.); Shang Liu (PKU-Wuhan Institute for Artificial Intelligence); Min Tang (PKU-Wuhan Institute for Artificial Intelligence); Quansheng Feng (PKU-Wuhan Institute for Artificial Intelligence)

72-3 11:00-12:20
Annotation-free Fine-tuning for Unsupervised Anomalous Sound Detection
Kai Guo (Beijing Institute of Technology)*; Xiang Xie (Beijing Institute of Technology); Fengrun Zhang (Beijing Institute of Technology)

72-4 11:00-12:20
Knowledge Augmented Attention Gating Embedding for Link Prediction
Zewei Chen (Shantou University); Chen Shuhong (Shantou University); Chen Li (Foshan University); Xianwei Zheng (Foshan University)*; Minfan He (Foshan University); xutao li (Shantou University)

72-5 11:00-12:20
Detecting Coronary Artery Stenosis from Cardiac CT Images using 3D CNNs
Masaki Aono (Toyohashi University of Technology)*

72-6 11:00-12:20
Effective Speech Data Augmentation Method To Improve Customer Service Representative Speech Recognition System Performance

Huiyong Bak (Lotte innovate)*; Changhyeon Jeong (Lotte innovate)

72-7 11:00-12:20
LLM as decoder: Investigating Lattice-based Speech Recognition Hypotheses Error Correction Using LLM
Sheng Li (National Institute of Information & Communications Technology (NICT))*; Yuka Ko (Nara Institute of Science and Technology); Akinori Ito (Tohoku University)

72-8 11:00-12:20
A Two-Stage Wall Parameters Estimation Algorithm for MIMO Through-the-Wall Radar
Zhirun Li (University of Electronic Science and Technology of China)*; Shisheng Guo (University of Electronic Science and Technology of China); Jiahui Chen (University of Electronic Science and Technology of China); Zhihao Zhu (University of Electronic Science and Technology of China); Chen Qiu (University of Electronic Science and Technology of China); Cui Guolong (UESTC); Yutao Xiang (University of Electronic Science and Technology of China)

72-9 11:00-12:20
Tiny Object Detection Enhancement for Large-Scale Remote Sensing Imagery
Tianwei Zhang (Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China)*; Lianru Gao (Aerospace Information Research Institute, Chinese Academy of Sciences)

72-10 11:00-12:20
Robust Watermarking via Dual Guidance
Yuhang Zhang (Shenzhen University); Yuanman Li (Shenzhen University)*; Li Dong (Ningbo University); Xia Li (Shenzhen University)

72-11 11:00-12:20
Region Aware Framework for Constrained Image Splicing Detection and Localization
Haokun Cao (Shenzhen University); Yuanman Li (Shenzhen University)*; Xia Li (Shenzhen University)

73. Poster Foyer 04-Dec

73-1 14:00-16:00
Handling Missing Data in Limited-View Photoacoustic Tomography Using Compressive Sensing Algorithm-Based Deep Learning
Mary John (UAE University); Imad Barhumi (United Arab Emirates University)*

73-2 14:00-16:00
Keyword spotting for dialectal speech and Introduction of wav2vec2.0
Tomohiro Ariga (Iwate Prefecture University)*; Reo Minakawa (Graduated from Iwate Prefectural University); Yoshiaki Itoh (Iwate Prefectural University); Shi-wook Lee (National Institute of Avanced Industrial Science and Technology); Kazunori Kojima (Iwate Prefecture University)

73-3 14:00-16:00
LCMV-based Scan-and-Sum Beamforming for Region Source Extraction
Aoto Yasue (Tokyo Institute of Technology)*; Benjamin Yen (Tokyo Institute of Technology); Katsutoshi Itoyama (Honda Research Institute Japan Co., Ltd.); Kazuhiro Nakadai (Tokyo Institute of Technology)

73-4 14:00-16:00
Run-Time Adaptation of Neural Beamforming for Robust Speech Dereverberation and Denoising
Yoto Fujita (Kyoto University)*; Aditya Arie Nugraha (RIKEN); Diego Di Carlo (RIKEN); Yoshiaki Bando (National Institute of Advanced Industrial Science and Technology); Mathieu Fontaine (Télécom Paris); Kazuyoshi Yoshii (Kyoto University)

73-5 14:00-16:00
Performance Evaluation of Acoustic Echo and Noise Canceller with Variable-Step-Size Shared-Error NLMS Algorithm under Double-Talk Conditions

Kenta Iwai (Ritsumeikan University)*; Takanobu Nishiura (Ritsumeikan University)

73-6 14:00-16:00
Augmented sound-image perception using pre-virtual-leading ultrasounds based on precedence effect
Ryota Imanaka (Ritsumeikan University)*; Yuting Geng (Ritsumeikan University); Masato Nakayama (Osaka Sangyo University); Takanobu Nishiura (Ritsumeikan University)

73-7 14:00-16:00
Virtual multi-boosted amplitude modulation toward high-pressure audible sound with parametric array loudspeakers
Yoto Ikezaki (Ritsumeikan University)*; Yuting Geng (Ritsumeikan University); Masato Nakayama (Osaka Sangyo University); Takanobu Nishiura (Ritsumeikan University)

73-8 14:00-16:00
Analyzing House Music: Relations of Audio Features and Musical Structure
Justin Tomoya Wulf (Nihon University); Tetsuro Kitahara (Nihon University)*

73-9 14:00-16:00
Teager Energy Cepstral Coefficients for Spoken Language Identification
Arth Shah (DA-IICT)*; Savita Yadav (Uka Tarsadia University); Hemant Patil (DAICT, Gujrat)

73-10 14:00-16:00
A Parameter-free model for long-term concrete creep prediction
Conghui Li (Monash university)*; Chern Hong Lim (Monash University); Xin Wang (Monash University Malaysia)

73-11 14:00-16:00
Voice Liveness Detection Using Linear Frequency Residual Cepstral Coefficients
Arth Shah (DA-IICT)*; Nandini Mandaviya (DA-IICT); Hemant Patil (DAICT, Gujrat)

73-12 14:00-16:00
An isolated Vietnamese Sign Language Recognition method using a fusion of Heatmap and Depth information based on Convolutional Neural Networks
Phuoc Nguyen (Phenikaa); Thi-Huong Nguyen (Hanoi College of High Technology); Duc-Tan Tran (Phenikaa University); Tien Son Bui (Hanoi University of Industry); VanToi Nguyen (Phenikaa University)*

73-13 14:00-16:00
GILED: Lesion Detection of Gastrointestinal Tract from Endoscopic Images and Medical Notes
Vu-An Hoang (HUST)*; Minh-Hanh Tran (HUST); Viet Hang Dao (HMU); Thanh-Hai Tran (HUST)

73-14 14:00-16:00
Wavetable Synthesis Using CVAE for Timbre Control Based on Semantic Label
Tsugumasa Yutani (University of Tsukuba)*

74. Poster Foyer 04-Dec

74-1 16:20-18:00
Gamma-VAE: Speech representation based on VAE assuming gamma distribution for both latent variables and observation
Nanako Imaichi (The Univercity of Electro-communications, Tokyo)*; Toru Nakashika (The University of Electro-Communications, Tokyo)

74-2 16:20-18:00
Does Brain Atlas Choice Matter? An Empirical Study in Alzheimer's Diagnosis Using FDG-PET Images
MINH TUAN Pham (Aix Marseille University); Mouloud Adel (Institut Fresnel); Nguyen Linh Trung (Vietnam National University, Hanoi)*; Eric Guedj (Assistance Publique Hôpitaux de Marseille)

74-3 16:20-18:00
Transformer Attention Matrix Multiplication Design using 4x4 Systolic Arrays
Muhammad Sayyid Afif (Bandung Institute of Technology)*; Infall Syafalni (Bandung Institute of Technology); Nana Sutisna (Institut Teknologi Bandung); Trio Adiono (Institut Teknologi Bandung)

74-4 16:20-18:00
Quefrency Approach to Audio Deepfake Detection
Kanishq Singhal (The LNM Institute of Information Technology); Aditya Goyal (The LNM Institute of Information Technology); Priyanka Gupta (The LNM Institute of Information Technology)*

74-5 16:20-18:00
A SEMI-SUPERVISED LOW-LIGHT IMAGE ENHANCEMENT WITH COLOR GUIDANCE
Yuxin Wang (Center for Optics Research and Engineering (CORE), Shandong University)*; Yang Yang (Shandong University)

74-6 16:20-18:00
Cloud Removal in Hyperspectral Satellite Images Using Low-rank Tensor Completion
Chuong Vo (Dongguk University)*; Mai Truong (Dongguk University); Chul Lee (Dongguk University)

74-7 16:20-18:00
Color Guided Disease Segmentation for Plant Images
Soyeon Jang (Korea University)*; Jong-Ok Kim (Korea University)

74-8 16:20-18:00
Performance Optimization in the Cascade of VAD and ASR Systems: A Study on Evaluation and Alignment Strategies
Zhentao Lin (Guangdong University of Technology); Zihao Chen (Guangdong University of Technology)*; Bi Zeng (Guangdong University of Technology); Leqi Chen (Guangdong University of Technology); Jia Cai (China Electronic Product Reliability and Environmental Testing Research Institute; Key Laboratory of MIIT for Intelligent Products Testing and Reliability)

75. Poster Foyer 05-Dec

75-1 10:20-12:00
StylebookTTS: Zero-Shot Text-to-Speech Leveraging Unsupervised Style Representation
Juhwan Yoon (Yonsei University)*; Hyungseob Lim (Yonsei University); Hyeonjin Cha (Yonsei University); Hong-Goo Kang (Yonsei University)

75-2 10:20-12:00
GENERATING PHONETIC TRANSCRIPTIONS FOR KOREAN ENGLISH L2 LEARNERS USING MULTIPLE SELF-SUPERVISED-MODEL-BASED ASR SYSTEMS AND ROVER METHOD
Jong In Kim (Seoul National University)*

75-3 10:20-12:00
Adaptive Time-Varying Graph Learning for Traffc Flow Data Based on Anomaly Moment Detection
Chen Shuhong (Shantou University); Zewei Chen (Shantou University); Chen Li (Foshan University); Xianwei Zheng (Foshan University)*; Minfan He (Foshan University); xutao li (Shantou University)

75-4 10:20-12:00
Cuisine Image Synthesis with Improved Multiscale GANs Guided by CLIP
Weiyi Xia (Hosei University)*; Satoru Fujita (Hosei University)

75-5 10:20-12:00
A Novel LLM-based Two-stage Summarization Approach for Long Dialogues
yuan jhe yin (National Taiwan Normal University)*; Bo-Yu Chen (National Taiwan University of Science and Technology); Berlin Chen (National Taiwan Normal University)

75-6 10:20-12:00
Impulse response transforming method to control distance perception based on direct-to-reverberant energy ratio
Toru Takahashi (Osaka Sangyo University)*; Masato Nakayama (Osaka Sangyo University)

75-7 10:20-12:00
Data-Driven Sound Field Reproduction for Higher-Order Mode Matching Using a Circular Loudspeaker Array
Keiko Kawase (Tokyo Denki University)*; Gen Sato (Tokyo Denki University); Izumi Tsunokuni (Tokyo Denki University); Yusuke Ikeda (Tokyo Denki University)

75-8 10:20-12:00
Layer-Wise Feature Distillation with Unsupervised Multi-Aspect Optimization for Improved Automatic Speech Assessment
Chung-Wen Wu (National Taiwan Normal University)*; Berlin Chen (National Taiwan Normal University)

75-9 10:20-12:00
Sparse Blind Deconvolution and Demixing via Block Majorization-Minimization
Mengting Chen (ShanghaiTech University)*; Ziping Zhao (ShanghaiTech University)

76. Poster Foyer 05-Dec

76-1 14:00-15:20
An End-to-End Two-Stream Network Based on RGB Flow and Representation Flow for Human Action Recognition
Songjiang Lai (CAiRS)*; Tsun-Hin Cheung (Centre for Advances in Reliability and Safety)

76-2 14:00-15:20
Learning a Sequence of Cursive-Style Japanese Characters in Classical Literary Works
Satoru Fujita (Hosei University)*; Keizo Oyama (National Institute of Japanese Literature)

76-3 14:00-15:20
Automatic Prompt Generation and Grounding Object Detection for Zero-Shot Image Anomaly Detection
Tsun-Hin Cheung (Centre for Advances in Reliability and Safety)*

76-4 14:00-15:20
Development of Simple Algorithm to Detect and Filter Motion Artifact Noise in Non-Invasive Blood Pressure (NIBP) Measurement
Trio Adiono (Institut Teknologi Bandung); Rd. Elviana La'salina Muhlis (Bandung Institute of Technology); Clarence Amadeus (Bandung Institute of Technology)*; Sindy Novaria Cicilya Sinaga (PT. Xirka Dama Persada)

76-5 14:00-15:20
MYMV: A Music Video Generation System with User-preferred Interaction
Kyungjune Lee (Yonsei University)*; Mingyu Jang (Yonsei University); Jungwoo Huh (Yonsei University); Jeonghaeng Lee (Yonsei University); Seok Keun Choi (Yonsei University); Sanghoon Lee (Yonsei University, Korea)

76-6 14:00-15:20
Text-guided Visual Prompt Tuning with Masked Images for Facial Expression Recognition
Rongkang Dong (The Hong Kong Polytechnic University)*; Cuixin Yang (The Hong Kong Polytechnic University); Kin-Man Lam (The Hong Kong Polytechnic University)

76-7 14:00-15:20
Fine-Grained Privacy-Preserving Image Retrieval in Cloud Environment
Jing Liang (JNU); Libo Wang (Hainan University); PEIYA LI (Jinan University)*

76-8 14:00-15:20
Measurement of Relative Transfer Function for Own Voice in Head-Mounted Microphone Array
Kyoka Kazama (Tokyo Metropolitan University)*; Taishi Nakashima (Tokyo Metropolitan University); Nobutaka Ono (Tokyo Metropolitan University)

76-9 14:00-15:20
Enhancing Early Plant Disease Detection: 1D to 2D Spectral Transformations
Mas Ira Syafila Mohd Hilmi Tan (Multimedia University)*; Lai-Kuan Wong (Multimedia University); Yuen Peng Loh (Multimedia University); Chih-Yang Pee (Multimedia University)

77. Poster Foyer 05-Dec

77-1 16:40-18:00
KhmerFormer: Multi-Scale CNNs-Transformer with External Attention for Ancient Khmer Palm Leaf Isolated Glyph Classification
Nimol Thuon (University of Science and Technology of China)*; Jun Du (University of Science and Technology of China)

77-2 16:40-18:00
DDPMVC: Non-parallel any-to-many voice conversion using diffusion encoder
Ryuichi Hatakeyama (The University of Electro-Communications, Tokyo)*; Toru Nakashika (The University of Electro-Communications, Tokyo)

77-3 16:40-18:00
MGVul: a Multi-Granularity Detection Framework for Software Vulnerability
Xiangyu Zhao (University of Science and Technology of China)*; Li Yanjun (University of Science and Technology of China); Zhengpeng Zha (University of Science and Technology of China); Zhen-Hua Ling (University of Science and Technology of China)

77-4 16:40-18:00
Dysarthria Severity Classification Using Phase Based Features of LP Residual
Rohini Mannepalli (Dhirubhai Ambani Institute of Information and Communication Technology)*; Aditya Pusuluri (DAIICT); Hemant Patil (DAICT, Gujrat)

77-5 16:40-18:00
A Joint Graph Signal and Laplacian Denoising Network Inspired by Majorization-Minimization
Zepeng Zhang (EPFL); Ziping Zhao (ShanghaiTech University)*

77-6 16:40-18:00
Comparative Analysis of Glottal and Vocal Tract Features in Dysarthria
Indukuri Geeta Sai Sahasra (KL University); Swapna Kadwasra (Koneru Lakshmaiah Education Foundation); Arushi Srivastava (DA-IICT)*; Aditya Pusuluri (DAIICT); Hemant Patil (DAICT, Gujrat)

77-7 16:40-18:00
Contrast-Aware DCT for Image Enhancement with JPEG Compatible Coding
Kohei Hayashi (Nagoya Institute of Technology)*; Soichiro Honda (Nagoya Institute of Technology); Hirokazu Kamei (Nagoya Institute of Technology); Yoshihiro Maeda (Shibaura Institute of Technology); Norishige Fukushima (Nagoya Institute of Technology)

77-8 16:40-18:00
Non-blind Deblurring Using Probabilistic Models and Spatial Adaptive Restoration
Chun-Lin Liao (National Taiwan University); Jian-Jiun Ding (National Taiwan University)*; Chun-Jen Shih (National Taiwan University)

77-9 16:40-18:00
Comparative Analysis of Glottal and Vocal Tract Features in Dysarthria
Indukuri Geeta Sai Sahasra (KL University); Swapna Kadwasra (Koneru Lakshmaiah Education Foundation);

Arushi Srivastava (DA-IICT)*; Aditya Pusuluri (DAIICT); Hemant Patil (DAICT, Gujrat)

77-10 16:40-18:00
Multi-band Satellite Image Analysis for Multi-label Classification
Sarah Abdul Rauf (Multimedia University); Mas Ira Syafila Mohd Hilmi Tan (Multimedia University); Yuen Peng Loh (Multimedia University)*

78. Poster Foyer 06-Dec

78-1 9:00-10:20
LoFLAT: Local Feature Matching using Focused Linear Attention Transformer
Naijian Cao (Northwestern Polytechnical University); Renjie He (Northwestern Polytechnical University)*; Yuchao Dai (Northwestern Polytechnical University); Mingyi He (Northwestern Polytechnical University)

78-2 9:00-10:20
Inference Efficient Source Separation Using Input-dependent Convolutions
Shogo Seki (CyberAgent, Inc.)*; Li Li (CyberAgent, Inc.)

78-3 9:00-10:20
High and Low Frequency Region Separation Method for Adaptive Image Expansion
Shao-Yun Luo (National Taiwan University); Kuei-Chen Chen (National Taiwan University); Jian-Jiun Ding (National Taiwan University)*; Cheng-Che Lee (National Taiwan University); Hsin-Jung Lee (National Taiwan University)

78-4 9:00-10:20
Unleashing Attributes-content Adaptation with Multi-color Spaces for Food Photo Aesthetic Assessment
Shintami Hidayati (Institut Teknologi Sepuluh Nopember)*; Muhammad Firdaus (Institut Teknologi Sepuluh Nopember); Riki Dianto (Institut Teknologi Sepuluh Nopember); Sarworsri Sarworsri (Institut Teknologi Sepuluh Nopember)

78-5 9:00-10:20
An Explainable Raman Spectral Classification Pipeline via NMF and SHAP: A Case Study of Pen Ink Colors
Pongpon Lapsatid (Sirindhorn International Institute of Technology, Thammasat University); Somrudee Deepaisarn (Sirindhorn International Institute of Technology, Thammasat University)*; Pitak Eiamchai (National Electronics and Computer Technology Center, National Science and Technology Development Agency)

78-6 9:00-10:20
Pressure Matching Using Data-Driven Estimation for Sound Fields and Transfer Functions
Koki Horikoshi (Tokyo Denki University)*; Gen Sato (Tokyo Denki University); Izumi Tsunokuni (Tokyo Denki University); Yusuke Ikeda (Tokyo Denki University)

78-7 9:00-10:20
Acoustic model adaptation in noisy and reverberated scenarios using multi-task learned embeddings
Aditya Raikar (Cerence Services India LLP); Meet Soni (Rakuten Institute of Technology); Ashish Panda (TCS Innovation Labs, Mumbai, India)*; Sunil Kumar Kopparapu (TCS Research)

78-8 9:00-10:20
Generalized SpecAugment: Robust Online Augmentation Technique for End-to-End Automatic Speech Recognition
Meet Soni (Rakuten Institute of Technology); Ashish Panda (TCS Innovation Labs, Mumbai, India)*; Sunil Kumar Kopparapu (TCS Research)

78-9 9:00-10:20
ComplexFace: A Public Visible-Thermal Face Dataset with Real-Life Complexity
Jiajin He (Kunming University of Science and Technology)*; Chengxi Dong (Kunming University of Science and

Technology); Yunqi Cai (Kunming University of Science and Technology); Dong Wang (Tsinghua University)

78-10 9:00-10:20
Development and Evaluation of a Semi-autonomous Parallel Attentive Listening System
Divesh Lala (Kyoto University)*; Koji Inoue (Kyoto University); Haruki Kawai (Kyoto University); Zi Haur Pang (Kyoto University); Mikey Elmers (Kyoto University); Tatsuya Kawahara (Kyoto University)

79. Poster Foyer 06-Dec

79-1 10:40-12:00
PPHiFi-TTS: Phonetic Preserved High-Fidelity Text-to-Speech for Long-Term Speech Dependencies
Ravindrakumar Purohit (DA-IICT)*; Dharmendra Vaghera (DA-IICT Gandhinagar); Arth Shah (DA-IICT); Hemant Patil (DAICT, Gujrat)

79-2 10:40-12:00
Noise-Robust Estimation of Early-part Room Impulse Responses based on Physics-Informed Neural Network with Dynamic Pulling Method
Ken Kurata (Tokyo Denki University)*; Gen Sato (Tokyo Denki University); Izumi Tsunokuni (Tokyo Denki University); Yusuke Ikeda (Tokyo Denki University)

79-3 10:40-12:00
Cross Lingual Speech Representation for Infant Cry Classification
Hiya Chaudhari (DA-IICT)*; Arth Shah (DA-IICT); Hemant Patil (DAICT, Gujrat)

79-4 10:40-12:00
Data-Driven Physics-Informed Neural Network for Sound Field Estimation in Rooms of Arbitrary Size
Gen Sato (Tokyo Denki University)*; Yusuke Ikeda (Tokyo Denki University)

79-5 10:40-12:00
GPGAN-VC: Enhancing Voice Conversion using Gradient Penalty
Ravindrakumar Purohit (DA-IICT)*; Dharmendra Vaghera (DA-IICT Gandhinagar); Hemant Patil (DAICT, Gujrat)

79-6 10:40-12:00
Improved Cassava Plant Disease Classification with Leaf Detection
Ming Xuan Chai (Multimedia University); Yao Deng Fam (Multimedia University); Quinito Norman Octaviano (Multimedia University); Chih-Yang Pee (Multimedia University)*; Lai-Kuan Wong (Multimedia University); Mas Ira Syafila Mohd Hilmi Tan (Multimedia University); John See (Heriot-Watt University Malaysia)

79-7 10:40-12:00
Teager Energy Cepstral Coefficients for Spoken Language Identification
Arth Shah (DA-IICT)*; Savita Yadav (Uka Tarsadia University); Hemant Patil (DAICT, Gujrat)

79-8 10:40-12:00
Agent Attention Feature Reconstruction Network for Fine-Grained Few-Shot Image Classification
Dongfei Chang (Lanzhou University of Technology); Jijie Wu (Lanzhou University of Technology); Xiaoxu Li (Lanzhou University of Technology)*

80. Tutorial Meeting Room 6 03-Dec

80-1 09:30-11:30
[T01] EEG Signal Processing and Machine Learning
Saeid (Saeed) Sanei

80-2 13:00-15:00
[T03] Human-Centric RF Sensing: Pose Estimation, ECG Monitoring and Self-Supervised Learning
Yan Chen, Dongheng Zhang, Zhi Lu

80-3 15:30-17:30
[T04] Emerging Topics for Speech Synthesis: Versatility and Efficiency
Yuki Saito, Shinnosuke Takamichi, Wataru Nakata

81. Tutorial Meeting Room 9 03-Dec

81-1 09:30-11:30
[T02] From Statistical to Causal Inferences for Time-Series and Tabular Data
Pavel Loskot

82. Winter School Meeting Room 9 03-Dec

82-1 13:00-14:00
Overview of Neural Network AI
Mingyi He

82-2 14:00-15:00
Hopfield Neural Network Fundamental for Machine Learning
Mingyi He

82-3 15:30-16:30
Deep Learning for Image forensics
Bonnie Law

82-4 16:30-17:30
Generative Modeling and Learning for Conversational AI
Jen-Tzung Chien

83. Keynote Merged Room 16+17 04-Dec

83-1 09:40-10:40
Rate-Distortion Optimization in Video/Image Compression: From Temporal Dependency Formulation to Learning-based Modeling
Zhu Ce

83-2 09:00-10:00
Learning from Unreliable Sources via Crowdsourcing
Georgios Giannakis

83-3 15:40-16:40
AI and Cognitive Health
Helen Meng

84. Women's Forum Meeting Room 6 05-Dec

84-1 12:20-12:40
Engineering Her Future, Engineering Our Future
Helen Meng

84-2 12:40-13:00
My working life as a women in Engineering
Sansanee Auephanwiriyakul

84-3 13:00-13:20
A few suggestions for our young women professionals
Hong (Vicky) Zhao

85. Industrial Forum Meeting Room 1 04-Dec

85-1 14:00-14:35
Research, Clinical, and Business Challenges of AI and Machine Learning Applications in Medicine – A Case Study in Metastatic Cancer and Infectious Diseases Detections by Microscopic Imaging.
Yusen Eason Lin

85-2 14:35-15:10
Smart Rings: Pioneering Biomedical Technologies for Transformative Healthcare Applications
Hao Wu

85-3 15:10-15:45
An Industry Perspective: Video Analytics meets Generative AI
Jianquan Liu

85-4 15:45-16:00
Panel Discussion: AI Frontiers: From Cloud to Edge and Biomedical
-

GMA: Green Multi-Modal Alignment for Image-Text Retrieval

Tsung-Shan Yang Yun-Cheng Wang Chengwei Wei Suya You C.-C. Jay Kuo
University of Southern California, USA
{tsungsha, yunchenw, chengwei, suya, jckuo}@usc.edu

***Abstract*—Image-text retrieval is a fundamental task in image understanding. This task aims to retrieve the most relevant information from another modality based on the given image or text. Recent approaches focus on training large neural networks to bridge the gap between visual and textual domains. However, these models are computationally expensive and not explainable regarding how the data from different modalities are aligned. End-to-end optimized models, such as large neural networks, can only output the final results, making it difficult for humans to understand the reasoning behind the model's predictions. Hence, we propose a green learning solution, Green Multi-Modal Alignment (GMA), for computational efficiency and mathematical transparency. We reduce trainable parameters to 3% compared to fine-tuning the whole image and text encoders. Experimental results show that our model can outperform the SOTA retrieval models in text-to-image and image-to-text retrieval on the Flick30k and MS-COCO datasets. Besides, our alignment process can incorporate visual and text encoder models trained separately and generalize well to unseen image-text pairs.**

***Index Terms*—Image-text retrieval, Multimodal Alignment, Green Learning, Image Understanding.**

I. Introduction

Image-text retrieval is a fundamental step of image understanding in computer vision. Real-world information can be represented as an array of pixels or textual descriptions. Texts and images are the interfaces to access stored knowledge. The image-text retrieval task aims to find the most relevant counterparts of the image-text pairs given either an image or a textual description. Fig. 1 shows an example of an image and its paired textual descriptions.



1. [x] A damaged building has a an excavator in front of it.
2. [x] A bulldozer works to demolish a decrepit building; in the background, another brick building waits for its demise, its face covered with a grid of blackened window-holes.
3. [x] Construction equipment at work.
4. [x] Heavy machinery in a construction zone.
5. [x] A crane operates amidst piles of rubble.
6. [x] A yellow construction vehicle is posed near two buildings, its arm engaged with a pile of rubble.
7. [HIT] A man in a yellow coat is on a blue industrial crane working on the side of a tall building.
8. [HIT] A worker in a yellow jacket is hoisted up high to work on a building.
9. [x] A man using a bulldozer on a construction site.
10. [HIT] A Heavy machine lifting up a worker.

Fig. 1. The example of image-to-text retrieval. By giving an image, we need to retrieve the paired captions from the candidate set.

One of the challenges in image-text retrieval is explainability. Humans will expect a complete reasoning procedure instead of a magic answer from the model. However, the complicated models hide the reasoning procedures in the numerical latent spaces, and the nonlinearities in the model make the whole inference process a black box. To this end, we propose a stagewise methodology, dividing the retrieval process into three stages: 1) Global Alignment, 2) Image Cluster Alignment, and 3) Text Cluster Alignment. Each alignment stage consists of three modules: a) alignment, b) subdomain clustering, and c) subdomain feature selection. More fine-grained information can be revealed in the module's feature selection process.

The second issue is the availability of paired image and text data. It is critical to have high-quality pairs in both domains when we train the dual encoders jointly. However, most datasets only contain high-quality data in a single modality. For example, ImageNet [1] and MS-COCO [2] contain diverse images but lack sentence-level textual descriptions associated with the images. On the contrary, in the textual datasets, the BooksCorpus (800M words) [3] and English Wikipedia (2,500M words) contain high-quality paragraphs yet without corresponding images. Instead of jointly training text and image encoders from scratch, we adopt the pre-trained encoders in the image and text domains. Then, we proposed a green-learning alignment process to deal with the lack of paired information.

We propose a new scheme called Green Multi-Modal Alignment (GMA). The method utilizes the frozen image and text encoder models and aligns the representations using the proposed alignment process. Our contributions are summarized as follows:

- Instead of fine-tuning the pre-trained encoders, we design the stagewise alignment procedure. The number of trainable parameters is around 3% compared to fine-tuning the whole encoders, making our model computationally efficient.
- The modularized design provides explainability in the retrieval process. The task can be divided into subproblems. Since the token importance can defined in the clustering and feature selection modules, we can understand the alignment statistically within the stages.
- The stagewise alignments are linear projections without any nonlinearity. Thus, the alignment process can be easily reversed from one to another.
- We conduct extensive experiments on two public multi-

modal datasets. The results demonstrate that our method can significantly improve the performance in text-to-image retrieval.

II. Related Work

The existing methods can be divided into two categories: 1) Cross-Modal retrieval and 2) Visual-Language models (VLMs). Cross-modal models consist of Convolutional Neural Networks (CNNs) for extracting features from images and Recurrent Neural Networks (RNNs) for processing text data. On the other hand, VLMs employ Large Language Models (LLMs) that work in tandem with the visual foundational models for optimal performance.

A. Cross-Modal Retrieval

The cross-modal retrieval algorithms consist of feature extractors and representation matching. Zheng et al. [4] adopt deep CNN as the backbone to extract the image features and deep RNN as the backbone to obtain text features. The instance loss optimizes the two feature extractors, which can project the representations from different modalities onto the joint latent space. Lee et al. [5] utilize bottom-up attention object detector [6] to obtain semantic representations of images and conduct the word-level matching in the captions.

Liu et al. [7] formulate the information as a graph and adopt the structural matching to retrieve the closest sub-graph. The finer image features can be obtained from the region of interest, and the finer word representations can be formulated from the part-of-speech(pos) tagging [8].

As the metric learning, Hadsell et al. [9] propose the idea of contrastive learning. The objective function aims to increase the distance between unpaired image and text representations while reducing the distance between paired representations. However, pairwise optimization relies heavily on the quality of paired data. We align two pre-trained encoders trained separately in the text and image domains to overcome the need for paired data.

B. Visual-Language Model

Transformers [10] have significantly succeeded in natural language processing and computer vision tasks. The image-text encoders can share similar architectures. CLIP [11] demonstrates the impressive visual representations jointly trained with the paired text descriptions. The model uses a contrastive learning scheme to project image and text representations onto a shared latent space. This shared space allows for a better understanding of the relationship between the two modalities. The dual-encoder(image-text) architecture is prevalent in multimodal applications.

Despite achieving state-of-the-art performance, large visual-language pre-trained models still have shortcomings in inference. The matching process is not transparent, and humans can't understand the decision-making within the fully connected layers as they lack semantic meanings. Apart from explainability, the jointly fine-tuning process is computationally expensive. To handle the transparency and efficiency, we introduce the Green Learning Alignment algorithm, which uses separately pre-trained image-text encoders. The idea of Green Learning was proposed by Kuo et al. [12] and aims to reduce the computational cost of backpropagation while providing a theoretically explainable learning process for various applications.

III. Method

The algorithm can be divided into three stages: 1) Global Alignment, 2) Image Cluster Alignment, and 3) Text Cluster Alignment. We adopt the stage-wise approach to approximate the complicated decision-making process rather than building the visual-language fundamental models from scratch. Starting from the pre-trained image and text feature extractors, we keep the pre-trained model frozen to maintain its ability to generalize with unpaired data in the matching process. We align the representations by training additional one-layered adaptor matrices to project the representations onto the joint latent space. Precisely, the alignment process consists of three modules: a) alignment, b) sub-domain clustering, and c) sub-domain feature selection, where sub-domain clustering and feature selection are conducted in both image and text domains, as shown in Fig.2.

A. Alignment

In the alignment process, we do not fine-tune the pre-trained encoders. We train a lightweight linear transformation in the visual and textual domains to align the two representation spaces. The visual and text embeddings can be formulated as:

$$\begin{aligned} e_{vis} &= \mathcal{F}(\text{Image}) \in \mathbf{R}^{d_{vis}} \\ e_{txt} &= \mathcal{G}(\text{Caption}) \in \mathbf{R}^{d_{txt}}, \end{aligned} \tag{1}$$

where e_{vis}, e_{txt} are the image and text embeddings, $\mathcal{F}, \mathcal{G}$ are the frozen image and text encoder models, and d_{vis}, d_{txt} are the dimensions of the image and text representations. With the deterministic representations, the matching process can be denoted as:

$$sim(Ae_{vis}, Be_{txt}) = sim(z_{vis}, z_{txt}), \tag{2}$$

where $A \in \mathbf{R}^{d_{joint} \times d_{vis}}$ and $B \in \mathbf{R}^{d_{joint} \times d_{txt}}$ represent the trainable image-text alignment matrices, $z \in \mathbf{R}^{d_{joint}}$ represents the vector in the joint space, and $sim(.,.)$ represents the similarity metric. We adopt the cosine similarity as the similarity metric, namely $sim(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$. We can further optimize the trainable parameters with the contrastive learning loss function [13].

$$\mathcal{L}_{con} = -\log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} exp(sim(z_i, z_k)/\tau)}. \tag{3}$$

Here, (i, j) denotes the paired image and sentence in the sampled batch, N denotes the batch size, and $\tau \in \mathbf{R}$ denotes the temperature hyperparameter. $\mathbb{1} \in \{0, 1\}$ is an indicator function, and the value is one while $[k \neq i]$. The objective function collects the representations of paired samples and pushes apart the distances of unpaired samples in the latent space.

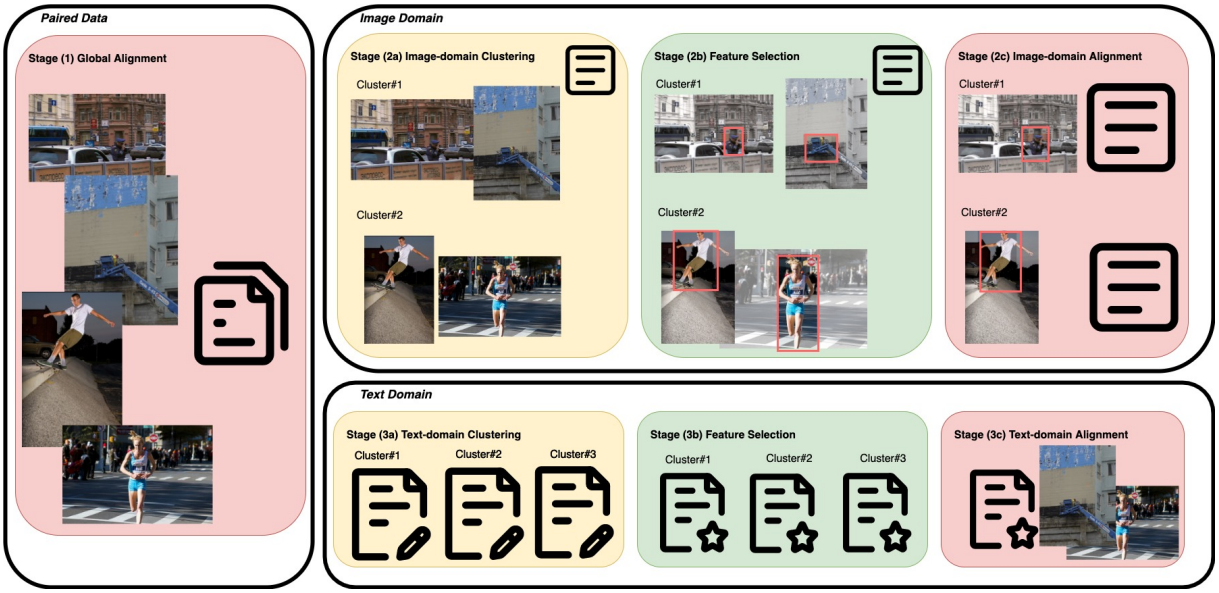


Fig. 2. The overall algorithm design of Alignment. The first stage is the global alignment. The second and third stages include fine-grained clustering and image- and text-domain feature selections.

Hence, the problem can be formulated as an optimization problem, and all the transformations are linear. We can define the inverse projection in the joint latent space without nonlinearities.

$$\begin{aligned} e_{vis} &= A^{-1} z_{vis} \\ e_{txt} &= B^{-1} z_{txt}, \end{aligned} \tag{4}$$

where $A^{-1} \in \mathbf{R}^{d_{vis} * d_{joint}}$ and $B^{-1} \in \mathbf{R}^{d_{txt} * d_{joint}}$ represent the inverse transformation from the joint space to the original image-text representations. We define the reconstruction loss for both the image and text modality.

$$\begin{aligned} \mathcal{L}_{recon} = &\left\| A^{-1} z_{vis} - e_{vis} \right\|_2 \\ &+ \left\| B^{-1} z_{txt} - e_{txt} \right\|_2, \end{aligned} \tag{5}$$

The alignment process is a linear transformation conducted by the A and B matrices. The objective function can be denoted as:

$$\mathcal{L} = \alpha \mathcal{L}_{con} + \beta \mathcal{L}_{recon}, \tag{6}$$

where α, β, and $\gamma \in \mathbf{R}$ represent hyperparameters in training. The linear alignment provides the invertible transformation from the image-text modality to the joint latent space and vice versa. However, the single-layered alignment is too simple to match all the samples. Thus, we cluster the data to form subdatasets and utilize the stagewise alignments for the detailed decision.

B. Sub-domain Clustering

With the alignment process, we can find similar representations by linear transformations. However, the transformation can only take the global representations, which means that the images/captions are represented as d_{vis}-/d_{txt}-dimensional vectors. In previous research, fine-grained information was also crucial in information-matching tasks. The image/sentence representations are the pooled output from the tokens in the prevailing transformer models.

Due to the complexity of the fine-grained token representations, it is challenging to train the tokenwise alignment in a brutal force manner. Thus, we adopt the clustering algorithms and use the clustering results to obtain the crucial tokens. The crucial token selection will be introduced in section III-C. We can reduce the feature dimension from the number of tokens and perform a second-stage alignment.

The clustering is based on the KMeans algorithm. To ensure the consistency of alignment and clustering, we use the l_2-norm of normalized representations as the distance metric.

$$\|\tilde{u} - \tilde{v}\|_2^2 = \|\tilde{u}\|_2^2 + \|\tilde{v}\|_2^2 - 2\tilde{u}\tilde{v} = 2 - 2sim(u, v), \tag{7}$$

where $\tilde{u}$ and $\tilde{v}$ are the normalized representations, namely $\tilde{u} = \frac{u}{\|u\|}$. The clustering probability can be denoted as

$$\begin{aligned} prob(u \in clus_i) &= \frac{e^{\epsilon'(2-2sim(u,cen_j))}}{\sum_{j=1}^{K} e^{\epsilon'(2-2sim(u,cen_i))}} \\ &= \frac{e^{\epsilon \cdot sim(u,cen_i)}}{\sum_{j=1}^{K} e^{\epsilon \cdot sim(u,cen_j)}}, \end{aligned} \tag{8}$$

where $clus_i$ and cen_i represent the i-th cluster and i-th centroid vector, respectively. K represents the number of clusters and ϵ is a hyperparameter. If the ϵ increases, the probability distribution will concentrate on a certain class. If the ϵ decreases, the probability distribution will become uniform.

C. Feature Selection

Clustering results provide pseudo-labels for further feature selections. We adopt Discriminant Feature Selection [14] (DFT) to select informative and reduce feature dimensions.

DFT is a supervised feature selection process that measures the dimension-wise importance. For a given 1D input feature, we can order the samples by the feature values and bind the feature dimension with the sample maximum and sample minimum. Then, we can partition the samples along the given dimension and calculate the partition purity by weighted cross-entropies with the pseudo-labels obtained from section III-B. A feature is more discriminant if it has a lower loss value. Then, we can plot the loss value curve from the lowest to the highest and use the elbow point to select discriminant features from the whole feature set.

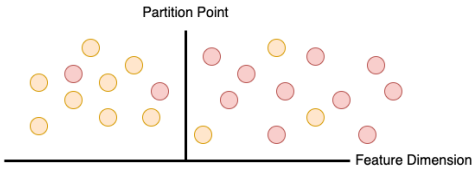


Fig. 3. Visualization of DFT. Red and orange dots represent the binary labels. The partition metric is the weighted sum of the left and right binary cross-entropy.

Separating the whole dataset into subsets allows us to conduct the discriminant feature test among the tokens with the pseudo-labels from the clustering results. Thus, the token-level alignments can be performed using the same procedure as the global-level alignment.

D. Mathematical Expression

The overall alignment process can be divided into three modules: 1) global matching, 2) sub-domain clustering, and 3) sub-domain matching. The sub-domain clustering and alignment will be conducted within the image and the text domain. We can aggregate the alignments in the sub-problem to approximate the overall alignment.

$$\begin{aligned} &sim(image, text) \\ &= sim(W^{vis}\mathcal{F}(\text{image}), W^{txt}\mathcal{G}(\text{text})) \\ &= sim(W^{vis}[G_{vis}; T_{vis}], W^{txt}[G_{txt}; T_{txt}]) \\ &\sim sim(W^{vis}_{global}G_{vis}, W^{txt}_{global}G_{txt}) \\ &+ sim(W^{vis}_{tokens}[G_{vis}; T_{vis}], W^{txt}_{g}G_{txt}) \\ &+ sim(W^{vis}_{global}G_{vis}, W^{txt}_{tokens}[G_{txt}; T_{txt}]), \end{aligned} \tag{9}$$

where G_{vis} and G_{txt} denote the pooled outputs from the feature extractors (global features), T_{vis} and T_{txt} denote the token, i.e. fine-grained, features, $W^{\cdot}_{\cdot}$ denote the alignment matrices corresponding to different subsets from the clustering results. Due to the computational cost, we cannot directly collect all token features. Therefore, we conduct the feature selection process based on the clustering results.

The feature selection process is an approximation based on the clustering results. The process is denoted as the combination of the conditional probabilities. For simplicity, we ignore the alignment matrix in the following representations.

$$\begin{aligned} &E[sim(\mathcal{F}(\text{image}), \mathcal{G}(\text{text}))] \\ &= E[E[sim(\mathcal{F}(\text{image}), \mathcal{G}(\text{text}))]|\text{image} \in C_1; \text{text} \in C_2] \\ &\sim E[sim(G_{vis}, G_{txt})] \\ &+ E[E[sim(DFT([G_{vis}; T_{vis}]), G_{txt})|\text{image} \in C_1]] \\ &+ E[E[sim(G_{vis}, DFT([G_{txt}; T_{txt}]))|\text{text} \in C_2]], \end{aligned} \tag{10}$$

where $DFT(.)$ represents the feature selection and dimension reduction process in section III-C, and C_1 and C_2 represent the KMeans cluster sets. Instead of training a complicated alignment process from the token level output of the feature extractor, we propose the stagewise decomposition on the dataset and train simpler structures for the subsets. Meanwhile, the alignments in the stages are linear, which provides the inversion operation and preserves the dual accessibility in both image and text domains.

IV. Experiments

A. Dataset

We conduct the image-to-text and text-to-image retrieval on the image-text benchmark: Flickr30k and MS-COCO. The Flickr30k dataset [15] contains 31,000 images, and every image has five paired captions. The training set contains 29,000 images; the validation and testing sets contain 1,000. The MS-COCO [2] is a larger-scale dataset with 123,287 images, each containing at least five captions. We follow the 'Karpathy' splitting for the experiments [16]: 113,287 images for training, 5,000 for validation, and 5,000 for testing. We use the two benchmarks with different sizes to demonstrate the scalability and generalizability of our approaches. The performance is evaluated using the Recall@K metric where $K \in \{1, 5, 10\}$. The notation K denotes the top-K matches from the candidate set. The retrieval will be considered a true positive once the predicted matches include the paired ground truth.

B. Retrieval

We conducted the experiments and compared our alignment approach to the SOTA retrieval models. The results are shown in Table. I. In the experiments, we extract information from the frozen CLIP image and text encoder. The CLIP encoder contains more than 428M parameters. However, we do not fine-tune the overall encoder in our alignment process; instead, we train additional alignment matrices. The trainable parameters can be reduced from 428M to 9.43M (2.2%).

In Flickr30k (1k testing set), our approach outperforms other image-to-text and text-to-image retrieval methods. The alignment can improve the recall@1 by 0.6% in the image-to-text retrieval. Meanwhile, our approach provides a 6% boost in text-to-image retrieval. RCAR [17] needs dual-way optimized models, namely image-to-text and text-to-image. Our method is optimized in a feed-forward manner and ensembles the substructures directly.

In MS-COCO (5k testing set), our method provides competitive performance in image-to-text retrieval and outperforms

TABLE I
THE FLICKR30K(1K TESTING SET) AND MSCOCO(5K TESTING SET) DATASET RETRIEVAL PERFORMANCE. WE COMPARE THE SINGLE-MODEL PERFORMANCE AMONG ALL MULTI-MODAL RETRIEVAL MODELS. THE NUMBERS ARE TAKEN FROM DIAO ET AL. [17] R@1 REPRESENTS RECALL@1 FOR SIMPLICITY.

	Flickr30k (1k testing set)						MS-COCO (5k testing set)					
	image-to-text			text-to-image			image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [5]	67.4	90.3	95.8	48.6	77.7	85.2	50.4	82.2	90.0	38.6	69.3	80.4
VSRN [18]	71.3	90.6	96.0	54.7	81.8	88.2	53.0	81.1	89.4	40.5	70.6	81.1
CAAN [19]	70.1	91.6	97.2	52.8	79.0	87.9	52.5	83.3	90.9	41.2	70.3	82.9
IMRAM [20]	74.1	93.0	96.6	53.9	79.4	87.2	53.7	83.2	91.0	39.7	69.1	79.8
MMCA [21]	74.2	92.8	96.4	54.8	81.4	87.8	54.0	82.5	90.7	38.7	69.7	80.8
GSMN [7]	76.4	94.3	97.3	57.4	82.3	89.0	–	–	–	–	–	–
SGRAF [22]	77.8	94.1	97.4	58.5	83.0	88.8	57.8	84.9	91.6	41.9	70.7	81.3
SHAN [23]	74.6	93.5	96.9	55.3	81.3	88.4	–	–	–	–	–	–
WCGL [24]	74.8	93.3	96.8	54.8	80.6	87.5	–	–	–	–	–	–
RCAR [17]	78.7	94.6	97.6	59.5	84.0	89.5	**59.6**	**85.8**	**92.4**	42.5	71.7	81.8
SGRAFS [8]	79.2	95.3	97.7	58.3	83.1	89.2	58.0	85.1	91.6	41.7	71.2	81.5
CLIP [11]	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
GMA(Ours)	**88.6**	**98.9**	**99.6**	**74.8**	**93.5**	**96.7**	58.6	83.2	90.0	**45.3**	**72.6**	**82.8**

the others in text-to-image retrieval by a boost of 2.1% in Recall@1. We achieve the best text-to-image retrieval performance among the two datasets, showcasing our approach's scalability.

C. Visual Feature Comparison

This section demonstrates the alignment between the visual/text encoders, which are trained separately. Starting from the jointly trained CLIP structure, we change the text encoders into the RoBERTa [25] and the visual encoder into a CNN-based object detector [6].

The results are shown in Table II. The best performance comes from the jointly train models, whose representations are preliminarily aligned in the pre-trained process. Compared to the CLIP visual encoder, the features from the object detector are weaker in the alignment process. On the other hand, the separately trained text encoder, RoBERTa [25], does not take harm from the unpaired training dataset. The representations from the CLIP visual encoder and the RoBERTa text encoder can provide a competitive performance in image-to-text retrieval and a better performance in text-to-image retrieval than the original CLIP. The encoder can be adapted to the retrieval application without fine-tuning with the paired text/text data.

V. CONCLUSION AND FUTURE WORK

Our approach can achieve outstanding performance in both image-to-text and text-to-image retrieval tasks. Furthermore, our method involves a step-by-step alignment process that maintains compatibility in the decision-making procedure. We divide the alignment into global and sub-domain matching and apply a feature selection method to decrease the input feature dimensions. All the sub-processes can be expressed statistically rather than the black-box outputs. To ensure computational efficiency, we have frozen the visual and text encoders and only trained the alignment matrices, accounting for only about 3% of the parameters compared to the original model.

We have tested our approach of aligning visual and text encoders trained separately. In the testing dataset, we found that the pre-trained text encoder can improve the performance of text-to-image retrieval. Replacing the text encoder can also lead to similar performance in image-to-text retrieval.

We are working on developing a purely green learning solution to image understanding in the visible future. By aiming not only for transparency but also computational efficiency, we can have a better understanding of the multimodal information representation.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[2] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[3] Y. Zhu, R. Kiros, R. Zemel, *et al.*, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[4] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–23, 2020.

[5] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 201–216.

TABLE II
THE EXPERIMENT RESULTS WITH DIFFERENT VISUAL AND TEXT FEATURES FOR THE ALIGNMENT PROCESS. ALL THE EXPERIMENTS ARE CONDUCTED IN THE FLICKR30K DATASET.

Visual *enc.*	Text *enc.*	Alignment (GMA)	Flickr30k (1k testing set)					
			image-to-text			text-to-image		
			Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
CLIP vis [11]	CLIP text [11]	x	88.0	98.7	99.4	68.7	90.6	95.2
Object detector [26]	RoBERTa [25]	v	36.0	65.3	77.6	53.2	69.1	70.2
Object detector [26]	CLIP text [11]	v	33.2	62.7	75.4	49.8	65.1	68.2
CLIP vis [11]	RoBERTa [25]	v	86.3	98.2	99.4	73.2	**93.5**	**96.9**
CLIP vis [11]	CLIP text [11]	v	**88.6**	**98.9**	**99.6**	**74.8**	**93.5**	96.7

[6] P. Anderson, X. He, C. Buehler, *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[7] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 921–10 930.

[8] B. Jawade, D. D. Mohan, N. M. Ali, S. Setlur, and V. Govindaraju, "Napreg: Nouns as proxies regularization for semantically aware cross-modal embeddings," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1135–1144.

[9] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, IEEE, vol. 2, 2006, pp. 1735–1742.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[12] C.-C. J. Kuo and A. M. Madni, "Green learning: Introduction, examples and outlook," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103 685, 2023.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[14] Y. Yang, W. Wang, H. Fu, C.-C. J. Kuo, *et al.*, "On supervised feature selection from high dimensional feature spaces," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[15] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[16] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[17] H. Diao, Y. Zhang, W. Liu, X. Ruan, and H. Lu, "Plug-and-play regulators for image-text matching," *IEEE Transactions on Image Processing*, 2023.

[18] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4654–4662.

[19] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3536–3545.

[20] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 655–12 663.

[21] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 941–10 950.

[22] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 1218–1226.

[23] Z. Ji, K. Chen, and H. Wang, "Step-wise hierarchical alignment network for image-text matching," *arXiv preprint arXiv:2106.06509*, 2021.

[24] Y. Wang, T. Zhang, X. Zhang, *et al.*, "Wasserstein coupled graph learning for cross-modal retrieval," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 2021, pp. 1793–1802.

[25] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

Speech Depression Recognition from the Self-reference Effect Using LSTM with ResNet

Xiaoyong Lu*, Yimin He, Daimin Shi and Yang Liu
*Northwest Normal University, Lanzhou, China
E-mail: luxy@nwnu.edu.cn Tel: 13893241411
F-mail: simonhe@nwnu.edu.cn Tel: 15984096104

***Abstract*— Deep neural networks (DNNs) have recently achieved great success in various learning tasks and have also been used for the classification of depression. While DNNs are showing their potential in the classification task, they cannot fully utilize the temporal information. To mine the relevance of signals in audios and increase the diversity of information, a neural network architecture for the purpose of using sequential information is proposed in this paper. The proposed structure is composed of two separated lower networks and one upper network. We refer to these as LSTM layers, Residual Networks (ResNet) layers, and connected layers, respectively. The LSTM layers can learn long-term dependencies from the learned local features. The ResNet layers learn the spectro-temporal locality from spectrogram images. Finally, the connected layers summarize the outputs of two networks to take advantage of the complementary features of the LSTM and ResNet by combining them. In addition to this, we experiment on our own designed dataset, which was designed based on the Self-reference Effect (SRE) experimental paradigm of psychology. Our results show the proposed combination structure can improve the accuracy of speech depression recognition.**

I. Introduction

Depressive disorder is a common affective disorder, also known as depression[1], which is characterized by sadness, loss of interest, feelings of guilt or low self-worth, and poor concentration [2]. As speech is easy to obtain non-offensively with low cost, many researchers have explored the possibility of depression prediction through speech [3]. In the past decade, predicting emotion using speech has been a success[4]. Corollary to that, the proof of speech brings useful in predicting clinical depression scores implies that speech is an important indicator for depression[5]. There are also reports stating that the verbal behavior of depressed patients should have "decreased verbal activity productivity, a diminished prosody and monotonous, 'lifeless' sounding speech" [6][7][8]. This paper aims to boost depression diagnostics accuracy by using speech as an objective measure.

However, as deep learning techniques have been widely used on various learning tasks, researchers have started to apply them to depression classification as well [9]. Recently, approaches based on deep learning have been proposed [10]. To overcome this shortcoming, recurrent neural networks (RNNs) and long short-term memory (LSTM), which is a special type of RNN, have been applied to sequence learning[11][12]. DNNs can only map from present input vector to output vector, whereas LSTM can map from sequence to output sequence or vector. Therefore, LSTM can learn the temporal information through consecutive input vectors. The authors in [13] and [14] proposed depression recognition techniques based on bi-directional LSTM which yielded higher performance compared to the DNNs. Unlike depression speech which occurs in a short time frame, depressions are maintained for relatively a longer range. Other approaches were proposed to use convolutional neural networks (CNNs) with spectrogram image features (SIF) [15]. In [16], the authors addressed the importance of spectro-temporal locality and proposed a CNN-based depression recognition algorithm. In this paper, we propose to combine the LSTM and ResNet in parallel as lower networks in order to exploit sequential correlation and local spectro-temporal information.

Most depressed individuals have cognitive biases, which are often manifested as self-negative biases in the processing of self-reference information. That is, with the increase in the severity of depression, the subjects are less and less inclined to associate self with positive information, and more and more inclined to associate self with negative information. In summary, this paper further studies the effects of self-referential processing of depressed individuals through modeling.

II. Parallel Combination of Lstm And ResNet

In this section, we describe our approach to improving the classification accuracy of depression. The schematic of the proposed neural network structure can be seen in Fig.1.

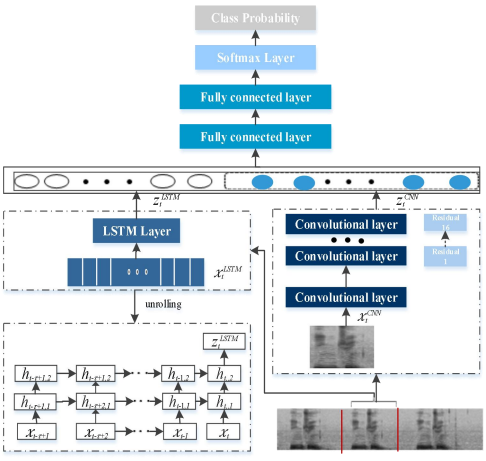


Fig.1. Neural network architecture for the proposed method.

A. Dataset

A total of 133 subjects were recruited for the study. All subjects, including healthy subjects and subjects with depression, were from Northwest Normal University. The study recruited 67 healthy and 66 depressed subjects, all of whom were between the ages of 18-26. From a gender perspective, there were 31 males and 35 females in the depression group, and in the health group, there were 33 males and 34 females. All subjects were excluded from other mental disorders. The subjects' BDI-II (Beck Depression Inventory, BDI-II) scores are shown in Fig.2.

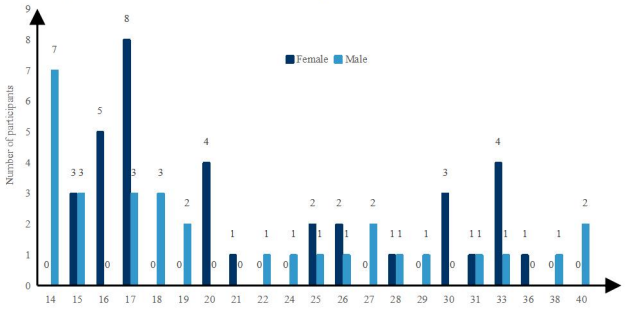


Fig.2. The BDI-II score distribution of depressed subjects.

This study is an experimental design of 3 (emotional state: positive, neutral, negative) × 4 (task type: vocabulary reading, interview, picture description, emotional sentence). Besides, depressed subjects have damages of self-processing and self-recognition, negative self-reference, over-attention to self, and use the classic paradigm of SER to the corpus. The picture of subjects recording scene is show in Fig.3.

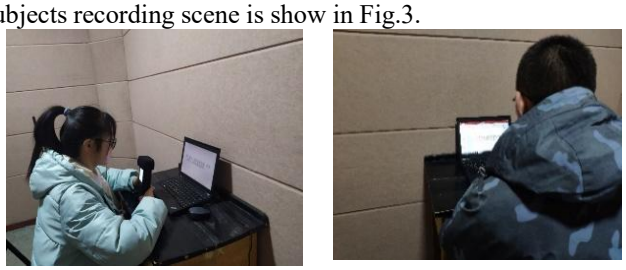


Fig.3. Picture of recording scene.

The experiment was conducted in a professional recording studio. Participants held professional recording equipment R-26. All tasks and instructions were in the same computer program. Subjects sat about 1m away from a 21-inch computer screen directly in front of the screen, and the task requirements were displayed on the screen. After each task material is presented, the participant needs to answer according to the prompts. The main examiner operates the program from the side and uses the recording device to record the subject's speech.

TABLE I. BDI-II SCALE DEPRESSION LEVEL DIVIDE

Score range	Level	Class
0-13	minimal	0
14-19	mild	1
20-28	moderate	2
29-63	severe	3

All audio recordings are annotated by the BDI-II scale. TABLE I shows the corresponding relationship between the BDI-II scale score and the severity level of depression.

B. Feature extraction

In the proposed system, different types of neural networks are combined in parallel. The LSTM layers utilize a sequence of depressive features, but the ResNet layers use spectrogram images. As inputs for the ResNet layers, the SIF [17][18] are extracted from the speech. Firstly, a spectrogram is generated by short-time Fourier transform. Given audio frame $s(n)$ segmented by length N and Hamming window $w(n)$, the short time spectral column $F(f,t)$ at time t is computed as,

$$F(\mathrm{f},\mathrm{t}) = \left|\sum_{\mathrm{n}=0}^{\mathrm{N}-1} \mathrm{s(n)w(n)e}^{\frac{-\mathrm{j}2\pi\mathrm{nf}}{\mathrm{N}}}\right| \quad (1)$$

for $f = 0, ..., \frac{N}{2}$. To generate a spectrogram image which has K-bin frequency resolution, down-sampling is performed by using a window of length $W = \frac{N}{2K}$ as follows:

$$F_{down}(f,t) = \sum_{i=0}^{W-1} \frac{F(f+i,t)}{W} \quad (2)$$

for $f = 0, \ldots, (K-1)$. Finally, a simple de-noising method is performed by subtracting each minimum frequency bin value in a frame-wise manner as follows:

$$\mathrm{F_{dn}(f,t)} = \mathrm{F_{down}(f,t)} - \min\{\mathrm{F_{down}(f,t)}\} \quad (3)$$

for $f = 0, \ldots, (K-1)$. In the proposed system, the extracted SIF has a size of $K \times \tau$, where τ represents the time resolution which is also identical to the timesteps in the LSTM layers.

C. LSTM layers

The hidden layers of LSTM have self-recurrent weights. These enable the cell in the memory block to retain previous information. In the proposed system, τ vectors are used for sequential learning. Previous $\tau - 1$ vectors and one present vector are forwarded to the recurrent layer sequentially. If the MFCC vectors from $x_{t-\tau+1}$ to x_t are used as the present inputs, vectors from $x_{t-\tau+2}$ to x_{t+1} will be used as the next input sequence. The output vector z_t^{LSTM} is extracted from the input MFCC sequence x_t^{LSTM} through the LSTM layers, where $z_t^{LSTM} = [x_{t-\tau+1}, ..., x_t]$.

D. ResNet layers

For ResNet layers, SIF x_t^{CNN}, which is a $F \times \tau$ matrix, is extracted. The convolutional layer performs 2-dimensional convolution between the spectrogram image and the pre-defined linear filters. Thus, if we apply K different filters to the spectrogram image, K different filtered images are generated in the convolutional layer. The flattened output vector z_t^{CNN} is extracted from x_t^{CNN} through the ResNet layers.

E. Connected layer of LSTM and ResNet

In [19], the long-term recurrent convolution network (LRCN) model was proposed for visual recognition, whose outputs are fed into the LSTM network, which finally predicts the class of the input. In [20], a cascade structure was used for speech search. Compared to the method mentioned above, the proposed network forms a parallel structure in which LSTM and ResNet accept different inputs separately. Concatenated vector z_t^{concat} is forwarded to the fully connected layer, where $z_t^{LSTM} = [z_t^{LSTM}, z_t^{CNN}]$. These enable the proposed model to learn the sequential information and spectro-temporal information, simultaneously. Finally, the class probability $\tilde{y}_i$ is predicted through the softmax layer.

III. Evaluation

The distribution of speech samples in the training set and validation set for different tasks and the number of subjects is shown in Table II. To assess the performance of the proposed method, we conducted several experiments on the corpus.

TABLE II. Distribution of speech segments in each task

Task	Section	Subjects	Segment
Vocabulary reading	train	106	2361
	validation	27	583
Interview	train	106	6963
	validation	27	1924
Image description	train	106	1737
	validation	27	419
Emotional sentence	train	106	38453
	validation	27	12865
Autobiographical	train	106	8719
	validation	27	2915

A. Neural Networks Setup

All networks in our experiments were trained using mean squared error as the loss function supervised by one-hot encoding class vectors. The randomly ordered minibatches in each epoch were set to be 16. After a minibatch was processed, the weights were updated using adadelta [21]. To mitigate the over-fitting problem in the training phase, we used the dropout technique which has already proved its regularization capability [22]. The output layer contained softmax nodes identical to the number of classifying.

- ResNet

The ResNet architecture for the deep learning method proposed in this paper uses ResNet structure to predict depression. The input features were F × τ size SIF, where F=224 and τ=224. The convolutional neural network has five residual blocks, Conv1, Conv2x, Conv3x, Conv4x, Conv5x, and FC is the final classification layer. The size of Conv1 convolution kernel is 7×7, and the step size is 2. Each of the remaining convolutional layers uses a 3×3 convolution kernel of step size 2.

- LSTM

The network had two hidden layers with 256 LSTM units each and one feed-forward layer with 512 ReLU units. The input sequence consisted of 40 frames of 60-dimensional MFCC features. Dropout was applied with a probability of 40% for all layers.

- Combination of LSTM and ResNet

As a proposed system, we built a combined structure of LSTM and ResNet in parallel. The network setup and structure of the LSTM part and ResNet part were identical. To combine and further train the two separated networks, we used fully connected layers. The connected layers onsisted of two hidden layers with 512 ReLU units each.

B. Results and Discussion

In this study, we not only predict whether the person is depressed but also estimate the severity of depression. In the designed corpus, the depression binary classification of an individual is given based on the severity of depression which is measured using BDI-II scores. These predictions are performed under the following section.

a) Depression Binary Classification

In the binary classification task, we aim to predict whether an individual is depressed. Table III shows the classification accuracy under different tasks. The overall accuracy achieved in predicting depression reaches 68.5%. F1 score is designed to deal with the non-uniform distribution of class labels by giving a weighted average of precision and recall. It can also be seen that the recognition accuracy of self-type tasks is significantly higher than that of reading-type tasks.

TABLE III. The results of Binary classification of each task(%)

Task	Model	Accuracy	F1 Score	Precision	Recall
Vocabulary reading	ResNet18_LSTM	68.5	70.7	66.1	76
	ResNet34_LSTM	62.5	62.3	62.6	62
Interview	ResNet18_LSTM	69	70.2	67.6	73
	ResNet34_LSTM	67	68.6	65.5	72
Image description	ResNet18_LSTM	59.5	70.3	55.5	96
	ResNet34_LSTM	68	75.4	61.3	98
Emotional sentence	ResNet18_LSTM	68	66.7	69.6	64
	ResNet34_LSTM	65.5	65	66	64
Autobiographical	ResNet18_LSTM	74	71.1	80	64
	ResNet34_LSTM	68.5	69.9	67	73

Table IV shows the influence of gender on results under the premise of self and others.

TABLE IV. RESULTS OF SELF AND OTHERS(%)

Model	Indicator	Others		Self	
		Male	*Female*	*Male*	*Female*
ResNet18_LSTM	accuracy	62	68	64.5	72.5
	F 1 score	60	67	64.7	70.9
	precision	63.3	69.1	64.4	75.3
	recall	57	65	65	67
ResNet34_LSTM	accuracy	64	67	67.5	68
	F 1 score	63.6	64.9	66.3	69.8
	precision	64.3	69.3	68.9	66.1
	recall	63	61	64	74

In the gender case, compared with normal individuals, patients too much associate negative emotions with self for processing, leading to self-negative prejudice. They negatively encode individual experiences, tending to summarize the negative experience and avoid the positive factors in the situation, which will distort the perception and bias the information. It can be seen from the table that the accuracy rates are all above 60%, and the highest accuracy rate can reach 72.5%. Besides, the recognition accuracy rate of females is higher than that of males regardless of self or others, which is in line with previous research results.

Under different emotional valences, the influence of gender on the recognition results was investigated. The experiment results are shown in the Table V.

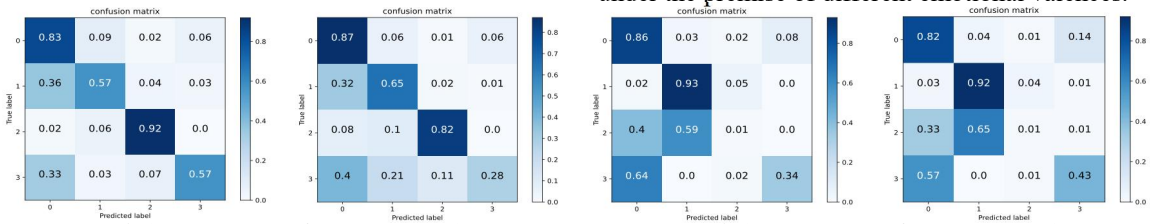
TABLE V. RESULTS UNDER DIFFERENT GENDER (%)

Model	Indicator	ResNet18_ LSTM	ResNet34_ LSTM
Positive (Male/Female)	accuracy	75.5/73.5	77/72
	F1 score	75.9/71.4	78.3/71.7
	precision	74.8/77.6	74.1/72.4
	recall	77/66	83/71
Neutral (Male/Female)	accuracy	82.5/73.3	78/64.5
	F1 score	82.1/75.8	75.8/97.8
	precision	84.2/71.1	84.1/69.8
	recall	80/81.1	69/82
Negative (Male/Female)	accuracy	81.5/74	83/76
	F1 score	82.1/75	82.8/76
	precision	91/72.2	83.7/76
	recall	85/78	82/76

With the help of different evaluation indicators, it can be seen from the table that the highest recognition accuracy rate is as high as 83%. Under negative emotions, the overall accuracy rate of female subjects is still higher than that of male subjects, indicating that female depressed patients are more inclined to express their emotions in speech. It also manifests that the depressed subjects have abnormal self-processing, and that the female subjects are more sensitive to emotions.

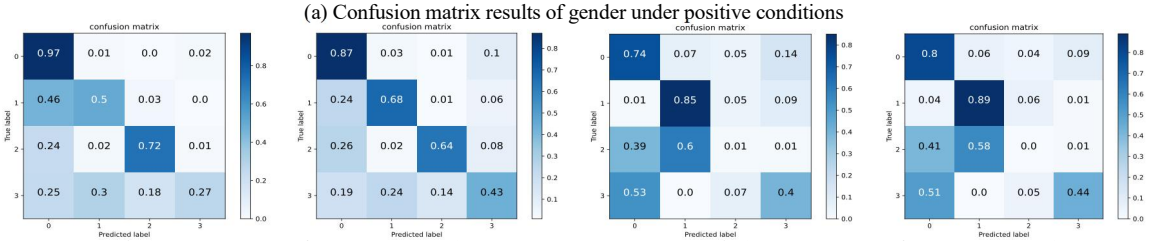
b) Performances on depression severity levels prediction task

The Fig.4 shows the confusion matrix diagrams of multi-class performance of gender in the depression recognition task under the premise of different emotional valences.



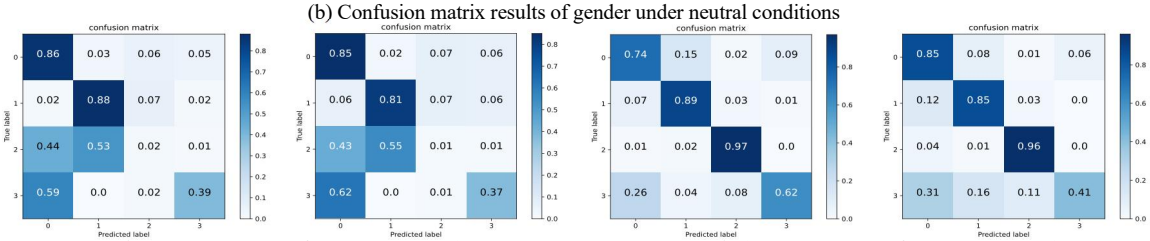
Male (ResNet18_LSTM/ResNet34_LSTM) Female (ResNet18_LSTM/ResNet34_LSTM)

(a) Confusion matrix results of gender under positive conditions



Male (ResNet18_LSTM/ResNet34_LSTM) Female (ResNet18_LSTM/ResNet34_LSTM)

(b) Confusion matrix results of gender under neutral conditions



Male (ResNet18_LSTM/ResNet34_LSTM) Female (ResNet18_LSTM/ResNet34_LSTM)

(c) Confusion matrix results of gender under negative conditions

Fig.4. Multi-class confusion matrix results of gender under different emotional valence

From the results of the confusion matrix, it can be seen that under negative emotions, the recognition accuracy rate of both male and female subjects is higher than that under other emotions. This also reveals that depressed individuals have a negative bias in the processing of self-reference information, which has its cognitive roots, that is, individuals have formed a negative self-schema in their minds. Besides, as can be seen from Fig.4 in the recognition rate of moderate, the model

performs well. The proposed method has the highest recognition rate of mild, while has a low recognition rate of severe. Similar to neutrality recognition, the mild and severe recognition rate of each algorithm is similar. In conclusion, the proposed method has satisfactory results in terms of minimal and mild recognition rate. Furthermore, for minimal and mild, because of the large number of samples and obvious characteristics of them, they are in a relatively high recognition rate in all kinds of conditions. In addition, in the experimental results, we can see the most emotions of moderation are misclassified as mild training sample is the largest, which results in many samples of moderate shifting to the mild samples. Looking at the confusion matrix, the method proposed in this paper can classify subjects with depression propensity and healthy subjects.

IV. Conclusion

In this paper, to enhance the recognition accuracy of depression, we proposed a novel neural network structure. By combining different networks in parallel, the proposed method was able to learn complementary information of LSTM and ResNet. From the experimental results, depressed individuals have a slightly higher recognition accuracy of negative cognitive activities or cognitive processes, therefore, depressed patients have self-negative bias.

Acknowledgment

This research has received funding from the academic requirements for the National Science Foundation of China (NSFC) under grant No. 31860285 and No. 62367007.

References

[1] H. Cai, Z. Qu, Z. Li, Y. Zhang, and B. J. I. F. Hu, "Feature-level Fusion Approaches Based on Multimodal EEG Data for Depression Recognition," vol. 59, 2020.

[2] T. B. M. D. Aaron and A. A. P. D. Brad, "Depression: Causes and Treatment," Clin. Geriatr. Med., vol. 14, pp. 765-786, 2014.

[3] P. Wei et al., "Depression recognition based on speech analysis," Chinese Science Bulletin, 2018, pp.2081-2092.

[4] K. Venkataramanan and H. R. Rajamohan, "Emotion Recognition from Speech," 2019.

[5] B. Stasak, J. Epps, and A. Lawson, "Analysis of Phonetic Markedness and Gestural Effort Measures for Acoustic Speech-Based Depression Classification," in International Conference on Affective Computing & Intelligent Interaction Workshops & Demos, 2017, pp. 165-170.

[6] X. W. Chen, R. Q. Yang, L. Zhang, J. C. Guo, and L. V. Rui-Xue, "Abnormal autonomic nervous system and electrophysiology in patients with depression," Chinese Mental Health Journal, 2016.

[7] C. W. Espinola, J. C. Gomes, J. Pereira, and W. Santos, "Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study," Research on Biomedical Engineering, vol. 37, no. 1, pp. 53-64, 2021.

[8] Y. Omiya, T. Takano, T. Uraguchi, M. Higuchi, and S. Tokuno, "Estimating depressive status from voice," in The 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM2018), 2018.

[9] M. S'Adan, A. Pampouchidou, and F. Meriaudeau, "Deep Learning Techniques for Depression Assessment," in 7th International Conference on Intelligent and Advanced System, 2018.

[10] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "[ACM Press the 7th Annual Workshop - Mountain View, California, USA (2017.10.23-2017.10.23)] Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC \"17 - Hybrid Depression Classification and Estimation from Audio Video and Text In," pp. 45-51, 2017.

[11] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech," 2019.

[12] T. Al Hanai, M.M. Ghassemi, and J.R. Glass, "Detecting Depression with Audio/Text Sequence Modeling of Interviews," In Interspeech, pp. 1716-1720, 2018.

[13] A. Salekin, J.W. Eberle, J.J. Glenn, B.A. Teachman, and J.A. Stankovic, "A weakly supervised learning framework for detecting social anxiety and depression," Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, vol. 2, no. 2, p. 81, 2018.

[14] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. ifti, H. Gle, A.A. Salah, and M. Pantic, "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," In Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, ACM, pp. 3-13, 2018.

[15] K. Chlasta, K. Wołk, and I. J. a. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," 2019.

[16] J. Li, X. F. J. C. Applications, and Software, "AUDIO DEPRESSION RECOGNITION BASED ON DEEP LEARNING," 2019.

[17] KK. Chlasta, K. Wołk, and I. J. a. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," 2019.

[18] Z. Wang, L. Chen, L. Wang, and G. J. I. A. Diao, "Recognition of audio depression based on convolutional neural network and generative antagonism network model," vol. PP, no. 99, pp. 1-1, 2020.

[19] J. Zhao, X. Mao, L. J. B. s. p. Chen, and control, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," vol. 47, no. JAN., pp. 312-323, 2019.

[20] R. Jain, "Improving performance and inference on audio classification tasks using capsule networks," 2019.

[21] M. D. Zeiler, "Adadelta: an adaptive learning rate method," arXiv preprint arXiv:1212.5701, 2012.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. J. J. o. M. L. R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," vol. 15, no. 1, pp. 1929-1958, 2014.

SDNet: Noise-Robust Bandwidth Extension under Flexible Sampling Rates

Junkang Yang*§, Hongqing Liu*, Lu Gan†, Yi Zhou*, Xing Li‡, Jie Jia‡ and Jinzhuo Yao*
* School of Communications and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China
† College of Engineering, Design and Physical Science, Brunel University, London, U.K.
‡ AI Lab, vivo Mobile Communication Co,.Ltd, Nanjing, China
E-mail: s220101187@stu.cqupt.edu.cn

***Abstract*—Bandwidth extension (BWE), also known as audio super-resolution (SR), aims to predict a high resolution (HR) speech signal from its low resolution (LR) corresponding part. Most neural BWE models work at a specific sampling rate but, producing the final result in a noise-free environment by recovering the spectrogram of high-frequency part of the signal and concatenating it with the original low-frequency part. Although these methods achieve high accuracy, they become less effective when facing the real-world scenario, where unavoidable noise is present and sampling rates are flexible. To address this problem, we propose Super Denoise Net (SDNet), a neural network for a joint task of BWE and noise reduction from a flexible low sampling rate signal. To that end, we design gated convolution and lattice convolution blocks to enhance the repair capability and capture information in the time-frequency axis, respectively. The experiments show our method outperforms all current state-of-the-art (SOTA) noise-robust BWE model in Valentini-Botinhao test set. Our model also outperforms other baselines on DNS 2020 no-reverb test set with higher objective and subjective scores.**

I. Introduction

Bandwidth extension (BWE) is the task to reconstructing the high resolution (HR) part of the speech signal from its low-resolution (LR) part, which is also termed as audio super-resolution (SR). As one of the important tasks in the front-end of speech processing, BWE is widely applied in wireless communication, speech recognition [1], text-to-speech [2], to name a few.

Although BWE has achieved considerable progress in past years, few studies have focused on the task of BWE in noisy environments, i.e., bandwidth extension along with the noise suppression. Most existing frequency domain based work basically keeps the low-frequency part and predicts only the high-frequency part, and finally concatenates the two parts in series. When noise exists in the low-frequency part, this pipeline not only fails to remove the noise, but also produces the biased prediction of high-frequency part due to noise interference. To address this issue, in [3], the authors introduced a multi-stage model to respectively conduct noise reduction and bandwidth extension. In [4], authors combine UNet+AFiLM [5] and an improved DTLN [6] to form a two-stage system. The authors in [7] simultaneously estimate the missing components and the noise distribution in degraded speech signal with a DNN. For the models of MTL-MBE [8] and EP-WUN [9], the authors claimed that they achieved the SOTA performance of this task. On the other hand, there are a lot of speech signals with flexible sampling rates or known sampling rates but small effective bandwidths (the high-frequency components are missing), and so far there is no noise-robust BWE model can deal with this situation.

In this paper, we propose Super Denoise Net (SDNet), a neural network that removes noise while extending the bandwidth. To that aim, we design a generator and discriminator network, where encoder-bottleneck-decoder structure is utilized in generator and multi-discriminator structure is developed, enabling the possibility of adversarial training to make the model robust. It is of interest to note that the proposed model does not require the prior information of the sampling rate of LR speech. Our main contributions are provided as follows.

- We develop a single-stage network, where the gated convolution and lattice convolution blocks are utilized to jointly perform noise reduction and SR. Our method has improved the performance of model and outperformed the baselines in both noisy and noise-free cases.
- Our approach is one of the first noise-robust BWE model that supports to process all the speeches whose sampling rates are from 4 kHz to 16 kHz, which means that the sampling rate of LR speech is flexible in our case.
- The quality of the speeches generated by our model from 8 kHz noisy samples are even better than those of some popular models only focusing on 16 kHz to 16 kHz noise reduction.

II. Methods

A. Problem Settings

In this paper, we address the problem of recovering HR clean speech from LR noisy speech. Given a HR clear speech $y \in R^T$, where T is the length, after downsampling s times and adding noise at the same sampling rate with the same length, the LR noisy speech $x \in R^{\frac{T}{s}}$ is generated. Our goal is to design a function G that can efficiently predict y from the observation x, i.e., $\hat{y} = G(x) \approx y$. The same representation as above will be used in the formulas below.

§ Part of work was done during internship at vivo AI lab.

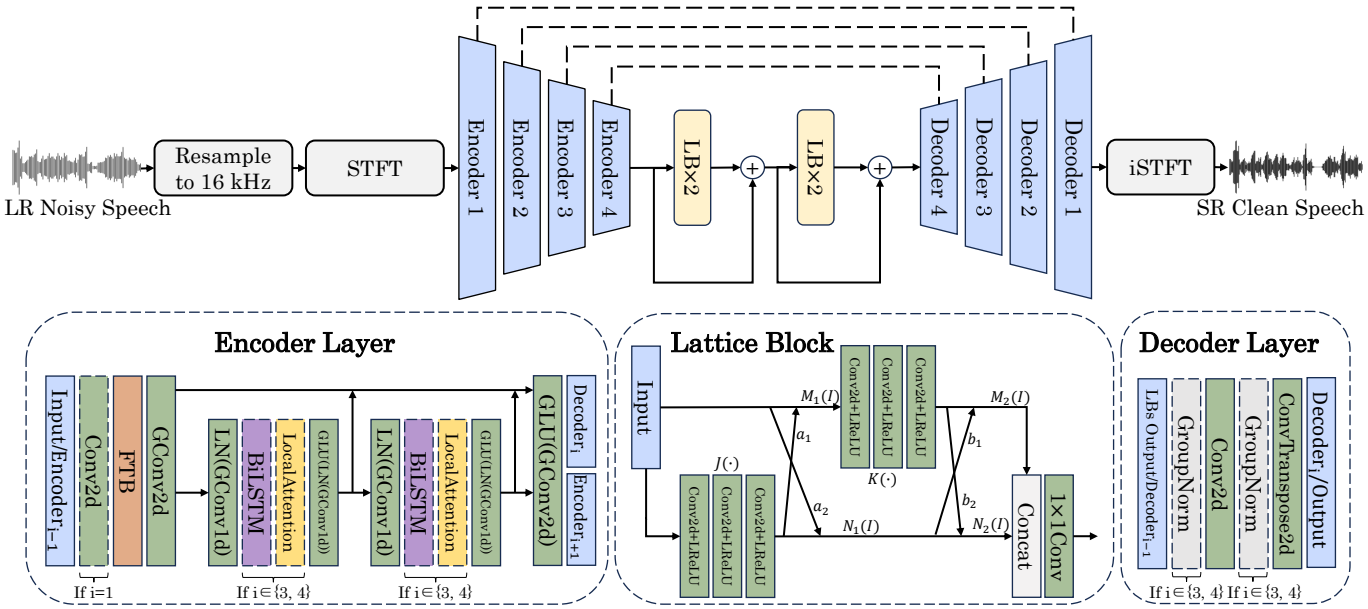


Fig. 1. The generator network architecture.

B. Network Architecture

Our model uses a U-shaped structure, containing encoders, decoders, and lattice convolution blocks (LBs) as bottleneck layers. The residual connections are set between encoders and decoders, also between the lattice convolution blocks. Model architecture is visualized in Figure 1. Due to space limit, we provide parameter settings, input size, and output size for each layer at our demo page[1].

1) Generator:

Encoder and decoder. As illustrated in Figure 1, there are 4 layers in encoder and decoder each. The input of encoder is the spectrogram of upsampled waveform and it will be reshaped at the first encoder layer with a 2D convolution, with the complex part moved to channel dimension. After that, a frequency transform block (FTB) [10] is applied to capture the non-local correlations in spectrogram along the frequency axis. Unlike previous work [11], we utilize gated convolution (GConv) instead of ordinary convolution in the later structure, which has been proved to enhance the model's generation ability by learning a dynamic feature selection mechanism for each channel and each spatial location [12]. Inside the encoder, there are two residual branches with two 1D gated convolutions at the beginning and end. In the middle, there are LSTM and temporal-based attention modules to capture long-distance relations. Followed by each encoder layer is a decoder layer that recovers the latent vectors equal to the size of spectrogram before passing the encoder. It is worth noting that there is concatenated residual connection between each encoder and decoder layer, while between two residual branches within the encoder layer, it is a summation residual connection.

[1]https://sdnetdemo.github.io/

Bottleneck layers. The bottleneck layers include 4 lattice convolution blocks (LBs), which were first proposed in image restoration task [13]. Combined with gated convolution, this structure can offer a blend of structured interpolation with adaptive and long-range context modeling. As shown in Figure 1, each LB includes paired butterfly-style structures. The input passes two branches that contain several convolution layers and LeakyReLU activation layer is followed for each layer. The two branches interact with each other through learnable combination coefficients. Specifically, given an input feature I, the first combination is

$$M_1(I) = I + a_1 J(I), \tag{1}$$

$$N_1(I) = a_2 I + J(I), \tag{2}$$

where $J(\cdot)$ denotes to the implicit non-linear function of several layers shown in Figure 1. Similarly, the second combination is

$$M_2(I) = b_1 N_1 + K(M_1(I)), \tag{3}$$

$$N_2(I) = N_1 + b_2 K(M_1(I)). \tag{4}$$

Afterwards, the output of two branches are merged in channel dimension and then compressed through a 1×1 convolution layer. The final output is

$$O = Conv(Concat(M_2(I), N_2(I))). \tag{5}$$

The combination coefficients are mainly determined in the following way. The mean and standard deviation in channel dimension are first obtained by global mean pooling in the upper branch and global standard deviation pooling in the lower branch. Then, those statistics in the two branches are passed through two fully connected layers, each followed by

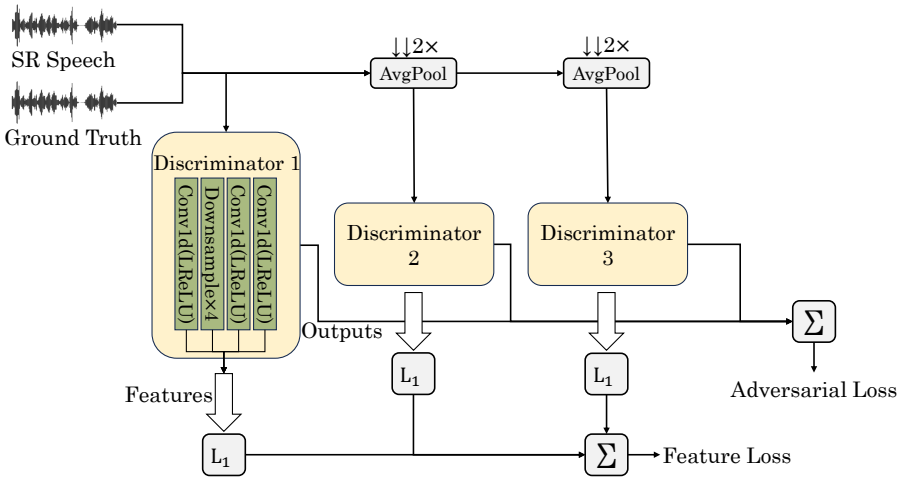


Fig. 2. The multi-scale discriminators.

ReLU and Sigmoid activation functions, respectively. Finally, the outputs of the two branches are averaged to obtain the combined coefficients.

2) Discriminators:

To implement multi-loss training in an adversarial manner, we employ multi-scale discriminators, whose structure is shown in Figure 3. The inputs to the discriminator system are super-resolution (SR) speech and high-resolution reference signals generated by the generator. It contains 3 discriminators (D_1, D_2, D_3), each with the same structure as in MelGAN. Specifically, there are 7 convolutional layers in each discriminator, 4 of which have downsampling capabilities. When the data pass through each layer, real and fake features at different scales will be produced, which are used to compute the feature loss. Also, the data passing through the discriminator will become an output, and the outputs will be used to compute the adversarial loss of the generator and discriminator. Furthermore, the inputs of D_1, D_2, and D_3 are original, 2-times down-sampled, and 4-times down-sampled waveforms, respectively. For interested readers, see more details of discriminators in [14].

C. Loss Function

The model is trained with an adversarial approach. We use a multi-scale STFT loss with FFT bins $\in \{512, 1024, 2048\}$ and hop length $\in \{50, 120, 240\}$ to form one part of the loss function. The window lengths are $\{240, 600, 1200\}$. On the other hand, the multi-scale adversarial and feature losses in time domain is added in. The total loss can be expressed as

$$\mathcal{L} = \mathcal{L}_{MSTFT} + \mathcal{L}_G^{adv} + \lambda_f \mathcal{L}_f, \tag{6}$$

where $\lambda_f = 100$, $\mathcal{L}_{MSTFT}$, $\mathcal{L}_G^{adv}$ and $\mathcal{L}_f$ are multi-scale STFT loss, adversarial loss of generator, and feature loss, respectively. Let $s(x, \theta_m)$ denote $|STFT(x)|$ with the mth hyperparameters θ_m, the multi-scale STFT loss is defined as

$$\mathcal{L}_{MSTFT} = E_{(x,y)\sim p_{data}} \left[\sum_{m=1}^{3} \left(\frac{||s(y,\theta_m) - s(x,\theta_m)||_F}{||s(y,\theta_m)||_F} + \frac{1}{N} ||log \frac{s(y,\theta_m)}{s(x,\theta_m)}|| \right) \right], \tag{7}$$

where $||\cdot||_F$ and $||\cdot||_1$ are Frobenius and ℓ_1-norms, N is the number of elements in the magnitude.

As shown in Figure 3, the latter two loss functions can be depicted as

$$\mathcal{L}_G^{adv} = E_{x\sim p_{data}} \left[\frac{1}{K} \sum_k max(0, 1 - D_k(G(x))) \right], \tag{8}$$

$$\mathcal{L}_f = E_{(x,y)\sim p_{data}} \left[\frac{1}{KL} \sum_{k,l} ||D_k^l(y) - D_k^l(G(x))||_1 \right], \tag{9}$$

where $k = 1, ..., K$ is the number of discriminators, $l = 1, ..., L$ is the number of layer in one discriminator.

III. Experiments

A. Dataset

We use the dataset provided by Deep Noise Suppression (DNS) Challenge at ICASSP 2023 [16] and Valentini-Botinhao [17] to generate training data containing noise. We synthesize clean and noisy speech pairs with 500 hours by randomly mixing the speech and noise, and each sample lasts for 5 seconds. The SNR of all samples are at -5 - 20 dB and the sampling rate is 16 kHz. We then downsample all the noisy speech by a factor of $s = 2$, i.e., turning the paired data into 8 kHz noisy speech and 16 kHz clean speech. For validation, we separate %10 (50 hours) of total data as validation set. For testing, no-reverb test set of DNS Challenge 2020 [18] and Valentini-Botinhao test set are used and the noisy samples are also downsampled if necessary.

TABLE I
TEST RESULTS OF NOISE-ROBUST BWE MODELS ON VALENTINI-BOTINHAO NOISY TEST SET DOWNSAMPLED TO 8 KHZ.

Method	PESQ-WB↑	STOI (%)↑	CSIG↑	CBAK↑	COVL↑	LSD↓
UEE [15]	2.23	93	2.27	2.39	2.17	2.72
MTL-MBE [8]	2.55	94	2.64	3.21	2.46	2.29
EP-WUN [9]	2.25	92	**3.50**	2.94	2.86	1.23
AFiLM + I-DTLN [4]	2.54	90	2.63	2.87	2.18	1.54
Ours	**2.67**	**95**	3.29	**3.32**	**2.92**	**1.16**

TABLE II
TEST RESULTS FOR DIFFERENT TASK ON DNS-CHALLENGE NO-REVERB TEST. "B" IS NOISE-FREE BWE, "D" IS DENOISE, AND "RB" IS NOISE-ROBUST BWE. "SOURCE" AND "NOISE" REPRESENT THE SAMPLING RATE OF INPUTS AND THE CASE WHETHER THE INPUTS CONTAIN NOISES.

Method	Task	Source	Noise	PESQ-NB↑	PESQ-WB	STOI(%)	CSIG	CBAK	COVL	LSD	MOS↑
WSRGlow	B	8 kHz	×	4.365	2.811	**99.4**	3.946	4.068	3.433	0.929	4.21
NU-Wave 2				4.353	2.646	**99.4**	3.663	2.869	3.209	1.328	4.08
AERO				4.369	3.295	98.5	**4.287**	4.273	3.844	0.802	4.27
Ours				**4.377**	**3.661**	98.6	4.103	**4.553**	**3.935**	**0.783**	**4.55**
DCCRN	D	16 kHz	✓	3.17	2.64	92.9	—	—	—	—	—
FullSubNet				3.28	2.72	95.3	—	—	—	—	—
DPT-FSNet				3.28	2.72	95.3	—	—	—	—	—
Ours				**3.29**	**2.80**	**96.0**	—	—	—	—	—
VoiceFixer	RB	8 kHz	✓	2.535	1.679	84.0	2.532	1.914	2.043	1.323	3.83
Ours				**3.554**	**2.777**	**97.1**	**3.313**	**3.532**	**3.063**	**1.218**	**4.38**
VoiceFixer	RB	4-16 kHz	✓	2.540	1.822	84.2	2.737	1.984	2.222	1.280	3.89
Ours				**3.550**	**3.013**	**97.3**	**3.657**	**3.726**	**3.355**	**1.112**	**4.43**

For the case of uncertain low sampling rate, we adopt same operation as above, except for downsampling. We use a filter with random parameters when doing downsampling, the types include *Chebyshev, Elliptic, Butterworth* and *Boxcar*, the order is a random integer from 2 to 10, the cutoff frequency is an integer from 2000 to 8000 Hz.

B. Training Details

Unlike the complex training policies such as multi-stage training, variable learning rate and warmming up used in previous work, our training method is single-stage and simple. We use an Adam optimizer ($\beta_1 = 0.8, \beta_2 = 0.999$) to optimize both generator and discriminator with a stable learning rate of 1×10^{-4}. We train the model on NVIDIA RTX3090 GPUs for 200 epochs and select the checkpoint which has the best performance on validation data to test. The FFT bins and hop length of the STFT operation in our network is 512 and 64 respectively.

C. Baselines

For baselines, we consider both cases of noise-robust and noise-free BWEs, and they are

1) noise-robust BWE: we compare our model with the previous SOTA methods of MTL-MBE and EP-WUN in [8], [9], in the same test set. Since the authors did not provide the source code, we have re-implemented the method proposed in [4] to produce the results. For uncertain input sampling rate case, the baseline is VoiceFixer [19].
2) noise-free BWE: we compare our model with WSRGlow [20], NU-Wave 2 [21], and AERO [11].

To further demonstrate the capability of our model, we also compare our model with 16 kHz to 16 kHz denoise-only models of DCCRN [22], FullSubNet [23], and DPT-FSNet [24]. For all the models on denoise task only, we have referred to the results provided in [25]. This comparison further showcases our model presents a high capacity of removing the noise and generating high quality speech.

D. Evaluation Metrics

To evaluate the quality of the generated speech, both objective and subjective evaluation metrics were used. The objective evaluation metrics used were PESQ [26] including narrow-band one (PESQ-NB, 0-8 kHz) and wide-band one (PESQ-WB, 8-16 kHz), STOI [27], CSIG, CBAK, COVL scores[28] and log spectral distance (LSD). The subjective evaluation metrics is overall Mean Opinion Score (MOS) [29]. We randomly selected 50 samples from the test set and asked 15 people to provide overall MOS values of each sample. The final MOS score is the average of these evaluations. For all metrics in this paper except LSD, the higher score means a better performance.

E. Results and Analysis

Table I summarizes the comparison between our model and noise-robust BWE methods. The results show that our SDNet performs best in all the metrics except for CSIG. This suggests that the audio generated with our method has a higher quality, which meets our expectation to our network and datasets

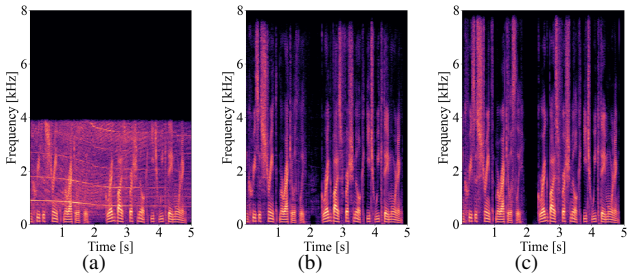


Fig. 3. Spectrograms of noise-robust BWE task results. (a) Input; (b) our method; (c) ground truth.

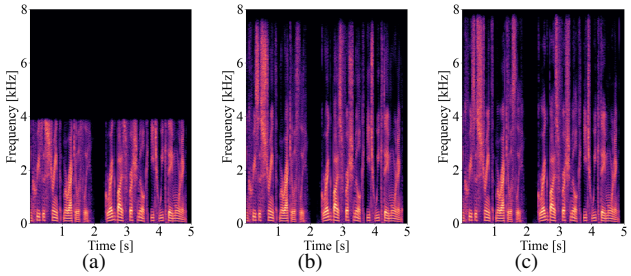


Fig. 4. Spectrograms of noise-free BWE task results. (a) Input; (b) our method; (c) ground truth.

designs. As for the mild drop in CSIG, this may because the over-suppression of noise hurts the speech component.

In 8 kHz to 16 kHz noise-free BWE task, our SDNet remains competitive in most metrics (see Table II), especially in PESQ-WB and CBAK, which shows that our generated wide-band signals have a better audibility and a more reasonable handling of background sounds. At the same time, it shows that our focus on BWE under noise does not sacrifice its performance in a noise-free environment.

In a comparison with VoiceFixer, our method outperforms it in all metrics for both 8 kHz to 16 kHz and 4-16 kHz to 16 kHz noise-robust BWE tasks. VoiceFixer aims to repair many distortions such as clipping, reverberation, and we find the speeches produced mismatch with the reference signal in terms of loudness, etc., which causes the degradation of its performance in objective metrics, but in subjective metrics, the scores of these speeches are still very high, which shows its repair is still very effective.

For the flexible sampling rate scenario (the last row in Table II), on the other hand, with a very limited bandwidth information, our model still improve the speech quality effectively, which is rare among the existing noise-robust BWE models. On 16 kHz to 16 kHz noise reduction task, our method performance better than the baseline denoise-only models, and surprisingly, the PESQ and STOI of the speeches generated by our model from 8 kHz noisy samples are even higher than the popular baseline models in Table II only focusing on 16 kHz to 16 kHz noise reduction.

For the same task, our model performs slightly better on the DNS test set than the Valentini-Botinhao's data set. This is because the speech lengths in the latter are generally very short, which prevents the model from utilizing contextual information to make more accurate predictions, resulting in a slight performance degradation.

We also tested our model with real-world data, and the results showed that the noise in these speeches was largely suppressed and the human voice was much clearer. Due to space limit, we provide the audio samples at our demo page. Figure 3 and 4 are the spectrograms of the results for different tasks.

TABLE III
ABLATION STUDY OF NETWORK STRUCTURE BASED ON 8 KHZ "RB" TASK OF DNS-CHALLENGE NO-REVERB TEST SET.

Method	PESQ-NB	PESQ-WB	LSD
SDNet	**3.554**	**2.777**	**1.218**
w/o GConv	3.445 (-0.109)	2.630 (-0.147)	1.256 (+0.038)
w/o LBs	3.442 (-0.112)	2.633 (-0.144)	1.262 (+0.044)
w/o Both	3.372 (-0.182)	2.538 (-0.239)	1.293 (+0.075)

F. Ablation Studies

In order to study the impact of network components on network performance, we conducted ablation studies by gradually removing some components of the original model. The experiments show that the original model produces the best results. When gated convolutions are replaced by general convolutions ("w/o GConv" in Table III), the performance of network declines due to the missing details at 6-8 kHz. When LBs only are removed, the performance is also degraded because of the lack of utilization of the time dimension information in the spectrogram. When both changes act together, the accuracy of the model drops dramatically.

IV. CONCLUSION & FUTURE WORK

In this paper, we proposed SDNet, a U-shaped encoder-decoder neural network to jointly handle BWE and denoise tasks in low sampling rate noisy environments. Experiments demonstrated that our model outperforms all baseline models in both objective and subjective metrics in different cases. The ablation studies have verified that our use of gated convolutions and LBs enhances the performance of the model. However, our model also have some limitations. When we try to train the model at a higher resolution such as 48 kHz, we find it is hard to deal with two tasks at one stage, and this is a common issue encountered by many models. In our future work, we will continue this task at higher resolutions and we will also consider adding music and other personalized datasets.

REFERENCES

[1] D. Haws and X. Cui, "Cyclegan bandwidth extension acoustic modeling for automatic speech recognition," in *Proc. ICASSP*, 2019, pp. 6780–6784.

[2] R. Yoneyama, R. Yamamoto, and K. Tachibana, "Non-parallel high-quality audio super resolution with domain adaptation and resampling cyclegans," in *Proc. ICASSP*, 2023, pp. 1–5.

[3] H. Seo, H.-G. Kang, and F. Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," in *Proc. ICASSP*, 2014, pp. 6087–6091.

[4] C.-W. Chen, W.-C. Wang, Y.-Y. Ou, and J.-F. Wang, "Deep learning audio super resolution and noise cancellation system for low sampling rate noise environment," in *2022 10th International Conference on Orange Technology (ICOT)*, 2022, pp. 1–5.

[5] N. C. Rakotonirina, "Self-attention for audio super-resolution," in *Proc. MLSP*, 2021, pp. 1–6.

[6] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," in *Proc. Interspeech*, 2020, pp. 2477–2481.

[7] T. Taher, N. Mamun, and M. A. Hossain, "A joint bandwidth expansion and speech enhancement approach using deep neural network," in *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2023, pp. 1–4.

[8] N. Hou, C. Xu, J. T. Zhou, E. S. Chng, and H. Li, "Multi-Task Learning for End-to-End Noise-Robust Bandwidth Extension," in *Proc. Interspeech 2020*, 2020, pp. 4069–4073.

[9] Y.-T. Lin, B.-H. Su, C.-H. Lin, S.-C. Kuo, J.-S. R. Jang, and C.-C. Lee, "Noise-Robust Bandwidth Expansion for 8K Speech Recordings," in *Proc. INTERSPEECH 2023*, 2023, pp. 5107–5111.

[10] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 9458–9465.

[11] M. Mandel, O. Tal, and Y. Adi, "Aero: Audio super resolution in the spectral domain," in *Proc. ICASSP*, 2023, pp. 1–5.

[12] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4471–4480.

[13] X. Luo, Y. Qu, Y. Xie, Y. Zhang, C. Li, and Y. Fu, "Lattice network for lightweight image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4826–4842, 2022.

[14] K. Kumar, R. Kumar, T. De Boissiere, *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[15] B. Liu, J. Tao, and Y. Zheng, "A novel unified framework for speech enhancement and bandwidth extension based on jointly trained neural networks," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2018, pp. 11–15.

[16] H. Dubey, A. Aazami, V. Gopal, *et al.*, "Icassp 2023 deep noise suppression challenge," in *ICASSP*, 2023.

[17] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks," in *Proc. Interspeech 2016*, 2016, pp. 352–356.

[18] C. K. Reddy, V. Gopal, R. Cutler, *et al.*, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech 2020*, 2020, pp. 2492–2496.

[19] H. Liu, X. Liu, Q. Kong, *et al.*, "VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration," in *Proc. Interspeech*, 2022, pp. 4232–4236.

[20] K. Zhang, Y. Ren, C. Xu, and Z. Zhao, "WSRGlow: A Glow-Based Waveform Generative Model for Audio Super-Resolution," in *Proc. Interspeech*, 2021, pp. 1649–1653.

[21] S. Han and J. Lee, "Nu-wave 2: A general neural audio upsampling model for various sampling rates," in *Proc. Interspeech*, 2022, pp. 4401–4405.

[22] Y. Hu, Y. Liu, S. Lv, *et al.*, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[23] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. ICASSP*, 2021, pp. 6633–6637.

[24] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *Proc. ICASSP*, 2022, pp. 6857–6861.

[25] Y. Li, Y. Sun, W. Wang, and S. M. Naqvi, "U-shaped transformer with frequency-band aware attention for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1511–1521, 2023.

[26] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[28] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[29] I.-T. P. 835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *ITU-T Recommendation*, 2003.

GLASS: Investigating Global and Local context Awareness in Speech Separation

Kuan-Hsun Ho *, En-Lun Yu †, Jeih-weih Hung ‡, Shih-Chieh Huang §, and Berlin Chen ¶
*†¶ Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
E-mail: {jasonho610, enlunyu, berlin}@ntnu.edu.tw
‡ Department of Electrical Engineering, National Chi Nan University, Taiwan
E-mail: jwhung@ncnu.edu.tw
§ Realtek Semiconductor Corp., Taiwan

***Abstract*—Previous speech separation systems commonly employ the Dual-Path (DP) mechanism. The DP mechanism addresses optimization challenges posed by considerable sequential input lengths, yet its compulsory interleaving pattern for local and global feature extraction raises concerns regarding optimal utilization of features across different layers. This study emphasizes the need for parallel processing of global and local information in speech separation, proposing the Global and Local context-Aware Speech Separation method (GLASS). GLASS integrates self-attention and convolutional layers into a parallel design, demonstrating state-of-the-art performance in both anechoic and noisy settings. The findings reveal patterns in the relevance of local and global information across layers, underscoring the significance of proper architecture in improving speech separation systems.**

***Index Terms*—Speech Separation, Global and Local Dependencies**

I. Introduction

In real-world environments, audio often contains parts where multiple speakers talk over each other. Therefore, accurately separating multiple speakers from a single-channel mixture would significantly facilitate many applications. Due to the outstanding modeling capability, deep-learning algorithms have been used as core models in state-of-the-art speech separation systems [1]–[9]. Moreover, most successful systems follow the encoder-decoder masking-based strategy [2], [3] and the Dual-Path (DP) mechanism [4].

The DP mechanism, proven effective in various speech-processing tasks such as separation [4], [7], [8] and enhancement [10], encompasses three stages: segmentation, block processing, and overlap-add. The segmentation stage splits a sequential input into overlapped chunks and concatenates all the chunks into a 3-D block. Afterward, the block is passed to local and global extractors in an interleaving fashion. Finally, the output from the last layer is transformed back to a sequential output with the overlap-add method. One of the objectives of the DP mechanism is to decrease the optimization difficulty that arises when the sequential input length is considerable.

However, the interleaving pattern of local and global feature extraction in the DP mechanism is compulsory. The fixed single-branch architecture translates a vague interpretation of how global and local relationships are utilized across different layers. Moreover, it hypothesizes that global and local information hold identical levels of importance since the numbers of global and local extractors are set the same. Without further exploration, one may naturally wonder if the arrangement is optimal, as questions like: Are global and local information equally crucial in every layer? If not, is it possible to identify a pattern?

Therefore, our intuition is to seek a scheme that retrieves global and local information in parallel and investigates the best possible way to utilize them while performing speech separation. Such an investigation is crucial yet not emphasized and discussed in previous works. Furthermore, the respective outcome can help researchers design better speech separation architectures. In light of this, we propose a novel Global and Local context-Aware Speech Separation method entitled GLASS. GLASS adapts Branchformer [11], which combines self-attention and convolutional layers into a parallel design, to the speech separation task. Self-attention is valuable for capturing long temporal relationships, while convolutional layers are competent to extract regional features within its receptive field.

Through extensive evaluation under anechoic and noisy settings, GLASS achieves state-of-the-art performance. Interestingly, we found a regular pattern of when local and global information matters and can even reform the model accordingly to gain improved results. This contribution underscores the significance of parallel processing and utilizing proper dependency for enhanced speech separation architectures, providing valuable insights for future research and design considerations.

II. GLASS

A. Overall Structure

The overall description of the proposed GLASS framework follows the encoder-decoder masking-based strategy, as depicted in Fig. 1. This strategy works as follows: Initially, the encoder transforms the mixture $x = \sum_{i=1}^{C} s_i + n \in \mathbb{R}^T$, which contains audio from C active speakers and additive noise n, into an STFT-like representation w that characterizes the signal:

$$w = \text{ReLU}(f_{\text{enc}} * x) \in \mathbb{R}^{N \times L}, \tag{1}$$

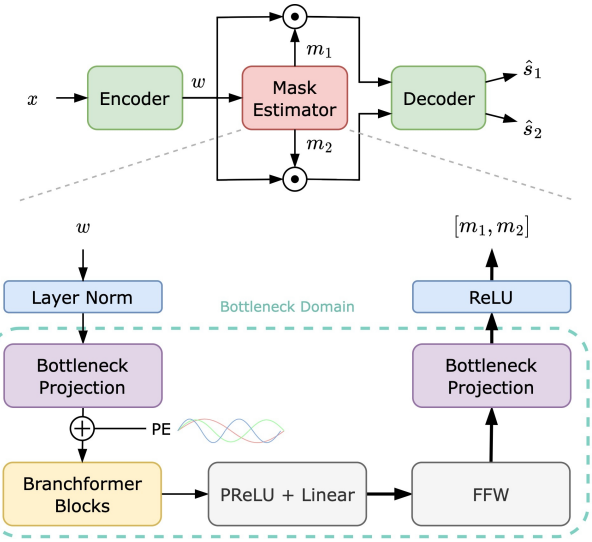


Fig. 1. The overall architecture of proposed method. This figure demonstrates an example of a mixture including two sources, and the bold arrows indicate that there are two vectors.

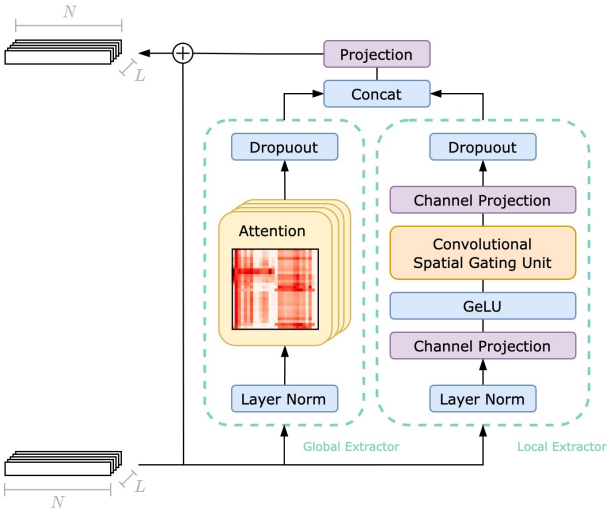


Fig. 2. A Branchformer block consists a global and a local context extractor.

where f_{enc}, $*$, and ReLU denote the encoding matrix, the convolution operation, and a rectified linear unit, respectively, and N and L are the feature size and the number of frames.

Then, by feeding w, the mask estimator produces C masks $\{m_i\}$ for each active speaker in the mixture. Finally, the decoder reconstructs each estimated source $\hat{s}_i$ by deconvolving the masked representation:

$$\hat{s}_i = f_{\text{dec}} * (w \odot m_i) \in \mathbb{R}^T, \tag{2}$$

where f_{dec} and $\odot$ denote the decoding matrix and element-wise multiplication. The analytic filters in f_{enc} and f_{dec} are learned from uni-dimensional convolution (Conv1d) and its transposition, respectively, both having N bases with a length of k. The objective is to minimize the distance between s_i and $\hat{s}_i$.

A more profound illustration of the mask estimator is also visualized in the lower part of Fig. 1. The mask estimator contains one pair of bottleneck projections, multiple Branchformer blocks, a linear layer, and feed-forward networks (FFWs). The representation w is first layer-normalized, and feature-wise downsized to a bottleneck dimension. Prior to the Branchformer blocks, positional encoding (PE) [25] is applied to inject information on the order of frame sequence. Afterward, the downsized representation feeds R Branchformer blocks to realize an enhanced vector space that alleviates the task difficulty for the subsequent linear layer. This linear layer is where the separation is carried out, transforming each of size N into a matrix of size $C \times N$. Ultimately, each of the C representations is transferred to FFWs, an upsizing projection, and a ReLU to generate a non-negative mask.

Branchformer was initially designed to tackle automatic speech recognition (ASR) and spoken language understanding (SLU) tasks. However, their objectives are quite different from ours. For example, the output dimension to which each task wishes to map differs. Speech separation or enhancement, in particular, remains the same dimension as the input audio, whereas ASR or SLU maps to tokens with fewer quantities. Therefore, we carefully modified Branchformer to meet our objectives.

B. Branchformer Blocks

Two parallel branches mainly constitute a Branchformer block, as depicted in Fig. 2. Both branches share the same input but focus on different relation ranges that complement each other. For an input w, the global branch outputs a globally-viewed tensor $u \in \mathbb{R}^{N \times L}$ produced by the global context extractor, whereas the local branch outputs a locally-tuned tensor $v \in \mathbb{R}^{N \times L}$ produced by the local context extractor. The outputs of two branches are then merged, with the original input added as a residual connection to smooth the output when sources are absent. At the beginning and end of each branch, layer normalization [12] and dropout [13] are employed, respectively. For brevity, we omit them in the following formulation.

Global Context Extractor. The global extractor exploits Multi-Head Self-Attention (MHSA) [25], which can be expressed as follows:

$$\begin{aligned} &Q_i = (\mathbf{W}_i^Q w)^T, K_i = (\mathbf{W}_i^K)^T, V_i = (\mathbf{W}_i^V)^T; \\ &\text{head}_i = \text{SA}(Q_i, K_i, V_i) = \mathcal{S}(\frac{Q_i K_i^T}{\sqrt{N}}) V_i; \\ &u = \text{MHSA}(w) = \mathbf{W}_{\text{head}}[\text{head}_0, ..., \text{head}_{h-1}]_{\mathbf{1}}^T, \end{aligned} \tag{3}$$

where $\mathbf{W}_i^{\{Q,K,V\}} \in \mathbb{R}^{N/h \times N}$ and $\mathbf{W}_{\text{head}} \in \mathbb{R}^{N \times N}$ are learnable linear layers, $\mathcal{S}(\cdot)$ denotes a softmax activation function, and h denotes the number of attention heads. The concatenation along the $(\mathbf{i}+\mathbf{1})$'th dimension is denoted by $[\cdots]_{\mathbf{i}}$.

Local Context Extractor. The local extractor exploits a convolutional-gated Multi-Layer Perceptron (cgMLP) [14], which contains one pair of feature-wise projections, a Gaussian

Error Linear Unit (GELU) [15], and a Convolutional Spatial Gating Unit (CSGU). The sequential procedure can be formulated as follows:

$$\begin{aligned}
&\tilde{w} = \text{GELU}(\mathbf{W}^{\text{up}}_{\text{mlp}} w), \quad \tilde{w} \in \mathbb{R}^{F \times L};\\
&[\tilde{w}_1, \tilde{w}_2]_{\mathbf{0}} = \tilde{w}, \quad \tilde{w}_i \in \mathbb{R}^{F/2 \times L};\\
&\tilde{w}'_2 = \text{CSGU}(\tilde{w}_2) = \text{DWConv}(\text{LayerNorm}(\tilde{w}_2));\\
&\tilde{w}'_1 = \tilde{w}_1 \odot \tilde{w}'_2;\\
&v = \mathbf{W}^{\text{down}}_{\text{mlp}} \tilde{w}'_1,
\end{aligned} \tag{4}$$

where F denotes the hidden dimension that is usually larger than that of input, DWConv denotes depth-wise convolution, and $\mathbf{W}^{\text{up}}_{\text{mlp}} \in \mathbb{R}^{F \times N}$ and $\mathbf{W}^{\text{down}}_{\text{mlp}} \in \mathbb{R}^{N \times F/2}$ are learnable linear layers as well.

Merging Branches. After obtaining u and v as in Eqs. (3) and (4), two merging methods are designed to gain an intermediate representation w' that captures global and local dependencies. The first is simply concatenating them and performing projection back to the original dimension by a linear layer $\mathbf{W}_{\text{cat}} \in \mathbb{R}^{N \times 2N}$:

$$w' = \mathbf{W}_{\text{cat}}[u, v]_0 \in \mathbb{R}^{N \times L}. \tag{5}$$

Concatenation-based merging is straightforward. However, due to its lack of flexibility, the alternative may be a weighted sum of two branches:

$$w' = \alpha_u u + \alpha_v v \tag{6}$$

The weights, α_u and α_v, indicate how global and local relationships are incorporated. This can be particularly useful in cases where the relative importance of the global and local information may vary depending on different layers. While there are multiple ways to learn α_u and α_v, attention-based pooling [16] is employed here as follows:

$$\begin{aligned}
\alpha'_u &= \mathcal{S}(\frac{\mathbf{W}^{u,1}_{\text{pool}} u}{\sqrt{N}})(\mathbf{W}^{u,2}_{\text{pool}} u)^T \in \mathbb{R};\\
\alpha'_v &= \mathcal{S}(\frac{\mathbf{W}^{v,1}_{\text{pool}} v}{\sqrt{N}})(\mathbf{W}^{v,2}_{\text{pool}} v)^T \in \mathbb{R};\\
[\alpha_u, \alpha_v]_{\mathbf{0}} &= \mathcal{S}([\alpha_u, \alpha_v]_{\mathbf{0}}) \in \mathbb{R}^2,
\end{aligned} \tag{7}$$

where $\mathbf{W}^{\{u,v\},\{1,2\}}_{\text{pool}} \in \mathbb{R}^{1 \times N}$ are learnable linear layers.

C. Comprehending Branches

If we regard every linear layer $\mathbf{W}$ as interpolation or subsampling, then the understanding of Branchformer's architecture can be comprehensible. For the global branch, each attention map, $\mathcal{S}(Q_i K_i^T/\sqrt{N}) \in \mathbb{R}^{L \times L}$, explains how the network adjusts its focus based on long-temporal information. In contrast, the local branch modifies features within a single frame. Specifically, CSGU processes the neighboring values via convolution kernels, resembling sub-band processing. At the same time, the gating design and up- and down-sizing projection layers can be interpreted as full-band processing.

III. Experiments

A. Datasets

We validate GLASS on the popular WSJ0-2mix dataset [3] under the anechoic setting and use the improvement of SI-SDR [17] and SDR as the evaluation metrics. WSJ0-2mix is generated from the Wall Street Journal (WSJ) dataset [18] and consists of mixed speech utterances from two distinct speakers with random SDR between 0 dB and 5 dB. The training, validation, and test sets contain 30, 10, and 5 hours of speech data. Furthermore, we perform experiments in noisy settings. We rely on WHAM! [19], where each two-speaker utterance from the WSJ0-2mix dataset is mixed with a unique noise sample recorded in ambient environments such as coffee shops, restaurants, and bars. The models are supposed to simultaneously perform speech separation and denoising to extract clean signals from such mixed data. Finally, all speech data are sampled at 8 kHz.

B. Model Configurations

GLASS is implemented using PyTorch [20], and the experiments are conducted using the Speechbrain toolkit [21]. For the encoder and decoder, we set the number of bases $N = 256$ and the kernel size $k = 16$ with a stride factor of 8. Regarding the mask estimator, the repetition of the Branchformer block R is 8, 12, and 16. We use $h = 8$ parallel attention heads inside each global branch and the hidden dimension $F = 2048$ for each local branch. The kernel size used in CSGU is 17. All the dropout layers are applied with a probability of 0.1. For model training, we optimize the model using the Adam optimizer [22] with a learning rate of $1.5e-4$, which is halved after 5 epochs without improvement. Finally, the models are trained over 150 epochs with Permutation Invariant Training [23] and SI-SDR losses [3].

IV. Results

A. Difference in Context Extractor

To demonstrate its effectiveness in modeling long- and short-term dependencies, we first compare GLASS with two context extractors: Transformer [25] and Conformer [24]. While Transformer is good at modeling long-range global context, it is less capable of extracting fine-grained local feature patterns. Conformer, on the other hand, employs a single-branch architecture, making it challenging to analyze how local and global interactions are utilized in different layers. Furthermore, similar to the DP mechanism, the fixed interleaving pattern between self-attention and convolution in Conformer might not be optimal.

The result is shown in Tab. I. We can observe that if the extractor readily possesses both local and global views, it is less favored by incorporating the DP mechanism. Positive examples are Conformer without DP mechanism and GLASS, while the negative example is Transformer without DP mechanism. Furthermore, while the DP mechanism is beneficial

TABLE I
THE SI-SDR AND SDR IMPROVEMENT SCORES OBTAINED BY GLASS AND DIFFERENT EXTRACTOR IN WSJ0-2MIX. GLASS USING CONCATENATION-BASED MERGING WITH 8 LAYERS IS ABBREVIATED AS "GLASS-C8", AND THE WEIGHTED-SUM ONE AS "GLASS-S8."

Extractor	Total layers	DP	SI-SDRi ↑	SDRi ↑
Transformer	8	×	17.6	18.0
		✓	18.8	19.1
Conformer	8	×	18.6	18.8
		✓	18.3	18.5
GLASS-c8	8	×	18.6	18.9
GLASS-s8	8	×	19.0	19.1

when integrated with Transformer (equivalent to a smaller-sized SepFormer [8]), probably due to the division of labor, GLASS using weighted-sum merging attains better outcomes.

B. Comparison with Previous Works

Here, we increase the number of the Branchformer blocks R in GLASS from 8 to 12 and 16 to gain a higher result. Tab. II compares the performance achieved by GLASS with the best results obtained by well-known methods tested on the WSJ0-2mix and WHAM!. As seen, GLASS families stand competitively; notably, "GLASS-s16" achieves state of the art. Although our method is slightly inferior to Wavesplit [9] for anechoic data, it still manages to discern noises and separate the clean sources better than Wavesplit on WHAM!, where source separation and speech enhancement are practiced simultaneously and thus closer to the real environment. The same explanation can also be given to DPTNet [7], which seems to be more parameter-wise efficient but fails for realistic data. It is noteworthy that Wavesplit and VSUNOS [6] leverage additional speaker information, either trained from scratch or with pre-trained embeddings.

Furthermore, the results from Tabs. I and II suggest that GLASS performs better with weighted-sum merging than with concatenation, which contradicts our initial expectations. We reckon the underlying cause is that concatenating the embeddings and projecting them back to the original dimension may lead to a certain loss of information. When performing one-to-more mapping as speech separation, as opposed to classification tasks like ASR and SLU, the loss of information is treated more severely and has to be addressed more carefully. The weighted-sum merging can mitigate this issue by providing a more controlled interpolation between the two embeddings.

C. Receptive Field in Local Branch

The choice of the sub-band range is crucial yet hard to determine in practice. As metaphorized in Sec II-C, the range of the sub-band translates to the receptive field in the context of Conv1d. We are curious about the impact of varying kernel sizes in Conv1d, and the findings are presented in Tab III. It shows diminishing rewards when the kernel size increases from 17 to 33, while performance is marginally worse when decreasing the kernel size from 17 to 9. A similar outcome can be found in [27].

D. Visualizing Branch Weights

As formulated in Eq. (7), the branch weights are determined by the given context and thus may be dynamic accordingly. This raises an interesting question of whether weights in the same layer are consistent or divergent regarding different inputs. To address such concern, we visualize each branch weights gained from 3000 testing trials via box plots, as plotted in Fig. 3.

We can observe four phenomena:

1) The first quarter of layers primarily retains local information while discarding global information (see data points cropped in the orange box).
2) Weights with high dynamic ranges are found in two to three levels preceding the last two layers, where many outliers exist (see data points cropped in the green box). It reveals that these layers adjust their priority to global or local embeddings when facing different inputs.
3) Aside from the layers mentioned above, the weights in the remaining layers are low-dispersive, indicating that the division of labor is distinct in each layer.
4) On average, the local information is utilized more in the overall percentage than the global information (see the dashed lines), which opposes to the hypothesis of equal importance.

We also conducted an experiment in which the global branches of the first four layers in GLASS-s12 were dropped since their weights were close to zero, according to Fig. 3. Relative to the original GLASS-s12, this arrangement yields comparable SI-SDRi performance (20.4 for WSJ0-2mix and 16.1 for WHAM), but converges faster. Due to the specific vision we have assigned for those layers, the subsequent layers can be more focused on their priorities and have a better grasp of what kind of embedding to employ. This highlights the importance of using global and local dependencies in an appropriate order, not to mention the reduction of parameters that leads to faster convergence. Furthermore, these findings suggest that an interleaving pattern of local and global extractors is suboptimal.

E. Ablation Study

Inspired by the visualization of the branch weights, a question naturally emerges: What if only one branch is used while the other is dropped? We conduct an ablation study, shown in Tab. IV, verifying whether both the global context extractor and the local context extractor are useful. It turns out that the performance of the model using only the local extractor is superior to the one using only the global extractor, even though Transformer is unarguably a more powerful design. The result gives hints on an essential and supplementary role for local and global features respectively, yet they complement each other when using both branches. The same finding can also be indicated from Fig. 3, where the overall weight for the local branch is greater than that of the global branch.

TABLE II
THE SI-SDR AND SDR IMPROVEMENT SCORES OBTAINED BY GLASS AND PREVIOUS WORKS IN WSJ0-2MIX AND WHAM!.

Method	Size	WSJ0-2mix		WHAM!	
		SI-SDRi ↑	SDRi ↑	SI-SDRi ↑	SDRi ↑
Chimera++ [26]	32.9M	11.5	12.0	9.9	-
Deep CASA [1]	12.8M	17.7	18.0	-	-
BiLSTM-TasNet [2]	23.6M	13.2	13.6	12.0	-
Conv-TasNet [3]	5.1M	15.3	15.6	12.7	-
DPRNN [4]	2.6M	18.8	19.0	13.9	-
SuDoRM-RF [5]	2.6M	18.9	-	-	-
VSUNOS [6]	7.5M	20.1	-	15.2	-
DPTNet [7]	2.7M	20.2	20.6	14.9	15.3
SepFormer [8]	26M	20.4	20.5	16.3	16.7
Wavesplit [9]	29M	21.0	21.2	15.4	15.8
GLASS-c12	14.8M	20.0	20.3	15.7	16.1
GLASS-s12	14.1M	20.3	20.5	16.1	16.2
GLASS-c16	19.9M	20.4	20.6	16.2	16.4
GLASS-s16	18.6M	20.8	21.0	16.5	16.8

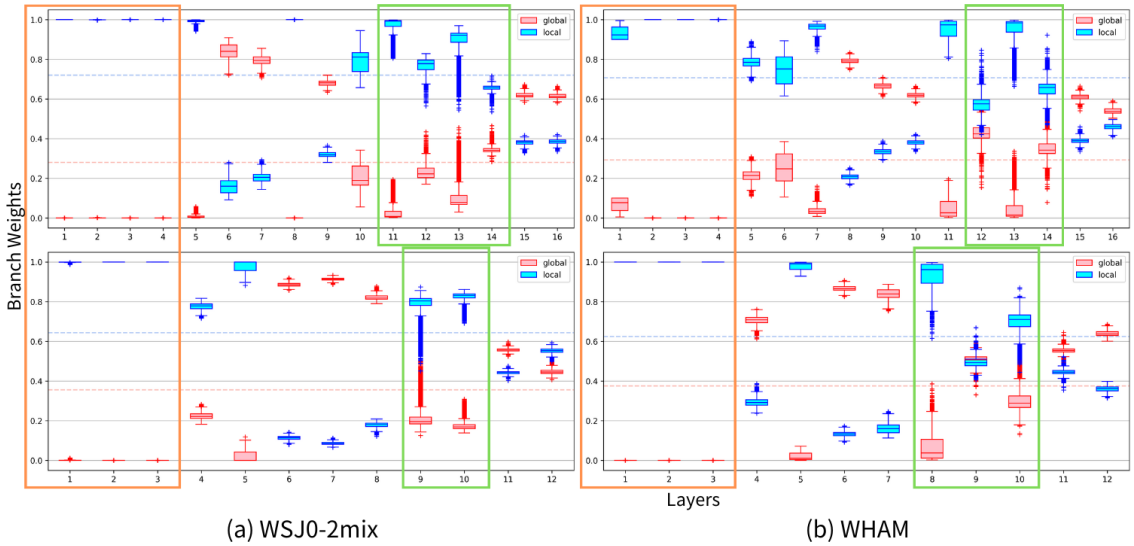


Fig. 3. The box plot of branch weights for model with 12 and 16 layers. '+' symbols denote outliers, and dashed lines plot the mean of all layer weights.

TABLE III
THE EFFECT ON DIFFERENT RESPECTIVE FIELD IN LOCAL BRANCH.

Model	kernel size			
	$k = 9$	$k = 17$	$k = 25$	$k = 33$
GLASS-s16	20.82	20.84	20.76	20.57

TABLE IV
ABLATION STUDY ON BRANCH DROPOUT.

local	global	SI-SDRi ↑	SDRi ↑
✓		18.75	18.91
	✓	18.48	18.72
✓	✓	20.84	21.01

F. Time complexity

In Fig. 4, we plot the execution time of each model. The first thing that comes into sight is the linear function against the step function. Being also categorized as a frame-online system, Conv-TasNet [3] has the same linear time complexity as GLASS. Regardless of the latency gap caused by their respective parameter size, GLASS holds a large margin of better separation quality than Conv-TasNet. On the other hand, the step function describes the chunked modeling of the DP-based method. Albeit this stepping characteristic helps reduce the time complexity if a long recording is readily in hand [28], the fact that it takes more time to separate short frames disqualifies the DP mechanism from being a frame-

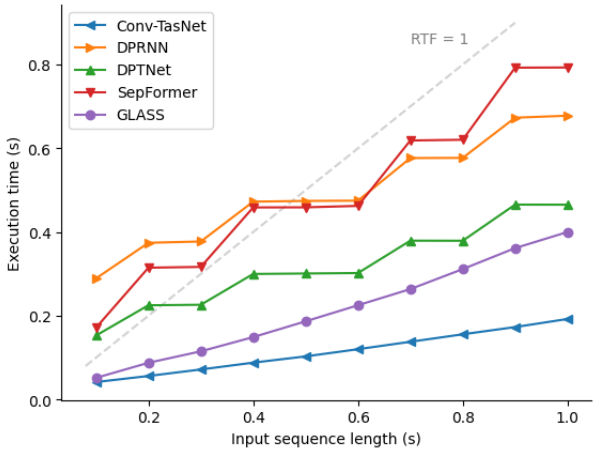


Fig. 4. Execution time, ran by GTX1080Ti, plotted as a function of the input length. RTF stands for real-time factor.

online application. Using GLASS, instead, reserves more spare time for downstream speech applications, say online speaker diarization [29], to operate.

V. Conclusions

In this study, we introduce GLASS, a novel approach that explores the use of global and local information in speech separation. GLASS employs a parallel design with branches extracting and merging features based on contextual importance. Evaluation results demonstrate the state-of-the-art performance of GLASS. The learned branch weights provide insights into an optimal pattern for integrating global and local information, enhancing interpretability and understanding in speech separation. Our work contributes to advance future endeavors in optimizing network architectures for speech separation.

VI. Acknowledgement

This work was supported in part by Realtek Semiconductor Corporation under Grant Number 112KK010. Any findings and implications in the paper do not necessarily reflect those of the sponsors.

References

[1] Y. Liu and D. Wang, "Divide and conquer: A deep casa approach to talker-independent monaural speaker separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, 2019.

[2] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network." in *Interspeech*, 2018.

[3] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal timefrequency magnitude masking for speech separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[4] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP*, 2020.

[5] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *MLSP*, 2020.

[6] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *ICML*, 2020.

[7] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Interspeech*, 2020.

[8] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention Is All You Need In Speech Separation," in *ICASSP*, 2021.

[9] N. Zeghidour and D. Grangier, "Wavesplit: End-to-End speech separation by speaker clustering," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

[10] Fu. Chao, J. Hung, T. Sheu, B. Chen, "A time-reversal enhancement network with cross-domain information for noise-robust speech recognition," *IEEE Multimedia*, vol. 29, no. 1, pp. 114-124, 2022.

[11] Y. Peng, S. Dalmia, I. R. Lane and S. Watanabe, "Branchformer: Parallel MLP-Attention Architectures to Capture Local and Global Context for Speech Recognition and Understanding," in *ICML*, 2022.

[12] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[14] J. Sakuma, T. Komatsu, and R. Scheibler, "MLP-based architecture with variable length input for automatic speech recognition," in *ICLR*, 2022.

[15] D. Hendrycks, and K. Gimpel, "Gaussian Error Linear Units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

[16] C. Wu, F. Wu, T. Qi, Y. Huang, and X. Xie, "Fastformer: Additive attention can be all you need," *arXiv preprint arXiv:2108.09084*, 2021.

[17] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *ICASSP*, 2019.

[18] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," *LDC93S6A. Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[19] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," in *Interspeech*, 2019.

[20] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] M. Ravanelli et al., "SpeechBrain: A General-Purpose Speech Toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, 2015.

[23] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[24] A. Gulati et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Interspeech*, 2020.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[26] Z.-Q. Wang, J. L. Roux, and J. R Hershey, "Alternative objective functions for deep clustering," in *ICASSP*, 2018.

[27] Y. Tsao, K. Ho, J. Hung, and B. Chen, "Adaptive-FSN: Integrating Full-Band Extraction and Adaptive Sub-Band Encoding for Monaural Speech Enhancement," in *SLT*, 2023.

[28] C. Li et al., "Dual-Path Modeling for Long Recording Speech Separation in Meetings," in *ICASSP*, 2021.

[29] E. Gruttadauria, M. Fontaine, S. Essid, "Online speaker diarization of meetings guided by speech separation," in *ICASSP*, 2024.

Temporal-Spatial Correlation Analysis for Ship-Radiated Noise Based on Random Matrix Theory

Qing Feng* and Zhiqiang Wu* Xuebin Li†and Heping Shen‡ and Shang Liu * and Min Tang* and Shengquan Feng‡
* PKU-Wuhan Institute for Artificial Intelligence
† Jilin Engineering Normal University
‡ Wuhan Puhui Ocean Opto-Electronic Technology Co., Ltd.

***Abstract*—Time-domain methods for temporal and spatial correlation estimation often face challenges due to background noise and environmental factors, leading to low time delay resolution. This limitation reduces the accuracy of underwater acoustic source localization and depth estimation. To tackle this issue, this paper introduces a statistical approach to estimate temporal and spatial correlations in ship-radiated noise. Using random matrix theory (RMT), we employ a factor model to determine the number of dynamic factors and the binary correlation structure of residuals from ship-radiated noise. This technique effectively captures the essential characteristics of underwater acoustic signal dynamics. Our proposed RMT-based modeling approach provides a robust analysis of ship-radiated noise which exhibits a wide frequency band and non-Gaussian nature. Results show that this method is highly sensitive and reliable in detecting variations in the temporal and spatial correlation of ship-radiated noise. The proposed method provides a novel insight into improving temporal and spatial correlation estimation and enhancing situation awareness in underwater environments.**

I. Introduction

The underwater noise received by passive sonar includes ambient noise, ship-radiated noise from vessels (ships, submarines, et al.), and self-noise originating from the vessel machinery. Ship-radiated noise, in particular, is generated by the vibration of the propeller and various mechanical parts of the vessel during navigation. These vibrations radiate sound waves through the ship's hull, converting them into underwater acoustic waves. As ship-radiated noise propagates through the complex marine environment, it exhibits non-stationary, non-Gaussian, and nonlinear characteristics, which make it challenging for traditional Gaussian-based methods to analyze the underwater acoustic signal effectively. Therefore, this paper investigates a statistical method for the analysis of ship-radiated noise with a wide frequency band and non-Gaussian.

This work is partially supported by the National Key R&D Program of China (No. 2022YFE0136800), the National Natural Science Foundation of China (No. U20A20162), Knowledge Innovation of Wuhan Science and Technology Bureau - Dawn Special Program (No. 2023010201020490), the Stable Supporting Fund of National Key Laboratory of Underwater Acoustic Technology (No. JCKYS2023604SSJS001), the Wuhan East Lake High-Tech Development Zone National Comprehensive Experimental Base for Governance of Intelligent Society.

Temporal and spatial correlations in underwater noise arise from the pathways of propagation and various noise sources[1]. Temporal correlation primarily results from the multipath propagation of underwater radiated noise. Variations in the ocean's physical properties, such as temperature, salinity, and pressure, affect the propagation process through mechanisms like refraction, reflection, and scattering. Environmental factors such as water depth, seafloor topography, and marine life further influence noise propagation. Additionally, ocean surface phenomena like breaking waves, wind, and rain also impact the propagation of underwater noise[1][2]. When a vessel emits underwater signals, the scattered and direct waves create multiple propagation paths. As a result, sound waves reach the receiver at different times and intensities. The echoes from these propagation paths may overlap and mix, causing the received signal to differ from the original in terms of time delay and amplitude. Consequently, the multi-path effect significantly influences the temporal correlation of underwater acoustic signals. Additionally, the interaction between various mechanical noises from the vessel and background noise further affects the correlative characteristics of ship-radiated noise.

The spatial correlation of ship-radiated noise is primarily generated by mutual interference between hydrophones in the underwater noise field. Studies have shown that in an underwater spatially inhomogeneous and anisotropic noise field, refraction can significantly impact spatial correlation performance based on receiver locations. This is because ship noise sources are distributed near the water surface and exhibit stratification[2][3]. The relative distance between receivers also affects the spatial correlation of underwater signals[3]. Another critical factor for the efficiency of the spatial correlation process is the stability of the receiver positions and the variability in receiver separation[3][4]. Correlation plays a crucial role in the study of underwater acoustic signals. By analyzing the spatial correlation matrix, researchers can assess the spectrum similarity measured at different times. Studies have demonstrated a certain degree of similarity between different frequency bands of underwater acoustic signals[2][5]. Spatial correlation is used for analyzing and estimating the travel time of acoustic signals[2][6].

Reference[7] employed spatial correlation to estimate the time Green's function. Additionally, reference[8] applied the spatial correlation technique to determine the ship noise spectrum and its modulation spectrum.

However, traditional methods of spatial correlation are primarily based on time-domain analysis. Due to the influence of complex background noise on underwater acoustic wave propagation, as well as the dependence of sound speed on temperature and seawater pressure, conventional methods for source localization and depth estimation typically rely on multi-path time delay and elevation angles obtained from hydrophone arrays. These time-domain analysis methods often result in low time delay resolution, thereby limiting the accuracy of the estimations.

To address this issue, we propose a modeling approach based on RMT. This method utilizes a factor model to analyze and construct the binary structure of residuals obtained from ship-radiated noise, revealing correlation variations in spatially inhomogeneous noise fields. This approach provides novel insights for underwater source localization and sonar detection. Implementing this method could potentially enhance the accuracy and reliability of these applications.

Recently, RMT has been a powerful mathematical tool for statistical data analysis and revealing correlations between observations. Such as detection techniques based on RMT have found applications in smart grids, intelligent healthcare, communications, and finance area[9][10][11][12]. Its robust performance under low signal-to-noise ratio conditions has been well-documented in communications[13]. In recent years, RMT has also been applied to underwater acoustic signal research. These applications include separating predominant single-scattering from weak multiple-scattering of ultrasound in weakly inhomogeneous media, isolating loud directional sources from the isotropic diffuse noise field in ocean environments, and distinguishing interference noise components from diffuse noise in weakly inhomogeneous ocean environments[14][15][16]. RMT is also utilized for noise field modelings, such as modeling sound scattering by random sea bottom roughness and simulating sound propagation in randomly inhomogeneous oceanic noise fields[17][18]. Furthermore, RMT has been employed to detect events using Marchenko-Pastur and Tracy-Widom distributions with random acoustic signal noise[19].

RMT is a data-driven and statistical method that offers advantages over traditional time-frequency-based analysis schemes for underwater acoustic signals in the complex ocean noise environment. By introducing RMT into underwater acoustic signal processing, especially for studying correlations in inhomogeneous noise fields and identifying targets in low signal-to-noise ratio environments, researchers can improve algorithm accuracy and enhance capabilities in weak signal detection. This approach also increases the robustness of the system, providing new research perspectives for passive sonar signal detection technology.

This paper presents a statistical method based on RMT for ship-radiated noise and proposes a modeling approach to estimate temporal and spatial correlations. Utilizing a factor model based on RMT, it estimates the number of dynamic factors and constructs a binary correlation structure of residuals from ship-radiated noise[20]. The key contributions of this paper are summarized as follows:

(1)Propose a statistical approach for analyzing temporal and spatial correlations of ship-radiated noise based on RMT. This method is robust against interference noise and data errors.

(2)Introduce a modeling technique using a factor model to enhance the accuracy and reliability of ship-radiated noise analysis in non-Gaussian underwater noise environments.

(3)The proposed method is highly sensitive to variations in the temporal and spatial correlations of ship-radiated noise, making it valuable for target detection and enhancing situational awareness in underwater environments.

(4))Provide new insight into temporal and spatial correlations in ship-radiated noise by extracting dynamic factors and constructing the binary correlation structure for residuals.

The remaining part of the paper is organized as follows: Section II provides a detailed problem description. Section III introduces the proposed method, including the theoretical spectral distribution of residuals and temporal-spatial correlation modeling and estimation. Experimental validation is presented in Section IV, followed by conclusions in Section V.

II. Problem Description

Assuming there are M hydrophones deployed in a vertical line array underwater, consider the ship-radiated noise $h_i(t)$ received by i^{th} hydrophone. $h_i(t)$ consists of the superposition of $\mathrm{m}_j(t)$, the noise radiated from N ships, and n_i, the uncorrelated ambient noise[8]. Thus, the ship-radiated noise received from i^{th} hydrophone can be represented as:

$$h_i(t) = n_i(t) + \sum_{j=1}^{N} \alpha_{ij} \mathrm{m}_j(t - \tau_{ij}) \tag{1}$$

Where α_{ij} represents the attenuation parameter, and τ_{ij} is propagation time from ship j to the hydrophone i, $\mathrm{i} \in 1 \ldots \mathrm{M}$, $\mathrm{j} \in 1 \ldots \mathrm{N}$.

Assume $h_i(t)$ exists temporal and spatial correlation in an inhomogeneous underwater environment. Consider two hydrophones separated by a distance L, from which signals $\mathrm{h}_u(t)$ and $\mathrm{h}_v(t)$ are collected from u^{th} and v^{th} hydrophones. The temporal correlation can be obtained as:

$$\mathrm{R}_{uu}(\tau) = \int_{-\infty}^{+\infty} \mathrm{h}_u(\mathrm{t}) \mathrm{h}_u(t - \tau) dt \tag{2}$$

Spatial correlation mainly results from mutual interference between two or more hydrophones in the underwater noise field. The spatial correlation R_{uv} of the underwater acoustic signals h_u and h_v is defined as:

$$\mathrm{R}_{uv}(\tau) = \int_{-\infty}^{+\infty} \mathrm{h}_u(\mathrm{t}) \mathrm{h}_v(t - \tau) dt \tag{3}$$

III. The Proposed Method

Traditional RMT-based studies for ship-radiated noise analysis typically assume Gaussian characteristics in homogeneous underwater noise fields. However, in spatially inhomogeneous noise fields, the noise exhibits non-stationary, non-Gaussian, and nonlinear characteristics, making it difficult for traditional Gaussian-based methods to effectively analyze the underwater acoustic signals. In this section, a factor model is applied to capture the temporal-spatial correlation structure of ship-radiated noise in inhomogeneous noise fields. Initially, the subspace of residuals is obtained by removing several factors from the data. Then, the model for temporal-spatial correlation is developed by deriving the theoretical spectral distribution of residuals. Finally, by minimizing the distance between the theoretical and empirical distributions, the temporal-spatial correlations of ship-radiated noise are determined.

A. Residual Formulation for Ship-Radiated Noise

The ship-radiated noise consists of mechanical noise, propeller noise, and hydrodynamic noise. Mechanical noise and parts of propeller have low-frequency line spectrum components that are periodic and stable. These components can be extracted using several principal components. Thus, the ship-radiated noise can be decomposed into principal components and residuals. Assume matrix D represents ship-radiated noise recorded by the hydrophone array, which can be formulated as: $D = U + \sum_{p=1}^{n} L_p F_p$, where L_p is the matrix of factor loading, F_p is the matrix of factor, and p is the number of factors to be subtracted from the underwater acoustic signal. These factors reflect the spatial correlation of the underwater acoustic signals. When the first P most important factors are removed from the passive acoustic signal, the remaining components, referred to as residuals, U, are represented as follows:

$$U = D - \sum_{p=1}^{n} L_p F_p \tag{4}$$

The residuals contain irregular noise components made up of various sound sources, such as ocean environmental noise and part of hydrodynamic noise. Unlike traditional Gaussian noise, the residuals typically exhibit the characteristics of nonlinear colored noise, which include acoustic information about the ocean environment and rich ship-related knowledge. This information reflects the interaction between the ships and the surrounding marine environment, such as the impact of ship turbulence, acoustic properties of the ship surface material, as well as the size and shape of the ship. Therefore, an in-depth analysis of the residual components can reveal complex dynamics of the interaction between the ocean environment and ships.

B. Temporal-spatial Correlation Estimation

The next step is to model the covariance structure of residual components U using (4) and then estimate the temporal-spatial correlation. Let A_N and B_N be N × N and T × T symmetric non-negative definite matrices, respectively. The matrix A_N represents temporal correlation, while B_N represents spatial correlation[20]. The binary structure of the residue is then characterized by:

$$U = A_N^{1/2} \varepsilon {B_T}^{1/2} \tag{5}$$

Where $A_N = \left\{ (A_N)_{ii} = 1, (A_N)_{ij, i \neq j} = \eta, i, j = 1, \cdots N \right\}$, and $B_T = \{ (B_T)_{\alpha} = \exp(-|h-t|/\alpha, h, t = 1, \cdots T \}$, ε is a $N \times T$ matrix with independent identical distribution. Thus, the covariance matrix of the residual component U using (5) is obtained as:

$$C = \frac{1}{T} U U^T = \frac{1}{T} {A_N}^{\frac{1}{2}} \varepsilon B_T \varepsilon^T A^{\frac{1}{2}} \tag{6}$$

To simplify the model, two assumptions are made: first, we assume that total spatial correlations are efficiently removed from p principal components. So $\mathrm{A}_N \approx \mathrm{I}_{N \times N}$ or $\eta = 0$; second, the temporal correlation is exponentially decreasing, represented by$\{\mathrm{B}_T\} = b^{|i-j|}$, with $|b| < 1$.

Subsequently, the N transform of the covariance matrix C is obtained by utilizing the cyclic property of the trace and the FRV multiplication law as follows:

$$\begin{aligned}
&\mathrm{N}_{\frac{1}{T} A^{\frac{1}{2}} \varepsilon B \varepsilon^T A^{\frac{1}{2}}}(z) = N_{\frac{1}{T} A^{\frac{1}{2}} A^{\frac{1}{2}} \varepsilon B \varepsilon^T}(z) \\
&= N_{\frac{1}{T} A \varepsilon B \varepsilon^T}(z) (cyclic \quad property \quad of \quad trace) \\
&= \tfrac{z}{1+z} N_A(z) N_{\frac{1}{T} \varepsilon B \varepsilon^T}(z) (FRV \quad multiplication \quad law) \\
&= \tfrac{z}{1+z} N_A(z) N_{\frac{1}{T} \varepsilon^T \varepsilon B}(z) (cyclic \quad property \quad of \quad trace) \\
&= \tfrac{z}{1+z} N_A(z) \tfrac{rz}{1+rz} N_B(\mathrm{rz}) N_{\frac{1}{T} \varepsilon^T \varepsilon}(z) (FRV \quad multiplication \quad law) \\
&= \tfrac{z}{1+z} N_A(z) \tfrac{rz}{1+rz} N_B(\mathrm{rz}) \tfrac{(1+z)(1+rz)}{z} \\
&= rz N_A(z) N_B(\mathrm{rz})
\end{aligned} \tag{7}$$

Where $N_{\frac{1}{T} \varepsilon^T \varepsilon}(z) = \frac{(1+z)(1+rz)}{z}$. By taking $M \equiv M(z)$ and $z = M = M(z)$ derived from N transformation $M_X(N_X(z)) = N_H(M_X(z)) = z$, the above equation (7) can be reformulated as follows:

$$\mathrm{N}_{\frac{1}{T} A^{\frac{1}{2}} \varepsilon B \varepsilon^T A^{\frac{1}{2}}}(M(z)) = rM N_A(M) N_B(\mathrm{rM}) = Z \tag{8}$$

Assuming that the spatial correlation of residual is eliminated, the spatial correlation matrix is satisfied $A = I_N$. Consequently, $N_A(z) = N_I(z) = 1 + 1/z$. Therefore, the equation (8) is conveniently reformulated as follows:

$$\mathrm{rM} = \mathrm{M}_B\left(\frac{z}{r(1+M)}\right) \tag{9}$$

Next, we need to determine M_B for equation (9). The Temporal covariance matrix B from equation (6) follows AR(1) process, which can be represented in the simple form $B = b^{|i-t|}$. By applying a Fourier transform to matrix B, the M transform of the matrix B can be obtained:

$$M_B(z) = -\frac{1}{\sqrt{1-z}\sqrt{1 - \frac{(1+b^2)^2}{1-b^2} z}} \tag{10}$$

where $a^2 = 1 - b^2$. Finally, a fourth-order polynomial is obtained as follows:

$$a^4c^2M^4 + 2a^2c(-(1+b^2)z + a^2c)M^3 + ((1-b^2)^2z^2 - 2a^2c(1+b^2) + (c^2-1)a^4)M^2 - 2a^4M - a^4 = 0 \quad (11)$$

Where $|b| < 1$, $a = \sqrt{1-b^2}$ and $c = \frac{N}{T}$.

To further determine the green's function of the binary structure model, it can be derived from the moment generating function $M(z)$ as follows:

$$G^{Model}(z) = \frac{M(z)+1}{z} \quad (12)$$

Finally, the spectral distribution of the covariance matrix given by equation(6), is estimated from the imaginary part of $G^{Model}(z)$:

$$\mathrm{H}^{Model}(b) = -\frac{1}{\pi}\lim_{\varepsilon\to 0^+} \Im m G(\lambda + i\varepsilon) \quad (13)$$

C. The Spectral Distribution Divergence

Assuming the underwater ambient noise remains relatively stable under normal conditions, we can use this typical ambient noise as a reference. The first step involves estimating an appropriate reference spectral distribution, denoted as H^{Model}, using the processes described in parts A and B. Next, we obtain the estimated spectral distribution, H^{Est} using online data. We then compare H^{Est} and H^{Model} by employing the Kullback-Leibler Divergence spectral distance as:

$$D_{KL}(H^{Est}||H^{Model}) = \frac{1}{2}\sum_i H_i^{Est}\log\frac{H_i^{Est}}{Q_i} + \frac{1}{2}\sum_i H_i^{Model}\log\frac{H_i^{Model}}{Q_i} \quad (14)$$

Where $Q_i = \frac{H_i^{Est}+H_i^{Model}}{2}$. The number of factors p and b approximate their true values when the divergence between the estimated and reference spectral distribution is sufficiently small. Here, p represents the spatial correlation parameter, and b represents the temporal correlation parameter.

IV. Experiment Validation

This section presents numerical results to illustrate the behavior of the temporal and spatial correlation in ship-radiated noise. Experiments are conducted to verify the performance of the proposed method.

A. Datasets

The dataset was collected in the South Sea of China between August 20 and 22, 2023. It was recorded using a vertical array of 64 hydrophones arranged in a line, as shown in Fig. 1, with a sampling rate of 32,000Hz. The hydrophone array is placed at a depth of approximately 400 meters. Data from the 64-channel array hydrophone was digitized and transmitted via a cable to a monitoring center for real-time processing and archival storage. The hydrophones were spaced approximately 6 meters apart along the vertical line. Then the original data underwent data cleaning and filtering processes. As a result, we obtained 28 hours of relatively pure ambient noise data as well as 40 hours of ship-radiated noise. In this section, we analyzed the correlation variation to gain a better understanding of the dynamic characteristics of underwater acoustic signals.

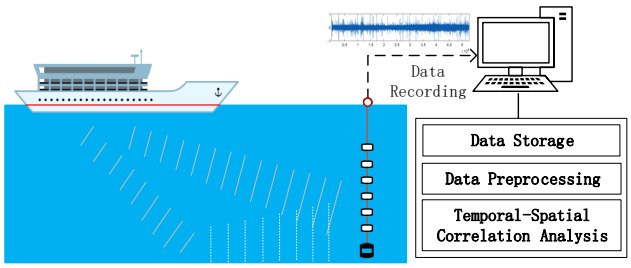


Fig. 1. Hydrophone setup is for capturing underwater ship-radiated noise.

B. Eigenvalue Distribution Analysis for Ship-Radiated Noise

In this subsection, we conducted experiments to validate the effectiveness of the proposed method for ship-radiated noise. Through detailed analysis and comparison of the eigenvalue distributions of the covariance matrix of residuals, we examined the statistical characteristics of the underwater acoustic signal and performed environmental perception. Fig. 2 illustrates the eigenvalue distribution before the ship entered the monitoring area. The window sizes are set to be 64x360. In contrast, Fig. 3 shows the eigenvalue distribution when the ship appeared in the monitoring area. The significant change in the eigenvalue distribution before and after the ship's appearance is evident. Therefore, the factor model effectively captures the statistical characteristics of ship-radiated noise in both states: before the ship's appearance and when a ship is present. The results indicate that during short-term events, such as ship appears, the eigenvalue distribution of the covariance matrix of residuals shows a sensitive response. This sensitivity enables the proposed method to detect variations in the ocean environment by analyzing the eigenvalue spectrum of the underwater acoustic signal.

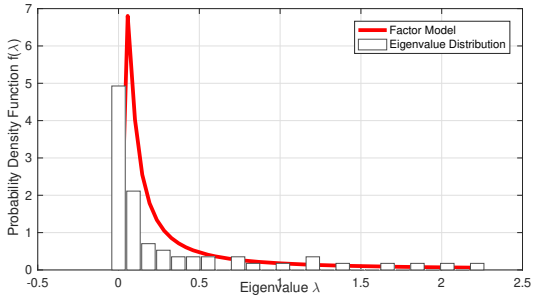


Fig. 2. Eigenvalue distribution before the ship's appearance.

C. Temporal and Spatial Correlation Estimation

In this subsection, we examine the temporal and spatial correlations in three scenarios: before, during, and after the

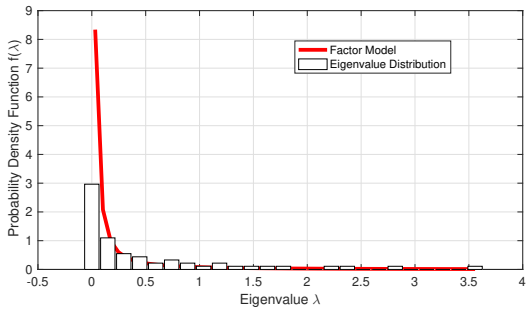


Fig. 3. Eigenvalue distribution when the ship appears.

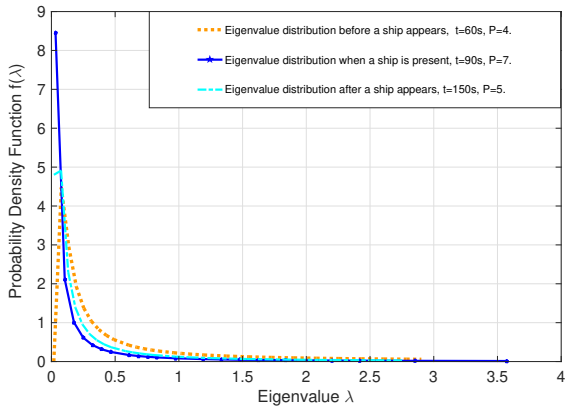


Fig. 4. Eigenvalue distribution and spatial correlation estimation.

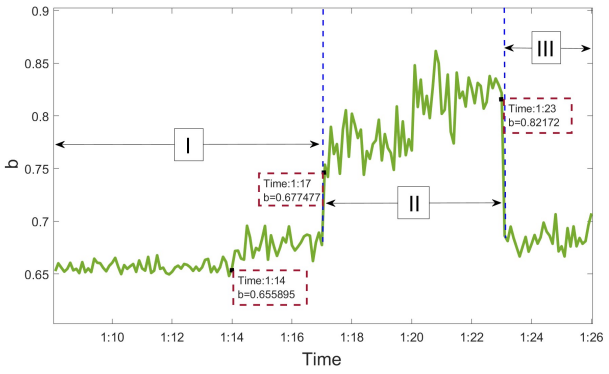


Fig. 5. Temporal correlation estimation for ship-radiated noise.

ship's appearance. We also estimate the temporal and spatial correlations in those three states, represented by the parameters b and p. In this experiment, we used a sliding window method to process the data with a window size of 64x320. First, we estimated the spatial correlation coefficient P at different times, as depicted in Fig. 4. The results indicate that spatial correlation parameter p increases when a ship is present at t=90 seconds. It is observed that p reflected spatial correlation and varies with different states of the underwater environment. As shown in Fig. 4, we compare the eigenvalue distribution of different states: before, during, and after the appearance of a ship. The findings demonstrate that the presence of the ship leads to an increase in spatial correlation and changes in the statistical characteristics of the data. This reveals that the proposed model can effectively capture the dynamic of the underwater acoustic signal.

In addition, we analyze the temporal correlation in different scenarios to reveal the characteristics of underwater acoustic signal dynamics, as shown in Fig. 5. A ship was present in the monitoring area at 1:17 AM on August 21, 2023. It is observed that the temporal correlation fluctuates with the appearance of the ship in the monitoring area. The temporal correlation parameter b is calculated for each set of correlation parameters and then averaged. The averaged values of b before (I), during (II), and after (III) the appearance of a ship are $\widehat{\mathrm{b}}$=0.6644, 0.7942, and 0.6842, respectively, shown in Fig.5. These results demonstrate that the temporal correlation significantly increases when a ship is present in the monitoring area. This indicates that the temporal correlation parameter b is sensitive to the changes in different states of the ocean environment.

V. Conclusion

This paper introduces a new statistical approach using RMT to estimate the temporal and spatial correlations in ship-radiated noise. It proposes a novel method for modeling ship-radiated noise using a factor model. The results demonstrate that the method effectively captures variations in temporal and spatial correlation of ship-radiated noise. By monitoring these correlations across different scenarios, we gain insights into the dynamic characteristics of underwater acoustic signals. This approach offers new perspectives for monitoring water areas and improving marine situational awareness. Future research should focus on evaluating how well subsequent pattern recognition systems can utilize the eigenvalue distribution characteristic derived from RMT as a feature for analyzing underwater acoustic signals.

References

[1] S. H. Kim, B. K. Choi, and B.-N. Kim, "Correlation between underwater noise and sea level at ieodo ocean research station," *Journal of Marine Science and Engineering*, vol. 9, no. 1, p. 1, 2020.

[2] M. J. Buckingham, "Cross-correlation in band-limited ocean ambient noise fields," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2643–2657, 2012.

[3] E. Skarsoulis and B. Cornuelle, "Cross-correlation of shipping noise: Refraction and receiver-motion effects," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. 3003–3010, 2019.

[4] B. Yang, X. Zhang, Y. Hou, P. Hu, and K. Wu, "Simulation research of ocean environment noise spatial correlation," in *2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, IEEE, 2019, pp. 1723–1727.

[5] S. M. Nichols and D. L. Bradley, "Use of noise correlation matrices to interpret ocean ambient noise," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2337–2349, 2019.

[6] C. Jin, F. Ye, H. Zhang, and X. Bao, "Travel time picking of ambient noise cross-correlation using a deep neural network combining convolutional neural networks and transformer," *Acta Geophysica*, vol. 72, no. 1, pp. 97–114, 2024.

[7] J. Zhou, "Analysis of ambient noise spectrum level correlation characteristics in the china sea," *IEEE Access*, vol. 8, pp. 7217–7226, 2019.

[8] L. Fillinger, A. Sutin, and A. Sedunov, "Acoustic ship signature measurements by cross-correlation method," *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 774–778, 2011.

[9] R. Qiu, L. Chu, X. He, Z. Ling, and H. Liu, "Spatio-temporal big data analysis for smart grids based on random matrix theory: A comprehensive study," *arXiv preprint arXiv:1708.04935*, 2017.

[10] P. Sarma and S. Barma, "Emotion recognition by distinguishing appropriate eeg segments based on random matrix theory," *Biomedical Signal Processing and Control*, vol. 70, p. 102 991, 2021.

[11] R. Qiu and M. Wicks, *Cognitive networked sensing and big data*. Springer, 2014.

[12] Q. Feng, G. Radman, and X. Li, "Early anomaly detection for power systems based on kullback-leibler divergence using factor model analysis," *American Journal of Electrical Power and Energy Systems*, vol. 10, no. 4, pp. 60–73, 2021.

[13] L. Zheng, R. C. Qiu, Q. Feng, and X. Li, "Shifting maximum eigenvalue detection in low snr environment," *arXiv preprint arXiv:1802.10325*, 2018.

[14] R. Menon, P. Gerstoft, and W. S. Hodgkiss, "Cross-correlations of diffuse noise in an ocean environment using eigenvalue based statistical inference," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3213–3224, 2012.

[15] A. Aubry and A. Derode, "Multiple scattering of ultrasound in weakly inhomogeneous media: Application to human soft tissues," *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 225–233, 2011.

[16] G. Li, J. Liu, and S. Zhang, "Enhancing cross correlations of ocean ambient noise in the time domain based on random matrix theory," *The Journal of the Acoustical Society of America*, vol. 152, no. 5, pp. 2849–2858, 2022.

[17] D. V. Makarov, P. S. Petrov, and M. Y. Uleysky, "Random matrix theory for sound propagation in a shallow-water acoustic waveguide with sea bottom roughness," *Journal of Marine Science and Engineering*, vol. 11, no. 10, p. 1987, 2023.

[18] D. Makarov, "Modeling of sound propagation in the ocean by means of random matrices," in *2017 Days on Diffraction (DD)*, IEEE, 2017, pp. 227–232.

[19] B. A. E. Bencharif, I. Ölçer, E. Özkan, B. Cesur, *et al.*, "Detection of acoustic signals from distributed acoustic sensor data with random matrix theory and their classification using machine learning," in *SPIE Future Sensing Technologies*, SPIE, vol. 11525, 2020, pp. 389–395.

[20] J. Yeo and G. Papanicolaou, "Random matrix approach to estimation of high-dimensional factor models," *arXiv preprint arXiv:1611.05571*, 2016.

Compressed Deepfake Video Detection Based on 3D Spatiotemporal Trajectories

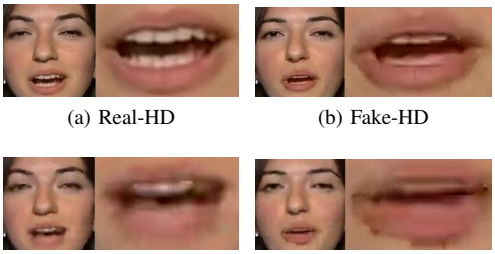
Zongmei Chen[1], Xin Liao[1†], Xiaoshuai Wu[1], Yanxiang Chen[2]
[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[2]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
E-mail: {chenzongmei, xinliao, shinewu}@hnu.edu.cn, chenyx@hfut.edu.cn
[†]Corresponding Author

***Abstract*—The misuse of Deepfake technology by malicious actors poses a potential threat to nations, societies, and individuals. However, existing methods for detecting Deepfakes primarily focus on uncompressed videos, such as noise characteristics, local textures, or frequency statistics. When applied to compressed videos, these methods experience a decrease in detection performance and are less suitable for real-world scenarios. In this paper, we propose a Deepfake video detection method based on 3D spatiotemporal trajectories. Specifically, we utilize a robust 3D model to construct spatiotemporal motion features, integrating feature details from both 2D and 3D frames to mitigate the influence of large head rotation angles or insufficient lighting within frames. Furthermore, we separate facial expressions from head movements and design a sequential analysis method based on phase space motion trajectories to explore the feature differences between genuine and fake faces in Deepfake videos. We conduct extensive experiments to validate the performance of our proposed method on several compressed Deepfake benchmarks. The robustness of the well-designed features is verified by calculating the consistent distribution of facial landmarks before and after video compression. Our method yields satisfactory results and showcases its potential for practical applications.**

I. Introduction

Deepfakes [1] is a compound word that is a combination of Deep Learning and Fake in the field of artificial intelligence (AI). It combine deep learning and fake media to create realistic but deceptive images and videos. This technology can alter a person's identity, mimic expressions, edit facial features, and even generate non-existent faces. Criminals exploit Deepfakes for telecom fraud, fake news, slander, and more, posing significant threats to individuals and society. Therefore, the research on Deepfake detection is of particular criticality.

Video compression on social networks is a common phenomenon. When users upload videos to social media platforms (such as Facebook, Instagram, Twitter, etc.), these platforms usually compress the video to reduce the file size and speed up the upload and playback. This compression usually uses different compression algorithms and parameters to balance video quality and file size. As depicted in Fig. 1, compression may cause the loss of video details, reduce the resolution, frame rate, and visual quality, causing the overlap of compression artifacts and tampering artifacts [2], thereby reducing the performance of existing Deepfake video detection methods.



(a) Real-HD (b) Fake-HD

(c) Real-Compressed (d) Fake-Compressed

Fig. 1: Visualization of real and fake videos in compressed and uncompressed states. The videos are from FaceForensics++. HD stands for high definition. Comparing the first and second columns, the lip borders and teeth of the fake video become blurred and tampering artifacts are present. Comparing the first and second rows, the lips and teeth lose their obvious shape in the compressed video, and compression artifacts appear. When compression artifacts and tampering artifacts coexist, the teeth are no longer visible and the shape of the lips has changed. These challenges lead to the low accuracy observed in current methods for detecting compressed Deepfake videos.

Current Deepfake video detection methods are mainly categorized into active defense and passive detection. Active defense methods prevent the generation of Deepfake videos by adding signal interference. However, the execution conditions of these methods are harsh, and the data source is not controllable. On the other hand, passive detection methods can be divided into learning from forged samples [3]–[6], learning without forged samples [7]–[9], task migration [10]–[12], and generate data-driven [13], [14]. These methods explore identifiable facial features from various perspectives, including spatial, frequency, multimodal domains, etc. However, due to the complexity of propagation scenarios and the presence of multiple adversarial factors, these methods show poor robustness and generalization ability in real forgery scenarios. Currently, there are also a few works targeting real-world scenarios, such as compression. F3-Net [15] enhances the frequency domain of frames affected by the Deepfake forgery process using frequency-aware decomposition and local frequency statistics.

This work is supported by the National Natural Science Foundation of China (Grant No. U22A2030), the National Key R&D Program of China (Grant No. 2022YFB3103500), and Hunan Provincial Funds for Distinguished Young Scholars (Grant No. 2024JJ2025).

However, this method achieves good detection results only for videos with a specific compression rate. FT-two-stream [16] proposes a dual-stream network for detecting compressed videos, but it obtains the compression dataset through hard coding and does not utilize compressed videos from real-world scenarios. To address this issue, Marcon et al. [17] download videos from social networks and fine-tune the network to detect compressed videos. Nevertheless, this method has a strong data dependency, and the influence of unknown platform types on model performance is significant. FAMM [18] characterizes facial muscle movements from a geometric perspective to detect compressed videos. However, these methods focus only on FaceForensics++ [19]. The visual quality of the videos in this dataset is generally poor. It is not comparable to the highly realistic faked data currently circulating in social networks.

To address compressed Deepfake videos in real-world scenarios, this paper proposes a detection method based on 3D spatiotemporal trajectories to enhance the detection performance and robustness of compressed Deepfake video detection. In detail, by studying facial motion, constructing motion features in the temporal domain and spatial domain, and performing time series analysis on phase space motion trajectories to realize the authenticity determination of Deepfake videos. The main contributions of this paper include:

1) We propose a spatiotemporal feature construction method based on the robust 3D model, which locates and tracks facial and head landmarks in the video, and combines the spatial dynamics and temporal characteristics of facial action units to construct the features.
2) A temporal feature analysis method based on phase space motion trajectories is proposed to model the facial change pattern between the first frame and each subsequent frame within a continuous period. It explores the temporal changes of facial coordinates in 3D space, and analyses the video's overall and global characteristics.
3) We conduct extensive experiments to verify the detection performance on compressed Deepfake videos and the results demonstrate our method achieves promising performance compared to the state-of-the-art methods. In addition, our method could avoid the effects of large head rotation angle or low illumination. It better resolves the task of detecting Deepfake videos in real scenes.

II. THEORETICAL ANALYSES

A. Analysis of compressed Deepfake videos

With the widespread dissemination of video content on social networks, a massive amount of unstructured video data has emerged. In this context, to ensure efficient transmission and storage of videos, social platforms commonly employ video compression technology. The primary purpose of video compression is to reduce the size of video files by removing redundant information (including spatial, temporal, and visual redundancies), thereby lowering transmission and storage costs. However, the compression process often introduces some noise and distortion. These not only result in compression artifacts overlapping with tampering artifacts in Deepfake videos but also hinder deep learning models from fully exploiting lost detail information. Moreover, extra noise interferes with model training, increasing the difficulty of detecting Deepfake videos.

TABLE I: Quality assessment of hard-coded video (FaceForensics++ LQ) and compressed video (FaceForensics++ Social).

Datasets	PSNR	SSIM	UQI	IEF	VIF	RECO
Raw vs LQ	32.06	0.84	0.99	0.91	0.48	0.85
Raw vs Social	31.06	0.75	0.82	0.93	0.80	1.00

This study randomly selected a total of 300 uncompressed videos, hard-coded compressed videos, and social network compressed videos from FaceForensics++. Six metrics, namely Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Universal Image Quality Index (UQI), Image Enhancement Factor (IEF), Visual Information Fidelity (VIF), and Relative Edge Change Ratio (RECO), were used to analyze the quality of videos before and after compression. PSNR is one of the commonly used metrics to measure image quality. It evaluates the distortion of images by calculating the mean squared error between the original image and the corrupted image, and then converts the result into a logarithmic ratio in decibels. Higher PSNR values indicate less image quality loss. SSIM is a metric used to measure the similarity between two images. It considers aspects such as brightness, contrast, and structure similarity, which are closely related to the human perception of image quality. The SSIM values range from -1 to 1, with values closer to 1 indicating higher similarity between the images. UQI is a universal image quality evaluation metric that comprehensively considers contrast, brightness, and structural information of images. Similar to SSIM, UQI evaluates image quality by comparing the similarity between the original and distorted images. IEF is used to evaluate the effectiveness of image enhancement algorithms. It compares the quality difference between the enhanced image and the original image to reflect the effectiveness of the enhancement algorithm. VIF is a metric used to evaluate video quality by comparing the structural similarity between the original video and the distorted video. RECO is also used to evaluate the impact of video distortion on edge information. It calculates the degree of video distortion by comparing the edge information between the original video and the distorted video. As shown in the Table I, the PSNR values of both compressed videos are below the common threshold of 40 dB, indicating significant visual quality loss and detail information missing after compression. SSIM, UQI, and IEF values below 1 indicate the introduction of noise in compressed videos. VIF values are 0.48 and 0.80, respectively, indicating the structural impact on compressed videos. RECO values are 0.85 and 1.00, respectively, indicating edge information loss in compressed videos.

B. Analysis of facial motion

In forged videos, there often exist unnatural facial movements, which may stem from multiple factors. Firstly, the forging process may lead to discontinuity or deformation of facial actions. In genuine videos, facial movements are typically

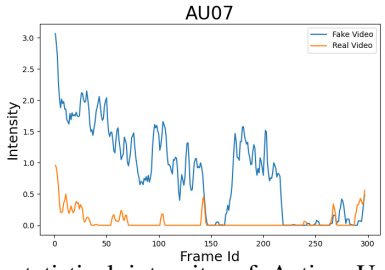


Fig. 2: The statistical intensity of Action Units (AU07) is tracked in 100 real videos and 100 fake videos, which describes upward eyelid movement. There is a significant contrast of the intensity between the real and fake videos.

continuous and natural, but in forged videos, due to potential abrupt changes, discontinuities, missing or distorted frames in each frame's synthesis, facial movements often appear discontinuous and lack smoothness. This lack of continuity is particularly evident in subtle facial movements such as eyelid movements. Secondly, the forging process results in a lack of coordination between facial movements and head postures. In genuine videos, facial movements often change with variations in head posture, reflecting the coordination between facial muscles and head movements. However, in forged videos, the limitations of synthesis algorithms may prevent accurate simulation of real head posture changes, or synchronization issues between facial and head movements may lead to a lack of coordination, resulting in unnatural facial expressions in the video. We utilized the open-source tool OpenFace 2.2.0 to track and locate facial features in the videos and analyze their temporal changes. We selected in 100 real videos and 100 fake videos from the FaceForensics++ (LQ) dataset for our study. Facial Action Coding System (FACS) is a system to taxonomize human facial movements by their appearance on the face. Movements of individual facial muscles are encoded by FACS from slight different instant changes in facial appearance. Using FACS it is possible to code nearly any anatomically possible facial expression, deconstructing it into the specific Action Units (AU) that produced the expression. It is a common standard to objectively describe facial expressions. In FACS, we focused on action units (AU07), which describes upward eyelid movement. Additionally, we examined head postures represented by Euler angles as

$$R = R_x \cdot R_y \cdot R_z \tag{1}$$

where R_x, R_y, R_z represent pitch, yaw, and roll, respectively. As depicted in Fig. 2, we observed relatively smooth temporal changes in genuine videos, while forged videos exhibited significant fluctuations and variations, highlighting their pronounced differences.

III. METHODOLOGY

A. Overview

Fig. 3 illustrates the proposed Deepfake video detection framework based on 3D spatiotemporal trajectories. The method consists of two modules: 3D spatiotemporal feature construction module and phase space motion trajectory analysis module. In the 3D spatiotemporal feature construction module, a robust 3D model is employed for facial landmark localization and tracking, concurrently tracking head movements decoupled from facial expressions. Subsequently, through the spatial dynamics and temporal combination of facial action units (AU), spatiotemporal motion features are constructed. In the phase space motion trajectory analysis module, we process motion features using time-delayed embedding techniques to reconstruct phase space trajectories. Subsequently, we use these reconstructed trajectory data to train a lightweight Transformer architecture for exploring spatiotemporal patterns of facial features. Finally, we apply Dempster-Shafer evidence theory to fuse the model results. The core idea of this approach hinges on employing a resilient 3D landmark localization and tracking for crafting spatiotemporal motion features. In addition, this approach decouples head movements from facial expressions, covering a wide range of head movements, to evaluate facial muscle movements in a more nuanced way. These enables effective resilience against the influence of compressed videos on model detection performance.

B. 3D spatiotemporal feature building module

The module utilizes a robust 3D model to localize and track facial landmarks and head movements, where head movements are decoupled from facial movements. Then the module selects facial points and head poses to construct phase space motion trajectories. These trajectories are used to characterize facial muscle movements in the temporal and spatial domains. Specifically, it first uses a 2D facial alignment algorithm to automatically locate 68 landmarks for each frame of the facial video. Secondly, the 3D facial model is used to estimate depth information from 2D frames, thereby enabling 3D landmark tracking. Finally, spatiotemporal domain motion features are constructed through the spatial dynamics and temporal combination of facial action units (AU).

1) Landmark localization and tracking: Landmark localization, a cascade of trained regressors is employed to achieve accurate positioning of facial landmarks in each video frame. The gradient tree boosting algorithm is utilized to train each regressor, employing an accumulated square error loss. The training dataset comprises pairs (I_i, S_i), where each I_i represents a facial image, and S_i denotes its corresponding shape vector. The initial shape estimation $S_i^{(0)}$ for each facial image is set to the mean shape of the training dataset: $S_i^{(0)} = \text{mean}(\{S_1, S_2, \ldots, S_n\})$. In each regression tree, the regression function r_t is learned using the gradient tree boosting algorithm, and the estimate for each shape is updated as follows:

$$S_i^{(t+1)} = S_i^{(t)} + r_t(I_i, S_i^{(t)}) \tag{2}$$

Notably, the initial shape selection at each level involves Histograms of Oriented Gradients (HOG) features for centering and scaling, ensuring comparability across all frames.

Landmark tracking, the facial landmark tracking algorithm achieves decoupling of facial expressions and head poses,

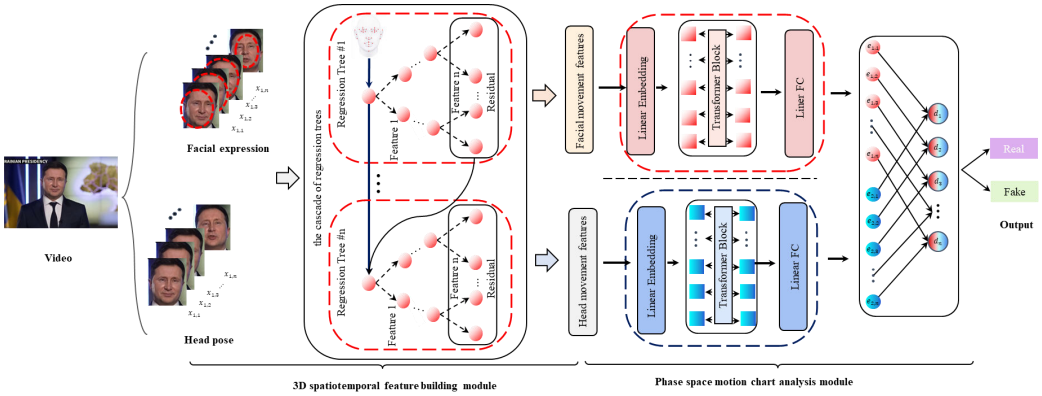


Fig. 3: The overview of compressed Deepfake video detection based on 3D spatiotemporal. It consists of the 3D spatiotemporal feature building module and the phase space motion chart analysis module.

eliminating interference from head movements such as translation, scaling (approaching or moving away from the camera), and rotation (rolling, yawing, pitching) on facial expression landmarks. First, depth information for each facial pixel is extracted from 2D video frames using a 3D morphable face model (3DMM). This model incorporates Principal Component Analysis (PCA) for facial shapes, allowing the reconstruction of a 3D face from a single 2D image. The PCA model comprises principal components $V = [\upsilon_1, \upsilon_2, \ldots, \upsilon_n]$, the mean value of all facial meshes $\bar{\nu}$, and their standard deviation σ_n. The shape of a novel face is then generated as follows:

$$S_i = \bar{\nu} + \sum_{n=1}^{N} \alpha_n \sigma_n \upsilon_n \tag{3}$$

where N is the number of principal components, and α_n represents S_i in the coordinates of the PCA shape space. The 3D face shapes are reconstructed by fitting 68 detected landmarks to a PCA shape model. In the 3D model fitting stage, we employ the gold standard algorithm to achieve a least squares approximation of the affine camera matrix given a set of 2D-3D point pairs. By applying a 3D geometric transformation matrix, we are able to convert each frame into a frontal face, enabling tracking and comparison of facial landmark movements throughout the entire video. Simultaneously, we introduce geometric constraints for facial landmarks to address issues with the low accuracy of facial landmark detection algorithms in frames with significant head rotation angles or insufficient lighting.

2) Construct spatial and temporal features based on 3D model: By analyzing the spatial and temporal sequences of Facial Action Units (AU), facial movements can be quantified more reliably and specifically, while also helping to reduce random noise in landmark positioning. We select facial landmarks related to facial expressions and head poses to construct the following features, with a focus on the movement patterns of blinking, eyebrow raising, eye movement, lip closure, and head poses. (1) Eyebrow Region: Vertical positions of the left and right eyebrows, angle difference within the eyebrows. (2) Eye Region: Differences in eye corners, horizontal and vertical

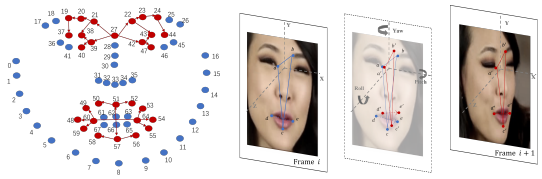


Fig. 4: Construct spatial and temporal features based on 3D model. Left: Distance features and angular features of the eyebrow, eye, and mouth regions. Right: Rigid displacement and rotation angle features of the head in 3D space.

distances between eye corners. (3) Mouth Region: Horizontal distance between mouth corners, vertical distance of the lips, average vertical position between both mouth corners. (4) Head Displacement: Rigid displacements of the head in the X, Y, and Z directions. (5) Head Rotation: Rigid rotations of the head in the roll, pitch, and yaw directions. Construct spatial and temporal features based on 3D model as shown in Fig. 4.

The method proposed in this paper presents several advantages over other approaches in constructing facial movement features. Firstly, it achieves the decoupling of head movements from facial expressions. This decoupling allows subsequent deep learning models to more effectively capture the characteristics of either head movements or facial expressions, enabling them to focus on specific learning tasks without being influenced by other factors. This not only reduces the overall complexity of the task and improves accuracy but also makes it difficult for creators of forged videos to evade analysis of head movements even if they attempt to modify facial expressions to deceive detection models. Secondly, traditional AU detection algorithms typically only apply to frontal views of the face, while the 3D model tracking method proposed in this paper can directly and continuously extract facial and head movement information from video data. It considers cases with significant head rotation angles and ensures continuous measurement of landmarks. This provides a foundation for a more comprehensive representation of AUs and their intensity, rather than just a few discrete values. Consequently, it effectively mitigates the impact of compressed videos on model detection performance.

C. Phase space motion chart analysis module

This method first processes the constructed features by introducing temporal delay embedding to reconstruct phase space trajectories. Secondly, a recursive graph (RP) is created to capture relationships between features. Finally, a lightweight Transformer architecture is employed to explore the differences in feature distributions between real and fake videos in both temporal and spatial domains. Existing detection methods mainly focus on temporal differences between adjacent frames, while our approach concentrates on modeling facial variation patterns between the first frame and each subsequent frame within a continuous time segment, emphasizing holistic and global video feature analysis.

The recursive graph (RP) is a visualization method that uses a binary square matrix to represent temporal dependencies between all states in time series data. Assuming that at times i and j the states of system X are represented by X_i and X_j, respectively, recurrence can be recorded through a binary function as follows:

$$R_{i,j}^{X} = \Theta(\epsilon_X - \|\mathbf{X}_i - \mathbf{X}_j\|_1), \quad \mathbf{X}_i \in R^m, \quad i,j = 1,\ldots,N \tag{4}$$

where Θ is a Heaviside function. For two time points i and j in the time series, if their similarity exceeds a predefined threshold ϵ_X, a point will be displayed at the corresponding position in the recursive graph (i.e., $R_{i,j}^{X} = 1$). By repeating this process over the entire time series, a matrix is created where each element represents the similarity between corresponding time points.

First, extract the feature motion trajectories between the first frame and subsequent frames of the video to construct a recursive graph. Secondly, convert the recursive matrix into the adjacency matrix of the network, representing the spatiotemporal neighborhood relationships between system states in the entire time series. Finally, to fully exploit spatiotemporal feature differences in real and fake videos, we design a lightweight Transformer classification model. The initial part of the model includes a linear layer to embed the input data into the hidden representation space. This allows the model to learn useful representations of the data without the need for manual feature engineering. Secondly, with the hidden dimension set to 128 and the number of encoder layers and attention heads to 2 each, a more lightweight Transformer model has been achieved. This model is suitable for practical deployment, particularly in scenarios with high task efficiency requirements and limited resources. Then, average pooling is employed to generate the final classification probability label instead of sequence-to-sequence output, thereby reducing the computational complexity of the model. Subsequently, calculate the loss between the predicted label and the actual label, and update the network parameters to complete the training. Finally, employ the Dempster-Shafer evidence theory to fuse the model results. In the testing phase, the algorithm receives the video to be tested as input and outputs the predicted label of the video. If the algorithm outputs 0, the video is predicted to be true; if the output is 1, the prediction is false.

TABLE II: Ablation study of the detection ACC (%) on facial expressions (FE) and head pose (HP).

Datasets	w/o FE	w/o HP	Ours
FF++HQ	94.10	92.59	**98.48**
FF++LQ	91.32	92.97	**97.47**
Celeb-DF-HQ	94.92	96.45	**99.37**
Celeb-DF-LQ	96.67	96.62	**98.28**

TABLE III: The comparison of the detection ACC (%) and AUC (%) with the state-of-the-art methods on FF++. The experimental setup strictly follows TALL-Swim [22] and adopts its experimental results.

Method	FF++HQ		FF++LQ	
	ACC	AUC	ACC	AUC
MesoNet [23]	83.10	-	70.47	-
Xception [4]	95.73	96.30	86.86	89.30
Face X-ray [7]	-	87.35	-	61.60
Two-branch [6]	96.43	98.70	86.34	86.59
Add-Net [24]	96.78	97.74	87.50	91.01
F3-Net [15]	97.52	98.10	90.43	90.43
FDFL [25]	96.69	99.30	89.00	92.40
Multi-Att [26]	97.60	99.29	88.69	90.40
FT-two-stream [16]	92.47	95.56	90.70	91.25
FInfer [27]	95.67	97.17	92.27	93.10
RECCE [28]	97.06	99.32	91.03	95.02
FAMM [18]	96.75	97.98	94.67	96.98
TALL-Swim [22]	**98.65**	**99.87**	92.82	94.57
Ours	98.48	98.82	**97.47**	**97.98**

TABLE IV: The comparison of the detection ACC (%) and AUC (%) with the state-of-the-art methods on Celeb-DF and DFDC. The experimental setup strictly follows BRCNet [29] and adopts its experimental results.

Method	Celeb-DF-V2-HQ		DFDC	
	ACC	AUC	ACC	AUC
Two-stream [3]	-	53.8	-	61.40
Xception [4]	97.9	99.73	79.35	80.92
Face X-ray [7]	-	-	72.65	89.5
Multi-Att [26]	97.92	99.94	76.81	90.32
Add-Net [24]	96.93	99.55	78.71	89.85
F3-Net [15]	95.95	98.93	76.17	88.39
FT-two-stream [16]	80.74	86.67	63.85	64.03
FInfer [27]	90.47	93.30	80.39	82.88
RECCE [28]	98.59	**99.94**	81.20	91.33
FAMM [18]	82.54	85.33	80.10	84.25
BRCNet [29]	98.73	99.94	81.65	91.89
Ours	**99.37**	99.56	**93.29**	**93.51**

IV. EXPERIMENTS

Datasets. We conducted experiments on FaceForensics++ (FF++), DFDC [20] and Celeb-DF [21] datasets. Videos from FF++ and Celeb-DF are compressed into two versions: medium compression (HQ) and high compression (LQ), using the H.264 codec with constant rate quantization parameters of 23 and 40 respectively. DFDC is the largest publicly available Deepfake detection dataset, which was released by Facebook. The implementation details, additional experiments, and analysis can be found at https://arxiv.org/abs/2404.18149.

Ablations. This paper characterizes the relative movement patterns of facial landmarks in both temporal and spatial domains by integrating facial expressions and head pose. We conducted ablation studies to assess the combination of facial expressions and head pose. Specifically, we utilized only facial expressions or head pose to represent facial motion, and the

results are presented separately in Table II. Experimental results indicate that the detection performance achieved by combining facial expressions and head pose surpasses that of using only facial expressions or head pose. The reason for this improvement may be attributed to their combination, which provides an excellent feature for facial motion.

Comparison with SOTA methods. Table III shows the ACC and AUC evaluation metrics of our method and existing methods on the FaceForensics++ dataset. We conducted comparative experiments on HQ and LQ. In the LQ dataset, our method is outperforming TALL-Swim [22] by 4.65% in terms of the ACC and 3.41% in terms of the AUC. And the performance on HQ is currently comparable to the existing methods. Table IV displays the ACC and AUC evaluation metrics of our method and existing methods on the Celeb-DF and DFDC datasets. Our method achieves an ACC value of 93.29 and an AUC value of 93.51 on DFDC, surpassing existing methods. And it demonstrates competitive detection performance on the Celeb-DF-V2-HQ.

Compared to the existing baselines, our method achieves nearly optimal performance in video compression experiments. We attribute this to the following reasons. On the one hand, almost no landmark errors are introduced during the video compression process, which does not affect the designed features. However, facial landmark errors are introduced during video tampering, altering facial motion patterns. Our method directly employs a robust 3D model to locate and track facial and head landmarks in videos, and then constructs features combining 2D and 3D frames. This enhances the robustness of the model. On the other hand, existing detection methods mostly rely on neural networks. The video compression introduces compression artifacts that coexist with tampering artifacts, which may mislead baseline learning. Therefore, our method performs better in detecting compressed videos.

V. CONCLUSIONS

This paper pioneers the migration of 3D models into the task of Deepfake facial landmark localization and tracking, constructing more robust facial motion features. Additionally, through a sequential analysis approach based on phase space motion trajectories, it explores the overall and global features of Deepfake videos. Finally, extensive experiments demonstrate that our method achieves state-of-the-art performance on compressed videos and performs well on uncompressed videos. At the same time, our method exhibits the highest detection efficiency and is more suitable for practical applications in real-world scenarios. Future work will focus on further improving the robustness of the algorithm and deploying it for widespread use in real-world scenarios.

REFERENCES

[1] "Deepfakes github [eb/ol]," in *https"//github.com/Deepfakes/faceSwap*, 2022.

[2] X. Zhang and X. Wu, "Multi-modality deep restoration of extremely compressed face videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, DOI:10.1109/TPAMI.2022.3157388.

[3] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, IEEE, 2017, pp. 1831–1839.

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[5] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[6] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, Springer, 2020, pp. 667–684.

[7] L. Li, J. Bao, T. Zhang, *et al.*, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[8] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 023–15 033.

[9] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 606–20 615.

[10] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[11] L. Shi, J. Zhang, and S. Shan, "Real face foundation representation learning for generalized deepfake detection," *arXiv preprint arXiv:2303.08439*, 2023.

[12] C. Kong, H. Li, and S. Wang, "Enhancing general face forgery detection via vision transformer with low-rank adaptation," *arXiv preprint arXiv:2303.00917*, 2023.

[13] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[14] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 710–18 719.

[15] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European conference on computer vision*, Springer, 2020, pp. 86–103.

[16] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1089–1102, 2021.

[17] F. Marcon, C. Pasquini, and G. Boato, "Detection of manipulated face videos over social networks: A large-scale study," *Journal of Imaging*, vol. 7, no. 10, p. 193, 2021.

[18] X. Liao, Y. Wang, T. Wang, J. Hu, and X. Wu, "Famm: Facial muscle motions for detecting compressed deepfake videos over social networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[19] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[20] B. Dolhansky, J. Bitton, B. Pflaum, *et al.*, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[21] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.

[22] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, and R. He, "Tall: Thumbnail layout for deepfake video detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 658–22 668.

[23] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: A compact facial video forgery detection network," in *Proceedings of IEEE International Workshop on Information Forensics and Security*, 2018, pp. 1–7.

[24] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2382–2390.

[25] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6458–6467.

[26] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.

[27] J. Hu, X. Liao, J. Liang, W. Zhou, and Z. Qin, "Finfer: Frame inference-based deepfake detection for high-visual-quality videos," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, pp. 951–959.

[28] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4113–4122.

[29] D. Zhang, C. Fu, D. Lu, J. Li, and Y. Zhang, "Bi-source reconstruction based classification network for face forgery video detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Low-resource Language Adaptation with Ensemble of PEFT Approaches

Chin Yuen Kwok*, Sheng Li†, Jia Qi Yip‡* and Eng Siong Chng*
* Nanyang Technological University, Singapore
† National Institute of Information & Communications Technology (NICT), Japan
‡ Alibaba Group, Singapore
E-mail: kwok0062@e.ntu.edu.sg

***Abstract*—Despite recent advances in automatic speech recognition (ASR) performance on common languages, a large fraction of the world's languages remain unsupported. Parameter-efficient fine-tuning (PEFT) methods are used to adapt these models to unseen languages by inserting language-specific modules into the models. To further improve adaptation performance, an ensemble of PEFT models can be formed, where the outputs of the ensemble can be aggregated to create the final prediction, and it has been shown that increasing the diversity of outputs from the ensemble produce can improve results. However, PEFT model ensembles have rarely been studied in the context of ASR despite its advantage of requiring significantly less memory for model storage, and the effect of using diverse PEFT methods to create diversity in PEFT ensemble model outputs remains unexplored. Specifically, it is unclear whether training with different PEFT methods improves diversity more than using the same PEFT method with different random seeds. To verify this, we examine whether a better model ensemble can be formed by combining models adapted by different PEFT methods instead of the same PEFT method. When adapting Whisper to 10 hours of data for each of the 3 unseen languages from Common Voice, results show that our ensemble with diverse PEFT methods consistently outperforms those that use the same PEFT method. Moreover, compared to the common approach of using fully fine-tuned models to form ensembles, our diverse PEFT ensemble can reduce the word error rate from 8.4% to 7.9% and reduces the memory consumption for model storage by 79%.**

I. Introduction

Automatic Speech Recognition (ASR) is the task of obtaining text transcriptions from a given segment of speech. Due to the availability of training data ASR models primarily only support commonly spoken languages such as English. This poses a significant challenge, as a large portion of the world's linguistic diversity remains unsupported by ASR models. To bridge this gap, researchers have explored techniques for adapting existing models to new languages. However, a major hurdle lies in the fact that most of these new languages are considered low-resource, meaning there's a scarcity of training data available for fine-tuning. This limited data makes it difficult to achieve optimal adaptation performance for these languages.

To efficiently adapt the ASR models to these low-resource languages, parameter-efficient fine-tuning (PEFT) methods [1] have been studied to add language-specific modules to the models, and multiple module designs have been previously explored. These include Sparse Subnetworks [2], which use modules in the form of a binary mask to selectively keep or remove each connection in a model; Low-Rank Modules [3], which are in the form of two low-rank matrices, and are applied to certain groups of model parameters like the linear projection; Prompting [4], which use modules in the form of a learnable vector, and the vector is merged with the model input; and Adapters [5], which use modules in the form of a small neural network, and the network is composed with selected model layers to modify the layers' outputs.

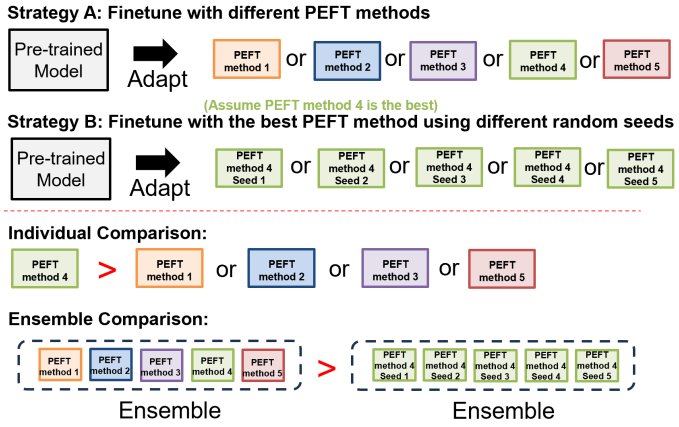


Fig. 1: *Overview of our PEFT model ensemble. Top: A pre-trained model is adapted using different PEFT methods. Bottom: Assume that PEFT method 4 is better than other PEFT methods when the models are used individually. However, an ensemble of models adapted by different PEFT methods produces better results than an ensemble of models adapted by the best PEFT method 4.*

As different PEFT methods have very different designs for the modules they use, previous works [6]–[8] have compared their effectiveness in the audio domain. Among these, a few works have studied PEFT for low-resource ASR. They include [9], which adds adapters and learnable bias into the ASR model, [10], which learns to fuse multiple sets of weights to create binary masks, and [11], which learns to perform fusion on multiple adapters.

To further improve adaptation performance, a PEFT model ensemble can be formed. Compared with the common approach of using fully fine-tuned models to form ensembles

[12]–[14], PEFT model ensembles has the extra benefit of requiring less memory for model storage as most of the model weights are identical among models. However, to the best of our knowledge, PEFT ensembles are rarely studied in the context of ASR and the interaction between the independently useful PEFT models within a model ensemble remains unexplored. Specifically, while previous studies have shown that diverse outputs can improve the performance of model ensembles [15], it is unclear whether training with different PEFT methods improves diversity more than using the same PEFT method with different random seeds. To verify this, we examine whether a better model ensemble can be formed by combining models adapted by diverse PEFT methods instead of the same PEFT method as shown in the bottom part of Figure 1.

Our contributions are four-fold. We show that: 1) An ensemble of models adapted by diverse PEFT methods consistently outperforms ensembles of models adapted by the same PEFT method. 2) When a model ensemble includes two models adapted by the same PEFT method, replacing one with another model adapted by a different PEFT method improves accuracy, even when the other model individually performs worse. 3) Our diverse PEFT ensemble can reduce the Word Error Rate from 8.4% to 7.9% while requiring about five times less memory compared with the common approach of using fully fine-tuned models to form ensembles. 4) PiggyBack (PB) [2] is the overall best PEFT method for low-resource language adaptation and only performs slightly worse than full fine-tuning.

II. Methodology

The following sections discuss five PEFT methods for adapting an ASR model to unseen low-resource languages. The PEFT methods include PB, SA, PA, LoRA, and LGate as shown in Figure 2. In addition, we propose an ensemble method to combine multiple models adapted by different PEFT methods as shown in Figure 1 to improve the adapted language performance.

A. Whisper Model for multilingual ASR

We use Whisper [16] as the base model as it has shown strong multilingual ASR (MASR) capability. The left part of Figure 2 shows an overview of its architecture. Whisper is a speech model trained with half a million hours of multilingual ASR and translation data. Its largest variant consists of 1.5 billion parameters. It supports MASR of 75 languages and also supports language identification (LID). It uses the Transformer [17] attention-based encoder-decoder architecture and decodes in an auto-regressive manner. Whisper is perferred for low-resource language adaptation because while [18] shows that Whisper performs similarly with XLS-R [19] in terms of WER averaged across seen and unseen languages, [20] shows that Whisper surpasses WavLM [21] overall in MASR adaptation and [22] shows that Whisper performs better than wav2vec 2.0 based models [23] and WavLM [21] for LID. In addition, Whisper decoder is preferred over the Large Language Model (LLM) based decoder [24] as it is found that a larger language model size does not benefit low resource languages much [25].

B. PiggyBack (PB)

To adapt an ASR model to new languages, one solution is to learn a Sparse Subnetwork [26]. As shown in Figure 2A, PB [2] learns a language-specific binary mask and derives a language-specific set of model parameters by taking a copy of the original model parameters and zeroing out part of it. More specifically, a trainable real-valued mask is first initialized to a value γ, and it is binarized using a threshold function:

$$y = \begin{cases} 1, & \text{if } x \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where τ is a hyper-parameter and affects the sparsity of the binary mask. The binary mask is then applied to the model to zero out partial model parameters. The gradients obtained through backpropagation of the training loss are used to directly update the real-valued mask weights.

C. Adapter

As an alternative to leaning a binary mask, as shown in Figure 2B, [27] learns a sequential adapter (SA) and adds it after the multi-head attention (MHA) or the feedforward network (FFN) module of a pre-trained network to modify the layer outputs. Specifically, SA consists of a feedforward (FF) layer to down-project the layer outputs into bottleneck features, a ReLU [28] activation function, and another FF layer to up-project the bottleneck features back to the original dimension. Similarly, Figure 2C shows parallel adapter (PA) [5], which adds the adapter in parallel to the MHA or FFN instead of inserting sequentially after them and behaves similarly to prefix-tuning [4], except that prefix-tuning uses Softmax instead of ReLU [28] as the non-linear activation function.

D. Low-Rank Adaptation (LoRA)

LoRA [3] is another approach to learning language-specific modules by adjusting the weight matrices of the FF layers in a model, and its structure is similar to PA. As shown in Figure 2D, LoRA learns a down and up projection layer like PA. Unlike PA, LoRA is placed in parallel to the FF layers in the model instead of the MHA and FFN. Also, LoRA does not have a non-linear activation function between the projection layers.

The forward pass of a FF layer modified with LoRA yields:

$$h = W_0 x + \Delta W x = W_0 x + W_{up} W_{down} x \tag{2}$$

where random Gaussian is used for initializing W_{down} and zero for W_{up}, so $\Delta W = W_{up} W_{down}$ is zero at the beginning of training.

To further adapt the FF layers, as shown in Figure 2E, [29] adds a gating mechanism before and after each FF layer to modulate their input and output. Specifically, each layer's input and output is multiplied with a separate learnable vector of the same dimension. The approach is inspired by previous work [30], which uses the language identity of the input

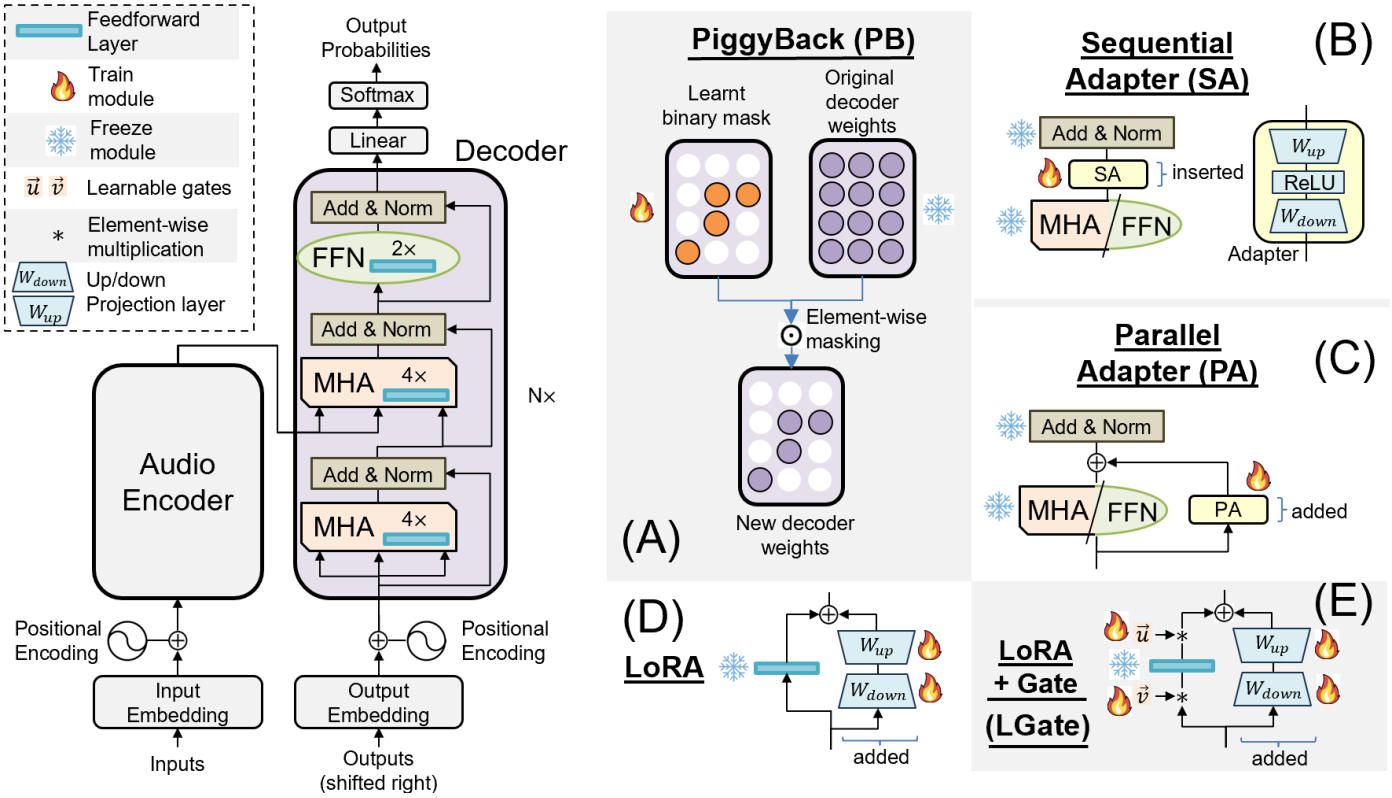


Fig. 2: *Overview of the attention-based encoder-decoder model architecture for ASR (left) and different PEFT methods (right).*

audio to select language-specific gates to modulate internal representations of a multilingual ASR model; and [31], [32], which learns a language code to gate the activity of neurons in an acoustic model. We refer to this method as "LGate".

E. Model Ensemble

To further enhance the adapted language performance, we study the strategy to perform model ensemble, which is a method to combine multiple models, such that the combined system can perform better than any of the individual models. Specifically, we study the strategy of combining multiple models adapted by different PEFT methods instead of models adapted by the same PEFT method to form a model ensemble, as shown in the bottom part of figure 1. As for the model ensemble method, we use ROVER [33]. It is a post-processing system that combines multiple ASR outputs, which uses a "voting" process to reconcile differences in ASR system outputs to obtain a composite ASR output with a lower error rate.

Our method is different from previous works [10], [11] in that we combine models adapted by different PEFT methods, instead of models adapted by the same PEFT method, to form model ensembles. In addition, compared with other approaches which combine models adapted by full fine-tuning to form model ensembles, our approach requires significantly less memory as we only need to store the extra language-specific PEFT modules for each model in the ensemble.

III. Experiments

A. Dataset and Model Details

We implement our methods based on the popular SpeechBrain [34] toolkit and CL-MASR [20].

Following previous work [10], [20], we evaluate our method using a subset of the widely used large-scale CommonVoice dataset[1] [35]. We follow the CL-MASR benchmark [20] and select three unseen languages Esperanto (eo), Interlingua (ia), and Frisian (fy) for Whisper adaptation. Following the benchmark, each language contains 10 hours of data for training, 1 for validation, and 1 for testing.

For each unseen language, we adapt whisper-small and whisper-large-v2 for two epochs with a train batch size of 4. Specifically, we freeze the Whisper encoder and only adapt the decoder to unseen languages. PEFT methods are applied to add language-specific modules to the decoder, and only the parameters of the modules and the text embedding layer at the decoder's side are updated. We use AdamW [36] as the optimizer and a variant[2] of the ReduceLROnPlateau[3] learning rate (LR) scheduler. Validation is done at an interval of $1/32$ epoch. We sweep through the hyper-parameters to tune them for all methods.

[1]https://commonvoice.mozilla.org/en

[2]https://speechbrain.readthedocs.io/en/latest/_modules/speechbrain/nnet/schedulers.html\#NewBobScheduler

[3]https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html

B. Results and Discussion

We show both the results of adapting whisper-small and whisper-large-v2 to new languages Esperanto, Interlingua, and Frisian as shown in Table I. To expand Whisper's language support, a naive solution is to duplicate the model, fully fine-tune (FT) it on new languages, and keep the original copy for old languages. However, this requires double the memory to store both models. To address this, PEFT methods are used to reduce the memory requirement, although the target language WER is compromised as a trade-off. Among the PEFT baseline methods, PB shows the lowest Average Word Error Rate (AWER) on new languages, where it achieves an AWER of 20.9% on whisper-small and 13.7% on whisper-large-v2, which is competent with the result of FT. In addition, our PEFT baseline methods requires more memory than usual as the text embedding layer weights are also stored. We will show later in Section III-C that adapting the embedding layer is necessary to improve the adaptation results for new languages.

For the model ensemble baseline results, three models are adapted by FT using different random seeds to form a model ensemble (ROVER-3a), and it slightly reduces the AWER from 13.7% to 13.4% for whisper-large-v2 compared with single-model FT. Similarly, three models are adapted by PB (ROVER-3b) and it slightly reduces the AWER from 13.7% to 13.5% for whisper-large-v2 compared with single-model PB. The results suggested that combining models that are adapted by the same method to form ensembles only slightly improves the WER. We further test with increasing the model ensemble size from three to five (ROVER-5a), and although AWER is reduced for whisper-small, the AWER for whisper-large-v2 is unaffected. This shows that the improvement made by adding more models to an ensemble may saturate at some point.

Next, we adapt 3 models using PB, PA and LoRA respectively to form a model ensemble (ROVER-3M) and adapt 5 models using PB, PA, LORA, SA and LGate to form an ensemble (ROVER-5M). For whisper-small, results show that our method performs worse than FT. We hypothesize that this is because applying PEFT methods to smaller models is less effective, as shown by the larger AWER gap between FT and the PEFT baseline methods. For whisper-large-v2, results show that we can reduce the AWER of the model ensembles from 13.4% to 13.2% if we use 3 models adapted by diverse PEFT methods, instead of 3 models adapted by PB, to form the model ensemble. The AWER can be further reduced from 13.2% to 12.9% as more models adapted by other PEFT methods are added to the ensemble.

Lastly, the results show that our method ROVER-5M reduces AWER from 13.4% to 12.9% and reduces memory consumption by 79% compared to ROVER-5a. In addition, our method reduces AWER from 13.7% to 12.9% and reduces memory consumption by 37% compared to FT. The results suggest that different PEFT methods can complement each other to improve the performance of model ensembles and reduce the memory consumption for storing the ensembles.

C. Ablation study

We further perform an ablation study on fine-tuning the text embedding layer at the decoder as shown in Table II. Results show that WER is increased for both whisper-small and whisper-large-v2 if the embedding layer is frozen during FT. This shows that adapting the embedding layer is necessary for low-resource language adaptation.

In addition, we perform an ablation study on combining different models to form model ensembles and show the results in Table III. First, we observe that the performance of a model ensemble heavily depends on the best single-performing model inside the ensemble. This is shown by the results in row 1, which is obtained by combining three models adapted by PA, SA, and LoRA. This model ensemble performs the worst among all other ensembles because it does not include models adapted by PB, which is the best single-performing model as shown in Table I, while the other ensembles do.

Second, results in rows 2-3 show that combining the models adapted by PB, PA, and SA or the models adapted by PB, LoRA, and LGate to form an ensemble have worse performance than other ensembles that have PB-adapted models included. We hypothesize that this is because the models used in rows 2-3 are adapted by less diverse PEFT methods. Specifically, there is less diversity as row 2 uses PA and SA, and row 3 uses LoRA and LGate. These PEFT method pairs are similar in design as shown in Figure 2, so the models adapted by these methods may acquire similar knowledge that complements less with each other. *The results suggested that more diverse PEFT methods are needed to form model ensembles with better accuracy performance.*

The hypothesis that using models with less diverse PEFT methods will lead to inferior results is further supported by the results in row 4 and 7. For row 4, a model ensemble is formed by combining two models adapted by PB and one model adapted by PA. After replacing one of the PB-adapted model to a LoRA-adapted model as shown in row 7, the AWER is reduced from 13.5% to 13.2%. However, LoRA has been shown to perform worse than PB and PA as shown in Table I. *This shows that a model ensemble can benefit from the diversity of PEFT methods even when one of the models in the ensemble is replaced by another model with worse performance.*

D. Increasing model size

We emphasize that simply increasing the model size for adaptation does not guarantee performance improvement. For example, PB has the parameter size fixed by definition, and increasing the bottleneck dimensions of the sequential adapters has diminishing effects on performance improvement [27]. Our paper studies how to obtain SOTA results beyond this limit using speech foundation models, and our diverse PEFT ensemble can reduce the Word Error Rate from 8.4% to 7.9% while requiring five times less memory than the common approach of using fully fine-tuned models to form ensembles.

TABLE I: *WER (%) of adapting whisper-small and whisper-large-v2 to three unseen languages: Esperanto (eo), Interlingua (ia), and Frisian (fy). The "None" method refers to not performing adaptation, so Whisper is not trained to transcribe unseen languages.*

Method	**whisper-small**					**whisper-large-v2**				
	Mem	**eo**	**ia**	**fy**	**avg**	Mem	**eo**	**ia**	**fy**	**avg**
FT	2x	17.6	12.3	29.0	19.6	2x	12.2	8.4	20.6	13.7
PEFT baselines										
PB [2]	1.18x	18.9	14.5	29.2	20.9	1.06x	12.0	8.4	20.8	13.7
SA [27]	1.18x	21.6	17.5	34.2	24.5	1.05x	13.9	9.4	23.3	15.5
PA [5]	1.18x	20.8	16.3	33.3	23.5	1.05x	13.3	9.4	22.4	15.0
LoRA [3]	1.18x	21.8	18.8	33.1	24.6	1.05x	14.2	12.6	23.0	16.6
LGate [29]	1.18x	20.9	18.5	31.8	23.7	1.05x	15.0	13.0	21.9	16.6
ensemble baselines										
ROVER-5a	6x	**17.3**	**11.8**	**27.9**	**19.0**	6x	11.7	8.4	20.0	13.4
ROVER-3a	4x	17.7	12.3	28.5	19.5	4x	11.8	8.2	20.1	13.4
ROVER-3b	1.54x	18.2	13.5	29.4	20.4	1.18x	11.7	8.3	20.5	13.5
our PEFT + ensemble methods										
ROVER-3M	1.54x	18.3	13.5	29.8	20.5	1.16x	11.3	8.0	20.3	13.2
ROVER-5M	1.90x	17.9	13.3	28.9	20.0	1.26x	**11.3**	**7.9**	**19.4**	**12.9**

TABLE II: *Ablation study of fine-tuning the text embedding layer at the decoder.*

Method	**eo WER (%)**
FT (*whisper-small*)	17.63
w/o fine-tune embedding layer	18.79
FT (*whisper-large-v2*)	12.15
w/o fine-tune embedding layer	13.43

TABLE III: *Ablation study of combining different models adapted by FT, PB, PA, SA, LoRA, or LGate to form a model ensemble. The last row shows the result of combining two models adapted by FT and one model adapted by PB to form an ensemble.*

Models for Ensemble						**eo,ia,fy**
FT	PB	PA	SA	LoRA	LGate	**AWER (%)**
		✓	✓	✓		14.3
	✓	✓	✓			13.9
	✓			✓	✓	13.8
	2	✓				13.5
	✓	✓			✓	13.5
	✓		✓	✓		13.4
	✓	✓		✓		13.2
2	✓					13.2

IV. Conclusion

To conclude, we propose a method to improve low-resource language adaptation performance by combining models adapted by different PEFT methods to form a model ensemble, and ablation study has shown that the performance of a model ensemble can be consistently improved if more diverse PEFT methods are used.

Acknowledgment

This research is supported by the National Research Foundation Singapore under its AI Singapore Programme grant number AISG2-GC-2022-005. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This research is also supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore. The computational work for this article was fully performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

[1] N. Ding, Y. Qin, G. Yang, *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, pp. 220–235, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257316425.

[2] A. Mallya, D. Davis, and S. Lazebnik, "Piggyback: Adapting a single network to multiple tasks by learning to mask weights," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 67–82.

[3] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[4] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[5] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," *arXiv preprint arXiv:2110.04366*, 2021.

[6] N. M. Selvaraj, X. Guo, A. Kong, B. Shen, and A. Kot, "Adapter incremental continual learning of efficient audio spectrogram transformers," *arXiv preprint arXiv:2302.14314*, 2023.

[7] L.-J. Yang, C.-H. H. Yang, and J.-T. Chien, "Parameter-efficient learning for text-to-speech accent adaptation," *arXiv preprint arXiv:2305.11320*, 2023.

[8] S. Radhakrishnan, C.-H. H. Yang, S. A. Khan, N. A. Kiani, D. Gomez-Cabrero, and J. N. Tegner, "A parameter-efficient learning approach to arabic dialect identification with pre-trained general-purpose speech model," *arXiv preprint arXiv:2305.11244*, 2023.

[9] D. Ng, C. Zhang, R. Zhang, *et al.*, "Adapter-tuning with effective token-dependent representation shift for automatic speech recognition,"

[10] Z. Yu, Y. Zhang, K. Qian, *et al.*, "Master-asr: Achieving multilingual scalability and low-resource adaptation in asr with modular learning," in *International Conference on Machine Learning*, PMLR, 2023, pp. 40 475–40 487.

[11] J. Qi *et al.*, "Parameter-efficient dysarthric speech recognition using adapter fusion and householder transformation," *arXiv preprint arXiv:2306.07090*, 2023.

[12] A. Arunkumar, V. N. Sukhadia, and S. Umesh, "Investigation of ensemble features of self-supervised pretrained models for automatic speech recognition," *arXiv preprint arXiv:2206.05518*, 2022.

[13] A. K. Parikh, L. ten Bosch, and H. van den Heuvel, "Ensembles of hybrid and end-to-end speech recognition.," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 6199–6205. [Online]. Available: https://aclanthology.org/2024.lrec-main.547.

[14] I. Gitman, V. Lavrukhin, A. Laptev, and B. Ginsburg, "Confidence-based ensembles of end-to-end speech recognition models," *arXiv preprint arXiv:2306.15824*, 2023.

[15] A. K. Parikh, L. ten Bosch, and H. van den Heuvel, "Ensembles of hybrid and end-to-end speech recognition.," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 6199–6205.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, PMLR, 2023, pp. 28 492–28 518.

[17] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] A. Rouditchenko, S. Khurana, S. Thomas, *et al.*, "Comparison of multilingual self-supervised and weakly-supervised speech pre-training for adaptation to unseen languages," *arXiv preprint arXiv:2305.12606*, 2023.

[19] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[20] L. Della Libera, P. Mousavi, S. Zaiem, C. Subakan, and M. Ravanelli, "Cl-masr: A continual learning benchmark for multilingual asr," *arXiv preprint arXiv:2310.16931*, 2023.

[21] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[22] K. Praveen, B. Radhakrishnan, K. Sabu, A. Pandey, and M. Shaik, "Language identification networks for multilingual everyday recordings," 2023.

[23] A. Babu, C. Wang, A. Tjandra, *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[24] C. Tang, W. Yu, G. Sun, *et al.*, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[25] Z. Liu, J. Spence, and E. Prud'Hommeaux, "Studying the impact of language model size for low-resource asr," in *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 2023, pp. 77–83.

[26] J. Pfeiffer, S. Ruder, I. Vulić, and E. M. Ponti, "Modular deep learning," *arXiv preprint arXiv:2302.11529*, 2023.

[27] N. Houlsby, A. Giurgiu, S. Jastrzebski, *et al.*, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*, PMLR, 2019, pp. 2790–2799.

[28] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[29] N.-Q. Pham, T.-N. Nguyen, S. Stüker, and A. Waibel, "Efficient weight factorization for multilingual speech recognition," *arXiv preprint arXiv:2105.03010*, 2021.

[30] S. Kim and M. L. Seltzer, "Towards language-universal end-to-end speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 4914–4918.

[31] M. Müller, S. Stüker, and A. Waibel, "Neural language codes for multilingual acoustic models," *arXiv preprint arXiv:1807.01956*, 2018.

[32] M. Müller, S. Stüker, and A. Waibel, "Neural codes to factor language in multilingual speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 8638–8642.

[33] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, IEEE, 1997, pp. 347–354.

[34] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[35] R. Ardila, M. Branson, K. Davis, *et al.*, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

Annotation-free Fine-tuning for Unsupervised Anomalous Sound Detection

Kai Guo*, Xiang Xie† and Fengrun Zhang*
* Beijing Institute of Technology, Beijing, China
E-mail: {guokai, zhangfengrun}@bit.edu.cn
† Beijing Institute of Technology, Zhuhai, Guangdong, China
E-mail: xiexiang@bit.edu.cn

***Abstract*—The goal of unsupervised anomalous sound detection (ASD) for industrial machines is to identify anomalous sounds using only normal sounds for training. A common method is to use attribute information as auxiliary labels to train ASD models. However, this approach often faces the challenge of obtaining auxiliary attribute information in complex industrial environments. This paper proposes a novel unsupervised anomalous sound detection method that fine-tunes pre-trained models without using attribute information, achieved by training a machine type classifier. The proposed method is called Annotation-Free Fine-tuning (AFF) for unsupervised anomalous sound detection. In addition, we propose an anomaly score calculation method that combines the machine type classifier with an unsupervised anomalous sound estimator, further improving the anomalous detection performance of AFF. Experiments on DCASE 2024 Task 2 development dataset indicate that our method outperforms other typical ASD methods that do not utilize attribute information.**

I. Introduction

The approach of using sound to assess the proper functioning of industrial machines has gained popularity and is commonly known as Anomalous Sound Detection (ASD). The goal of Detection and Classification of Acoustic Scenes and Events (DCASE) task 2 is to train a model using only normal sounds to achieve ASD [1]. By 2023, the challenge focused on the first-shot problem, aiming to achieve ASD without tuning model hyperparameters [2]. In the past few years, methods based on autoencoders have been used as the baseline system for DCASE Task 2 [3]. Autoencoders reconstruct sound features to learn the distribution of normal machine sounds. Popular ASD methods also include those based on GAN [4] or GMM [5], for example. A shared characteristic of these methods is that they do not require any auxiliary information to assist in model training.

On the other hand, one approach for ASD that has remained popular in recent years is the use of self-supervised learning methods based on metadata of audio files [6]. In past competitions, the metadata available for model training included machine ID, machine section, and attribute information. Attributes are parameters that define states of machines or types of noise [2]. In the 2023 competition, researchers typically used these attributes as labels to train a sound feature extractor and classifier, enabling them to calculate anomaly scores by clustering latent features or indirectly using the classifier's output. However, obtaining attribute information for real-world machine sounds can be challenging. The DCASE 2024 Challenge Task 2 [7] has been designed to address the issue of the difficulty in obtaining attribute information in real-world scenarios. In both the development and evaluation datasets, attribute information for some machine types has been completely removed, making typical self-supervised ASD methods that rely on attribute information infeasible.

Recently, [8] proposed an ASD method based on large-scale pre-trained models. They utilize attribute information to fine-tune pre-trained models like HuBERT [9] and adopt transformer pooling to gather embeddings from the same recording into a single embedding, in order to discover effective information. They use k-NN [10] to compute the anomaly score. However, the previous method is not suitable for situations where attribute information is missing. Inspired by the aforementioned work, we propose an ASD method that fine-tunes pre-trained models without the need for attribute information, aiming to extend its applicability to real-world anomalous sound detection scenarios. Additionally, inspired by the model ensemble algorithm [11], we proposed a backend anomaly score calculation method that combines embedding outlier detectors with a machine type classifier, enhancing the anomalous sound detection performance of pre-trained models. The main contributions of this paper are as follows:

- We propose a method using machine type as a condition to fine-tune pre-trained models. Since our method does not use any auxiliary attribute information from the dataset, we call it Annotation-Free Fine-tuning (AFF) for anomalous sound detection.
- We use the AFF method to fine-tune the pre-trained model CED [12], and the results of the experiment on the DCASE 2024 task 2 development dataset validate the feasibility of fine-tuning large pre-trained models for anomalous sound detection.
- We propose an anomaly score calculation method that combines k-NN, PCA [13], and the output of the machine type classifier, enhancing the performance of AFF. Ablation experiments and visualization analysis further demonstrate the effectiveness of the proposed method.

†Corresponding author

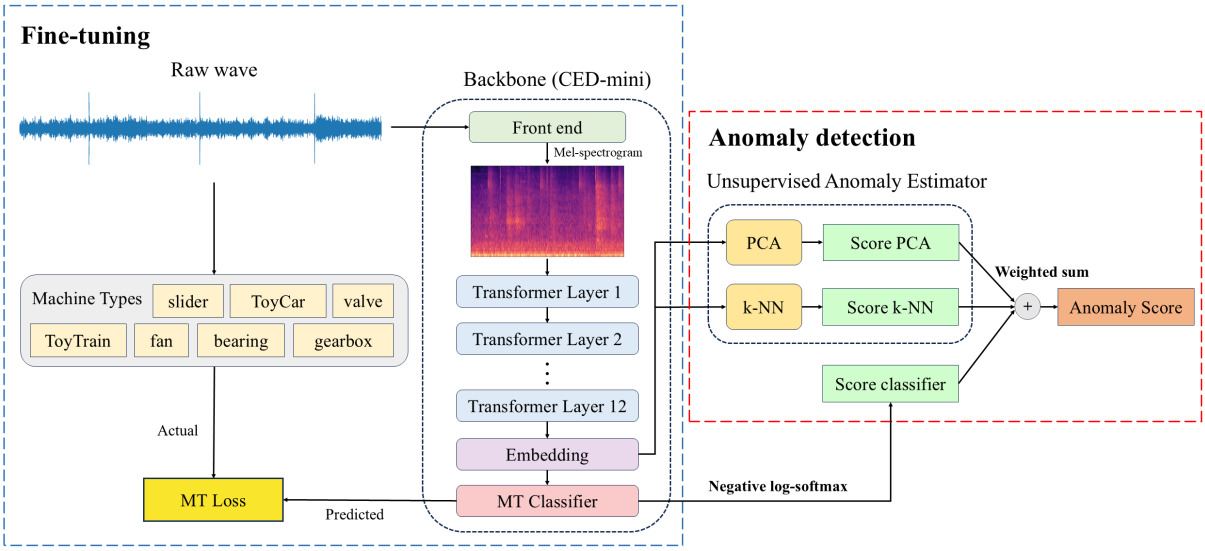


Fig. 1. Overview of the proposed method. The raw audio is fed into the front end of the CED to extract the Mel-spectrogram, which is then passed through several layers of transformer to obtain the embedding. A classifier is used to predict the machine type of the input audio. For anomalous sound detection, the negative log-softmax of the classifier's output is used in conjunction with the output of the unsupervised anomaly estimator.

II. Proposed Method

This section introduces our Annotation-Free Fine-tuning framework designed for anomalous sound detection, with an overview of our proposed method shown in Fig. 1. The system mainly consists of two parts: one part is a backbone network based on a pre-trained model (i.e., CED), and the other part is a backend for anomaly sound detection, which includes an unsupervised anomaly estimator and the output of a Machine Type (MT) classifier.

A. Annotation-Free Fine-tuning

The anomalous sound detection method based on fine-tuning pre-trained models has been proven to be effective [8]. In DCASE 2024 task2, it is no longer possible to obtain attribute information for all samples in order to simulate real-world scenarios. Besides, previous work based on attribute information needs to train the model from scratch for each machine type individually, resulting in an increase in the training cost. Additionally, some works set different hyperparameters for different machine types to achieve better performance, which is difficult to achieve in practice [6]. To address the above issues, we propose an Annotation-Free Fine-tuning method for anomalous sounds detection. Our approach differs from the previous approach in that we treat all samples of the same machine type in the dataset as a class, regardless of whether the sample has an attribute annotation that can be used for classification. We chose the Consistent Ensemble Distillation model (CED) as the pre-trained model for fine-tuning because CED achieves audio tagging performance similar to large ensembles while having significantly fewer parameters [12].

During the model training phase, raw audio is fed into the front end of the CED, and Mel-spectrogram features are extracted. After passing through several transformer layers, the model extracts the embeddings of the input, which are used to train a MT classifier as well as an unsupervised anomaly estimator. The classifier consists of a layer normalization layer followed by a linear layer. We use the cross-entropy loss for model training. Assuming the training set contains a total of N machine types, then the loss is

$$L_{MT} = \sum_{c=1}^{N} y_{o,c} \log(p_{o,c}) \tag{1}$$

where $y_{o,c}$ is a binary indicator, and $p_{o,c}$ is the predicted probability that observation o is of machine type c. Based on our design, the model learns to correctly classify the machine type of normal machine sound samples. During the inference phase, when an incorrect sample is fed into the model, the confidence of the MT classifier in making a correct classification decreases. If the sample is determined to belong to an incorrect machine type, it is more likely to be considered an anomalous one.

B. Anomaly Score Calculator Based on Model Ensemble

Previous self-supervised ASD methods typically use a single approach for anomaly score computation, such as using the classifier's output [14] or estimating the distribution of embeddings [8]. According to our research, combining the above methods to calculate anomaly scores is feasible. Specifically, our anomaly score calculator consists of two main components: one part is the unsupervised anomaly estimator, and the other part is the negative log-softmax of the output from the MT classifier. The unsupervised anomaly estimator is composed of a PCA model and a k-NN model. We train them for each type of machine using embeddings of normal sounds. The final anomaly score is defined as

$$\mathcal{A} = \alpha S_{PCA} + \beta S_{kNN} + \gamma S_{Classifier} \tag{2}$$

where S_{PCA}, S_{kNN}, and $S_{Classifier}$ represent the anomaly scores calculated by PCA, KNN, and machine type classifier models, respectively, while α, β, and γ are the corresponding

weight coefficients. We set the same weight coefficients for all machine types, specifically setting α, β, and γ to 1e-3, 1e-3, and 1, respectively, merely to ensure that the anomaly scores of the three parts are on the same order of magnitude without conducting a more detailed hyperparameter search.

III. Experiments

A. Dataset

We conduct experiments on the development dataset in DCASE 2024 Challenge task2 [7]. The dataset includes seven types of machines: bearing, fan, gearbox, slider, ToyCar, ToyTrain, and valve. It is noteworthy that the training data for four machine types—ToyCar, fan, bearing, and valve—include attribute information such as machine model or working conditions. In contrast, the training sets for ToyTrain, gearbox, and slider do not contain any attribute information. However, no attribute information was used to assist model training in the experiments. Each machine type in the dataset contains: 990 clips of normal sounds in the source domain for training, 10 clips of normal sounds in the target domain for training, and 100 clips each of normal and anomalous sounds including data from both domains for testing.

B. Evaluation Metrics

The Area Under the Receiver Operating Characteristic Curve (AUC) is the most widely used metric in ASD, given that anomaly detection is essentially a binary classification task [1]. In line with the DCASE 2024 Challenge task 2 [7], we use source-AUC (sAUC), target-AUC (tAUC), and partial-AUC (pAUC) as evaluation metrics, with the parameter p set to 0.1 for pAUC. The final system score is determined by calculating the harmonic mean (Hmean) across all machine domains and types. Compared to the arithmetic mean, the harmonic mean is more sensitive to low values, making it a better measure for assessing the overall performance of ASD systems.

C. Training Configuration

The model takes Mel-spectrograms as the input. We follow the default configuration of CED [12], setting the window length to 512, the hop length to 160, and extracting the 64-dimensional Mel-filter bank features. The audio is sampled at 16 kHz, and the chunk length of the model is 10 seconds. We choose CED-Mini as the backbone model because it is a good compromise between performance and model size. The model was pre-trained on AudioSet [15]. We use the Adam optimizer [16] to fine-tune the CED model with a learning rate of 1e-4. Training runs for 200 epochs, and we save the model every 20 epochs for each machine type. For the PCA model, we take only one component for anomaly detection. For the k-Nearest Neighbors model, the number of neighbors is set to 2.

IV. Results

A. Performance Comparison

The results on the DCASE 2024 Task 2 development dataset are shown in Table I. To ensure a fair comparison of model performance, we have chosen to compare models that do not use any attribute information for training. AE (MAHALA) is an autoencoder model that operates in selective mahalanobis mode, and the mahalanobis distance is introduced to calculate the anomaly scores. It is the baseline model for DCASE 2024 Task 2 [3], and it was part of the first-place submission of DCASE 2022 Task 2. IDNN is an anomalous sound detection method based on Interpolation Deep Neural Network [17], which has been shown in previous studies to perform better than autoencoders based on reconstruction errors. TWFR-GMM [5] is the core method of the submission that achieved 3rd place in the DCASE 2022 Challenge Task 2. None of the methods use additional data augmentation techniques, except for TWFR-GMM, which uses SMOTE [18] to enhance domain generalization ability.

As can be seen from Table I, the harmonic mean of the AUC across all machine types for our proposed AFF-CED method exceeded that of other typical ASD methods that do not use attribute information. The AUC performance of the proposed method exceeds that of the baseline method by 2.62%. The proposed method performs exceptionally well in detecting anomalous sounds in bearing, slider, and valve. It is worth mentioning that TWFR-GMM also achieved good ASD performance across multiple machine types. However, its overall Hmean is not remarkable, as poor performance on individual machine types can significantly lower the overall system performance, which highlights the great robustness of the proposed method. Moreover, the proposed method achieved the best anomaly detection performance in the target domain across various machine types, including bearing, fan, ToyTrain, and valve. This indicates to some extent that our proposed method has better domain generalization ability compared to baseline method, even though we did not specifically design the system for domain generalization.

B. Ablation Studies

To further validate the proposed method, we conducted a series of ablation experiments on the DCASE 2024 Task 2 development dataset. The results of the ablation experiments are shown in Table II, where "w/o Fine-tune" indicates that we do not fine-tune the CED model and directly use the model pre-trained on AudioSet [15] for ASD. "w/o MT score" means that we fine-tune the model but do not use the classifier's output to calculate anomaly scores. "w/o PCA" and "w/o k-NN" mean that we do not use the PCA or k-NN models as unsupervised anomaly estimators. Based on the results, it is evident that without fine-tuning, the MT classifier will not function effectively and the model will essentially perform as a random anomaly detector. By fine-tuning, the model can fully learn the characteristics of normal sounds for each type of machine, thereby achieving ASD. When we do not utilize

TABLE I
PERFORMANCE (%) COMPARISON WITH ANNOTATION-FREE METHODS ON DCASE 2024 TASK 2 DEVELOPMENT DATASET.

Machine	AE (MAHALA) [3]			IDNN [17]			TWFR-GMM [5]			**AFF-CED**		
	sAUC↑	tAUC↑	pAUC↑	sAUC↑	tAUC↑	pAUC↑	sAUC↑	tAUC↑	pAUC↑	sAUC↑	tAUC↑	pAUC↑
bearing	56.92	51.58	58.68	55.00	62.42	54.89	53.24	58.60	**59.11**	**60.94**	**63.98**	54.63
fan	79.26	43.64	**55.42**	75.84	38.96	52.21	**79.44**	32.40	50.84	65.56	**47.32**	52.53
gearbox	81.82	73.66	55.57	61.30	63.66	53.16	**83.50**	**79.12**	57.79	71.88	68.32	**58.42**
slider	74.88	66.22	49.31	60.60	56.20	50.21	80.90	**75.21**	**58.32**	**87.02**	63.14	52.21
ToyCar	**64.42**	39.02	50.74	54.44	**54.10**	49.84	62.62	33.70	**51.58**	47.02	53.84	47.58
ToyTrain	59.98	40.84	48.00	61.18	48.20	48.37	**73.30**	45.14	49.16	59.06	**55.20**	**51.05**
valve	58.66	53.92	52.31	61.42	60.94	51.74	52.12	46.44	51.11	**64.06**	**70.38**	**57.37**
Hmean		55.61			54.95			54.96			**58.23**	

TABLE II
ABLATION STUDIES ON DCASE 2024 TASK 2 DEVELOPMENT DATASET.

Setting	Hmean (%)↑	mean (%)↑
AFF-CED	**58.23**	**59.60**
w/o Fine-tune	50.58	51.56
w/o MT score	56.50	57.61
w/o PCA	56.09	57.93
w/o k-NN	56.92	58.01

MT classifier for calculating anomaly scores, the model's Hmean and mean AUC performance drop by 1.73% and 1.99%, respectively. This indicates that MT classifier plays a significant role in anomaly detection. When we remove either the PCA model or the k-NN model for calculating anomaly scores, the system's Hmean AUC performance drops by 2.14% and 1.31%, respectively, and the mean AUC performance decreases by 1.67% and 1.59%, respectively. This indicates the effectiveness of both estimators for anomalous sound detection. Moreover, the performance drop is more significant when the PCA model is removed compared to the k-NN model. This indicates to some extent that the PCA model plays a more important role in anomaly detection in our system compared to the k-NN model.

C. Visualisation Analysis

As illustrated in Fig. 2, we present confusion matrices for each type of machine, providing a more intuitive demonstration of the stability and robustness of our proposed method in detecting anomalous sounds, particularly for bearing, slider, and ToyTrain. To further validate our method, we also visualized the latent features of the sound samples in the test dataset, which were extracted by the fine-tuned CED model. Fig. 3 shows the t-SNE [19] visualization of all samples in the test set of the DCASE 2024 Task 2 development dataset. From the visualization, it can be seen that our proposed method can essentially classify the seven types of machines in the dataset. We highlight some obvious outliers with red circles. It can be observed that most of these points are anomalies. In such cases, the machine type classifier tends to misclassify these points as incorrect machine types. Consequently, the $S_{Classifier}$ value in (2) will significantly increase, thereby achieving anomalous detection. Additionally, some normal samples may also be misclassified, which we believe is caused by domain shift. In our future work, we will apply data augmentation to the target domain samples for Annotation-Free Fine-tuning, aiming to further enhance the model's domain generalization performance.

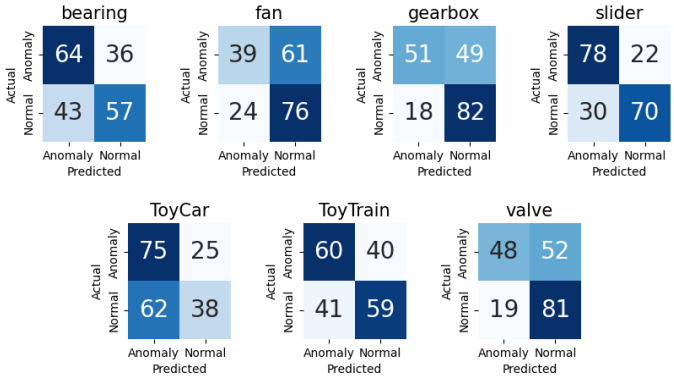


Fig. 2. Confusion matrices of each machine type. The threshold for classifying instances as anomalies is the optimal threshold, which maximizes the difference between the true positive rate and the false positive rate.

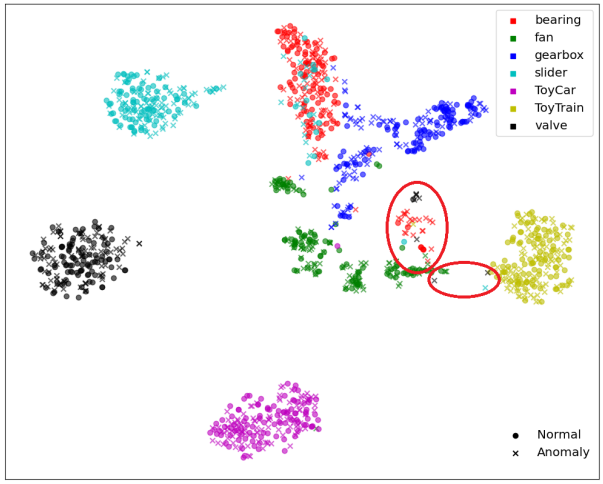


Fig. 3. The t-SNE visualization of all samples in the test dataset. Different colors in the visualization represent different machine types. The "·" symbol represents normal sounds, while the "x" symbol denotes anomalous sounds.

V. CONCLUSIONS

In this paper, we propose a pre-trained model fine-tuning method based on machine type classification that does not rely on attribute information for unsupervised anomalous sound detection. We also propose a method for calculating anomaly scores that combines classifier outputs with unsupervised estimations of embedding distributions. Results on the DCASE 2024 Task 2 development dataset indicate that our proposed method achieves better anomalous detection performance compared to other typical methods that do not utilize attribute information. In the future, we will aim to enhance the domain generalization performance of the proposed method.

References

[1] Y. Koizumi, Y. Kawaguchi, K. Imoto, *et al.*, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Nov. 2020, pp. 81–85. [Online]. Available: http://dcase.community/documents/workshop2020/proceedings/DCASE2020Workshop%5C_Koizumi%5C_3.pdf.

[2] K. Dohi, K. Imoto, N. Harada, *et al.*, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2305.07828*, 2023.

[3] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 191–195.

[4] A. Jiang, W.-Q. Zhang, Y. Deng, P. Fan, and J. Liu, "Unsupervised anomaly detection and localization of machine audio: A gan-based approach," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096813.

[5] J. Guan, Y. Liu, Q. Zhu, T. Zheng, J. Han, and W. Wang, "Time-weighted frequency domain audio representation with gmm estimator for anomalous sound detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096356.

[6] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, Nov. 2020, pp. 46–50.

[7] T. Nishida, N. Harada, D. Niizumi, *et al.*, "Description and discussion on DCASE 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2406.07250*, 2024.

[8] B. Han, Z. Lv, A. Jiang, *et al.*, "Exploring large scale pre-trained models for robust machine anomalous sound detection," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1326–1330. DOI: 10.1109/ICASSP48485.2024.10447183.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021. DOI: 10.1109/TASLP.2021.3122291.

[10] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *SIGMOD Rec.*, vol. 29, no. 2, pp. 427–438, May 2000, ISSN: 0163-5808. DOI: 10.1145/335191.335437. [Online]. Available: https://doi.org/10.1145/335191.335437.

[11] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006. DOI: 10.1109/MCAS.2006.1688199.

[12] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "Ced: Consistent ensemble distillation for audio tagging," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 291–295. DOI: 10.1109/ICASSP48485.2024.10446348.

[13] F. L. Gewers, G. R. Ferreira, H. F. D. Arruda, *et al.*, "Principal component analysis: A natural approach to data exploration," *ACM Comput. Surv.*, vol. 54, no. 4, May 2021, ISSN: 0360-0300. DOI: 10.1145/3447755. [Online]. Available: https://doi.org/10.1145/3447755.

[14] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 816–820. DOI: 10.1109/ICASSP43922.2022.9747868.

[15] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.

[16] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1412.6980.

[17] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 271–275. DOI: 10.1109/ICASSP40776.2020.9054344.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002, ISSN: 1076-9757.

[19] V. D. M. Laurens and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2605, pp. 2579–2605, 2008.

GSBIQA: Green Saliency-guided Blind Image Quality Assessment Method

Zhanxuan Mei*, Yun-Cheng Wang*, and C.-C. Jay Kuo*
* University of Southern California
E-mail: zhanxuan@usc.edu

***Abstract*—Blind Image Quality Assessment (BIQA) is an essential task that estimates the perceptual quality of images without reference. While many BIQA methods employ deep neural networks (DNNs) and incorporate saliency detectors to enhance performance, their large model sizes limit deployment on resource-constrained devices. To address this challenge, we introduce a novel and non-deep-learning BIQA method with a lightweight saliency detection module, called Green Saliency-guided Blind Image Quality Assessment (GSBIQA). It is characterized by its minimal model size, reduced computational demands, and robust performance. Experimental results show that the performance of GSBIQA is comparable with state-of-the-art DL-based methods with significantly lower resource requirements.**

I. Introduction

Objective image quality assessment (IQA) is pivotal in various multimedia applications. It can be categorized into three distinct types: Full-Reference IQA (FR-IQA), Reduced-Reference IQA (RR-IQA), and No-Reference IQA (NR-IQA). FR-IQA directly compares a distorted image against a reference or original image to assess quality. RR-IQA, on the other hand, uses partial information from the reference images to evaluate the quality of the target images. NR-IQA, also known as blind image quality assessment (BIQA), becomes essential in scenarios where reference images are unavailable, such as at the receiver's end or for user-generated content on social media. The demand for BIQA has surged with the increasing popularity of such platforms.

Research in BIQA has gained significant momentum over recent years, branching into two primary approaches: conventional methods and deep-learning-based (DL-based) methods. Conventional BIQA methods typically follow a structured pipeline involving quality-aware feature extraction followed by regression to map these features to quality scores. Over the past two decades, various conventional BIQA methods have been developed, including those based on Natural Scene Statistics (NSS) [1] and codebook-based approaches [2]. More recently, inspired by the success of DNNs in computer vision, researchers have developed DL-based methods [3], [4] to solve the BIQA problem. However, the high cost of collecting large-scale annotated IQA datasets and the tendency of DL-based methods to overfit on limited-sized datasets pose significant challenges. Some advanced DL-based methods have adopted large pre-trained models from external datasets, such as ImageNet [5] to address this. Based on that, a further direction [6] is enhancing dataset variability by sampling patches randomly from quality-annotated images and incorporating saliency-guided approaches to improve accuracy, reflecting the intrinsic link between perceptual quality prediction and saliency prediction.

However, due to their large model sizes and high computational complexity, DL-based BIQA methods face substantial challenges, particularly in their deployment on mobile or edge devices. Moreover, the conventional saliency predictors often used in these methods do not significantly enhance perceptual quality predictions. In response to these challenges, this work introduces a novel, efficient, and saliency-assisted BIQA method named Green Saliency-guided Blind Image Quality Assessment (GSBIQA). It is particularly suited for real-world applications with authentic distortions and offers a modular design with a straightforward training pipeline.

We should point out that green image saliency detection (GreenSaliency) and green blind image quality assessment (GreenBIQA) were initially presented in [7] and [8], respectively. This work offers a significant expansion of these initial efforts by combining and extending these two lightweight models into a comprehensive and advanced BIQA method.

II. Related Works

A. Blind Image Quality Assessment Methods

Over the past two decades, several conventional BIQA methods have been developed, including those based on Natural Scene Statistics (NSS) [1] and codebook approaches [2]. The remarkable success of DNNs has inspired a surge in research focused on employing DL-based methods to address the BIQA challenge. Early-stage DL-based BIQA methods, such as BIECON [3] and WaDIQaM [4], constructed neural networks consisting of convolutional layers, max pooling, and fully connected layers. To mitigate the issue of limited annotated datasets in IQA, most contemporary DL-based approaches leverage advanced pre-trained models for feature extraction. A probabilistic quality representation was introduced in PQR [9], utilizing a robust and optimal loss function to more accurately reflect the score distribution provided by different evaluators, thereby enhancing prediction accuracy and expediting the training phase. HyperIQA [10] employed a self-adaptive hypernetwork architecture that adjusts the quality prediction parameters. This architecture can manage a broad spectrum of distortions through a local distortion-aware module and

adapts to diverse content through perceptual quality patterns. DBCNN [11] utilizes networks pre-trained on synthetic-distortion datasets and ImageNet for distortion and general classification, respectively. Features from both models are combined using bilinear pooling for final quality prediction. Addressing the large model size issue common in DL-based methods, GreenBIQA [8] demonstrates a successful approach using a non-DL-based method to achieve high accuracy with a smaller model size and reduced computational complexity.

B. Saliency-guided BIQA methods

Image saliency detectors aim to elucidate the mechanisms underlying human observation, a goal that aligns closely with the objectives of IQA, particularly in the context of BIQA. Consequently, some researchers have incorporated saliency detection techniques within BIQA frameworks to improve their efficacy. Within DL-based frameworks, saliency models enhance data pre-processing, such as in [6], where these models guide data cropping by identifying and selecting patches containing salient objects from the original images. Furthermore, saliency detectors are integrated into the data cropping and decision fusion processes, as demonstrated in [12]. Rather than merely applying saliency models as either a pre-processing or post-processing tool, [13] introduces SGDNet. This network is built on an end-to-end multi-task learning framework that addresses image saliency prediction and quality assessment.

III. Proposed Method

An overview of the proposed GSBIQA method is illustrated in Figure 1. The GSBIQA approach features a modularized architecture comprising five distinct modules: (1) green saliency detection, (2) saliency-guided data cropping, (3) green BIQA feature extraction, (4) local patch prediction, and (5) saliency-guided global prediction. They are elaborated below.

A. Green Saliency Detection

Drawing inspiration from GreenSaliency [7], a simplified version is implemented within the proposed green saliency detection model. Contrary to the original method, which extracts features from five distinct layers or scales, this model selects only the three lowest layers: d4, d8, and d16. Overlapped patches are collected from each of these layers, and two consecutive Saab transforms of sizes 2×2 and 4×4 are employed to extract features, resulting in d4, d8, and d16 features derived from the corresponding layers. It is important to note that the Saab transforms [14] are applied independently across different layers, unlike the interconnected Saab transform modules found in GreenSaliency.

In the saliency map prediction phase, features from all three layers are concatenated. The most significant subset of these features is identified through a Relevant Feature Test (RFT) [15], and subsequently, an XGBoost regressor is utilized to estimate the saliency values for each patch. An up-sampling operation is then executed to generate the predicted saliency map. The predicted saliency map facilitates data cropping and decision ensemble modules, as detailed in Sections III-B and III-E, respectively. Given the absence of ground truth saliency maps in most IQA datasets, the green image saliency prediction model was pre-trained on the SALICON [16] dataset, a widely used resource for image saliency prediction studies.

B. Saliency-guided Data Cropping

In the saliency-guided data cropping process, input images are uniformly cropped into overlapped sub-images of size 256×256, each retaining the original image's quality score. Then, the predicted saliency map is used to calculate an Average Saliency Score (ASS) for each sub-image, defined as:

$$ASS_i = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} S(n, m), \tag{1}$$

where $S(n, m)$ represents the saliency value at pixel coordinates (n, m) in the ith sub-image, and $N \times M$ represents the size of sub-images. The ASS indicates how well sub-images attract human attention, with higher scores indicating more significant impacts on perceptual judgments. Sub-images with the highest ASS are selected for further analysis, ensuring focus on the most salient parts of the image to optimize subsequent quality assessments.

C. Green BIQA Feature Extraction

In this part, three sets of features are extracted and detailed.

1) Spatial Features: The pipeline of spatial feature extraction is depicted in Figure 2. Initially, input sub-images are segmented into non-overlapping 8×8 blocks, and Discrete Cosine Transform (DCT) coefficients are computed via a block DCT transform. These coefficients are then ordered in a zigzag pattern, producing one DC coefficient and sixty-three AC coefficients labeled AC1 to AC63, then divided into 64 channels. For further processing and to decorrelate DC coefficients while deriving higher-level representations, a dual-stage Saab transform approach is utilized, consisting of two successive transformations, referred to as Hop1 and Hop2:

- Hop1 Processing: The 32×32 DC coefficients are subdivided into non-overlapping 4×4 blocks. These blocks undergo a Saab transform, resulting in one DC channel and 15 AC channels for each block. Each block's 8×8 DC coefficients are then advanced to Hop2.
- Hop2 Processing: A second Saab transform is applied to non-overlapping 4×4 blocks from DC coefficients from Hop1, yielding one DC and 15 AC channels.

After maximum pooling of Saab coefficients from Hop1 and DCT coefficients from the top layer, we take further steps to reduce the number of representations. This process includes:

- Computing the standard deviation as shown by the green downward arrow in the diagram.
- Conduct the PCA transform on spatially adjacent regions of these coefficients as indicated by the orange downward arrow in the diagram.

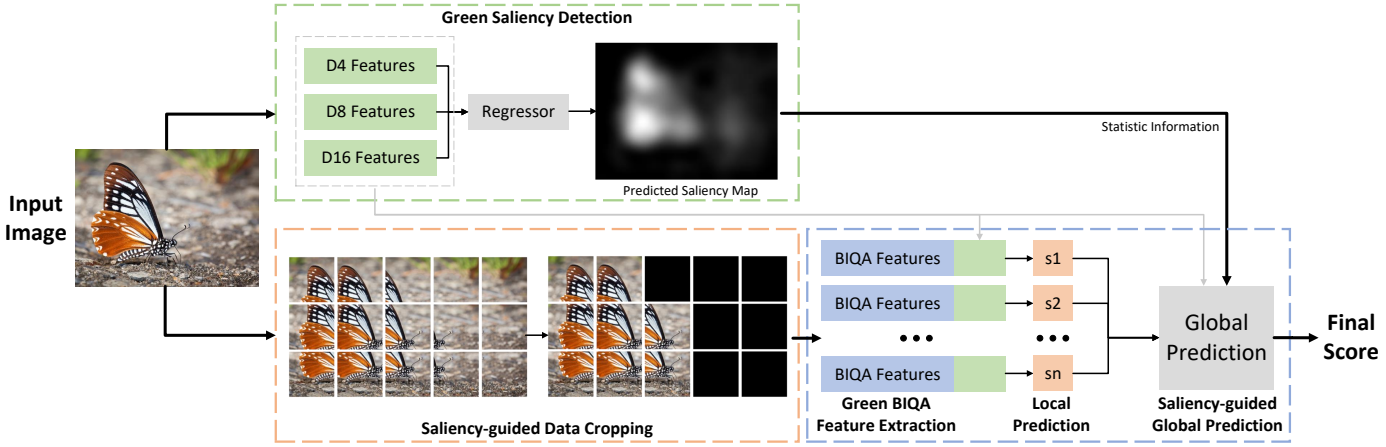


Fig. 1: An overview of the proposed GSBIQA method.

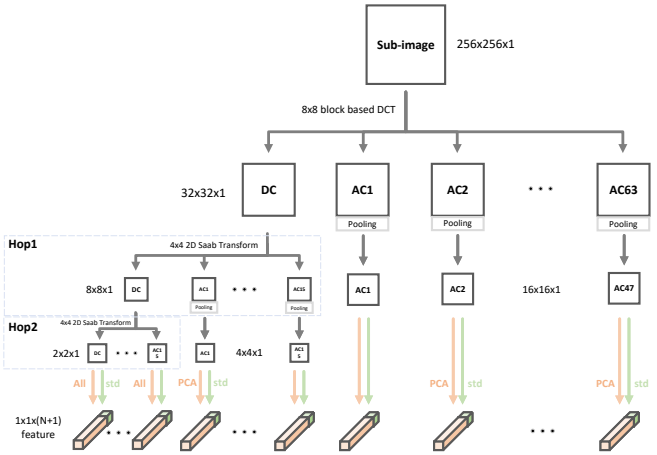


Fig. 2: Structure of spatial features extraction.

2) Spatio-color Features: The extraction process for spatial-color features begins by converting sub-images from the YUV color to the RGB color space. These converted sub-images form spatial-color cuboids with dimensions $H \times W \times C$, where H and W represent the height and width of the sub-image, respectively, and $C = 3$ denotes the number of color channels. These cuboids are then input into a two-hop hierarchical structure, similar to that in the spatial features extraction part.

- In the first hop, the input cuboids are segmented into non-overlapping smaller cuboids of size $4 \times 4 \times 3$. A 3D Saab transform is applied to each cuboid independently, resulting in one DC channel and 47 AC channels labeled AC1 through AC47. Each channel maintains a spatial resolution of 64×64.
- Given the spatial correlation of the DC coefficients, a 2D Saab transform is utilized in Hop2. Here, the DC channel, sized 64×64, is further subdivided into 16×16 non-overlapping blocks, each 4×4 in dimension.

Similar to those in the spatial features extraction part, all channels of coefficients in each layer are first downsampled by max pooling. Then, PCA coefficients and standard deviation values are computed.

3) Saliency Features: Building on the insights that multi-level features can boost the accuracy of IQA methods, our approach integrates saliency features from multiple layers into the green BIQA feature extraction process. These features, initially extracted in the green image saliency detection module and shown as d4 and d8 features in Figure 1, are resized for compatibility with BIQA dimensions and seamlessly concatenated with BIQA features. This integration enhances the feature set by leveraging the spatial coherence and contextual relevance of the saliency data, offering a more comprehensive understanding of image quality and potentially enhancing prediction precision without additional computational overhead.

D. Local Patch Score Prediction

The feature set derived from the green BIQA feature extraction module is characterized by its large dimensionality. To address this, we adopt the Relevant Feature Test (RFT), as proposed in [15], to identify and select the most discriminant features. RFT is a supervised feature selection technique that evaluates the significance of each feature based on its ability to improve the predictive model's performance. It uses statistical methods to ascertain the relevance of features, ensuring that only the most powerful features are retained for model training.

Due to the lack of Mean Opinion Scores (MOS) for individual sub-images and only a global MOS for the entire image, directly assigning the global MOS to all local sub-images is ineffective. To overcome this, we employ an indirect method that allows each sub-image to predict more adaptive local scores while maintaining the average global image score close to the ground truth. The local quality score of ith sub-image in the jth image and the global quality score of jth image are denoted as q_{ij} and Q_j, respectively. We initiate the training process by uniformly distributing the global quality score across all local sub-images. An XGBoost regressor is trained to predict q_{ij} for each sub-image. Q_j for the entire image is computed by the average of all q_{ij} from the same image. The discrepancy between Q_j and its corresponding label is

used to update the gradient, which is then propagated back to each local patch. This iterative process, characterized by successive relaxation and adjustment, converges to a stable set of local quality scores q_{ij}. This approach effectively captures the diverse quality variations across sub-images, providing a more nuanced image quality assessment.

E. Saliency-guided Global Quality Score Prediction

Upon calculating local quality scores for each sub-image, we integrate these scores to form a global quality assessment using an additional XGBoost regressor, which utilizes a detailed set of input features. The feature collection and preparation involve:

- Collection of all local quality scores from the same image to capture quality variance.
- Computation of statistical metrics such as average, standard deviation, and maximum value of saliency local scores from the same image.
- Integration of downsampled d16 features from the green saliency detection module refined by the RFT.

The prepared features enable the XGBoost regressor to provide an accurate global image quality prediction, offering a thorough evaluation that reflects both local detail and global perceptual impact.

IV. Experiments

A. Experimental Setup

1) Datasets: The green image saliency detection module was pre-trained using the SALICON [16] dataset. For the evaluation of the GSBIQA method, we utilized the LIVE-C [17] and KonIQ-10K [18] datasets. These datasets are particularly relevant for authentic IQA as they encompass various real-world images captured by users featuring assorted distortions.

2) Evaluation Metrics: The effectiveness of the proposed method is quantified using two widely recognized metrics: the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank Order Correlation Coefficient (SROCC). PLCC is computed as follows:

$$PLCC = \frac{\sum_i (p_i - p_m)(\hat{p_i} - \hat{p_m})}{\sqrt{\sum_i (p_i - p_m)^2}\sqrt{\sum_i (\hat{p_i} - \hat{p_m})^2}}, \quad (2)$$

where p_i represents the predicted scores, $\hat{p_i}$ denotes the subjective scores, and p_m and $\hat{p_m}$ are the mean values of the predicted and subjective scores, respectively. SROCC is calculated as:

$$SROCC = 1 - \frac{6\sum_{i=1}^{L}(m_i - n_i)^2}{L(L^2-1)}, \quad (3)$$

where m_i and n_i are the ranks of the predicted and actual subjective scores, respectively, and L represents the total number of samples, which corresponds to the number of images evaluated.

3) Implementation Details: In the training stage, we crop and select 35 sub-images of size 256×256 for each image in the two authentic datasets. The RFT selects 3,000 dimensions of features to do local patch score prediction. To perform global score prediction, the input features include 35 predicted scores from local patch score prediction, 165 dimensions of features from the statistics of the saliency map, and 600 dimensions of features from saliency features selected by RFT. We adopt the standard evaluation procedure by splitting each dataset into 80% for training and 20% for testing. Furthermore, 10% of training data is used for validation. We ran experiments ten times and reported median PLCC and SROCC values.

B. Experimental Results

1) Benchmarking Methods: BRISQUE [1] and CORNIA [2] are conventional BIQA methods. BIECON [3] and WaDIQaM [4] are DL-based BIQA methods without pre-trained models on external datasets. PQR [9], DBCNN [11], SGDNet [13], HyperIQA [10], and TReS [19] are DL-based BIQA methods with pre-trained models on external datasets. GreenBIQA [8] is our previous method.

2) Comparison among Benchmarking Methods: The evaluation of two datasets featuring authentic distortions, LIVE-C and KonIQ-10K, reveals that our GSBIQA method surpasses conventional BIQA, simple DL-based, and our previous methods, as detailed in Table I. This superior performance underscores the efficacy of the quality-aware features extracted by GSBIQA, highlighting its potential for robust and accurate image quality assessment for real-world image distortions. While GSBIQA exhibits promising performance, a performance gap remains compared to state-of-the-art DL-based methods. However, it is essential to consider the significantly lower model complexity of GSBIQA as a beneficial trade-off. The details of model complexity are discussed in Section IV-C.

3) Qualitative Analysis on Exemplary Images: We conducted a qualitative analysis to evaluate the performance of the GSBIQA method by comparing the predicted MOS against the ground truth. Figure 3 presents exemplary images demonstrating successful outcomes. The ground truth MOS and GSBIQA-predicted MOS values are denoted as MOS(G) and MOS(P), respectively. These images are characterized by salient objects, effectively detected by the green image saliency detector. The saliency guidance provided by this component enables GSBIQA to accurately capture these features, significantly enhancing the model's performance. These images confirm that GSBIQA is proficient in scenarios where salient objects dominate the visual field. Conversely, GSBIQA exhibits limitations in images that lack distinct salient objects. Examples of such images include those predominantly featuring full backgrounds or textures, as shown in Figure 4. In these cases, the absence of transparent, distinguishable objects leads to challenges in saliency detection, resulting in poorer performance.

C. Model Complexity

We examine the complexity of BIQA methods from three perspectives: model size, computational complexity as mea-

TABLE I: Performance comparison in PLCC and SROCC metrics between our GSBIQA method and ten benchmarking methods on two IQA databases. The best performance numbers are shown in boldface and "X" denotes the multiple number.

BIQA Method	LIVE-C		KonIQ-10K		Model Size (MB)↓	GFLOPs↓	Inference Time (ms)↓
	SROCC↑	PLCC↑	SROCC↑	PLCC↑			
BRISQUE [1]	0.608	0.629	0.665	0.681	-	-	-
CORNIA [2]	0.632	0.661	0.780	0.795	7.4 (3.1X)	-	-
BIECON [3]	0.595	0.613	0.618	0.651	35.2 (14.6X)	2.82 (2.2X)	26 (1.4X)
WaDIQaM [4]	0.671	0.680	0.797	0.805	25.2 (10.5X)	4.38 (3.5X)	26 (1.4X)
PQR [9]	0.857	**0.882**	0.880	0.884	235.9 (98.3X)	157 (125.6X)	100 (5.3X)
DBCNN [11]	0.851	0.869	0.875	0.884	54.6 (22.7X)	242 (193.6X)	55 (2.9X)
SGDNet [13]	0.851	0.872	0.897	0.917	323.6 (134.8X)	260 (208X)	141 (7.4X)
HyperIQA [10]	**0.859**	**0.882**	0.906	0.917	104.7 (43.6X)	145 (116X)	54 (2.8X)
TReS [19]	0.846	0.877	**0.915**	**0.928**	582 (242.5X)	290 (232X)	260 (13.7X)
GreenBIQA [8]	0.801	0.809	0.858	0.870	**1.82 (0.76X)**	**0.85 (0.68X)**	23 (1.21X)
GSBIQA(ours)	0.830	0.839	0.875	0.883	2.4 (1X)	1.25 (1X)	**19 (1X)**

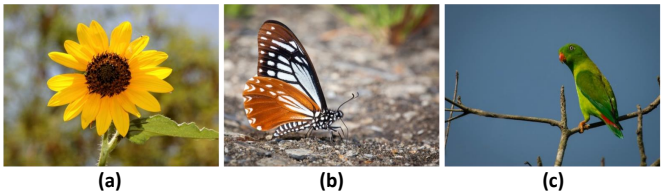


Fig. 3: Successful cases in GSBIQA.
(a) MOS(G)=3.80, MOS(P)=3.80, (b) MOS(G)=3.76, MOS(P)=3.76, (c) MOS(G)=3.70, MOS(P)=3.70.

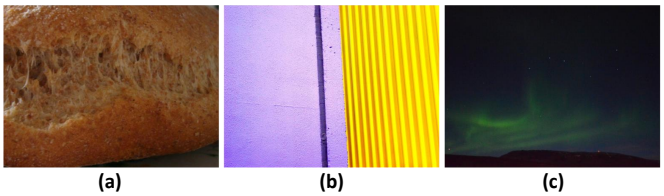


Fig. 4: Failed cases in GSBIQA.
(a) MOS(G)=2.61, MOS(P)=2.41, (b) MOS(G)=3.23, MOS(P)=3.00, (c) MOS(G)=2.30, MOS(P)=2.07.

sured by floating-point operations (FLOPs), and inference time, as shown in Table I.

1) Model Size: Compared to conventional methods such as CORINA, GSBIQA demonstrates superior performance and maintains a smaller model footprint. Additionally, GSBIQA outperforms two early-stage DL-based methods, BIECON and WaDIQaM, while achieving a significantly reduced model size. In comparison with advanced DL-based methods, namely PQR, DBCNN, SGDNet, HyperIQA, and TReS, GSBIQA offers competitive performance on the LIVE-C and KonIQ-10K datasets, yet with a considerably smaller model size. Notably, advanced DL-based methods typically rely on extensive pre-trained networks, often exceeding 100MB, which underscores the efficiency of GSBIQA in maintaining lower resource usage.

2) Computational Complexity: The "GFLOPs" column in Table I details the GFLOPs required to process a single image. Among early-stage DL-based methods such as WaDIQaM and BIECON, although the networks are smaller and require fewer FLOPs, their performance still falls short of GSBIQA. Conversely, more advanced DL-based methods like TReSN and HyperIQA, while surpassing GSBIQA in accuracy, also possess considerably larger model sizes and higher computational complexities. Specifically, the FLOPs for TReS and HyperIQA are 232 and 116 times greater than those for GSBIQA, respectively. It is crucial to highlight that non-DL-based methods like GSBIQA might benefit less from GPU acceleration due to a lack of optimized hardware-software integration compared to DL-based methods. However, the notably low computational complexity of GSBIQA underscores its potential for effective operation in GPU-supported environments, pending further improvements in third-party libraries and coding optimizations.

3) Inference Time: Inference time is another critical factor, particularly for applications on mobile or edge devices. Table I presents the inference times, measured in milliseconds per image, across various methods on two datasets. Each method was evaluated in the same environment using a single CPU. GSBIQA offers significant advantages over benchmarked methods by jointly considering performance and inference time. It is noteworthy that GSBIQA can process approximately 52 images per second using a single CPU, thus meeting the real-time requirements for processing videos at 30 frames per second. Furthermore, one can reduce the inference time for GSBIQA through code optimization and the integration of more advanced library support.

TABLE II: Ablation study for GSBIQA on KonIQ-10K dataset.

Saliency Guidance		Local prediction	SROCC	PLCC
Data Cropping	Global Prediction			
			0.832	0.845
✓			0.848	0.859
	✓		0.850	0.860
		✓	0.862	0.866
✓	✓		0.862	0.868
✓	✓	✓	0.875	0.883

D. Ablation Study

We conduct an ablation study in Table II by focusing on three critical components of GSBIQA: data cropping, global quality score prediction, and local patch prediction. First, the study assessed the impact of each component on its own, as shown in the second to fourth rows. Both SROCC and PLCC metrics on the KonIQ-10K dataset improved when each component was implemented independently, underscoring the significant role each plays in enhancing model performance. Next, a combined implementation of data cropping and global quality score prediction, guided by saliency detection, was analyzed. The results in the fifth row demonstrated substantial improvements, validating the effectiveness of integrating saliency guidance into these processes. Finally, we use all components in the sixth row and see that SROCC and PLCC can be further improved to reach the highest value.

V. Conclusion and Future Work

A novel and lightweight saliency-guided BIQA method, called GSBIQA, was proposed. GSBIQA outperforms conventional and early-stage DL-based methods on two authentic distortion datasets. It also offers competitive performance with a significantly smaller model size than state-of-the-art methods. It processes 52 images per second using only CPU resources, making it well-suited for mobile and edge devices. Future research directions include enhancing the adaptability of the saliency guidance system to improve performance across varied image types and expanding GSBIQA for blind video quality assessment (BVQA).

References

[1] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[2] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 1098–1105.

[3] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 1, pp. 206–220, 2016.

[4] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on image processing*, vol. 27, no. 1, pp. 206–219, 2017.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[6] A. Chetouani, "Convolutional neural network and saliency selection for blind image quality assessment," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 2835–2839.

[7] Z. Mei, Y.-C. Wang, and C.-C. J. Kuo, "Greensaliency: A lightweight and efficient image saliency detection method," *arXiv preprint arXiv:2404.00253*, 2024.

[8] Z. Mei, Y.-C. Wang, C.-C. J. Kuo, *et al.*, "Lightweight high-performance blind image quality assessment," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 1, 2024.

[9] H. Zeng, L. Zhang, and A. C. Bovik, "Blind image quality assessment with a probabilistic quality representation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 609–613.

[10] S. Su, Q. Yan, Y. Zhu, *et al.*, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.

[11] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[12] L. Shen, C. Zhang, and C. Hou, "Saliency-based feature fusion convolutional network for blind image quality assessment," *Signal, Image and Video Processing*, vol. 16, no. 2, pp. 419–427, 2022.

[13] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1383–1391.

[14] C.-C. J. Kuo, M. Zhang, S. Li, J. Duan, and Y. Chen, "Interpretable convolutional neural networks via feedforward design," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 346–359, 2019.

[15] Y. Yang, W. Wang, H. Fu, C.-C. J. Kuo, *et al.*, "On supervised feature selection from high dimensional feature spaces," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[16] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1072–1080.

[17] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015.

[18] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[19] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1220–1230.

A Document Presentation Attack Detection Scheme with Optical Flow under a Flashlight

Changsheng Chen*, Wenyu Chen, Ximin Chen, Haodong Li
Guangdong Provincial Key Laboratory of Intelligent Information Processing and
Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China
*E-mail: cschen@szu.edu.cn

***Abstract*—The presentation attack of a document image can be launched by simple printing/displaying and capturing operations. It has posed a serious threat to the existing document authentication system. However, the generalization performance of existing document presentation attack detection (DPAD) methods is not satisfactory over document images with diverse contents and qualities. In this work, we propose a novel DPAD method that exploits a new data modality observable in the video recorded under a flashlight, named the optical Flow-based Authentication Network (F-AutheNet). Our F-AutheNet extracts both the spatial and optical flow information from video frames. The spatial data carries information on low-level image features, while the optical flow data shows the variations of reflective light under different materials. The features extracted from both data modalities are fused through a cross-modal focal loss. In the experiment, we collect a comprehensive student card video database (SCVD) under a flashlight, comprising more than 18K genuine, print-recaptured, and screen-recaptured video frames. Experimental results show that, by incorporating features extracted from the optical flow data, the generalization performance of generic DPAD backbones has improved by a large margin under some challenging experimental protocols. Specifically, the average EER metrics have been reduced by 2.73 and 2.84 percentage points under cross-domain experiments covering samples of different capturing distances (*i.e.*, qualities) and contents, respectively.**

I. Introduction

In recent years, the popularity of online business systems has led to a significant rise in the utilization of digital documents. Major service platforms require digital document image submissions, including resident ID and student cards, to confirm the identity and qualifications of users. Using such digital documents is convenient for users to handle business online and significantly improves work efficiency. However, verifying the authenticity of document images online is challenging. Illegal users can utilize image editing software (*e.g.*, Adobe Photoshop) or generative networks [1] to tamper with some essential information in document images and conceal the tampering traces of documents through presentation attacks (by displaying/printing and imaging operations) to deceive the system. Therefore, performing the document presentation attack detection (DPAD) task in a document authentication system is imperative.

Most existing works for recaptured image detection focus on natural and biometric images. The keys to the presentation attack detection of natural and biometric images, respectively, are to distinguish the scene depth [2] and living cues [3] (*e.g.*, Specular reflections from the iris region). However, no such forensic trace is available in the DPAD task. Document images are typically captured from a flat paper surface, leading to great difficulty in identifying the recaptured version.



Fig. 1. Some document images taken by smartphones with the flash turned on. (a) Samples of British residence permit. Left: genuine. Right: screen-recaptured. (b) Samples of ID card images from DLC-2021 dataset [4]. Left: genuine. Right: print-recaptured.

Lately, DPAD has attracted extensive research attention. Li *et al.* [5] proposed a two-branch deep network that leverages a designed frequency filter bank and a multi-scale cross-attention fusion module to extract more comprehensive recapture artifacts. Chen *et al.* [6] proposed a frequency-domain halftoning augmentation strategy to enhance the halftoning features associated with recapturing distortion. These works provide reliable solutions for the DPAD task, but they do not consider the samples recaptured through a display.

Further, some DPAD methods have investigated the generalization performance under different recapturing channels. Mudgalgundurao *et al.* [7] proposed a DenseNet-based pixel-wise supervision method to detect document presentation attacks by leveraging the subtle texture differences between genuine and recaptured samples. Chen *et al.* [8] proposed a three-input Siamese network structure paired with a forensic similarity concept for comparing the features extracted from recaptured and genuine document images. These methods mainly rely on low-level features (*e.g.*, textures and edges) of the document. Thus, they are sensitive to the quality and content of the document images.

In this work, we improve the generalization performance of the existing DPAD backbones by introducing a new data modality, the optical flow. Unlike existing methods solely

based on RGB images, the proposed method operates on videos of documents recorded under a flashlight. Based on the differences in reflective lights produced by genuine and recaptured documents, we propose an optical **F**low-based **Authe**ntication **Net**work (F-AutheNet) for detecting the recaptured document images by characterizing the intrinsic difference in the rendering materials. Specifically, our F-AutheNet extracts both the spatial and optical flow information from video frames. The spatial data carries information on low-level image features, while the optical flow data shows the variations of reflective light under different materials. As shown in Fig. 1, we observe that the genuine document images under flash show light spots since many important documents (such as ID cards) are with glossy surfaces. Meanwhile, the recaptured document images produced by displaying or printing devices show reflective lights with different shapes and sizes due to the distinct reflective properties of display surfaces and printing substrates. The features extracted from both data modalities are fused through a cross-modal focal loss.

In the experiment, we collect a comprehensive student card video database (SCVD) under a flashlight, comprising more than 18K genuine, print-recaptured, and screen-recaptured video frames. To evaluate the generalization performance of the proposed method, our experimental protocols consider training and testing data across different capturing environments, distances, and document types. By incorporating optical flow data, our F-AutheNet outperforms the generic deep learning methods, which are based on the same backbones but use only spatial information. In particular, the proposed method shows better robustness under different capturing distances (*i.e.*, qualities) from 10 to 30 cm with an average improvement of EER and AUC by 2.73 p.p. and 0.024, respectively, over deep models trained by spatial data only. Moreover, the effectiveness of our F-AutheNet is also demonstrated by the evaluation of the DLC-2021 dataset (consisting of passport and ID card images) shared by the industry.

II. Proposed Method

A. Physical Model

We discover that the properties of optical flow can be utilized for identity documents that possess reflective surfaces. As shown in Fig. 2, the surface of many important document documents is usually composed of glossy material, which reflects the ambient light sources and captures environmental reflections. When the light from the flash is illuminated on the card surface, it results in a spot area. During video capture, there is movement between the imaging device and the document being recorded due to an unsteady hand. Such movement changes the position of the spot region in different video frames. As the spot area corresponds to the virtual image of the flash on the camera, it is not moving in the same direction as the document under inspection. Thus, the optical flow direction of the spot area is significantly different from that of other document regions.

Different reflective characteristics can be employed as a forensic feature in the DPAD task. Specifically, as shown in

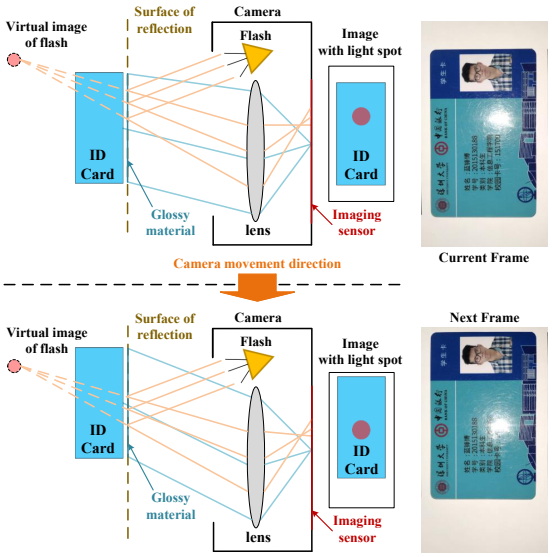


Fig. 2. Illustration of optical flow difference in an identity document with two video frames recorded under a flashlight. The relative movement between a camera and an ID card changed the position of the ID card in the next video frame, while the position of the spot area remained unchanged.

Fig. 3, two frames from the video clip recorded with the flashlight turned on are input into an optical flow algorithm, and the optical flow can be generated. The magnitude and direction of each pixel's displacement can be reflected in the flow field color coding. Color encodes flow direction, and intensity encodes flow magnitude, while white means no motion. If the document is a genuine identity document, an area of light spot will appear on the document. The estimated optical flow will display a spot-shaped area with a distinct color that sets it apart from other regions. This indicates the difference in movement directions of the spot area in genuine document images. We denote this spot-shaped area as peaks in optical flow data. Conversely, such obvious reflective phenomena are not present in the recaptured documents rendered on paper due to the diffusive property of paper. It results in optical flow data without peaks. For documents shown on displays, the recaptured images show light spots with different shapes and sizes than those observed in genuine document images.

B. Data Preprocessing

1) Spatial Data Generation: In our approach, we verify the authenticity of a questioned document by recording a video clip with a flashlight. A video clip is a sequence of frames $V = \{F_1, \ldots, F_T\}$, where T is the total number of frames. To crop out the document region in each frame, we employ the Scale-Invariant Feature Transform (SIFT) algorithm [9] to match the feature points of the samples in V with their original template. Subsequently, we segment these frames $\{F_1, \ldots, F_T\}$ into numerous image patches of 224×224 pixels, yielding the required spatial data.

2) Optical Flow Data Generation: We first match the n-th ($n \in [1, 30]$) video frame in the first and the last seconds, respectively, and compute the bidirectional flow separately for

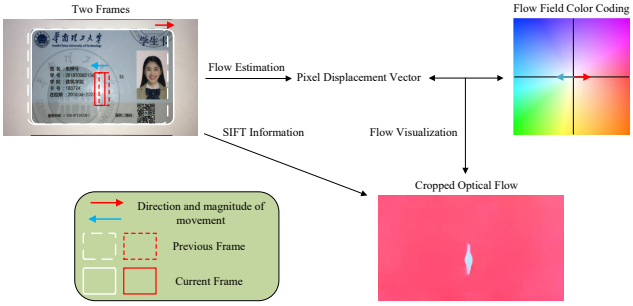


Fig. 3. Illustration of optical flow difference in an identity document with two video frames recorded under a flashlight.

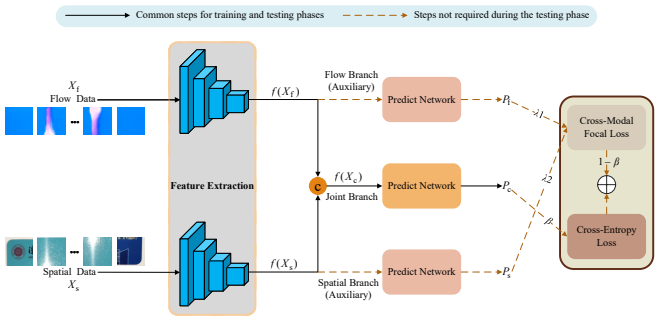


Fig. 4. The network architecture of the proposed F-AutheNet. The training process consists of three main stages: feature extraction, feature concatenation, and classification decision.

each pair. Then, to ensure the optical flow and spatial data are spatially aligned, we preserve the coordinate information generated by the SIFT algorithm [9] during the spatial data cropping process and use this information to crop the optical flow of the corresponding frame. Similarly, we segment these images into numerous image patches of 224×224 pixels, obtaining the required optical flow data. The spatial data and flow data are denoted as X_{s_i} and X_{f_i}, respectively, where i is the index in a minibatch.

C. Architecture of the Proposed F-AutheNet

Our F-AutheNet processes the spatial and optical flow data with a two-stream network and cross-modal focal loss. As shown in Fig. 4, we feed the spatially matched X_{s_i} and X_{f_i} into the two branches of the backbone networks simultaneously and extract a 256-dimensional feature vector from each branch. $f(X_\mathrm{s})$ focuses on low-level distortion features in RGB images, while $f(X_\mathrm{f})$ is used to model the motion characteristics of document content and light sources. The forensic features carried out in both data modalities are complementary. $f(X_\mathrm{s})$ detects signs of capturing distortion in samples that lack optical flow motion features, while $f(X_\mathrm{f})$ identifies the materials of documents when low-level distortion feature extraction is challenging (*e.g.*, image blur). Therefore, both features should be jointly considered to leverage their complementary advantages.

To take advantage of both data modalities in addressing difficult samples in the DPAD task, we employ the cross-modal focal loss (CMFL) [10], that is

$$\mathcal{L}_{CM}(P_\mathrm{s}, P_\mathrm{f}) = -\alpha(1-\omega(P_\mathrm{s}, P_\mathrm{f}))^\gamma \log(P_\mathrm{s}), \tag{1}$$

where α is a weight for the amount of genuine and recaptured samples and γ is a tunable focusing parameter utilized to control the loss weight of samples [11]. P_s and P_f are the spatial and optical flow branches' classification probabilities, respectively. The probability P_s can be written as

$$P_\mathrm{s} = \begin{cases} P & \text{if } y=1, \\ 1-P & \text{if } y=0, \end{cases} \tag{2}$$

where $y \in \{0,1\}$ denotes the class label (0 represents recaptured, 1 represents genuine), and $P \in [0,1]$ is the prediction probability for the labeled class from the spatial branch. The symbol P_f can be defined similarly.

The function $\omega(P_\mathrm{s}, P_\mathrm{f})$ outputs a modulating factor whose magnitude depends on the probabilities obtained from the two single-branch. Specifically, the values of $\omega(P_\mathrm{s}, P_\mathrm{f})$ and $\omega(P_\mathrm{f}, P_\mathrm{s})$ increase with respect to the probabilities of P_f and P_s, respectively, that is

$$\omega(P_\mathrm{s}, P_\mathrm{f}) = P_\mathrm{f}\frac{2P_\mathrm{s}P_\mathrm{f}}{P_\mathrm{s}+P_\mathrm{f}}, \quad \omega(P_\mathrm{f}, P_\mathrm{s}) = P_\mathrm{s}\frac{2P_\mathrm{s}P_\mathrm{f}}{P_\mathrm{s}+P_\mathrm{f}}. \tag{3}$$

Combining Eq. (1) and Eq. (3), we can find that when one branch correctly classifies the samples with high confidence, the loss contribution from the other branch can be reduced.

To be specific regarding our task, we add two branches to the single-channel decision process and incorporate a fully connected layer into all three decision branches to predict their respective probabilities. For the two auxiliary branches, we apply the CMFL. Moreover, decision information from $f(X_\mathrm{c})$ can also be fused to take advantage of their common strengths when two sets of feature decisions are consistent. In this branch, we apply the CE loss.

Therefore, the overall loss function for training the proposed F-AutheNet can be written as

$$\mathcal{L} = (1-\beta)\cdot\left[\lambda_1\cdot\mathcal{L}_{CM}(P_\mathrm{s}, P_\mathrm{f}) + \lambda_2\cdot\mathcal{L}_{CM}(P_\mathrm{f}, P_\mathrm{s})\right] + \beta\cdot\mathcal{L}_{CE}(P_\mathrm{c}), \tag{4}$$

where λ_1 and λ_2 are the adaptive weights between the two types of features, and $\lambda_1 + \lambda_2 = 1$. β is used to balance the two losses. P_c is the classification probabilities of the joint branch and can be defined using a similar fashion as P_s.

In the inference stage, the probability obtained from the joint branch is chosen as the decision probability.

III. Database and Experiment Results

A. Experimental Datasets

1) SCVD: As shown in Sec. II-A, samples collected under a flashlight yield distinctive optical flow features. However, many document image datasets are usually collected without a flashlight. In this section, we build a document image video dataset under a flashlight suitable for our research. We choose the student cards of 10 universities (*e.g.*, Sun Yat-sen University, Shenzhen University) as the original document images. Specifically, we manufacture 40 student cards based on the templates of 10 universities, and each template has

TABLE I

THE DEVICES USED FOR COLLECTING SCVD. THE SYMBOL 'V' INDICATES THAT THESE DEVICES ARE USED FOR VIDEO SHOOTING.

	Imaging Devices		Printing Devices	Displaying Devices
Phone	HONOR 50SE (108 MP) (V)	iQOO Z5 (V)	HP LaserJet m176n (1200 DPI)	Dell-U2718Q (3840 × 2160)
	OPPO K9x (108 MP) (V)	iPhone 6s (V)	Canon C3530 (1200 DPI)	Lenovo-T2324C (1920 × 1080)
Scanner	Canon 5600f (1200 DPI)		HP 258 (600 DPI)	Rz09-0288 (1920 × 1080)
	Epson V850 (3200 DPI)		Epson L805 (5760 × 1440 DPI)	iPad-Mini4 (2048 × 1536)



(a) (b) (c) (d)

Fig. 5. Some video frames with light spots. (a) Genuine samples. Top: from SCVD-I10. Bottom: from SCVD-O10. (b) Print-recaptured samples. Top: from SCVD-I^{P}10. Bottom: from SCVD-O10. (c) Screen-recaptured samples from SCVD-I^{S}10. (d) Print-recaptured samples covered with transparent plastic from SCVD-I^{T}10. The final imaging device is identical for samples within the same row.

4 student cards. Compared with our prior work [8], which only has 12 templates of 5 universities, we have enriched and improved the diversity of the samples.

We follow the rules in [8] to gather recaptured documents. The complete list of devices used in the acquisition process is tabulated in Tab. I. To increase the variety of recaptured samples, we cover the spoofed documents with transparent plastic. We consider transparent plastics of different materials (*i.e.*, glossy and matte) to diversify the reflection spots. The frame resolution is set to 1920×1080 pixels, and the frame rate is set to 30 FPS (frames per second). Each video clip has a minimum length of 5 seconds. The samples are collected at a distance of about 10, 20, and 30 cm from the camera to the student card. To sum up, we construct a practical video database of recaptured ID documents comprising 420 video clips. The genuine samples and the print-recaptured samples consist of 120 video clips each, the print-recaptured samples covered with transparent plastic comprise 20 video clips, and the screen-recaptured samples contain 160 video clips. Some video frames with light spots are shown in Fig. 5.

We divide the SCVD into six subsets according to the capturing environments, recapturing types, and shooting distances, as shown in Tab. II. The abbreviations 'I' and 'O' represent indoor and outdoor shooting, respectively. 'P' represents print-recaptured samples, 'T' represents print-recaptured samples covered with transparent plastic, and 'S' represents screen-recaptured samples. Since the samples taken indoors at a distance of 10cm contain three types of presentation attack media, we collectively refer to the three subsets as SCVD-I10.

2) DLC-2021: The DLC-2021 dataset [4] consists of 1424 video clips captured under different acquisition devices and shooting conditions, encompassing multiple national identity documents. Some video clips in this dataset are recorded with the flashlight turned on, and genuine samples within these clips also exhibit light spots. Therefore, we select one video clip from the genuine sample and the color-print recaptured sample, respectively, for each type of ID card. This has resulted in 20 video clips that will be used in our testing process. It contains 640 genuine and 640 print-recaptured video frames.

TABLE II

THE DIVISION OF SCVD AND THE CORRESPONDING RECAPTURING TYPES AND FRAME COUNTS OF EACH SUBSET. THE SYMBOL 'G' REPRESENTS GENUINE SAMPLES, 'R' REPRESENTS RECAPTURED SAMPLES.

Env.	Attack Media	Dist.	# Video Frames	Purpose	Notation
Indoor	Paper	10cm	6000 G, 6000 R	train&test	SCVD-I^{P}10
		20cm	1280 G, 1280 R	test	SCVD-I^{P}20
		30cm	1280 G, 1280 R	test	SCVD-I^{P}30
	Transparency	10cm	1280 R	train&test	SCVD-I^{T}10
	Screen	10cm	6000 R	train&test	SCVD-I^{S}10
Outdoor	Paper	10cm	6000 G, 6000 R	train&test	SCVD-O10

TABLE III

EXPERIMENTAL RESULTS IN EVALUATION OVER DIFFERENT CAPTURING ENVIRONMENTS. I10 → O10 DENOTES TRAINING AND TESTING WITH DATASET SCVD-I10 AND SCVD-O10, RESPECTIVELY, WHILE O10 → I10 MEANS OTHERWISE. THE BEST PERFORMANCE FOR EACH BACKBONE NETWORK IS **BOLD-FACED**.

Methods	I10 → O10		O10 → I10	
	EER	AUC	EER	AUC
Single modality: Optical Flow Only				
ResNet50	18.30%	0.8930	32.17%	0.7268
DenseNet121	20.62%	0.8615	30.68%	0.7347
ViT-B16	15.89%	0.9166	29.23%	0.7437
Swin-B	14.70%	0.9343	30.55%	0.7458
Single modality: Spatial Only				
ResNet50	**0.01%**	**1.0000**	11.35%	0.9432
DenseNet121	**0.01%**	**1.0000**	12.59%	0.9360
ViT-B16	**0.01%**	**1.0000**	4.14%	0.9817
Swin-B	**0.01%**	**1.0000**	3.22%	0.9841
Multi-modalities: Optical Flow + Spatial				
Li *et al.* [12]	0.98%	0.9997	42.88%	0.6231
ResNet50 (ours)	**0.01%**	**1.0000**	**9.25%**	**0.9665**
DenseNet121 (ours)	**0.01%**	**1.0000**	**10.14%**	**0.9584**
ViT-B16 (ours)	**0.01%**	**1.0000**	**3.06%**	**0.9879**
Swin-B (ours)	**0.01%**	**1.0000**	**2.65%**	**0.9909**

B. Implementation Details

In this work, we select ARFlow [13] as our model for generating optical flow data. To demonstrate the advantage of the proposed approach, we employ some popular CNN-based models (ResNet50 [14], DenseNet121 [15]) and transformer-based models (ViT-B16 model [16], Swin-B model [17]). These models are pre-trained with the ImageNet-1K dataset, which is subsequently fine-tuned with our document recapturing dataset. We also compare with a method based on optical flow [12]. These models are trained using the Adam optimizer, and the batch size is 128. The training period lasts for 20 epochs with a learning rate of $1e^{-4}$. The values of λ_1 and β in our F-AutheNet are assigned as 0.5, whereas the value of γ is set to 3. Performance is evaluated using the Area Under the ROC Curve (AUC) and Equal Error Rate (EER) metrics at the image-level. Our implementation is based on Pytorch 1.12.1.

Given that most methods achieve good performance under intra-dataset settings, we elaborate on the experimental results of cross-dataset experiments, which involve different capturing environments, capturing distances, and document contents.

TABLE IV
EXPERIMENTAL RESULTS OVER DIFFERENT CAPTURING DISTANCES FROM 10 TO 30 CM.

Methods	$I^P10 \rightarrow I^P20$		$I^P10 \rightarrow I^P30$		$O10 \rightarrow I^P20$		$O10 \rightarrow I^P30$	
	EER	AUC	EER	AUC	EER	AUC	EER	AUC
Single modality: Optical Flow Only								
ResNet50	27.50%	0.7961	19.96%	0.8899	28.75%	0.7906	**20.53%**	0.8494
DenseNet121	25.11%	0.8223	23.90%	0.8546	27.53%	0.8024	**21.99%**	0.8614
ViT-B16	20.39%	0.8679	21.17%	0.8919	22.49%	0.8476	22.89%	0.8780
Swin-B	19.33%	0.8759	15.07%	0.9429	23.90%	0.8331	19.99%	0.8905
Single modality: Spatial Only								
ResNet50	11.64%	0.9674	21.01%	0.8765	11.40%	0.9692	26.32%	0.8060
DenseNet121	18.04%	0.9130	27.30%	0.8088	12.50%	0.9454	36.01%	0.7119
ViT-B16	8.12%	0.9835	12.92%	0.9417	7.18%	0.9826	18.71%	0.8910
Swin-B	3.67%	0.9966	9.60%	0.9674	6.64%	0.9863	18.71%	0.9119
Multi-modalities: Optical Flow + Spatial								
Li *et al.* [12]	12.54%	0.9444	36.79%	0.6886	20.15%	0.8481	41.60%	0.6109
ResNet50 (ours)	**9.53%**	**0.9723**	**16.48%**	**0.9070**	**7.34%**	**0.9840**	22.07%	**0.8696**
DenseNet121 (ours)	**16.87%**	**0.9192**	**21.60%**	**0.8750**	**10.85%**	**0.9689**	25.15%	**0.8664**
ViT-B16 (ours)	**7.68%**	**0.9854**	**11.05%**	**0.9467**	**6.32%**	**0.9838**	**15.58%**	**0.9338**
Swin-B (ours)	**2.85%**	**0.9976**	**7.57%**	**0.9778**	**5.82%**	**0.9895**	**12.34%**	**0.9535**

C. Evaluation over Different Capturing Environments

As shown in Fig. 5(a), the shape and size of light spots may vary across different capturing environments due to the differences in the intensity and distribution of light sources under indoor and outdoor environments. Consequently, it is necessary to evaluate the performance of our F-AutheNet over different capturing environments.

In this experiment, we select the samples of SCVD-I10 and SCVD-O10 for training and testing, respectively. As shown in Tab. III, most models exhibit a significant performance degradation in the experimental protocol SCVD-O10 $\rightarrow$ SCVD-I10. This is because the SCVD-O10 does not contain screen-recaptured samples, and these samples do not have discernible moiré patterns. The poor performance of the method in [12] is because it assigns higher weights to regions with smaller motion magnitudes in the optical flow of recaptured samples through the attention module, while document images are captured from flat surfaces, which makes the method unable to effectively integrate spatial and optical flow features.

The results in Tab. III demonstrate the robustness of our approach under the unknown samples forged by different recapturing environments and channels. While these recaptured samples may have obvious reflection spots on their surfaces, the differences in the shape and size of the peaks in the optical flow compared to genuine samples allow our method to classify them as recaptured samples accurately. Given the error on unknown samples caused by the spatial branch, the optical flow branch provides additional information to improve the detection performance.

D. Evaluation over Different Capturing Distances

With large capturing distances, the quality of the samples would be poor due to shaking, blurring, and low resolution. If a model exclusively relies on noisy spatial data to distinguish genuine samples from recaptured samples, it encounters difficulty in making a reliable decision. Under such scenarios, the peaks-based feature in the optical flow data plays a crucial role in our F-AutheNet.

TABLE V
EXPERIMENT RESULTS IN EVALUATION OVER DIFFERENT TYPES OF DOCUMENTS.

Methods	$I^P10 \rightarrow$ DLC		$O10 \rightarrow$ DLC	
	EER	AUC	EER	AUC
Single modality: Optical Flow Only				
ResNet50	30.62%	0.7654	38.43%	0.6100
DenseNet121	32.42%	0.7146	39.37%	0.6165
ViT-B16	38.84%	0.6924	40.07%	0.6438
Swin-B	25.23%	0.7889	35.93%	0.6207
Single modality: Spatial Only				
ResNet50	14.68%	0.9475	10.46%	**0.9709**
DenseNet121	21.48%	0.8794	11.71%	0.9694
ViT-B16	16.71%	0.9323	7.49%	0.9859
Swin-B	10.47%	0.9654	11.95%	0.9588
Multi-modalities: Optical Flow + Spatial				
Li *et al.* [12]	28.36%	0.8199	53.51%	0.5367
ResNet50 (ours)	**12.18%**	**0.9519**	**8.28%**	0.9663
DenseNet121 (ours)	**19.60%**	**0.8970**	**9.92%**	**0.9746**
ViT-B16 (ours)	**15.85%**	**0.9401**	**1.01%**	**0.9983**
Swin-B (ours)	**10.31%**	**0.9695**	**5.07%**	**0.9922**

To demonstrate the advantage of F-AutheNet, we first evaluate the models trained by SCVD-I^P10 with testing sets acquired under different distances, *i.e.*, SCVD-I^P20, and SCVD-I^P30. As shown in Tab. IV, it can be seen that the performance of generic deep models has dropped significantly, and the farther the distance, the worse the performance. In contrast, the proposed F-AutheNet yields lower error rates. The average EERs are reduced by 1.14 and 3.53 percentage points (p.p.) for SCVD-I^P10 $\rightarrow$ SCVD-I^P20 and SCVD-I^P30 protocols, respectively. Experimental results demonstrate that using our F-AutheNet in the forensic stage can enhance detection accuracy, even in challenging scenarios.

Subsequently, we employ SCVD-O10 as the training set and SCVD-I^P20 and SCVD-I^P30 as the testing sets. As shown in Tab. IV, the performance of our F-AutheNet based on different backbones has been significantly improved in most experiments. In particular, the AUC of the DenseNet121 model has increased by 0.024 and 0.155, while the EER has decreased by 1.65% and 10.86%. Although differences in samples may arise due to diverse capturing environments, our F-AutheNet

shows strong generalization performance.

E. Evaluation over Different Types of Documents

A practical document authentication system should be robust under cross-domain experiments covering different document images. The genuine documents in the DLC-2021 dataset are laminated by an office laminator. Gray-spoofed documents are made by office HP Laser B/W printer from images on ordinary office paper, and color-spoofed documents are made by laser color printer on photo-paper matte photo paper. The acquisition and imaging devices used differ from those employed in SCVD. Simultaneously, the samples included in the DLC-2021 dataset consist of ID cards and passports. To validate the robustness of our algorithm, we select 1280 video frames from the DLC-2021 dataset according to Sec. III-A2. The models trained by SCVD-I^{P}10 and SCVD-O10 are evaluated by the testing set from DLC-2021, respectively. Thus, this protocol covers the differences in document types, printing substrates, and acquisition devices.

As shown in Tab. V, our F-AutheNet achieves the best performance in most cases. The average EERs are reduced by 1.35 and 4.33 percentage points (p.p.) for SCVD-I^{P}10 and SCVD-O10 $\rightarrow$ DLC-2021 protocols, respectively. For example, the EER of ViT-B16 decreases from 7.49% to 1.01%. The generalizability of the proposed F-AutheNet is successfully demonstrated by achieving a lower EER under such a challenging protocol.

IV. Conclusions

In this work, we propose F-AutheNet to improve the generalization performance of the existing DPAD methods. Our F-AutheNet takes advantage of optical flow information from video clips to capture the variations of reflective light under different rendering materials while extracting low-level features from the images. Additionally, we employ a cross-modal focal loss to better leverage these two features. Experimental results show that the generalization performance of the generic DPAD backbones has improved by reducing average EERs by 2.73 p.p. and 2.84 p.p. under cross-domain evaluation with document images of different qualities and contents, respectively.

In the future, we plan to exploit a more stable optical flow estimation algorithm suitable for documents with reflection and spot areas. A new overlapping model to characterize spot area on the document surface should be established instead of employing the existing occlusion model for scenarios where foreground objects move before background regions. Moreover, we aim to detect stable optical flow information without the aid of a flashlight. Keeping a good DPAD performance without additional lighting will make the user experience in the document authentication process even better.

Acknowledgment

This work was supported in part by the Nature Science Foundation of China (NSFC) under Grant 62371301 and 62072313, and in part by the Guangdong Provincial Key Laboratory under Grant 2023B1212060076.

References

[1] L. Zhao, C. Chen, and J. Huang, "Deep Learning-based Forgery Attack on Document Images," *IEEE Transactions on Image Processing*, vol. 30, pp. 7964–7979, 2021.

[2] H. Li, S. Wang, and A. Kot, "Image recapture detection with convolutional and recurrent neural networks," *Electronic Imaging*, vol. 2017, pp. 87–91, Jan. 2017.

[3] A. F. Ebihara, K. Sakurai, and H. Imaoka, "Efficient face spoofing detection with flash," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 535–549, 2021.

[4] D. V. Polevoy, I. V. Sigareva, D. M. Ershova, *et al.*, "Document liveness challenge dataset (dlc-2021)," *Journal of Imaging*, vol. 8, no. 7, p. 181, 2022.

[5] J. Li, C. Kong, S. Wang, and H. Li, "Two-branch multi-scale deep neural network for generalized document recapture attack detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] C. Chen, B. Li, R. Cai, J. Zeng, and J. Huang, "Distortion model-based spectral augmentation for generalized recaptured document detection," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 1283–1298, 2024.

[7] R. Mudgalgundurao, P. Schuch, K. Raja, R. Ramachandra, and N. Damer, "Pixel-wise supervision for presentation attack detection on identity document cards," *IET Biometrics*, 2022.

[8] C. Chen, S. Zhang, F. Lan, and J. Huang, "Domain-Agnostic Document Authentication Against Practical Recapturing Attacks," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2890–2905, 2022.

[9] D. G. Lowe, "Object Recognition from Local Scale-invariant Features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 1999, pp. 1150–1157.

[10] A. George and S. Marcel, "Cross modal focal loss for rgbd face anti-spoofing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[12] L. Li, Z. Xia, J. Wu, L. Yang, and H. Han, "Face presentation attack detection based on optical flow and texture analysis," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1455–1467, 2022.

[13] L. Liu, J. Zhang, R. He, *et al.*, "Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation," in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2020.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[17] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 012–10 022.

Robust Image Watermarking Scheme under Halftone Distortion with Surrogate Model

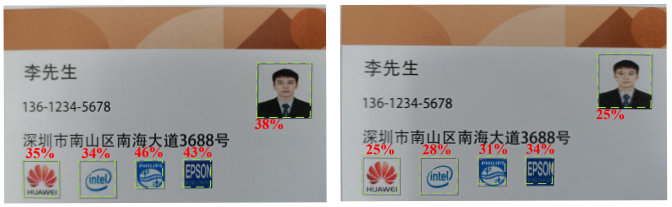
Changsheng Chen* and Xijin Li
Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, China
Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen, China
*E-mail: cschen@szu.edu.cn

***Abstract*—Recently, document images have been widely used in various online applications. Digital watermarking is an important forensic technique to verify the authenticity of a document image. However, the recapturing operation leads to a significant risk in document images since the extraction accuracy of recaptured digital watermarking drops significantly. In this work, we propose robust watermark against halftone distortion by utilizing surrogate models and end-to-end watermarking frameworks. Firstly, we employ differentiable surrogate models to generate the halftone distortion. Then, surrogate models are incorporated into end-to-end watermarking frameworks to enhance the robustness of the watermark in the print-camera scenario. To evaluate the robustness of the proposed method, we conduct a series of experiments in real-world scenarios. The experimental results confirm that our method can improve the robustness of the watermark across different devices, datasets, printing sizes, and watermark capacities.**

I. Introduction

The document image is an important information carrier widely used in e-business and e-government applications. However, due to the development of image editing tools and image generation methods, document image security is facing great threats. Active image forensic techniques like digital watermarking and passive image forensic techniques like tamper detection are commonly utilized to ensure the security of document images. However, researchers have found that recapturing operations significantly reduces the effectiveness of existing document image forensic methods. For instance, Zhao *et al* [1]. utilize a deep learning scheme to edit practical document images and conceal the traces of tampering by recapturing operations. Moreover, as shown in Fig. 1, recaptured watermarked images suffer severe distortions, resulting in a significant decline in extraction accuracy. Specifically, the error bit rate of watermarks in small printing sizes is close to 50%. Therefore, research on document image forensics techniques in recapturing scenarios is of practical significance and demands our attention. In this work, we focus on improving the robustness of digital watermarking in the print-camera scenario.

Previous research in digital watermarking has mainly focused on robustness against digital channel distortion, which mainly consists of Gaussian noise, Gaussian blur, and other factors. Recently, capturing a printed image with a mobile phone has gained popularity as information acquisition. Compared to digital channel distortion, print-camera channel distortion is much more complex, including perspective distortion, color distortion, and environmental noise, which brings significant challenges to digital watermarking. Therefore, robust watermark in the print-camera scenario has gained great attention in recent years. However, most existing works focus on the distortions introduced by the shooting process while neglecting halftone distortion. Through our research on the distortion produced by the print-camera process, we find that halftoning technology is widely used in current mainstream printers, which means that the halftone distortion is a key part of the print-camera distortion.



(a) StegaStamp (b) StegaStamp+S

Fig. 1. The practical application of robust watermark against halftone distortion. The images are embedded with 200 bits with printing sizes of 1.5×1.5 cm^2 and 0.75×0.75 cm^2. The bit error rate is shown next to the corresponding watermark image. (a) The application example of StegaStamp [2]. (b) The application example of our method.

Given the above limitations, we propose robust watermark against halftone distortion. This approach aims to enhance the robustness of digital watermarking and improve the security of current image forensic systems. We train a CycleGAN-based simulation network for the halftone distortion and employ end-to-end watermarking frameworks based on deep learning. To evaluate the robustness of the proposed method, we conducted a series of experiments in real-world application scenarios, including different devices, datasets, printing sizes, and watermark capacities. The experimental results demonstrate that the proposed method achieves better performance than SOTA methods, and the average bit error is lower than 20% in most scenarios.

The main contributions of this work are as follows:

1) We employ a simulation network of the halftone distortion based on CycleGAN to enhance the robustness of the watermark. Moreover, this simulation network can be incorporated into different deep learning-based watermarking frameworks.

2) Experimental results confirm that our method achieves better robustness in several real-world scenarios, including different devices, datasets, printing sizes, and watermark ca-

pacities.

II. The Proposed Method

A. Surrogate Model for Halftone Distortion

1) The Printing Process (Inkjet, LaserJet): The halftoning technique is utilized by printers to simulate a color image in continuous tone through halftone dots with variable sizes or spacing. Dispersed dot and cluster dot halftoning [3] are two classes of halftoning patterns that are widely used by inkjet printers and laserjet printers, respectively.

Mainstream inkjet printers employ the error diffusion halftoning technique to produce the dispersed dot halftoning pattern. During the error diffusion process, each color channel is processed with the following steps [3]. First, each pixel value of the input image is set to 0 or 255 according to the threshold value of 128. Second, the error value is calculated between the assigned value in the first step and the actual value. Finally, according to a predefined error filter kernel, the error is accumulated and diffused to the neighborhood pixels. In our implementation, we adopt the error filtering kernel defined by Floyd and Steinberg [4], and weights in this filter are [0, 0, 7/16; 3/16, 5/16, 1/16].

Mainstream laserjet printers produce the cluster dot. When we consider a single-channel image patch, the representation of the halftone can be described as follows

$$I_L(\mathbf{x}) = \sum_m \sum_n \delta(\mathbf{x} - m\mathbf{a} - n\mathbf{b}) \otimes H(\mathbf{x}), \tag{1}$$

where $\mathbf{x} = (x, y)$ represents the horizontal and vertical coordinates of each pixel of the image, 'δ' denotes the Dirac delta function, spatial halftone vectors **a** and **b** are the density and direction of the halftone array, m, n represent indices of halftone dots along the direction of **a** and **b**, respectively, '$\otimes$' denotes the convolution operation, and $H(\mathbf{x})$ is a binary masking function that defines the shape of halftone dots. In our implementation, we adopt the ordered dithering technique with a halftone dot size of 4×4 pixels. When considering a color image, the halftoning process described in Eq. (1) operates independently in each color channel.

2) A Surrogate Model for Halftone Distoriton: The noise layer is vital in our method. Moreover, the end-to-end watermarking framework requires a differentiable noise layer. Chen *et al*. [5] propose a distortion model guided surrogate model based on CycleGAN structure with two training stages to generate distortions introduced by the recapturing process. The trained surrogate model can be utilized to improve the generalization performance of networks. What's more, the surrogate model consists of a series of differentiable operations, which meets the requirement of the end-to-end watermarking framework.

Therefore, inspired by [5], we utilize CycleGAN-based surrogate models for generating halftone distortions, aiming to improve the robustness of the watermark. We collect three types of images for the surrogate model training, including original images I_O, simulated halftone images I'_H yield by the process described in Sec. II-A1, and real halftone images I_H. The training process of the surrogate model is shown in Fig. 2. In the first stage, I_O and I'_H are employed to train the surrogate model for the basic ability to generate halftone distortions. In the second stage, real halftone images I_H are utilized to finetune the discriminator. In addition, I_O and I'_H are still participating in training to provide a reliable source for learning halftone distortion.

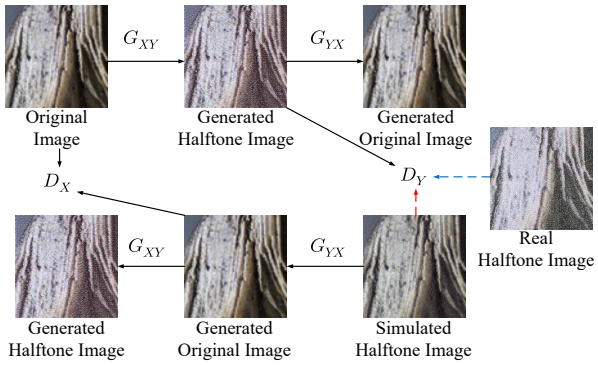


Fig. 2. The surrogate model for halftone distortion. The black arrows indicate the steps for two stages. The red dashed line represents the step only for the first stage. The blue dashed line represents the step only for the second stage.

B. Watermarking Framework

We adopt the end-to-end watermarking framework to realize robust watermark against halftone distortion, which consists of an encoder, noise layer, decoder, and discriminator. Specifically, we improve the noise layer by utilizing surrogate models for halftone distortion. The generic watermarking framework is shown in Fig. 3. The UNet-like encoder embeds the watermark and maintains the quality of the watermarked image. The encoder converted the concatenated watermark and original image to a residual image, which is overlaid on the original image to finish the watermark embedding. The decoder consisting of convolution layers and fully connected layers recovers the watermark from the distorted image by converting the image to a vector with the same length as the original watermark. The discriminator outputs a binary sequence to judge whether the image is watermarked. Similar to generative adversarial networks, adversarial training between the discriminator and encoder can further improve the visual quality of the watermarked image. The overall loss function of the generic watermarking framework can be written as

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_I + \lambda_2 \cdot \mathcal{L}_W + \lambda_3 \cdot \mathcal{L}_D, \tag{2}$$

where $\mathcal{L}_I$, $\mathcal{L}_W$, and $\mathcal{L}_D$ denote the image reconstruction loss, watermark loss, and discriminator loss. λ_1, λ_2, λ_3 are the weights for three loss components. Image reconstruction loss can be MSE loss, LPIPS perceptual loss [6], etc, which are utilized to improve the visual quality of the watermarked image. Discriminator loss aims to distinguish the original image and the watermarked image and usually adopts Wasserstein loss [7], which is commonly utilized in generative adversarial

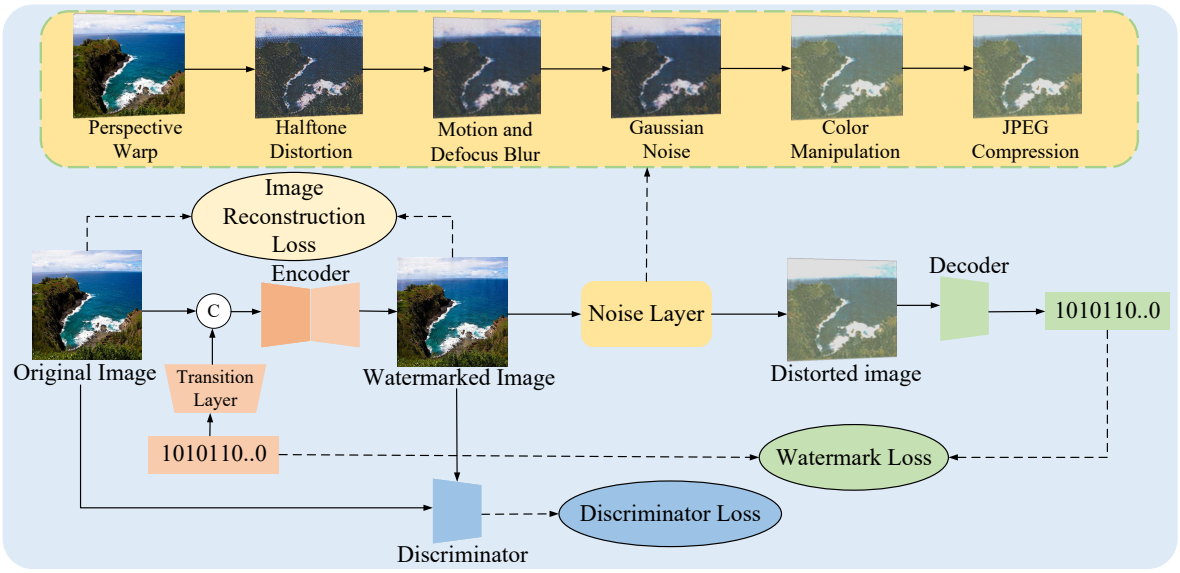


Fig. 3. The framework of robust watermark against halftone distortion.

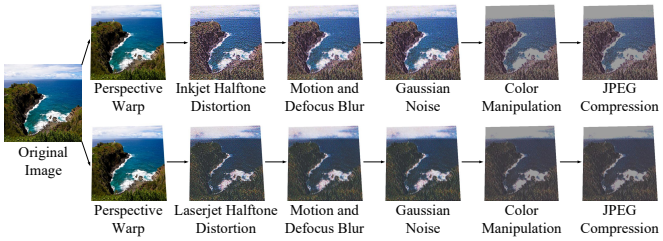


Fig. 4. The visualization of each step of distortion layers.

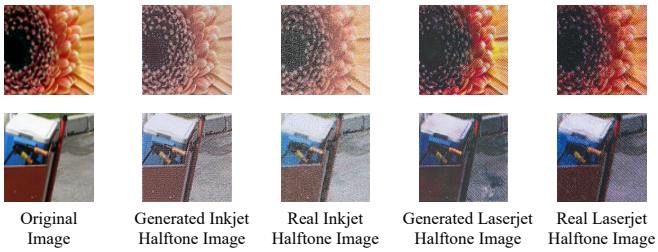


Fig. 5. Halftone images generated by surrogate models and collected in the real environment.

networks. Watermark loss aims to minimize the difference between the original watermark and the recovered watermark, utilizing cross entropy loss.

The noise layer is the key to robust watermark against halftone distortion. The noise layer in StegaStamp mainly focuses on the distortion introduced by the shooting process, while ignoring the halftone distortion in the printing process. Our surrogate model focuses on generating the halftone pattern during the printing process. Therefore, we combine the halftone distortion surrogate model with the noise layer proposed by StegaStamp. The visualization of each step of distortion layers is shown in Fig. 4. The details of the noise layer are shown as follows:

1) Perspective Warp: The unevenness of the paper during the printing process and the deviation of the lens during the shooting process can cause perspective deformation. To simulate this distortion, a random perturbation is employed to the corner locations of watermarked images ($\pm$ 40 pixels), and homographic transformation is adopted to map the watermarked images to new locations.

2) Halftone Distortion: Two surrogate models are trained to simulate inkjet and laserjet halftone distortion, respectively. As shown in Fig. 5, surrogate models can generate halftone distortions similar to real halftone images. To avoid the watermark network overfit halftone distortion, both of them are added to the noise layer and the probability is both set as 50%. The halftone distortion enables the encoder to adjust the embedding area, intensity, etc automatically. At the same time, it also improves the decoder's ability to recover messages from watermarked images with halftone distortions.

3) Motion and Defocus Blur: The shooting process can introduce blur because of the camera motion and inaccurate autofocus. A random angle between 0 and 2π and a straight line blur kernel with a width between 3 and 7 pixels are employed to simulate motion blur. In addition, a Gaussian blur kernel with a standard deviation sampling between 1 and 3 pixels is adopted to simulate defocus blur.

4) Gaussian Noise: The electronic components in the camera will introduce various noises during the imaging process. A Gaussian noise with the standard deviation sampling in [0, 0.02] is employed to simulate the noise.

5) Color Manipulation: The camera uses white balance, exposure settings, and chroma correction to adjust the output image. These distortions can be approximated with random color transformations, including hue shift, desaturation, and brightness adjustment. the hue shift can be achieved by adding a random color offset sampled uniformly from [-0.1, 0.1] to each RGB channel. The desaturation can be achieved by randomly linearly interpolating between the full RGB image

and its grayscale equivalent. The brightness adjustment can be achieved by a linear transformation $mx + b$ with m in [0.5, 1.5] and b in [-0.3, 0.3].

6) JPEG Compression: The images captured by the camera are usually saved in the JPEG format. Since the rounding step of JPEG compression is not differentiable at zero, the method proposed by Shin *et al* [8] is employed to approximate the quantization step near zero with a piecewise function. The JPEG quality is sampled uniformly in [50, 100].

III. Experimental Results

To train the surrogate model for halftone distortion, we collect the training data with two printers and a scanner. We choose an inkjet printer (Canon G3800) and a laser printer (Canon iR-ADV C3530), which are two main types of printers. Considering that the camera-shooting process may introduce various factors, such as lighting and noise, we utilize a high-quality scanner (CanoScan 5600F) to capture the halftone distortion of the printed images minimizing the factors that affect the halftone distortion capturing as possible. First, We randomly select 300 images from the MIRFLICKR dataset [9] and print them with the size of 5 × 5 cm^2. Then, we scan the printed images and crop them to 400 × 400 resolution. Finally, we collect 15000 patches with halftone distortions for the training of surrogate models.

To train the watermarking network against halftone distortion, we randomly select 20,000 images from the MIRFLICKR dataset [9] as the training dataset. The training images are resampled to 400 × 400 resolution and are embedded with 200 bits random message. The implementation is based on Tensorflow 1.13.1.

We evaluate the robustness performance of the proposed method in the real-world environment. Firstly, 20 images not included in the training set are randomly selected from the MIRFLICKR dataset for watermark embedding. The length of the watermark is 200 bits. Secondly, the watermarked images are printed by inkjet and laserjet with the printing size of 2 × 2 cm^2 on A4 standard paper. Thirdly, the printed images are captured by mobile phones with the shooting distance and angle of 10 cm and 0°, respectively. Finally, the decoder recovers messages from captured images.

The bit error rate (BER) is employed to evaluate the robustness of different methods. A lower bit error rate indicates better robustness. Two SOTA print-camera resistant watermarking methods, StegaStamp [2] and NSN [10], are used for comparison. Our surrogate models are incorporated into these SOTA methods. The details of our experimental configurations are shown below.

StegaStamp: This approach proposes a set of differentiable operations to simulate distortions in the print-camera scenario, including perspective warp, motion and defocus blur, color manipulation, noise, and JPEG compression.

NSN: This approach proposes a noise simulation network. In the earlier training stage, the math model is utilized to simulate distortions in the print-camera scenario. In the later training stage, the math model is replaced with the noise simulation network.

StegaStamp+S: Our noise layer consists of **s**urrogate (S) models for the halftone distortion and differentiable operations proposed by StegaStamp, which generate printing distortion and shooting distortion, respectively. The other configurations (encoder, decoder, training loss, and hyperparameters) of *StegaStamp+S* are the same as those of *StegaStamp*.

NSN+S: In the later training stage, the noise simulation network is replaced with our noise layer which contains surrogate models for halftone distortion. The other configurations (encoder, decoder, training loss, and hyperparameters) of *NSN+S* are the same as those of *NSN*.

To validate the effectiveness of the proposed method, we conduct a series of robustness experiments in different real-world scenarios, including different devices, datasets, printing sizes, and watermark capacities. Detailed experimental results are shown as follows.

A. Visual Quality Evaluation

In this part, we evaluate the visual quality of watermarked images after printing and shooting both quantitatively and qualitatively.

The quantitative comparisons are shown in Tab. I. We employ Peak Signal Noise Ratio (PSNR), Structural Similarity (SSIM) [11], and Frechet Inception Distance (FID) [12] to evaluate the visual quality of watermarked images. It can be seen that watermarked images after the print-camera process perform poorly in visual quality performance indicators. This is because watermarked images undergo various distortions, including halftone distortion, noise, blurring, etc. Moreover, *StegaStamp* and *NSN* achieve slightly better performance than *StegaStamp+S* and *NSN+S* in visual quality. This is because the surrogate model generates obvious halftone distortion, which leads to severe damage to the watermark. Therefore, the encoder continuously adjusts the embedding area and intensity to ensure correct watermark decoding after halftone distortion while sacrificing some visual quality.

The qualitative comparisons are shown in Fig. 6. It can be seen that watermarked images of *StegaStamp+S* and *NSN+S* contain foggy traces visually. This is because the encoder adjusts embedding strength and area of the watermark to combat halftone distortion.

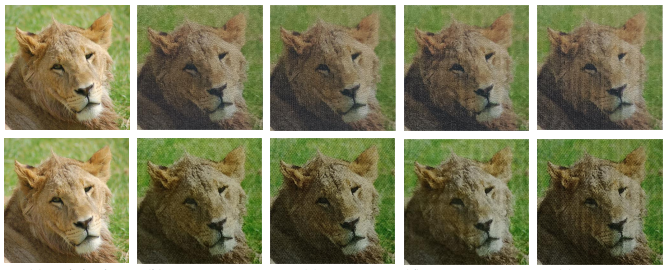
The watermarked images in practical scenarios are usually printed in a small size. Moreover, they undergo various distortions after printing and shooting. The slight decrease in visual quality after using the distortion surrogate model has a relatively small impact and is within the acceptable range. Therefore, we mainly focus on the robustness of the watermark in practical scenarios.

B. Robustness Evaluation

1) Robustness to Different Printing and Shooting Devices: In practical scenarios, watermarked images are commonly captured by various printers and mobile phones with different brands and imaging qualities. Therefore, it is essential to

TABLE I
THE VISUAL QUALITY OF DIFFERENT METHODS. THE BEST PERFORMANCES ARE **BOLD-FACED**.

Methods	Metrics		
	PSNR(dB)	SSIM	FID
StegaStamp	**14.14**	0.2435	**262.26**
NSN	**14.20**	0.2514	**266.02**
StegaStamp + S	14.05	**0.2501**	265.82
NSN + S	14.03	**0.2554**	276.80



(a) Original (b) StegaStamp (c) NSN (d) StegaStamp+S (e) NSN+S

Fig. 6. The visualization of watermarked images after printing and shooting. Top: the images printed with inkjet; Bottom: the images printed with laserjet.

evaluate the robustness performance of watermarks across different devices. In this part, the watermarked images are printed by 4 printers, including inkjet (EPSON L805 and HP OfficeJet 258) and laserjet (Konica Minolta C6500 and HP LaserJet M176n). Then they are captured by 4 mobile phones with different pixel resolutions (100 MP, 64 MP, and 13 MP).

Experimental results are shown in Tab. II. It can be seen that the proposed method has achieved better robustness across different devices. The average bit error rates of our method are close to 10%. This indicates that the surrogate model effectively improves the robustness of the watermark to halftone distortion, although different printers have different parameters for the halftoning process.

2) Robustness to Different Datasets: In practical scenarios, watermarks are commonly embedded in various types of images. Identification photos and logos are two typical images widely used in important documents and are usually the key to verifying the authenticity of document images. For example, Identification photos and logos are used to identify personal and legal identities, respectively, and are high-risk areas for tampering. Therefore, we evaluate the robustness performance of watermarks across different datasets. We randomly select 20 images from the Identification Photo dataset [13] and Logo-2k+ [14], respectively. In this part, "EPSON L805" and "HP LaserJet M176n" are utilized for printing, and "Oppo K9x" is utilized for shooting.

Experimental results are shown in Tab. III. It can be observed that the proposed method has achieved better performance of each dataset with bit error rates dropping below 20%. This demonstrates the robustness of our method across different datasets, even if the content of testing images is significantly different from training images.

3) Robustness to Different Printing Sizes: In practical scenarios, watermarked images are commonly printed in different

TABLE II
BIT ERROR RATE (BER) WITH DIFFERENT PRINTING AND SHOOTING DEVICES. THE BEST PERFORMANCES ARE **BOLD-FACED**.

Devices	Methods	Honor 50se	Oppo K9x	iQoo Z5	Meizu Metal
EPSON L805	StegaStamp	26.25%	27.42%	27.55%	28.12%
	NSN	28.85%	28.55%	28.80%	29.77%
	StegaStamp + S	**11.15%**	**11.42%**	**11.67%**	**11.45%**
	NSN + S	**11.07%**	**11.22%**	**11.20%**	**12.13%**
HP OfficeJet 258	StegaStamp	22.60%	23.90%	22.87%	22.50%
	NSN	25.20%	24.55%	25.43%	24.87%
	StegaStamp + S	**9.40%**	**9.05%**	**9.25%**	**9.77%**
	NSN + S	**9.77%**	**10.28%**	**10.00%**	**11.07%**
Konica Minolta C6500	StegaStamp	21.95%	23.05%	21.52%	22.00%
	NSN	23.25%	22.30%	23.10%	22.82%
	StegaStamp + S	**10.05%**	**10.77%**	**10.50%**	**10.47%**
	NSN + S	**11.42%**	**10.82%**	**11.30%**	**11.35%**
HP LaserJet M176n	StegaStamp	25.80%	26.02%	26.80%	25.02%
	NSN	25.35%	25.15%	24.90%	24.98%
	StegaStamp + S	**10.58%**	**11.10%**	**12.17%**	**11.27%**
	NSN + S	**11.23%**	**11.55%**	**11.55%**	**11.20%**

TABLE III
BIT ERROR RATE (BER) WITH DIFFERENT DATASETS. THE BEST PERFORMANCES ARE **BOLD-FACED**.

Devices	Methods	Datasets	
		Photo	Logo
EPSON L805	StegaStamp	26.12%	26.32%
	NSN	28.78%	29.10%
	StegaStamp + S	**15.15%**	**18.07%**
	NSN + S	**13.22%**	**16.20%**
HP LaserJet M176n	StegaStamp	22.62%	22.77%
	NSN	22.02%	22.32%
	StegaStamp + S	**13.92%**	**15.77%**
	NSN + S	**11.52%**	**13.27%**

sizes according to specific requirements. For example, there is limited available space for embedding information on a personal business card, while multiple watermarked images need to be included. Therefore, it is necessary to resize watermarked images according to the importance of information or minimize the size of watermarked images as much as possible. In this part, we evaluate the robustness performance of our method across different printing sizes. The watermarked images are printed with 2×2 cm^2, 1.5×1.5 cm^2, and 1×1 cm^2. "EPSON L805" and "HP LaserJet M176n" are utilized for printing, and "Oppo K9x" is utilized for shooting.

Experimental results are shown in Tab. IV. It can be seen that the proposed method can improve the watermark robustness across different printing sizes. Specifically, our method can maintain bit error rates below 20% and 30% in printing sizes of 1.5×1.5 cm^2, and 1×1 cm^2, respectively. This indicates that the proposed method is suitable for the scenario with small and multiple watermarks. Therefore, our method enables a single image to accommodate watermarked images with different sizes, which means more information can be embedded.

4) Robustness to Different Watermark Capacities: In this part, we evaluate the robustness performance of the proposed method to different watermark capacities. We select the watermark capacities of 50 bits, 100 bits, 150 bits, and 200 bits for retraining and testing the network. "EPSON L805" and "HP LaserJet M176n" are utilized for printing, and "Oppo K9x" is utilized for shooting.

Experimental results are shown in Tab. V. It can be observed that our method can achieve better performance across differ-

TABLE IV
BIT ERROR RATE (BER) WITH DIFFERENT PRINTING SIZES. THE BEST PERFORMANCES ARE **BOLD-FACED**.

Devices	Methods	Printing size (cm^2)		
		2 × 2	1.5 × 1.5	1 × 1
EPSON L805	StegaStamp	27.42%	36.35%	43.55%
	NSN	28.55%	35.58%	42.40%
	StegaStamp + S	**11.42%**	**16.15%**	**27.65%**
	NSN + S	**11.22%**	**15.57%**	**25.48%**
HP LaserJet M176n	StegaStamp	26.02%	31.47%	40.10%
	NSN	25.15%	31.31%	41.25%
	StegaStamp + S	**11.10%**	**14.95%**	**24.37%**
	NSN + S	**11.55%**	**15.15%**	**22.85%**

TABLE V
BIT ERROR RATE (BER) WITH DIFFERENT WATERMARK CAPACITIES. THE BEST PERFORMANCES ARE **BOLD-FACED**.

Devices	Methods	Capacity (bits)			
		50	100	150	200
EPSON L805	StegaStamp	14.50%	20.30%	25.97%	27.42%
	NSN	13.50%	20.85%	22.43%	28.60%
	StegaStamp + S	**4.30%**	**11.15%**	**11.20%**	**11.42%**
	NSN + S	**2.90%**	**3.75%**	**4.57%**	**11.22%**
HP LaserJet M176n	StegaStamp	12.90%	19.40%	23.43%	26.02%
	NSN	11.50%	19.00%	20.97%	25.15%
	StegaStamp + S	**2.50%**	**8.80%**	**11.17%**	**11.10%**
	NSN + S	**3.20%**	**3.85%**	**6.27%**	**11.55%**

ent watermark capacities. Specifically, the proposed method can decrease the bit error to below 5% when the watermark capacity is 50 bits. Therefore, the watermark capacity of 50 bits can be selected when the watermark is required to be completely recovered as possible.

IV. CONCLUSION

This work proposes robust watermark against halftone distortion, utilizing surrogate models and end-to-end watermarking frameworks. The experimental results confirm that our method can achieve good robustness performance under many practical scenarios, including different devices, datasets, and printing sizes.

In the future, we plan to extend our research to robust watermark in the screen-camera scenario. By modeling distortions of the screen-camera process, especially the moiré pattern, and training the corresponding surrogate model, we may address the challenge of digital watermarking encountered in the screen-camera scenario.

ACKNOWLEDGMENT

This work is sponsored by NSFC, China 62072313, 62371301, and Guangdong Provincial Key Laboratory (Grant 2023B1212060076).

REFERENCES

[1] L. Zhao, C. Chen, and J. Huang, "Deep learning-based forgery attack on document images," *IEEE Transactions on Image Processing*, vol. 30, pp. 7964–7979, 2021.

[2] M. Tancik, B. Mildenhall, and R. Ng, "Stegastamp: Invisible hyperlinks in physical photographs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2117–2126.

[3] J. M. Guo and S. Sankarasrinivasan, "Digital halftone database (dhd): A comprehensive analysis on halftone types," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2018, pp. 1091–1099.

[4] R. W. Floyd, "An adaptive algorithm for spatial gray-scale," in *Proc. Soc. Inf. Disp.*, vol. 17, 1976, pp. 75–77.

[5] C. Chen, X. Li, B. Chen, and H. Li, "A distortion model guided adversarial surrogate for recaptured document detection," *Pattern Recognition*, p. 110 433, 2024.

[6] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[7] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, PMLR, 2017, pp. 214–223.

[8] R. Shin and D. Song, "Jpeg-resistant adversarial images," in *NIPS 2017 Workshop on Machine Learning and Computer Security*, vol. 1, 2017, p. 8.

[9] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 39–43.

[10] C. Qin, X. Li, Z. Zhang, F. Li, X. Zhang, and G. Feng, "Print-camera resistant image watermarking with deep noise simulation and constrained learning," *IEEE Transactions on Multimedia*, 2023.

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[13] C. Chen, W. Chen, X. Chen, and H. Li, "A document presentation attack detection scheme with optical flow under a flashlight," *Pattern Recognition, Submitted*, 2023.

[14] J. Wang, W. Min, S. Hou, *et al.*, "Logo-2k+: A large-scale logo dataset for scalable logo classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 6194–6201.

SRC-gAudio: Sampling-Rate-Controlled Audio Generation

Chenxing Li*, Manjie Xu* and Dong Yu†
* Tencent AI Lab, Beijing, China
† Tencent AI Lab, Bellevue, WA, USA
E-mail: lichenxing007@gmail.com

***Abstract*—We introduce SRC-gAudio, a novel audio generation model designed to facilitate text-to-audio generation across a wide range of sampling rates within a single model architecture. SRC-gAudio incorporates the sampling rate as part of the generation condition to guide the diffusion-based audio generation process. Our model enables the generation of audio at multiple sampling rates with a single unified model. Furthermore, we explore the potential benefits of large-scale, low-sampling-rate data in enhancing the generation quality of high-sampling-rate audio. Through extensive experiments, we demonstrate that SRC-gAudio effectively generates audio under controlled sampling rates. Additionally, our results indicate that pre-training on low-sampling-rate data can lead to significant improvements in audio quality across various metrics.**

I. Introduction

The advent of text-to-audio (TTA) generation has marked a significant milestone in the realm of multimedia content creation, offering a novel paradigm where textual descriptions are used to generate audio content that aligns with the semantic intent of the text. This technology holds immense potential for enhancing the auditory experience in applications such as audio novels and video dubbing, where the congruence between audio and textual content is paramount. Despite its promise, the field of TTA generation is fraught with challenges, particularly in the fidelity, quality, and versatility of generated audio.

Current methodologies in text-driven audio generation predominantly bifurcate into two streams: Autoregressive (AR) and Non-Autoregressive (NAR) approaches. Audiogen [1], [2] and Uniaudio [3] are based on the AR transformer-decoder language model, which condition textual inputs to predict discrete audio tokens step-by-step. For NAR methods, diffusion-based and flow-based techniques have shown promise in generating high-fidelity audio. Diffsound [4] applies diffusion probabilistic models to predict mel-spectrogram tokens. Spectrorgam decoder and vocoder gradually transform tokens into waveform. Notably, diffusion models such as AudioLDM [5], AudioLDM2 [6], Tango [7], Make-an-audio [8], and Make-an-audio2 [9], leverage latent variable generation coupled with pre-trained variational auto-encoder (VAE)s [10] and HiFi-GAN [11] for audio reconstruction, achieving notable successes. Stable Audio [12], [13] directly encodes waveform into latent features and estimates noise updated from U-Net-based [14] to DiT-based [15] diffusion model. This solution can simplify the generation process and reduce error accumulation. It also supports longer audio generation. Similarly, flow-based methods, exemplified by Audiobox [16], employ continuous transformations from simpler to complex data distributions, offering an alternative pathway for audio synthesis. Besides, MAGNet [17] adopts a NAR transformer to perform generation directly over several streams of audio tokens. The Hybrid-MAGNet fuses AR and NAR models, which generates the beginning of the tokens in an AR manner while the rest of the sequence is being decoded in parallel. However, there is a performance gap between MAGNet and the methods above in the TTA task.

Despite these advancements, both methodologies encounter inherent limitations, particularly in handling diverse sampling rates and maintaining audio quality. (1) AR-based Audiogen [1] often suffers from spectral inconsistencies and detail loss due to the not completely accurate prediction of audio tokens. Also, the pre-trained sampling rate specified EnCodec [18] limits the sampling rate of audio generated by Audiogen. (2) The current NAR methods have guaranteed generation performance, but they all generate audio at a sampling rate of 16 kHz, which sometimes can not meet the high-resolution requirements. Under the popular diffusion-based pipeline, like AudioLDM2 [6], modeling different sampling rates requires training VAE, U-Net, and vocoder separately, which will increase the complexity of training and deployment. These limitations not only impact the listening experience but also underscore the need for a more flexible and robust framework capable of addressing the multifaceted challenges of audio generation.

An intuitive method for generating audio with varying sampling rates involves training a model at a high sampling rate, for instance, 48 kHz, and then down-sampling as per the needs of real-world applications. However, the scarcity of high-quality, high-sampling-rate annotated data often hampers adequate training for high-sampling-rate models. Moreover, under high-resolution conditions, the model may struggle to capture the wider range and increased variability of frequency features. Consequently, such models often underperform in practical scenarios.

This situation underscores the need for a versatile approach to overcome these limitations. Recognizing these challenges, we propose the SRC-gAudio, a multi-sampling-rate controlled audio generation model that uses the sampling rate as a condition to control generation within a single model. More-

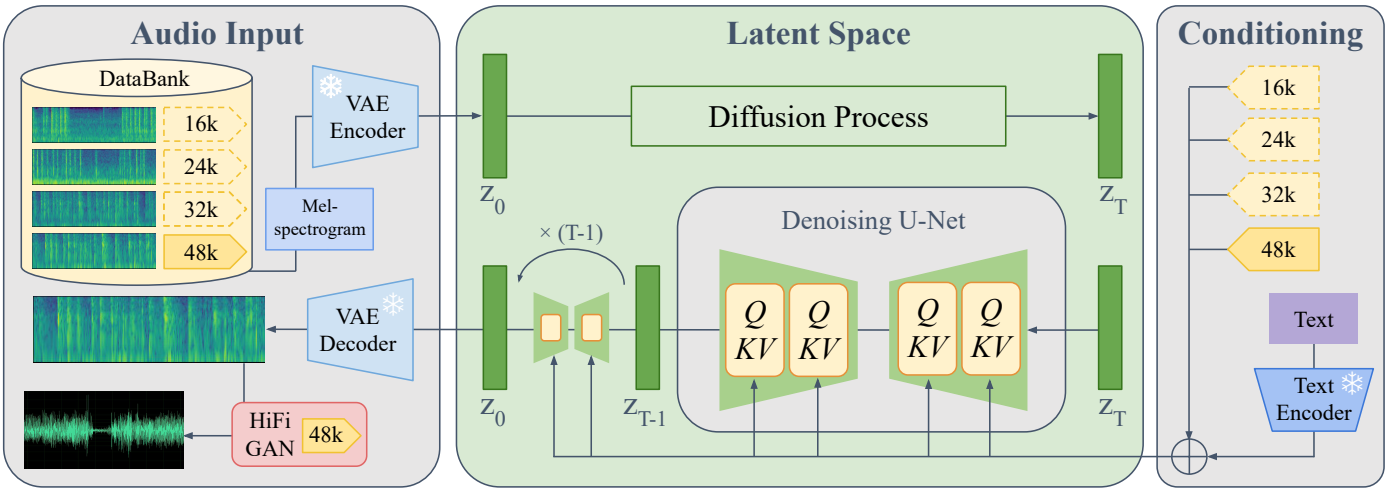


Fig. 1: **The overview of the proposed Sample-Rate Controlled Audio Generation (SRC-gAudio).**

over, SRC-gAudio leverages the strengths of low-sampling-rate pre-trained models to enhance generation quality at higher sampling rates. In detail, SRC-gAudio comprises the following components:

1) Diffusion-based SRC-gAudio fuses the sampling rate and text prompt as a condition to control audio generation. Through joint training under multi-sampling-rate conditions, SRC-gAudio shares the pre-trained text encoder, VAE, and trainable diffusion model. There is no need to train these models separately for different sampling rates, reducing the complexity of the overall system. The HiFi-GAN-based vocoder still needs to be trained separately.
2) Due to the scarcity of high-resolution data and the increased difficulty of training high-resolution models, we use low-sampling-rate data to help train high-sampling-rate models. We first employ large-scale low-sampling-rate data to train the model and then conduct multi-sampling-rate training based on the pre-trained model. A model trained with low-sampling-rate data can help the model reach a good initial point, and large-scale data can assist the model in achieving better generalization performance.

Our model offers a versatile solution that adapts to various sampling rates without compromising audio fidelity. Through rigorous experimentation, based on objective and subjective evaluations, SRC-gAudio demonstrates significant improvements, paving the way for a new era in sampling-rate-controlled audio generation technology.

II. System Overview

A. Conditional LDM-based audio generation

The Latent Diffusion Model (LDM) has emerged as a promising approach for generating high-quality and diverse audio samples in TTA task [5], [7], [8]. Given a data point x_0 sampled from the real data, LDMs focus on the efficient, low-dimensional latent space, with a trained perceptual compression model mapping the input to a hidden continuous feature space Z. In TTA generation, $z \in Z$ is the latent representation of the mel-spectrogram of the audio x. We leverage a pre-trained audio VAE from AudioLDM [5] to help compress the mel-spectogram of an audio sample x into the latent space.

In the forward process, Gaussian noise is gradually added to the input data according to a variance schedule β:

$$q(z_t|z_{t-1}) := \mathcal{N}\left(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I\right). \quad (1)$$

In reverse diffusion, diffusion models iteratively refine a randomly sampled noise input from $z_t \sim \mathcal{N}(0, I)$ to z_0. There have been methods that incorporate guidance into the diffusion process in order to "guide" the generation. The guidance refers to conditioning a latent of the prior data distribution $p(z)$ with a condition c, i.e., the class label or an image/text embedding, resulting in $p(z|c)$. In our model, the generation guidance c is included in the reverse process through cross-attention, where K and V embeddings are replaced with the condition embedding. In existing TTA models like AudioLDM [5] and Tango [7], the condition is the text embedding of the textual prompt. In the SRC-gAudio, the model functions by also conditioning the generation process on the desired sampling rate. We concatenate the sampling rate label, denoted as Sr, with the text condition $\psi(P)$ to form a new condition $c = concat(\psi(P), Sr)$. To turn a diffusion model p_θ into a conditional diffusion model, we add conditioning information at each diffusion step:

$$p_\theta(z_{0:T}|\psi(P), Sr) = p_\theta(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1}|z_t, c). \quad (2)$$

With a given text description P, to perform sequential denoising, a network ϵ_θ is often trained to predict artificial

noise, following the objective:

$$\min_{\theta} \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,I),t\sim\text{Uniform}(1,T)} \|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2, \quad (3)$$

where the $\psi(P)$ is the text embedding of the description. We leverage a pre-trained FLAN-T5 model [19] as the text encoder to obtain the text embedding.

We use U-Net [14], [20] with a cross-attention component as the backbone model for noise estimation. In detail, the U-Net model is conditioned on both the time step, sampling rate, and text embedding. We map the time step and sampling rate into a one-dimensional embedding and then concatenate these with text embedding as conditioning information. This approach allows for the generation of high-quality audio across a wide range of resolutions, all within a unified framework. The generation process is guided by these conditioning variables, ensuring that the output is in line with the attributes of the desired sampling rate.

We also leverage classifier-free guidance [21] to guide the diffusion toward the target audio. To achieve that, let $\varnothing = \psi(\text{" "})$ be the null text embedding, we define $c_\varnothing = concat(\psi(\text{" "}), Sr)$ as the unconditional case, and thus the generation can be defined by:

$$\tilde{\epsilon}_\theta = w \cdot \epsilon_\theta(z_t, t, c) + (1 - w) \cdot \epsilon_\theta(z_t, t, c_\varnothing), \quad (4)$$

where w denotes the guidance scale.

B. The pipeline of SRC-gAudio

The schematic diagram of SRC-gAudio is shown in Figure 1, which contains modules including a mel-spectrogram extraction module, an audio VAE, a text encoder, a LDM, and sampling-rate-based vocoders.

In the training stage, the mel-spectrogram is first extracted from the audio, and then the audio VAE is used to convert the mel-spectrogram into audio latents. The U-Net is trained based on Equ.(3) under text description and sampling rate control. In the generation stage, the audio latent is first initialized from $\mathcal{N}(0, I)$. Through the diffusion process, under the guidance of the text prompt and sampling rate conditions, the latent of target audio is gradually generated. The decoder of VAE and HiFi-GAN vocoder restore audio latent to mel-spectrogram and audio waveform step by step.

We mainly train the LDM under the control of different sampling rates. We also train vocoders so that they can convert the mel-spectrogram of audios with different sampling rates to audio.

C. Pre-training on low-sampling-rate data

Compared with low-sampling-rate data, high-sampling-rate data contains more high-frequency details and greater energy differences among frequencies, which may barricade model convergence. High sampling rate data, such as 32 kHz or 48 kHz, is difficult to obtain. The limited size of high-sampling rate data further leads to insufficient training of high-sampling rate-based generation models. These may lead to problems such as poor generation quality and less similarity between the generated audio and text prompt.

TABLE I: The evaluation results of the SRC-gAudio in two training paradigms: sampling rate as the generation condition (top) or training separately on different sampling rates (below).

Models	Sample rate	FD↓	IS ↑	KL↓	FAD↓	CLAP↑
Ground-truth	-	-	11.24	-	-	0.501
SRC-gAudio	16k	26.63	7.12	1.34	1.93	0.505
	24k	35.15	6.77	1.44	2.07	0.611
	32k	45.40	6.85	1.52	2.09	0.592
	48k	56.79	4.99	1.65	5.90	0.569
gAudio	16k	28.76	7.49	1.74	2.70	0.402
gAudio	24k	35.91	6.70	1.60	2.66	0.427
gAudio	32k	44.57	6.08	1.81	3.53	0.540
gAudio	48k	56.87	4.88	1.80	9.95	0.549

SRC-gAudio aims to generate audio with different sampling rates. In practice, the amount of data at different sampling rates is different. The amount of data in 16 kHz is much larger than the amount of data in 32 kHz or 48 kHz. We intuitively consider using a large amount of low-sampling-rate data to help generate high-sampling-rate audio. Specifically, we use low-sampling-rate data to pre-train a model with a fixed sampling rate. After the model training is completed, SRC-gAudio model training is continued based on the pre-trained model.

III. Experimental setup

A. Dataset

In this study, we primarily utilize the Audiocaps dataset [22] as the foundation for training. We obtain the audio data at a 48 kHz sample rate, which consists of 41,597 files[1]. The dataset comprises 40259 clips (120 hours in total) in the training set, 381 clips (1 hour in total) in the validation set, and 957 clips (3 hours in total) in the test set. During the model pre-training phase, we collect data from WavCaps [23], VGGSound [24], and ESC [25], amounting to approximately 4,000 hours of audio with paired captions. The training dataset for the 16k, 24k, and 32khz-based HiFi-GAN vocoder is the same as the pre-train dataset. For the 48khz-based vocoder, the train data is selected from freesound[2], which is extracted from WavCaps. In the test set, the caption for each clip is consistent with Tango [7] and AudioLDM [5]. This consistency ensures a fair comparison with their respective works.

B. Model architecture

We use the pre-trained audio VAE from AudioLDM [5] and the frozen FLAN-T5-LARGE [19] text encoder during training. For mel-spectogram extraction, we perform feature extraction in the (fftSize, hopSize, melDim) format for 16k, 24k, 32k, and 48khz-based configurations as follows: (1024, 160, 64), $(2048, 240, 64)$, $(2048, 320, 64)$, and $(2048, 480, 64)$, respectively. The 16k, 24k, 32k, and 48khz-based HiFi-GAN

[1]A few audio clips are excluded due to broken download links.
[2]https://freesound.org/

TABLE II: The evaluation results of the pre-train-based SRC-gAudio and the comparisons with other baseline methods. pre-AC16k+ft-AC and pre-Full16k+ft-AC indicate SRC-gAudio first pra-trains on Audiocaps or full pre-training data with fixed 16 kHz sampling rate and then fine-tunes on Audiocaps with sampling-rate condition, respectively. For baselines, 48 kHz-based gAudio-FT first pre-trains on 48 kHz upsampled full pre-training data and fine-tunes on 48 kHz-based Audiocaps. FT means the model is fine-tuned on the Audiocaps dataset. AC and AS refer to the Audiocaps and Audioset [26] dataset.

Models	Dataset	Params	Sample rate	Objective metrics					Subjective metrics		
				FD ↓	IS ↑	KL↓	FAD↓	CLAP↑	OGL ↑	RL ↑	AQ ↑
Ground-truth	-	-	-	-	11.24	-	-	0.501	93.72	93.64	94.45
SRC-gAudio-FT	pre-AC16k+ft-AC	561 M	16k	23.45	8.14	1.31	1.87	0.534	92.06	92.33	93.57
			24k	30.03	7.85	1.32	2.44	0.627	92.16	91.62	94.02
			32k	38.70	7.75	1.43	2.24	0.618	92.93	92.43	93.09
			48k	53.29	5.37	1.58	4.64	0.578	92.34	91.26	93.02
SRC-gAudio-FT	pre-Full16k+ft-AC	561 M	16k	20.63	8.91	1.21	2.10	0.529	92.81	92.56	93.56
			24k	26.83	8.50	1.30	3.00	0.633	93.58	92.74	94.36
			32k	35.80	9.15	1.38	2.62	0.625	92.59	92.21	94
			48k	49.67	5.83	1.51	4.51	0.568	91.26	93.34	92.75
gAudio-FT	pre-Full48k+ft-AC	561 M	48k	51.96	3.45	2.29	4.18	0.533	89.14	90.64	90.58
Tango	AC	866 M	16k	25.50	7.04	1.35	1.82	0.491	92.2	91.78	94.3
Tango-Full-FT	AS+AC+6 others		16k	18.43	8.01	1.16	2.77	0.554	92.89	92.46	94.05
AudioLDM-L-Full	AS+AC+2 others	739 M	16k	30.90	7.59	1.64	4.61	0.427	90.27	89.04	91.17
AudioLDM-L-Full-FT	AS+AC+2 others		16k	23.31	8.13	1.59	1.96	-	-	-	-
AudioLDM2-Full	AS+AC+6 others	346 M	16k	25.75	8.28	1.58	3.39	0.435	92.1	89.84	92.27
AudioLDM2-48k	-	262 M	48k	62.75	5.91	2.19	4.75	0.526	89.65	88.54	91.39

vocoders adopt the same structure as the vocoder in AudioLDM [5] but are trained separately with different sample rates.

We adopt the U-Net backbone of StableDiffusion [20] as the basic architecture of LDM of SRC-gAudio. The U-Net backbone we use has four encoder blocks, a middle block, and four decoder blocks. The channel dimensions of encoder blocks are $[320, 640, 640, 1280]$. The channel dimensions of decoder blocks are the reverse of encoder blocks. We add a cross-attention block in the last three encoder blocks and the first three decoder blocks, in which the number of heads is $[5, 10, 10, 20]$. The LDM of SRC-gAudio occupies 561M parameters.

We train the model for 40 epochs and report results for the checkpoint with the best validation loss. In sampling, we employ the DDIM [27] sampler with 200 sampling steps. For classifier-free guidance, a guidance scale w of 3.0 is used.

C. Evaluation metric

1) Objective metric: We employed various metrics, including Fréchet distance (FD), Kullback–Leibler (KL), Inception Score (IS), Fréchet audio distance (FAD), and contrastive language-audio pretraining (CLAP)[3] [28] to evaluate its performance. FD and FAD are used to measure the similarity distance between the generated audio and the ground truth audio, while KL calculates the Kullback–Leible divergence between them. Lower values for these metrics indicate better performance. On the other hand, IS is employed to measure the diversity and quality of the generated audio. CLAP evaluates the similarity between the generated audio and the text description.

2) Subjective metric: For subjective evaluation, we recruit 8 human participants to conduct a rating process. Following a similar approach to [4], the generated samples are assessed based on the overall generation quality (OGL), relevance to the input text (REL), and audio quality (AQ) using a scale of 1 to 100. Specifically, OGL examines the overall impact of the generated sound effect, encompassing factors such as semantic consistency between audio and text, and audio quality. REL measures the completeness and sequential consistency between the generated audio and the text prompt. AQ evaluates the quality of the generated audio, including aspects like audio clarity and intelligibility. We randomly select 100 test audio samples from the AudioCaps test set. Each participant is asked to evaluate each audio sample.

D. Baselines

To validate the effectiveness of the proposed SRC-gAudio approach, we initially train the generation model, named gAudio, using data with fixed sampling rates. The model structure and training strategy of gAudio mirror those of SRC-gAudio. However, without sampling rate control, gAudio can only generate audio at its pre-determined sampling rate.

For model comparisons, we employ diffusion-based generative models, including Tango [7], AudioLDM [5], and AudioLDM2 [6]. Currently, these generation models mainly generate audio with a 16 kHz sampling rate, with only AudioLDM2 offering a 48 kHz version. We adopt the results from AudioLDM, denoted as AudioLDM-L-Full-FT. For Tango and Tango-Full-FT[4], AudioLDM-L-Full[5], AudioLDM2-Full, and

[3]630k-best checkpoint from https://github.com/LAION-AI/CLAP

[4]https://github.com/declare-lab/tango

[5]https://github.com/haoheliu/AudioLDM

AudioLDM2-48k[6], we evaluate the results using the models and code released by the authors. The baseline results may differ from those in the corresponding papers, which could be attributed to the randomness of the generation process. Furthermore, to assess only the effectiveness of the generation model, we do not include clap filtering [6], which also accounts for the discrepancies between the baseline results and those reported in the papers.

IV. Experiments results

A. Experimental results of SRC-gAudio

The evaluation results of the SRC-gAudio are presented in Table I. We train our model under two paradigms: using the sampling rate as the generation condition to do joint training (top portion of the table) and training separately for different sampling rates (bottom portion of the table). Generally, lower sampling rates yield better generation results in evaluation metrics, indicating that audios with higher sampling rates are more challenging to learn. The results demonstrate that the SRC-gAudio can achieve competitive generation outcomes when trained collectively across various sampling rates, as compared to training a single model specifically for each sampling rate. A key observation is that in the case of high sampling rates, such as 32 kHz and 48 kHz, the joint-training SRC-gAudio outperforms the model trained specifically for the sampling rate in most evaluation metrics. These results suggest that training on lower sampling rates, which are considered easier to learn, can help the model fit better in more difficult generation tasks.

B. Experimental results on pre-training

The outcomes of the SRC-gAudio with pre-training are displayed in Table II. Compared Table II with Table I, it's evident that pre-training on data with a low sampling rate significantly enhances the generation of audio for both sampling rates. Pre-training was conducted on the identically distributed dataset (Audiocaps-16k) and large-scale pre-training dataset (Full16k) followed by fine-tuning. Based on objective evaluation, both tests reveal that a SRC-gAudio pre-trained on 16 kHz data yields superior generation results, particularly at higher sampling rates such as 32 kHz and 48 kHz.

After pre-training on the upsampled large-scale pre-training dataset and fine-tuned on Audiocaps, 48 kHz-based gAudio-FT still performes worse than the corresponding SRC-gAudio-FT. This illustrates the difficulty of training high-sampling rate models using low-quality data and further demonstrates the effectiveness of our proposed pipeline.

Additionally, we contrast the generation outcomes of SRC-gAudio with other cutting-edge baseline models such as Tango, AudioLDM, and AudioLDM2. By making the sampling rate an essential aspect of the generation condition, from both the objective and subjective metrics, our model delivers competitive results across different sampling rates in comparison to state-of-the-art baselines. It is observed that SRC-gAudio surpasses AudioLDM2-48k in generating high-sampling-rate audio. Even in the absence of pre-training (as depicted in Table I), SRC-gAudio excels in generating 48 kHz audio. We emphasize that training on low sampling-rate data can be beneficial.

[6]https://github.com/haoheliu/AudioLDM2

V. Conclusion

In this study, we present SRC-gAudio, an audio generation model designed to accommodate multiple sampling rates within a single, unified model framework. SRC-gAudio is capable of producing audio outputs at various sampling rates, conditioning on the specific generation configurations provided. By incorporating the sampling rate and text prompt as conditions, SRC-gAudio is jointly trained under multi-sampling-rate conditions. Furthermore, SRC-gAudio leverages the strengths of low-sampling-rate pre-trained models to enhance the generation quality at higher sampling rates. Our approach offers a scalable solution that adapts to diverse sampling rates without compromising audio fidelity. Rigorous experimentation has demonstrated significant improvements in TTA generation, paving the way for advancements in sampling-rate-controlled audio generation technology.

References

[1] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "Audiogen: Textually guided audio generation," in *The Eleventh International Conference on Learning Representations*, 2022.

[2] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al., "Uniaudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023.

[4] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[5] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," *Proceedings of the International Conference on Machine Learning*, 2023.

[6] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *arXiv preprint arXiv:2308.05734*, 2023.

[7] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria, "Text-to-audio generation using instruction guided latent diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3590–3598.

[8] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," *arXiv preprint arXiv:2301.12661*, 2023.

[9] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao, "Make-an-audio 2: Temporal-enhanced text-to-audio generation," *arXiv preprint arXiv:2305.18474*, 2023.

[10] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[12] Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons, "Fast timing-conditioned latent audio diffusion," *arXiv preprint arXiv:2402.04825*, 2024.
[13] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons, "Long-form music generation with latent diffusion," *arXiv preprint arXiv:2404.10301*, 2024.
[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
[15] William Peebles and Saining Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.
[16] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al., "Audiobox: Unified audio generation with natural language prompts," *arXiv preprint arXiv:2312.15821*, 2023.
[17] Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "Masked audio generation using a single non-autoregressive transformer," 2024.
[18] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.
[19] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei, "Scaling instruction-finetuned language models," 2022.
[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.
[21] Jonathan Ho and Tim Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.
[22] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.
[23] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.
[24] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.
[25] Karol J Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.
[26] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.
[27] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2020.
[28] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

On the Importance of Time and Pitch Relativity for Transformer-Based Symbolic Music Generation

Tatsuro Inaba*, Kazuyoshi Yoshii*, and Eita Nakamura†
*Kyoto University, Japan
E-mail: inaba@sap.ist.i.kyoto-u.ac.jp, yoshii.kazuyoshi.3r@kyoto-u.ac.jp
†Kyushu University, Japan
E-mail: nakamura@inf.kyushu-u.ac.jp

***Abstract*—This paper describes experimental investigation of music representations to draw the full potential of the Transformer with the self-attention mechanism for symbolic music generation. To use sequence-to-sequence model like the Transformer originally proposed for natural language processing, one typically serializes a musical score into a sequence of event- or note-based tokens without concern for the impact on the quality of generated music. The semantic invariance of music with respect to the time and pitch shifts is attributed to the positional relativity of musical notes over the time-pitch plane in which beats and pitch classes are repeated at intervals of bars and octaves, respectively. We here hypothesize that the capability of the self-attention mechanism to learn the musically meaningful rhythm, melody, and harmony is limited because the relativity and cyclicity of time and pitch information are not explicitly represented in the token sequence. To solve this problem, we propose a cyclicity-aware relative time and pitch encoding unique to music for the attention mechanism. Comprehensive evaluation using the POP909 dataset demonstrated that the proposed Transformer works better with event- or note-based score tokenization[1].**

I. Introduction

The Transformer [1], a sequence-to-sequence learning model originally proposed for natural language processing (NLP), has revolutionized generative tasks in music information processing [2]. To use the Transformer for symbolic music generation, one needs to represent a musical score as a sequence of *tokens* in an invertible manner because the input and output of the Transformer must be token sequences. In the event-based score tokenization (e.g., MIDI [3] and REMI [4]), each token corresponds to a musical event representing the onset time, instrument, pitch, or duration of a musical note, where multiple events collectively represent a musical note. In the note-based tokenization (e.g., OctapleMIDI [5] and MuMIDI [6]), each token corresponds to a musical note, i.e., a tuple of the onset time, instrument, pitch, and duration of the note. Note that the event-based representation does not guarantee the consistency of token sequences, i.e., token sequences cannot always be inverted to musical scores due to the possibility of missing or overlapping note attributes.

In these representations, the relationships between musical notes over time and pitch are not explicitly considered (Fig. 1). The time relativity is crucial in forming the semantic invariance of music to time shifts. Similarly, the pitch relativity is another key feature of music because a musical piece remains semantically identical with respect to key transposition. The relative index, pitch, and onset (RIPO) attention mechanism [7] is a noticeable approach that considers the time and pitch relativity for sinusoidal encoding of note-based tokens.

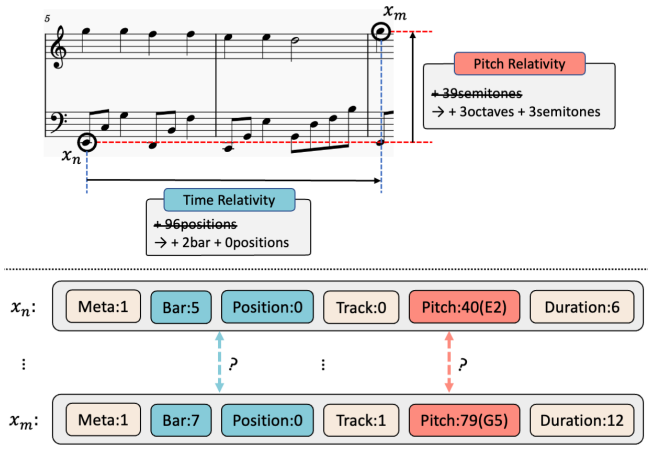


Fig. 1. A token x_n in the note-based representation holds the absolute pitch value of 40 (E2), and another token x_m holds the absolute pitch value of 79 (G5). On the other hand, the relative pitch distance 39 (3 octaves and 3 semitones) is not explicitly captured.

As a piece-agnostic nature of music, beat and pitch circulations play a vital role in music. The bar serves as a fundamental unit of time. Musical themes, for example, tend to be repeated at an interval of 1, 2, 4, or 8 bars. Similarly, the octave serves as a fundamental unit of pitch. Musical sounds whose pitches differ by octaves are perceived as having similar resonance because the harmonic structures of these sounds significantly overlap with each other. As illustrated in Fig. 1, it is musically meaningful to interpret 96 time steps as 2 bars and 39 semitones as 3 octaves plus 3 semitones (a minor 3rd). Appropriate treatment of these fundamental units would thus be considered as a key to achive musically-meaningful music generation compatible with standard music theory.

In this paper, we propose a variant of the attention mechanism named Circular Relative Attention (CirRelAttn) mechanism. Time relativity is decomposed into bar units and their remainders, and pitch relativity into octave units and their remainders. This method can effectively learn the repetitive structures at specific intervals and generates music with en-

[1]The demo page is available at https://tatsuropfgt.github.io/CirRelAttn.

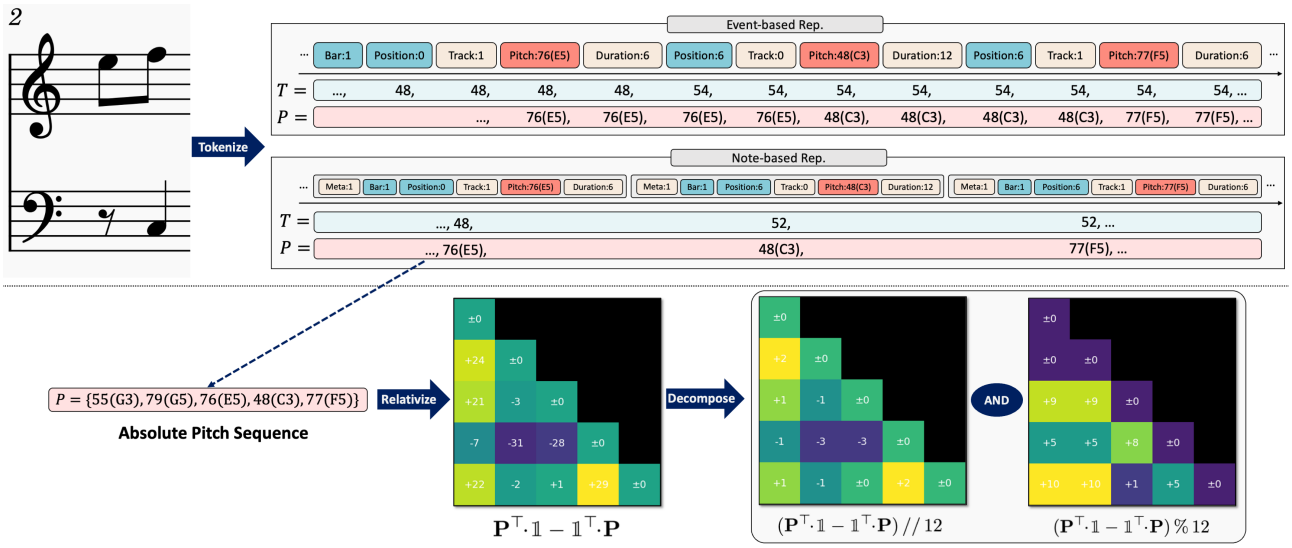


Fig. 2. The top figure shows an example of event-based and note-based representations, along with the absolute time $\mathbf{T}$ and the pitch sequence $\mathbf{P}$. The bottom figure illustrates decomposed relative pitch matrices, which are designed to emphasize the circularity of octaves

hanced coherence. We evaluated our approach using the POP-909 dataset[8], a collection of Chinese popular music rich in repetitive structures. Objective evaluation results showed that our method continually generates musical pieces closer to test data than previous approaches. Additionally, subjective evaluations revealed that human listeners favored the pieces produced by our method.

II. Related Work

A. Music Representation

Music representations must be designed carefully to convert a symbolic score of polyhonic music into a one-dimensional sequence of *tokens* that sequence-to-sequence models can directly deal with [2]. This choise is crucial because it has a strong impact on the model ability to learn musical structures. The event-based representations [3], [4] represents musical events such as note-on, note-off (or duration), pitch as tokens. The compound (CP) word representation [9] aggregates musical events based on their specific roles into single tokens. This was extended to the note-based representation [5], [6], where each note is represented by a single token, enhancing the efficiency of sequence length.

B. Music Structure Modeling

A few approaches have been explored to model the time and pitch structure of music. Chuan and Herremans [10] proposed a method that captures musical relationships between pitches inspired by the Tonnetz method. They encoded these relationships into a 2D representation and utilized a convolutional neural network (CNN) and a long short-term memory (LSTM) network for generating musical scores. Music Transformer [11] claimed the importance of relative information in music and incorporated the index intervals between tokens into the self-attention mechanism. This approach was extended to jointly consider the time and pitch relativity of music with the RIPO attention mechanism [7], which embeds the pitch and onset distances between notes into self-attention mechanism with sinusoidal encoding. Recently, Structure-PE [12] has incorporated the time, section, chord, and melodic pitch relationships between time steps into the self-attention mechanism.

III. Proposed Method

We define music representations and review the existing self-attention mechanisms (RelAttn and RIPO). We then propose the Circular Relative Attention (CirRelAttn).

A. Music Representations

We represent a musical score as a sequence of event-based (Section III-A1) or note-based (Section III-A2) tokens denoted by $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_L)$. For simplicity, we represent a musical score using only essential attributes obtained by omitting performance-related ones (e.g., tempo and velocity) and additional ones (e.g., chord). We also define a time resolution as the 12 steps per quarter note.

1) Event-based Representation: We use the events in Table I based on REMI+ [13] representation. The variable k associated with Bar, Position, Track, or Pitch tokens takes an integer within the range. In contrast, k associated with the Duration token takes one of 27 integers specified by the following expression as follows:

$$\begin{aligned} k = &\{1, ..., 12\} \\ &\cup \{12 + 3i \,|\, i \in [1, 4]\} \cup \{12 + 4i \,|\, i \in [1, 3]\} \\ &\cup \{24 + 6i \,|\, i \in [1, 4]\} \cup \{48 + 12i \,|\, i \in [1, 4]\}. \end{aligned} \quad (1)$$

Durations are represented as follows: up to quarter notes for time step lengths; up to half notes for sixteenth notes and triplets; up to whole notes for eighth notes; and up to double whole notes for quarter notes. Unlike REMI+, time-signature, tempo, chord, and velocity events are omitted. In this study, we use a total of $1 + 1 + 16 + 8 + 3 + 128 + 27 = 184$ types of tokens.

TABLE I
TOKEN LIST IN EVENT-BASED REPRESENTATION.

Event Type	Description	Value Range
BOS	Beginning of a song	
EOS	End of a song	
Bar:k	Start of the kth bar	$1 \leq k \leq 16$
Position:k	Position of the note in the bar, k	$0 \leq k \leq 47$
Track:k	Track number of the note, k	$1 \leq k \leq 3$
Pitch:k	Pitch of the note, k	$0 \leq k \leq 127$
Duration:k	Duration of the note, k	$1 \leq k \leq 96$

The top of Fig. 1 shows an example of the event-based representation. The token sequence begins with a BOS token and ends with an EOS token. Musical notes are sorted in an ascending order of time and pitch to form a series of four tokens: Position, Track, Pitch, and Duration, which are then added to the token sequence in order.

We define the absolute index sequence $\mathbf{I} = \{I_1, ..., I_L\}$, the absolute time sequence $\mathbf{T} = \{T_1, ..., T_L\}$, and the absolute pitch sequence $\mathbf{P} = \{P_1, ..., P_L\}$ in Algorithm 1, which are used to compute the relativity in Sections III-B, III-C, and III-D. *BarRes* represents the timesteps per bar.

2) Note-based Representation: This representation is defined as in the Multitrack Music Transformer [6], where each token (note) is expressed as a tuple of 6 variables:

$$\mathbf{x}_i = \{x_i^{\text{meta}}, x_i^{\text{bar}}, x_i^{\text{position}}, x_i^{\text{p}}, x_i^{\text{duration}}, x_i^{\text{track}}\}.$$

Each variable corresponds to Meta, Bar, Position, Track, Pitch, and Duration tokens. Except for Meta and Bar tokens, the definitions are the same as those in the event-based representation (see Section III-A1).

Meta token indicates the beginning of a song (0), a note (1), or the end of a song (2). Bar token represents the bar including a note, with a value range from 1 to 16, consistent with the event-based representation. During encoding, each variable is separately transformed by linear transformation layers, and their sum represents the latent representation of the token. Conversely, during decoding, the latent representation of the output is decoded into each variable separately by six types of linear transformation layers.

The top of Fig. 1 also shows an example of the note-based representation. After converting all musical notes into tokens, tokens are sorted based on time and pitch. A token representing the start of the piece (Meta:0) is added at the beginning, and a token representing the end of the piece (Meta:2) is added at the end of the sorted token sequence. We also define the absolute index, time, and pitch sequences for the note-based representation as follows.

$$\mathbf{I} = \{1, ..., L\}, \tag{2}$$

$$\mathbf{T} = \{x_i^{\text{bar}} \times \textit{BarRes} + x_i^{\text{position}}\}_{i=1}^{L}, \tag{3}$$

$$\mathbf{P} = \{x_i^{\text{pitch}}\}_{i=1}^{L}. \tag{4}$$

B. Relative Attention

In general, the self-attention layer (Attn) transforms a feature sequence $X \triangleq \{X_1, ..., X_L\} \in \mathbb{R}^{L \times D}$ to another feature sequence $Y \triangleq \{Y_1, \ldots, Y_L\} \in \mathbb{R}^{L \times D}$ while keeping the length L and the dimension D. Fist, X is transformed into queries $Q = XW_Q$, keys $K = XW_K$, and values $V = XW_V$, where $W_Q, W_K, W_V \in \mathbb{R}^{D \times D_h}$ are learnable parameters. Here, D_h represents the dimension in an attention head. The original transformer uses the multi-head attention mechanism for capturing the multifaceted relationships between different positions. For simplicity, we here focus on a single attention head. The output of the self-attention layer is given by

$$Y = \text{Softmax}\left(\frac{QK^\top}{\sqrt{D_h}}\right)V \triangleq \text{Attn}(X). \tag{5}$$

On the other hand, relative attention [11], [14] incorporates positional embeddings derived from relative index distances between tokens into the attention mechanism.

$$\mathbf{R}_{\text{idx}} = \text{LPE}(\mathbf{I}^\top \cdot \mathbb{1} - \mathbb{1}^\top \cdot \mathbf{I}) \in \mathbb{R}^{L \times L \times D_h}, \tag{6}$$

$$S_{\text{rel}}^{\text{idx}} = Q\mathbf{R}_{\text{idx}}^\top \in \mathbb{R}^{L \times L}, \tag{7}$$

$$Y = \text{Softmax}\left(\frac{QK^\top + \alpha S_{\text{rel}}^{\text{idx}}}{\sqrt{D_h}}\right)V \triangleq \text{RelAttn}(X), \tag{8}$$

where $\mathbb{1}$ is an all-one vector of length L and α is a hyperparameter used for scaling. The relative index matrix $\mathbf{I}^\top \cdot \mathbb{1} - \mathbb{1}^\top \cdot \mathbf{I}$ is embedded using learnable positional encoding LPE. Subsequently, the relative index logit $S_{\text{rel}}^{\text{idx}}$, computed by the product of the embedding $\mathbf{R}_{\text{idx}}$ and the query Q, is scaled by α and added to $QK^\top$. We refer to relative attention as RelAttn.

C. RIPO Attention

RIPO attention incorporates the relative time and pitch distances between notes in original music as well as the relative index distances between tokens.

$$\mathbf{R}_{\text{t}} = \text{SPE}(\mathbf{T}^\top \cdot \mathbb{1} - \mathbb{1}^\top \cdot \mathbf{T}) \in \mathbb{R}^{L \times L \times D_h}, \tag{9}$$

$$\mathbf{R}_{\text{p}} = \text{SPE}(\mathbf{P}^\top \cdot \mathbb{1} - \mathbb{1}^\top \cdot \mathbf{P}) \in \mathbb{R}^{L \times L \times D_h}, \tag{10}$$

Algorithm 1 Event-based $\mathbf{I}, \mathbf{T}, \mathbf{P}$

```
cbar ← None #current bar
cpos ← None #current position
cpit ← None #current pitch
I, T, P ← [], [], []
for idx, event ∈ enumerate(sequence) do
    if type of event is Bar then
        cbar ← Bar value of event
    end if
    if type of event is Position then
        cpos ← Position value of event
    end if
    if type of event is Pitch then
        cpit ← Pitch value of event
    end if
    I.append(idx)
    T.append(cbar × BarRes + cpos)
    P.append(cpit)
end for
```

$$S^{\mathrm{t}}_{\mathrm{rel}} = Q\mathbf{R}^{\top}_{\mathrm{t}} \in \mathbb{R}^{L\times L}, \tag{11}$$

$$S^{\mathrm{p}}_{\mathrm{rel}} = Q\mathbf{R}^{\top}_{\mathrm{p}} \in \mathbb{R}^{L\times L}, \tag{12}$$

$$S_{\mathrm{rel}} = S^{\mathrm{idx}}_{\mathrm{rel}} + S^{\mathrm{t}}_{\mathrm{rel}} + S^{\mathrm{p}}_{\mathrm{rel}} \in \mathbb{R}^{L\times L}, \tag{13}$$

$$Y = \mathrm{Softmax}\left(\frac{QK^{\top} + \alpha S_{\mathrm{rel}}}{\sqrt{D_h}}\right)V \triangleq \mathrm{RIPOAttn}(\mathbf{X}). \tag{14}$$

The relative time matrix $\mathbf{T}^{\top}\cdot\mathbb{1} - \mathbb{1}^{\top}\cdot\mathbf{T}$ and the relative pitch matrix $\mathbf{P}^{\top}\cdot\mathbb{1} - \mathbb{1}^{\top}\cdot\mathbf{P}$ are embedded with the sinusoidal positional encoding SPE. The embeddings $\mathbf{R}_{\mathrm{t}}$ and $\mathbf{R}_{\mathrm{p}}$ are then utilized to compute the relative time logit $S^{\mathrm{t}}_{\mathrm{rel}}$ and the relative pitch logit $S^{\mathrm{p}}_{\mathrm{rel}}$ by performing the matrix product with the query Q. These logits are subsequently scaled by α and added to $QK^{\top}$.

D. Circular Relative Attention

Leveraging the inherent cyclicity in time and pitch, the proposed CirRelAttn mechanism effectively models the repetitive and cyclic structures that weren't captured by RIPO attention. Initially, we decompose the relative time distance matrix into bar units and their remainders, and then separately embed these matrices using learnable positional encodings.

$$\mathbf{R}_{\mathrm{bar}} = \mathrm{LPE}((\mathbf{T}^{\top}\cdot\mathbb{1} - \mathbb{1}^{\top}\cdot\mathbf{T}) \,//\, \textit{BarRes}), \tag{15}$$

$$\mathbf{R}_{\mathrm{position}} = \mathrm{LPE}((\mathbf{T}^{\top}\cdot\mathbb{1} - \mathbb{1}^{\top}\cdot\mathbf{T}) \,\%\, \textit{BarRes}). \tag{16}$$

Similarly, we decompose relative pitch distance into octave units and their reminders and separately embed these matrices using learnable positional encoding.

$$\mathbf{R}_{\mathrm{octave}} = \mathrm{LPE}((\mathbf{P}^{\top}\cdot\mathbb{1} - \mathbb{1}^{\top}\cdot\mathbf{P}) \,//\, 12), \tag{17}$$

$$\mathbf{R}_{\mathrm{semitone}} = \mathrm{LPE}(\mathbf{P}^{\top}\cdot\mathbb{1} - \mathbb{1}^{\top}\cdot\mathbf{P}) \,\%\, 12). \tag{18}$$

Here, the reason for not using sinusoidal positional encoding is that proximity in relative distance does not necessarily imply a strong relationship. For example, as pointed out in [15], similar musical structures tend to appear in specific previous bars (1, 2, 4, 8, 16, ... bars). Moreover, the harmonic relationship between the pitches is intricate, resulting in each interval having a unique resonance. We utilize learnable positional encoding LPE to capture the features corresponding to these relative distances separately. The relative time and pitch logits from the aforementioned embeddings are represented by one of the following equations.

$$\begin{cases} S^{\mathrm{t}}_{\mathrm{rel}} = Q\left(\mathbf{R}_{\mathrm{bar}} + \mathbf{R}_{\mathrm{position}}\right)^{\top}, \\ S^{\mathrm{p}}_{\mathrm{rel}} = Q\left(\mathbf{R}_{\mathrm{octave}} + \mathbf{R}_{\mathrm{semitone}}\right)^{\top}, \end{cases} \tag{19}$$

$$\begin{cases} S^{\mathrm{t}}_{\mathrm{rel}} = Q\left(\mathbf{R}_{\mathrm{bar}} \odot \mathbf{R}_{\mathrm{position}}\right)^{\top}, \\ S^{\mathrm{p}}_{\mathrm{rel}} = Q\left(\mathbf{R}_{\mathrm{octave}} \odot \mathbf{R}_{\mathrm{semitone}}\right)^{\top}. \end{cases} \tag{20}$$

Here, we define equation (19) as CirRelAttn-S and equation (20) as CirRelAttn-H.

IV. Evaluation

In this section, we describe the objective and subjective evaluation results from the continuous generation experiments and analyze the effectiveness of the proposed method.

A. Experimental Setup

We used POP909 dataset [8], which contains popular music with a wide variety of repetitive structures. We excluded songs with inconsistent beat annotations and split the available 896 songs into 10% for validation, 10% for testing, and the remaining 80% for training. We extracted sequences of 16 measures from each song with a stride of 1 measure and used only sequences with 4/4 time signature. Finally, we obtained 4,749 sequences for validation, 4,137 for testing, and 35,452 for training. We quantized the data with a resolution of 12 steps per quarter note. We used MusPy [16] for data pre-processing. For training, we performed data augmentation by randomly transposing each data between -6 to +5 semitones.

For the proposed method, we used a decoder-only Transformer based on a self-attention mechanism with 4 layers, 8 heads, a hidden dimension of 256, and a dropout rate of 0.2. For comparison, we tested the vanilla attention mechanism (Attn) [1], the relative attention mechanism (RelAttn) [11], and the RIPO attention mechanism (RIPOAttn) [7], where the same architecture was used as the proposed method. We conducted validation every 1,000 steps and terminated training after 200,000 steps or if no improvement in loss was observed after 20 validation checks. We fixed the batch size at 8, with a warm-up step count of 10,000 and a peak learning rate of 2e-5. We set the hyperparameter to $\alpha = 0.1$.

B. Objective Evaluation

We evaluated how accurately the model can continuously generate musical pieces that resemble the test data. The model was given the first 15 bars of each test dataset and generated a continuation for 1 bar. We calculated the similarity between the generated musical pieces and the ground truth using the following five objective metrics.

- NoteF1 ($F1_{\mathrm{note}}$) or OnsetF1 [17] measures the accuracy of the generated pieces against the ground truth on a per-note basis using the F1 score. A true positive is identified when the onset timing, pitch, and track of the generated note all match the note in the ground truth.
- PianorollF1 ($F1_{pr}$) or FrameF1 [17] measures the accuracy of the generated pieces against the ground truth on the piano roll representation by computing F1 score by evaluating each time step and pitch class.
- Grooving Similarity (GS) [18], [19] is the cosine similarity between the grooving vectors of the generated music and the ground truth. Each grooving vector represents the number of onsets at each time step within a measure and is a 48-dimensional vector in this experiment.
- Chroma Similarity (CS) [18] is the cosine similarity between the chroma vectors [20] of the generated pieces and the ground truth. The chroma vector [20] is a 12-dimensional vector representing the number of onsets for each of the 12 pitch classes (C, C#, ..., B) within a specific time range. In this experiment, the cosine similarity is calculated for each half-bar (24 time steps) and the average is computed.

TABLE II
OBJECTIVE EVALUATION RESULTS.

No.	Rep.	Methods: Attn Type	params	time (ms/note)	Objective Evaluation: Loss	$F1_{note}$	$F1_{pr}$	GS	CS	PRS
1		Attn [1]	4.32M	8.46	0.979	0.174	0.239	0.846	0.620	0.955
2		RelAttn [11]	4.84M	9.73	0.937	0.218	0.291	0.848	0.650	0.955
3	event	RIPOAttn [7]	4.85M	18.3	0.904	0.233	0.305	0.855	0.660	0.956
4		CirRelAttn-S	4.88M	22.5	0.876	0.268	0.341	0.855	0.673	0.957
5		CirRelAttn-H	4.88M	20.8	**0.856**	**0.293**	**0.361**	**0.858**	**0.685**	**0.958**
6		Attn [1]	3.53M	4.29	4.42	0.188	0.267	0.784	0.619	0.941
7		RelAttn [11]	3.66M	5.79	4.38	0.186	0.271	**0.798**	0.628	**0.946**
8	note	RIPOAttn [7]	3.67M	5.93	4.40	0.180	0.257	0.786	0.612	0.942
9		CirRelAttn-S	3.70M	8.58	4.37	0.192	0.273	0.785	0.623	0.943
10		CirRelAttn-H	3.70M	6.54	**4.29**	**0.215**	**0.294**	0.787	**0.632**	0.944

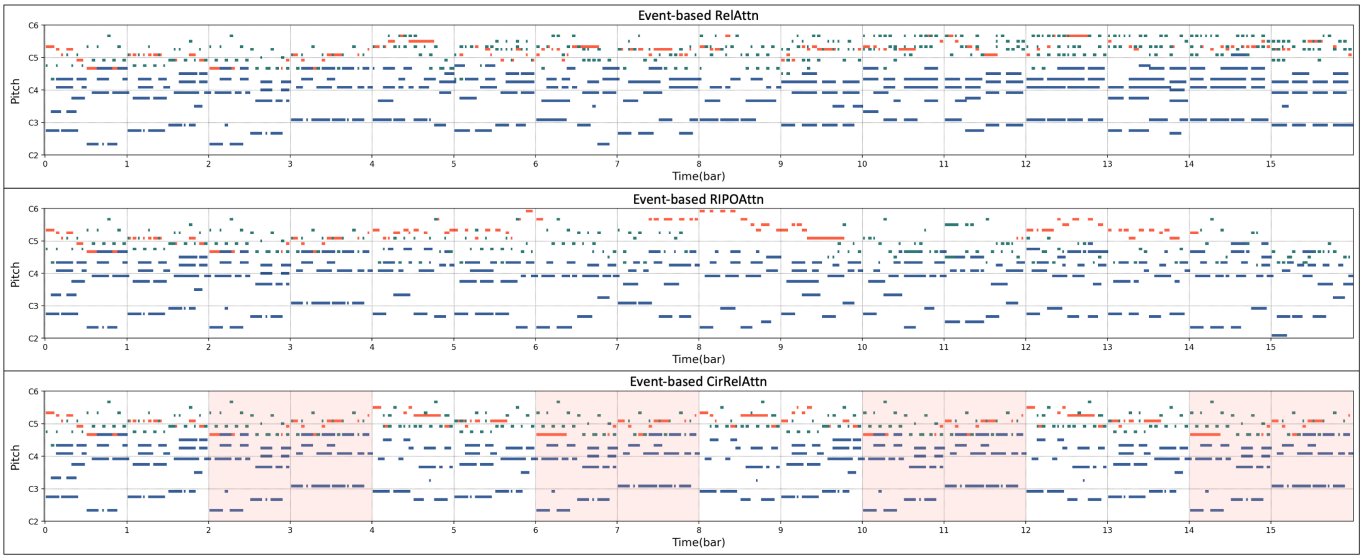


Fig. 3. Musical pieces generated by the event-based RelAttn, RIPOAttn, and CirRelAttn-H. Given the first four bars, each model generates a continuation for the next twelve bars. While the RelAttn and RIPOAttn generate entirely new phrases, CirRelAttn generates phrases that resemble the initial input, resulting in more coherent music. The areas shaded in red indicate a repetitive structure.

TABLE III
SUBJECTIVE EVALUATION RESULTS.

Methods: Rep.	Attn Type	Subjective Evaluation: Coherence	Musicality	Overall
Ground Truth		3.93	4.17	4.03
event	RelAttn [11]	2.93	2.69	2.79
	RIPOAttn [7]	3.00	3.21	3.03
	CirRelAttn-H	**4.31**	**3.41**	**3.69**
note	RelAttn [11]	2.28	2.69	2.45
	RIPOAttn [7]	**2.69**	**2.90**	**2.90**
	CirRelAttn-H	2.10	2.51	2.42

- Pitch Range Similarity (PRS) measures the similarity of the pitch range within a single bar between the generated pieces and the ground truth.

$$\text{PRS} = 1 - \frac{|\text{PR}_{\text{gen}} - \text{PR}_{\text{gt}}|}{128}, \tag{21}$$

where PR_{gen} and PR_{gt} represent the pitch range (the difference between the highest and lowest pitches) of the generated pieces and the ground truth, respectively.

Table II shows the objective evaluation results on the continuous generation task. Our methods (No. 5, 10) demonstrated superior performance to the baselines on most objective metrics within both event-based and note-based representations. In other words, we confirmed that the proposed methods generate musical pieces that more closely resemble the test data compared to the baselines.

C. Subjective Evaluation

We conducted a listening test to evaluate the quality of musical pieces generated by our method. We compared the generated music pieces by six models (event-based RelAttn, RIPOAttn, and CirRelAttn and note-based RelAttn, RIPOAttn, and CirRelAttn) and the ground truth. We provided only the initial four bars to each model, and the model then generated the continuation for twelve bars. Each evaluator listened to seven pieces and scored them based on three criteria: coherence, musicality, and overall quality, using a scale from 1 to 5. We recruited 30 evaluators to evaluate 10 sets of musical pieces, with each set scored by three evaluators.

Table III shows the result of the listening test. Our method with the event-based representation (CirRelAttn-H) outperformed other methods, particularly in coherence. We found

that CirRelAttn-H with the event-based representation tended to generate repetitive structures by analyzing the generated musical pieces. Fig. 3 shows an example of musical pieces generated by RelAttn, RIPOAttn, and CirRelAttn. While RelAttn and RIPOAttn constantly generated new phrases, CirRelAttn repeatedly generated the phrase given in the first four bars. We hypothesized that music containing a high amount of repetitive structures received particularly high ratings for coherence from human evaluators. The proposed attention mechanism enabled the model to learn the distinctive repetitive structure of music well and generate consistent music.

V. Conclusion

This paper presented a novel approach that decomposes the relative time and pitch distances based on bar and octave units and subsequently integrates these components into the self-attention mechanism via learnable positional encodings. Objective evaluation demonstrated that our method effectively learns music structures and outperforms previous methods in a music continuation task. We also confirmed that our method generates coherent music that contains repetitive structures through a listening test.

Acknowledgment

This work was partially supported by JST FOREST Nos. JPMJFR2270 and JPMJFR226X, JST PRESTO JPMJPR20CB, and JSPS KAKENHI Nos. 24H00742, 24H00748, and 22H03661.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Neural Information Processing Systems*, 2017, pp. 5998–6008.

[2] S. Ji, X. Yang, and J. Luo, "A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges," *ACM Comput. Surv.*, vol. 56, no. 1, 2023.

[3] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, "This time with feeling: Learning expressive musical performance," *Neural Computing and Applications*, vol. 32, pp. 955–967, 2018.

[4] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *ACM International Conference on Multimedia*, 2020, pp. 1180–1188.

[5] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T.-Y. Liu, "MusicBERT: Symbolic music understanding with large-scale pre-training," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 791–800.

[6] H.-W. Dong, K. Chen, S. Dubnov, J. McAuley, and T. Berg-Kirkpatrick, "Multitrack music transformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[7] Z. Guo, J. Kang, and D. Herremans, "A domain-knowledge-inspired music embedding space and a novel attention mechanism for symbolic music modeling," in *AAAI Conference on Artificial Intelligence*, 2023.

[8] Z. Wang, K. Chen, J. Jiang, *et al.*, "Pop909: A pop-song dataset for music arrangement generation," in *International Society for Music Information Retrieval Conference ISMIR*, 2020.

[9] W.-Y. Hsiao, J.-Y. Liu, Y.-C. Yeh, and Y.-H. Yang, "Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs," in *AAAI Conference on Artificial Intelligence*, 2021.

[10] C.-H. Chuan and D. Herremans, "Modeling temporal tonal relations in polyphonic music through deep networks with a novel image-based representation," *AAAI Conference on Artificial Intelligence*, 2018.

[11] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, *et al.*, "Music transformer," in *International Conference on Learning Representations ICLR*, 2019.

[12] M. Agarwal, C. Wang, and G. Richard, "Structure-informed positional encoding for music generation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[13] D. von Rütte, L. Biggio, Y. Kilcher, and T. Hofmann, "FIGARO: Controllable music generation using learned and expert features," in *International Conference on Learning Representations ICLR*, 2023.

[14] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 464–468.

[15] B. Yu, P. Lu, R. Wang, *et al.*, "Museformer: Transformer with fine- and coarse-grained attention for music generation," in *Neural Information Processing Systems*, 2022.

[16] H.-W. Dong, K. Chen, J. McAuley, and T. Berg-Kirkpatrick, "Muspy: A toolkit for symbolic music generation," in *International Society for Music Information Retrieval Conference ISMIR*, 2020.

[17] J. P. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel, "MT3: Multi-task multitrack music transcription," in *International Conference on Learning Representations ICLR*, 2022.

[18] S.-L. Wu and Y.-H. Yang, "MuseMorphose: Full-song and fine-grained piano music style transfer with one Transformer VAE," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[19] S. Dixon, F. Gouyon, and G. Widmer, "Towards characterisation of music via rhythmic patterns," in *International Society for Music Information Retrieval Conference ISMIR*, 2004.

[20] T. Fujishima, "Realtime chord recognition of musical sound : A system using common lisp music," *International Computer Music Conference ICMC*, pp. 464–467, 1999.

AFSDet: Video Small Object Detection Based on Adaptive Focused Slicing

Kangjian Huang†, Yan Yang*, Yongquan Jiang‡, Xiaobo Zhang§ and Zhuyi Angelina Li¶
* School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China
Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu, China
E-mail: yyang@swjtu.edu.cn
† School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China
Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu, China
‡ School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China
Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu, China
§ School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China
¶ Business School, Renmin University of China, Beijing, China

***Abstract*—The detection of video small objects is challenging in images due to their limited presence, which makes the information more susceptible to loss. In this paper, we propose an Adaptive Focused Slicing Detection (AFSDet) method inspired by the way humans observe small objects. AFSDet involves first localizing to a region with potential targets and then scrutinizing the small objects within that region. It is a two-stage approach utilizing the proposed model called Deformable Head YOLO Network (DHYNet) for object detection. The first stage performs a coarse-detection of the original image and utilizes clustering and Adaptive Scale-up Slicing methods to obtain patches, leading to a significant reduction in the number of patches compared to previous slicing-based detection methods.The second stage feeds each patches into the fine-detection model and passes the results to the next frame as a reference to assist coarse-detection. Additionally, taking advantage of the slight variations between consecutive video frames, a superimposing process is employed for coarse-detection, which also contributes to reducing the computational overhead. The results show that the proposed method, tested on the VisDrone-MOT dataset, improves accuracy by 9.6% compared to the baseline model, and improves the accuracy by 3.3 % compared to the previous slicing-based small object detection method.**

I. Introduction

The video small object detection holds a wide array of applications, including people and vehicle recognition and tracking in surveillance videos [1]. While significant research has been conducted on object detection at the image level, video object detection specifically targets objects across continuously captured frames. Video object detection encounters challenges such as motion blur, occlusion, and rare poses, distinct from typical image data [2]. Moreover, consecutive video frames exhibit correlation, enabling the exchange and supplementation of information between them, thereby enhancing detection accuracy. Additionally, these frames demonstrate redundancy, with minimal differences between neighboring frames, which can be leveraged to reduce unnecessary computations and improve detection efficiency. Overall, video-level object detection has new challenges than image-level object detection, but utilizing the characteristics of video can help the model achieve better detection results.

Small objects refer to objects that are small relative to the size of the image or the type of the object, and the definition may vary in different application scenarios. Video small object detection has many challenges [3], such as : (1) Susceptibility of information loss. (2) Vulnerability to noise interference. (3) Limited bounding box tolerance. And the fundamental reason is the small percentage of objects in the image. A straightforward and effective approach involves cropping the original image and enlarging it for detection, which can enhance accuracy. SAHI [4] proposed a method involving sliding traversal of the original image, dividing it into smaller sub-images, which generally improves model accuracy by approximately 5%. However, these methods, as illustrated in Fig. 1, often incur significant computational overhead and may lead to object truncation. Recognizing that small objects tend to cluster in specific regions within an image, this study draws inspiration from human observation techniques for densely packed small objects. Specifically, we first roughly observe the image, lock the region with potential targets, and then carefully identify the small objects within the region.

The Adaptive Focused Slicing Detection (AFSDet) method proposed in this paper aims to only cut the area with potential objects for fine-detection, which reduces the waste of computing resources while improving the detection accuracy. Simultaneously, it utilizes the feature of small changes in consecutive frames of the video data to interact the information between consecutive frames to further enhancing the detection accuracy and efficiency. The network uses the proposed Deformable Head YOLO Network (DHYNet) model as the detection model. As shown in Fig. 2, firstly, a single-class training DHYNet is used to perform coarse-detection on the original image, and the coarse-detection results are adaptively clustered and slicing. Then the patches is sent to the fine-detection model. Finally, the detection results from multiple patches

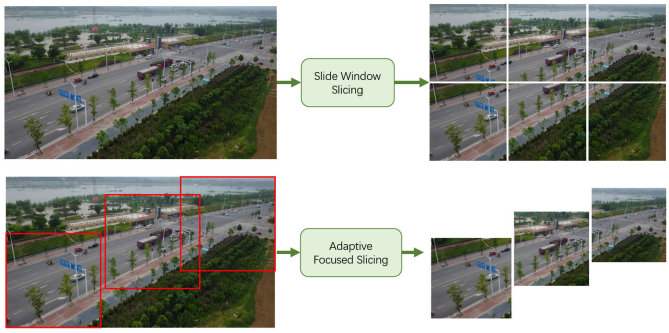


Fig. 1. The first row is the Sliding Window Slicing and the second row is the Adaptive Focused Slicing.

undergo weighted box fusion (WBF) [5], and the results of that frame are passed to the next frame to help the slicing. Meanwhile, since there is little variation between consecutive video frames, this paper proposes that adjacent frames can be fusion and coarse-detection at the same time to further enhance the detection efficiency. The main contributions of this paper are as follows:

- An improved model based on YOLOv8 [6] named DHYNet is proposed. A new module called CSPDarknet53 to FPN with DCNv3(CFD) is integrated into the detection head, fully capturing information in the feature map and alleviating the issue of losing deep semantic information regarding small objects.
- An adaptive focused slicing framework for video small objects detection is proposed, which employs adaptive scale slicing based on clustering to enhance detection. Additionally, the detection results are propagated to the next frame to enhance the slicing effect.
- The method of fusing adjacent frames is proposed, which can be regarded as coarse-detection of multiple frames at the same time to enhance efficiency without sacrificing accuracy.

II. Related Work

A. Video object detection

Due to the temporal dimension inherent in videos compared to 2D images, numerous algorithms have been developed to enhance detection performance by leveraging temporal information. DFF (Deep Feature Flow For Video Recognition) [7] considers that adjacent frames or nearby frames with similar appearances have similar features, and the feature network only acts on the feature extraction of key frames, and the optical flow is utilized to obtain the feature maps of non-key frames. FGFA (Flow-Guided Feature Aggregation For Video Object Detection) [8] considers that the features of some frames in a video sequence are subject to changes in appearance, and these changes can be improved by aggregating the features of adjacent frames or neighboring frames. Trackformer [9] uses Track Query, which is generated by the DETR detector and integrates the position information of the corresponding target over time, and the attention mechanism ensures that the model takes into account the position, occlusion, and recognition features of the object simultaneously.

B. Small object detection

Small objects, due to their limited pixel occupancy, reduced available information, and susceptibility to environmental influences, pose challenges in extracting discriminative features. Moreover, their tendency to aggregate often results in mutual occlusion among targets. To solve the above difficulties, the researcher proposed the following method.

Sample-Oriented Methods: The main problem addressed is how to sample the training data efficiently. Rrnet [10] augmented small targets by copying a small object and pasting it into different locations within the same image with random transformations. Feature-Imitation Methods: Rabbi et al [11] used GAN to perform super-resolution processing on low-resolution remote sensing images, filtering the edge details in order to avoid the loss of high-frequency information during the reconstruction process. Context-Modeling Methods: Objects' semantic or spatial relationships may provide more information when the objects themselves are of poor quality, SINet [12] assumes that primitive RoI pooling operations destroy the structure of small objects, and therefore introduces a context-aware RoI pooling layer to maintain contextual information. Slicing-and-Detect Methods: Akyon [4] et al. proposed the sliding window slicing framework, which can widely improve the detection performance by only simple slicing of the original image, and if fine-tuning on slicing will dramatically improve the performance.

III. Methods

A. Detector Improvement

YOLOv8 developed by Ultralytics in 2023, represents a significant advancement in both detection accuracy and speed compared to its predecessors in the YOLO series. This advanced object detection model has integrated many ideas that favor small objects detection in the design of its network structure. However, YOLOv8 still requires further enhancements. The research indicates that due to the insufficient representation of small object information on the feature map obtained through neck extraction, ordinary convolution is unsuitable. This paper employs deformable convolution [13] to enhance the capability of the detection head in capturing features.

For any given image, let its basic feature be denoted as $x \in R^{C\times H\times W}$, where C represents the number of channels, H and W signify the height and width of the feature map, respectively. The standard 2D convolution can be expressed as follows: Convolution is performed on a point, and the regular grid R is used to sample with as the origin, and then the weight w of the convolution kernel is weighted and summed with the sampling value. R can be expressed as (1):

$$R = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\} \tag{1}$$

Deformable convolution can be computed by (2):

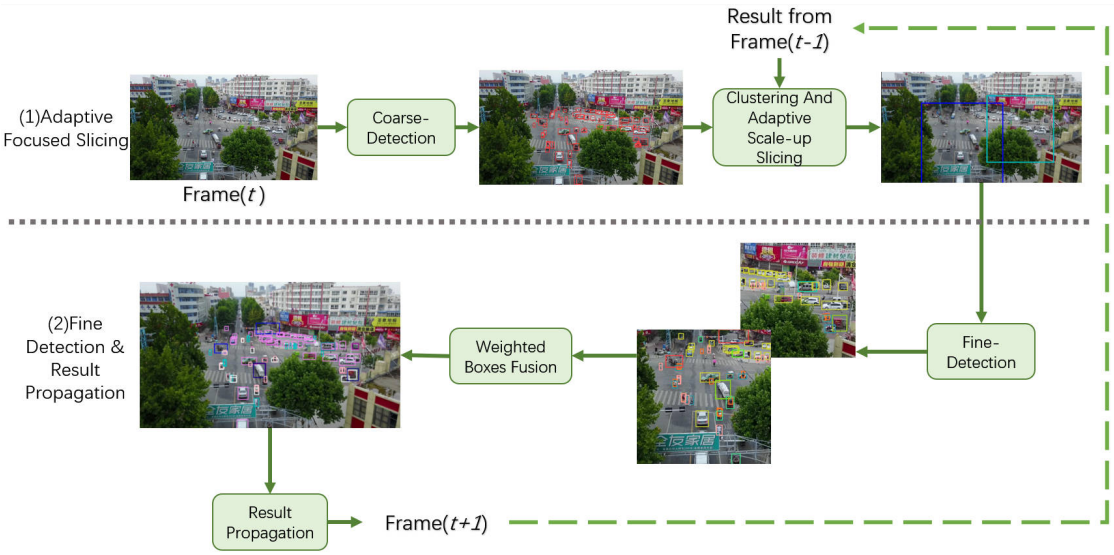


Fig. 2. The framework of AFSDet comprises two phases: (1) Adaptive Focused Slicing and (2) Fine Detection & Result Propagation. In stage 1, the video frames at time t are initially processed by the coarse-detection model. Subsequently, the coarse results at time t are clustered with the final results at time t-1 to obtain the patches through Adaptive Scale-up on the cluster. In stage 2, each patch is sequentially inputted into the fine-detection model. The outputs undergo a Weighted Boxes Fusion (WBF) process to obtain the final results, which are subsequently propagated to the frame at time t+1.

$$y\left(p_0\right)=\sum_{g=1}^{G}\sum_{k=1}^{|R|}w_g m_{gk} x_g\left(p_0+p_k+\Delta p_{gk}\right) \tag{2}$$

where G denotes the division of the spatial aggregation process into G groups, each containing a learnable offset parameter Δp_{gk} and a modulation parameter m_{gk}.

This paper presents a new model named DHYNet that integrates a new module called CFD, as shown in Fig. 3. The CFD module is designed by incorporating the concept of ELAN from YOLOv7 [14] into DCNv3. This module takes the input feature map, processes it through a CONV layer, and then splits it into two parts using the Split operation, The first part of the split feature map undergoes multiple BD modules. The output of the initial BD module, the final BD module, and the features extracted by Split are merged together. Subsequently, the combined features pass through a CONV module with a 1x1 kernel size and a step size of 1. The BD module sequentially integrates a standard convolutional layer with DCNv3, and then performs a residual connect on the result, which is formulized as follows:

$$f=\mathrm{BN}\left(\mathrm{DCNv3}\left(\mathrm{CONV}\left(f_{in}\right)\right)\right) \tag{3}$$

$$f_{out}=\mathrm{CONV}\left(f_{in}\right)\oplus\mathrm{SiLU}\left(f\right) \tag{4}$$

where f_{in} represents the input feature map, while f_{out} represents the output feature map. Additionally, BN refers to batch normalization, and SiLU (Sigmoid Linear Unit) functions as the activation function. The symbol $\oplus$ signifies the connection operation.

The entire process can be expressed as follows:

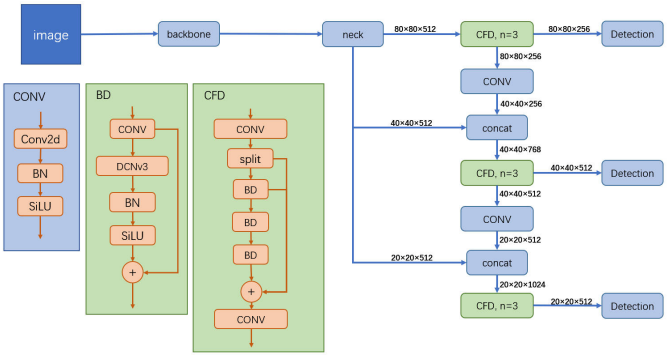


Fig. 3. The network structure of DHYNet.

$$f_a, f_b=\mathrm{split}\left(\mathrm{CONV}\left(f_{in}\right)\right) \tag{5}$$

$$f=\mathrm{BD}\left(\mathrm{BD}\left(\mathrm{BD}\left(f_a\right)\right)\right) \tag{6}$$

$$f_{out}=\mathrm{CONV}\left(f\right)\oplus\mathrm{BD}\left(f_a\right)\oplus f_b \tag{7}$$

where f_a, f_b denote the two feature map components of the original feature map split.

The CFD module harnesses the strong feature-capturing ability of deformable convolution, providing improved computational and memory efficiency through sparse sampling. This overcomes the constraints of standard convolution related to long-range dependencies and adaptive spatial aggregation, resulting in enhanced accuracy and a more thorough gradient flow analysis. By integrating the CFD module into the detection head, it facilitates the retrieval of feature maps at three different scales, effectively reducing the loss of intricate object

characteristics during downsampling and achieving superior object detection performance.

B. Adaptive Focused Slicing

Traditional image slicing methods typically involve dividing the original image into equal parts or using a sliding window approach to extract patches, leading to considerable redundant computations. In this study, we propose a novel slicing method that initially treats all labeled categories in the training set as "foreground" classes and employs the aforementioned DHYNet model for training, thereby yielding a model with improved recall for detecting potential targets.

Affinity Propagation(AP) is an algorithm that obviates the need to specify the number of clusters, instead determining the appropriate number of clusters to assign the samples based on a distance measure from the sample points. Following the extraction of the foreground from the original image in the coarse-detection stage, the centroid of the foreground bounding box serves as a sample point for unsupervised clustering employing the AP algorithm to obtain clusters. Every foreground box is allocated to a cluster, from which the top-left-bottom-right extreme values of all boxes within each cluster are derived to define the initial slicing region. Owing to the aggregation of small objects, this paper contends that there is a significant probability of potential targets being present near the foreground detected during coarse-detection, necessitating an appropriate enlargement of the slicing region.

In this paper, we propose an Adaptive Scale-up Slicing method for scaling the slice region, while mitigating the difficulty of scale variation, i.e., the same image may contain objects of widely varying scales that hinder detection. Denote the width of the original slicing region as w , and express the width of the scaled-up slicing region as in (8):

$$l^{'} = l\frac{w^2}{l^2} + 2\lambda w \tag{8}$$

where w represents the average width of all object boxes within each cluster. Multiply l by the scaling factor w^2/l^2 , and subsequently extend both sides of the slicing region by λw. Then the percentage α of foreground objects in the slicing region can be represented by (9):

$$\alpha = \frac{w}{l^{'}} = \frac{w}{l\frac{w^2}{l^2} + 2\lambda w} = \frac{1}{\frac{w}{l} + 2\lambda} \tag{9}$$

With $w/l \in (0,1)$, we have $\alpha \in (1/(1+2\lambda), 1/2\lambda)$, ensuring that the foreground percentage in the slicing region falls between two constant values. Illustrated in Fig. 4, the approach adjusts slicing regions to accommodate large-scale objects with greater areas and small-scale objects with reduced areas, effectively addressing the issue of scale variation.

C. Fine Detection & Result Propagation

Each patch obtained from the above process is detected utilizing the DHYNet, and the resulting detections are translated into the coordinate system of the original image. Post-processing is necessary to filter out duplicate detection boxes due to overlapping regions in the patches. In this paper, the Weighted Boxes Fusion (WBF) [5] method is employed instead of the conventional NMS method. WBF uses IoU to determine if a detection box corresponds to a specific object and conducts weighted fusion rather than suppression when dealing with multiple detection boxes of the same object.This paper argues that utilizing WBF effectively leverages the results from multiple detections, similar to performing ensemble learning on multiple detectors.

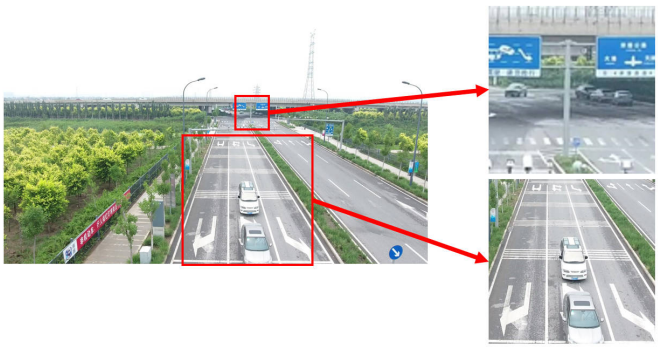


Fig. 4. On the left side, the original image with a notable scale difference between the objects in the two clusters. On the right side, the image after the application of Adaptive Scale-up Slicing, which effectively reduce the scale difference among the foreground objects.

The full detection results of the current frame is acquired and subsequently carried over to the next frame for joint clustering with the coarse-detection output of that frame. This is because the variations between consecutive video frames are usually minimal, and the high-quality detection result from the previous frame can be used as a reference for slicing patches in the next frame. The process of coarse-detection is still maintained, because coarse-detection can detect potential targets outside the existing slicing region.

D. Superimposing Adjacent Frame Detection

Consider using the characteristics of video data to further improve the detection efficiency. Between consecutive frames of a video, there is usually very little change in the background and significant change in the foreground, and it is possible to superimpose the pictures of the two neighboring frames, as shown in Fig. 5.

Consider a video sequence $S = \{(x_t, y_t)\}_{t=1}^{n}$, where x_t represents the image at index t and y_t represents the corresponding label. Then the video sequence $\tilde{S} = \{\tilde{x_t}, \tilde{y_t}\}_{t=1}^{n-1}$ after superimposing of neighboring frames are defined as:

$$\tilde{x_t} = \frac{1}{2}x_t + \frac{1}{2}x_{t+1}, \tilde{y_t} = y_t \cup y_{t+1} \tag{10}$$

the superimposed video sequence $\tilde{S}$ conforms to the distribution P_{fusion}. To ensure that the probability distribution P_{model} of the model closely approximates P_{fusion}, mixup [15] is employed as an augmentation technique during the training of the coarse-detection model.

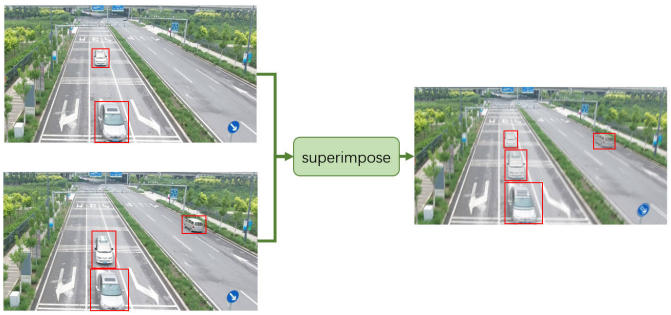


Fig. 5. This figure illustrates the superimposition of adjacent frames. It showcases the rendering of two images with foreground cars into a single composite image, whereby detecting objects in the fused image is tantamount to detecting objects in both original images simultaneously. Notably, the selected images are taken from frames at distinct time intervals rather than adjacent frames in order to show the effect of the method more clearly.

IV. Experiments

A. Datasets and Evaluation Metrics

This paper uses the VisDrone-MOT dataset collected by the AISKYEYE team at the Lab of Machine Learning and Data Mining, Tianjin University, China. The dataset comprises recordings obtained from diverse drone cameras, capturing scenes across varying weather and lighting conditions. It is a small object detection dataset with a high level of difficulty.

This paper adopts the evaluation protocol established by MSCOCO [16]. AP_{50} refers to the AP measured when the IoU threshold is set to 0.5, similarly AP_{75} refers to the AP when the IoU threshold is set to 0.75, and AP refers to mean Average Precision. The $\#img$ is used to record the number of slicing patches, and the FPS indicates how many original images the model can detect per second.

B. Implementation Details

In this paper, AFSDet is implemented with PyTorch 1.7 and the detection model is based on the open source Ultralytics 8.0.195 YOLOv8 framework. The experiments are completed on NVIDIA GeForce RTX 3070, Ubuntu system. In training the coarse-detection model, this paper uses Mosaic, Mixup, Horizontal Flip, Random Crop, and Color Transform data augmentation. The other parameters were kept as the default parameters recommended by the Ultralytics framework. The AP clustering algorithm uses the interface provided by the Scikit-learn library.

C. Comparison Algorithm

This paper compares seven popular and powerful algorithms. YOLOv8 is a highly popular single-stage end-to-end detection algorithm known for its exceptional performance in industrial applications. RT-DETR [17] represents a high-performance Transformer-based object detection model. FCOS [18] is an anchor-free single-stage detection algorithm. Cascade-RCNN [19] is a powerful two-stage object detection algorithm. VFNet [20] is specifically designed for detecting dense small targets. SAHI [4] is a sliding slicing framework, which is applied to the YOLOv8n detector in this experiment.

D. Results

The results on the VisDrone-MOT test set are shown in Table I. The results show that the method presented in this paper provides a large improvement over common object detections, sacrificing detection speed for a larger improvement in detection accuracy relative to the baseline model, YOLOv8, and the sacrificed speed is still usable in real time in some scenarios. In comparison to the previous slicing-based method SAHI, the approach proposed in this study effectively minimizes the number of patches by introducing an additional coarse-detection step, which not only greatly improves the detection speed, but also improves in the detection accuracy.

E. Ablation Experiments

The results of the ablation experiments as shown in Table II demonstrate that the enhanced model achieves superior results as a result of its ability to robustly capture features through the deformable convolutional layer and extract more informative feature maps for the detection head. The proposed adaptive focused slicing framework extracts patches from the original image containing potential targets and conducts separate fine-detection. The experiments demonstrate its significant impact on enhancing detection accuracy, however, it also leads to a decrease in detection efficiency. By combining the DHYNet with the adaptive focused slicing framework, a significant improvement in accuracy is observed on the validation set compared to the baseline model, thus validating the effectiveness of the proposed method in this paper.

V. Conclusion

This paper introduces a novel method called Adaptive Focused Slicing Detection for small object detection in videos. This method slices and scales up the regions in the original video frames containing potential targets, then feeds them into the fine-detection model to aid in small object detection. The DHYNet is proposed, utilizing this model for initial coarse-detection to identify potential targets. Subsequently, the set of targets is clustered and dynamically divided into multiple patches. Each patch is independently processed by the fine-detection model, and the resultant outputs are post-processed. To account for the traits of video data, this paper proposes inter-frame propagation of the results to enhance accuracy, and superimposing adjacent frames to reduce computations. Comparative experiments were conducted on the VisDrone-MOT dataset to showcase the method's superiority over other powerful approaches. Additionally, ablation experiments were performed to validate all components of the method.

References

[1] S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Trajectory-based surveillance analysis: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 29, no. 7, pp. 1985–1997, 2018.

TABLE I
COMPARISON OF OUR METHOD WITH OTHER METHODS FOR OBJECT DETECTION ON VISDRONE-MOT TEST SET.

Method	Evaluation Metrics							
	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	FPS	$\#img$
YOLOv8n	8.8	18.6	73	3.6	11.0	18.9	**88.3**	-
RT-DETRn	8.7	18.3	7.4	4.2	9.4	18.2	16.7	-
FCOS	7.8	19.5	4.3	3.8	6.3	15.3	27.1	-
VFNet	10.6	20.6	10.7	4.1	10.1	21.9	22.2	-
Cascade-RCNN	10.9	19.4	11.5	4.8	11.0	18.3	15.9	-
YOLOv8n+SAHI	12.8	24.9	11.9	4.5	16.5	22.5	2.1	46732
AFSDet(Ours)	**14.0**	**28.2**	**12.3**	**6.3**	**17.1**	**22.6**	12.3	**27121**

TABLE II
AN ABLATION EXPERIMENTS ON THE VISDRONE-MOT VALIDATION SET. AFS DENOTES THE PROPOSED ADAPTIVE FOCUSED SLICING FRAMEWORK. DN DENOTES THE PROPOSED DHYNET.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
AFSDet	**18.7**	**35.5**	**17.0**	**15.0**	**30.2**	**40.8**
AFSDet-DN	14.9	30.1	13.0	7.9	24.6	33.4
AFSDet-AFS	8.8	18.6	7.3	6.1	13.4	19.2
AFSDet-AFS-DN	8.7	18.1	7.1	3.6	11.0	18.9

[2] L. Jiao, R. Zhang, F. Liu, *et al.*, "New generation deep learning for video object detection: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3195–3215, 2021.

[3] G. Cheng, X. Yuan, X. Yao, *et al.*, "Towards large-scale small object detection: Survey and benchmarks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[4] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 966–970.

[5] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, p. 104 117, 2021.

[6] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics yolov8*, version 8.0.0, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[7] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2349–2358.

[8] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 408–417.

[9] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8844–8854.

[10] C. Chen, Y. Zhang, Q. Lv, *et al.*, "Rrnet: A hybrid detector for object detection in drone-captured images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 100–108.

[11] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network," *Remote Sensing*, vol. 12, no. 9, p. 1432, 2020.

[12] X. Hu, X. Xu, Y. Xiao, *et al.*, "Sinet: A scale-insensitive convolutional neural network for fast vehicle detection," *IEEE transactions on intelligent transportation systems*, vol. 20, no. 3, pp. 1010–1019, 2018.

[13] W. Wang, J. Dai, Z. Chen, *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 408–14 419.

[14] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7464–7475.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[16] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[17] Y. Zhao, W. Lv, S. Xu, *et al.*, "Detrs beat yolos on real-time object detection," *arXiv preprint arXiv:2304.08069*, 2023.

[18] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: A simple and strong anchor-free object detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1922–1933, 2020.

[19] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[20] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, "Varifocalnet: An iou-aware dense object detector," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8514–8523.

Diverse Time-Frequency Attention Neural Network for Acoustic Echo Cancellation

Jinzhuo Yao*, Hongqing Liu*, Yi Zhou*, Lu Gan† and Junkang Yang*
* School of Communications and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China
† College of Engineering, Design and Physical Science, Brunel University London, U.K.
E-mail:s220101190@stu.cqupt.edu.cn

***Abstract*—Acoustic echo cancellation (AEC) aims to eliminate echoes from near-end microphone signals and recover the near-end speech at the same time. In this work, we propose a Diverse Time Frequency Attention Neural Network (DTFAN) for AEC that operates in a full network-based manner. To that end, we first utilize a network aiming at aligning the features of the far-end reference signal and the near-end microphone signal. After that, the diverse time-frequency attentions capturing the intrinsic connections of the features in the time and frequency domains are developed. Since the alignment of the reference signal is conducted by the network, the requirement of traditionally pre-processing the far-end signal is avoided, and the whole network is end-to-end. The experimental results show that the proposed framework performs well and robustly on the synthetic test set and the blind test dataset compared to other recent approaches, especially in double-talk scenarios.**

I. Introduction

In a full-duplex voice communication system, the far-end user will receive a modified version of his/her speech due to the acoustic coupling that occurs between the near-end microphone and the near-end speaker. Acoustic Echo Cancellation (AEC) systems are designed to eliminate echoes from the microphone signal while minimizing distortion of the near-end speaker's voice [1], [2].

In addition to traditional methods [1], [2], recently, there have been many successes in deep learning-based AEC studies due to the great potential demonstrated by deep learning in modeling complex nonlinear problems. In [3], frequency-domain adaptive filters [4] is first used to remove linear echoes, then a complex GCCRN network is utilized to eliminate residual echoes and background noise. It seems that a use of traditional filter on top of network is cumbersome. Some studies that only utilize neural networks for the entire process have also achieved good performance. For example, DTLN [5] uses LSTM [6] networks to learn speech features. F-T-LSTM [7] employs a complex encoder-decoder structure, and after extracting the signal features, F-LSTM and T-LSTM are used for time-frequency domain modeling separately, and the final estimation yields the near-end microphone signal. In DeepVQE [8], the alignment module is designed to estimate the time delay of the far-end reference signal and the near-end microphone as a way to align the features of the two signals. Additionally, the Complex convolving mask block (CCM) module is also used for the final output of the estimated speech, and using three-vector components instead of the conventional two-vector components (real and imaginary) provides more stable output results and prevents low noise and echo leakage.

Self-attention mechanisms [9], [10] are good at capturing global information and contextual relationships as a way to enable better modeling. The authors in [11] propose a lightweight Axial Self-Attention (ASA) module to model along the frequency and time axis separately, which improves the network's ability to capture the global internal relationships between time domain features and frequency domain features separately with a smaller number of parameters and operations. Moreover, PCNN [12] proposes a Self Channel-Time-Frequency Attention (Self-CTFA) Module to capture the connections between signal features from the channel, frequency, and time domain dimensions, respectively, and finally performs feature extraction and fusion. In addition, the study reported after ablation experiments that F-attention has a great impact on the model effect. This shows that for processing speech signals, there is a great potential for a multidimensional self-attention mechanism, leveraging cross domain information.

Noting that current approaches tend to use one time-frequency attention for modeling, we propose a diverse time-frequency attention neural network in designing a AEC system. After initially extracting the features of the far-end reference signal and the near-end microphone signal, we estimate the time delay and align them, and then two different time-frequency attention modules (ASA and Frequency Transformation Block (FTB)) to capture the contextual and global information of the features are developed. We observe that multiple temporal-frequency attention outperforms single temporal-frequency attention in learning the contextual relationships between the time domain and the frequency domain simultaneously, which is particularly important for the model to differentiate between the far-end speaker's voice, the near-end speaker's voice, and the background noise. This is also the reason why our model presents an excellent performance in double-talk scenarios.

II. Problem formulation

The microphone signal $y(n)$ is composed of the echo signal $d(n)$, the near-end speech signal $s(n)$, and the background noise $v(n)$, given by

$$y(n) = d(n) + s(n) + v(n), \tag{1}$$

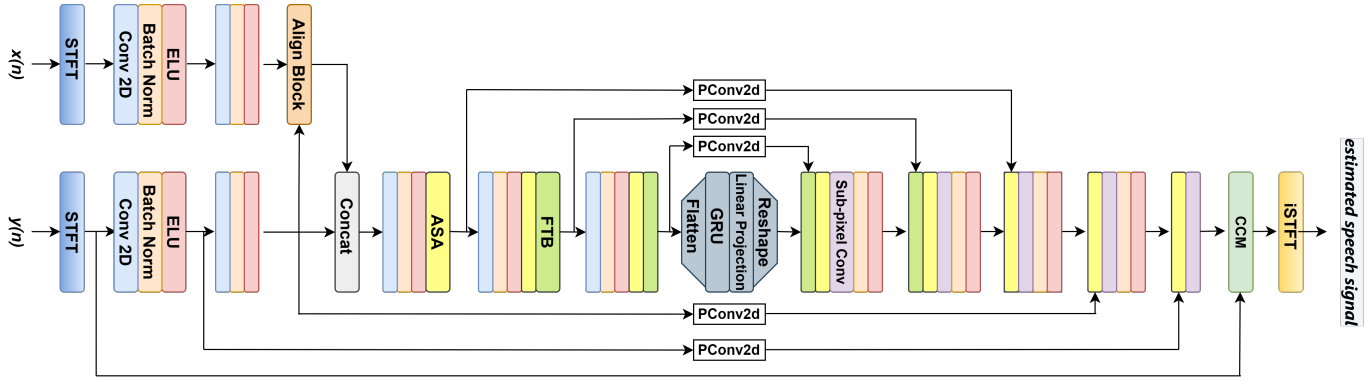


Fig. 1: The proposed DTFAN, where yellow represents ASA and green represents FTB (Best viewed in color).

where n is sample index, $d(n)$ is obtained by convolving the far-end reference signal $x(n)$ with the room impulse response $h(n)$. The task of AEC is to estimate $s(n)$ with $x(n)$ and $y(n)$ provided and the whole process can be formalized as

$$\hat{S}_r, \hat{S}_i = f_\varphi\left(X_r, Y_r, X_i, Y_i\right), \tag{2}$$

where f and φ represent our network and its parameters, and X_r, X_i, Y_r, Y_i are the real and imaginary parts of the complex spectrum obtained by short-time Fourier transform (STFT) of $x(n)$ and $y(n)$. Similarly, $\hat{S}_r$, $\hat{S}_i$ are the real and imaginary parts of the complex spectrum of the estimated clean speech.

A. Network Architecture

The structure of the whole network is depcited in Figure 1. The main body of our model is an encoder and decoder structure with a bottleneck layer. The encoder has two branches to extract features from the far-end reference signal and the near-end speech signal, respectively. After aligning the signal features, the ASA [11] and FTB [13] continue to capture the global information of the features alternatively. After the features passing through the decoder, the real and imaginary parts of the estimated speech spectrum are output by a CCM block [8].

1) Encoder: The encoder consists of a near-end mic branch and a far-end reference signal branch with five and two encoding blocks, respectively. The first two encoding blocks of the two branches are used to initially extract the signal features and feed the features into the Align Block[8], which learns the intrinsic relationship between the two signal features by estimating the delays between the two signals, and then feeds the features into the microphone branch after alignment. The purpose of doing so is that after alignment the network can better cancel out the echos. After that, three encoders further extract the features. The encoders consist of a convolutional block (all have), ASA (last three have), and FTB (last two have), respectively. The convolution block consists of a down-sampled convolutional layer, a batch normalization (BN) layer, and an ELU function. The downsampling convolution layer has a convolution kernel size of 4×3 and a step size of 1×2 to reduce the number of bins along the frequency values. In addition, the convolution is padded to maintain causality.

2) Bottleneck: In the bottleneck layer, we first expand the feature tensor $\mathbf{X} \in \mathbb{R}^{b\times c\times t\times f}$ along the channel and frequency axes as $\mathbf{X} \in \mathbb{R}^{b\times t\times (c\cdot f)}$, where b, c, t, $f \in \mathbb{N}$ denote the batch size, the lengths of the channel, time, and frequency axes. After that, $\mathbf{X}$ is fed into the GRU layer for contextual modeling and then into the linear projection, and the resulting tensor is reshaped back to its original shape. To reduce the number of hidden units, we added a linear injection after the recurrent layer, which additionally improves the training stability and model performance.

3) Decoder: The network contains five decoders and each decoder consists of FTB (first two have), ASA (all have), sub-pixel convolution block [14] (all have). Since the features are down-sampled in the encoder part and the length of the feature tensor is decreasing along the frequency axis, the decoder part needs to be up-sampled. As opposed regular up-scaling method based on transposed convolution, we use sub-pixel convolution instead. For $\mathbf{X} \in \mathbb{R}^{b\times c\times t\times f}$, it is transformed to $\mathbf{X} \in \mathbb{R}^{b\times 2c\times t\times f}$ by a regular convolution with $2c$ filters, and then reshaped to produce $\mathbf{Y} \in \mathbb{R}^{b\times c\times t\times 2f}$. This is done by learning a series of filters to up-sample the feature map. A sub-pixel convolution block consists of a sub-pixel convolution layer, a BN layer, and an ELU function stacked together. The last sub-pixel convolution block has no BN layer and no ELU function. In addition, we use point-wise convolution-based skip connection instead of summing or concatenation. The encoder features are point-wise projected and added to the corresponding decoder outputs. In this way, the number of channels in the encoder can be chosen more flexibly. At the end of the decoder, the CCM module processes the last obtained feature map to reconstruct the spectrum of the estimated clean speech.

4) ASA and FTB: To reduce computation and memory and to better handle long sequence signals represented by speech signals, we leverage the modeling ability of ASA and FTB. The structure of ASA is shown in Figure 2, where Ci denotes the number of input feature channels and C denotes the number of attention channels. ASA computes the attention scores sequentially along the frequency and time axes, which can be abbreviated as F-attention and T-attention. These at-

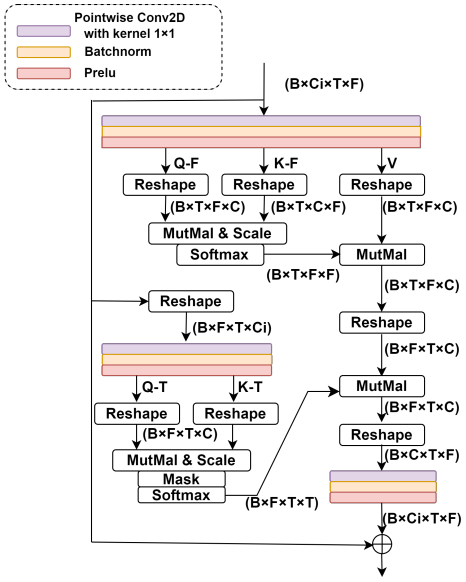


Fig. 2: Axial Self-attention (ASA).

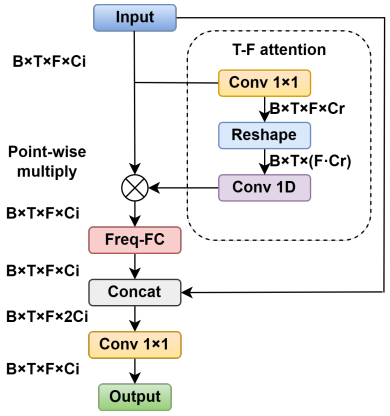


Fig. 3: Frequency Transformation Block (FTB).

tention matrices help the model capture potential relationships between signal feature long distances in both the frequency and time domains. In our network, the ASA module is applied to the last three decoders and all encoders.

In addition, consider fact that ASA is a purely convolutional module, to further diversify the utilization of time-frequency attention mechanisms, we introduce the FTB module into the network. We observe that acquiring time-frequency attention through multiple methods is more conducive to learn contextual information in the time and frequency domains. The structure of FTB is shown in Figure 3, where C and Cr are the number of input channels and the number of attention channels, respectively, and it is worth noting that Freq-FC is a key part of FTB. It contains a trainable frequency transformation matrix for feature map slicing at each time point, which is also an important difference from ASA. FTB solves the problem that 2D convolutional has a small perceptual field. Therefore, we embed FTB into the last two encoders and the first two decoders to produce the feature output of a frequency-aware field to further attend different frequencies.

B. Loss Function

We employ a joint loss function as the optimization objective. Specifically, we use the mean square error (MSE) loss in the frequency domain and the scale-invariant signal-to-noise ratio (SI-SNR) [15] as the loss function in the time domain. The final loss function is the summation of them, given by

$$Loss = -SI - SNR + \log_{10}(MSE(S, \hat{S})), \qquad (3)$$

where S, $\hat{S}$ are the spectrum of the target speech and the estimated clean speech.

III. Experiments

A. Dataset

To train our model, four types of signals need to be prepared: the near-end speech signal, background noise signal, far-end speech signal, and the corresponding echo signal.

We use the synthesized dataset provided by the AEC Challenge [16] for the far-end speech signals and the corresponding echo signals. This dataset consists of 10,000 10-second clips. The first 500 of these clips were used to create the test set and the last 9500 clips were used to create the training set.

For the near-end speech signal and background noise, we used the clean speech signal and background noise signal provided by DNS challenge [17]. Approximately 6.8 hours of clean speech and 5.5 hours of background noise signals were used to create test sets. In addition, we cleaned the background noise using the open-source model sliro-vad [18], cropping out potential segments containing human voices. Finally, we produced a 300-hour dataset for training and validation of the model, with a 4:1 ratio of the training set to the validation set and 10 seconds per audio clip. To improve the model's generalization ability, the signal-to-return ratio (SER) was randomly set from -15 dB to 15 dB and the SNR was randomly set from 5 dB to 25 dB.

B. Experimental setup

All audio sampling rates are 16 kHz. For the network, the STFT window length is 512 points, the window shift is 256, the FFT size is 512, and the Hanning window is used. Our model is trained using Adam's optimizer with an initial learning rate of 1e-4 for 150 epochs. If there is no improvement in performance for 3 epochs, the learning rate is halved.

For the encoder, the down-sampled convolution blocks for the far-end reference signal branch have 64 and 128 filters, respectively, while the number of filters for the near-end microphone signal branch is 32, 128, 128, 128, and 128. The number of filters in the Sub-pixel convolution in the decoder is 128, 128, 128, 64, and 27. The number of attention channels in all ASA modules is one-fourth of the number of input channels. The attention channel of the two FTBs closer to the

TABLE I: Ablation experiment. DT: doubletalk, ST: single-talk, FE: far-end, NE: near-end, T-att: T-attention, F-att: F-attention.

				DT-Noisy		DT-Clean		ST-FE	ST-NE	
CASE	T-att	F-att	FTB	PESQ	STOI	PESQ	STOI	ERLE	PESQ	STOI
1	✓	✓	✓	**2.46**	**0.86**	**2.72**	**0.89**	**46**	**3.27**	**0.95**
2	×	✓	✓	2.29	0.83	2.56	0.87	42	3.10	0.94
3	✓	×	✓	2.33	0.84	2.57	0.87	39	3.15	0.94
4	×	×	✓	2.14	0.81	2.39	0.85	42	3.04	0.94
5	✓	✓	×	2.19	0.82	2.45	0.86	24	3.15	0.94

bottleneck layer is set to 8, and in the remaining two FTBs, the attention channel is set to 15. The hidden unit of the GRU in the bottleneck layer is 320.

C. Performance metrics

We use perceptual evaluation of speech quality (PESQ) [19] and short-time objective intelligibility (STOI) [20] to evaluate the model's performance in double-talk and near-end single-talk scenarios. In addition, we use the echo return loss enhancement (ERLE) [21] to measure the model's performance in far-end single-talk scenarios. Finally, the AEC challenge provides subjective evaluation results based on the average P.808 mean opinion scores [22].

IV. Results and Analysis

To validate the role of diverse attention in the network, we first conducted ablation experiments for five different cases, and the results are shown in Table I. We observe that the model performance improves substantially in all scenarios after adding each type of attention. Notably, when CASE 2 is compared to CASE 3, T-attention is found to have a greater impact on performance than F-attention. We believe this is due to the down-sampling of features by the model in the encoder stage, which leads to a reduction in the feature dimensions along the frequency axis and objectively reduces the F-attention perceptual field. The impact is also highlighted when comparing CASE 4, CASE 5 with CASE 2 and CASE 3. Comparing Case 4 and Case 5, we can see that ASA and FTB have different impacts on performance in different scenarios. The reason is the FTB focuses more on modeling along the frequency axis, while the ASA models along both the time and frequency axes.

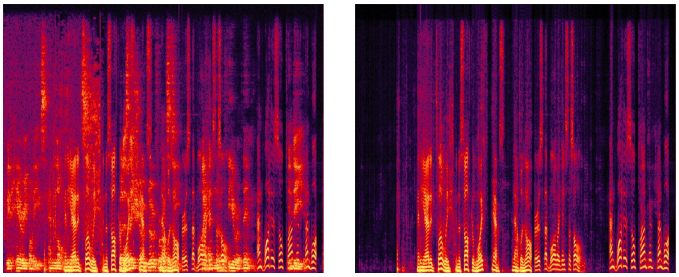
TABLE II: Objective performance metrics on the synthetic test set.

		DT-Noisy		DT-Clean		ST-FE	ST-NE	
Methed	Para	PESQ	STOI	PESQ	STOI	ERLE	PESQ	STOI
Unprocessed	-	1.66	0.70	1.81	0.73	-	2.64	0.92
Baseline23	-	2.05	0.78	2.42	0.84	39	2.77	0.92
DTLN	10M	2.14	0.81	2.37	0.84	33	3.01	0.94
DeepVQE	6.1M	2.40	0.87	2.63	0.90	**50**	3.14	0.95
Ours	5.1M	**2.64**	**0.88**	**2.88**	**0.91**	48	**3.38**	**0.96**

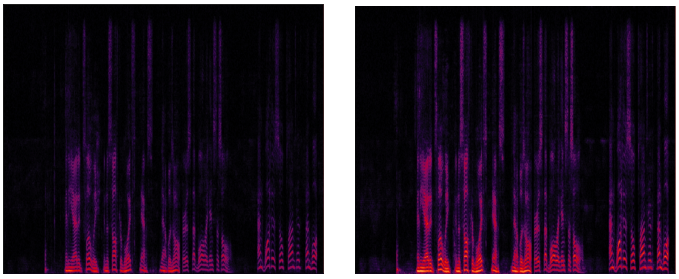
For objective performance evaluation, we compare our network with three other methods, the open-source baseline model of the 2023 AEC Challenge, DTLN, and DeepVQE, where the DTLN and DeepVQE models are retrained with our dataset, and the number of hidden units of the LSTM in the DTLN is 512. From the results in Table II, our method outperforms the other methods in almost all scenarios, especially in the double-talk scenario.

TABLE III: AECMOS of the blind test set in interspeech2021. DT ECHO means more associated with residual echo, DT Other means more related to other degradation.

Method	DT Echo DMOS	DT Other DMOS	ST FE Echo DMOS	ST NE MOS	Overall
Baseline21	4.04	3.45	3.82	4.18	3.87
Baseline23	4.32	3.93	4.41	4.19	4.21
F-T-LSTM	4.44	3.90	4.44	3.78	4.14
GCCRN	4.36	4.23	4.34	4.26	4.29
DeepVQE	4.57	4.15	**4.53**	4.05	4.32
Ours	**4.63**	**4.24**	4.46	**4.27**	**4.40**



(a) Microphone signal (b) DTLN



(c) DeepVQE (d) DTFAN

Fig. 4: A double-talk utterance in synthetic test set.

We also conducted experiments using the blind test set Interspeech2021 and the results are shown in Table III. We compare with four methods, the Interspeech 2021 AEC Challenge Baseline [16], the ICASSP 2023 AEC Challenge baseline [23], F-T-LSTM [7], GCCRN [3] and DeepVQE[8]. The F-T-LSTM employs a complex neural network to better mine the phase information for modeling. At the same time, the F-LSTM and T-LSTM are set up at the bottleneck layer to scan the time and frequency axes, respectively, for

temporal modeling. The GCCRN first estimates the linear echo using the partitioned block frequency domain least mean square (LMS) algorithm [4]. It thereafter performs residual echo cancellation and noise suppression using the GCCRN network. The GCCRN employs gated convolution instead of ordinary convolution, which learns the features of the channel and different spatial locations as a means of selecting the generating mechanism, employs a partitioned LSTM at the bottleneck layer for temporal modeling, and finally decodes the real and imaginary parts for output. In DeepVQE, a self-attention based alignment module is designed to align the near-end microphone signal features with the reference signal features, and subsequently the signal features are fed into the encoder-decoder structure network to finally obtain the estimated near-end speech. Our model performs well in the double talk scenario, maximizing the quality of the near-end speech while eliminating echoes, producing the highest overall score.

Finally, to visually see the differences, we provided the produced spectrum in synthetic dataset, depicted in Figure 4. Compared to DTLN and DeepVQE, the proposed DTFAN produces more details in harmonics and preserves speech signal better in high frequency. This further showcases the benefits using diverse attention mechanism.

V. Conclusions

In this paper, we observed that diverse time-frequency attention is better than single time-frequency attention in learning the contextual relationship between time and frequency domains, which is important for modeling the distinction between far-end speaker speech, near-end speaker speech, and background noise. Based on that, the proposed DTFAN integrates ASA and FTB to diversify learn global information extraction and modeling. Experiments show that diverse time-frequency attention leads to a superior AEC performance compared to single time-frequency attention.

References

[1] J.-S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[2] B. Farhang-Boroujeny, *Adaptive filters: theory and applications*. John wiley & sons, 2013.

[3] R. Peng, L. Cheng, C. Zheng, and X. Li, "Acoustic echo cancellation using deep complex neural network with nonlinear magnitude compression and phase information.," in *Interspeech*, 2021, pp. 4768–4772.

[4] K. Eneman and M. Moonen, "Iterated partitioned block frequency-domain adaptive filtering for acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 143–158, 2003.

[5] N. L. Westhausen and B. T. Meyer, "Acoustic echo cancellation with the dual-signal transformation lstm network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7138–7142.

[6] L. S.-T. Memory, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 2010.

[7] S. Zhang, Y. Kong, S. Lv, Y. Hu, and L. Xie, "Ft-lstm based complex network for joint acoustic echo cancellation and speech enhancement," *arXiv preprint arXiv:2106.07577*, 2021.

[8] E. Indenbom, N.-C. Ristea, A. Saabas, T. Parnamaa, J. Guzvin, and R. Cutler, "Deepvqe: Real time deep voice quality enhancement for joint acoustic echo cancellation, noise suppression and dereverberation," *arXiv preprint arXiv:2306.03177*, 2023.

[9] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Müller, S. Stüker, and A. Waibel, "Very deep self-attention networks for end-to-end speech recognition," *arXiv preprint arXiv:1904.13377*, 2019.

[10] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7867–7871.

[11] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9122–9126.

[12] X. Xu, W. Tu, and Y. Yang, "Pcnn: A lightweight parallel conformer neural network for efficient monaural speech enhancement," *arXiv preprint arXiv:2307.15251*, 2023.

[13] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 9458–9465.

[14] W. Shi, J. Caballero, F. Huszár, *et al.*, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[15] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[16] R. Cutler, A. Saabas, T. Pärnamaa, *et al.*, "Interspeech 2021 acoustic echo cancellation challenge.," in *Interspeech*, 2021, pp. 4748–4752.

[17] C. K. Reddy, H. Dubey, K. Koishida, *et al.*, "Interspeech 2021 deep noise suppression challenge," *arXiv preprint arXiv:2101.01902*, 2021.

[18] S. Team, *Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier*, https://github.com/snakers4/silero-vad, 2021.

[19] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, IEEE, vol. 2, 2001, pp. 749–752.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2010, pp. 4214–4217.

[21] S. Theodoridis and R. Chellappa, *Academic press library in signal processing: Image, video processing and analysis, hardware, audio, acoustic and speech processing*. Academic Press, 2013.

[22] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "Aecmos: A speech quality assessment metric for echo impairment," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 901–905.

[23] R. Cutler, A. Saabas, T. Parnamaa, *et al.*, "Icassp 2022 acoustic echo cancellation challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9107–9111.

LDMSE: Low Computational Cost Generative Diffusion Model for Speech Enhancement

Yuki Nishi*,Koji Iwano‡ and Koichi Shinoda†
* Tokyo Institute of Technology, Tokyo
E-mail: y-nishi@ks.c.titech.ac.jp
† Tokyo City University, Tokyo
E-mail: iwano@tcu.ac.jp
‡ Tokyo Institute of Technology, Tokyo
E-mail: shinoda@c.titech.ac.jp

***Abstract*—Recently, a generative model called diffusion model has attracted attention. Compared to GANs, it can be trained stably but has a high computational cost in the generation stage. This paper proposes a method called Low computational cost Generative Diffusion Model for Speech Enhancement (LDMSE). It reduces its computational cost with comparable quality by compressing speech signals to a latent space using an autoencoder and removing noise in the latent space using the diffusion model. In our evaluation using VOICBANK-DEMAND and WSJ0-CHiME3 datasets, the proposed method reduced the generation time by more than 35% without any degradation in speech quality.**

I. Introduction

Speech enhancement technology for noise reduction is needed in various situations. These include speech assistance with hearing aids, noise reduction in speech recognition, and processing for speech-containing content on the Web. In recent years, neural network-based technologies have been widely used to enhance speech with higher quality. Among neural network (NN) techniques, the Diffusion Model has attracted much attention because of its high quality in generation tasks and stable training. It has also been studied in the field of speech enhancement. However, this method has a drawback in that it is computationally expensive due to the need to pass data through the network repeatedly during the generation phase.

In this study, we propose a method called Low computational cost generative Diffusion Model for Speech Enhancement(LDMSE), aiming to reduce the computational cost of the generation stage in the diffusion model without sacrificing enhancement quality. The input speech data is dimensionally reduced into a latent space using an Encoder and Decoder. Then, the computational cost of each generation step is reduced accordingly. In our evaluation, we proved that LDMSE reduces the total computational cost by more than 35% without any degradation in speech quality.

II. Previous studies

A. Speech Enhancement

Speech enhancement has been utilized for various situations. In this study, we estimate clean speech data x given noisy speech $y = x + n$, a mixture of environmental noise n and x. Additionally, we only consider single-channel data. Recently, NN-based methods have achieved higher quality than those without using NN, such as the method of suppressing certain frequency regions [1] and the method of modeling human voices with hidden Markov models [2]. Some methods directly estimate a mask in the frequency domain [3], [4]. For example, Kingma et al. [5] utilize Variational Autoencoder (VAE) for this purpose. Moreover, Generative Adversarial Networks (GAN) have achieved high quality in speech enhancement [6]. For instance, MetricGAN+ [7] applies GAN to speech enhancement. However, GAN is unstable in its training. The following diffusion model solves this problem and achieves better quality.

B. Diffusion Model

One of the most popular generative models is the diffusion model. It models the diffusion process of the target data to which Gaussian noise is added and regenerates the original clean data from the target noisy data reversely. Diffusion models are divided into two categories: score-independent and score-based diffusion models.

The score-independent diffusion model, the so-called Denoising Diffusion Probabilistic Model (DDPM) [8], estimates a Gaussian noise added to the input data. It first defines the noise addition formula with N_{DDPM} steps. The transition of data from n to $n+1$ $(0 \leq n \leq N_{\mathrm{DDPM}} - 1)$, which we call it forward direction, is defined as $x_{n+1} = \sqrt{1-\beta_n}x_n + \sqrt{\beta_n}z$, where β_n is a hyperparameter and z is sampled from normal Gaussian distribution. The transition from x_{n+1} to x_n is also computed explicitly, which is the same as the inference process for, for example, generating images and sounds.

The score-based diffusion model [9] is formulated based on the stochastic differential equation (SDE). First, Gaussian noise diffusion is formulated as:

$$\mathrm{d}x_t = f(x_t)\mathrm{d}t + g(t)\mathrm{w}(t), \quad 0 \leq t \leq T \tag{1}$$

where x_t is the state of the data at t and x_0 corresponds to the clean data, and $\mathrm{w}(t)$ is a Wiener process. The functions f, g are defined in various ways depending on the experimental

setup. The SDE in the reverse direction is formulated as [9], [10]:

$$\mathrm{d}x_t = \left[-f(x_t) + g(t)^2 \nabla_{x_t} \log p_t(x_t)\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathrm{w}}(t) \quad (2)$$

where $0 \leq t \leq T$ and $\nabla_{x_t} \log p_t(x_t)$ is called a score, $p_t(x_t)$ is the probability distribution of x_t, and $\bar{\mathrm{w}}(t)$ is the standard Wiener process when t is traced backward from T to 0. The score is computed from f and g.

Several score-based diffusion models have been proposed to obtain x_0 by simulating backward transitions from noisy data x_T [10]. The simplest method is Euler-Maruyama method [11], which discretizes the range from 0 to T into N steps. Another high-quality method is the Predictor-Corrector (PC) samplers [9] where the process is divided into a Predictor and a Corrector. The Predictor solves the SDE, and the Corrector refines it to achieve high quality.

C. Reduction of Computational Cost of Diffusion Model

In the diffusion model, the process of gradually removing noise by passing the data through NN multiple times (e.g., 30 times for SGMSE+ [12], which is the baseline for this study) during the generation stage during the generation stage results in high computational cost. Several studies solve this problem by reducing the number of steps in the generation phase. As for DDPM, Denoising Diffusion Implicit Models (DDIM) [13] generalizes the definition formula of the diffusion process so that the inverse process is designed with arbitrary step intervals.

This approach that reduces the number of steps cannot be applied to the score-based diffusion model because its reverse process simulates SDEs during inference. Instead, Fast sampler [14] trains an NN sampler, which is applied to various diffusion models. Denoising MCMC (DMCMC) [15] introduces a relatively small NN that estimates the noise magnitude at each step. These two methods can generate with a small number of steps, though they slightly degrade the quality.

D. Speech Enhancement with Diffusion Model

Several studies have applied the diffusion model to speech enhancement. However, the widely used DDPM [8] removes only Gaussian noise, which limits its applicability to speech enhancement tasks where the noise types are various, such as the crowds at train stations and birdsong [16]. CDiffuSE [17] addresses this limitation by incorporating not only Gaussian noise but also non-Gaussian noise (e.g., environmental noise) in the forward step.

A study of speech enhancement using the score-based diffusion model is SGMSE+ [12], which is used as the baseline in this study. In this method, the functions f and g in Eq. 2 are defined as

$$f(x_t, y) = \gamma(y - x_t) \quad (3)$$

$$g(t) = \sigma_{\min} (\sigma_r)^t \sqrt{2 \log (\sigma_r)} \quad (4)$$

$$\sigma_r = \sigma_{\max} / \sigma_{\min}. \quad (5)$$

Then, the score in Eq.2 can be derived as follows.

$$\nabla_{x_t} \log p_t(x_t) = -\frac{x_t - \mu(x_0, y, t)}{\sigma(t)^2} \quad (6)$$

$$\mu(x_0, y, t) = e^{-\gamma t} x_0 + (1 - e^{-\gamma t} y) \quad (7)$$

$$\sigma(t)^2 = \frac{\sigma_{\min}^2 \left(\sigma_r^{2t} - e^{-2\gamma t}\right) \log (\sigma_r)}{\gamma + \log(\sigma_r)}, \quad (8)$$

where y is the sound mixed with speech and environmental noise, $\sigma_{\max}$ and $\sigma_{\min}$ are hyperparameters set to 0.5 and 0.05 respectively. γ is a hyperparameter to control the strength of the transition ("hardness") from x_0 to y and is set to 1.5. By defining the hyperparameters in this way, we obtain

$$p_{0t}(x_t | x_0, y) = N_C(x_t; \mu(x_0, y, t), \sigma(t)^2 \mathrm{I}) \quad (9)$$

where the center of x_t asymptotically moves toward y as t increases. See [12] and [18] for a more detailed explanation of these formulas.

This SGMSE+ [12] achieves higher quality than CDiffuSE [17]. Authors in SGMSE+ [12] suggest its reason is the architecture and fine hyperparameter tuning.

E. Diffusion model with latent space

The term "latent space" in AutoEncoder [19] and VAE [5] refers to the space formed by the Encoder's mapping of data. Latent Diffusion Model (LDM) [20] is a diffusion model that applies the diffusion process to the latent space formed by a VAE for image generation. It assumes that the image generation process is divided into Perceptual stage with a low compression ratio and a Semantic stage with a high compression ratio. The dimension of the latent space is compressed in the spatial direction and expanded in the channel direction. The total dimension becomes smaller because the reduction rate of the spatial dimension is greater than the expansion rate of the channel dimension.

AudioLDM [21] uses LDM to process audio data. This model is designed for Text To Audio (TTA) tasks. In this method, the diffusion model is executed in the latent space defined by VAE. It achieved State-Of-Art in the TTA. The architecture of Encoder and Decoder used in AudioLDM is shown in Fig.1. Another study [22] utilizes LDM for text-to-speech tasks and achieved state-of-the-art performance. LSGM [23] combines a score-based diffusion model with VAE for image generation tasks. Our study differs from these studies in that it is developed for speech enhancement, in which speech quality is directly evaluated by some metrics such as PESQ and ESTOI.

III. LDMSE

A. Use of Encoder-Decoder

We propose a method to reduce the computational cost of the diffusion model using a latent space with a smaller dimension than that of the original input data. We call this method Low computational cost Generative Diffusion Model for Speech Enhancement (LDMSE). This method has an encoder before the diffusion process and a decoder after it (hereafter, these two

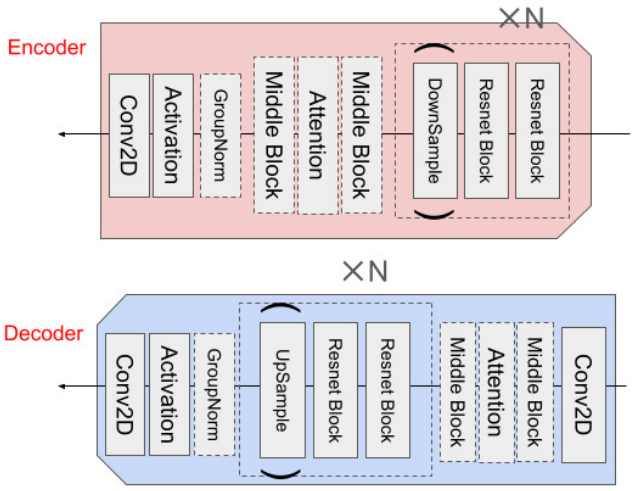


Fig. 1. Encoder-Decoder Architecture. Resnet Block is the same as AudioLDM [21]. Blocks with dashed lines mean they have been removed. Dashed rectangles denote processing groups that are executed N times. The Up/Down Sample layers in parentheses with dashed rectangles are not executed in the final N times loop.

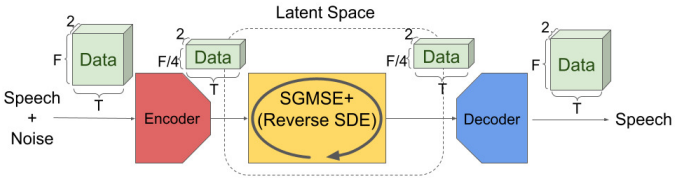


Fig. 2. Proposed process when $\xi = 1/4$

are abbreviated as EncDec). If the dimensionality of the data handled in the inverse diffusion is smaller, the computational cost will be reduced, which is the purpose of the processing process.

Let ξ represent the compression rate. Fig. 2 shows the proposed process flow when ξ is $1/4$. Also, F and T represent the number of frequency and time dimensions after STFT, respectively. The theoretical computational cost required for reverse diffusion is ξ (< 1) times lower than the Baseline.

In this study, EncDec is not a VAE but an ordinary AutoEncoder. We applied a tanh function to the Encoder's output because we found empirically that the variation of the data in the latent space would be very different depending on the seed value at the time of training if we did not do this.

The training process of LDMSE is divided into two phases. The first phase is to train an encoder to project data into the latent space and a decoder to restore it from the latent space to the real space. Only clean speech data were used for training. And loss function was L2 loss. The second stage is to fix the parameters of the EncDec and train the Diffusion Model to execute the inverse diffusion for the inference.

The EncDec was trained with 100 epochs, 8 minibatch size, and the Diffusion part was done with 300 epochs, 16 minibatch size. The learning rate in both trainings is 8.0×10^{-6}.

B. Architecture

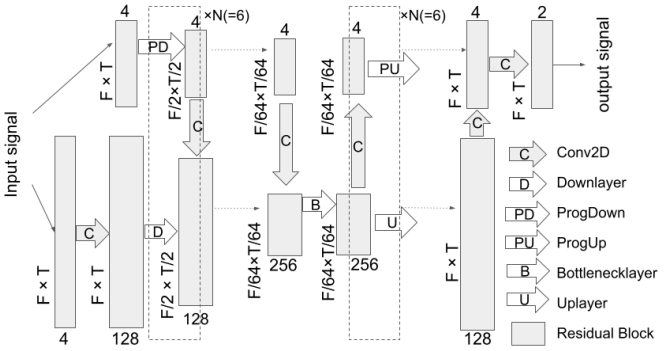


Fig. 3. Model architecture for the Diffusion part. The dimension of the input signal is $4 \times F \times T$. The numbers above or below the rectangles representing the Residual Blocks indicate the channel dimension. Four on the axis of the channel in the input dimension means the real and imaginary parts of the clean and noisy data, respectively. Its details are described in [12].

1) Diffusion part: We adopt the same architecture NCSN++ that is proposed in SGMSE+ [12] (Fig. 3). It has a similar structure to U-Net [24]. The components in Fig. 3, Residual Block, Uplayer, Downlayer, Bottlenecklayer, ProgUp, ProgDown, and Bottlenecklayer, are the same as used in NCSN++. The input and output signals are almost the same as the signals processed by STFT, but with scaling described in Sec. 3.2 in [18].

Depending on ξ, dimension-related issues may arise during the processing of NCSN++. For example, if $\xi = 1/8$, the dimension of the input data of the diffusion model becomes $F/8 \times T$, where $F \times T$ is the dimension before processing by the Encoder. Since $F = 256$ and $N = 6$ in the original SGMSE+, the dimension of the data processed in the last iteration of the iterative structure on the left side in the Fig. 3 becomes $(256/8)/(2^6) < 1$, indicating that the vector dimension becomes zero. In such cases, we refrain from further dimension reduction.

2) AutoEncoder part: The architecture of the EncDec module is adapted from AudioLDM [21]. It is shown in Fig. 1. However, the EncDec architecture utilized in LDMSE incorporates three modifications compared to AudioLDM. First, there are alterations in the architecture within the EncDec module itself. The middle block and normalization layer used in AudioLDM outside the Resnet block loop have been removed. The middle block calculates Attention for the time and frequency dimensions. This is because the computational cost of this block was very high when we used STFT as input, whereas AudioLDM used a mel-spectrogram. In our preliminary experiments, we found that EncDec's accuracy without the normalization layer was better.

Second, UpSample and DownSample are applied only to the frequency dimension and not to the time dimension. NCSN++ used in the diffusion process also compresses the time dimension. In the case of $\xi = 1/4$, the time dimension is compressed up to $1/256$ at minimum. For example, if we input data with a time dimension of 65, the time dimension becomes 32 after the first compression, as shown in Fig. 3, and 64 after

the restoration. Then, for $\xi = 1/4$, the time dimension in the output is a multiple of 256, corresponding to 2 seconds. This is too long to have a desirable time resolution in the output.

Third, the number of loops in the ResNet Block varies with ξ. For $\xi = 1/2, 1/4, 1/8$, $N = 2, 3, 4$ in Fig. 1. This is to make this architecture compatible with multiple ξ. This architecture consists of a stack of layers, each of which is a set of Down/Up Sample, Resnet Block, and Conv2D in the frequency dimension. The data is down/up sampled with each layer. Therefore, to support multiple ξ, the number of layers is changed.

IV. Experiment

A. Dataset

For our evaluation, we employ two datasets: VOICEBANK-DEMAND and WSJ0-CHiME3.

VOICEBANK-DEMAND serves as a benchmark, aligning with previous studies of speech enhancement. This is a mixture of VOICEBANK [25] and DEMAND [26]. VOICEBANK has two sets, one with 28 speakers and the other with 56 speakers, of which 28 speakers are selected. Each speaker has about 400 utterances. There are a total of 11572 utterances for training and 824 for evaluation. In the training data, each noise is mixed with one noise per utterance, with the signal-to-noise ratios of 15 dB, 10 dB, 5 dB, and 0 dB for the training data and 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB for the evaluation data, respectively. Each utterance is used exactly once.

WSJ0-CHiME3 is utilized for comparison to SGMSE+. This is a mixture of WSJ0 [27], comprising readings from the Wall Street Journal, and CHiME3 [28], prepared in the 3rd CHiME Challenge, a speech recognition competition and hereafter abbreviated as WSJ-C3. In WSJ0, the subset si_tr_s with a total of 12,776 utterances was used for training, and a subset of si_dt_05 with 1,026 utterances was used for the valid split. The signal-to-noise ratios are uniformly distributed from 0 dB to 20 dB. Each utterance is used only once. WSJ0-CHiME3 is divided into train, valid, and test sets. This paper presents the results of the evaluation of the Valication set.

We use the same consistent datasets for training and evaluation, not changing the dataset in these two stages.

B. Metrics

The Baseline for this research is SGMSE+ [12]. We mainly compare the quality with it. For the indicators of how close the emphasized speech is to the clean speech, we measured PESQ [29], ESTOI [30], SI-SDR [31]. All of these mean that the higher the value, the better the quality.

- PESQ [29]: The Perceptual Evaluation of Speech Quality. recommended by ITU-T P. 862 in 2001. The value range is from 1.0 to 4.5.
- ESTOI [30]: The Extended Short-Time Objective. The Extended Short-Time Objective Intelligibility. It takes a value range from 0.0 to 1.0.
- SI-SDR [31]: Scale-Invariant Signal-to-Distortion Ratio. It is a measure of sound distortion calculated in a way that is invariant to the volume of the two data being compared.

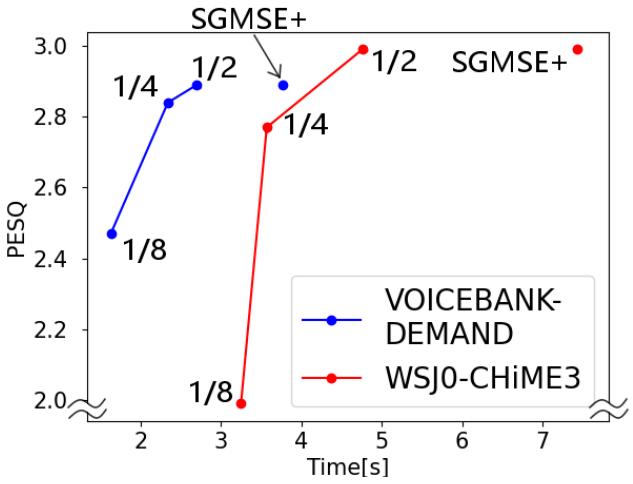


Fig. 4. Line graph for computational time and PESQ. "Time" indicates the average computational time required for speech enhancement per utterance, including the processing time of EncDec. Fractions represent the value of ξ.

We also measure the time that the GPU uses for generation. We use torch.cuda.Event for recording time. The GPU we used for generation is one RTX 3090. The CPU is Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz. The RAM has 32GB.

C. Results

TABLE I
Results on VB-DMD. "Time" in the table indicates the average time required for speech enhancement per utterance, including the processing time of EncDec. $\xi = 1/2, 1/4, 1/8$ are proposed method, LDMSE. Values after $\pm$ mean standard deviation. MGAN+ means MetricGAN+. The results of metrics and generation time in MGAN+ and CDiffuSE are from our preliminary experiment because [12] does not show the standard deviation of results.

	PESQ↑	ESTOI↑	SI-SDR↑	Time[s]
SGMSE+	2.89±0.60	0.86±0.10	16.8±3.4	3.77
$\xi = 1/2$	2.89±0.63	0.86±0.10	17.2±3.0	2.70
$\xi = 1/4$	**2.84±0.65**	**0.86±0.10**	**17.3±2.9**	**2.34**
$\xi = 1/8$	2.47±0.62	0.81±0.12	14.8±3.3	1.63
CDiffuSE	2.53±0.57	0.77±0.11	12.4±2.7	0.033
MGAN+	3.15±0.57	0.82±0.12	8.0±3.2	0.0012

TABLE II
Results on WSJ-C3. The meanings of each value and word are the same as in VB-DMD. The results of metrics in MGAN+(=MetricGAN+) and CDiffuSE are reported value in [12].

	PESQ↑	ESTOI↑	SI-SDR↑	Time[s]
SGMSE+	2.99±0.55	0.91±0.07	18.2±4.3	7.43
$\xi = 1/2$	**2.99±0.55**	**0.91±0.06**	**17.6±4.0**	**4.76**
$\xi = 1/4$	2.77±0.57	0.90±0.07	17.2±3.5	3.57
$\xi = 1/8$	1.99±0.41	0.84±0.09	12.4±3.6	3.25
CDiffuSE	2.27±0.51	0.83±0.09	9.2±2.3	0.099
MGAN+	3.03±0.45	0.88±0.08	10.5±4.5	0.0024

Fig. 4 shows the comparison of our proposed methods, LDMSE, with different ξ values. For $\xi = 1/2$, our method

succeeded in reducing the time without losing quality in VB-DMD, and in $\xi = 1/4$ with WSJ-C3 it did not lose quality in the $\xi = 1/2$ case, nor did it lose much. In these ξ, the effectiveness of the proposed method was confirmed. For $\xi = 1/8$, the quality has been noticeably reduced for both data sets. This may be because the degree of compression was too large, and thus, the necessary dimension in the real space was not obtained in the latent space.

Table I and II show the detailed results of the quality evaluation of speech enhancement for VB-DMD and WSJ-C3, respectively. The PESQ and Time values in Table I and II are the same as the values in Fig. 4. Results of MetricGAN+ and CDiffuSE are shown for comparison. With SI-SDR and ESTOI, up to $\xi = 1/4$, the speedup was achieved without losing quality. The bold values in Table I, II are those considered to have been particularly successful. Compared to the other methods, LDMSE is slower than CDiffuSE in generation time but better in quality. Compared to CDiffuSE, LDMSE outperforms in quality in SI-SDR and ESTOI, while MetricGAN+ outperforms in quality in PESQ. This may be because the objective of MetricGAN+ during training is to maximize PESQ. In all quality metrics, there is no significant difference in t-test at a 5% significance level.

V. Conclusion

We have proposed a method called LDMSE to reduce the computational cost of diffusion model based speech enhancement, it is shown that the quality of the diffusion model with SGMSE+ can be maintained while reducing the computational cost by using the latent space. In particular, if it is $\xi = 1/4$ in VB-DMD and $\xi = 1/2$ in WSJ-C3 we can achieve a speedup of more than 35% without quality loss.

In the future, we would like to aim to use VAE for EncDec [20], [21]. Additionally, we would like to investigate the degree of data dispersion in latent space. Finally, we want to examine how we can apply the proposed method to other diffusion model-based denoising methods, such as CDiffuSE, MGAN+, and complex number-based frameworks.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979. DOI: 10.1109/TASSP.1979.1163209.

[2] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. DOI: 10.1109/5.18626.

[3] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016. DOI: 10.1109/TASLP.2015.2512042.

[4] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," Sep. 2017, pp. 1–6. DOI: 10.1109/MLSP.2017.8168119.

[5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. arXiv: http://arxiv.org/abs/1312.6114v10 [stat.ML].

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14, Montreal, Canada: MIT Press, 2014, pp. 2672–2680.

[7] S.-W. Fu, C. Yu, T.-A. Hsieh, *et al.*, "Metricgan+: An improved version of metricgan for speech enhancement," Aug. 2021, pp. 201–205. DOI: 10.21437/Interspeech.2021-599.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20, Vancouver, BC, Canada: Curran Associates Inc., 2020, ISBN: 9781713829546.

[9] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS.

[10] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982, ISSN: 0304-4149. DOI: https://doi.org/10.1016/0304-4149(82)90051-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0304414982900515.

[11] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations* (Stochastic Modelling and Applied Probability). Springer Berlin Heidelberg, 2011, ISBN: 9783540540625. [Online]. Available: https://books.google.co.jp/books?id=BCvtssom1CMC.

[12] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023. DOI: 10.1109/TASLP.2023.3285241.

[13] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP.

[14] D. Watson, W. Chan, J. Ho, and M. Norouzi, "Learning fast samplers for diffusion models by differentiating through sample quality," in *International Conference*

on Learning Representations, 2022. [Online]. Available: https://openreview.net/forum?id=VFBjuF8HEp.

[15] B. Kim and J. C. Ye, "Denoising MCMC for accelerating diffusion-based generative models," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 23–29 Jul 2023, pp. 16 955–16 977. [Online]. Available: https://proceedings.mlr.press/v202/kim23z.html.

[16] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, pp. 146–152. DOI: 10.21437/SSW.2016-24.

[17] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7402–7406. DOI: 10.1109/ICASSP43922.2022.9746901.

[18] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Interspeech 2022*, 2022, pp. 2928–2932. DOI: 10.21437/Interspeech.2022-10653.

[19] D. Bank, N. Koenigstein, and R. Giryes, *Autoencoders*, Mar. 2020.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 10 684–10 695.

[21] H. Liu, Z. Chen, Y. Yuan, *et al.*, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 23–29 Jul 2023, pp. 21 450–21 474. [Online]. Available: https://proceedings.mlr.press/v202/liu23f.html.

[22] Z. Liu, Y. Guo, and K. Yu, "Diffvoice: Text-to-speech with latent diffusion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095100.

[23] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=P9TYG0j-wtG.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[25] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4. DOI: 10.1109/ICSDA.2013.6709856.

[26] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 035 081, Jun. 2013, ISSN: 1939-800X. DOI: 10.1121/1.4799597. eprint: https://pubs.aip.org/asa/poma/article-pdf/doi/10.1121/1.4799597/14782232/035081_1_online.pdf. [Online]. Available: https://doi.org/10.1121/1.4799597.

[27] e. a. P. L. D. C. Garofolo John S., *Csr-i (wsj0) complete ldc93s6a. web download*, Sep. 1993. DOI: 10.1109/MLSP.2018.8516711.

[28] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511. DOI: 10.1109/ASRU.2015.7404837.

[29] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, 749–752 vol.2. DOI: 10.1109/ICASSP.2001.941023.

[30] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016. DOI: 10.1109/TASLP.2016.2585878.

[31] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630. DOI: 10.1109/ICASSP.2019.8683855.

Continual Learning with Self-Organizing Maps: A Novel Group-Based Unsupervised Sequential Training Approach

Gaurav Hirani[1]*, Kevin I-Kai Wang[2] and Waleed Abdulla[3]
[123]The University of Auckland, Auckland, New Zealand
[1]E-mail: ghir325@aucklanduni.ac.nz Tel: +64 22 307 2932
[2]E-mail: kevin.wang@auckland.ac.nz Tel: +64 (9) 373 7599 Ext.89226
[3]E-mail: w.abdulla@auckland.ac.nz Tel: +64 (9) 373 7599 Ext.88969

Abstract – Continual learning is essential for intelligent systems to thrive. While deep learning models based on Convolutional Neural Networks are robust, they can be complex and resource-intensive, requiring a long and computationally intensive training phase due to Back-propagation (BP). On the other hand, unsupervised methods are favored for their lightweight, speed, and ease of management, making them ideal for ongoing learning. Self-organizing maps (SOMs) have gained attention in computer vision over the past decade due to their excellent feature extraction abilities and transparency. However, SOMs encounter challenges in learning, such as having a fixed map size that limits new feature accommodation and the risk of interference when retraining a pre-trained map with the latest data. Although some studies have explored using SOMs to tackle these issues, obstacles like catastrophic forgetting and lifelong learning hurdles remain prevalent. To combat forgetting issues, we propose an approach called learning by grouping. In this method, each group generates a trained map stored in memory before the model is reset and trained on the group to create another map. This sequential process facilitates learning while preserving previously acquired knowledge without disruptions. The suggested approach has been tried out on MNIST digit, Fashion MNIST, and CIFAR 10 datasets. The proposed method demonstrates final accuracy improvement of 60-70% across all three datasets when compared to traditional SOM, indicating the effectiveness of grouping in overcoming the limitations of SOM-based continual learning methods.

I. INTRODUCTION

Convolutional Neural Network (CNN)-based deep architectures have achieved significant success in the field of computer vision, demonstrating exceptional efficiency across various applications [1] [2]. These deep learning (DL) models are typically trained on labeled data for all available classes or target groups. When new data from a previously unseen class is introduced, the existing trained model cannot adapt without requiring a complete retraining of the entire architecture. The capability to continuously learn over time by incorporating new knowledge while retaining previously acquired experiences is known as continual or lifelong learning [3]. Unlike humans, who seamlessly integrate new information while preserving old knowledge, traditional algorithms face challenges in this area. The complexity and large size of DL architectures result in substantial time and computational power requirements for retraining.

To address the challenge of making deep architectures adaptive, significant approaches have been proposed for continual learning (CL) [4]. Models designed with CL functionality to prevent forgetting old tasks are primarily categorized into three types: retraining with regularization, additional training with network expansion, and selective network retraining and expansion [5] [6]. Introducing new sub-networks prevent catastrophic forgetting, the parameters learned for existing tasks remain unchanged, while new parameters are added and trained [7]. Combining both concepts in a balanced manner constitutes the third category of architecture. Additional categories include replay-based approaches, optimization-based approaches, representation-based approaches, and memory-based methods [4] [8]. These advancements have predominantly been applied to supervised models, particularly deep neural networks. DL models often suffer from limited tractability, making their outcomes difficult to interpret, which poses a significant bottleneck to major improvements. To transparently study the effects of adaptive learning, a few studies have focused on unsupervised approaches. Although these models are not as efficient as their supervised counterparts for large datasets, they are lightweight, tractable, and faster.

Self-organizing map (SOM) is an unsupervised clustering technique widely used for dimensionality reduction, feature extraction, and clustering in image datasets. Several key studies have attempted to implement SOM for continual learning. Two SOM-based models were proposed for incremental learning: (i) a modified self-organizing map (GeppNet) and (ii) a SOM extended with a short-term memory (GeppNet+STM) [3][9]. In GeppNet, task-relevant feedback from a regression layer determines whether learning in the self-organizing hidden layer should occur. In GeppNet+STM, the STM stores new knowledge and occasionally replays it to the GeppNet layer during sleep phases interleaved with training phases. This latter approach demonstrated better performance and faster convergence in incremental learning tasks with the MNIST dataset. However, the STM's limited capacity means that new knowledge can overwrite old information [10]. To address the significant memory requirements, a SOM-based memory-less architecture (SOMLP) was proposed, which combines standard

supervised neural networks with SOM to tackle the continual learning problem [11]. Another approach, Dendritic SOM (DendSOM), was proposed as a single-layer SOM for feature extraction using a set of hit-matrices [12]. The input image was divided into patches and projected onto independent SOMs, each accompanied by hit-matrices to associate labels with the units. However, the formation of hit-matrices remains opaque [12]. Additionally, a memory-based model called continual SOM (CSOM) was introduced to generalize the standard SOM with minimal memory allocation. The structural design of DendSOM was adopted to manage the spatial dimension of the data [13]. Despite these advancements, all these models suffer from the stability–plasticity dilemma, a well-known issue in lifelong learning aimed at preventing catastrophic forgetting in computational models [14] [15].

Summarizing the research on SOM, it is evident that SOM offers greater tractability in the learning process with new information. However, the introduction of additional classes necessitates excessively large maps for accommodation, leading to computational inefficiency and reduced feasibility. For self-organizing networks, catastrophic interference escalates rapidly with oversized maps, leading to performance degradation. Even stringent predefined map size limitations hinder the integration of new data eventually. Irrespective of map size, retraining exacerbates forgetting, as new features overwrite existing ones. To address these issues, this study proposes handling data in groups and processing them sequentially, "storing" previously learned maps before using the model for the next group. The architecture is a memory-based model that produces a bank of trained map weights as the outcome. The core contributions of the study are outlined below.

1. The proposed model handles new data by creating a new group and producing a new map without updating or interfering with the old maps. This approach eliminates the issue of forgetting, as no new information overwrites the old ones. As each map only represent fewer class, larger map size is not required.
2. The testing process only utilizes trained weight and labeled array to predict the label which makes the SOM available to get trained on new data with a new class independently and simultaneously.
3. The model offers the flexibility to be retrained on new data for which it has already been trained without interfering with any other group. This is achieved by using the current trained weights as filter initialization and performing training on the new data.

The rest of the paper is structured as follows: Section II explains the proposed methods including training, labelling and testing phases. Experimental details and results are provided in section III. Section IV has discussion and comparison of the results to the close studies. The paper is concluded with conclusion and future work in section V.

II. PROPOSED MODEL

The proposed method utilizes the datasets into groups and the concept is explained in the first subsection. The following subsections describe training of the SOM, labelling and testing.

A. Concept of Grouping

Traditionally, the training of nearly all machine learning models involves using the entire training dataset, including SOM. However, when the complete training dataset encompassing all labels is not initially available and new data associated with new classes is introduced over time, recent studies on SOM for continual learning employ various strategies to minimize the impact on previously trained weights. The challenge with SOM lies in its constrained output space: the map size is fixed initially, and adding more nodes to a 2D structure in terms of columns and rows can distort cluster shapes. New clusters for new data cannot simply be linear, requiring them to occupy space previously learned by other samples. The SOM's neighborhood learning attempts to symmetrically shape clusters, but whether the map size is increased or not, the process of retraining with new data inevitably affects existing knowledge. As more data arrives, the phenomenon of catastrophic forgetting becomes unavoidable. Using an excessively large map from the outset also fails to resolve the problem, as it would exponentially increase training time, and at some point, it would hit the bottleneck of accommodating more classes. Also, there is no surety of preventing interference between previously learned nodes and new ones. The issue of uncontrollable learning in SOM stems from the absence of a clear objective function for managing map updates.

To address the issue of forgetting, we propose a novel approach that involves processing the initially available data in groups rather than as a single set. These groups consist of samples belonging to specific labels, although the labels themselves are not used during training to maintain an unsupervised approach. Initially, the labels are only employed to partition the large dataset into these groups and can be considered as data-preprocessing. Fig. 1 illustrates the overall structure of our proposed method for handling data over time. Each group includes datasets associated with two different labels. Number of classes per group is a hyperparameter, and chosen as two to facilitate comparison with previous studies.

During the training phase, labels are not employed. As depicted in Fig. 1, after training each group, the corresponding map is generated and stored in a memory bank. Subsequently, the weights are re-initialized for the next group, and a new map is generated and stored in the same manner. This process iterates through all the groups. Once training is completed, the labeling process utilizes these trained maps alongside labels to create labeled arrays. In the testing phase, predictions are made using the trained maps, labeled arrays, and test samples with their actual labels, allowing comparison between predicted and actual labels. Detailed explanations of the training, labeling, and testing phases follow in subsequent subsections.

B. Training of the SOM

The SOM training follows the original method introduced by Kohonen [16]. As explained in previous subsection, the input dataset is divided into groups. Each group sequentially trains the SOM according to Algorithm 1 without any labeled information. As illustrated in Fig. 2, the Euclidean distance is computed from

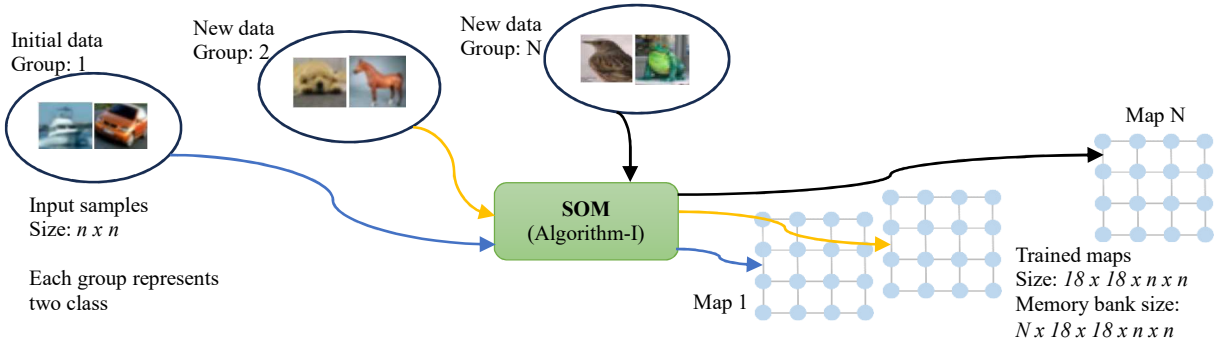


Fig. 1. The overview of the proposed structure. Introduction of new data in groups and generation of trained weights to store as learnt maps.

each image to all nodes within the assigned map. The node with the shortest distance is identified as the Best Matching Unit (BMU), representing the node most closely resembling the image in terms of data distribution or pattern similarity (lines 3-7, Algorithm I).

Once the BMU is determined, the next step involves updating the weights to move the BMU closer to the input image. Simultaneously, the weights of neighboring nodes to the BMU are adjusted based on a neighborhood function (lines 8-10, Algorithm I). $N(t)$ defines the extent of the neighborhood, which decreases over epochs; roughly halfway through training, only the BMU undergoes weight updates.

The training concludes upon reaching the maximum number of epochs. The main hyperparameters for SOM are map size, maximum number of Epochs (t), which defines the training cycles, learning rate $\eta(0)$ and initial neighborhood width or radius $\delta(0)$. η and δ are linear and monotonically decreasing functions respectively (lines 13 and 14, Algorithm I). The Fig. 1 mimics training process as continual learning which generates map banks as outcome.

TABLE I
TRAINING OF SOM

	Algorithm I. SOM Training
	Inputs: A batch of training dataset (No labels)
	Output: Trained SOMs
1:	*Weight Initialization:* Random [0,1]
2:	**for** t in epochs
3:	**for** s in train samples per group
	//find the BMU
4:	**for** i in nodes in SOM
5:	calculate $d(s, w_i)$ //$d(s, w_i) = \|\|s - w_i\|\|^2$
6:	**end**
7:	$\arg\min_i d(s, w_i)$ → BMU // n_{BMU}
	// weight update
8:	**for** i in nodes in SOM
9:	Compute the neighborhood function $N(t)$
	// $N(t) = exp\left(-\frac{\|n_{BMU} - n\|^2}{2\delta(t)^2}\right)$
10:	Weight update $w_i(t+1)$
	//$w_i(t+1) = w_i(t) + \eta(t)N(t)(s - w_i(t))$
11:	**end**
12:	**end**
13:	Update learning rate $\eta(t)$// $\eta(t) = 0.49\left(1 - \frac{t}{epoch}\right) + 0.01$
14:	Update neighborhood width $\delta(t) = \delta(0)\, exp^{\left(-\frac{t}{epoch}\right)}$
15:	**end**

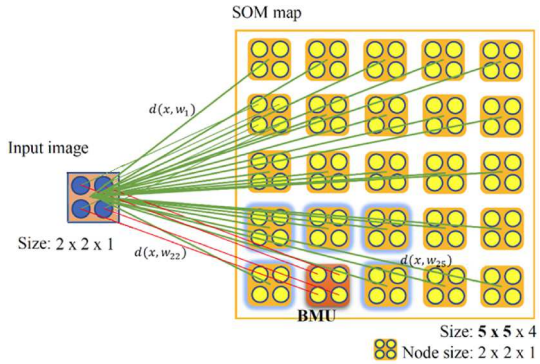


Fig. 2. BMU selection in SOM. For demonstration, node 22 is depicted as the BMU for the highlighted patch considering $d(x, w_{22})$ is the minimum distance. The neighborhood nodes of the BMU are highlighted with blue backlight.

C. *Labelling of the arrays*

The labeling process commences immediately after the training phase concludes, following the steps outlined in Algorithm II. The inputs to these sections include subsets of the labeled training dataset and the trained SOM maps. These inputs generate labeled arrays as outputs, where each array aligns spatially with the SOM map grid. Each element within the array functions as a unit storing a label represented by the node at the corresponding location on the SOM map.

TABLE II
LABELING OF THE ARRAYS

	Algorithm II. SOM Labeling
	Inputs: A subset (s) of the batch of the training dataset (with labels) and Trained SOMs
	Output: Arrays with label information
1:	The number of patches (=*SOMs)* defines the number of arrays.
2:	**for** s in a subset of train samples
	// Find the BMU with designated SOM(k)
3:	**for** n in group branches
4:	Index (BMU) → j
5:	Label[s] → array[n, j]
6:	**end**
7:	**end**

For each image, the Best Matching Unit (BMU) is identified using the trained map based on Euclidean distance. Once the BMU is determined, its index is recorded, and the corresponding label for that sample is assigned to the corresponding index in the respective array. Each group maintains its own map and associated array to represent the labels associated with that map.

Unlike during training, no further weight updates occur once the BMU is determined. Each sample is processed once, and this process repeats for all samples. As each group is sequentially trained over time, the memory bank accumulates trained maps along with their respective labeled arrays.

D. Testing and decision making

The model evaluation utilizes the test dataset, which includes labels. Each test image is projected onto all trained maps. We have implemented a multi-BMU technique, where instead of identifying just one closest distance, we determine the three closest distances per map. These distances are summed for each map, and the map with the smallest total distance determines the winning map and corresponding group. The three node indexes contributing to the decision of the winning group are recorded, and the labels associated with these indexes are extracted from the array belonging to the winning group. The mode of these three labels is then compared against the actual label to assess the model's performance. Since each map and array represent only two classes, the mode operation consistently identifies a clear winner.

III. EXPERIMENTS

This section addresses the datasets utilized in the experiments, the hyperparameters selection and evaluation metrices. The algorithm implementation was carried out in Python 3.8.8, utilizing the MiniSOM library for SOM calculations within the Keras framework [17]. The experiment was conducted on a Core i7-10700 processor without a GPU acceleration.

A. Datasets

The experiments utilized three datasets: MNIST-digit [18], Fashion-MNIST [19] and CIFAR-10 [20]. The data characteristics are detailed in Table III. MNIST-digit and Fashion-MNIST datasets consist of 8-bit grayscale images depicting digits and clothing items, respectively. Both datasets have comparable sizes for their training and testing subsets, with images uniformly sized at 28 x 28 pixels. In contrast, CIFAR-10 is a more complex dataset containing 24-bit color images representing real-world objects. Each image in CIFAR-10 measures 32 x 32 pixels, with objects randomly positioned within the spatial domain.

As part of the data preparation, the dataset undergoes normalization. Furthermore, the conversion of RGB to grayscale is achieved by assigning equal weight to each color channel. No additional data processing steps are applied. Training exclusively utilizes the clean training dataset.

For the colored datasets like CIFAR-10, using only the blue channel for both training and testing results in a performance enhancement of 2 - 2.5% compared to using grayscale or other channels. This finding was incorporated into subsequent experiments on these datasets, where all operations were conducted exclusively using the blue channels.

B. Hyperparameters selection

The experiments primarily involve two types of hyperparameters: 1) Key SOM parameters, as detailed in section II, and 2) parameters related to data selection. The SOM parameters used in the current experiments are specified in Table IV. As the study does not focus extensively on fine-tuning hyperparameters and their impact on model performance, they undergo moderate tuning and are selected following a few trial-and-error iterations.

C. Performance metrics

1. **Accuracy:** The model's performance is assessed as groups are sequentially added to the model, with each addition considered a task. Initially, the model begins with only task 1, where it is trained using the first group of data. Adding subsequent groups results in tasks labeled as Nth, corresponding to the Nth group. Two types of performance metrics are observed:
   1) *Task-specific accuracy*: it evaluates performance using only the test dataset from the current task against all previously learned maps [12],

$$Accuracy_t = \frac{1}{N_t} \sum_{n=1}^{N_t} AL_n \cap PL_n \quad (1)$$

Where, N_t is the size of the test dataset per task, AL_n is actual label, PL_n is the predicted label over all previous learnt maps.

2) *Average accuracy:* The performance on task j after applying it on all previous task n_j is denoted as $a_{n_j,j}$. The average accuracy at task T after testing it on all previous tasks from 1 to T-1 is as follows [21]:

$$Average\ Accuracy_T = \frac{1}{T} \sum_{j=1}^{T} a_{n_j,j} \quad (2)$$

2. **Average Interference:** It is defined as the average decrease in performance for each task from its highest accuracy to the accuracy attained at the end of training, computed as follows: [21]:

$$F_T = \frac{1}{T-1} \sum_{j=1}^{T-1} f_T^j \quad (3)$$

Where, f_T^j is forgetting term for task j. after model is applied on task T and is defined as follows:

$$f_T^j = \max_{\tau=[1,2,\dots T]} \left(a_{\tau,j} - a_{T,j}\right) \quad (4)$$

TABLE III
DATASET STATISTICS

Dataset	Train dataset	Classes	Images per class	Test dataset	Size of a sample	Used train dataset for training	Used train dataset for labeling	Used test dataset for testing
MNIST-digit	60000	10	6000	10000	28 x 28	4000	4000	10000
Fashion-MNIST	60000	10	6000	10000	28 x 28	4000	4000	10000
CIFAR-10 (RGB)	50000	10	5000	10000	32 x 32 x 3	7000	7000	10000

TABLE IV
HYPERPARAMETERS FOR THE PROPOSED STRUCTURE

Parameter	Value
Class per group	2
Number of SOM per dataset	5
SOM Map Size	15 x 15 x 28 x 28
	15 x 15 x 32 x 32*
Array size	15 x 15
Iteration	27000
Neighborhood radius	1
Learning rate	0.5

*Denotes the modification of hyperparameter for CIFAR-10

IV. RESULTS AND DISCUSSION

The model was applied to all three datasets using the hyperparameters outlined in Table IV. The accuracy for the task-specific dataset (Equation 1) and the overall average accuracy across all test datasets up to a given task (Equation 2) were recorded at each task, as detailed in Table V. With each task, a group representing two classes is added to the model, increasing the size of the memory bank by one map and one array. During the decision-making process at each task, all maps learned thus far, including the current one, are considered to determine the winning group.

A low task-specific accuracy at a particular task indicates that the group containing two classes has poorly learned a map, which does not emerge as the winner in the decision-making process. For instance, Group 5 for MNIST-digit, Group 4 for Fashion-MNIST, and Group 3 for CIFAR-10 were identified as the worst performers individually.

Typically, forgetting occurs when new features learned from new data overwrite or modify previously learned information. Previous SOM-based studies have employed similar maps to accommodate new knowledge or have retrained models using previously learned maps as initial weight states. However, in our approach, when new data is introduced, learning results in the creation of a new map while leaving old maps untouched. Consequently, there is technically no forgetting or interference during the training process itself. However, during the decision-making phase, all trained maps must be considered to identify the most closely related map to the sample. If multiple maps representing classes closely resemble each other, an incorrect group might be selected as the winner, which fits the definition of "interference." This can lead to a decline in accuracy at each task. The performance based average interference or forgetting

TABLE V
TASK SPECIFIC ACCURACY OVER PREVIOUSLY LEARNT TASKS

	MNIST-digit	Fashion-MNIST	CIFAR-10
Task 1	100%	97.44%	75.05%
Task 2	97.55%	92%	47.77%
Task 3	98.11%	91.16%	25.22%
Task 4	96.56%	73.16%	35.72%
Task 5	92.4%	91.33%	37.11%

AVERAGE ACCURACY OVER PREVIOUSLY LEARNT TASKS

	MNIST-digit	Fashion-MNIST	CIFAR-10
Task 1	100%	97.44%	75.05%
Task 2	98.78%	94.72%	60.16%
Task 3	98.55%	93.53%	48.51%
Task 4	98.06%	83.44%	45.32%
Task 5	96.92%	84.14%	43.67%

TABLE VI
AVERAGE INTERFERENCE AT EACH TASK

	MNIST-digit	Fashion-MNIST	CIFAR-10
Task 1	0.00%	0.00%	0.00%
Task 2	0.08%	3.68%	7.9%
Task 3	1.91%	12.22%	13.25%
Task 4	2.07%	19.48%	36.38%
Task 5	2.21%	18.19%	39.2%

(Equation 3) is recorded in Table VI. At the first task, there is only one map and no interference in selecting winner group.

The results obtained from our study were compared with previous SOM-based research. Fig. 3. displays the overall average accuracy at each task for MNIST-digit, with numbers extracted from relevant papers. Our proposed study achieves the highest accuracy throughout all tasks, including the final task encompassing all classes. Regarding Fashion-MNIST, we did not find specific results from the mentioned studies. However, results from previous studies on CIFAR-10 have been plotted alongside our proposed study in Fig. 4. DendSOM performs better for all tasks due to its effective patch-based learning for much diverse CIFAR-10. However, by the end of Task 5, the proposed method catches up with DendSOM, suggesting lower interference during incremental learning. It is worth noting that our proposed approach utilizes a single SOM model, whereas DendSOM employs 225 SOM models. Encompassing all classes, a single tradition SOM attains an accuracy of 58.95% for the MNIST-digit dataset, 52% for Fashion-MNIST, and 25.7% for CIFAR-10. In contrast, the proposed method achieves 96.92%, 84.14%, and 43.67% for these datasets, respectively. This indicates an improvement in final accuracy by grouping over the traditional SOM by 60-70%.

In summary, the proposed approach of learning classes in groups and storing learned maps in a memory bank effectively eliminates the limitation of catastrophic forgetting. By utilizing a single SOM that is reinitialized and trained with new data, the model achieves higher efficiency and lower computational load compared to previous methods. This method also supports lifelong learning, where each new group of data adds a map and a labeled array to the memory bank for future use.

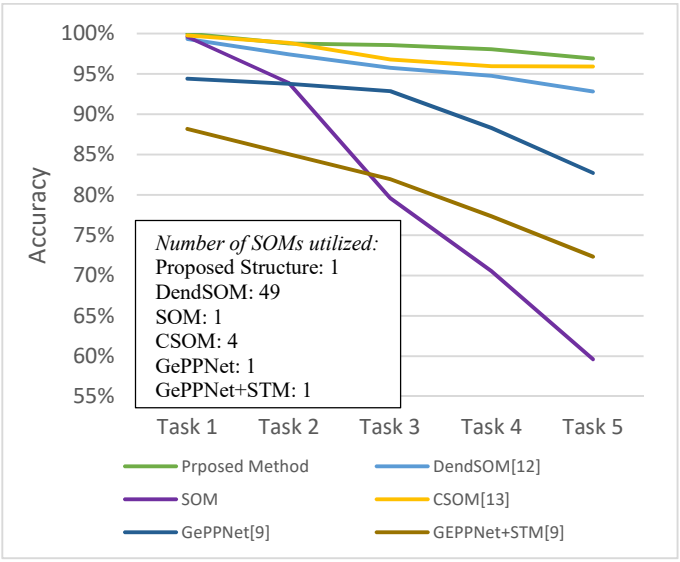


Fig. 3. Average accuracy for MNIST-digit per task

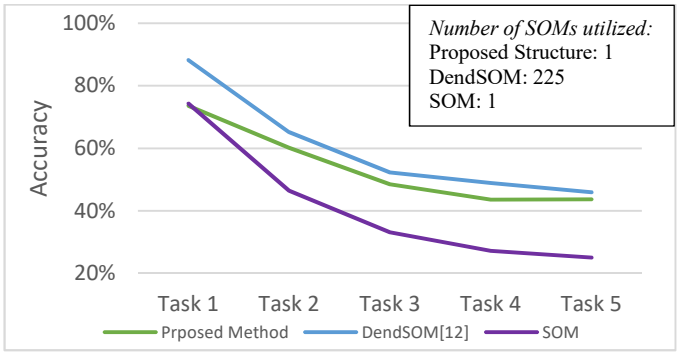


Fig. 4. Average accuracy for CIFAR-10 per task

For groups with additional available data, the SOM can be initialized with the existing trained weights and further trained with new data to enhance the map for that group. This is crucial if a group map performs poorly providing BMUs for own test dataset (e.g., Group 4 for Fashion-MNIST). The labeling and testing phases only require the trained weights and labels, not the entire model itself, which ensures independence between training and real-time testing operation parallelly.

An additional advantage is that labeled arrays can be optimized independently without modifying the weights, potentially improving accuracy. One limitation associated with the grouping concept is interference in winning group selection. As the number of maps increases, traditional distance measurement techniques become less effective in accurately identifying the correct group, which can result in incorrect predictions. To mitigate this issue, we intend to explore the implementation of additional metrics alongside Euclidean distance to determine the winning nodes.

V. Conclusion

This research presents a back-propagation free, unsupervised memory-based approach that utilizes a single Self-Organizing Map (SOM) as a tool for continuous learning. Unlike SOM-based approaches, this method offers several key benefits;

1. Prevention of Forgetting: Training individual maps helps prevent the issue of forgetting.

2. Separate Testing: This allows the model to learn from data while being used in real-time applications.

3. Efficiency: The model operates faster due to its design.

The model adjusts to data using multiple maps while keeping previously learned maps intact. Each new dataset creates a map without impacting existing ones. During testing, only the learned maps and labeled arrays are required to predict labels, enabling the SOM model to train on available data with fresh class labels simultaneously with testing phases.

This study confirms the effectiveness of group learning utilizing a Vanilla SOM, demonstrating substantially faster processing times compared to studies that employ multiple SOMs and various data processing modifications. Upon presenting all classes in the dataset, the proposed methods achieved accuracies of 96.92% for MNIST-digit, 84.14% for Fashion MNIST, and 43.67% for CIFAR-10, representing an improvement of nearly 60-70% over traditional SOM.

Future research will concentrate on improving class selection within each group and introducing methods to improve the correct group selections.

VI. Reference

[1] A. Khan *et al*, "A survey of the recent architectures of deep convolutional neural networks," *Artif Intell Rev,* vol. 53, *(8),* pp. 5455, 2020. DOI: 10.1007/s10462-020-09825-6.

[2] A. Shrestha and A. Mahmood, "Review of Deep Learning Algorithms and Architectures," *IEEE Access,* vol. 7, pp. 53040, 2019. DOI: 10.1109/access.2019.2912200

[3] G. I. Parisi *et al*, "Continual lifelong learning with neural networks: A review," *Neural Networks,* vol. 113, pp. 54, 2019. DOI: 10.1016/j.neunet.2019.01.012.

[4] L. Wang *et al*, "A Comprehensive Survey of Continual Learning: Theory, Method and Application," *IEEE Trans. Pattern Anal. Mach. Intell.,* pp. 1, 2024. DOI: 10.1109/tpami.2024.3367329.

[5] M. K. Benna and S. Fusi, "Computational principles of synaptic memory consolidation," *Nat Neurosci,* vol. 19, *(12),* pp. 1697, 2016. DOI: 10.1038/nn.4401.

[6] G. Hinton, O. Vinyals and J. Dean, "Distilling the Knowledge in a Neural Network," 2015.

[7] A. A. Rusu *et al*, "Progressive Neural Networks," *arXiv Preprint arXiv:1606.04671, 2016*.

[8] B. L. Üders, M. Schläger and S. Risi, "Continual Learning through Evolvable Neural Turing Machines," 2016.

[9] A. Gepperth and C. Karaoguz, "A Bio-Inspired Incremental Learning Architecture for Applied Perceptual Problems," *Cogn Comput,* vol. 8, *(5),* pp. 924, 2016. DOI: 10.1007/s12559-016-9389-5.

[10] A. Gepperth and C. Karaoguz, "A Bio-Inspired Incremental Learning Architecture for Applied Perceptual Problems," *Cogn Comput,* vol. 8, *(5),* pp. 924, 2016. DOI: 10.1007/s12559-016-9389-5.

[11] P. Bashivan *et al*, "Continual learning with self-organizing maps," *arXiv Preprint arXiv:1904.09330,* 2019.

[12] K. Pinitas, S. Chavlis and P. Poirazi, "Dendritic Self-Organizing Maps for Continual Learning," *arXiv Preprint arXiv:2110.13611,* 2021.

[13] H. Vaidya *et al*, "Neuro-mimetic Task-free Unsupervised Online Learning with Continual Self-Organizing Maps," *arXiv Preprint arXiv:2402.12465,* 2024.

[14] G. I. Parisi *et al*, "Lifelong learning of human actions with deep neural network self-organization," *Neural Networks,* vol. 96, pp. 137, 2017. DOI: 10.1016/j.neunet.2017.09.001.

[15] F. M. Richardson and M. S. C. Thomas, "Critical periods and catastrophic interference effects in the development of self-organizing feature maps," *Developmental Science,* vol. 11, *(3),* pp. 371, 2008. DOI: 10.1111/j.1467-7687.2008.00682.x.

[16] T. Kohonen, "The self-organizing map," *Proc IEEE,* vol. 78, *(9),* pp. 1464-1480, 1990.

[17] V. Giuseppe, "MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map," unpublished.

[18] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[19] H. Xiao, K. Rasul and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," 2017.

[20] Alex Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," 2009. Available: http https://www.cs.toronto.edu/~kriz/cifar.html

[21] P. Kumar and D. Toshniwal, "Lifelong learning gets better with MixUp and unsupervised continual representation," *Appl Intell,* vol. 54, *(7),* pp. 5235, 2024. DOI: 10.1007/s10489-024-05434-w

Improving Semi-Supervised Object Detection by ROI-Enhanced Contrastive Learning

Teng-Kuan Huang and Mei-Chen Yeh
Department of Computer Science and Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
E-mail: kensai1402@yahoo.com.tw, myeh@ntnu.edu.tw

***Abstract*—Semi-supervised object detection has emerged as a promising paradigm to alleviate the data annotation burden by utilizing a small set of labeled data in conjunction with a larger pool of unlabeled data. Current state-of-the-art methods commonly employ self-training strategies, using pseudo labels to learn from unlabeled data. However, pseudo labels are inherently noisy, particularly in the early stages of training. In this paper, we propose a contrastive learning approach to enhance semi-supervised object detection. Departing from conventional box-level predictions, our method introduces consistency regularization at the feature-level representation. Specifically, we leverage candidate boxes selected by the Region Proposal Network (RPN) for Region of Interest (RoI)-based contrastive learning and introduce pixel-level comparisons for spatial-aware loss calculation. Our experiments demonstrate that the proposed RoI-enhanced contrastive learning effectively enables the model to extract additional information from unlabeled data.**

I. Introduction

In recent years, remarkable progress in object detection has been driven by supervised learning methods, leveraging abundant labeled training data [1]. Yet, the resource-intensive process of obtaining accurate annotations has led to a growing interest in semi-supervised approaches. These approaches combine limited labeled data with a wealth of unlabeled data to optimize object detection performance [2], [3]. Within this domain, the Teacher-Student Mutual Learning framework, renowned for its effectiveness in self-training for semi-supervised object detection [4], has gained significant attention.

Despite the various methods for utilizing pseudo-labels generated by evolving teacher models, their universal applicability across diverse datasets remains a challenge, encompassing fixed confidence thresholds to sophisticated filtering mechanisms [5]–[7]. This paper proposes an alternative solution for leveraging unlabeled data in semi-supervised object detection.

The Teacher-Student framework introduces varying degrees of visual augmentation to unlabeled data, creating both strongly and weakly augmented image versions. Pseudo-labels are generated by the teacher model on weakly augmented images, while the student model learns from filtered pseudo-labels on strongly augmented counterparts. This process mirrors the contrastive learning approach commonly used in self-supervised learning, such as BYOL [8] and SimpleCL [9]. The distinction lies in the computation of the supervised loss or the similarity—where the former calculates the loss from prediction boxes, the latter does so from feature-level representations, as illustrated in Fig. 1.

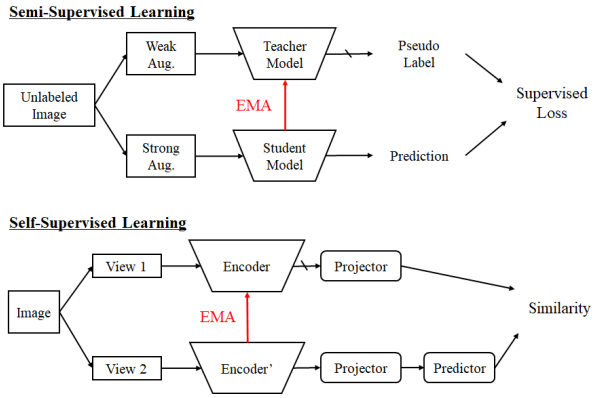


Fig. 1. Analogy between Semi-Supervised Learning and Self-Supervised Learning.

Despite the commonality in the self-training process, pseudo-labels generated by the teacher model inevitably introduce label noise. Nevertheless, deep neural networks have shown the ability to effectively memorize arbitrary noisy labels during training [10], emphasizing the importance of a strategy to enhance the utilization of unlabeled data. The current landscape of self-supervised learning methodologies has demonstrated the efficacy of learning visual representations from unlabeled data. Specifically, approaches grounded in instance discrimination, treating each image as a distinct class and utilizing a contrastive learning objective, have achieved significant success [8], [9], [11], [12]. In particular, contrastive learning seeks to encode image features by mapping multiple augmented views of the same image to a trainable embedding space. This label-free approach significantly diminishes the reliance on manual annotation in datasets, presenting a cost-effective solution for model training. While self-supervised learning methods exhibit success in downstream tasks even in the absence of a filtering process, our motivation lies in incorporating this strategy to foster feature learning independent of (pseudo) labels in the context of semi-supervised object detection.

However, when transitioning these learning methods to object detection datasets such as MS-COCO [13], a discernible decline in learning performance becomes evident. Most con-

trastive self-supervised methods are predicated on the assumption of semantic consistency in images. In these methods, the entire image is regarded as object-centric, assuming that a single instance predominantly occupies the image space, as is often the case in datasets like ImageNet [14]. While this assumption facilitates leveraging semantic consistency for effective representations in single-label classification tasks, real-world images deviate from this premise. In practice, images may comprise diverse semantic content, with instances varying in size and appearing at various locations within the image.

In this paper, we present a contrastive learning approach for semi-supervised object detection. Our approach involves processing images through the Faster-RCNN backbone [1] to generate feature maps, which are then subjected to the Region Proposal Network (RPN) to identify candidate boxes likely to be foreground objects. Unlike using image-level representation, we perform box-level contrastive learning by Region of Interest (RoI) pooling on feature maps generated from different augmented images at the same locations. Furthermore, we avoid global pooling during the contrastive loss calculation. Instead, we choose to compute the loss based on features extracted from corresponding locations in two augmentations. This spatially aware approach retains crucial feature information, as will be explored further in the ablation study (Section IV-C). The main contributions of this paper are summarized as follows:

- We present a contrastive learning approach to improve semi-supervised object detection, performing consistency regularization not only by aligning the box predictions to pseudo boxes but also by considering feature-level representations.
- To address the challenge of object detection, the contrastive loss is computed at the box level, rather than on the entire image. Furthermore, the loss computation is spatially aware.
- Through experiments, we demonstrate that contrastive learning on RoI features can enhance the model's ability to gain additional information from unlabeled data.

In the following sections, we begin by reviewing related studies in Section II. Section III outlines the main investigation method. Finally, experiments and results are discussed in Section IV, and the paper concludes in Section V.

II. Related Work

A. Semi-Supervised Learning

In the realm of semi-supervised learning, two predominant methodological approaches have emerged: one emphasizes data augmentation and perturbation of the input data, while the other centers around consistency regularization. Noteworthy examples include MixMatch [15], which applies multiple augmentations to unlabeled images, averages the predictions of these augmented images, and sharpens them as pseudo-labels for model training. Similarly, FixMatch [16] employs both strong and weak augmentations on unlabeled images, requiring the model's predictions to remain consistent across different versions of the same image. Both methodologies strive to maintain stable predictions across varied inputs, ensuring consistency even in the face of augmentation or perturbation [17].

However, the architectural intricacies of object detection models surpass those of image classification, demanding a more nuanced approach when adapting strategies from the image classification domain.

B. Semi-Supervised Object Detection

Recent advancements in semi-supervised object detection have been propelled by pseudo-labeling methods. STAC [3] initiates the process by generating pseudo-labels using a pre-trained model, subsequently fine-tuning the detection model by integrating these labels with strongly augmented data. Methods such as Unbiased Teacher [2], Instant Teaching [18], and Soft Teacher [19] dynamically generate pseudo-labels on weakly-augmented data while training the model on strongly-augmented data. This approach not only produces higher-quality pseudo-labels but also enhances the model's detection capabilities.

Pseudo-labeling strategies vary: Unbiased Teacher [2] employs a fixed confidence threshold to dynamically filter pseudo-labels, Active Teacher [7] utilizes three metrics derived from confidence values across the entire image, and Dense Learning [5] applies an Adaptive Filtering operation to segment pseudo-labels into foreground, background, and ignored areas. Label Matching [6] assumes that the class distribution in the unlabeled data should resemble that in the labeled data, adjusting thresholds dynamically to create high-quality pseudo-labels accordingly.

However, the generation of pseudo-labels inherently introduces label noise. Deep neural networks have demonstrated the ability to effectively learn from noisy labels during training [10], emphasizing the need for robust strategies to utilize unlabeled data in the context of semi-supervised object detection.

C. Self-Supervised Learning

Self-supervised learning has emerged as a compelling research area within computer vision, showcasing notable achievements in recent years. Diverging from traditional supervised learning, self-supervised approaches eliminate the need for manual data labeling. Instead, these methods capitalize on inherent data structures, providing implicit supervision through pretext tasks designed to learn meaningful representations. Pioneering works such as RotNet [20], jigsaw puzzles [21], and predicting relative patch positions [22] have laid the foundation for self-supervised learning, producing transferable embeddings that exhibit efficacy in tasks like image classification.

In recent advancements, contrastive learning has played a pivotal role. Methods like SimCLR [9] transform images into multiple views, minimizing the distance between identical views and maximizing distances between different views in the feature map. SimCLR, for instance, employs the InfoNCE loss [23] to align similar views in the embedding space. Distinctively, BYOL [8] utilizes asymmetric neural networks, eliminating the need for negative pairs during training. SwAV

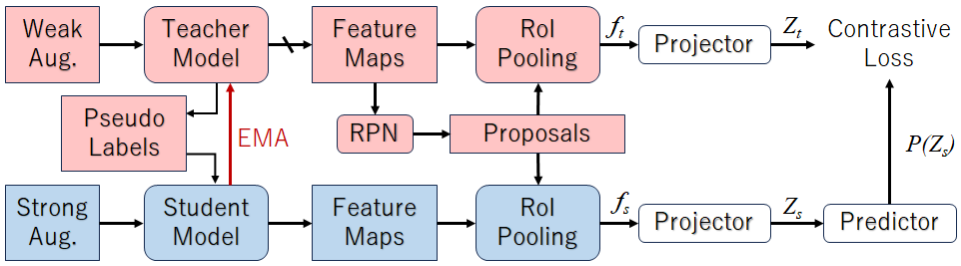


Fig. 2. Flowchart illustrating our semi-supervised object detection framework trained with RoI-enhanced contrastive learning.

[12] introduces an online clustering component to prevent training collapse, leveraging soft encoding and prototype-based predictions, all achieved without relying on a memory bank [11].

In summary, both contrastive learning in self-supervised fields and semi-supervised learning share the common goal of making consistent predictions across different augmented images. However, the distinction lies in the calculation of loss: self-supervised learning assesses feature similarity, while semi-supervised learning employs model-predicted labels to calculate loss. This key difference is underscored by the presence of a small amount of labeled data in semi-supervised learning, enabling the use of meaningful pseudo-labels in subsequent training. Incorrect pseudo-labels can lead to the model learning erroneous information, particularly in semi-supervised object detection. Thus, effective strategies for pseudo-label filtering and generation become crucial research directions. Combining the feature-based learning concept from self-supervised learning could offer a novel direction for pseudo-labeling methods, enabling the model to extract more valuable information from unlabeled data while striking a balance between the quality and quantity of pseudo-labels.

III. Method

In this section, we begin by outlining the semi-supervised object detection problem. Subsequently, we introduce our proposed semi-supervised object detection model, which leverages two approaches for learning from unlabeled data: one based on pseudo labels and another based on box-level feature representations.

A. Problem Definition

The semi-supervised object detection problem involves a set of labeled data $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and a set of unlabeled data $D_u = \{x_i^u\}_{i=1}^{N_u}$ used for training the model, where N_s and N_u represent the numbers of labeled and unlabeled images, respectively. Each labeled image x^s is accompanied by corresponding annotations y^s, encompassing the positions, dimensions, and classes of bounding boxes capturing foreground objects.

B. Semi-Supervised Object Detection with Contrastive Learning

1) Background: Our object detection model builds upon Faster-RCNN [1], incorporating a Feature Pyramid Network (FPN) and ResNet-50 [24]. Initially trained with a limited set of labeled data (e.g., 1% of the COCO training set), the model optimizes the supervised loss L_{sup} involving box-regression and box-classification losses——a stage referred to as the burn-in.

Following the Unbiased Teacher approach [2], the model transitions to the teacher-student mutual learning stage. Duplicating the detector generates two models (teacher θ_t and student θ_s). In this stage, the teacher model generates pseudo-labels on weakly-augmented images, filtered through a confidence threshold. The student model then employs these filtered pseudo-labels for supervised training on strongly-augmented images, computing the object detection loss L_{unsup}. This iterative process creates a symbiotic relationship, with the teacher guiding the student and the student enhancing the quality of pseudo-labels.

To address potential inaccuracies in pseudo-labels, especially during early training, we propose leveraging feature-level representations from the feature maps to enhance the utilization of unlabeled data.

2) RoI-Enhanced Contrastive Learning: We extend Unbiased Teacher with contrastive learning, introducing two projectors and one predictor (Fig. 2). After generating proposals, we perform RoI pooling on the feature maps of both teacher and student models. A projector is added to the teacher branch, and a projector and a predictor are added to the student branch. Exploiting the inherent strong and weak augmentations, we compute additional contrastive learning loss L_{CL}. Utilizing Faster-RCNN, we focus contrastive learning on positions selected through the Region Proposal Network (RPN), likely to be foreground objects. Specifically, we consider only the top 10 proposals based on the likelihood of being in the foreground returned from the teacher RPN.

Given these top 10 proposals, we perform RoI pooling on the student model's feature map, obtaining corresponding RoI features f_t and f_s. We then feed f_t to the projector to obtain Z_t, and feed f_s to the projector and the predictor to obtain Z_s. Utilizing an asymmetric network for contrastive learning [8], which enhances feature-level representations, we compute the contrastive loss as follows:

$$L_{CL} = 2 - 2 \cdot \frac{\langle Z_t, P(Z_s) \rangle}{\|Z_t\|_2 \cdot \|P(Z_s)\|_2}. \quad (1)$$

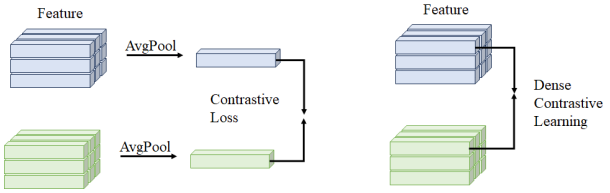


Fig. 3. Difference between the spatial-unaware (left) and the spatial-aware (right) loss calculation.

3) Dense Contrastive Learning: Contrary to applying global pooling to RoI features for contrastive loss calculation, we retain spatial information in the RoI features. We compute the contrastive loss across all positions in both teacher and student model features, as in [25]. Figure 3 illustrates the spatial-aware loss calculation.

The projector and predictor are implemented as multiple-layer perceptrons (MLPs), following [25], [26]. All features entering the projector avoid global pooling, with the projector consisting of three 1×1 convolutional layers. Each layer, except the final one, is followed by a batch normalization layer and a ReLU layer. The predictor comprises two 1×1 convolutional layers, with the first having an output dimension of 64, followed by batch normalization and ReLU, and the second having an output dimension of 256.

Finally, the parameters of the student model are updated through the supervised, unsupervised detection and contrastive loss:

$$L_{total} = L_{sup} + L_{unsup} + L_{CL}. \tag{2}$$

The exponential moving average (EMA) strategy updates the teacher model from the trained student model for each iteration:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s. \tag{3}$$

During inference, the teacher model is utilized for predictions.

IV. Experiments

A. Experimental Settings

This section presents the results of our experiments. To ensure a fair comparison across methods, we adopted Faster-RCNN with ResNet50-FPN as the backbone for object detection, following the setup in [2]. A confidence threshold of $\delta = 0.7$ was utilized. Data augmentation techniques included random horizontal flips for weak augmentation, along with random color jittering, grayscale, Gaussian blur, and cut-out patches for strong augmentations. The exponential moving average (EMA) rate α was set to 0.99. However, due to computational resource constraints, our model was trained with a smaller batch size (16).

We conducted our experiments on the MS-COCO dataset [13]. In line with [2], we randomly sampled 1%, 5%, and 10% of images from the approximately 118,000 images in the `train2017` subset to serve as labeled training data. The remaining images were utilized as unlabeled data. The test set comprised 5,000 images from `val2017`. Evaluation metrics were based on $AP_{50:95}$, denoted as mAP, and the performance was assessed on the teacher model.

B. Results

Our proposed method is designed as a versatile, plug-and-play framework compatible with various teacher-student architectures. Developed initially based on the Unbiased Teacher [2] framework, we exclusively compare our method with Unbiased Teacher in this study to showcase the adaptability within this well-established teacher-student paradigm.

Table I displays the detection performance of the supervised baseline, Unbiased Teacher [2], and our method. Our proposed method outperformed Unbiased Teacher across all settings, achieving improvements of 2.93%, 1.89%, and 1.01% for the 1%, 5%, and 10% labeled data settings, respectively. The mAPs at different iterations are visualized in Fig. 4. It's important to note that both methods underwent the same burn-in process.

TABLE I
Experimental results on COCO-standard.

Method	1%	5%	10%
Supervised	9.05	18.47	23.86
Unbiased Teacher [2]	20.19	28.20	31.46
Ours	**20.78**	**28.73**	**31.77**

C. Ablation Study

To gain a deeper understanding of the proposed method, we conducted additional experiments focusing on the impact of RoI-enhanced contrastive learning, spatial-aware loss computation, and the selection of the top-scored RoI. In these experiments, 10% of the images from the training set were used to train the models, and mAP values were computed on COCO `val2017`. To examine the effect of different contrastive learning choices, we disabled the computation of consistency between the teacher and the student models based on pseudo-labels and focused solely on feature-level consistency learned from unlabeled data.

1) Effect of RoI-Enhanced Contrastive Learning: In contrastive learning, we considered features from the candidate boxes returned by RPN, rather than directly using features obtained from the backbone network. We hypothesized that, during the training of the Faster R-CNN-based object detector, the RPN also learns to discover foreground objects. Intersection over Union (IoU) was calculated between the candidate boxes and the ground truth boxes to categorize all boxes into highly overlapped (positive) or lowly overlapped/non-overlapped (negative) samples. Confusing candidate boxes that did not sufficiently overlap with ground truth were excluded from loss calculation. This approach ensures that the contrastive learning process focuses exclusively on object-level information. Figure 5 (a) illustrates the performance comparison of using RoI features versus the entire feature map. RoI-based features consistently outperformed the alternative after reaching 20,000 iterations, validating the effectiveness of RoI-based contrastive learning.

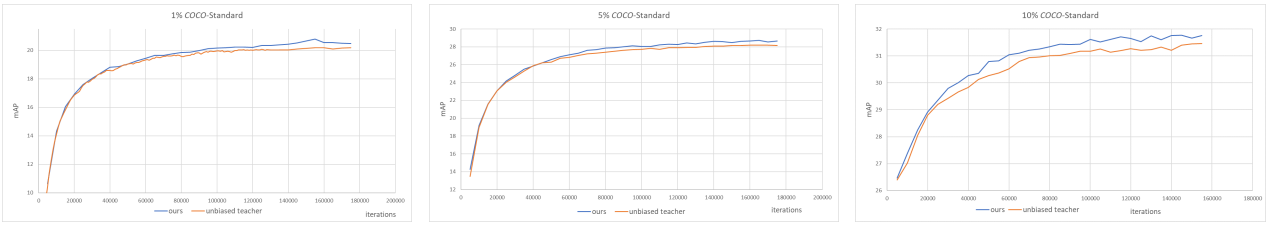
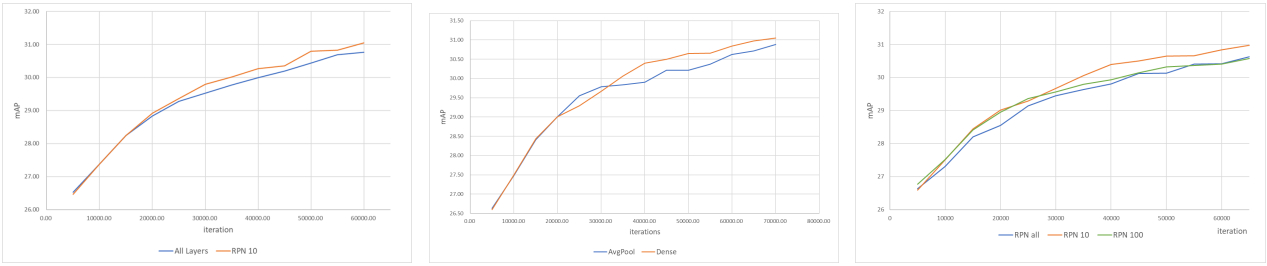


Fig. 4. Comparison with Unbiased Teacher [2] on COCO-Standard using 1%, 5%, and 10% labeled training data, showcasing the mAP values at different training iterations.



(a) (b) (c)

Fig. 5. Performance comparison of (a) contrastive learning with RoI features and the entire feature map; (b) spatial-unaware and spatial-aware contrastive learning; (c) using different numbers of RoIs.

2) Effect of AvgPool: When computing the contrastive loss, we consider that if the features extracted from the candidate boxes undergo global pooling, as in conventional methods, spatial information would be lost, potentially impacting the true similarity between two boxes. Figure 5 (b) displays the performance comparison of spatial-unaware (using global pooling on the features) and the proposed spatial-aware methods. Our approach consistently achieved higher mAP values after 30,000 iterations.

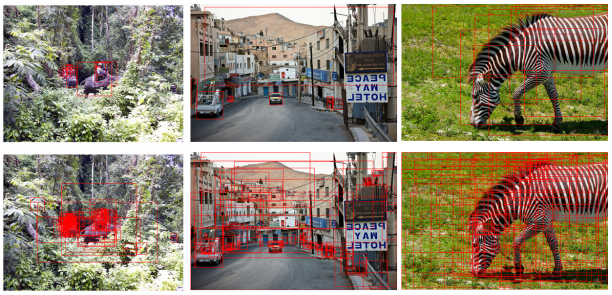


Fig. 6. Visualization of top 10 (top row) and top 100 (bottom row) high-scoring RoIs.

3) Number of RPN Proposals: In the proposed method, we utilized the top 10 boxes in contrastive learning. Alternatively, we experimented with other choices, including using the top 100 and all candidate boxes. Figure 5 (c) shows the comparison result. In this experiment, Non-Maximum Suppression (NMS) with an IoU threshold of 0.5 was used to eliminate overlapping boxes. After NMS, each image retained around 300 candidate boxes out of the initial 1000 boxes. Figure **??** illustrates that when using all candidate boxes, the model's mAP is generally lowest. When using top 100 candidate boxes, the model's early-stage mAP is roughly comparable to that of using the top 10; however, it falls behind after 30,000 iterations. To further investigate, we randomly selected three samples and visualized the RoIs in Figure 6, revealing that many boxes in the top 100 selection set do not effectively capture objects. Consequently, the effectiveness of performing contrastive learning on these boxes is questionable. Therefore, we adopted the top 10 high-scoring candidate boxes for contrastive learning.

V. Conclusion

In this paper, we have presented a contrastive learning approach to enhance semi-supervised object detection. Specifically, we leverage the candidate boxes selected by the Region Proposal Network (RPN) to facilitate RoI-based contrastive learning. Additionally, we incorporate pixel-level comparisons to enable spatial-aware loss calculation. Our experimental results showcase the effectiveness of the proposed RoI-enhanced contrastive learning, coupled with pseudo-labeling, in extracting valuable information from unlabeled data. This integration not only enhances model performance but also mitigates the impact of noisy pseudo-labels during consistency regularization in semi-supervised object detection.

One future direction we have been pursuing is the validation of the proposed plug-and-play method on alternative detection frameworks beyond Faster-RCNN [1]. Expanding our approach to different models will provide valuable insights into its adaptability and generalizability across various architectures. By extending our methodology to a broader spectrum of detectors, we aim to offer a robust and flexible solution that can be tailored to different detection paradigms,

ultimately advancing the state-of-the-art in semi-supervised object detection.

ACKNOWLEDGMENT

This work was supported by the National Science and Technology Council of Taiwan (MOST 111-2221-E-003-016-MY2, MOST 110-2634-F-002-050).

REFERENCES

[1] R. Girshick, "Fast r-cnn," in *IEEE International Conference on Computer Vision*, 2015.

[2] Y.-C. Liu *et al.*, "Unbiased teacher for semi-supervised object detection," *arXiv preprint arXiv:2102.09480*, 2021.

[3] K. Sohn *et al.*, "A simple semi-supervised learning framework for object detection," *arXiv preprint arXiv:2005.04757*, 2020.

[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[5] B. Chen *et al.*, "Dense learning based semi-supervised object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[6] B. Chen *et al.*, "Label matching semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[7] P. Mi *et al.*, "Active teacher for semi-supervised object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[8] J.-B. Grill *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, pp. 21 271–21 284, 2020.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, PMLR, 2020.

[10] D. Arpit, S. Jastrzebski, N. Ballas, *et al.*, "A closer look at memorization in deep networks," in *International Conference on Machine Learning*, PMLR, 2017.

[11] K. He *et al.*, "Momentum contrast for unsupervised visual representation learning," in *IEEE/CVF International Conference on Computer Vision*, 2020.

[12] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *arXiv:2006.09882*, 2020.

[13] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[14] O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge," in *arXiv:1409.0575*, 2014.

[15] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2019.

[16] K. Sohn *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, 2020.

[17] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, 2017.

[18] Q. Zhou *et al.*, "Instant-teaching: An end-to-end semi-supervised object detection framework," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[19] M. Xu *et al.*, "End-to-end semi-supervised object detection with soft teacher," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[20] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *International Conference on Learning Representations*, 2018.

[21] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*, 2016.

[22] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[23] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in *arXiv:1807.03748*, 2018.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[25] X. Wang *et al.*, "Dense contrastive learning for self-supervised visual pre-training," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[26] X. Chen and K. He, "Exploring simple siamese representation learning," in *IEEE/CVF International Conference on Computer Vision*, 2021.

Dual Motion Attention and Enhanced Knowledge Distillation for Video Frame Interpolation

Dengyong Zhang*, Runqi Lou*, Xiaoping Hna§, Jiaxin Chen*, Xin Liao†, Gaobo Yang†, and Xiangling Ding‡
* School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China
E-mail: Dengyong Zhang, zhdy@csust.edu.cn; Runqi Lou, lourunqi@stu.csust.edu.cn; Jiaxin Chen, jxchen@csust.edu.cn
† College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
E-mail: Xin Liao, xinliao@hnu.edu.cn; Gaobo Yang, yanggaobo@hnu.edu.cn
‡ School of Computer and Communication Engineering, Hunan University of Science and Technology, Xiangtan, 411201, Hunan, Chin
E-mail: xianglingding@163.com
§ Army Academy of Armored Forces, Beijing, 100072, China
E-mail: 15811060576@163.com

***Abstract*—Video frame interpolation presents a formidable challenge in the domain of video generation, primarily due to the intricate motion dynamics exhibited by objects within video frames. With the advancements in deep learning, numerous flow-based methods for video frame interpolation have emerged. These methods aim to predict intermediate frames by leveraging the estimation of motion information between frames. In this paper, we propose a novel framework for modeling input video frames, which employs a coarse-to-fine structure to extract motion information between frames. Additionally, it incorporates a Bidirectional Correlation Volume and a complementary module of contextual features, specifically designed to pay attention to the symmetry of the optical flow and shallow motion features. We incorporate this dual attention to the knowledge distillation part of the model, which further improves the performance of the model. Leveraging this framework, our model demonstrates the ability to accurately predict motion information between frames, consequently producing visually appealing intermediate frames.The code is available at https://github.com/famt0531/DAEK.**

I. Introduction

Video frame interpolation(VFI) serves as a low-level task in computer vision, utilized across various applications including video post-processing and surveillance. Its objective is to enhance the frame rate of a video sequence by generating intermediate frames between input frames. This process aims to achieve smoother video playback and mitigate motion blur[1], [2]. Traditional approaches typically involve estimating the optical flow between frames and subsequently interpolating or extrapolating along the optical flow vector. While effective with accurate optical flow, this method often yields significant artifacts and blurring when the optical flow estimation is imprecise.Moreover, they exhibit poor performance in scenarios involving occlusion and variations in luminance.

Recent years, numerous flow-based VFI techniques have emerged as prominent contenders in this domain[3]–[7]. These methods employ bidirectional optical flow to discern motion information across frames and map corresponding pixel positions. Utilizing estimated optical flow, they guide the warping process by transforming input images to interpolated frame positions, and blending them to preserve spatio-temporal correlations within the expected motion[4], [5]. Despite their promising outcomes, two main issues persist:

- Overlooking the correlation when estimating bidirectional optical flow, resulting in accumulated optical flow errors.
- Paying few attention to deep network features, which leads to missing detailed textures in the inference results.

To address the above limitations, in this paper, we propose a novel network architecture termed DAEK, which extracts coarse motion information, comprising intermediate optical flow and features, via three CNN-based Motion Estimation Blocks(MEB) in a coarse-to-fine fashion. Subsequently, these coarse motion information undergoes refinement through the Motion Refinement Blocks(MRB). During training, akin to RIFE, we employ intermediate supervision. However, we prioritize the accuracy of the generated optical flow of teacher model so that we use a combination of MEB and MRB, instead of directly employing the teacher module in RIFE. Our contributions are summarized as follows:

- We employed bidirectional correlation volumes to focus on the symmetry of motion at different stages of the network. Additionally, we utilized feature point multiplication to address the loss of motion information in the deeper layers of the network.
- We applied our proposed dual motion attention mechanism to a smaller model, which we trained as a teacher model for knowledge distillation. The teacher model with dual motion attention demonstrates a superior ability to guide the student model in accurately inferring optical flow information, compared to conventional teacher models
- We proposed a novel video frame interpolation framework, which adopts a coarse-to-fine structure to capture the motion information and considers the motion symmetry and texture loss through double motion focus.

II. Related Work

A. VFI

VFI, a complex task in low-level computer vision, has seen the emergence of numerous innovative methods in recent years, broadly categorized into flow-based and non-optical flow-based approaches for motion simulation. In recent research endeavors, optical flow estimation has emerged as an indispensable component. For instance, Kong et al.[6] proposed IFRNet, a video interpolation network with a single encoder-decoder structure. It extracts feature pyramids from two input frames, refines bidirectional intermediate optical flow fields, and restores the desired output at the input resolution. Huang et al.[8] introduced RIFE, which estimates intermediate flows efficiently from coarse to fine, enhancing speed. They also developed a privileged distillation scheme for training for boosting performance. Jin et al.[9] created a compact model that estimates bidirectional motion simultaneously using a flexible pyramid recurrent framework, fine-tuning components within optical flow research for improved performance.

B. Correlation Volume

Correlation volumes play a fundamental role in representing matching costs across various computer vision tasks. In the domain of VFI, several studies [9]–[11] have proposed construction schemes inspired by PWC-Net [12]. However, these schemes often focus solely on the local region's cost volume during matching, leading to inaccuracies when the region undergoes distortion. In contrast, AMT[13], building upon RAFT[14], transforms the unidirectional correlation volume into a bidirectional one, effectively capturing multiscale correspondence between frames. This compact global motion representation facilitates precise optical flow prediction, particularly for large motions. In this work, we adopt the design of AMT for the correlation volume construction.

C. Self Attention

Given the remarkable success achieved by Transformer in natural language processing, there has been a surge in introducing Transformer to computer vision tasks, yielding promising results. In VFI, Liu et al. proposed ConvTransformer[15] to model long-term dependencies between 2D video frames in both temporal and spatial domains. Similarly, Lu et al.[16] utilized Transformer as an encoder in their model, enabling it to learn relationships between each frame and others via a self-attentive mechanism. However, many VFI methods based on Transformer employ the self-attention mechanism as a feature extractor, which can be computationally expensive due to the large number of parameters involved. Therefore, rather than directly using the Transformer for feature extraction, we simulate its self-attention mechanism to supplement contextual information during optical flow updating.

III. Proposed Method

A. Problem Description

Given two frames I^0 and I^1, the objective of Video Frame Interpolation (VFI) is to generate the intermediate frame I^t, where $t \in [0, 1]$ represents the time step and is typically set to 0.5 by default. Our proposed model follows an end-to-end optical-flow based approach, illustrated in Fig. 1. We use the same residual structure to extract appearance features and context features, which are used to construct bidirectional cost volumes and enrich the texture information in the deeper layers of the network, respectively. The coarse motion information is then refined in MRB.

B. MEB

MEB iterates to estimate intermediate flows utilizing a coarse-to-fine structure, enabling effective handling of large motions. For each MEB, the inputs consist of I^0 and I^1 which obtained after bilinear interpolation downsampling, intermediate flows $flow_i \in \mathbb{R}^{4\times H^{'}\times W^{'}}$, mask $M_i \in \mathbb{R}^{1\times H^{'}\times W^{'}}$, intermediate features $Feat_i \in \mathbb{R}^{C^{'}\times H^{'}\times W^{'}}$ (where $i = 1, 2$) which is predicted by the previous MEB, and the $\tilde{I}_0, \tilde{I}_1$ obtained by warping I^0 and I^1 with $flow_i$ (for the first MEB, we merely input I^0 and I^1). Specifically, we employ three consecutive motion estimation modules, each processing motion information at different scales. Following the multi-scale strategy outlined in RIFE[8], we set the scales to $[1/8, 1/4, 1/2]$.

All MEBs share the same structure, realized entirely by convolutional layers. Additionally, we output intermediate features for each MEB, as shown in Fig. 2, which play a crucial role in the subsequent optical flow update phase and motion estimation phase. By including intermediate features, our model retains more motion information during the iterative process of generating optical flow at multiple scales.

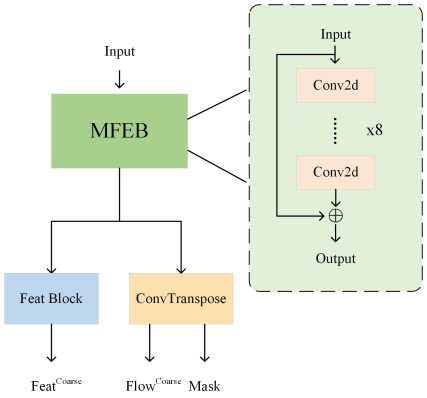


Fig. 2. The structure of MEB. The $Conv$ comprises 8 layers of convolutional layers and PRELU units. Each convolutional layer utilizes a 3x3 convolutional kernel with a stride of 1 and padding of 1.

C. MRB

After obtaining the initial bi-directional optical flow and intermediate features, we utilize MRB to generate residuals for the bi-directional optical flow and intermediate features. These residuals are then used to update both the optical flow and intermediate features. We use bidirectional cost volumes to obtain the bidirectional matching scores between the two real frames. Our feature matching is consistently performed at

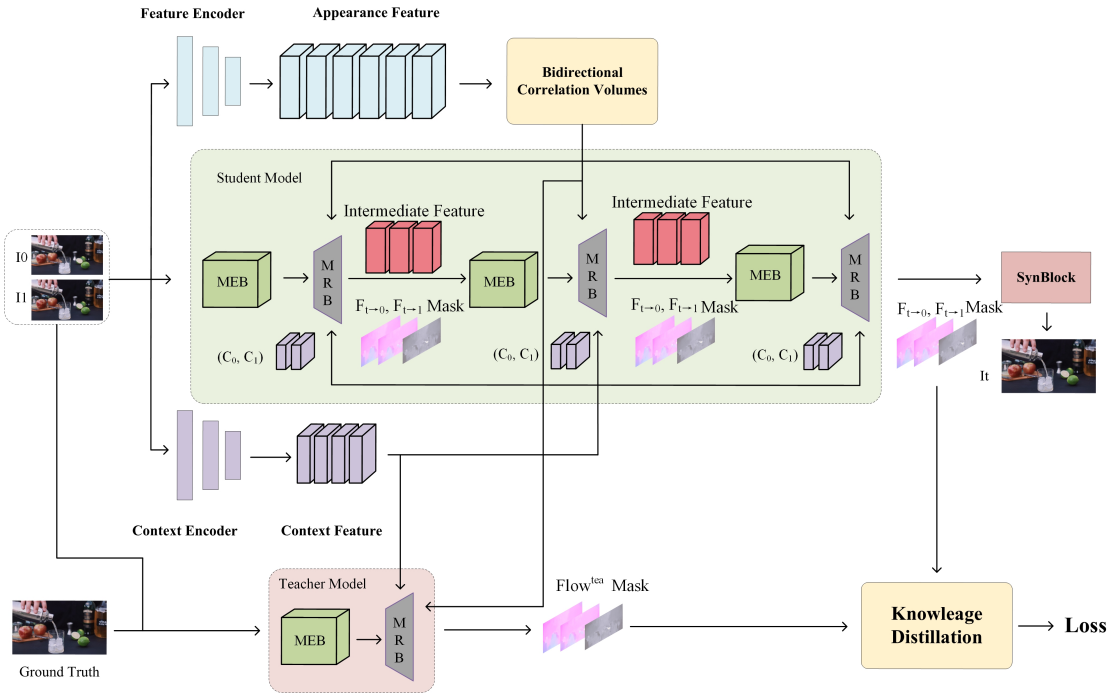


Fig. 1. The overview of the proposed model. The input frames are processed through MEB and MRB across three scales to generate bidirectional optical flow ($F_{t\to0}$, $F_{t\to1}$) and $Mask$. The synthesis module then outputs the intermediate frame. Our teacher module generates distillation loss only during the training phase and is discarded during inference.

a resolution of $\frac{1}{8}$. For each scale output of the MEB, we first downsample them to $\frac{1}{8}$ resolution using linear interpolation. After updates, we then restore them to the original resolution. We employ several convolutional layers to extract bidirectional correlation features $BCF \in \mathbb{R}^{C\times\frac{H}{8}\times\frac{W}{8}}$ and coarse flow features, then concatenate them along the channel dimension to obtain the coarse motion information CMF, as follows:

$$CMF = Cat(Conv(Flow^{coarse}), Conv(BCF)) \quad (1)$$

Where Cat represents the concat operation, and $Conv$ for convolution operation.

We posit that features at the shallow end of the network encompass more textural details, whereas deeper layers tend to encapsulate semantic information. However, as the network depth increases, texture features become inevitably attenuated. Therefore, we introduce Contextual Feature Complementary Module(CFCM) that uses an attention mechanism to complement feature information at deeper layers of the network which serves as a crucial component of MRB.

As shown in Fig.3, the context feature undergoes transformation by a linear layer into $Query, Key \in \mathbb{R}^{C\times H*W}$. Subsequently, the resulting outputs are multiplied and passed through a $Softmax$ unit to obtain $Score$. Meanwhile, the CMF is transformed into $Value \in \mathbb{R}^{C^v\times H^v\times W^v}$ through a convolutional layer with a kernel size of 1×1. The CFCM we propose is essentially a feature enhancement method, so we abandon the tedious calculation of multiplying the result of Q and K and then multiplying it with V, and instead add it to V. In order to better achieved feature fusion and feature

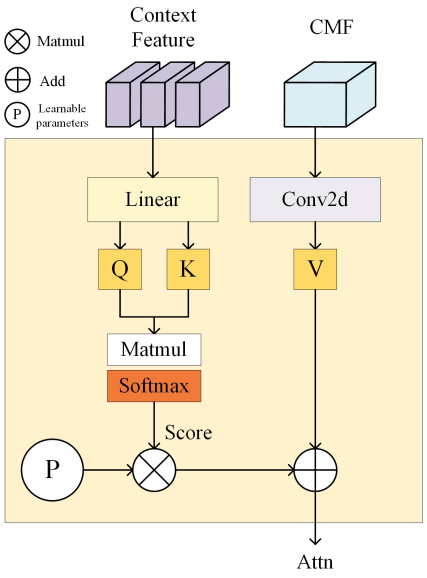


Fig. 3. The structure of CFCM, where the convolutional layer has a kernel size of 1x1.

enhancement, we introduce a learnable parameter, initialized to 0, which is multiplied with the Score and added to the Value to obtain Attn, This step can be expressed as:

$$Q, K = linear(contextfeature) \quad (2)$$

$$Score = Softmax(Q \bigodot K) \quad (3)$$

$$V = Conv(CMF) \tag{4}$$

$$Attn = V + (Para \bigodot Score) \tag{5}$$

Where $linear$ represents fully connected layer, $Softmax$ represents Softmax unit, $Conv$ represents convolution operation, $\bigodot$ represents multiplication.

Finally, we concatenate $Attn$ with $Flow^{coarse}$ and intermediate feature along the channel dimension and apply convolutional layers to obtain the residuals which is added with coarse flow and intermediate feature to achieve the goal of refining.

D. Enhanced Knowleage Distillation

We stack an extra MEB as the optical flow estimation module for the teacher model, the MEB for the teacher model has the same structure as the MEB for the student model, except that extra ground truth is added as an input. In order to obtain a more accurate labeling of the optical flow, we add a corresponding MRB for the teacher model, which produces residuals only for the flow involved in knowledge distillation, but not for the intermediate feature, since the intermediate feature here are not involved in subsequent parameter updates during the training phase. After obtaining the optical flow labels generated by the teacher model again, we use it and the intermediate flows generated by the student model at different scales to generate distillation losses, as follows:

$$L_{dis} = \sum ||Flow^{tea}_{t\to 0}, Flow_{t\to 0}||_1 + ||Flow^{tea}_{t\to 1}, Flow_{t\to 1}||_1 \tag{6}$$

Where $Flow^{tea}_{t\to i}(i = 0, 1)$ represents flow generated by teacher module, $Flow_{t\to i}(i = 0, 1)$ represents flow fields generated by student module, and $||\cdot||_1$ is the operation of $L1$ loss.

IV. Experiments

A. Datasets

1) Training Dataset: We train our DAEK on the Vimeo-90K[17] training set which comprises a total of 51,312 triples. Each triple contains three consecutive video frames with a resolution of 448 × 256. In training stage, we crop patches from the original images with a resolution of 224 × 224 and randomly augment it by performing horizontal and vertical flipping, chronological flipping, and rotating by 90 degrees, similar to other methods such as RIFE[8], AdaCoF[18].

2) Evaluation settings: We test the performance of DAEK on the Vimeo-90K test set, UCF101[19] dataset, and SNU-FILM[20] dataset, and for all the tables in this section, the best performance is labeled by us in bold font.

B. Implementation Details

1) Loss function: DAEK uses three loss functions, which including reconstruction loss L_{rec}, teacher reconstruction loss L^{tea}_{rec} and distillation loss L_{dis}. The total loss L:

$$L = L_{rec} + L^{tea}_{rec} + \lambda L_{dis} \tag{7}$$

where we set λ=0.01.

Our reconstruction loss quantifies the quality of the reconstructed I_t and I^{gt}_t using the $L1$ loss function with the Laplacian operator[21] which could be defined as:

$$L_{rec} = Lap(I_t, I_{gt}) \tag{8}$$

$$L^{tea}_{rec} = Lap(I^{tea}_t, I_{gt}) \tag{9}$$

where Lap is the Laplacian operator, I_t is the intermediate frame of student module, I_{gt} is the ground truth, and I^{tea}_t is the intermediate frame of teacher module.

2) Training Details: DAEK was implemented using the Pytorch framework and trained on a NVIDIA A30 Tensor Core GPU. We used AdamW[22] as the optimizer with a weight decay of 10^{-3}. The initial learning rate is 10^{-4}, which is decayed to 10^{-6} by cosine annealing strategy during the training process, and the whole training process requires 300 epochs with a batch size set to 32.

C. Performance Evaluation

1) Objective results: We compare DAEK to current state-of-the-art methods including, IFRNet[23], AdaCoF[18], BMBC[11], EBME[9], RIFE[8], M2M[24], as shown in TabI.

We conducted a subjective performance comparison on the highly challenging extreme dataset of SNU-FILM, as shown in Fig.4 and Fig.5. When interpolating across multiple video frames, the motion within the frames becomes significantly more extensive and highly nonlinear. DAEK demonstrates excellent performance when dealing with these complex motions. Other optical flow-based methods exhibit noticeable blurring and distortion, while the kernel-based method AdaCoF, although maintaining motion continuity, suffers from pixel loss. In contrast, DAEK preserves rich texture details and accurately estimates motion continuity.

D. Ablation Study

1) Structural Ablation Experiment: We use models without dual motion attention as the baseline, as shown in TabII. Here, F indicates the use of intermediate features, while its absence indicates no use. CA represents correlation attention, SFA stands for shallow feature attention, and DA denotes dual attention. The use of these two motion attention individually brings certain performance improvements. When used together, the model's ability to represent bidirectional motion is further enhanced. Additionally, the introduction of intermediate features provides the intermediate frames with realistic texture details and accurate motion representation.

TABLE II
VALIDITY OF DIFFERENT COMPONENTS OF OUR MODEL

Setting	Vimeo90K PSNR	UCF01 PSNR
Baseline	35.32	35.29
DAEK-F-CA	35.43	35.39
DAEK-F-SFA	35.31	35.35
DAEK-F-DA	**35.66**	**35.45**
DAEK-DA	35.49	35.30

TABLE I
THE EVALUATION OF VARIOUS INTERPOLATION METHODS ON VIMEO-90K TEST SET, UCF101 DATASET AND SNU-FILM DATASE.

Method	Vimeo-90K	UCF101	SNU-FILM				Parameters (M)	Runtime (S)
			Easy	Medium	Hard	Extreme		
IFRNet	**35.80/0.979**	35.29/0.969	**40.03**/0.990	**35.94**/0.979	30.41/0.935	25.05/0.858	**5.0**	0.02
AdaCoF	32.00/0.971	35.08/0.966	39.80/0.990	35.05/0.975	29.46/0.924	24.31/0.843	21.8	0.02
BMBC	35.01/0.976	35.15/0.969	39.90/0.990	35.31/0.977	29.33/0.927	23.92/0.843	11.0	0.3
EBME	35.58/0.978	35.30/0.969	40.01/**0.991**	35.80/0.979	30.42/0.935	**25.25/0.861**	3.9	0.02
RIFE	35.32/0.976	35.29/**0.973**	38.06/0.983	35.14/0.975	29.72/0.925	24.31/0.843	9.8	**0.01**
M2M-PWC	35.49/0.978	35.32/0.970	39.66/**0.991**	35.74/**0.980**	30.32/**0.936**	25.07/0.860	7.6	0.03
Ours	35.66/**0.979**	**35.42**/0.971	36.13/0.981	33.97/0.971	29.91/0.933	25.13/**0.861**	22.48	0.03



(a)Ground Truth (b)AdaCoF (c)IFRNet (d)M2M (e)RIFE (f)DAEK

Fig. 4. Comparison of Visualized Test Results on Extreme set of SNU-FILM



(a)Ground Truth (b)AdaCoF (c)IFRNet (d)M2M (e)RIFE (f)DAEK

Fig. 5. Comparison of Visualized Test Results on Extreme set of SNU-FILM

2) Knowledge Distillation Ablation Experiment: Knowledge distillation can refine model parameters during the training phase, guiding it to produce more accurate results. As shown in Tab.III, the model's performance significantly declines without knowledge distillation. However, when we apply the knowledge distillation proposed in RIFE, the model's performance improves. Since we incorporate dual motion attention in the knowledge distillation process, the flow labels generated by the teacher model include richer bidirectional motion information, thereby guiding the student model to produce more accurate outputs. To validate the robustness of our enhanced knowledge distillation, we replaced the teacher model in RIFE. This substitution resulted in a performance improvement for RIFE as well.

TABLE III
ABLATION EXPERIMENTS FOR DISTILLATION SCHEMES.

Setting	Vimeo90K
	PSNR
DAEK w/o distill	35.20
DAEK w/ RT	35.40
DAEK w/ OT	**35.66**
RIFE w/ OT	35.49

V. CONCLUSIONS

In this study, we introduce a novel model named DAEK, which initially generates the base motion flow field using MEB. Subsequently, the original motion flow field is refined using MRB which integrates dual motion attention. These designs aids DAEK in addressing complex motions in video

frames resulting in significant results. Moving forward, we aim to optimize the construction of bidirectional correlation volumes, for instance, by employing the PatchMatch[25], to reduce the computational load and inference time of the model. Additionally, we plan to enhance the model's architecture to improve its learning capacity, as a higher initial learning rate facilitates parameter optimization, ultimately boosting the model's learning effectiveness.

VI. Acknowledgements

This work is supported in part by the Postgraduate Scientiic Research Innovation Project of Hunan Province (CSLGCX23089)

References

[1] J. He, G. Yang, X. Liu, and X. Ding, "Spatio-temporal saliency-based motion vector refinement for frame rate up-conversion," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–18, 2020.

[2] W. Bao, W. S. Lai, C. Ma, X. Zhang, Z. Gao, and M. H. Yang, "Depth-aware video frame interpolation," *IEEE*, 2019.

[3] Z. Zhang, L. Song, R. Xie, and L. Chen, "Video frame interpolation using recurrent convolutional layers," in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, IEEE, 2018, pp. 1–6.

[4] S. Yu, B. Park, and J. Jeong, "Posnet: 4x video frame interpolation using position-specific flow," *IEEE*, 2019.

[5] H. Li, Y. Yuan, and Q. Wang, "Fi-net: A lightweight video frame interpolation network using feature-level flow," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2019.

[6] L. Kong, B. Jiang, D. Luo, *et al.*, "Ifrnet: Intermediate feature refine network for efficient frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1969–1978.

[7] S. Niklaus and F. Liu, "Softmax splatting for video frame interpolation," *IEEE*, 2020.

[8] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *European Conference on Computer Vision*, Springer, 2022, pp. 624–642.

[9] X. Jin, L. Wu, G. Shen, *et al.*, "Enhanced bi-directional motion estimation for video frame interpolation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5049–5057.

[10] Q. Hou, A. Ghildyal, and F. Liu, "A perceptual quality metric for video frame interpolation," in *European Conference on Computer Vision*, Springer, 2022, pp. 234–253.

[11] J. Park, K. Ko, C. Lee, and C.-S. Kim, "Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, Springer, 2020, pp. 109–125.

[12] D. Sun, X. Yang, M. Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Z. Li, Z.-L. Zhu, L.-H. Han, Q. Hou, C.-L. Guo, and M.-M. Cheng, "Amt: All-pairs multi-field transforms for efficient frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9801–9810.

[14] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 2020, pp. 402–419.

[15] Z. Liu, S. Luo, W. Li, *et al.*, "Convtransformer: A convolutional transformer network for video frame synthesis," *arXiv preprint arXiv:2011.10185*, 2020.

[16] L. Lu, R. Wu, H. Lin, J. Lu, and J. Jia, "Video frame interpolation with transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3532–3542.

[17] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.

[18] H. Lee, T. Kim, T. Y. Chung, D. Pak, and S. Lee, "Adacof: Adaptive collaboration of flows for video frame interpolation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *Computer Science*, 2012.

[20] M. Choi, H. Kim, B. Han, N. Xu, and K. M. Lee, "Channel attention is all you need for video frame interpolation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 10 663–10 671.

[21] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1701–1710.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[23] L. Kong, B. Jiang, D. Luo, *et al.*, "Ifrnet: Intermediate feature refine network for efficient frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[24] P. Hu, S. Niklaus, L. Zhang, S. Sclaroff, and K. Saenko, "Video frame interpolation with many-to-many splatting and spatial selective refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[25] Z. Zheng, N. Nie, Z. Ling, *et al.*, "Dip: Deep inverse patchmatch for high-resolution optical flow," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8925–8934.

Detecting Coronary Artery Stenosis from Cardiac CT Images using 3D CNNs

Masaki Aono*, Tetsuya Asakawa†, Kazuki Shimizu‡, Masashi Hashimoto§, Takeshi Miyaji¶, and Kei Nomura‖
* Toyohashi University of Technology, Toyohashi, Aichi, JAPAN
E-mail: masaki.aono.ss@tut.jp
† E-mail: asakawa.tetsuya.um@tut.jp
‡ Toyohashi Heart Center, Toyohashi, Aichi, JAPAN
E-mail: shimizu@heart-center.or.jp
§ E-mail: m.hashimo6123@gmail.com
¶ E-mail: miyaji@heart-center.or.jp
‖ E-mail: k.nomura@heart-center.or.jp

***Abstract*—In this paper, we propose an innovative method for fine-tuning a video pretrained model and introducing 3D representations to predict coronary artery stenosis. The input data is a collection of both contrast and non-contrast CT images. Both types of CT images are converted into 3D representations, which we call "sandwiches" of size k.**

We divide the sandwiches into training and validation data, and then apply 3D CNNs to detect stenosis. The experimental results are very promising, and we believe that our fine-tuned 3D CNN is practical for stenosis detection.

I. Introduction

Recent advances in artificial intelligence in the medical domain make it possible to predict a variety of serious diseases based on a collection of medical images obtained from diagnostic devices such as CT, MRI, and ultrasound. Coronary artery stenosis is one of the diseases that has been very difficult to identify the existence, location and severity of the disease.

In this paper, we propose an innovative method for fine-tuning a video pretrained model and introducing 3D representations to predict coronary artery stenosis. The input data is a collection of both contrast and non-contrast CT images. Both types of CT images are converted into 3D representations, which we call "sandwiches" of size k. We divide the sandwiches into training and validation data, and then apply 3D CNNs for stenosis detection.

In the following, we briefly survey related work in Section II. In Section III, we describe preprocessing data, followed by focusing on the proposed methods in Section IV. Experimental results are described in Section V, and Section VI concludes the paper.

II. Related Work

In this section, we briefly describe recent research activities and their objectives related to the prediction of stenosis.

Most of the approaches to date have employed CCTA (Coronary Computed Tomography Angiography), as a reliable noninvasive modality to assess coronary artery stenosis.

Dan Han et al. [1] used CCTA to assess the coronary artery stenosis, where U-Net [2] was adopted for image segmentation and V-Net [3] was employed for stenosis detection. Andrew Lin et al. [4] used CCTA and ICA (Invasive Coronary Angiography) to predict stenosis.

Abdul Rahman Ihdayhid et al. [5] published their work on the assessment of coronary artery stenosis and high-risk plaque with a deep learning technique. Specifically they used approximately 600 patients and translated the problem into four-class classification, where the degree of stenosis severity was divided into four ranges; $> 0\%$ (class 1), 1 to 49% (class 2), $\geq 50\%$ (class 3), and $\geq 70\%$ (class 4). Similarly, Jingjing Yan et al. [6] translated the problem into 3 class classification.

In our previous work on stenosis severity [7], we defined the problem as a multi-label multi-class classification problem and a regression problem, using 2D DNNs.

To our knowledge, no known approaches have used cardiac CT slice images to predict coronary artery stenosis, using video-pretrained 3D CNNs.

III. Preprocessing Data

In this section, we will first discuss what we call "annotation data" which includes anonymous patients' diagnostic data as shown in Fig. 1. The total amount of stenosis-positive data we use is approximately 900, while that of stenosis-negative is more than 2,000.

A. Annotation data

Fig. 1 illustrates the data we have received from the Toyohashi Heart Center for the detection of stenosis, where **No** denotes the patient ID (or number), **YN** denotes stenosis (Y) or non-stenosis (N). **CA** indicates the name of coronary artery; RCA, LAD, LCX, or LMT. **segment** is the segment number, **QCA** (Quantitative Coronary Angiography) represents the severity of the stenosis. **sLesion** is the "simple" lesion number of CT images, referring to the center of the CT slice where stenosis is evident. Note that a "non-contrast" CT image is sometimes referred to as "simple", as a subject does not have to ingest a contrast agent before the CT scan.

On average, we obtain 56 non-contrast CT slices, scanned approximately 2.0 mm apart for each subject. **sLesionImage** is a DICOM image corresponding to the **sLesion**. Similarly, **cLesion** stands for the "contrast" lesion number of the CT images regarding as the center of the CT slices where stenosis is evident. **cLesionImage** is a DICOM image corresponding to the **cLesion**. On average, we have 330 contrast CT slices for each subject, scanned approximately 0.5 mm apart.

No	YN	CA	segment	QCA	sLesion	sLesionImage	cLesion	cLesionImage
2	Y	LCX	14	70.21	23	0917950d.dcm	103	0dc29017.dcm
3	Y	RCA	2	71.27	30	0917bb10.dcm	144	0dc2ed05.dcm
4	Y	LMT	5	63.6	7	0918dc06.dcm	51	0dc3540e.dcm
6	Y	LAD	67	100	15	09193f0e.dcm	66	0dc40b27.dcm
7	Y	LCX	11	60.02	19	09196607.dcm	91	0dc4681d.dcm
8	Y	RCA	2	59.68	40	0919871c.dcm	112	0dc4f322.dcm
8	Y	LAD	6	78.3	16	09198704.dcm	272	0dc4f806.dcm
10	Y	RCA	1	72.33	25	0919f414.dcm	169	0dc5ae0c.dcm
11	Y	RCA	1	69.12	13	091a150c.dcm	69	0dc60704.dcm
11	Y	RCA	2	81.56	20	091a1513.dcm	101	0dc60724.dcm
11	Y	LAD	7	60.15	9	091a1508.dcm	159	0dc6090c.dcm
12	Y	RCA	1	61.66	22	091a4615.dcm	76	0dc67109.dcm
12	Y	LAD	6	59.47	15	091a460e.dcm	118	0dc6720a.dcm

Fig. 1. Annotation data outlook

B. 3D data representation for CT slices

Among the annotation data, we focus primarily on **sLesion** and **cLesion** to represent 3D data for detecting stenosis. As described in the previous section, there is a large difference in the number of CT slices between contrast and non-contrast CT images, acquired at the Radiology Department of the Toyohashi Heart Center. As illustrated in Fig. 2, for the contrast CT image data, we have calculated the C_k number of consecutive slices (which we call "sandwich" of size C_k) as the unit for the diagnosis of stenosis, while for non-contrast CT image data, we have adopted S_k number of consecutive slices ("sandwich" of size S_k) as the unit for diagnosing stenosis. Usually S_k is much smaller than C_k, because the distance between the adjacent CT scans is smaller in non-contrast CT, compared to contrast CT. In our implementation, each sandwich is represented by NumPy ndarray [8].

It should be noted that for *positive* data, we coincide the center of consecutive CT slices with the lesion number, assuming that the location should be around the center of the lesion. Meanwhile, for *negative* data, we randomly generate a healthy center somewhere near the median of all CT slices per subject.

IV. Proposed Methods

Here we propose novel methods for the detection of coronary artery stenosis by taking advantage of 3D convolutional neural networks (CNNs) with 3D data representations ("sandwiches") of cardiac CT images.

A. Sandwich structure

Given a stenosis lesion image within a series of both contrast and non-contrast CT images, we hypothesize that the adjacent

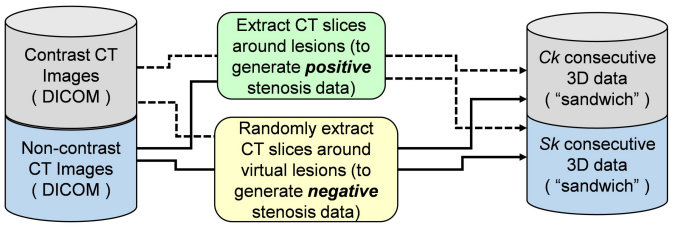


Fig. 2. Preprocessing of contrast and non-contrast cardiac CT images

consecutive CT images should also have stenosis. Specifically, for contrast CT images where the average number of CT slices is approximately 330, at least 10 anterior and posterior to the central lesion should have stenosis. On the other hand, for non-contrast CT images, where the average number of CT slices is approximately 56, at most 6 before and after the central lesion should have stenosis. These numbers (10 and 6) are empirically determined.

Fig. 3 illustrates a central "lesion" image with 10 consecutive *previous* contrast CT slices and with 10 consecutive *subsequent* contrast CT slices, resulting in $Ck = 21$ consecutive contrast CT slices (sandwich of size 21) in total. Similarly, Fig. 4 illustrates a central "lesion" image with 6 consecutive *previous* contrast CT slices and with 6 consecutive *subsequent* contrast CT slices, resulting in $S_k = 13$ consecutive contrast CT slices (sandwich of size 13) in total. In other words, in our implementation, for contrast CT, the sandwich size is assumed to be 21,while for non-contrast CT, the sandwich size is assumed to be 13.

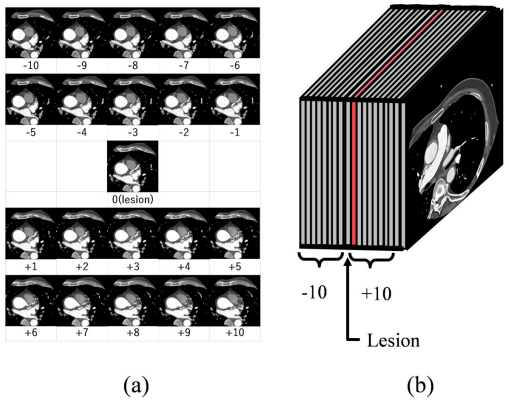


Fig. 3. Representations for contrast CT images (a) 2D representation of the 21 consecutive contrast CT slices, where 10 slices are hypothesized to inherit stenosis before and after the central lesion image (b) 3D representation of the C_k (=21) consecutive slices stick together.

B. Overall flow of proposed system

The overall flow of our proposed system is illustrated in Fig. 5. After converting DICOM data into sandwich structures, we randomly divide the sandwiches into training and validation data, with the ratio of approximately 5:1. Since a sandwich

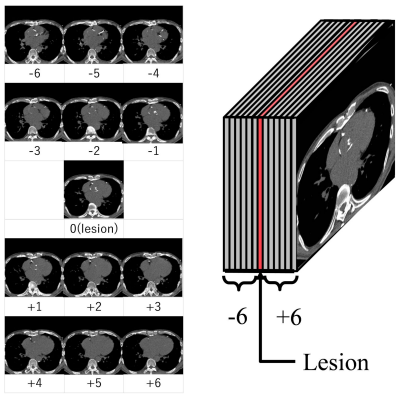


Fig. 4. Representations for non-contrast CT images (a) 2D representation of the 21 consecutive contrast CT slices, where 6 slices are hypothesized to inherit stenosis before and after the central lesion image (b) 3D representation of the S_k (=13) consecutive slices stick together.

structure is a 3D structure by nature, it is natural to construct a neural network that takes advantage of this 3D structure.

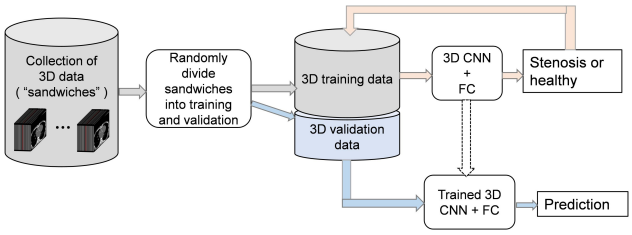


Fig. 5. Overall flow of our proposed methods using 3D representation of data ("sandwiches") with 3D CNN

One possibility might be to adopt a neural network that allows an LSTM [9] or Transformer [10] based sequence of 2D images. This approach has a small problem in that there are no pretrained weights available. Another possibility might be to utilize 3D PointNet [11] and its siblings, which have been primarily used in the robotics and CAD industries for retrieving and segmenting 3D shape objects for years. Unfortunately, the PointNet family requires us to prepare 3D shape objects before training. To the best of our knowledge, there are no 3D cardiovascular objects that are reconstructed from CT slices that can be used with the PointNet family.

Therefore, in this paper, we introduce video-based 3D Convolutional Neural Networks (CNNs). The main reason behind this idea is that we can take advantage of pretrained models with kinetics 400 [12], which is a popular dataset for video classification. Fig. 6 shows five 3D CNNs that we use to evaluate the the feasibility of the application to medical domains, specifically for stenosis detection. As shown in Fig. 6, five 3D CNNs are (a) R3D [13], (b) I3D [12], (c) MC3 [13], (d) R2Plus [13], and (e) S3D [14].

For each 3D CNN, we fine-tune and train the model with training data by adding extra FC (Fully Connected) layers

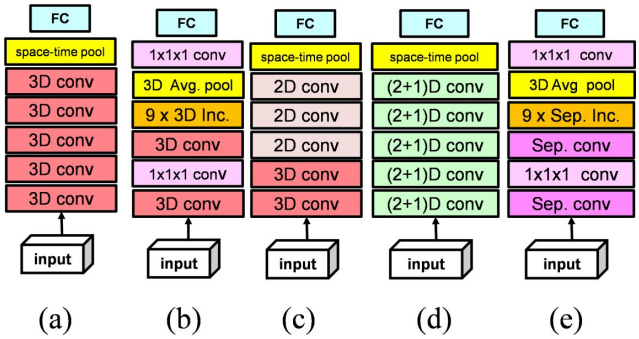


Fig. 6. Five 3D CNN models applied; (a) R3D, (b) I3D, (c) MC3, (d) R2DPlus, and (e) S3D

TABLE I
EVALUATION CRITERIA OF FIVE 3D CNNS WITH CONTRAST CT IMAGES

Model	Precision	Recall	F1	Specificity	Accuracy	Time
R3D	0.757	0.838	0.796	0.778	0.806	4.89
I3D	**0.972**	**0.886**	**0.927**	**0.969**	**0.924**	2.91
MC3	**0.972**	**0.886**	**0.927**	**0.969**	**0.924**	5.16
R2Plus	0.854	0.809	0.831	0.846	0.826	7.28
S3D	0.828	0.730	0.776	0.802	0.761	2.47

to produce two outputs (stenosis positive and negative (i.e. healthy)). After training the 3D CNN, we feed the validation data as unknown data to evaluate the prediction results.

V. EXPERIMENTAL RESULTS

In this section, we will describe the experimental results, and their analysis in terms of the performance with contrast and non-contrast CT images, as well as in terms of the performance with five 3D CNNs.

A. Overall results

The results of five 3D CNNs are summarized in Tables I and II with evaluation criteria; Precision, Recall (Sensitivity), F1, Specificity, Accuracy, and Time in 10^3 seconds for training with 15 epochs. Table I summarizes "Contrast" CT image data, while Table II summarizes "Non-Contrast" CT image data. Which 3D CNN is better depends on the kind of CT image data. In general, contrary to our intuition, the results with "Non-Contrast" CT image data yielded good performance.

We have compared the results with normalized confusion matrices as shown in Fig. 7 for the case of contrast CT images, and as shown in Fig. 8 for the case of non-contrast CT images. In addition, we demonstrate ROC (Receiver Operating Characteristic) curves [15] as well as DET (Detection Error Trade-off) curves [16] to compare the five 3D CNN models.

For the sake of comparison with 2D DNNs, including EfficientNetB2 [17], Vision Transformer [18], and MobileNet-V3-Large [19], Table III shows the results of these 2D DNNs on non-contrast CT image data. If we compare Table III with

TABLE II
EVALUATION CRITERIA OF FIVE 3D CNNS WITH NON-CONTRAST CT IMAGES

Model	Precision	Recall	F1	Specificity	Accuracy	Time
R3D	**0.993**	0.953	0.972	**0.993**	0.972	3.13
I3D	0.988	0.964	0.976	0.988	0.976	2.22
MC3	0.972	0.993	0.982	0.986	0.982	4.32
R2Plus	0.988	**0.994**	**0.991**	0.988	**0.991**	5.62
S3D	**0.993**	0.953	0.972	**0.993**	0.972	2.45

TABLE III
EVALUATION CRITERIA OF THREE 2D DNNS WITH NON-CONTRAST CT IMAGES

Model	Precision	Recall	F1	Specificity	Accuracy	Time
EFB2	0.774	0.813	0.793	0.784	0.798	8.27
ViT32	0.610	0.716	0.659	0.661	0.684	4.84
MBV3L	0.798	0.830	0.814	0.790	0.809	6.61

Table II, it is obvious that our proposed 3D CNN-based models outperformed 2D CNN-based models for stenosis detection in all the evaluation metrics.

B. Analysis of the results in terms of the performance with contrast and non-contrast CT images

As shown in Tables I and II, confusion matrices in Figs. 7 and 8, ROC curves in Figs. 9 and 10, and DET curves in Figs. 11 and 12, non-contrast CT slices (or sandwiches of size S_k) showed a much better performance in all evaluation criteria. A possible interpretation could be that a non-contrast CT image has less noise than a contrast CT image, while a contrast CT image has apparently more detailed and clearer coronary arteries, whereas it may have more noisy information, which may make it difficult to isolate coronary arteries.

C. Analysis of the results in terms of the performance with five 3D CNNs

Among five 3D CNN models, R2Plus performed best on non-contrast CT slices, although it took more time for training, while I3D and MC3 exhibited good performance on contrast CT slices. In particular, if we look at the ROC curve (d) as in Fig. 10 and the DET curve (d) as in Fig. 12 of R2Plus, both exhibit nearly ideal curves.

For contrast CT images, I3D and MC3 have almost the same best results with confusion matrices as shown in Fig. 7, while MC3 is slightly better than I3D in terms of DET curves as shown in Fig. 11 as the false positive rate (horizontal axis) of MC3 is less than that of I3D, even though MC3 took more time for training as shown in Table I. S3D has the worst result with respect to contrast CT images, while S3D and R3D have almost the same results with respect to non-contrast CT images.

VI. CONCLUSIONS

In this paper, we proposed a new method for applying video-pretrained 3D CNNs to predict the coronary artery stenosis. The input data of our proposed method is a collection of contrast and non-contrast cardiac CT images, while 3D CNNs we tested are R3D, I3D, MC3, R2Plus, and S3D.

We have introduced a "sandwich" 3D representation for a sequence of CT slices with a "lesion" image in the center. Specifically, for contrast CT images, we represent the lesion-centered images as a sandwich of size C_k, where in our implementation, we set $C_k = 21$, while for non-contrast CT images, we represent the lesion-centered images as a sandwich of size S_k, where in our implementation, we set $S_k = 13$. It should be noted that we have carried out other k for these sandwich structures. In fact, for S_k, we have tested with $k = 4, 5, 6, 7$, and turned out that $k = 6$ ($S_k = 13$) yielded the best among these four different sizes of sandwich structures.

To detect stenosis using the combination of 3D CNN models for the input data as contrast and non-contrast CT images, we have found that in terms of precision, recall (sensitivity), F1, accuracy, specificity, confusion matrix, ROC curve, and DET curve, the results with non-contrast CT images were better than those with contrast CT images.

Regarding the combination of a 3D CNN and a sandwich, the R2Plus 3D CNN resulted in the best performance with a sandwich of size 13, while MC3 resulted in the best performance with a sandwich of size 21.

ACKNOWLEDGMENT

A part of this research was carried out with the support of the Grant for Toyohashi Heart Center Smart Hospital Joint Research Course and the Grant-in-Aid for Scientific Research (C) (issue numbers 22K12149 and 22K12040).

REFERENCES

[1] D. Han, J. Liu, Z. Sun, Y. Cui, Y. He, and Z. Yang, "Deep learning analysis in coronary computed tomographic angiography imaging for the assessment of patients with coronary artery stenosis," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105 651, 2020, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2020.105651.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[3] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79.

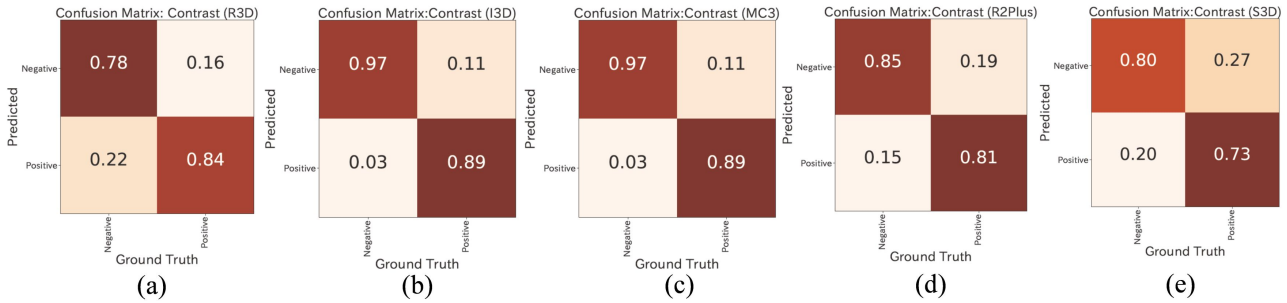


Fig. 7. Confusion matrices of five 3D CNNs with contrast CT images

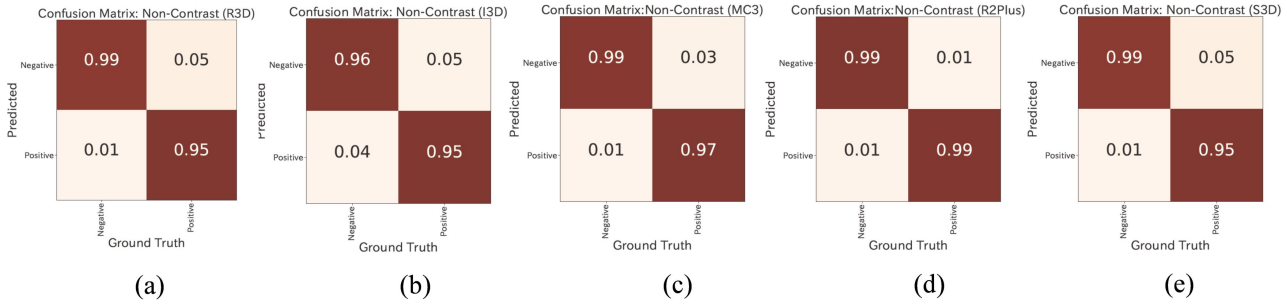


Fig. 8. Confusion matrices of five 3D CNNs with non-contrast CT images

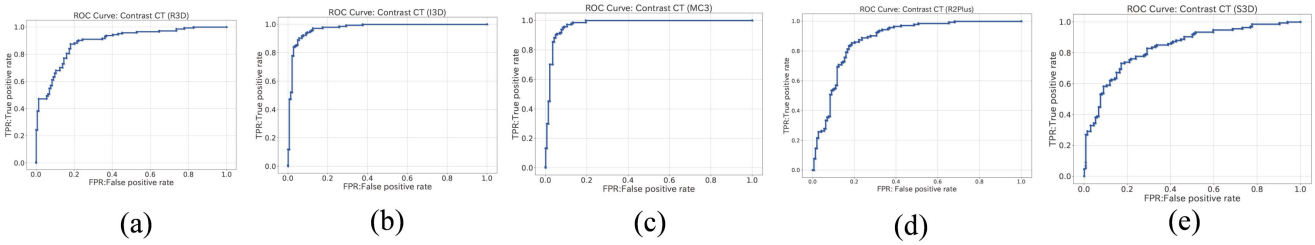


Fig. 9. ROC curve of five 3D CNNs with contrast CT images; (a) R3D, (b) I3D, (c) MC3, (d) R2Plus, and (e) S3D

[4] A. Lin, N. Manral, P. McElhinney, *et al.*, "Deep learning-enabled coronary ct angiography for plaque and stenosis quantification and cardiac risk prediction: An international multicentre study," *The Lancet Digital Health*, vol. 4, no. 4, e256–e265, 2022, ISSN: 2589-7500. DOI: https://doi.org/10.1016/S2589-7500(22)00022-X.

[5] A. R. Ihdayhid, A. Sehly, A. He, *et al.*, "Coronary artery stenosis and high-risk plaque assessed with an unsupervised fully automated deep learning technique," *JACC: Advances*, p. 100 861, 2024, ISSN: 2772-963X. DOI: https://doi.org/10.1016/j.jacadv.2024.100861.

[6] J. Yan, J. Tian, H. Yang, *et al.*, "A clinical decision support system for predicting coronary artery stenosis in patients with suspected coronary heart disease," *Computers in Biology and Medicine*, vol. 151, p. 106 300, 2022, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2022.106300.

[7] M. Aono, T. Asakawa, H. Shinoda, K. Shimizu, T. Togawa, and K. Nomura, "Predicting stenosis severity and localization in coronary artery using deep learning," in *Proceedings of the 2024 7th International Conference on Machine Vision and Applications*, ser. ICMVA '24, Singapore, Singapore: Association for Computing Machinery, 2024, pp. 117–125, ISBN: 9798400716553. DOI: 10.1145/3653946.3653964.

[8] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020. DOI: 10.1038/s41586-020-2649-2.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA:

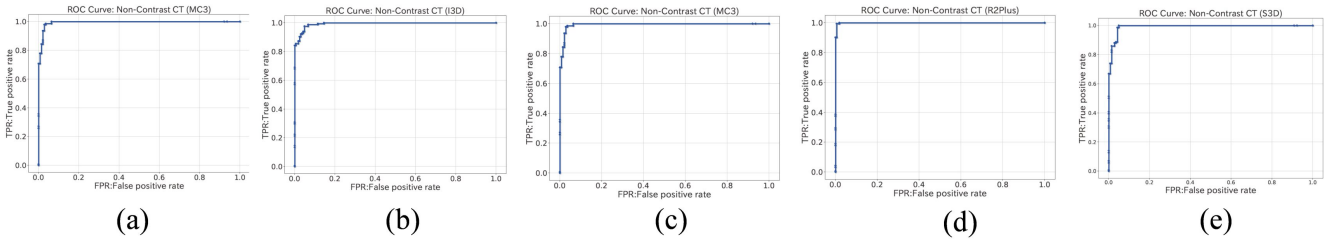


Fig. 10. ROC curve of five 3D CNNs with non-contrast CT images; (a) R3D, (b) I3D, (c) MC3, (d) R2Plus, and (e) S3D

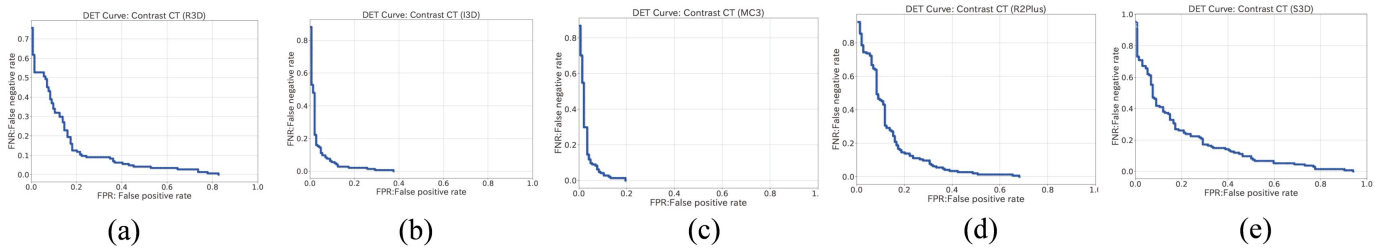


Fig. 11. DET curve of five 3D CNNs with contrast CT images; (a) R3D, (b) I3D, (c) MC3, (d) R2Plus, and (e) S3D

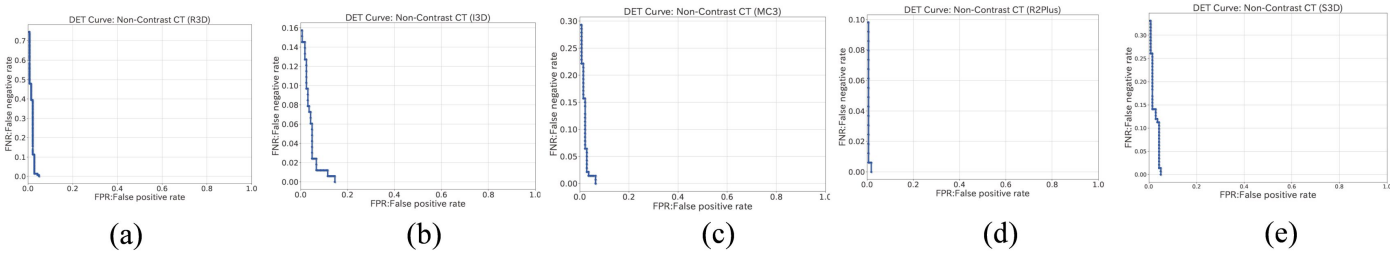


Fig. 12. DET curve of five 3D CNNs with non-contrast CT images; (a) R3D, (b) I3D, (c) MC3, (d) R2Plus, and (e) S3D

Curran Associates Inc., 2017, pp. 6000–6010, ISBN: 9781510860964.

[11] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[12] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[13] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[14] S. B. Hong, Y. H. Kim, S. H. Nam, and K. R. Park, "S3d: Squeeze and excitation 3d convolutional neural networks for a fall detection system," *Mathematics*, vol. 10, no. 3, 2022.

[15] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, ROC Analysis in Pattern Recognition, ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010.

[16] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The det curve in assessment of detection task performance.," in *EUROSPEECH*, G. Kokkinakis, N. Fakotakis, and E. Dermatas, Eds., ISCA, 1997, pp. 1895–1898.

[17] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, Sep. 2019, pp. 6105–6114.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[19] A. Howard, R. Pang, H. Adam, *et al.*, "Searching for mobilenetv3.," in *ICCV*, IEEE, 2019, pp. 1314–1324, ISBN: 978-1-7281-4803-8.

Real-time Segmentation of Coronary Artery Calcification Using Spatial Attention and Parallel Convolution

Tetsuya Asakawa* Yuki Sugimoto** Masashi Hashimoto‡ Takeshi Miyaji§ Kazuki Shimizu¶ Kei Nomura‖ and Masaki Aon
*†** Toyohashi University of Technology, Aichi, Japan
‡§¶‖ Toyohashi Heart center, Aichi, Japan
* E-mail: asakawa@kde.cs.tut.ac.jp
** E-mail: sugimoto.yuki.fd@tut.jp
‡ E-mail: m.hashimo6123@gmail.com
§ E-mail: miyaji@heart-center.or.jp
¶ E-mail: K
‖ E-mail: k.nomura@heart-center.or.jp
† E-mail: masaki.aono.ss@tut.ac.jp

***Abstract*—Semantic segmentation is an image recognition technique that classifies each pixel in an image. This technique is crucial in the fields of automated driving and medicine. Despite its significance, limited computational resources is available for its application. The current challenge involves to developing models that are accurate and fast, secondary lightweight. DDRNet (Deep Dual-Resolution Net) is a real-time segmentation method, which is a highly accurate method with a multi-resolution substructure. DDRNet is a model that to achieve efficiently both accuracy and real-time performance reducing the number of resolutions used in the subnetworks. However, DDRNet is less accurate than non-lightweight models, especially in small object recognition. Therefore, this research aimed to improve the accuracy of DDRNet while maintaining its light-weight and high computation speed. We conducted experiments on a coronary calcification dataset using DDR-SAPC-U as our proposed model, which is a modified version of the proposed method specialized for medical imaging with the addition of a skip connection. Overall, DDR-SAPC-U as our proposed demonstrated the highest accuracy compared to conventional medical methods and was relatively lightweight in terms of computational complexity. DDR-SAPC-U as our proposed model improved the accuracy compared to conventional methods.**

I. Introduction

Semantic segmentation is an image recognition technique that classifies each pixel in an image. This technology is widely applied role in various fields. Such as object recognition in autonomous driving, abnormal part detection in industrial inspection, and detection of organs and lesions in the medical field. In cardiac CT, this technology is applied for detailed detection of lesions called coronary artery calcification, which are strongly associated with coronary artery disease. However, the limited computational resources in cars and medical settings have prompted advancements in semantic segmentation using deep learning wherein the development of a model focuses not only on accuracy but also on lightness and speed.

HRNet (High-Resolution Net) [1] [2] is a powerful backbone in image recognition tasks, and methods applying it have recorded high accuracy in semantic segmentation, pose estimation, and object detection. Owing to the structure in which the sub-networks with multiple resolutions are connected in parallel, the computation amount is large and the processing speed is slow. DDRNet [3] is a real-time segmentation method inspired by HRNet and has a structure in which sub-networks with two different resolutions are connected in parallel. DDRNet is a model that to achieve efficiently both accuracy and real-time performance reducing the number of resolutions used in the subnetworks. However, DDRNet is less accurate than non-lightweight models, especially in small object recognition, especially in images. Therefore, this research aims to improve the accuracy of DDRNet by using coronary artery data while maintaining its lightweight nature and high computation speed.

The remainder of this paper is organized as follows. In Section 2, as related to research, we will explain a semantic segmentation method based on deep learning, a method using attention, one using multiple resolutions, and another for medical images. The describes the proposed method is described in Section 3. Results of comparative experiments between the proposed method and the conventional method are discussed in Section 4. In Section 5, we discuss the details of the experiment and conclusions and suggestions for further work are summarized in Section 6.

II. Related Work

In this Section, we review research literature related to the topic of this study. First, we describe research on medical images in semantic segmentation using deep learning. Thereafter, research on methods using attention and methods using multiple resolutions are described. Finally, we describe DDRNet, which is the baseline of this research.

A. Semantic segmentation

Semantic segmentation uses deep learning developed from FCN [4]. Thus far, DeepLab [5], [6], which improved accuracy by convolution using U-Net with Encoder-Decoder structure and multiple dilated rates. In addition, several methods have been proposed, such as PSPNet [7], which improves accuracy by using multiple sizes of Average Pooling.

Currently, numerous segmentation methods that pursue real-time performance. For example, SegNet [8], which is accelerated by using a small network structure and skip coupling, and spatial information is extracted from two networks called Spatial Path and Context Path, BiSeNet [9], and BiSeNet v2 [10], etc. The other studies, E-Net [11], ICNet [12], and SFNet [13] have proposed fast and lightweight semantic segmentation methods.

B. Medical images in semantic segmentation

Semantic segmentation has a great impact in the medical field, and is useful for detecting organs and lesions. UNet [14] is a simple Encoder-Decoder model featuring skip joins. FCN, the main semantic segmentation method proposed before UNet, yielded only low-resolution output.

On the other hand, UNet has an Encoder-Decoder structure, and an output with the same resolution as the input can be obtained. For this reason, it is often used in tasks in the medical field that require dense output, and has recently been applied to the detection of the brain, lungs, and heart. Some models based on the structure of UNet have also been proposed. Examples include UNet++ [15], which has a dense structure with multiple convolutions and skip connections added between the encoder and decoder, and TransUNet [16], which incorporates a transformer into the structure of UNet. These encoder-decoder structure methods have recorded high accuracy in tasks in the medical field.

C. Attention mechanism

In recent years, attention mechanisms [17] have been introduced not only in natural language processing but also in semantic segmentation. DANet [18] and PSA [19] are methods that focus on this. These methods have a network structure that uses both spatial and channel-oriented attention. In particular, PSA recorded high accuracy by combining OCR [20], which is a method that also applies attention, and HRNet, which has multiple resolution sub-networks.

In addition, Tao et al. [21] proposed a method that outputs spatial direction attention at multiple resolutions, and by using spatial attention, it was possible to extract information more appropriately for object regions than conventional convolutions. In this method, the element product of each attention map and feature map is taken to emphasize the area of the object and improve the accuracy. From this research, we obtained findings showing that the object to be focused on differs depending on the resolution of the image used.

III. Proposed Method

We propose two methods, Dual Attention Module (DAM) and Parallel Basic Block (PBB), which effectively extract features with less computation. Furthermore, we propose ding DDR-Spatial Attention Parallel Convolution (DDR-SAPC) to DDRNet-23-slim.

A. DDR-SAPC-U

We improved DDR-SAPC for application to medical image datasets. The improved network is called DDR-SAPC-U. Figure 1 is a schematic of the network of DDR-SAPC-U. There are two improvements from DDR-SAPC. The first point is the resolution of convolution. Even a small error in medical image segmentation can greatly change the diagnostic result. In addition, detection of lesions and abnormalities requires high accuracy. Therefore, we modified the Encoder-Decoder structure to output at the original resolution such as U-Net [14] and UNet++ [15]. This structure enables the fine detection of objects, and further improvement in accuracy can be expected. The second point is the addition of skip joins. In the encoder-decoder structure, the loss of features obtained on the encoder side occurs during convolution and upsampling. To prevent this, a skip join is introduced in the red-lined part of the figure. The introduction of skip coupling prevents the loss of detailed features. Furthermore, the detection performance is expected to improve.

1) Dual Attention Module: The Dual Attention Module (DAM) is a module inspired by Tao's method [21]. Figure 2 displays the network structure.

This module is decomposed into the part that generates spatial attention in Figure 2(a) and the part that extracts features by convolving feature maps in Figure 2(b). In DDR-Net, fusion of low and high-resolution was performed only in the part displayed in Figure 2(b). Therefore, we added a spatial attention generation part emphasizing the feature values of objects obtained at low and high-resolution. In the spatial attention generator, low-resolution and high-resolution attention maps AS and AL are generated by two layers of 3 × 3 conv, respectively. Then, $\mathrm{A_H}$+(1-A_L) for each element on the high-resolution side and $\mathrm{A_L} + (1 - \mathrm{A_H})$ on the low-resolution side is processed. The element product considers the resulting attention map and the original feature map. Generate a feature map by passing it through two layers of convolution and adding it to the inverse resolution.

2) Parallel Basic Block: PBB is a module devised to increase the types of resolutions in the convolutional layers inside DDRNet. DDRNet-23-slim uses the Basic Block in Figure 2(a) for the internal convolution block. This method is used in ResNet18 [22] etc. It is a very lightweight block that extracts features with two layers of convolution.

In contrast, the proposed method, Parallel Basic Block (PBB), has a structure in which two layers of convolution are performed for three kinds of resolutions [1/1, 1/2, 1/4]. As resolution is halved and area is squared, the amount of processing is exponentially small. The purpose of PBB is to

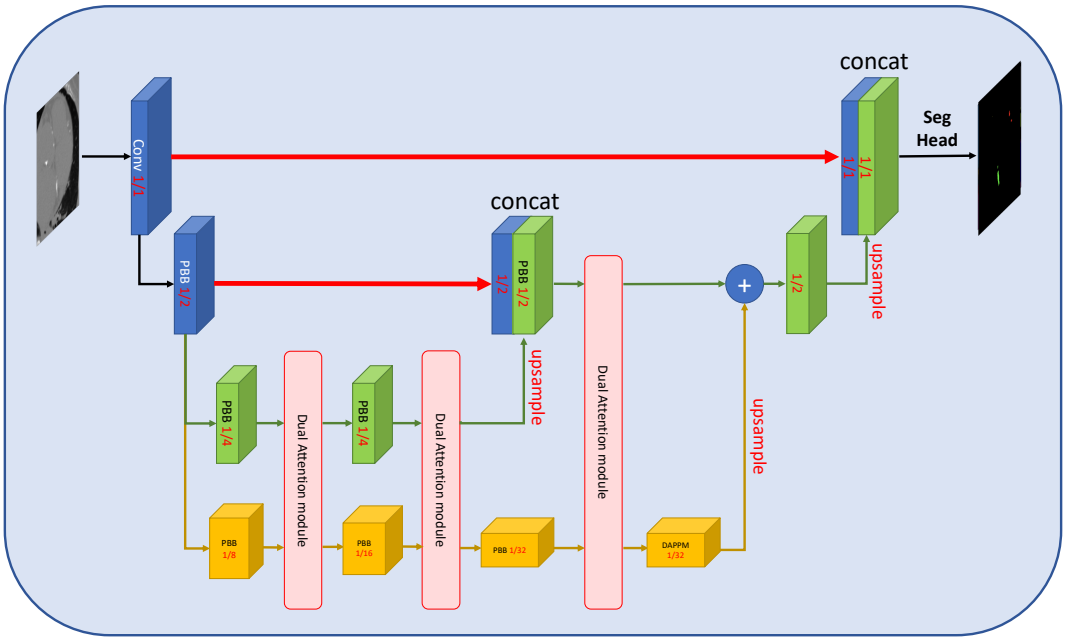


Fig. 1. DDR-SAPC-U schematic diagram Changes from DDRNet are shown in red. The stride of the convolution is reduced and the feature resolution is increased.

effectively extract features using multiple resolutions while the increase in processing amount is maintained to a small amount.

IV. Experiment

A. Dataset

In this paper, we used a cardiac CT coronary artery calcification data set to verify that the proposed method can be applied to medical images depicted in Figure 3. This dataset was created for detecting coronary artery calcification [23] [24] [25], which is considered to be highly related to coronary artery disease. We manually performed the annotation. The objects of label are the left anterior descending artery and calcifications present in three coronary arteries: left anterior descending artery (LAD, LMT), left circumflex artery (LCX), and right coronary artery (RCA) In this dataset, there are four classes including the class without calcification. The number of training verification data is 5600 (obtained from 100 patients), and the number of test data is 1240 (obtained from 21 patients).

B. Pretreatment and training method

The coronary artery calcification dataset was not preprocessed during training. The CT image was input as is. SGD was used as the optimization method, and momentum and weight decay were set to 0.9 and 0.0005. The initial value of the learning coefficient was 0.01, and training was performed with a batch size of 4 and an epoch number of 50. OHEM CE loss was the loss function.

The model was evaluated by inputting test data for the coronary artery calcification data set and examining the detection accuracy and detection speed, computational load, and the number of parameters. The mean Intersection over Union (mean IoU) was considered as an evaluation index for detection accuracy.

C. Visualization of detection results at coronary artery calcification dataset

Visualization example of the detection results in the coronary artery calcification test data set shows as below. Figure 3(b) shows the classification of classes in the correct and output images. From the upper left of the figure, the input image, the output of UNet, the output of UNet++, and from the lower left the correct image, the output of TransUNet, and the output of DDR-SAPC-U. First, in Figure 4(a), UNet, UNet++, and TransUNet incorrectly detect aortic calcification as RCA calcification for a correct image with LAD and LCX calcification. In comparison, DDR-SAPC-U improved this false detection. Furthermore, LCX calcification was also detected in a form close to the correct image. Figure 4(b) displays an output diagram for images with LAD, LCX, and RCA calcifications. UNet and UNet++ failed to detect LCX calcification, however TransUNet and DDR-SAPC-U detected it correctly.

D. Experiment results

1) Testing Environment: For the proposed models, all the training and testing pipelines, as well as baselines were implemented using PyTorch 1.6.0+cu92 framework in a Python 3.8.5 virtual environment. The graphics processing unit used in the training pipeline was NVIDIA Quadro P6000, and the RAM was 24 GB. Jupyter notebooks were utilized to conduct the experiments.

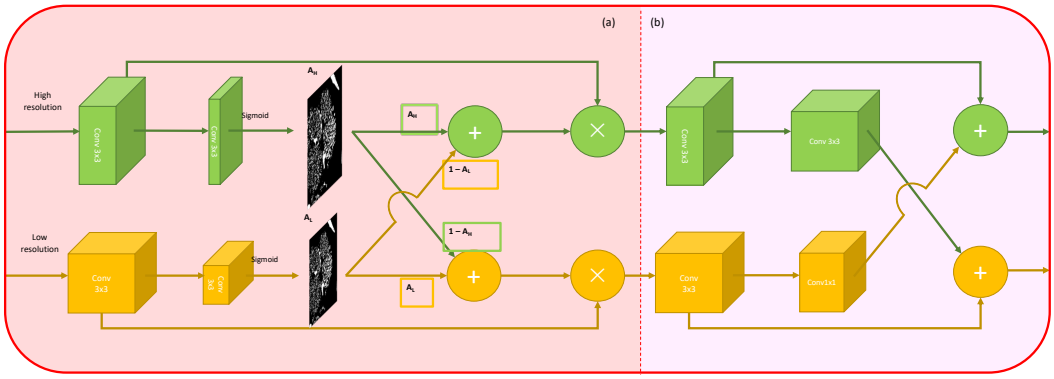


Fig. 2. Dual Attention Module. Learn more about the Dual Attention Module. This method adds the structure of part (a) to part (b) of the conventional method. Part (a) generates spatial attention at both high and low resolutions for the purpose of emphasizing various objects.

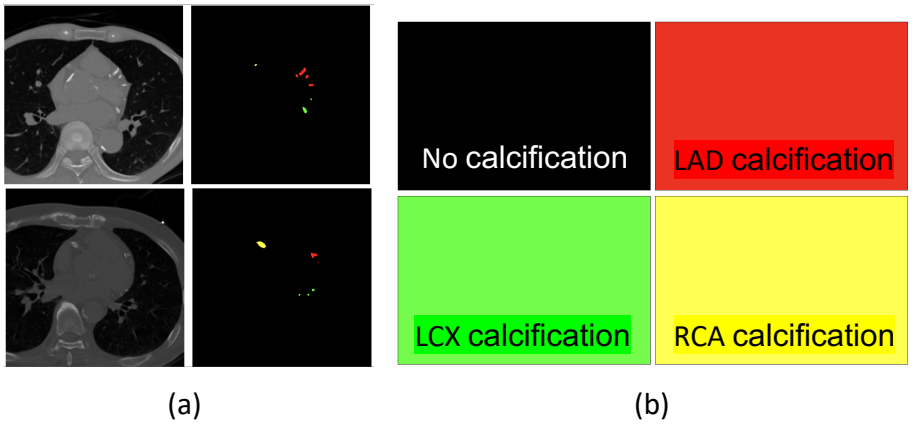


Fig. 3. Coronary Artery Calcification Dataset Image Example. (a) Left side: Input images. Right side: Correct images.(b) Coronary calcification dataset color map. Black indicates no calcification, red indicates LAD calcification, green indicates LCX calcification, and yellow indicates RCA calcification class.

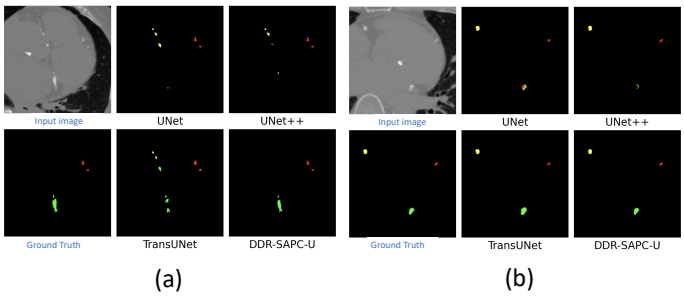


Fig. 4. Output images by model in coronary artery calcification dataset Black indicates no calcification, red indicates LAD calcification, green indicates LCX calcification, and yellow indicates RCA calcification class.

2) Experiment results: Using the proposed method DDR-SAPC-U, we conducted experiments according to the method described above. We added UNet [14], UNet++ [15], and TransUNet [16] as conventional methods for semantic segmentation tasks in the medical field and comparatively evaluated them.

The results are listed in Table I. The proposed method recorded high accuracy in all classes compared to other methods. FLOPs were relatively low values, indicating that the proposed method is superior in terms of accuracy, computational complexity, and lightness.

V. Discussion

DDR-SAPC-U recorded relatively fewer FLOPs and maximum accuracy in the Table I . In particular, the recognition accuracy has improved in the calcification classes located in the LAD and LCX. These calcifications tend to be difficult to recognize, especially at the branching points of the coronary arteries. The feature quantity was effective in the spatial direction by DAM. RCA improved accuracy less than that other methods. This is probably because of the shallow convolutional layers and the lightweight structure of the proposed method, DDR-SAPC. Therefore, the accuracy can be improved by adding more channels in the convolution and deepening the convolution layers.

VI. Conclusion

This research proposed a fast segmentation method, DDR-SAPC, which adds spatial attention and convolution with low resolution. Compared with UNet, UNet++, and TransUNet,

TABLE I
CORONARY ARTERY CALCIFICATION TEST DATASET RESULTS

Architecture	IoU by class				mIoU	GFLOPs	Params
	no calcification	LAD	LCX	RCA			
Unet [14]	99.99	80.74	52.55	73.28	76.64	**54.62**	**7.76M**
Unet++ [15]	99.99	76.58	55.97	80.61	78.29	137.95	9.16M
TransUnet [16]	99.98	66.41	42.83	63.67	68.23	62.19	20.07M
DDR-SAPC-U (Our proposed model)	99.99	**88.33**	**61.64**	**80.65**	**82.65**	60.15	13.42M

which are conventional methods in the medical field, the proposed method demonstrated the highest accuracy. Furthermore, it is relatively lightweight in terms of computational complexity and has been validated to be effective in the medical field as well. And we contribute that the proposed model is effective for segmentation in a wide range of segmentation applications. Future tasks include adding channel direction attention to the network, improving the accuracy of difficult classes such as walls and fences, and changing the network structure.

ACKNOWLEDGMENT

A part of this research was carried out with the support of the Grant for Toyohashi Heart Center Smart Hospital Joint Research Course and the Grant-in-Aid for Scientific Research (C) (issue numbers 22K12149 and 22K12040) and Knowledge Hub Aichi in Priority Research Project DX "Realization of a smart hospital that brings together IT and AI technologies".

REFERENCES

[1] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[2] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *CoRR*, vol. abs/1908.07919, 2019. [Online]. Available: http://arxiv.org/abs/1908.07919

[3] Y. Hong, H. Pan, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," *CoRR*, vol. abs/2101.06085, 2021. [Online]. Available: https://arxiv.org/abs/2101.06085

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: http://arxiv.org/abs/1606.00915

[6] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017. [Online]. Available: http://arxiv.org/abs/1706.05587

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016. [Online]. Available: http://arxiv.org/abs/1612.01105

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *CoRR*, vol. abs/1511.00561, 2015. [Online]. Available: http://arxiv.org/abs/1511.00561

[9] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," *CoRR*, vol. abs/1808.00897, 2018. [Online]. Available: http://arxiv.org/abs/1808.00897

[10] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet V2: bilateral network with guided aggregation for real-time semantic segmentation," *CoRR*, vol. abs/2004.02147, 2020. [Online]. Available: https://arxiv.org/abs/2004.02147

[11] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *CoRR*, vol. abs/1606.02147, 2016. [Online]. Available: http://arxiv.org/abs/1606.02147

[12] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," *CoRR*, vol. abs/1704.08545, 2017. [Online]. Available: http://arxiv.org/abs/1704.08545

[13] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, and Y. Tong, "Semantic flow for fast and accurate scene parsing," *CoRR*, vol. abs/2002.10120, 2020. [Online]. Available: https://arxiv.org/abs/2002.10120

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[15] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *CoRR*, vol. abs/1807.10165, 2018. [Online]. Available: http://arxiv.org/abs/1807.10165

[16] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR*, vol. abs/2102.04306, 2021. [Online]. Available: https://arxiv.org/abs/2102.04306

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[18] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," *CoRR*, vol. abs/1809.02983, 2018. [Online]. Available: http://arxiv.org/abs/1809.02983

[19] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise regression," *CoRR*, vol. abs/2107.00782, 2021. [Online]. Available: https://arxiv.org/abs/2107.00782

[20] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," *CoRR*, vol. abs/1909.11065, 2019. [Online]. Available: http://arxiv.org/abs/1909.11065

[21] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *CoRR*, vol. abs/2005.10821, 2020. [Online]. Available: https://arxiv.org/abs/2005.10821

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[23] M. J. Budoff and K. M. Gul, "Expert review on coronary calcium," *Vascular Health and Risk Management*, vol. 4, no. 2, pp. 315–324, 2008, pMID: 18561507. [Online]. Available: https://www.tandfonline.com/doi/abs/10.2147/vhrm.s1160

[24] S. Matsuoka, T. Yamashiro, A. Diaz, R. S. J. Estépar, J. C. Ross, E. K. Silverman, Y. Kobayashi, M. T. Dransfield, B. J. Bartholmai, H. Hatabu, and G. R. Washko, "The relationship between small pulmonary vascular alteration and aortic atherosclerosis in chronic obstructive pulmonary disease: quantitative CT analysis," *Acad. Radiol.*, vol. 18, no. 1, pp. 40–46, Jan. 2011.

[25] M. J. Blaha, J. Yeboah, M. Al Rifai, K. Liu, R. Kronmal, and P. Greenland, "Providing evidence for subclinical cvd in risk assessment," *Global Heart*, vol. 11, no. 3, pp. 275–285, 2016, legacy of Multiethnic Study of Atherosclerosis. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211816016307086

Effective Speech Data Augmentation Method To Improve Customer Service Representative Speech Recognition System Performance

Huiyong Bak* and Changhyeon Jeong†
* LOTTE INNOVATE Corp, Republic of Korea
E-mail : huiyong.bak@lotte.net
† LOTTE INNOVATE Corp, Republic of Korea
E-mail : Changhyeon.Jeong@lotte.net

***Abstract*—In this paper, we propose two data augmentation methods that effectively improve the performance of the customer service representative (CSR) speech recognition system. Building an CSR speech recognition system includes high transcription costs for speech data. Additionally, for use in training, long form CSR speech data is split into one sentence. The performance of the CSR speech recognition system is low because the split data does not contain conversational features between the CSR and the customer. The data augmentation method proposed in this paper uses untranscribed CSR speech data by transcribing it with WhisperX. Additionally, the two CSR speech data divided into one sentence are combined into speech data in the form of a conversation between two people. To verify the proposed data augmentation method, two CSR speech data from AI-HUB were used. As a result of the verification, it was confirmed that the performance of the CSR speech recognition system was improved when the proposed speech data augmentation method was applied.**

I. Introduction

Speech recognition is an important technology for communication between people and machines. It is a technology that converts speech into text and is used in various applications such as AI call center (AICC) and meeting minutes systems. Research to build a speech recognition system is actively being conducted [1][2]. Automatic speech recognition (ASR) can be divided into traditional hybrid ASR based on machine learning and end-to-end method. Hybrid ASR system consists of an acoustic model, a language model, and a pronunciation dictionary. Hybrid ASR uses each model by training it and merging it into a single weighted finite state transducer (WFST) graph. Since the models are trained separately, each model ensures optimal performance during training, but cannot guarantee optimal performance when merging them into one WFST graph. Additionally, it has the limitation of having to forcefully align speech and text for training [3]. End-to-end speech recognition is a technology that simplifies and improves the hybrid speech recognition process. Unlike hybrid ASR, end-to-end speech recognition converts input speech into text using a single model without an intermediate model carefully designed by humans. In addition, training is possible only with speech and transcribed text pair corresponding to the speech, so there is no need for forced alignment of speech and text. Several studies reported that end-to-end speech recognition outperforms traditional speech recognition [4][5][6]. With these reasons end-to-end speech recognition is widely used in business areas that require speech recognition, such as CSR speech recognition.In most businesses, customer services are operated to resolve customer complaints. Customer service have played an important role in maintaining smooth relationships between companies and customers. In a customer service, CSR process requests through calls with customers. At this time, since the CSR writes conversation information while talking on the phone with the customer, the CSR work's fatigue increases. Recently, speech recognition technology has been applied to customer services, reducing the fatigue of worker. By applying ASR technology to customer services, CSR workers can concentrate more on the call and perform other tasks while on the phone because they do not write the contents of the consultation. Additionally, because the call details are automatically written, workers can effectively understand the details of the consultation when viewing the consultation details later. Additionally, in CSR speech recognition result is used as input to the consultation quality evaluation system. Since the introduction of a CSR speech recognition system has the advantage of increasing the CSR work efficiency, research on CSR speech recognition has been conducted.

Ha et al. [7] released ClovaCall, a Korean CSR dataset, to build a Korean CSR speech recognition system. ClovaCall consists of 60,000 pairs of text and corresponding speech by more than 11,000 people in a restaurant reservation domain. Ha et al. built a CSR speech recognition system by training the end-to-end speech recognition models Deep speech2 and LAS with ClovaCall. Valizada et al. [8] built the CSR speech recognition system by training a hybrid speech recognition system from data collected from an actual emergency call center. Additionally, Valizada et al. improved ASR performance by modifying the transcribed text before training. Parra-Gallego et al. [9] confirmed that noise in data collected from an actual customer service deteriorates the performance of speech recognition and improved the performance of speech recognition by removing noise from the data using a GRU-based noise remover. Despite these previous studies, there are

many difficulties in building a CSR speech recognition system. To utilize CSR speech data in a speech recognition system, the data must be transcribed. speech data transcription requires a lot of money and time because it is transcribed manually. Additionally, CSR speech data is long form data of about 3 minutes. This long form speech data is divided into one sentence each for training the speech recognition model. When CSR speech data is split into a one sentence, the conversational speech data between the CSR and customer becomes speech data from a single speaker. For this reason, the data used for training does not have the characteristics of data in which two speakers speak and converse. Therefore, performance is not good when recognizing CSR speech data uttered by two speakers.

In this paper, we propose a new speech data augmentation method to effectively build a CSR speech recognition system. The first method is to transcribe the CSR speech data with WhisperX and then use the transcribed speech data as training data. Through the first augmentation method, it is expected that the cost required to transcribe CSR speech data can be reduced. Because WhisperX uses a fixed structure, the structure of WhisperX cannot be changed to suit each task. Therefore, in this study, WhipserX was used only for data augmentation. CSR speech data is long form data over 3 minutes. When building an actual CSR speech recognition system, CSR speech data split into one sentence is used. This is because long form audio data (3+ minutes) is difficult to use for training because it uses a very large amount of GPU-VRAM. When CSR speech data is split into a one sentence, the conversational speech data between the CSR and customer becomes speech data from a single speaker. This split data does not have the characteristics of data in which two speakers speak and converse. Therefore, a CSR speech recognition model trained using CSR speech data split into one sentence has a lower recognition rate on actual CSR speech data. To improve this limitation, in this study propose a second speech data augmentation method. The second method is to select two CSR speech data split into one sentence and then merge them into one data containing the conversation of two speakers. To achieve this purpose, each of the two CSR speech is split into word-unit speeches. Then, the word-level speech from the two sentences is randomly mixed and combined into one speech sentence. By training a speech recognition model using the data generated by the second augmentation method, the conversation characteristics of the two speakers can be trained.

II. Materials and Methods

Chapter 2 covers the overall process of the augmentation method proposed in this paper. Section A describes WhisperX used in the proposed augmentation method. Section B describes the speech recognition model used in the experiment, and Section C describes the proposed augmentation method. Section D introduces the settings used in the experiment, and Section E describes the dataset used in the experiment and evaluation metric.

A. WhisperX

Recently, in the field of speech recognition, the Self-Supervised Learning (SSL) method has achieved good performance and is used in many speech recognition models. The self-supervised method pre-trains speech representations from unlabeled speech data and fine-tunes them with a small amount of labeled speech data. But this method has the limitation that it builds a speech recognition model that is only suitable for small amounts of labeled data. To overcome these shortcomings, OpenAI released Whisper, which learned with a dataset of 680,000 hours on a basic Transformer structure without using SSL. Whisper learns from big data sets, enabling robust voice recognition even in environments where it is difficult to recognize with existing voice recognition methods. But Whisper only can use speech data of 30 seconds or less. Additionally, it provides timestamps for utterance units and does not provide timestamps for word units [10].

WhisperX was announced to overcome the limitations of Whisper mentioned above. As shown in Fig. 1, WhisperX divide input audio through Voice Activity Detection (VAD) module. The divided audio is grouped into one batch through Cut and Merge, and the batch are transcribed and merged into one text through Whisper. Additionally, the start and end positions of each word are obtained through Phoneme Model and Forced Alignment. Phoneme Mode classifies the input speech phoneme. Optimal time path for the phoneme obtains the using the Dynamic Time Warping (DTW) algorithm of Forced Alignment. Using the optimal time path and phoneme classification results, the time of the first and last phonemes of the word is obtained [11]. The two augmentation methods proposed in this paper used WhisperX. This is explained in detail in Section C.

B. Fast Conformer

The speech recognition model used to build the CSR speech recognition system in this study is Fast Conformer. Conformer is a speech recognition model that is currently widely used in speech recognition. Conformer is a model that combines Transformer, which is good at capturing the global information of data, and CNN, which is good at capturing the local information. Conformer consists of a multi-head self-attention module to capture global information of data and a convolution module to capture local information. Multi-head self-attention module uses relative positional embedding with self-attention. Relative positional embedding provides relative positional information according to the length of the input data and performs self-attention. This allows ASR model to generalize location information from different lengths of data for effective training.

Fast Conformer is an improved model of Conformer and follows the structure of Conformer. Fast Conformer reduces input data in half compared to Conformer. This allows for faster training and inference compared to Conformer. Additionally, by applying local attention to only a certain part than the entire, GPU V-RAM can be used effectively [12].

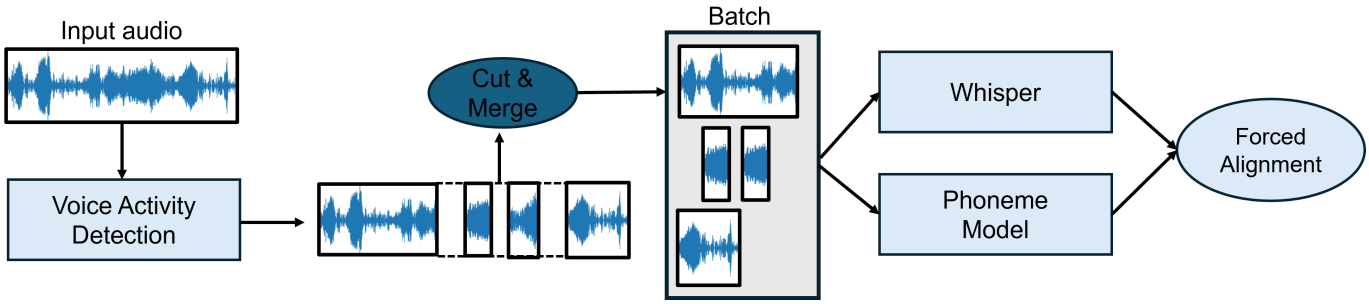


Fig. 1. Structure of WhisperX.

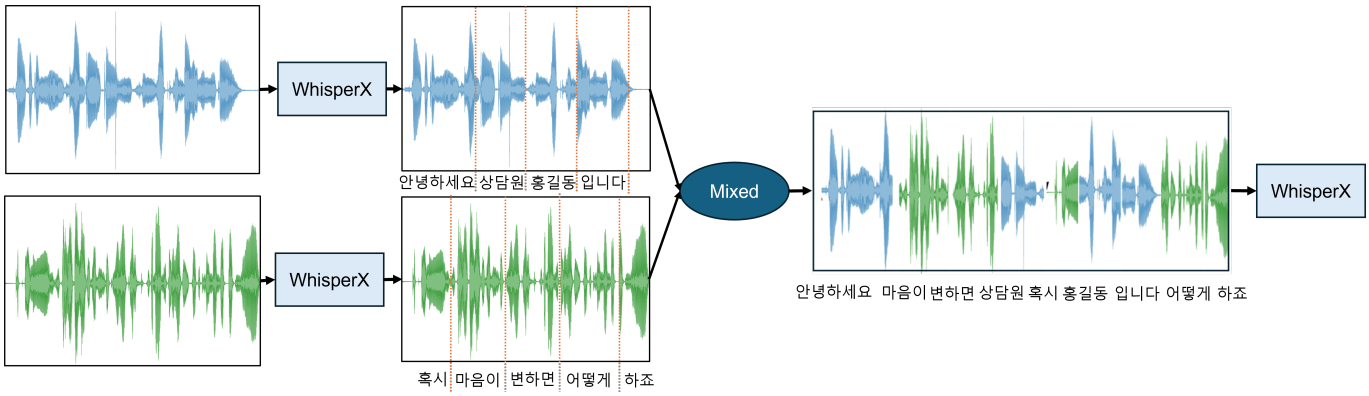


Fig. 2. Structure of the second proposed speech augmentation method.

C. Proposed speech data augmentation method

This paper proposes two speech data augmentation methods using WhisperX. The first data augmentation method utilizes untranscribed speech data. untranscribed speech data is transcribed through WhisperX. Speech data with transcribed texts through WhisperX are used as training data. Radford, Alec, et al. By comparing Whisper's transcription results with those of humans, they confirmed that Whisper has a transcription ability like that of humans [10]. This study expects to reduce the cost of CSR data transcription through the first data augmentation method. The second data augmentation method is shown in Fig. 2. The second data augmentation method uses two speech data uttered by different speakers. Each speech data is segmented data that has been cut into sentences in the original recording data. The second method inputs speech data uttered by two speakers into WhisperX, as shown in Fig. 2. Through WhisperX, transcription results with word-level timestamps are obtained. The two pieces of speech data are splited into word speech units using word-level timestamps. Divided word-level speech data is randomly mixed to generate new single sentence of speech data. The generated speech data is transcribed again with WhisperX to increase transcription accuracy, and the data generated through this method become training data for the ASR model. The speech data generated by the second augmentation method contains the characteristics of data in which two speakers utter and conversation because the speech data of the two speakers is mixed word-by-word.

D. Experimental setup

The speech used in the experiment were converted to Mel Spectrogram. To convert speech to spectrogram, a window size of 25 ms and a window shift of 10 ms were used. The speech recognition model used in the experiment is Fast Conformer. Fast Conformer consists of a model dimension of 512, 8 attention heads, and 17 layers. Fast Conformer was trained using AdamW optimizer and a cosine scheduler with weight decay of 0.01. Additionally, we set the batch size to 24 and trained with four RTX 3090s.

E. Datasets and evaluation metric used in the experiment

For the experiments, in this study used the CSR dataset and the CSR data in the welfare field. Both data are publicly available on AI-Hub [13][14]. Both datasets provide CSR speech data split into one sentence units from long speech files longer than 3 minutes. In this paper, CSR data is now denoted as the dataset-1, and CSR data in the welfare field is denoted as the dataset-2. dataset-1 is speech data recorded by contacting the customer service via wired or wireless. Dataset-1 consists of 1,000 hours of speech files from the education domain, 1,000 hours from the finance domain, and 1,000 hours from the telemarketing domain. Dataset-2 consists of 2,945 hours of audio and labeled data collected from university hospitals, regional mobility support centers, and mental health counseling centers. Dataset-2 is 2,276,065 pieces of speech data, uttered by 76 men and 202 women. In this study, 1000 hours of data

were used in each dataset. Of the 1,000 hours used, 500 hours used speech and text corresponding to the speech, and the remaining 500 hours used only speech data.

Test data from each dataset was used for testing in the experiment. The test data in each dataset is data that split from long form to one sentence. Therefore, the test dataset is different from the real-world CSR speech data form. In this study re-combined the split test data into the originally long form. To do this, in this study utilized the meta information of the test data. The meta information has the split information from the originally form. From now on, in this paper, 500 hours of speech and corresponding text to will be denoted as train-data. And the 500 hours of speech-only data at 500 hours will be denoted as method-1-data. Additionally, the data generated by applying the second augmentation method proposed to train-data and method-1-data will be denoted as method-2-data.

The evaluation metric used in this experiment is character error rate (CER). CER is the syllable error rate, which measures the error rate based on the number of substitutions, deletions, and insertions for each syllable in a sentence. Korean uses prepositional particles, has a complex morpheme structure, and spacing boundaries are ambiguous. Due to these characteristics of the Korean, it is difficult to evaluate the error rate at the word level. Therefore, CER, which measures errors at the syllable level, is widely used in Korean speech recognition. The recognition rate was measured by measuring the edit distance from the correct text and the recognition results text of the speech recognition model.

III. Results and Discussion

Section A describes the experiments and results of experiments.

A. Experiment

To validate the effectiveness of the proposed speech data augmentation method, this study conducted three experiments. The experimental results are shown in Table 1. The first experiment assumes a situation where a lot of CSR data cannot be used because transcription costs are high. Therefore, the first experiment builds a CSR speech recognition system by training Fast Conformer with 500 hours of data. The second experiment builds the CSR speech recognition system by adding CSR speech data generated by the first proposed augmentation method. The third experiment is to verify the impact of conversational data from two speakers on CSR speech recognition performance.

The first experiment builds a basic CSR speech recognition system by training with Fast Conformer using train-data. The recognition rates of the two test data were measured using the built CSR speech recognition system. As a result of the first experiment, dataset-1 was recognized as 14.23% and dataset-2 as 13.02%. In the first experiment, only 500 hours of data were used, so the recognition rate was not good.

The second experiment is to verify the first data augmentation method. Fast Conformer was trained with train-data and method-1-data. As a result of the second experiment, dataset-1 was recognized as 11.44% and dataset-2 as 7.94%. The CSR speech recognition system built through the second experiment improved the performance of dataset-1 by 19.6% and the performance of dataset-2 by 39.0%. This shows that using the data generated by the first proposed augmentation method for training improves CSR speech recognition performance. The results of this study are expected to enable CSR speech recognition systems to be built with relatively low transcription costs.

TABLE I
Experiment result

	Dataset-1	Dataset-2
Exp 1	14.23%	13.02%
Exp 2	11.44%	7.94%
Exp 3	9.63%	3.98%

The third experiment is to verify the second data augmentation method. In the third experiment, Fast Conformer was trained with train-data, method-1-data, and method-2-data. As a result of the third experiment, dataset-1 was recognized as 9.63% and dataset-2 as 3.98%. The CSR speech recognition system built through the third experiment improved the performance of dataset-1 by 15.8% and the performance of dataset-2 by 49.8%. This shows that the conversational data between the two speakers improves the performance of the CSR speech recognition system.

IV. Conclusions

A CSR speech recognition system is necessary because it improves the CSR work efficiency. In this paper, we propose two data augmentation methods that effectively increase the performance of the CSR speech recognition system. The first proposed augmentation method transcribes the untranscribed CSR speech with WhisperX and then uses the pair of speech-transcribed text data for training. The second method of augmentation combines the two split CSR speech into a single sentence. The combined data is used for training. The performance of the CSR speech recognition system was improved when the proposed data augmentation methods are applied. In this study, we proposed a method that utilize untranscribed CSR speech and a method that utilize CSR speech data split into single sentence. It is expected to reduce the cost of CSR speech data transcription and improve the performance of the CSR speech recognition system built with data divided into one sentence.

Acknowledgment

This research was conducted with support from LOTTE INNOVATE.

REFERENCES

[1] G. Hinton, L. Deng, D. Yu, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, 2019.

[3] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The kaldi speech recognition toolkit. inieee 2011 workshop on automatic speech recognition and understanding. ieee signal processing society, dec. 2011," *IEEE Catalog No.: CFP11SRW-USB*,

[4] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[5] W. Han, Z. Zhang, Y. Zhang, *et al.*, "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," *arXiv preprint arXiv:2005.03191*, 2020.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[7] J.-W. Ha, K. Nam, J. Kang, *et al.*, "Clovacall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers," *arXiv preprint arXiv:2004.09367*, 2020.

[8] A. Valizada, N. Akhundova, and S. Rustamov, "Development of speech recognition systems in emergency call centers," *Symmetry*, vol. 13, no. 4, p. 634, 2021.

[9] L. F. Parra-Gallego, T. Arias-Vergara, and J. R. O. Arroyave, "Robust automatic speech recognition for call center applications," in *Workshop on Engineering Applications*, Springer, 2021, pp. 72–83.

[10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.

[11] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747*, 2023.

[12] D. Rekesh, N. R. Koluguri, S. Kriman, *et al.*, "Fast conformer with linearly scalable attention for efficient speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–8.

[13] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=100.

[14] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=470.

LLM as decoder: Investigating Lattice-based Speech Recognition Hypotheses Error Correction Using LLM

Sheng Li*, Yuka Ko† and Akinori Ito‡
* National Institute of Information and Communications Technology (NICT), Kyoto, Japan
† Nara Institute of Science and Technology (NAIST), Nara, Japan
‡ Tohoku University, Sendai, Japan
E-mail: sheng.li@nict.go.jp, ko.yuka.kp2@is.naist.jp, akinori.ito.a2@tohoku.ac.jp

***Abstract*—With the strong representational power of large language models (LLMs), generative error correction (GER) for automatic speech recognition (ASR) aims to provide semantic and phonetic refinements to address ASR errors. However, the previous LLM GER method used n-best lists, which are simpler, more efficient, and more vulnerable to the issue of hypothesis space collapse. Hypothesis space collapse is a phenomenon in speech recognition where the range of possible interpretations becomes too narrow. This paper proposes replacing n-best hypotheses with lattice, a more flexible ASR output format, to improve the LLM GER and reduce the likelihood of hypothesis space collapse. Experiments on CSJ corpus show that compared with using n-best hypotheses, using lattice can improve the performance of Japanese speech recognition.**

I. Introduction

Automatic speech recognition (ASR) is one of the most natural human-machine interfaces for various spoken language applications. However, various factors, such as background noise, accents, speech clarity, and the quality of the recording equipment, always influence ASR results. These influences can lead to inaccuracies in transcriptions, affecting the overall effectiveness and reliability of speech recognition systems.

The automatic generative error correction (GER) task uses pre-trained models to automatically revise and enhance written content, including documents, articles, and translations. This process involves correcting grammar, syntax, style, and coherence errors and tailoring language to specific requirements or preferences. GER can effectively refine ASR outputs by correcting transcription errors. This improves the quality and usability of transcribed text for applications like captioning, transcription services, and voice-controlled systems.

Pretrained large neural network-based language models (LMs) have been applied to GER tasks linked to ASR in recent years. To lower the substitution error in Mandarin speech recognition, Zhang et al. [1] suggested a spelling corrector based on the transformer [2]. The effectiveness of BERT [3] in identifying spelling errors was enhanced by [4], using the soft-masking technique as a link between the error detector and corrector. Using the BERT distilling knowledge, Futami et al. [5] created soft labels for ASR training. Additionally, certain studies have been conducted [6], [7] to enhance ASR rescoring by BERT. Furthermore, BERT has also been effectively used in multi-modal research in voice-language pretraining [8], [9], [10] or vision-language pretraining [11], [12], [13], [14].

More recently, Chen et al. [15] proposes combining large language models (LLMs) into a speech recognition system. However, n-best lists are more prone to hypothesis space collapse due to their fixed and limited number of alternatives. In speech recognition, "hypotheses space collapse" refers to a phenomenon where the range of possible interpretations of an input (the hypothesis space) becomes too narrow or restricted. This can lead to several issues: The system might consider only a limited set of possible transcriptions, ignoring other possible alternatives. This can reduce the recognition system's accuracy and robustness. As for the n-best list used in [15], once the list is generated, any transcription not in the top-N hypotheses is ignored, which can significantly narrow the hypothesis space. While n-best lists are simpler and more efficient, they are more vulnerable to hypotheses space collapse due to their limited representation ability.

Compared to the n-best list, lattices are less likely to hypothesize space collapse because they maintain a broader and more flexible representation of possible transcriptions. The lattice structure allows for more diverse hypotheses, reducing the likelihood of collapsing the hypothesis space. This paper addresses the issue of n-best hypotheses and enhances the LLM GER. As depicted in Figure 1, we replace n-best hypotheses with lattice, a more compact ASR output format.

II. Related Work

A. LLMs for ASR task

Chen et al. [15] have suggested integrating LLMs into a voice recognition system more recently. In this research, LLM is used for second-pass rescoring in the output transcriptions produced by the ASR system (n-best decoding hypotheses) for GER. This work presents LoRA [16], which allows for efficient learning of the n-best to transcription mapping without affecting the pre-trained parameters of the LLM. LoRA works by inserting a neural module with a small number of extra trainable parameters to approximate the full parameter updates, thereby avoiding the need to tune the entire set of parameters

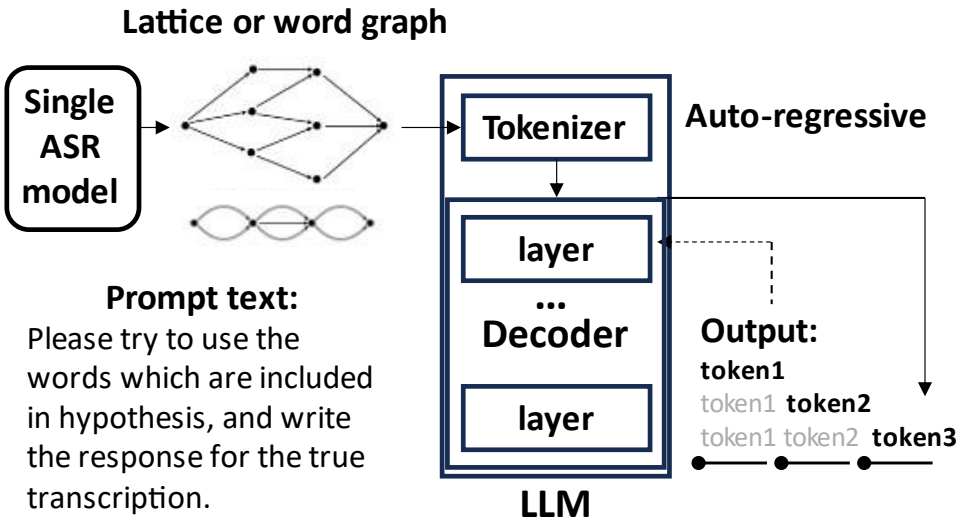


Fig. 1: The flowchart of the proposed LLM GER method using lattice output from ASR system instead of n-best hypotheses.

of a pre-trained model. Additionally, Li et al. [17] investigated knowledge transfer when fine-tuning a multilingual LLM to correct 1-best hypothesis errors generated by different speech foundation models across multiple languages. Ko et al. [18] benchmarked Japanese Speech Recognition on ASR-LLM Setups and proposed Multi-Pass augmented GER.

By adding trainable low-rank decomposition matrices to the current layers of LLMs, this technique allows the model to adjust to fresh input while maintaining the original LLMs' fixed structure to preserve prior knowledge. In particular, LoRA injects low-rank decomposition matrices to reparameterize each model layer, which is described as a matrix multiplication. The representations produced by the LLM are, therefore, not warped by task-specific tailoring. Simultaneously, the adapter module gains the capacity to forecast the actual transcription based on the n-best estimates. Thanks to effective training, the method makes use of a large-scale language model that should be able to comprehend the task description and identify correlation in the n-best list.

III. Improving LLM GER using Lattice for ASR

As mentioned in previous work [15], a significant challenge arises when previous LLM GER fails to yield improvements at relatively low Character Error Rates (CER). Our proposed method addresses this issue by enhancing LLM GER. As depicted in Figure 1, we replace n-best hypotheses with lattice, a more compact ASR output format. In speech recognition, a lattice is a graph representation of multiple potential word sequences generated during decoding. It captures various hypotheses of the spoken utterance, allowing for more accurate recognition by considering alternative paths and word sequences. Lattices are useful for handling ambiguity and improving recognition accuracy, especially in noisy or complex speech environments.

LLMs obtain the entire hypothesis space, which is compactly stored in lattice format from the speech recognition system, and then derive the 1-best hypothesis. In a sense, this accomplishes the traditional function of a speech recognition decoder. Therefore, we can say that the LLM is the decoder.

There are two lattice formats in speech recognition and machine translation areas as follows:

- Kaldi's lattice format is a representation of possible transcriptions from speech recognition, structured as directed acyclic graphs. Each node signifies a point in time, and each arc between nodes represents a hypothesized word with associated acoustic scores and transition probabilities. This format allows multiple hypotheses to be stored efficiently, enabling further processes like rescoring with more sophisticated models to improve transcription accuracy. These lattices can be pruned, determinized, and minimized for efficiency.

- The PLF (Python Lattice Format) used in Moses represents word lattices as directed acyclic graphs with nodes and edges labeled with words and weights. Nodes are ordered topologically and assigned numerical IDs, with edges indicating the word, probability, and distance between nodes. The format allows efficient decoding of multiple sentence hypotheses by Moses.

For LLM inferences, the PLF lattice format is generally more friendly than Kaldi's native lattice format. This is because PLF is designed to be easily readable and manipulable using Python, which is commonly used in LLM development. PLF's structure (nodes and edges with word and weight labels) aligns

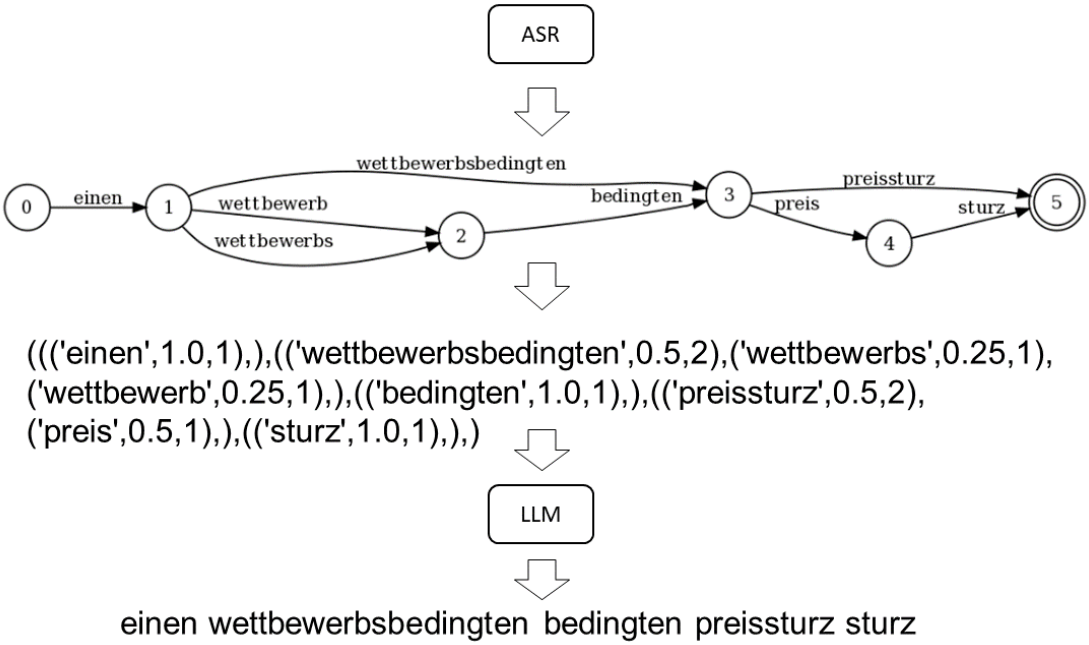


Fig. 2: Convert graph representation of lattice from ASR system to text-based format for LLM (example cited from Moses Users Manual [19], and the meaning is: a competition-related price drop).

well with typical data processing pipelines used in LLMs. In contrast, while efficient for speech recognition tasks, Kaldi's format is more complex and less intuitive for direct use with LLMs, as shown in Figure 2.

IV. Experiments

A. Experimental Settings

1. LLM used: The experiment employs the Japanese-Llama-2-7b model[1]. The Japanese-Llama-2-7b model is presumably a variant of the LLaMA (Large Language Model by Meta AI) adapted for Japanese language processing.
2. Dataset: We use the "Corpus of Spontaneous Japanese (CSJ)" [20]. Three official evaluation sets ($Eval_1$, $Eval_2$, and $Eval_3$), each containing ten lecture recordings [21], are used to evaluate the speech recognition results.
3. ASR model: We use the very deep resnet tdnn CTC model (similar to the Chain model in Kaldi) in our previous work [22], [23] for generating lattice for this paper. For training ASR models, we use approximately 250 hours of lecture recordings as the training set (Train APS (academic)) with similar settings according to other benchmarks [24], [25], [26], [21].
4. Training LLM GER models: Implementing our LLM GER models is based on the repository[2] for GER task [27]. For the LLM test set, we extracted each 424 utterances from each $Eval_1$ (2918), $Eval_2$ (3067), and $Eval_3$ (2484). The left $Eval_1$, $Eval_2$, and $Eval_3$ not included in the LLM test set were gathered and used as our LLM train set (7197 utterances in total). In CSJ data, the training is performed on an NVIDIA Tesla A6000 GPU using 8-bit training. The hyperparameters for finetuning are 17 epochs, learning rate 1e-4, batch size 64, and LoRA rank 4. We used an English-Japanese prompt based on automatic translation following the previous default English prompt.[3]
5. Evaluations tool: We use NIST-SCTK to evaluate the Word Error Rate (WER%).

B. Experimental Results

In this paper, we made four systems for experimental comparisons as follows:

- 1-best: This is the direct 1-best hypothesis output from the above-mentioned ASR model in the last subsection.
- N-best: We use the LLM GER method described in [15] to restore the n-best hypotheses (n=10). The top 10 hypotheses are extracted from the lattice using the default acoustic weight by the lattice-to-nbest tool [4] of Kaldi.
- Lattice: We use the LLM GER method with the same setting as the N-best experiments, just replacing the n-best hypotheses with the PLF-formated lattice hypotheses. Since the CSJ words are with part-of-speech (POS) and

[1]huggingface.co/elyza/ELYZA-japanese-Llama-2-7b
[2]https://github.com/Hypotheses-Paradise/Hypo2Trans

[3]We observed roughly lower loss and lower CER trends in English-Japanese prompt compared to the default English prompt (https://github.com/Hypotheses-Paradise/Hypo2Trans/blob/main/H2T-LoRA/templates/H2T-LoRA.json).
[4]https://kaldi-asr.org/doc/lattice-to-nbest_8cc.html

TABLE I: WER in CSJ Eval test sets. Each evaluation data consists of 424 utterances sampled from the original CSJ evaluation data.

	$Eval_1$ (424)	$Eval_2$ (424)	$Eval_3$ (424)
1-best	**5.6**	10.9	10.9
N-best	5.9	**10.4**	10.9
Lattice	6.0	**10.4**	**10.4**
Lattice + POS + Kana	9.4	18.9	12.3

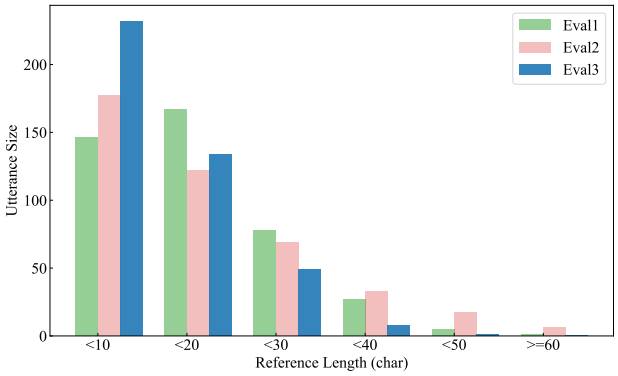


Fig. 3: The reference length distribution in $Eval_1$, $Eval_2$ and $Eval_3$.

Katakana pronunciations (Kana), we remove these POS and Kana from the PLF-format lattice.

- Lattice + POS + Kana: The LLM GER experimental settings are the same as the Lattice experiment but keep the POS and Kana from the PLF-format lattice.

Table I shows WERs in each CSJ evaluation set. We observe that the LLM GER lattice (without POS and Kana) results improved in eval02 and eval03 and resulted in overall improvement compared to n-best hypotheses. On the other hand, there was no improvement in the LLM GER lattice with POS and Kana for leveraging lattice rescoring in LLM. We expect that this tag information will become confusing for the LLM model we expect.

C. Further Discussions

1) Performance in each reference length regime: We further analyze the reason that n-best and lattice-based LLM GER methods could not outperform the 1-best hypotheses in $Eval_1$. Figure 3 shows the utterance numbers of $Eval_1$, $Eval_2$, and $Eval_3$ in each reference length regime. We see that $Eval_3$ consists of relatively shorter sentences compared to $Eval_1$. The $Eval_3$ has the most number of sentences, which are shorter than 10. Figures 4, 5 and 6 show the averaged WER scores in each reference length regime. These results show that the WERs by the LLM-GER with lattice were lower than those of 1-best for samples when the reference lengths are shorter than 10 in all evaluation sets. In $Eval_3$ and $Eval_2$, all WER scores were lower in length regimes shorter than 50. We expect that recovering errors from long sentences remains challenging for the LLM GER method. We also expect that using a small LLM (7B) may make it more difficult for the LLM GER to process longer inputs because it is trained with data. We expect that because there are more small size outputs in $Eval_2$ and $Eval_3$ (especially when the reference length is shorter than 10) than $Eval_1$, the LLM GER lattice results became better in $Eval_2$ and $Eval_3$.

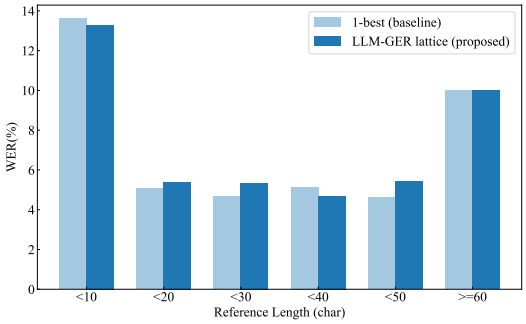


Fig. 4: WERs in $Eval_1$

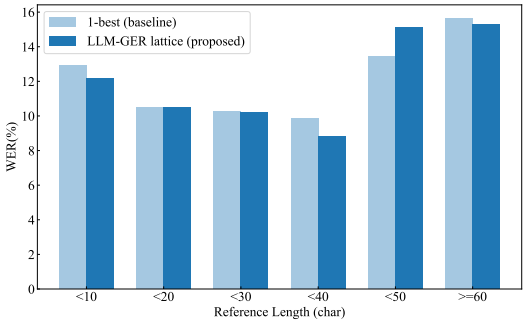


Fig. 5: WERs in $Eval_2$

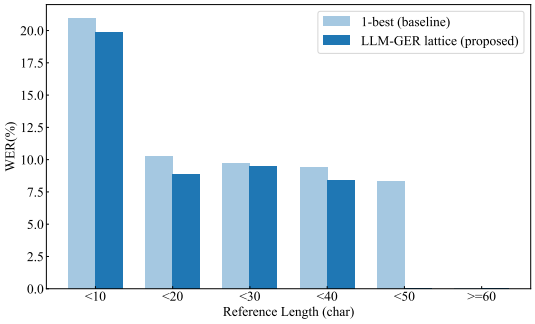


Fig. 6: WERs in $Eval_3$

Fig. 7: Averaged WERs in each reference length regimes in $Eval_1$, $Eval_2$ and $Eval_3$.

2) Output examples: To see the trends by LLM GER correction, Table II shows some output examples both in the proposed LLM GER lattice and baseline 1-best. In Example 1, 2, and 3, we see all of the 1-best outputs include error, which is caused by phonetic similarity. These 1-best outputs are not as natural as Japanese and have major problems, such as the lack of meaning in words. On the other hand, in LLM GER Lattice, we see more recovery with correction, and the corrected outputs became more similar to references. In Example 4, we see some unknown tokens were included in

TABLE II: Example Japanese sentences in baseline 1-best and proposed LLM GER lattice.

Example 1: Phonetic similar errors ($Eval_3$)	
1-best (baseline)	えっとニセコの あのっぷり スキー場というところに行ってきました
LLM GER Lattice (proposed)	えとニセコの **アンヌ** スキー場というところに行ってきました
Reference	えとニセコの **アンヌ**プリスキー場というところに行ってきました
Example 2: Phonetic similar errors ($Eval_3$)	
1-best (baseline)	当然全域 に慣れてない
LLM GER Lattice (proposed)	**と全然雪** に慣れてない
Reference	**と全然雪** に慣れてない
Example 3: Phonetic similar errors ($Eval_2$)	
1-best (baseline)	今回はディバイドと呼ばれる周波数 へへん環境 を用いた為に周波数および
LLM GER Lattice (proposed)	今回はディバイドと呼ばれる周波数 **変換** を用いた為に周波数および
Reference	今回はディバイダーと呼ばれる周波数 **変換** 器を用いた為に周波数および
Example 4: Correcting unknown tokens ($Eval_2$)	
1-best (baseline)	関係してきてるのではないでしょ <UNK> <UNK> <UNK> <UNK> か
LLM GER Lattice (proposed)	関係してきてるのではないでしょ う か
Reference	関係してきてるのではないでしょ う か

the 1-best output, but they were converted to the correct word. In $Eval_2$, we observed some of the 1-best outputs included unknown tokens. After the correction by the LLM GER lattice, we observed all of the unknown tokens were removed or substituted with corrected words.

3) The advantage for using lattice: We find that the process of lattice-based data using LLM resembles that of the traditional ASR decoder. Here's the process within a traditional decoder for language model rescoring, which involves re-evaluating the paths in a lattice using more sophisticated models. The initial decoding uses a simpler language model to assign scores to different paths in the lattice. In rescoring, these initial scores are replaced with scores from a more complex, typically larger, language model (i.e., LLM in this paper). The words with new scores reflect a more accurate evaluation of word sequences, improving overall transcription accuracy.

To sum up, the method we proposed in this paper leverages the strengths of advanced LLMs to refine the transcription output by reassessing the probabilities of more diverse hypotheses in the lattice.

V. Conclusion

This paper proves that using LLM-based GER can directly work on lattice-structured output from the speech recognition system. Experiments on CSJ corpus show that compared with using n-best hypotheses, using lattice can improve the performance of Japanese speech recognition. In the future, we will conduct more experiments to demonstrate the method's effectiveness in a broader range of settings.

Acknowledgment

We want to thank the Tohoku-NICT matching funding's support.

References

[1] S. Zhang, M. Lei, and Z. Yan, "Investigation of transformer based spelling correction model for ctc-based end-to-end mandarin speech recognition." in *Proc. Interspeech*, 2019, pp. 2180–2184.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *CoRR abs/1706.03762*, 2017.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[4] S. Zhang, H. Huang, J. Liu, and H. Li, "Spelling error correction with soft-masked bert," *arXiv preprint arXiv:2005.07421*, 2020.

[5] H. Futami, H. Inaguma, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Distilling the knowledge of BERT for sequence-to-sequence ASR," *CoRR*, vol. abs/2008.03822, 2020.

[6] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," *arXiv:1910.14659*, 2019.

[7] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using BERT for speech recognition," in *Proc. ACML*, 2019, pp. 1081–1093.

[8] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *Proc. IEEE-ICASSP*, 2020, pp. 7694–7698.

[9] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?" in *Proc. IEEE-ICASSP*, 2021, pp. 6533–6537.

[10] C. Wang, Y. Wu, S. Liu, M. Zhou, and Z. Yang, "Curriculum pre-training for end-to-end speech translation," *arXiv preprint arXiv:2004.10093*, 2020.

[11] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *arXiv preprint arXiv:1908.02265*, 2019.

[12] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[13] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *Proc. AAAI*, 2020, pp. 13 041–13 049.

[14] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proc. ECCV*, 2020, pp. 121–137.

[15] C. Chen, Y. Hu, C.-H. H. Yang, S. M. Siniscalchi, P.-Y. Chen, and E. S. Chng, "Hyporadise: An open baseline for generative speech recognition with large language models," *ArXiv*, vol. abs/2309.15701, 2023.

[16] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *ArXiv*, vol. abs/2106.09685, 2021.

[17] S. Li, C. Chen, C. Kwok, C. Chu, E. Chng, and H. Kawai, "Investigating asr error correction with large language model and multilingual 1-best hypotheses," in *Proc. Interspeech*, 2024, pp. 1315–1319.

[18] Y. Ko, S. Li, C.-H. H. Yang, and T. Kawahara, "Benchmarking japanese speech recognition on asr-llm setups with multi-pass augmented generative error correction," 2024.

[19] P. Koehn, "MOSES Statistical Machine Translation System User Manual and Code Guide," 2014. [Online]. Available: http://www.statmt.org/moses/manual/manual.pdf

[20] K. Maekawa, "Corpus of spontaneous japanese: Its design and evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[21] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the corpus of spontaneous japanese," in

Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.

[22] S. Li, X. Lu, R. Takashima, P. Shen, T. Kawahara, and H. Kawai, "Improving CTC-based Acoustic Model with Very Deep Residual Time-delay Neural Networks," in *Proc. Interspeech 2018*, 2018, pp. 3708–3712.

[23] ——, "Improving very deep time-delay neural network with vertical-attention for effectively training ctc-based asr systems," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 77–83.

[24] T. Moriya, T. Shinozaki, and S. Watanabe, "Kaldi recipe for Japanese spontaneous speech recognition and its evaluation," in *Autumn Meeting of ASJ*, no. 3-Q-7, 2015.

[25] N. Kanda, X. Lu, and H. Kawai, "Maximum a posteriori based decoding for CTC acoustic models," in *Proc. INTERSPEECH*, 2016, pp. 1868–1872.

[26] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint CTC-Attention based End-to-End speech recognition with a deep CNN Encoder and RNN-LM," in *Proc. INTERSPEECH*, 2017.

[27] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "Generative speech recognition error correction with large language models and task-activating prompting," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

EavaNet: Enhancing Emotional Facial Expressions in 3D Avatars through Speech-Driven Animation

Seyun Um*, Yongju Lee*, WooSeok Ko*, Yuan Zhou†, Sangyoun Lee*, Hong-Goo Kang*
* Department of Electrical and Electronic Engineering , Yonsei University, Seoul, Korea
E-mail: [syum,wsko]@dsp.yonsei.ac.kr, [paulyongju,syleee,hgkang]@yonsei.ac.kr
† 01.AI, Beijing, China
E-mail: zhouyuan@01.ai

***Abstract*—Speech-driven 3D facial animation models are essential for creating human-like avatars that synchronize lip movements realistically with speech. Despite these advancements, it is still difficult to effectively convey a wide range of emotional facial expressions that align with the voice. This issue arises due to the lack of clearly labeled datasets for individual emotions as well as insufficient inputs to adequately describe these emotions. To overcome this challenge, we propose a re-categorization process that reduces the data into four emotional groups: angry, sadness, happy, and neutral. We use the re-categorized datasets to estimate style embeddings, which serve to distinctly express emotions and control their intensity. Additionally, we tackle the challenge of slow inference speed in autoregressive models by introducing EavaNet, a non-autoregressive model utilizing gated activation units (GAUs) and bidirectional long short-term memory (BLSTM) modules for efficient prediction of 3D face mesh vertices. Our proposed model outperforms previous state-of-the-art models in terms of emotional expressiveness and lip synchronization accuracy in both subjective and objective evaluations.** [1]

I. Introduction

Speech-driven 3D facial animation systems are now widely used in personal assistants, computer games, film production, and virtual communication systems. Recent developments in artificial intelligence and 3D face modeling technologies have made it easier to create lifelike avatars capable of accurately synchronizing voice and lip movements[1]. However, most 3D facial animation models have primarily focused on improving lip synchronization rather than the full range of facial movements associated with emotions conveyed by the voice. As a result, these systems can generate unnatural facial expressions that do not match the emotions expressed in the voice, which limits their usefulness.

In this work, we identify two primary reasons for this issue: low-quality training data labels and the lack of style embeddings with facial expression information. Previous state-of-the-art models [2], [3] have used 3D subject IDs as supplementary input to produce speaker-dependent facial animations. Although these models can create 3D avatars with high lip synchronization accuracy, they often generate monotonous facial expressions and fail to properly convey the emotional nuances present in the input audio.

In our experiments, we utilize the BIWI dataset [4], which includes emotional speech paired with corresponding expressive facial movements, aligning with our research objectives. However, this dataset has limitations; the emotion labels lack consistency across different subjects, and its size is insufficient for effectively training. To address these challenges, we reclassify the BIWI dataset into four emotions (angry, sadness, happy, and neutral) to improve the model's emotional expressiveness within the constraints of the dataset. Additionally, we carefully select appropriate 3D subjects for each emotion for clear emotional expression.

Our model extracts a style embedding, which encapsulates the emotional expressions conveyed in the input speech utterances. This style embedding is subsequently utilized to control the expressions of the avatar, making its facial movements more expressive even for subjects not previously seen during training. Furthermore, we are able to adjust the intensity of emotions by interpolating between style embeddings obtained from the target emotion and the neutral one. This capability enables us to generate avatars capable of displaying a broader spectrum of emotional expressions than those present solely in the existing BIWI dataset.

Our main contributions are as follows:

- To enhance facial expressions, we reclassify emotion labels into four distinct categories and employ style embeddings trained through intercross training[5].
- Our framework allows us to control the intensity of an emotion by interpolating between the target emotional style embedding and the neutral one.
- We use efficient short-term and long-term modeling approaches, such as GAU and BLSTM, to address the issue of slow inference speed.
- We show that our proposed model outperforms previous state-of-the-art models in terms of emotional expression and lip synchronization accuracy in both objective and subjective evaluation experiments.

II. Related Work

The goal of a 3D face animation model is to create a 3D avatar face from a given speech input, which can include text, speaker ID, language ID, and emotion [2], [6], [7], [11]. [6] introduce VOCA, along with the VOCASET dataset, which incorporates FLAME [11]. MeshTalk [7] is trained on an in-house dataset of 250 subjects, featuring more active facial expressions than VOCASET, including blinking and eyebrow

[1]We provide a demo page with sample videos and reclassified emotion labels at https://sam-0927.github.io/eavenet-demo/

TABLE I: Dataset of 3D facial animation.

Name	Speakers	Emotion	Duration	Release
VOCASET[6]	12	1	32m	o
Multiface[7] (subset)	13	1	21m	o
BIWI[4]	14	15	43m 35s	o
[8]	2	N/A	8m 46s	x
CREMA-D[9]	91	6	7442 clips	x
AdaIN[10]	1	8	38m 57s	x

movements. FaceFormer [2], based on the transformer [12], employs two biased attention masks and a periodic position encoding strategy. By combining FaceFormer with a pre-trained wav2vec 2.0 [13] to obtain speech representations, the model auto-regressively predicts vertices. CodeTalker [3] estimates facial movements as a discrete codebook and regressively predicts motion codes to enhance vividness and reduce the over-smoothed problem of facial movements.

[8] analyze formants and map emotional speech to 3D vertices by adding emotional embeddings to the articulation embeddings. They adopt a data-driven approach where the model automatically learns the representation of emotional states. [9] propose a model that predicts FLAME parameters to control the type and intensity of the emotion while generating expressive facial animations. During training, the model's emotion control module recognizes emotions from images, and it recognizes emotions from audio during inference. EmoTalk [14] generates emotional 3D facial animations by disentangling content and emotion in speech using cross-reconstruction loss with different emotion labels. [15] apply adaptive instance normalization (AdaIN) to separate the two aforementioned features, removing emotional information from speech and converting it to content features, including specific emotions using labels [10].

Previous studies on expressing emotions in 3D avatars have used abundant datasets containing with rich emotional information [8]–[10]. However, these datasets are often unavailable to the public, which hinders further research as described in Table I. To address this limitation, this study uses existing publicly accessible datasets.

III. Proposed Model

A. Emotion classification

The BIWI dataset [4] contains 15 emotion types and 5 scales of intensity (1:not at all $\sim$ 5:very) that indicate the level of emotional expression. Due to the small amount of training data and the uncertainty between 15 emotion types, we have decided to reclassify them into four distinct and broad categories (angry, happy, sadness, and neutral) that are still applicable to a wide range of service scenarios.

Five people were asked to re-label the emotion types using Algorithm 1 by observing facial expressions in video samples and listening to speech signals. Similar to the previous dataset, each sample was evaluated by an average of three individuals. As with the original dataset, people were asked to rate the degree of emotional expression on a scale of 1 to 5. To ensure the reliability of the labeling process and the consistency of the training data, we selected 3D subjects when the average emotional expression scores are greater than 3. For each sample of the selected subject mesh, emotion types were labeled based on the majority vote from the evaluators, which further enhances the quality of the dataset.

Algorithm 1 Emotion labeling

Input: number of samples for each 3D subject n, evaluators $m = 3$, Evaluated emotion label set $V_m^{id,index}$=$\{v_1^{1,1}, v_2^{1,1}, v_3^{1,1}, ..., v_3^{14,n}\}$, subject set S=$\{s_1, ..., s_{14}\}$, emotion set E=$\{h:0, a:0, s:0, ne:0\}$ of each sample, average expression score Scr=$\{scr_1, ..., scr_{14}\}$
Output: training subject set S_{tr}, emotion label L=$\{l^{1,1}, ..., l^{14,n}\}$

```
for i = 1 to 14 do
    if scr_i > 3 then
        S_tr ← s_i
    end if
end for
for i = 1 to 14 do
    for j = 1 to n do
        initialize values of E to 0
        for k = 1 to 3 do
            E[v_k^{i,j}] ← E[v_k^{i,j}] + 1
        end for
        l^{i,j} ← E's index with maximum value
    end for
end for
```

B. EavaNet: Emotional avatar generator

Overview. Our proposed model EavaNet, shown in Fig. 1, takes a speech signal $S_{1:N} = s_1, ..., s_N$ and a speaker ID as input, and predicts the sequences of 3D face vertices $V_{1:T} = v_1, ..., v_T$ using an encoder-decoder architecture. The encoder is composed of an audio encoder and a style encoder, which respectively predicts the contextual and style embeddings. To address the issue of insufficient training data, we use a pre-trained speech representation model, wav2vec2.0, for the audio encoder. We freeze the weights of feature extractor in this module during training. The extracted contextual embeddings are first passed through a 1D convolution layer (Conv1D). Then, they are combined with a style embedding and a speaker ID before being fed to the decoder. The Conv1D operation with a kernel size of 4, padding size of 2, and stride size of 1, performs a downsampling operation, which helps align the timing information of the acoustic features with that of the target visual vertices.

Style encoder. The style encoder computes a style embedding $e \in \mathbb{R}^{1\times d}$ that represents the facial expressions linked with each emotion from style vertices $V_{spk^*,utt^*}^{emo^{target}}$, where spk and utt denote speaker and utterance, respectively. To ensure that the style embedding exclusively encapsulates emotional information rather than contextual cues, we select style vertices solely from data within the same emotion category, excluding those that contain the target vertices ($spk^* \in S_{tr}$, $spk^* \neq spk^{target}$ and $utt^* \neq utt^{target}$, $*$ means "any"). The style encoder comprises a style extractor and a style token layer (STL), resembling the architecture of GST-Tacotron [16]. The style extractor extracts style information from the style vertices, while the STL encodes this style information into a token that is then used by the decoder to generate the output. The style extractor is made up of a projection layer,

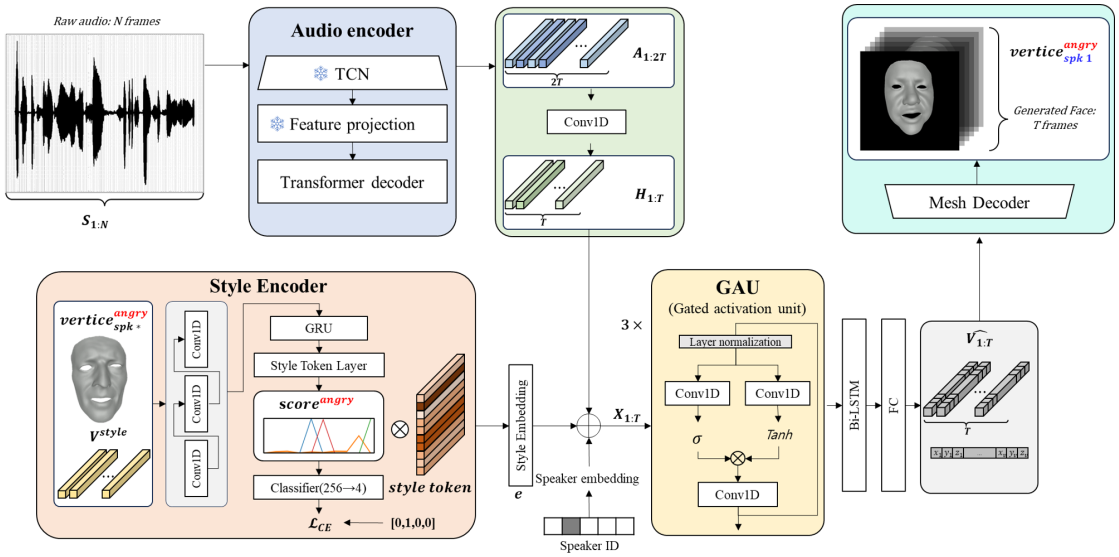


Fig. 1: Architecture of EavaNet. The snow symbols in the audio encoder denote freezing the weights.

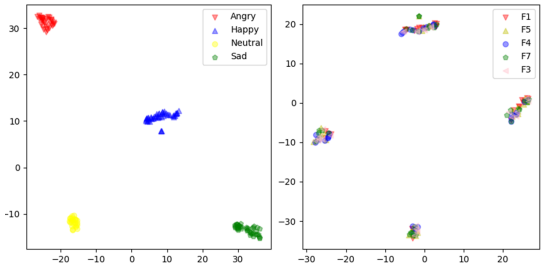


Fig. 2: t-SNE [18] plots of style embeddings from five subjects in training data (left: emotion, right: speaker).

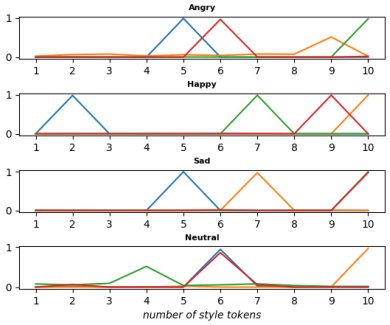


Fig. 3: The score of 4-head attention in the style token layer. Each color represents the index of each head, and each emotion has a different distribution of style tokens.

three Conv1D layers with layer normalization, LeakyReLU activation functions, and a bidirectional gated recurrent unit (GRU) [17]. The projection layer transforms the input data into a feature vector. The Conv1D layers, with a channel size of 256 and kernel size of 4, along with a stride of 2 and padding size of 1, extract features from the feature vector. Subsequently, layer normalization is applied to normalize these features. The GRU then processes the features to extract a 256-dimensional style embedding.

$$\mathbf{e} = \text{STL}(\text{style}_{\text{extractor}}(V^{emo^{target}}_{spk^*,utt^*})). \quad (1)$$

In the 4-head attention layer of STL, the output of the style extractor is multiplied by the style tokens $V_{style\,token}$ and each token's score. The attention score is calculated by multiplying the query Q from the output of the style extractor with the transposed trainable parameter key K^T, then dividing it by the square root of the key's dimension d_k. Finally, the weighted sum of style tokens for each head is concatenated to form the style embedding.

$$\text{Att}(Q, K, V_{style\,token}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V_{style\,token}. \quad (2)$$

To separate the style embeddings depending on the type of emotion, we compute the cross-entropy $\mathcal{L}_{CE}$ between the style embedding and the emotion labels rather than using contrastive learning technique in embedding space [19]. This allows us to identify the different styles that are associated with each emotion, as shown in Fig. 2. In the figure, style embeddings that belong to the same emotion category become closer to each other, forming independent clusters regardless of the speaker variations. This means that the style of speech can be used to identify the emotion expressed, even if the speaker's voice is different. The style embedding characteristics can be determined by the score since the style tokens are shared for all training data.

As shown in Fig. 3, the score of each emotion highlights different style tokens. This means that the style embedding can be represented by the combination of style tokens, as they are tailored to their specific characteristics. In other words, the style embedding can be created by combining different style tokens, each of which represents a specific emotion. To use the style embeddings during inference, we first compute the centroid of the generated style embeddings in each emotion cluster, then store them in a lookup table during training. Our model directly retrieves the style embeddings from the pre-trained lookup table during inference.

Decoder. The decoder is composed of GAU, BLSTM and

projection layers motivated by [20]. The style embedding and speaker embedding $e_{spk} \in \mathbb{R}^{1\times d}$ are added to the hidden representations of the Conv1D layers and they are passed to the GAU.

$$\begin{aligned} \mathbf{X_{1:T}} = \mathbf{e} \oplus \mathbf{H_{1:T}} \oplus \mathbf{e_{spk}}, \qquad \mathbf{e_{spk}} = F_\theta(i), \\ \mathbf{H_{1:T}} = \text{Conv1D}(\mathbf{A_{1:2T}}), \\ \mathbf{A_{1:2T}} = \text{Enc}_{\text{audio}}(\mathbf{S_{1:N}}), \end{aligned} \tag{3}$$

where $X \in \mathbb{R}^{T\times d}$ denotes the input of GAU and $\oplus$ means adding. $H \in \mathbb{R}^{T\times d}$ and $A \in \mathbb{R}^{2T\times d_A}$ are hidden representations and audio encoder outputs, respectively. F_θ and i depict fully connected layers and a speaker ID, respectively. The GAU used in the generative model [21] has three residual blocks, each consisting of two paths: a filter and a gate, with residual connections. Each path has a single 1D convolution layer and two activation functions (hyperbolic tangent and sigmoid). The filter path estimates local features in adjacent frames based on the receptive field size, and the gate path determines how much to pass through. After the outputs of the two gates are multiplied, they are passed through an output 1D convolution layer and are added to the residual connections, which have 256 channel size with 1 kernel size. The convolution layers have channel sizes of (256, 512) and kernel sizes of (5, 3, 5) with padding sizes of (2, 1, 2).

$$\begin{aligned} z_k = \tanh(W_{f,k} * x_k + b_{f,k}) \otimes \sigma(W_{g,k} * x_k + b_{g,k}), \\ x_{k+1} = (W_{o,k} * z_k + b_{o,k}) \oplus x_k, \end{aligned} \tag{4}$$

where tanh, σ, $*$ and $\otimes$ denote hyperbolic tangent, sigmoid, convolution operator, and an element-wise multiplication operator, respectively. The f, g and, o denote filter, gate, and output layer, respectively. $W_{*,k}$ and $b_{*,k}$ are trainable weights and the bias of the kth layer.

All of these processes are conducted concurrently. Since the GAU uses only CNNs to capture local features, we also incorporate two BLSTM layers with a channel size of 256 to capture temporal characteristics. Finally, the output of the BLSTM layers is projected to the vertices:

$$\hat{V}_{1:T} = P_\theta(\text{BLSTM}(\mathbf{Y}_{1:T})), \tag{5}$$

where $Y_{1:T}$ and $\hat{V}_{1:T}$ are the outputs of GAU and predicted vertices from fully connected layers P_θ, respectively.

Training criterion and details. We compute mean square errors (MSEs) between the predicted vertices $\hat{V}_{1:T}$ and the target ones $V_{1:T}$.

$$\mathcal{L}_{\text{vert}} = \frac{1}{T}\sum_{i=1}^{T} \|V_i - \hat{V}_i\|_2, \quad V_i^j = V_i^j - m^j, \tag{6}$$

where j represents subject id. The total training loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{vert}} + \lambda_{CE}\mathcal{L}_{\text{CE}}, \tag{7}$$

where the scale factor λ_{CE} is set to 0.001.

We trained EavaNet for 200 epochs on an NVIDIA GeForce RTX 3090 GPU using the Adam optimizer. The initial learning rate is set to 1e-4 and is reduced by half every 40 epochs. The mini-batch size is set to 1. For the reference models for comparison, we adhere to their respective official codes for training.

TABLE II: Quantitative evaluation on BIWI-A.

Model	Vertex error ↓ ($\times 10^{-5}$mm)	Lip vertex error ↓ ($\times 10^{-4}$mm)	FDD ↓ ($\times 10^{-5}$mm)
FaceFormer [2]	3.9876	7.5916	5.0914
CodeTalker [3]	4.7354	8.2971	**3.6702**
EavaNet	**3.4603**	**6.3628**	3.9958

TABLE III: A/B results of EavaNet and previous models.

Model		Lip sync↑	Expression↑
EavaNet	vs GT	40.86	45.21
	vs FaceFormer	51.73	66.23
	vs CodeTalker	53.52	60.41

IV. Experiments

A. Experiment setting

We use the BIWI dataset [4], which contains emotional speech accompanied by dense dynamic 3D facial geometries. The dataset consists of 14 subjects, each with 40 sentences, for an average sequence length of 4.67 seconds. The 3D facial meshes are captured at a rate of 25 frames per second, with each mesh consisting of 23,370 vertices. We eliminate outlier samples by following the BIWI guidelines; instances with zero scores across all 15 emotion categories are considered outliers that should be removed. The dataset is divided into three parts: a training set with 146 sentences spoken by five speakers (F1, F3, F4, F5, F7); a validation set with 18 sentences spoken by the same five speakers; and two separate test sets, BIWI-A and BIWI-B. BIWI-A has 19 sentences spoken by five subjects who were previously seen, while BIWI-B has 30 sentences spoken by nine subjects who were not seen before. We use BIWI-A for objective evaluation, and BIWI-B for user studies and ablation studies.

To stabilize the training, we removed the silence sections before and after the main audio segments using the *librosa* library, aligned the video accordingly, and normalized the audio scale to -3dB. The audio samples were captured at a rate of 16,000 samples per second.

B. Quantitative Evaluation

To measure the performance of our model, we use two metrics: lip vertex error and the *upper-face dynamics deviation* (FDD), following the methodologies of FaceFormer and CodeTalker. The lip vertex error measures the accuracy of lip-syncing, while the FDD assesses the fidelity of facial expressions. Both of these metrics measure the discrepancy between the predicted and actual vertices. The lip vertex error is a measure of the maximum distance between the actual and predicted lip positions for each frame, averaged over all frames. The FDD measures the difference in standard deviation between the vertices of ground-truth and predicted facial expressions for each frame, which are associated with upper facial expressions. Furthermore, we compute the L2 error for all vertices in each frame and then average these

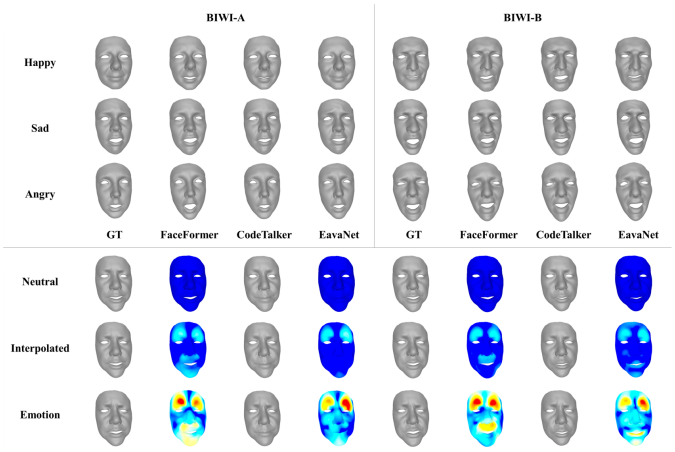


Fig. 4: Visual comparison of sampled 3D faces generated by different models (top). 3D faces and their heatmaps animated by EavaNet with neutral, interpolated and, emotional style embedding (bottom). We use the *e38* sentence in BIWI-B (*angry* emotion) in this sample.

values across all frames. These metrics are used to fully assess the overall quality of the generated 3D avatars.

Table II presents the quantitative findings for BIWI-A. Our model outperforms previous state-of-the-art models in both lip accuracy and overall quality. This result highlights the fact that our model can accurately generate 3D facial animations while conveying distinct pronunciations. Our model also has a better FDD score than FaceFormer, which indicates its superiority in facial expression representations. While CodeTalker outperforms EavaNet in terms of FDD score, the overall quality of its generated avatars is inferior to that of other models.

C. Qualitative Evaluation

Our model's ability to create expressive 3D facial animation with precise lip synchronization has been confirmed through qualitative evaluation. A visual comparison of EavaNet and competing models is presented in Fig. 4. To ensure a fair comparison, we use the same randomly selected subject ID as the conditional input for FaceFormer, CodeTalker, and our proposed model. In both cases, the samples produced by EavaNet express the emotion more clearly through dynamic facial movements, while other samples do not, especially with drooping eye tails and wrinkles between the brows when expressing *sadness*. One interesting finding is that our model expresses *sadness* and *angry* more strongly than the ground truth (GT) in both BIWI-A and BIWI-B (the 2nd and 3rd rows of Fig. 4). Unlike other models, EavaNet can successfully convey emotions even when they cannot be discerned solely from the prosody of the speech input because it uses style embeddings effectively. We suggest that you zoom in and examine the figure more closely.

D. User study

We conducted preference A/B tests with GT, FaceFormer, and CodeTalker to evaluate the quality of the generated avatars from a human perspective. Twenty-three people were asked to choose samples with high lip-sync accuracy from the outputs of two different models. In addition, participants were asked to choose samples that effectively conveyed the emotions expressed in the audio input. We created avatars for unseen subjects in BIWI-B by using five seen subjects' IDs as conditions, and then picked 30 samples from all the conditioned IDs. Using this setup, we created 90 A/B pairs (30 samples $\times$ 3 comparison models).

TABLE IV: Results in terms of inference speed.

State	FaceFormer	CodeTalker	EavaNet
RTF	0.1356	1.5684	**0.0097**
Ratio	$\times$ 11.56	$\times$ 1	$\times$ **161.55**

TABLE V: MOS (mean opinion score) results with 95% confidence intervals in terms of emotion manipulation.

State	Neutral	Interpolated	Emotion
MOS	2.44$\pm$0.29	3.13$\pm$0.21	4.11$\pm$0.19

Table III demonstrates a considerable improvement in our model's emotional expression and lip synchronization accuracy with audio. Our approach has been shown to be more accurate in terms of lip synchronization and emotional expression than state-of-the-art models, with a preference of over 50% and 60%, respectively. In particular, our model demonstrates a 66% preference in expressing emotions compared to avatars generated by FaceFormer, and it outperforms CodeTalker, which uses vector quantization (VQ) to address smoothing issues, by 60%. Additionally, EavaNet performs similarly to the ground truth (GT), with a preference of 45%. This implies that by training the model on the five subjects with the richest emotional expression from the 14 subjects in the BIWI dataset (as mentioned in Section III-A), we can effectively overcome limitations and create authentic facial expressions, even for subjects with lower emotional expression.

E. Inference speed

When deploying the model in real-world applications, it is essential to generate high-quality 3D facial animations with human-like emotional expressions in a time-efficient manner. To assess the generation speed of each model, we randomly selected a sample from the test set and measured the generation time using an NVIDIA GeForce RTX 3090 GPU. We calculated the average time required after conducting 100 synthesis operations. The RTF (Real Time Factor) denotes the ratio of the time taken to predict the positions of 3D vertices for a 4-seconds length sequence.

As detailed in Table IV, our proposed model demonstrates the capability to generate 3D avatars at speeds 14 and 161 times faster than FaceFormer and CodeTalker, respectively. FaceFormer and CodeTalker rely on autoregressive (AR) modeling, wherein predictions of vertices are conditioned on previous predicted frames. Consequently, these models demand a significant amount of time during inference due to the sequential nature of their predictions. In contrast, EavaNet adopts a non-AR structure that enhances inference speed by enabling parallel generation of vertices. Although the model incorporates BLSTM layers after GAUs, it processes data in latent space sequentially, mitigating computational overhead.

F. Ablation study

Emotion intensity. Style embeddings are beneficial for producing high-quality 3D avatars and controlling the intensity of emotional expression. We use linear interpolation to blend the style embedding between neutral and (happy, sadness, anger) to control the intensity of the expression.

$$e_{\alpha}^{emotion} = \alpha e^{emotion} + (1-\alpha)e^{neutral}, \tag{8}$$

where $e^{emotion}$ and $e^{neutral}$ are style embeddings extracted from the trained lookup table (see Section III-B). We set the ratio α to 0.3 to perform exemplary experiments. We randomly selected 45 samples (15 samples $\times$ 3 expression strength; neutral, weak emotion, and emotion) from BIWI-B by conditioning the subject *F4*, and then asked 25 people to rate the strength of emotional expression on a scale of 1 to 5. The higher the score, the more intense the emotional expression. As shown in Table V, the avatar generated using the interpolated style embedding $e_{\alpha}^{emotion}$ has a weaker emotional representation than those generated using the emotion style embedding $e^{emotion}$, but a stronger emotional representation than those generated using the neutral style embedding $e^{neutral}$. Fig. 4 shows an example. The wrinkles between the brows become more pronounced and the facial expression changes as the emotion of anger intensifies, gradually strengthening the emotional expression.

V. Conclusion

In this work, we presented EavaNet, a generative model that creates 3D facial animations with emotional expression given speech signals. EavaNet is non-autoregressive, and is able to quickly predict 3D face vertices showing accurate emotional facial expressions. To address limited training data, we re-categorized BIWI dataset labels into four emotions. Additionally, EavaNet estimates style embeddings and interpolates between them enables fine-tuned control over emotional expression strength. Our model outperforms previous state-of-the-art models in terms of emotion expression and lip sync accuracy in both subjective and objective evaluations.

References

[1] H. X. Pham, S. Cheung, and V. Pavlovic, "Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach," in *Proc. CVPR*, 2017.

[2] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *Proc. CVPR*, 2022.

[3] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong, "Codetalker: Speech-driven 3d facial animation with discrete motion prior," in *Proc. CVPR*, 2023.

[4] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-d audio-visual corpus of affective communication," *Transactions on Multimedia*, vol. 12, 2010.

[5] Y. Bian, C. Chen, Y. Kang, and Z. Pan, "Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis," in *interspeech*, ISCA, 2019.

[6] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *Proc. CVPR*, 2019.

[7] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," in *Proc. CVPR*, 2021.

[8] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics*, vol. 36, 2017.

[9] Y. Chen, J. Zhao, and W.-Q. Zhang, "Expressive speech-driven facial animation with controllable emotions," *Proc. ICMEW*, 2023.

[10] C.-J. Chang, L. Zhao, S. Zhang, and M. Kapadia, "Disentangling audio content and emotion with adaptive instance normalization for expressive facial animation synthesis," *Computer Animation and Virtual Worlds*, vol. 33, 2022.

[11] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans.," *ACM Transactions on Graphics*, vol. 36, 2017.

[12] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Proc. NIPS*, vol. 30, 2017.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NIPS*, vol. 33, 2020.

[14] Z. Peng, H. Wu, Z. Song, *et al.*, "Emotalk: Speech-driven emotional disentanglement for 3d face animation," *Proc. ICCV*, 2023.

[15] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. ICCV*, 2017.

[16] Y. Wang, D. Stanton, Y. Zhang, *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018.

[17] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS*, 2014.

[18] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, 2008.

[19] J. Kim and H.-G. Kang, "Contrastive learning based deep latent masking for music source separation," in *Proc. ISCA*, 2023.

[20] D. Kim, S.-W. Chung, H. Han, Y. Ji, and H.-G. Kang, "Hd-demucs: General speech restoration with heterogeneous decoders," in *Proc. ISCA*, 2023.

[21] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," in *Proc. ISCA*, 2016.

Optimal Investment with Incomplete Information and Herd Effect

Huisheng Wang, Mingxiao Liu, Ji Qi, and H. Vicky Zhao
Department of Automation, Tsinghua University, Beijing, China
E-mail: {whs22,mx-liu21,qij21}@mails.tsinghua.edu.cn, vzhao@tsinghua.edu.cn

***Abstract*—Due to incomplete information in financial markets, asset returns and volatility are generally not observable to normal investors. To address this issue, normal investors can use historical price data to estimate these parameters. In this process, they often exhibit bounded rationality, demonstrating their limited cognitive and information-processing capabilities. Also, given the investment expertise of experts, normal investors can refer to and align with the experts' decisions, which is called the herd effect. To better understand normal investors' behaviors, in this work, we modify the general market parameter estimation methods and propose sliding window-based estimation methods to model their bounded rationality. We then formulate the optimal investment problem, which involves one expert and one normal investor whose decisions are influenced by the expert, and jointly consider incomplete information and the herd effect. We derive the analytical solution for the normal investor's optimal decision and theoretically analyze the influence of bounded rationality, incomplete information, and herd effect on his/her decisions. We validate our analysis using numerical experiments on real data.**

I. Introduction

In financial markets, investors allocate assets offering high returns and low volatility to maximize their expected utility of terminal wealth over the investment period [1]. However, the asset returns and volatility are generally not observable for normal investors due to incomplete information [2]. Normal investors can use price data to derive optimal estimates of the returns and volatility [3]. Additionally, with their expertise in accessing and interpreting market information, experts' decisions offer valuable insights into the returns and volatility [4]. When experts anticipate high returns or low volatility, they are more likely to buy the asset, and vice versa. Therefore, normal investors can also refer to and align with the experts' decisions, known as the herd effect [5]. To better understand normal investors' behaviors and devise effective mechanisms to guide their decisions, we first need to study how incomplete information and herd effect influence their decisions.

There exists extensive research on estimating asset returns and volatility with incomplete information. The CAPM [6] and the GARCH model [7] are two classic models for estimating the return and volatility, respectively. The Black-Scholes model simultaneously considers the returns and volatility [8], based on which, the work in [9] provides general methods to jointly estimate the returns and volatility. However, these methods assume that normal investors are fully rational and use all available data for estimation. In reality, normal investors often exhibit bounded rationality with limited cognitive and information-processing capabilities. Due to bounded rationality, they tend to focus on price data within a specific timeframe [10]. To better model their behavior, it is crucial to incorporate bounded rationality into the estimation methods.

Additionally, there have been many qualitative analyses confirming the significant influence of herd effect on investment decisions [11], and our prior work in [12] quantitatively studied the influence of herd effect on the normal investors' decision. However, these studies assume complete information in financial markets and ignore parameter estimation.

In the literature, there are limited works that jointly consider the influence of incomplete information and herd effect on normal investors' decisions. In this work, we address this problem by formulating an optimal investment problem involving one expert and one normal investor. Following the work in [1], we assume that the expert makes rational decisions. For the normal investor, at each time step, we assume that he/she first estimates the asset return and volatility with bounded rationality. Then, he/she makes the optimal decision to maximize the expected utility of the estimated terminal wealth while minimizing the distance between the two investors' decisions due to the herd effect. Finally, the normal investor updates his/her wealth and repeats the above estimation and decision-making process for the next time step.

The rest of this paper is organized as follows. In Section II, we propose sliding window-based market parameter estimation methods considering bounded rationality, and formulate the optimal investment problem for the normal investor, which jointly considers incomplete information and herd effect. In Section III, we derive the analytical solution of the normal investor's optimal decision and quantitatively analyze the influence of bounded rationality, incomplete information, and herd effect on normal investor's decisions. In Section IV, we conduct numerical experiments on a real asset dataset to validate our analysis. Section V is the conclusion.

II. Problem Formulation

In this section, we first define the problem, introduce the expert's rational decision, and extend the general estimation methods for the asset returns and volatility considering the normal investor's bounded rationality. Then, we formulate the optimal investment problem for the normal investor.

The important notations used in this work are in Table I.

A. Problem Definition

Following the work in [1], we consider the financial market with one risk-free asset and one risky asset. Extend-

TABLE I
NOTATIONS

Notation	Meaning	Notation	Meaning
r	Interest rate	μ	Return
v	Excess return	σ	Volatility
α_1, α_2	Risk aversion coefficients	$\mathcal{T}, T$	Investment period
N	Sample size	$\hat{\mu}$	Estimated return
$\hat{v}$	Estimated excess return	$\hat{\sigma}$	Estimated volatility
w	Sliding window width	θ	Herd coefficient
X_1	Wealth	ϕ	Utility function
ρ	Decay rate	D	Average deviation
η	Integral constant	Z_1	Opinion

Notation	Meaning
$\hat{\bar{P}}_1$	Rational decision with incomplete information (*IRD*)
$\bar{P}_1, \bar{P}_2$	Rational decision with complete information (*CRD*)
P_1^*	Optimal decision (*OPD*)

ing this analysis to encompass multiple assets is straightforward. We denote the investment period as $\mathcal{T} := [0, T]$. Let $(\Omega, \mathscr{F}, \{\mathscr{F}(t)\}_{t\in\mathcal{T}}, \mathbb{P})$ be a complete probability space on which a standard Brownian motion $\{B(t)\}_{t\in\mathcal{T}}$ with $B(0) = 0$ is defined. We utilize $\{B(t)\}_{t\in\mathcal{T}}$ to model the randomness of the risky asset price. The market parameters include the interest rate r of the risk-free asset, the return μ, and the volatility σ of the risky asset. Furthermore, we define the excess return of the risky asset over the risk-free asset as

$$v := \mu - r > 0. \tag{1}$$

Typically, investors can easily access the current interest rates from bank counters or websites. Following the work in [1], we assume that both the normal investor and the expert can obtain the true value of r. Compared to the risk-free asset, the risky asset exhibits a higher return and comes with volatility. Following the work in [1], we assume the price process of the risk-free asset $\{S(t)\}_{t\in\mathcal{T}}$ satisfies a geometric Brownian motion (*GBM*), and we assume that the expert can obtain the true values of μ and σ during the whole investment period due to their investment expertise. Conversely, the normal investor needs to estimate μ and σ at each time t from $\{S(u)\}_{u\in[0,t]}$, denoted as $\hat{\mu}(t)$ and $\hat{\sigma}(t)$, respectively.

We use subscripts $i = 1, 2$ to denote the normal investor and the expert, respectively. We define the i-th investor's *decision* $\{P_i(t)\}_{t\in\mathcal{T}}$ as the wealth invested in the risky asset. Following the work in [1], given the true values of r, v, and σ, the expert's rational decision with complete information (*CRD*) $\{\bar{P}_2(t)\}_{t\in\mathcal{T}}$ without considering other's influence is

$$\bar{P}_2(t) := \frac{v}{\alpha_2\sigma^2}\mathrm{e}^{r(t-T)}, t \in \mathcal{T}, \tag{2}$$

where $\alpha_2 > 0$ is called the expert's risk aversion coefficient [13], and a larger α_2 indicates that he/she is less risk-taking.

The normal investor's decision process is in Fig. 1. At each time t, the normal investor with wealth $X_1(t)$ first estimates $\hat{v}(t)$ and $\hat{\sigma}(t)$ using historical price data $\{S(u)\}_{u\in[0,t]}$. Then, he/she makes the optimal decision (*OPD*) $\{P_1^*(u)\}_{u\in[t,T]}$ to maximize the expected utility of the estimated terminal wealth (*EUETW*) $\mathbb{E}_t\phi(\hat{X}_1(T))$ while minimizing the distance between his/her decision and that of the expert $D_t(P_1, \bar{P}_2)$, which we call the average deviation. Finally, he/she invests $P_1^*(t)$ in the risky asset and $X_1(t) - P_1^*(t)$ in the risk-free asset at time t,

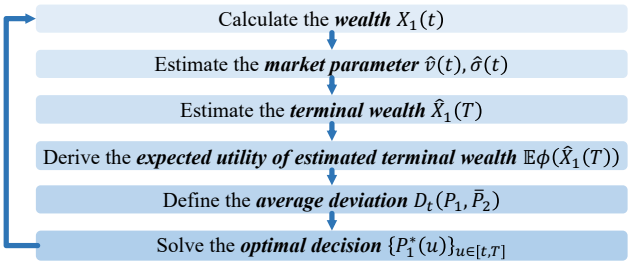


Fig. 1. The normal investor's decision process.

and then proceeds to the next time step $t + \mathrm{d}t$ and updates his/her wealth $X_1(t + \mathrm{d}t)$ using $P_1^*(t)$ and $S(t)$.

B. Market Parameter Estimation Methods

1) The general estimation methods: Based on the statistical properties of *GBM*, the work in [9] provides the following two general estimation methods for the market parameters. To simplify the notation, we denote N as the sample size of the *GBM* and denote the sample value at sample point n as

$$S_n(t) := S\left(\frac{n}{N}t\right), n = 0, \ldots, N. \tag{3}$$

General Method Using Differences (*GEM-diff*): *GEM-diff* is based on the probability distribution of the *difference* between successive observations, i.e., $S_n(t) - S_{n-1}(t), n \in \{1, \ldots, N\}$, and the maximum likelihood estimators (*MLE*s) of $\hat{\mu}_{\text{diff}}^{\text{GEM}}(t)$ and $\hat{\sigma}_{\text{diff}}^{\text{GEM}}(t)$ at time t are

$$\hat{\mu}_{\text{diff}}^{\text{GEM}}(t) = \frac{1}{t}\sum_{n=1}^{N}\frac{S_n(t) - S_{n-1}(t)}{S_{n-1}(t)}, \quad \text{and} \tag{4}$$

$$\hat{\sigma}_{\text{diff}}^{\text{GEM}}(t) = \sqrt{\frac{1}{t}\sum_{n=1}^{N}\left[\frac{S_n(t) - S_{n-1}(t)}{S_{n-1}(t)} - \frac{t}{N}\hat{\mu}_{\text{diff}}^{\text{GEM}}(t)\right]^2}.$$

General Method Using Ratios (*GEM-ratio*): *GEM-ratio* is based on the probability distribution of the ratio between successive observations, i.e., $S_n(t)/S_{n-1}(t), n \in \{1, \ldots, N\}$, and the *MLE*s of $\hat{\mu}_{\text{ratio}}^{\text{GEM}}(t)$ and $\hat{\sigma}_{\text{ratio}}^{\text{GEM}}(t)$ at time t are

$$\hat{\mu}_{\text{ratio}}^{\text{GEM}}(t) = \frac{1}{t}\ln\frac{S(t)}{S(0)} + \frac{[\hat{\sigma}_{\text{ratio}}^{\text{GEM}}(t)]^2}{2}, \quad \text{and} \tag{5}$$

$$\hat{\sigma}_{\text{ratio}}^{\text{GEM}}(t) = \sqrt{\frac{1}{t}\sum_{n=1}^{N}\left[\ln\frac{S_n(t)}{S_{n-1}(t)} - \frac{1}{N}\ln\frac{S(t)}{S(0)}\right]^2}.$$

These two general estimation methods assume that normal investors are fully rational and use *all* historical price data $\{S(t)\}_{t\in\mathcal{T}}$ when estimating μ and r. To model their bounded rationality, we modify these methods and propose the following sliding window-based methods.

2) The sliding window-based methods: From [10], investors often focus on price data within a certain timeframe. Thus, we assume a sliding window with width $w < t$, and the estimators only use $\{S(u)\}_{u\in[t-w,t]}$. We denote the sample value at sample point n with sliding window width w as

$$S_n^w(t) := S\left(\frac{nw}{N} + t - w\right), n = 0, \ldots, N. \tag{6}$$

TABLE II
THE MEAN AND VARIANCE OF DIFFERENT ESTIMATORS

$\hat{\mu}(t)$	$\mathbb{E}\hat{\mu}(t)$	$\mathbb{D}\hat{\mu}(t)$
$\hat{\mu}_{\text{diff}}^{\text{GEM}}(t)$ in (4)	μ	$\frac{\sigma^2}{t}$
$\hat{\mu}_{\text{diff}}^{\text{SWM}}(t)$ in (5)	μ	$\frac{\sigma^2}{w}$
$\hat{\mu}_{\text{ratio}}^{\text{GEM}}(t)$ in (7)	$\mu - \frac{\sigma^2}{2N}$	$\frac{\sigma^2}{t} + \frac{(N-1)\sigma^4}{2N^2}$
$\hat{\mu}_{\text{ratio}}^{\text{SWM}}(t)$ in (8)	$\mu - \frac{\sigma^2}{2N}$	$\frac{\sigma^2}{w} + \frac{(N-1)\sigma^4}{2N^2}$
$\hat{\sigma}^2(t)$	$\mathbb{E}\hat{\sigma}^2(t)$	$\mathbb{D}\hat{\sigma}^2(t)$
$[\hat{\sigma}_{\text{diff}}^{\text{GEM}}(t)]^2$ in (4), $[\hat{\sigma}_{\text{diff}}^{\text{SWM}}(t)]^2$ in (5), $[\hat{\sigma}_{\text{ratio}}^{\text{GEM}}(t)]^2$ in (7), $[\hat{\sigma}_{\text{ratio}}^{\text{SWM}}(t)]^2$ in (8)	$\frac{N-1}{N}\sigma^2$	$\frac{2(N-1)\sigma^4}{N^2}$

Based on *GEM-diff/ratio*, we propose the **Sliding Window-Based Method Using Differences (*SWM-diff*)**, where

$$\hat{\mu}_{\text{diff}}^{\text{SWM}}(t) = \frac{1}{w}\sum_{n=1}^{N}\frac{S_n^w(t) - S_{n-1}^w(t)}{S_{n-1}^w(t)}, \quad \text{and} \tag{7}$$

$$\hat{\sigma}_{\text{diff}}^{\text{SWM}}(t) = \sqrt{\frac{1}{w}\sum_{n=1}^{N}\left[\frac{S_n^w(t) - S_{n-1}^w(t)}{S_{n-1}^w(t)} - \frac{w}{N}\hat{\mu}_{\text{diff}}^{\text{SWM}}(t)\right]^2},$$

and the **Sliding Window-Based Method Using Ratios (*SWM-ratio*)**, where

$$\hat{\mu}_{\text{ratio}}^{\text{SWM}}(t) = \frac{1}{w}\ln\frac{S(t)}{S(t-w)} + \frac{[\hat{\sigma}_{\text{ratio}}^{\text{SWM}}(t)]^2}{2}, \quad \text{and} \tag{8}$$

$$\hat{\sigma}_{\text{ratio}}^{\text{SWM}}(t) = \sqrt{\frac{1}{w}\sum_{n=1}^{N}\left[\ln\frac{S_n^w(t)}{S_{n-1}^w(t)} - \frac{1}{N}\ln\frac{S(t)}{S(t-w)}\right]^2}.$$

Detailed derivations of (7)–(8) are in [14].

3) Comparison of different estimation methods: We theoretically analyze the statistical properties of the above estimation methods. The mean and the variance of different estimators, in (4)–(8) are in Table II. We use $\mathbb{E}X$ and $\mathbb{D}X$ to denote the mean and variance of an estimator X. The proofs are in [14]. Note that it is challenging to derive $\mathbb{E}\hat{\sigma}(t)$ and $\mathbb{D}\hat{\sigma}(t)$, so we analyze $\mathbb{E}\hat{\sigma}^2(t)$ and $\mathbb{D}\hat{\sigma}^2(t)$ instead.

From Table II, for the estimation of μ, $\hat{\mu}_{\text{diff}}^{\text{GEM}}(t)$ and $\hat{\mu}_{\text{diff}}^{\text{SWM}}(t)$ are unbiased, while $\hat{\mu}_{\text{ratio}}^{\text{GEM}}(t)$ and $\hat{\mu}_{\text{ratio}}^{\text{SWM}}(t)$ are biased. This suggests that estimating the return using differences gives higher accuracy than ratios. Furthermore, when studying the variance of different estimators, we have $\mathbb{D}\hat{\mu}_{\text{diff}}^{\text{GEM}}(t) < \mathbb{D}\hat{\mu}_{\text{diff}}^{\text{SWM}}(t)$ and $\mathbb{D}\hat{\mu}_{\text{ratio}}^{\text{GEM}}(t) < \mathbb{D}\hat{\mu}_{\text{ratio}}^{\text{SWM}}(t)$, which suggests that estimating the return with bounded rationality results in higher uncertainty compared to the scenario with full rationality.

Additionally, from Table II, for the estimation of σ^2, the mean and the variance of different estimators are the same, which suggests that a bounded rational normal investor's estimation of volatility is as accurate and certain as those of fully rational investors. As the sample size N approaches infinity, $\mathbb{E}\hat{\sigma}^2(t)$ converges to the true value σ^2, and $\mathbb{D}\hat{\sigma}^2(t)$ approaches zero. Therefore, $\hat{\sigma}^2(t)$ consistently converges to σ^2, i.e., $\lim_{N\to\infty}\mathbb{P}(\hat{\sigma}^2(t) = \sigma^2) = 1$, and in this case, the normal investor can almost surely obtain the true value of σ^2. In summary, the accuracy and uncertainty of volatility estimation depend on the sample size and are independent of whether the investor is bounded rational or not.

4) Summary: We assume that the expert can acquire the true values of r, μ, v, and σ, while the normal investor can acquire the true value of r, and use (4)–(9) to estimate μ and σ for $t \in \mathcal{T}$. In the following, when not referring to a specific estimator, we simplify the notation by using $\hat{\mu}(t)$ and $\hat{\sigma}(t)$ to denote the estimated parameters in general. We will specify each estimator when comparing their performances. From (1), the *MLE* of the excess return $\hat{v}(t)$ at time t is

$$\hat{v}(t) = \hat{\mu}(t) - r. \tag{9}$$

C. The Normal Investor's Optimal Investment Problem

Next, we analyze the normal investor's wealth, the *EUETW*, and the average deviation, and formulate the problem.

1) The wealth: Following the work in [1], at time t, given the historical price data $\{S(u)\}_{u\in[0,t]}$ and the normal investor's historical *OPD* $\{P_1^*(u)\}_{u\in[0,t]}$, his/her wealth process $\{X_1(u)\}_{u\in[0,t]}$ satisfies

$$\mathrm{d}X_1(u) = r[X_1(u) - P_1^*(u)]\mathrm{d}u + \frac{P_1^*(u)}{S(u)}\mathrm{d}S(u), \tag{10}$$

subject to $X_1(0) = x_1$, where x_1 is his/her initial wealth.

2) The EUETW: Following the work in [1], at time t, given the market parameter estimates $\hat{\mu}(t)$, $\hat{v}(t)$, and $\hat{\sigma}(t)$, the normal investor estimates that his/her wealth $\{\hat{X}_1(u)\}_{u\in[t,T]}$ satisfies

$$\mathrm{d}\hat{X}_1(u) = [r\hat{X}_1(u) + \hat{v}(t)P_1(u)]\mathrm{d}u + \hat{\sigma}(t)P_1(t)\mathrm{d}B(u), \tag{11}$$

subject to $\hat{X}_1(t) = X_1(t)$, if he/she allocates $\{P_1(u)\}_{u\in[t,T]}$ to the risky asset. Based on (11), the normal investor chooses his/her optimal decision $\{P_1^*(u)\}_{u\in[t,T]}$ to maximize his/her *EUETW*. The utility function $\phi(\hat{X}_1(T))$ satisfies the characteristics of the diminishing marginal returns and concavity. Following the work in [13], we consider the following form:

$$\phi(\hat{X}_1(T)) := -\frac{1}{\alpha_1}\mathrm{e}^{-\alpha_1\hat{X}_1(T)}, \tag{12}$$

where $\alpha_1 > 0$ is the normal investor's risk aversion coefficient.

3) The average deviation: Following our prior work in [12], we use the average deviation to measure the distance between the two investors' decisions. At time t, the average deviation between the normal investor's decision $\{P_1(u)\}_{u\in[t,T]}$ and the expert's *CRD* $\{\bar{P}_2(u)\}_{u\in[t,T]}$ is

$$D_t(P_1, \bar{P}_2) := \frac{1}{2}\int_t^T \mathrm{e}^{\rho r(T-u)}[P_1(u) - \bar{P}_2(u)]^2\mathrm{d}u, \tag{13}$$

where $\rho \geqslant 0$ is the decay rate. A larger ρ implies that deviations occurring later carry less weight. When $\rho = 0$, deviations for all times are equally weighted.

4) Optimal investment problem: In summary, the optimal investment problem for the normal investor at time t becomes

$$\sup_{\mathcal{U}_t \ni \{P_1(u)\}_{u\in[t,T]}} \mathbb{E}_t\phi(\hat{X}_1(T)) - \theta D_t(P_1, \bar{P}_2) \tag{14}$$

$$\text{s.t.} \quad \mathrm{d}\hat{X}_1(u) = [r\hat{X}_1(u) + \hat{v}(t)P_1(u)]\mathrm{d}u + \hat{\sigma}(t)P_1(t)\mathrm{d}B(u),$$
$$\mathrm{d}X_1(u) = r[X_1(u) - P_1^*(u)]\mathrm{d}u + \frac{P_1^*(u)}{S(u)}\mathrm{d}S(u),$$
$$\hat{X}_1(t) = X_1(t), \quad X_1(0) = x_1,$$

where $\mathcal{U}_t$ represents the set of admissible decisions, which is a subset of $\mathcal{L}_t^1(\mathcal{T}) := \left\{ \{P(u)\}_{u\in[t,T]} \middle| \mathbb{E}\int_t^T |P(u)|\mathrm{d}u < +\infty \right\}$. The herd coefficient $\theta > 0$ is to address the tradeoff between the two different objectives, i.e., maximizing the *EUETW* $\mathbb{E}_t\phi(\hat{X}_1(T))$ and minimizing the average deviation $D_t(P_1, \bar{P}_2)$. Specifically, as θ approaches infinity, the normal investor copies the expert's *CRD*:

$$\lim_{\theta\to+\infty} P_1^*(t) = \bar{P}_2(t), t \in \mathcal{T}. \tag{15}$$

When $\theta = 0$, the normal investor's *OPD* is entirely independent of the expert's decision, and becomes the rational decision using the estimated $\hat{v}(t)$ and $\hat{\sigma}(t)$, i.e.,

$$\hat{\bar{P}}_1(t) := \lim_{\theta\to 0} P_1^*(t) = \frac{\hat{v}(t)}{\alpha_1\hat{\sigma}^2(t)}\mathrm{e}^{r(t-T)}, t \in \mathcal{T}, \tag{16}$$

which we call the normal investor's rational decision with incomplete information (*IRD*). Furthermore, we define the normal investor's *CRD*, i.e., the rational decision using the true values of v and σ as

$$\bar{P}_1(t) := \frac{v}{\alpha_1\sigma^2}\mathrm{e}^{r(t-T)}, t \in \mathcal{T}. \tag{17}$$

From (16) and our prior work in [12], we have

$$\{\hat{\bar{P}}_1(t)\}_{t\in\mathcal{T}} = \mathrm{argmax}_{\mathcal{U}_0\ni\{P_1(t)\}_{t\in\mathcal{T}}}\mathbb{E}_0\phi(\hat{X}_1(T)), \text{ and} \tag{18}$$

$$\{\bar{P}_1(t)\}_{t\in\mathcal{T}} = \mathrm{argmax}_{\mathcal{U}_0\ni\{P_1(t)\}_{t\in\mathcal{T}}}\mathbb{E}_0\phi(X_1(T)). \tag{19}$$

That is, the *IRD* $\{\hat{\bar{P}}_1(t)\}_{t\in\mathcal{T}}$ maximizes the expected utility of the *estimated* terminal wealth $\hat{X}_1(T)$ calculated using (11), while the *CRD* $\{\bar{P}_1(t)\}_{t\in\mathcal{T}}$ maximizes the expected utility of the *real* terminal wealth $X_1(T)$ calculated using (10). From the work in [1], the normal investor's objective is to maximize $\mathbb{E}_0\phi(X_1(T))$. However, due to incomplete information, the normal investor can only obtain $\hat{X}_1(T)$ but not $X_1(T)$ in the decision-making process. From (16) and (17), when the estimators $\hat{v}(t)$ and $\hat{\sigma}(t)$ are more accurate and certain, the *IRD* $\{\hat{\bar{P}}_1(t)\}_{t\in\mathcal{T}}$ aligns better with the *CRD* $\{\bar{P}_1(t)\}_{t\in\mathcal{T}}$, and the normal investor's expected utility of the real terminal wealth $\mathbb{E}_0\phi(X_1(T))$ increases. As mentioned above, the presence of bounded rationality in estimators leads to higher uncertainty, which in turn diminishes the expected utility of the real terminal wealth.

III. Problem Solution and Analysis

In this section, we derive the analytical solution of the normal investor's *OPD* in (14) and theoretically analyze the influence of incomplete information and herd effect on the normal investor's *OPD*. To simplify the notation, we denote

$$\vartheta(t) := \frac{\theta}{\alpha_1\hat{\sigma}^2(t)}, \quad \text{and} \quad \varrho := 2 - \rho \tag{20}$$

as the modified herd coefficient and the modified decay rate, respectively. All proofs are in [14].

A. The Normal Investor's Optimal Decision

Theorem 1. *The optimal decision* $\{P_1^*(t)\}_{t\in\mathcal{T}}$ *in* (14) *is*

$$P_1^*(t) = \frac{\eta(t)\hat{v}(t)\mathrm{e}^{(\varrho-1)r(T-t)} + \theta\bar{P}_2(t)}{\eta(t)\alpha_1\hat{\sigma}^2(t)\mathrm{e}^{\varrho r(T-t)} + \theta}, t \in \mathcal{T}, \tag{21}$$

where the integral constant $\eta(t)$ *is defined as*

$$\eta(t) := -\alpha_1\mathbb{E}_t\phi(\hat{X}_1(T)) = \exp\left\{-\alpha_1 X_1(t)\mathrm{e}^{r(T-t)} - \frac{\hat{v}^2(t)(T-t)}{2\hat{\sigma}^2(t)} + \int_t^T \frac{\vartheta^2(t)\hat{v}^2(t)[\lambda\alpha_1\hat{\sigma}^2(t)/\hat{v}(t) - 1]^2\mathrm{d}u}{2\hat{\sigma}^2(t)[\eta(t)\mathrm{e}^{\varrho r(T-u)} + \vartheta(t)]^2}\right\}, \tag{22}$$

and his/her wealth process $\{X_1(t)\}_{t\in\mathcal{T}}$ *is*

$$X_1(t) = x_1\mathrm{e}^{rt} + \int_0^t \frac{\eta(u)\hat{v}(u)\mathrm{e}^{\varrho r(T-u)} + \lambda\theta}{\eta(u)\alpha_1\hat{\sigma}^2(u)\mathrm{e}^{\varrho r(T-u)} + \theta}\left(\frac{\mathrm{d}S(u)}{S(u)} - r\mathrm{d}u\right), \tag{23}$$

where the portfolio parameter $\lambda := \frac{v}{\alpha_2\sigma^2} \equiv \bar{P}_2(t)\mathrm{e}^{r(T-t)}$ *for* $t \in \mathcal{T}$ *can be calculated from the expert's CRD.*

Note that it is challenging to derive the closed-form solution of the normal investor's *OPD* due to the complexity of (21), (22), and (23). To address this problem, we provide a numerical method. We discretize the investment period into several subintervals, each of which has a length of τ, which satisfies that $M := T/\tau$ is an integer, i.e., $\mathcal{T} = \bigcup_{i=1}^M [(i-1)\tau, i\tau]$. Starting from the initial time with a given initial wealth $X_1(0) = x_1$, we first estimate the market parameters $\hat{\mu}(\tau)$ and $\hat{\sigma}(\tau)$, and solve for the integral constant $\eta(\tau)$ using (22). Note that when τ is sufficiently small, we can approximate $X_1(\tau)$ as $X_1(0)$ in (22), i.e.,

$$\eta(\tau) \approx \exp\left\{-\alpha_1 X_1(0)\mathrm{e}^{r(T-\tau)} - \frac{\hat{v}^2(\tau)(T-\tau)}{2\hat{\sigma}^2(\tau)} + \int_\tau^T \frac{\vartheta^2(\tau)\hat{v}^2(\tau)[\lambda\alpha_1\hat{\sigma}^2(\tau)/\hat{v}(\tau) - 1]^2\mathrm{d}u}{2\hat{\sigma}^2(\tau)[\eta(\tau)\mathrm{e}^{\varrho r(T-u)} + \vartheta(\tau)]^2}\right\}. \tag{24}$$

Then, we calculate the *OPD* $P_1^*(\tau)$ and wealth $X_1(\tau)$ using (21) and (23), respectively. We iteratively calculate $\hat{v}(i\tau)$, $\hat{\sigma}(i\tau)$, $\eta(i\tau)$, $P_1^*(i\tau)$, and $X_1(i\tau)$, for $i \in \{2, \ldots, M\}$, until reaching the terminal time T.

B. Influence of Incomplete Information and Herd Effect on the Normal Investor's Optimal Decision

Due to the complexity of (21), it is challenging to theoretically analyze the influence of incomplete information and herd effect on the *OPD*. Note that the normal investor's *IRD* in (16) reflects the influence of incomplete information, while the expert's *CRD* in (2) reflects the influence of herd effect. Therefore, we can decompose the normal investor's *OPD* into his/her *IRD* and the expert's *CRD*, and use the weight function to measure the influence of incomplete information and herd effect on the *OPD*. Following our prior work in [12], the normal investor's *OPD* in (21) is a convex linear combination of his/her *IRD* and the expert's *CRD*.

Theorem 2. *The normal investor's optimal decision* $\{P_1^*(t)\}_{t\in\mathcal{T}}$ *in* (14) *is a convex linear combination of the two investors' rational decisions* $\{\hat{\bar{P}}_1(t)\}_{t\in\mathcal{T}}$ *and* $\{\bar{P}_2(t)\}_{t\in\mathcal{T}}$*, i.e.,*

$$P_1^*(t) = Z_1(t)\hat{\bar{P}}_1(t) + [1 - Z_1(t)]\bar{P}_2(t), t \in \mathcal{T}, \tag{25}$$

where the weight function $\{Z_1(t)\}_{t\in\mathcal{T}}$ *is*

$$Z_1(t) = \frac{\eta(t)\mathrm{e}^{\varrho r(T-t)}}{\eta(t)\mathrm{e}^{\varrho r(T-t)} + \vartheta(t)} \in [0,1], t \in \mathcal{T}. \tag{26}$$

From (25), $\{Z_1(t)\}_{t\in\mathcal{T}}$ measures the extent to which the normal investor adheres to his/her *IRD*. A larger $\{Z_1(t)\}_{t\in\mathcal{T}}$ indicates that the normal investor's *OPD* aligns more with his/her *IRD*. Conversely, a smaller $\{Z_1(t)\}_{t\in\mathcal{T}}$ indicates a preference for aligning with the expert's *CRD* rather than relying on his/her market parameter estimates when making decisions. Thus, we define $\{Z_1(t)\}_{t\in\mathcal{T}}$ as the normal investor's *opinion* to measure the influence of estimations with incomplete information and herd effect on the *OPD*.

From (20), $\vartheta(t)$ is proportional to θ. A larger θ indicates a higher level of herd effect, and the normal investor is more inclined to the experts' *CRD* when making decisions. In this case, from (26), the opinion $\{Z_1(t)\}_{t\in\mathcal{T}}$ will decrease, and from (25), the normal investor's *OPD* $\{P_1^*(t)\}_{t\in\mathcal{T}}$ will align more with the expert's *CRD* $\{\bar{P}_2(t)\}_{t\in\mathcal{T}}$.

From (22), $\eta(t)$ is negatively proportional to the *EUETW* $\mathbb{E}_t\phi(\hat{X}_1(T))$. Considering the characteristics of the diminishing marginal returns of the utility function, if the *EUETW* $\mathbb{E}_t\phi(\hat{X}_1(T))$ is smaller, i.e., the integral constant $\eta(t)$ is larger, the normal investor is more motivated to increase his/her *EUETW*. In this case, from (26), the opinion $\{Z_1(t)\}_{t\in\mathcal{T}}$ will increase, and thus from (25), the normal investor's *OPD* $\{P_1^*(t)\}_{t\in\mathcal{T}}$ will align more with his/her *IRD* $\{\hat{\bar{P}}_1(t)\}_{t\in\mathcal{T}}$, and he/she is more inclined to estimate the return and volatility himself/herself from historical price data.

IV. Numerical Experiments

In this section, we validate our proposed estimation methods and the analysis of the normal investor's *OPD* in Section II and Section III using numerical experiments.

A. Comparison of Different Estimation Methods

We consider a time frame spanning fifty years from 1974 to 2023, i.e., $T = 50$. We set the true values of the return and the volatility as $\mu = 0.07$ and $\sigma = 0.16$ according to the Dow Jones Industrial Average [15]. We generate 500 trajectories following the *GBM* with $\mu = 0.07$ and $\sigma = 0.16$. We use all the mentioned methods, including *GEM-diff*, *GEM-ratio*, *SWM-diff*, and *SWM-ratio*, to estimate $\hat{\mu}(t)$ and $\hat{\sigma}(t)$, and we compare their performance. We set $N = 1,000$ and $w = 3$, and observe the same trend for other values of N and w.

Fig. 2 shows the estimated return and volatility averaged over $1,000$ simulation results, and we also plot the true values of the return and volatility for comparison. The curves of *SWM-diff* and *SWM-ratio* in Fig. 2 overlap almost completely. From Fig. 2, when $t \in [0, 20)$, the accuracy of parameter estimation across all methods is relatively low due to limited data. Therefore, in the following experiment, we assume that during $t \in [0, 20)$, the normal investor observes data only, while during $t \in [20, T]$, he/she observes data and makes decisions. This assumption also satisfies our problem formulation in Section II, which is explained in [14].

When $t \in [20, T]$, for the return, the estimators of *GEM-diff/ratio* exhibit lower uncertainty than those of *SWM-diff/ratio*. It indicates that a bounded rational investor has higher uncertainty in estimating the return than a fully rational investor. Additionally, it can be observed that $\hat{\mu}_{\text{diff}}^{\text{GEM}}(t)$ is closer to the true value of μ than $\hat{\mu}_{\text{ratio}}^{\text{GEM}}(t)$. Therefore, estimating the return using differences and with full rationality leads to higher accuracy and certainty than using ratios and with bounded rationality, respectively. For the volatility, the four estimators demonstrate comparable performance. In [14], we show that as the sample size increases, the estimator of volatility has higher accuracy and certainty. The above results verify our analysis in Section II.

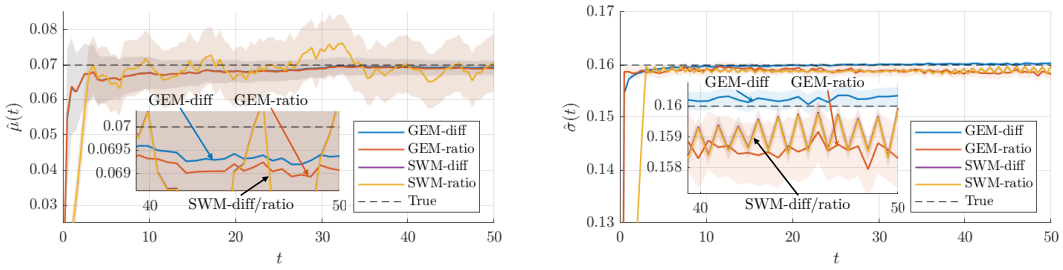
B. Influence of Bounded Rationality on the Normal Investor's Rational Decision and Expected Utility

Next, we simulate the influence of bounded rationality on the *IRD*. We set $T = 50$, $\mu = 0.07$, and $\sigma = 0.16$ as mentioned above. The interest rate r is approximated as 0.04 using the daily average of U.S. 1-Year T-Bills' interest rates in 2022 and 2023 [16]. Following our prior work in [12], we set the normal investor's risk aversion coefficient as $\alpha_1 = 0.2$. Note that the initial wealth can be any positive number [12], and we set $x_1 = 1$ as an example. We set the sliding window width w as 10 and 20. A larger w indicates that the level of bounded rationality is weaker, and thus the estimation method is less uncertain. We observe the same trend for other parameters' values. We repeat the experiment $1,000$ times. We calculate the *IRD* and the *CRD* using (16) and (17), and their corresponding expected utilities of the real terminal wealth $\mathbb{E}_0\phi(X_1(T))$ using (23) by averaging $\phi(X_1(T))$ in all experiments.

We analyze the influence of bounded rationality on the normal investor's *IRD* in [14]. The experiment results show that when the estimation method is more accurate, the *IRD* aligns better with the *CRD*. Table III shows the expected utility of the real terminal wealth with different estimation methods. From Table III, we can observe that the expected utility with complete information of μ and σ is the highest, and among all estimators, *GEM-diff* achieves the highest expected utility due to its highest accuracy and certainty; while *SWM-ratio* ($w = 10$) gives the lowest expected utility as its accuracy and certainty are the lowest. The above results verify our analysis in Section II.

C. Influence of Incomplete Information and Herd Effect on the Normal Investor's Optimal Decision

Next, we simulate the influence of incomplete information and herd effect on the *OPD*. Apart from the aforementioned parameters, we set the expert's risk aversion coefficient as $\alpha_2 = 0.4$ and the decay rate as $\rho = 2$. We let the herd



(a) The estimated return $\hat{\mu}(t)$ (b) The estimated volatility $\hat{\sigma}(t)$

Fig. 2. Comparison of different estimation methods: the estimated return $\hat{\mu}(t)$ and the estimated volatility $\hat{\sigma}(t)$.

TABLE III
INFLUENCE OF BOUNDED RATIONALITY ON THE NORMAL INVESTOR'S EXPECTED UTILITY OF THE REAL TERMINAL WEALTH

Estimation method	*CRD*	*GEM-diff*	*GEM-ratio*	*SWM-diff* ($w = 20$)	*SWM-ratio* ($w = 20$)	*SWM-diff* ($w = 10$)	*SWM-ratio* ($w = 10$)
$\mathbb{E}_0\phi(X_1(T))$	-3.9137	-4.3601	-4.3643	-4.6937	-4.6950	-6.7842	-6.7980

coefficient θ be 10^{-4}, 10^{-3}, and 10^{-2}. We observe the same trend for other parameters' values. We assume that the normal investor estimates the market parameters using *GEM-diff*, and calculates the *OPD*s and opinions using Theorem 1 and Theorem 2. The experiment results are in [14], which show that as the herd coefficient increases, the normal investor's *OPD* aligns more with the expert's *CRD* and deviates further from his/her *IRD*, and his/her opinion decreases, which verifies Theorem 2.

V. CONCLUSION

In this work, we formulate the optimal investment problem jointly considering incomplete information and the herd effect and derive the normal investor's optimal decision. We modify the general market parameter estimation methods and propose the sliding window-based methods considering bounded rationality. Our theoretical analysis shows that while the estimation of the return is more uncertain with bounded rationality, the uncertainty of volatility estimation remains unchanged. As parameter estimation methods become more accurate and certain, the normal investor's rational decision with incomplete information converges to that with complete information, leading to a higher expected utility of real terminal wealth. As the herd effect intensifies, the normal investor's optimal decision aligns more with the expert investor's decision and diverges from his/her own rational decision.

REFERENCES

[1] R. C. Merton, "Lifetime portfolio selection under uncertainty: The continuous-time case," *The Review of Economics and Statistics*, pp. 247–257, 1969.

[2] G. Gennotte, "Optimal portfolio choice under incomplete information," *The Journal of Finance*, vol. 41, no. 3, pp. 733–746, 1986.

[3] J. Lewellen, "Momentum and autocorrelation in stock returns," *The Review of Financial Studies*, vol. 15, no. 2, pp. 533–564, 2002.

[4] J. R. Brown, Z. Ivković, P. A. Smith, and S. Weisbenner, "Neighbors matter: Causal community effects and stock market participation," *Journal of Finance*, vol. 63, no. 3, pp. 1509–1531, 2008.

[5] V. M. Eguiluz and M. G. Zimmermann, "Transmission of information and herd behavior: An application to financial markets," *Physical Review Letters*, vol. 85, no. 26, p. 5659, 2000.

[6] W. F. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," *Journal of Finance*, vol. 19, no. 3, pp. 425–442, 1964.

[7] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986.

[8] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, vol. 81, no. 3, pp. 637–654, 1973.

[9] J. C. Hull and S. Basu, *Options, Futures, and Other Derivatives*. Pearson Education India, 2016.

[10] J. Conlisk, "Why bounded rationality?" *Journal of Economic Literature*, vol. 34, no. 2, pp. 669–700, 1996.

[11] K. Yu, Y. Li, and Z. Cao, "Detection and estimation of herd behavior in futures market," in *2019 International Conference on Economic Management and Model Engineering (ICEMME)*, IEEE, 2019, pp. 213–223.

[12] H. Wang and H. V. Zhao, "Optimal investment with herd behaviour using rational decision decomposition," in *2024 43rd Chinese Control Conference (CCC)*, IEEE, 2024, pp. 1645–1651.

[13] J. W. Pratt, "Risk aversion in the small and in the large," in *Uncertainty in economics*, Elsevier, 1978, pp. 59–79.

[14] *The supplementary file*, https://drive.google.com/file/d/1pa3vucxlabQAzdicu9Na5ZC6Yyr0Kw8Z/view?usp=drive_link.

[15] https://www.spglobal.com/spdji/.

[16] https://home.treasury.gov/policy-issues/financing-the-government/interest-rate-statistics.

A Two-Stage Wall Parameters Estimation Algorithm for MIMO Through-the-Wall Radar

Zhirun Li*, Shisheng Guo*†, Jiahui Chen*, Zhihao Zhu*, Chen Qiu*, Cui Guolong*† and Yutao Xiang*
* University of Electronic Science and Technology of China, Chengdu
† Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Quzhou 324000, China
E-mail: lzr18842200408@163.com, ssguo@uestc.edu.cn, chenjiahui@uestc.edu.cn, zzh_991227@163.com,
qiu_chen0522@163.com, cuiguolong@uestc.edu.cn, e1ysium@qq.com

***Abstract*—This paper proposes a two-stage wall parameter estimation method based on multi-channel to retrieve the unknown position, thickness, and relative permittivity of wall. In the first stage, it aims to estimate the slope and intercept of the front surface of the wall. This is achieved by formulating a cost-minimization problem, where the position of the front wall is modeled using a linear analytical expression. Subsequently, in the second stage, it focuses on calculating the positions of equivalent arrays that are aligned parallel to the wall for each transmitting antenna, so as to enable the estimation of the wall's thickness and its relative permittivity. The effectiveness of the algorithm is validated through electromagnetic simulations.**

***Index Terms*—Two-stage wall parameter estimation, through-the-wall imaging, equivalent array position approximation, linear MIMO array**

I. Introduction

Through-the-wall imaging radar (TWIR) uses electromagnetic (EM) waves to penetrate building walls for concealed target detection, localization, tracking and imaging. It has significant applications in counter-terrorism, surveillance and disaster rescue. During the imaging process, neglecting wall penetration effects lead to defocused target images and positional shifts, posing difficulties in subsequent target localization and identification [1]–[3]. Therefore, the precise estimation of wall parameters is crucial for effective TWIR.

Existing methods for estimating unknown wall parameters fall into two categories: wall feature value matching-based and time-delay-based methods. The former feature value based methods involve extracting parameters from radar echoes and matching them with known wall parameters. For example, in [4], matching is achieved by altering the array or its position relative to the wall. However, this method is operationally complex and constrained by the site. The time-delay methods estimate wall parameters based on the delay difference between echoes from the front and back walls, and they involve using various array structures or adjusting the distance from the array to the wall [5], including algorithms based on the eigenvalue decomposition of correlation matrices [6] and sparse reconstruction-based estimation algorithms [7]. The accuracy of these methods depends on the time-delay estimation results, and require the array structure to be parallel to the wall, which may limit its further application. This paper introduces a two-stage algorithm to estimate unknown parameters of inclined walls using a multiple-input-multiple-out (MIMO) array. In the initial stage, the algorithm iterates through various slopes and intercepts corresponding to the wall's geometry. During this process, it identifies the mirror points that correlate with the position of the transmitting antenna. Subsequently, a cost function is constructed to determine the set of slope and intercept values that most closely approximate the actual parameters of the wall. In the second stage, it calculates the equivalent array positions parallel to the wall, so as to estimate the thickness and relative permittivity of the wall. EM simulation validates the effectiveness of the proposed algorithm.

II. The Propagation of EM Waves in a Single Layer Wall

As shown in Fig. 1, the transmit antenna *(TX)* and receive antenna *(RX)* are horizontally placed. Assuming that the wall is homogeneous and smooth, the position of the front wall is expressed by equation $y = kx+b$. The line parallel to the front wall passing through *TX* is labeled as $y_2 = kx + c_1$, with h representing the distance from *TX* to the front wall. The mirror point of *TX* with respect to the front wall is *TX'*. Connecting *TX'* to *RX*, the intersection points at y and y_2 are defined as P_1 and Q, respectively. The angle between the line of the transmitting and receiving antennas and y is θ. The thickness and relative permittivity of the wall are denoted as d and ε, respectively. The red solid line represents the reflection path of EM on the front surface of the wall. The propagation path of EM on the rear surface of the wall is $TX \rightarrow P_2 \rightarrow P_3 \rightarrow P_4 \rightarrow RX$, intersecting y_2 at point D, whose coordinate is (x_D, y_D). D serves as the virtual equivalent receiving antenna for *RX* on this parallel line.

III. Estimation of Wall Parameters

A. Estimation of the Front Wall Position

Defining the coordinate origin as o, for the m-th (m=1,2,...,M) transmitting antenna and the n-th (n=1,2,...,N) receiving antenna, their coordinates are defined as $TX_m\,(x_m, y_m)$ and $RX_n\,(x_{rn}, y_{rn})$, respectively. The symmetric point of the m-th transmitting antenna with respect to the front wall is TX'_m, which is denoted as $(x_m{}', y_m{}')$, where

$$x_m{}' = (2k \cdot (y_m - b) - x_m(k^2 - 1))/(k^2 + 1) \quad (1)$$

$$y_m{}' = (2 \cdot (b + kx_m) + y_m(k^2 - 1))/(k^2 + 1) \quad (2)$$

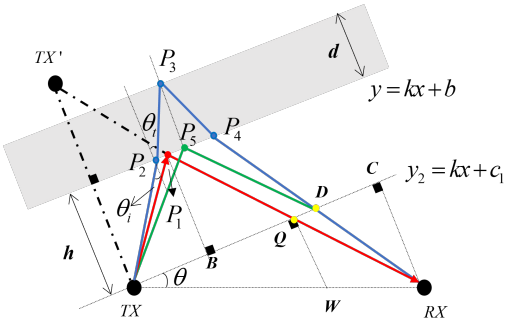


Fig. 1. Electromagnetic wave paths

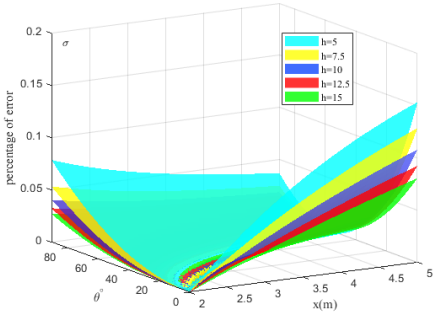


Fig. 2. Error percentage of (12)

The actual reflection path of the EM wave with respect to the front wall is denoted as $TX_m \rightarrow P_1 \rightarrow RX_n$, which can be equivalently represented by the propagation from the symmetric point $TX_m{}'$ to RX_n. The real and estimated time delay from TX to RX can be defined as t_n and $\hat{t}_n$. Specifically, $\hat{t}_n$ can be expressed as

$$\hat{t}_n(k,b) = \frac{\left\| TX_m{}'(k,b) - RX_n \right\|}{c} \tag{3}$$

where c represents the speed of light, $\|\cdot\|$ represents Euclidean distance. $TX_m{}'(k,b)$ represents the position of $TX_m{}'$ with different k and b.

The mean square error between t_n and $\hat{t}_n$ is defined as the cost function for estimating k and b. The parameters k and b can be estimated by minimizing the following cost function using Nelder-Mead algorithm [8].

$$f(k,b) = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \left| t_n - \hat{t}_n(k,b) \right|^2 \tag{4}$$

where $f(k,b)$ represents the amplitude of the function above with different k and b. The fundamental principle of this algorithm lies in the assumption that minimizing the amplitude of the cost function results in the theoretical time delay closely matching the measured time delay. This alignment is crucial as it enables the accurate estimation of the values for k and b, which are key parameters in the algorithm's process.

B. Estimation of Thickness and Permittivity of the Wall

For the model depicted in Fig. 1, the reception at *RX* is primarily composed of three main parts: direct coupling between *TX* and *RX*, reflection on the front surface of the wall (NSR), and reflection on the rear surface of the wall (FSR). We use l to denote electrical lengths. All lengths mentioned below refer to electrical lengths. The length of NSR can be expressed as

$$l_{NSR} = l_{TXP_1} + l_{P_1RX} = 2l_{TXP_1} + l_{QRX} \tag{5}$$

Refraction occurs at point P_2, where the incident angle θ_i and refracted angle θ_t follow Snell's Law and can be expressed as

$$\frac{\sin\theta_i}{\sin\theta_t} = \sqrt{\varepsilon} \tag{6}$$

where ε represents the relative permittivity of the wall. The equivalent electrical length of FSR can be expressed as

$$l_{FSR} = 2(l_{TXP_2} + \sqrt{\varepsilon} l_{P_2P_3}) + l_{DRX} \tag{7}$$

According to geometric relationships, we can obtain

$$l_{TXP_2} = l_{TXP_5} - d\tan\theta_t \sin\theta_i \tag{8}$$

$$l_{P_2P_3} = d/\cos\theta_t \tag{9}$$

Furthermore, we can derived the following equation by putting (8) and (9) into (7)

$$l_{FSR} = 2(l_{TXP_5} + d\sqrt{\varepsilon - \sin^2\theta_i}) + l_{DRX} \tag{10}$$

The time delay difference between the front and rear surfaces is the function of the incident angle for ε and d.

$$\begin{aligned} t_d(\theta_i) &= (l_{FSR} - l_{NSR})/c \\ &= \left(2\left(l_{TXP_5} - l_{TXP_1} + d\sqrt{\varepsilon - \sin^2\theta_i}\right) + l_{DRX} - l_{QRX}\right)/c \\ &= \left((l_{TX'D} + l_{DRX}) - l_{TX'RX} + 2d\sqrt{\varepsilon - \sin^2\theta_i}\right)/c \end{aligned} \tag{11}$$

Assuming the mirror point of the transmitting antenna is sufficiently far from the receiving antenna, we have

$$l_{TX'RX} - l_{TX'D} \approx l_{DRX} \tag{12}$$

As shown in Fig. 1, the distance between *TX* and *RX* is defined as x. Keeping the position of the equivalent receiving antenna D unchanged, place *RX* at any point on the extension line of P_4D. As the position of *RX* varies, θ, x, and the position of Q also vary accordingly. The error percentage of (12) can be represented by

$$\sigma = \left| \frac{l_{TX'D} + l_{DRX} - l_{TX'RX}}{l_{TX'D} + l_{DRX}} \right| \tag{13}$$

Fig. 2 illustrates the percentage of error at different $\theta\,(0^\circ \sim 90^\circ)$ and $x\,(l_{TXD} \le x \le h_{\min})$ as h varies. When the distance from *TX* to the front wall is 1.5m, the error rate doesn't exceed 15.3%. Furthermore, as *TX* moves farther from the front wall, the error percentage decreases. At this point, $l_{TX'D} + l_{DRX}$ approximates a straight line, implying that point D can be approximated as the point Q.

For a MIMO array with *M* transmitting elements and *N* receiving elements, building upon (11), the time delay between the first reflection and the second reflection for the *m*-th transmitting element and the *n*-th receiving element is defined as $t_d(m,n)$, which is given by

$$t_d(m,n) \approx \left(2d\sqrt{\varepsilon - \sin^2\theta_i(m,n)}\right)/c \tag{14}$$

where

$$\theta_i(m,n) \approx \arctan\left(|x_Q - x_m|/2h\right) \tag{15}$$

The line through the *m*-th *TX* parallel to the front wall is defined as $y_2^m = kx + c_1^m$, and the distance from TX_m to the front wall is expressed as h_m. Since k can be obtained from (4), by substituting the coordinate of the TX_m, we can obtain

$$c_1^m = y_m - kx_m \tag{16}$$

$$\theta = \arctan k \tag{17}$$

h_m can be expressed as

$$h_m = |y - y_2^m|\cos\theta = |b - c_1^m|\cos\theta \tag{18}$$

According to [11], the following equation is hold

$$0.25c^2{t_d}^2(m,n) = d^2\varepsilon - d^2\frac{(x_D - x_m)^2}{(x_D - x_m)^2 + 4{h_m}^2} \tag{19}$$

As shown in Fig. 1, the ratio of *h* and distance from RX to y_2 is defined as γ, which is given by

$$\gamma = \frac{h}{|RX - C|} \tag{20}$$

Since *TX'* is the mirror point of *TX*, it can be deduced that

$$|TX - TX'| = 2h = 2\gamma|RX - C| \tag{21}$$

The triangle *TX' - TX - Q* is similar to the triangle *Q - C - RX*, and the same as the triangle *TX - Q - W* to the triangle *TX - C - RX*. Based on the similarity of triangles, the relationships among their sides can be derived as follows:

$$\frac{|TX - TX'|}{|RX - C|} = \frac{|TX - Q|}{|Q - C|} \tag{22}$$

$$\frac{|TX - Q|}{|TX - C|} = \frac{|Q - W|}{|RX - C|} \tag{23}$$

Subsequently, the length of *QW* can be determined as

$$l_{QW} = \frac{2\gamma}{2\gamma + 1}|RX - y| \tag{24}$$

The coordinate on the Y-axis of point Q is

$$y_Q = y_m + l_{QW}\cos\theta = y_m + l_{QW}\cdot\arctan k \tag{25}$$

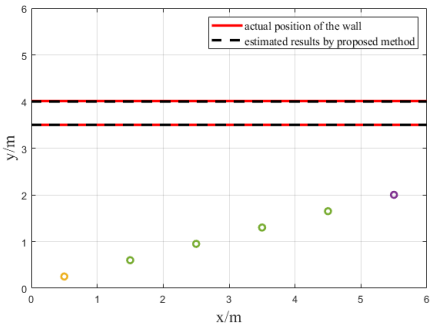


Fig. 3. Estimation results for wall position

Substituting (24)(25) into $y_2 = kx + c_1$, the coordinates of point Q can be determined as (x_Q, y_Q). Plugging these coordinates into (19) and integrating parameter values from each channel, we obtain

$$\mathbf{A} = \begin{bmatrix} 1 & -\frac{\left(x_Q^1 - x_1\right)^2}{\left(y_Q^1 - y_1\right)^2 + 4{h_1}^2} \\ 1 & -\frac{\left(x_Q^2 - x_1\right)^2}{\left(y_Q^2 - y_1\right)^2 + 4{h_1}^2} \\ \vdots & \vdots \\ 1 & -\frac{\left(x_Q^n - x_m\right)^2}{\left(y_Q^n - y_m\right)^2 + 4{h_m}^2} \end{bmatrix}_{MN\times 2} \tag{26}$$

$$\mathbf{p} = \begin{bmatrix} d^2\varepsilon \\ d^2 \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \tag{27}$$

where (x_Q^n, y_Q^n) $(n = 1, 2, \ldots, N)$ represents the position of the equivalent virtual receiving antenna for the *n*-th receiving antenna corresponding to each transmitting antenna.

In that case

$$\mathbf{Ap} - \mathbf{b} = \mathbf{0} \tag{28}$$

Therefore

$$\mathbf{p} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b} \tag{29}$$

$$\mathbf{b} = \begin{bmatrix} 0.25c^2{t_d}^2(1,1) \\ 0.25c^2{t_d}^2(1,2) \\ \vdots \\ 0.25c^2{t_d}^2(m,n) \end{bmatrix}_{MN\times 1} \tag{30}$$

$$\hat{d} = \sqrt{p_2} \tag{31}$$

$$\hat{\varepsilon} = p_1/p_2 \tag{32}$$

By minimizing the cost equation, we obtain d and ε.

$$f(d,\varepsilon) = |\mathbf{Ap} - \mathbf{b}|^2 \tag{33}$$

IV. Simulation Results and Analysis

To validate the effectiveness of the proposed algorithm, we deploy a UWB MIMO radar with a wall parallel to the ground, which has a relative permittivity of 6 and a thickness of 0.5m. The slope and intercept of the wall are 0 and 3.5, respectively. Using a linear array with two transmitters and

TABLE I
ESTIMATION RESULTS UNDER DIFFERENT SNRS

SNR/db	proposed algorithm		algorithm in [12]	
	d/m	ε	d/m	ε
30	0.515	6.05	0.523	5.863
25	0.517	6.06	0.523	5.86
20	0.52	6.08	0.527	5.793
15	0.523	5.89	0.53	5.6
10	0.53	5.84	0.533	5.6

four receivers, positioning in a straight inclined line. The transmitting antennas are located at coordinates (0.5m, 0.25m) and (5.5m, 2m). The simulation utilizes GprMax [9] to generate the echoes, where the ricker pulse with 1.5 GHz center frequency is set as the transmitting signal. In the estimation procedure, the searching space of slope and intercept are set to $k \in [-0.5, 0.5]$, $b \in [0, 5]$. After the first stage simulation, we can obtain the estimated results are $k = 0.0087$ and $b = 3.5$.

Next, using k and b as prior conditions to estimate thickness and relative permittivity of the wall. The search ranges are set to $d \in [0, 1]$, $\varepsilon \in [5.5, 8.5]$. The gaussian white noise is added to achieve a signal-to-noise ratio (SNR) of 30 dB. After the second stage stimulation, we can get the estimated wall thickness is 0.515m, and the relative permittivity is 6.05. The simulation results are depicted in Fig. 3, where the black line represents the estimated wall position.

In order to assess the algorithm's performance under different SNRs, we choose to quantitatively evaluate the estimated results using the error rate.

$$error = \frac{|\alpha_{real} - \alpha_{est}|}{\alpha_{real}} \times 100\% \tag{34}$$

where α_{real} is the true value of the parameter, and α_{est} is the estimated value of the parameter.

The error rates of thickness and relative permittivity for the proposed method compared with the method proposed in [12] at different SNRs are shown in Fig. 4(a) and Fig. 4(b), respectively. It's obvious that the method presented in this paper exhibits lower error rates under different SNRs. Table I presents the estimated parameters using both the method from reference [12] and the approach proposed in this paper under various SNRs. At a SNR of 10 dB, the estimation errors for wall thickness and relative permittivity are 0.03m and 0.16, respectively, which demonstrate the superiority of the algorithm.

V. CONCLUSION

This paper introduces a two-stage algorithm to estimate parameters of inclined walls. Initially, it employs mirror reflection for approximate calculations to find time delays between the front surface of the wall and potential parameters. Then, by minimizing a cost function, it accurately estimates the slope and intercept of the front wall, thus establishing the linear equation of the front wall. Using this equation, the linear array is transformed into a virtual array parallel to the wall for each channel. After combining all the channels, it enabls the extraction of unknown wall thickness and relative permittivity.

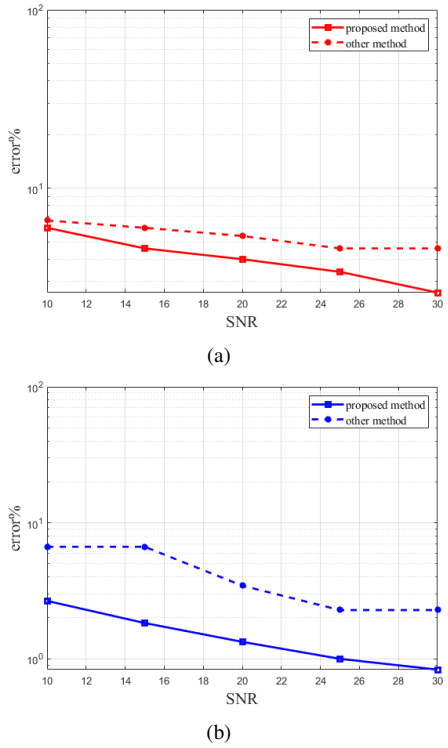


Fig. 4. Error rates of wall parameters for two methods at different SNRs (a) Wall thickness error rate (b) Relative permittivity error rate.

The method is compared with other approaches, and simulation results validate its robust performance.

VI. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62371110, 62401098, in part by Postdoctoral Innovation Talent Support Program under Grant BX20230056, and in part by the Municipal Government of Quzhou under Grant 2023D032.

REFERENCES

[1] F. Ahmad, M. Amin and G. Mandapati, " Autofocusing of Through-the-Wall Radar Imagery Under Unknown Wall Characteristics," *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1785-1795, July 2007.

[2] J. Liu, L. Kong, X. Yang and Q. Liu, "Refraction Angle Approximation Algorithm for Wall Compensation in TWRI," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 7, pp. 943-946, July 2016.

[3] K. Browne, R. Burkholder and J. Volakis, "Fast Optimization of Through-Wall Radar Images Via the Method of Lagrange Multipliers," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 1, pp. 320-328, Jan. 2013.

[4] C. Thajudeen and A. Hoorfar, "A Hybrid Bistatic–Monostatic Radar Technique for Calibration-Free Estimation of Lossy Wall Parameters," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 1249-1252, 2017.

[5] L. Qu, Z. Fang, T. Yang, et al, "Sparse Recovery Method for Estimation of Wall Parameters in Through-the-Wall Radar," *2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019, pp. 3558-3561.

[6] G. Wang and M. Amin, "A new approach for target location of through the wall radar imaging in the presence of wall ambiguities," *Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology*, Rome, Italy, 2004, pp. 183-186.

[7] P. Protiva, Pavel, Mrkvica, et al, “Estimation of Wall Parameters From Time-Delay-Only Through-Wall Radar Measurements,” *IEEE Transactions on Antennas and Propagation*, vol. 59, no. 11, pp. 4268-4278, Nov. 2011.

[8] X. Jun, B. L, D. Yang, et al, “Review of Wall Parameter Estimation and Compensation Methods of TWR,” *Radio Engineering*, vol. 52, pp. 1003-3106, 2022.

[9] C. Warren, A. Giannopoulos and I. Giannakis, “An advanced gpr modelling framework: The next generation of gprmax,” in *Int. Workshop Adv. Ground Penet. Radar*, Florence, Italy, Jul. 2015, pp. 1–4.

[10] S. Zhou, H. Zhang, Y. Zhang, et al, “A Robust Estimation Method of the Wall Parameters Based on Machine Learning,” *2019 International Applied Computational Electromagnetics Society Symposium - China*, Nanjing, China, 2019, pp. 1-2.

[11] T. Jin, Z. Lu, B. Chen, et al, “Estimation of wall parameters for cognitive imaging in through-the-wall radar,” *2012 IEEE 11th International Conference on Signal Processing*, Beijing, China, 2012, pp. 1936-1939.

[12] Y. Song, T. Jin, B. Lu, et al, “A Novel Linear MIMO Array Through-the-wall Imaging Model and Its Associated Environmental Parameters Estimation,” *Journal of Electronics and Information Technology*, vol. 36, pp. 2980-2985, 2014.

[13] J. Chen, Y. Zhang, Y. Wang, X, et al, “A Two-Stage Wall Parameter Estimation Algorithm for UWB Through-the-wall Radar,” *2021 CIE International Conference on Radar*, Haikou, Hainan, China, 2021, pp. 3041-3044.

[14] Y. Jia, L. Kong, X. Yang and K. Wang, “A new approach for wall parameters estimation in through-the-wall radar imaging,” *2011 IEEE RadarCon* , Kansas City, MO, USA, 2011, pp. 095-098.

Bluemarble: Bridging Latent Uncertainty in Articulatory-to-Speech Synthesis with a Learned Codebook

Seyun Um*, Miseul Kim*, Doyeon Kim*, Hong-Goo Kang*
* Department of Electrical and Electronic Engineering , Yonsei University, Seoul, Korea
E-mail: syum, miseul4345, ehyeon24@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

***Abstract*—In this paper, we present Bluemarble, a novel neural articulation-to-speech (ATS) system designed to generate high-quality speech from articulatory recordings. Traditional ATS methods often demand a substantial dataset of corresponding articulatory and speech signal pairs for effective model training. This requirement arises because these approaches treat the estimation of acoustic features as a regression task, which can pose difficulties for accurately mapping latent articulatory features to the desired speech-related features. To tackle this challenge, we propose a method that involves employing a finite learned codebook to constrain the uncertainty latent space. Operating within a framework consisting of a speech mel-spectrogram encoder, quantizer, and decoder, our model establishes a mapping rule between the latent embeddings derived from electromagnetic articulography (EMA) signals and quantized discrete speech tokens. During the inference stage, EMA embeddings are first transformed into speech-related discrete tokens, which are then input into a neural vocoder to synthesize the speech waveform. Experimental results show that our approach outperforms existing state-of-the-art methods in both qualitative and quantitative assessments. Audio samples are available online.**[1]

I. Introduction

Articulation-to-speech (ATS), the task of utilizing articulatory movements in speech synthesis, is a field of research that holds significant importance with numerous practical applications [1]. It provides essential communication tools for individuals facing challenges in producing audible speech freely, such as those who have undergone laryngectomy. In addition, ATS systems can play a crucial role in establishing a connection between neural activity in the brain and speech signals, making it a valuable intermediate step in the development of brain-to-speech systems [2]–[4].

Previous works have used a wide variety of approaches for modeling ATS systems. Gaussian mixture model (GMM)-based ATS synthesizers have been designed to estimate the vocal tract spectrum from articulatory movements [5]. Hidden Markov model (HMM)-based methods have also been adopted for this task, combined with a codebook-based network [6]. More recently, the emergence of deep neural models has led to significant improvements in the speech synthesis quality of ATS systems compared to traditional statistical-based approaches [7]–[11]. For example, ATS models often incorporate fully connected layers along with uni- and bi-directional LSTM-based networks [7], [12]. The inclusion of attention-based transformer modules [13] have further contributed to advances in the field; in [8], transformer blocks are employed to estimate mel-frequency cepstral coefficients (MFCCs) from articulatory features. Based on the CARGAN vocoder [14], [9] directly predicts waveforms from EMA signals without intermediate acoustic features using an adversarial training criterion. Text-to-speech (TTS)-based methods have been adopted [10] to extract pitch and energy information from articulatory recordings and generate high-quality mel-spectrograms. Voice conversion-based schemes have also been utilized to synthesize speech from articulated content using the voices of different speakers for vocalization [11].

Although the aforementioned frameworks have demonstrated impressive speech synthesis capabilities, they have typically required a substantial number of parallel articulation and speech recording pairs for model training. Furthermore, they often treat the ATS task as a regression problem, minimizing the mean-squared error (MSE) between estimated and target acoustic features. Mean-squared error primarily aims to minimize the energy difference between two features, resulting in a lack of fine-grained feature estimation (coarse granularity) in complex latent domains [15]–[17]. This can make it challenging to precisely reconstruct intricate target acoustic features from the complex latent embeddings.

In this paper, we propose a neural network-based ATS system designed to decode high-quality covert speech from articulatory kinematics by adopting acoustic tokens extracted from pretrained codebooks. Our overall system utilizes two network training flows: one for mel-spectrogram codebook training and another for articulation-to-acoustic token modeling. Initially, acoustic tokens are estimated from a neural speech compression model through adversarial training [18], utilizing mel-spectrograms. Upon the completion of mel-spectrogram codebook training, an electromagnetic articulograph (EMA) encoder is trained to predict the corresponding acoustic tokens from articulatory recordings, estimating the appropriate codebook indices using cross-entropy loss. The mel-spectrograms are then reconstructed using the estimated acoustic tokens. In the final step, raw speech waveforms are generated by passing the synthesized mel-spectrograms through a pretrained vocoder

[1] https://sam-0927.github.io/Bluemarble/

(e.g. HiFi-GAN [19]). The key advantages of our framework can be summarized as follows:

- **Discretized prediction task:** Thanks to the simpler prediction task involving discrete, quantized acoustic tokens rather than regression in a continuous latent space, our model can more effectively generate speech from articulatory movements.
- **Less paired data required:** Our proposed articulation-to-acoustic token modeling framework can be trained using a significantly smaller amount of paired recordings between articulatory and speech signals compared to previous methods.

II. Related work

A. Style modeling for multi-speaker ATS systems

The most closely related work to ours is [10], which addresses the task of estimating ground-truth mel-spectrograms from EMA signals [20]. Our baseline approach is centered around the reconstruction of mel-spectrograms in a multi-speaker setting, simultaneously estimating both the target speaker's speaking style and contextual information from the provided EMA signals. The entire framework is built using modules that are integrated using convolutional neural network (CNN) and self-attention-based layers. It then utilizes the HiFi-GAN neural vocoder [19] to reconstruct raw speech waveforms from mel-spectrograms.

B. Acoustic feature tokenization in generative models

There has recently been an increasing focus on utilizing discrete acoustic tokens as speech representations for generative speech modeling [21], [22]. The key idea is to predict a sequence of tokens from a codebook, which can then be decoded into more concrete speech representations (e.g. raw audio). High-quality speech can be generated by splitting the task into multiple stages that model semantic and acoustic details successively. This approach simplifies the generation task due to the inherent constraints on the codebook's size, and allows models to be trained using abundant unlabeled audio-only data. AudioLM [23], WavLM [24], CLaM-TTS [25], and LM-VC [26] are key examples of this modeling approach. SPEAR-TTS [27] is an extension to this approach that learns a mapping between text and acoustic tokens to perform TTS. This work leverages the benefits of the aforementioned token-based acoustic modeling in the development of an articulation-to-speech framework.

III. Proposed model

In this section, we provide a comprehensive overview of our proposed network, Bluemarble. Our model employs a two-step learning process to predict mel-spectrograms from the EMA signals, shown in Figure 1. First, we train an autoencoder-based neural speech compression model to learn a codebook of acoustic tokens. These tokens are used to encode a mel-spectrogram into a set of discrete representations. The autoencoder's decoder is trained to reconstruct the mel-spectrogram given a set of codebook indices. Second, we train an EMA encoder to predict discrete acoustic tokens in the codebook from articulatory signals while freezing the codebook and mel-spectrogram decoder. During inference, we input the acoustic tokens predicted by the EMA encoder into the pretrained decoder to generate the target mel-spectrogram. Then, we synthesize a speech waveform from the mel-spectrogram using a pretrained HiFi-GAN vocoder [19].

A. Mel-spectrogram codebook training

Since speech signals contain complex and intertwined information, including content and speaker characteristics, it can be challenging to directly predict these intricate continuous spaces from articulatory signals. We alleviate this problem by changing the prediction domain from a continuous latent space to a discrete latent space, bridging the two distinct domains with reduced uncertainty.

In our proposed model, we encode the mel-spectrograms $S_{1:T} = s_1, ..., s_T$, where $s \in \mathbb{R}^{1\times F_s}$, of a target speech signal to create a discrete codebook. The model has an autoencoder structure consisting of encoder $\mathbf{E_{Mel}}$, codebook $\mathbf{Q}$, and decoder $\mathbf{D_{Mel}}$ [28]. Both the encoder and decoder that down-samples and up-samples the acoustic features [29] are equipped with an attention module and residual blocks. The encoder in our model is comprised of five residual blocks, each consisting of a 2D convolutional layer with dropout and group normalization, an attention layer, and five additional 2D convolutional layers. The decoder is composed of seven residual blocks, an attention layer, and five 2D convolutional layers. The encoder estimates hidden features, denoted as $\hat{Z} \in \mathbb{R}^{F'\times T'\times d}$, from the mel-spectrogram and subsequently converts them into discrete tokens Z_q through a codebook, where d represents the dimension of the codebook entries, and $F' \times T'$ corresponds to a reduced frequency and time dimension. Meanwhile, the decoder predicts mel-spectrograms based on the quantized representations.

$$\begin{aligned} \hat{Z} = \mathbf{E_{Mel}}(S_{1:T}), \quad Z_q = \mathbf{Q}(\hat{Z}), \\ \hat{S}_{1:T} = \mathbf{D_{Mel}}(Z_q). \end{aligned} \tag{1}$$

We train our model using the LSGAN [30] method to enhance the intelligibility of the reconstructed output. The discriminator is built with a 2D convolutional layer featuring skip connections and the Exponential Linear Unit (ELU) activation function. The overall loss for training the autoencoder is defined as follows:

$$\mathcal{L}^{AE}_{total} = \mathcal{L}_G + \mathcal{L}_D + \mathcal{L}_{VQ}, \tag{2}$$

where $\mathcal{L}_G$ and $\mathcal{L}_D$ are generator and discriminator losses and the VQ loss is defined as:

$$\mathcal{L}_{VQ} = \lambda||\hat{Z} - sg(Z_q)||^2_2 + ||sg(\hat{Z}) - Z_q||^2_2, \tag{3}$$

where λ is 0.25. The sg means stop gradient.

A crucial consideration in the process of codebook design is how to determine the number of codebook entries. As the codebook size expands, the task of accurately predicting acoustic tokens from EMA signals becomes progressively

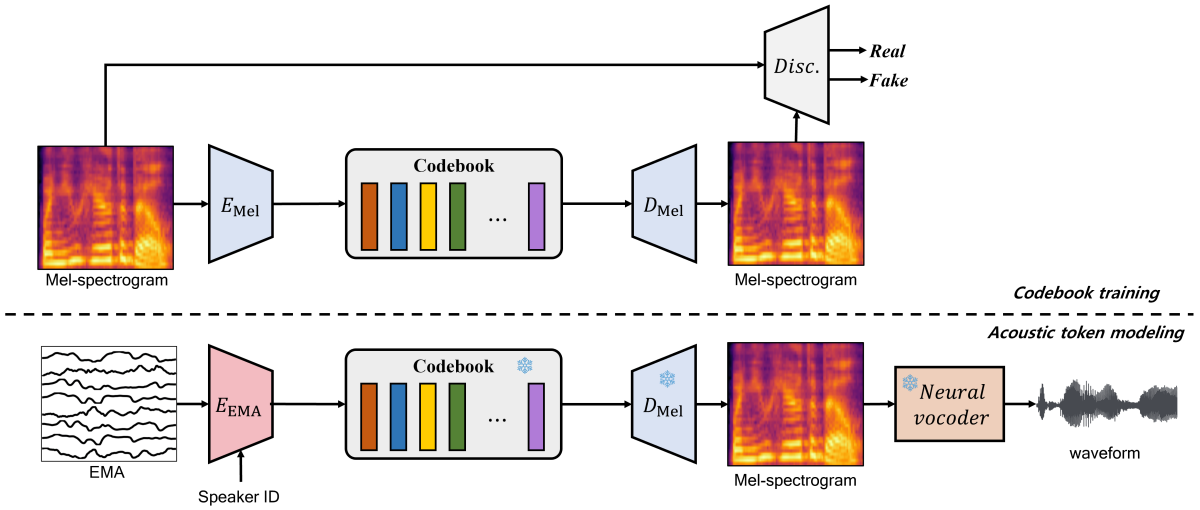


Fig. 1: Architecture of Bluemarble. During codebook training, the target mel-spectrogram is fed into the mel encoder, and it is converted to discrete acoustic tokens by utilizing vector quantization (VQ) codebooks. These tokens are subsequently decoded back into the original mel-spectrogram. The codebook training is carried out concurrently with the reconstruction of the input mel-spectrogram. During acoustic token modeling, the electromagnetic articulography (EMA) encoder utilizes speaker embeddings to predict the target indices of acoustic tokens in the pretrained codebook. The snow symbol means that the module's weights are frozen.

more challenging. Conversely, when the codebook size is too small, the decoder may encounter difficulties in learning how to reconstruct high-quality acoustic features. Taking these factors into account, we chose a codebook size of 32. We conducted an ablation experiment to assess how variations in codebook size affect the model's performance in Section IV.

B. Articulatory to acoustic token modeling

After the codebook has been trained, we proceed to freeze it, as well as the decoder. We then train an encoder that takes in EMA signals and predicts the acoustic tokens for the corresponding target acoustic feature. Constructing a codebook as we do above enables us to map EMA signals to discrete acoustic tokens instead of directly predicting acoustic features.

To predict acoustic tokens, we employ a 2D convolutional module with residual connections and group normalization, maintaining a structure identical to that of the mel encoder. We also apply a 1D convolution to the EMA signal input $A_{1:T} = \{a_1, a_2, ..., a_T\}$, where $a \in \mathbb{R}^{1\times F_a}$, with the aim of increasing the channel size to align it with the target mel-spectrogram. Additionally, we incorporate speaker information into the model because the characteristics of articulatory signals differ from one person to another. The EMA encoder processes the input to predict the acoustic token index within the pretrained codebook, then the decoder estimates the acoustic features.

$$\hat{Z}' = \mathbf{E_{EMA}}(EMA_{1:T} + e_{spk}), \quad e_{spk} = \mathbf{H}(i), \tag{4}$$

where e_{spk} denotes speaker embedding created by the fully connected layer $\mathbf{H}$. We train the EMA encoder using cross-entropy loss between predicted and target acoustic token indices.

$$\mathcal{L}_{CE} = -\frac{1}{L}\sum_{i=1}^{L} y_i \log(\mathbf{F}(\hat{Z}')_i), \tag{5}$$

where $\mathbf{F}$ denotes the classifier consisting of a fully connected layer and softmax. y and L denote the target and codebook size, respectively.

IV. Experimental Setup

A. Dataset

We utilize the publicly available Haskins dataset **haskins**, which consists of paired speech and EMA signals collected from 4 male and 4 female speakers. Each speaker's data covers approximately 55 minutes of recordings, with the EMA signals recorded at a sampling rate of 100 Hz and the speech signals at 44.1 kHz. We excluded silence regions and applied a pretrained denoising model [31] among speech enhancement models [32]–[35] to eliminate background noise from all speech samples. For computational convenience, we downsampled the speech samples from 44.1 kHz to 16 kHz. For further processing, the speech data was transformed into 40-dimensional mel-spectrograms using Short-Time Fourier Transforms (STFTs) at intervals of 10 ms, using a 25 ms window. The EMA data is represented as an 18-dimensional vector, where each dimension corresponds to one of the 6 sensors (TR, TB, TT, UL, LL, jaw), and each sensor comprises 3 trajectory orientations (X: posterior to anterior, Y: right to left, Z: inferior to superior).

In Table I, both our model and baseline models undergo training with a paired dataset spanning 5 hours. In contrast, we train the codebook and decoder using a speech-only dataset lasting 5 hours, while the EMA encoder is trained using paired datasets of 5 and 1 hour as detailed in Section V-C.

TABLE I: Objective test results.

Model	Objective	
	NISQA-T↑	NISQA-E↑
Reference	2.63±0.02	3.86±0.03
[8] (ACL 21')	2.08±0.02	2.9±0.03
[9] (INTERSPEECH 22')	1.84±0.02	2.22±0.02
[10] (ICASSP 23')	2.29±0.02	3.02±0.03
Bluemarble (Ours)	**2.66±0.02**	**4.22±0.02**

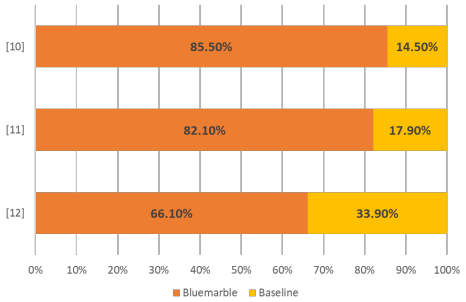


Fig. 2: A/B preference test results.

B. Training

We performed training on a single NVIDIA RTX 3090 GPU for a total of 150 epochs. We employed the Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.9$, and $\epsilon = 1e-8$) for training both the EMA encoder and EMA discriminator, setting their respective learning rates to $1e-4$ and $1e-5$. We utilized a HiFi-GAN neural vocoder pre-trained on the VCTK dataset [36].

V. Experimental Results

A. Objective measurement

To evaluate the clarity and overall quality of synthesized speech, we use speech quality assessment metrics, NISQA-TTS (NISQA-T) [37] and NISQA-Enhance (NISQA-E) [38]. NISQA-TTS is relevant to evaluating the naturalness of the synthesized speech, while NISQA-Enhance focuses on measuring fundamental audio clarity and quality. As shown in Table I, our proposed model shows superior performance compared to all baseline models. Notably, Bluemarble achieves NISQA-TTS scores of 2.66 vs. 2.29 and NISQA-Enhance scores of 4.22 vs. 3.02 compared to the best baseline [10], demonstrating its capabilities for generating high quality speech that is both natural and intelligible. Furthermore, these results show that our framework of predicting discrete acoustic tokens from a codebook, rather than directly estimating acoustic features, significantly enhances ATS performance.

B. Subjective measurement

To assess generated sound quality as perceived by human listeners, we conducted an A/B preference listening test between each baseline and our proposed model. 14 listeners participated in the test. After listening to pairs of speech samples, consisting of one from Bluemarble and one from a baseline, participants were asked to choose the preferred speech sample based on the quality of sound. Results are shown in Figure 2. For all comparisons, participants showed much higher preference for samples synthesized by Bluemarble, demonstrating its superiority in terms of subjective evaluation in addition to objective metrics.

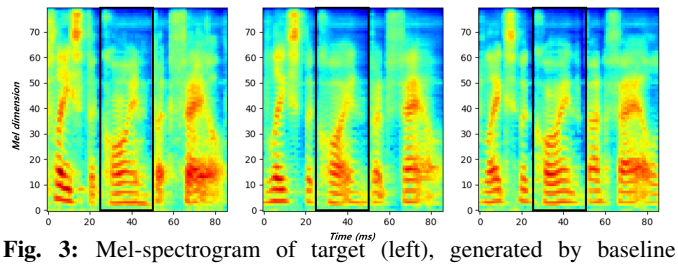


Fig. 3: Mel-spectrogram of target (left), generated by baseline (middle) and Bluemarble (right). The Bluemarble's sample has better quality than that of the baseline. The distinct pitch is especially evident in the region delineated by a black box. To facilitate a detailed comparison, we generate a mel-spectrogram of synthesized speech using an 80-dimensional resolution from a pre-trained vocoder, as opposed to the 40-dimensional output from the decoder.

TABLE II: Performance evaluation in terms of a codebook size and training dataset size.

Codebook	Data	NISQA-T↑	NISQA-E↑
32	5h	**2.66±0.02**	**4.22±0.02**
64	5h	2.59±0.02	4.10±0.02
16	5h	2.63±0.02	4.15±0.02
32	1h	**2.55±0.02**	**4.17±0.03**
64	1h	2.52±0.02	3.77±0.03
16	1h	**2.55±0.02**	4.08±0.03

C. Ablation study

To assess the influence of codebook and training dataset size on Bluemarble's performance, we conducted model training by varying codebook and training dataset size.

Codebook size. As shown in Table II, compared to the standard model with codebook size 32, model performance decreases when adopting smaller (16) or larger (64) codebook sizes. A larger codebook size allows for a more comprehensive representation of the signal, facilitating the reconstruction of acoustic features during modeling. However, it also means that the EMA encoder must learn to predict a larger number of tokens, which is more difficult. We conjecture that this caused a reduction in the quality of synthesized speech due to the more challenging token prediction task. Conversely, when the codebook size is too small, the acoustic tokens fail to adequately encode all of the information in the acoustic features, leading to degradation in reconstructing them. These findings highlight the importance of selecting an appropriate codebook size that strikes a balance between encoding sufficient information to reconstruct the mel spectrogram features and allowing the model to effectively predict acoustic tokens for high-quality speech synthesis.

Dataset size. Table II shows the NISQA-T and NISQA-E scores for Bluemarble trained on 1 hour vs. 5 hours of paired EMA and speech data. We can see that our model still demonstrates reasonable performance even when trained with only 1 hour of data, achieving 2.55 and 4.17 NISQA scores. This demonstrates that decoupling the training of the codebook and mel-spectrogram autoencoder with that of the

EMA encoder allows for high-quality speech synthesis even with only a small amount of paired data.

VI. Conclusion

In this paper, we proposed Bluemarble, an articulation-to-speech (ATS) system that uses electromagnetic articulography (EMA) signals. Utilizing concepts from recent trends in generative audio modeling, our model predicts a sequence of discrete acoustic tokens in a codebook rather than predicting continuous acoustic features. Specifically, it learns a mapping from the EMA signals to discrete latent acoustic embeddings represented by vector quantization codebooks, which are then decoded to produce mel-spectrograms. Experiments demonstrate that our model surpasses previous state-of-the-art models in terms of the clarity and naturalness of synthesized speech, as confirmed by both objective and subjective evaluation metrics.

VII. Acknowledge

This work was supported by the 'Alchemist Project' (Fully implantable closed loop Brain to X for voice communication) funded By the Ministry of Trade, industry & Energy (MOTIE, Korea), under Grant 20012355 and NTIS 1415181023.

References

[1] B. Cao, M. J. Kim, J. R. Wang, J. P. van Santen, T. Mau, and J. Wang, "Articulation-to-speech synthesis using articulatory flesh point sensors' orientation information.," in *INTERSPEECH*, 2018, pp. 3152–3156.

[2] P. Li, S. Cai, E. Su, and L. Xie, "A biologically inspired attention network for eeg-based auditory attention detection," in *IEEE Signal Processing Letters*, 2022, pp. 284–288.

[3] M. Angrick, C. Herff, E. Mugler, *et al.*, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *Journal of neural engineering*, vol. 16, p. 036 019, 2019.

[4] M. Kim, Z. Piao, J. Lee, and H.-G. Kang, "Braintalker: Low-resource brain-to-speech synthesis with transfer learning using wav2vec 2.0," in *BHI*, 2023, pp. 1–5.

[5] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *ISCA Workshop on Speech Synthesis*, 2004, pp. 31–36.

[6] S. Hiroya and M. Honda, "Determination of articulatory movements from speech acoustics using an hmm-based speech production model," in *ICASSP*, 2002, pp. I-437-I–440.

[7] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech & Language*, vol. 36, pp. 260–273, 2016.

[8] D. Gaddy and D. Klein, "An improved model for voicing silent speech," in *ACL/IJCNLP*, 2021, pp. 175–181.

[9] P. Wu, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Deep speech synthesis from articulatory representations," in *INTERSPEECH*, 2022, pp. 779–783.

[10] M. Kim, Z. Piao, J. Lee, and H.-G. Kang, "Style modeling for multi-speaker articulation-to-speech," in *ICASSP*, 2023, pp. 1–5.

[11] K. Scheck and T. Schultz, "Multi-speaker speech synthesis from electromyographic signals by soft speech unit prediction," in *ICASSP*, 2023, pp. 1–5.

[12] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, vol. 385, pp. 37–45, 2012.

[13] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Proc. NIPS*, vol. 30, 2017.

[14] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive gan for conditional waveform synthesis," in *ICLR*, 2021.

[15] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *ICLR*, 2015.

[16] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *ICLR*, 2016.

[17] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in atari games," *NIPS*, vol. 28, 2015.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, pp. 139–144, 2020.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *NIPS*, vol. 33, pp. 17 022–17 033, 2020.

[20] T. Rebernik, J. Jacobi, R. Jonkers, A. Noiray, and M. Wieling, "A review of data collection practices using electromagnetic articulography," *Laboratory Phonology*, vol. 12, p. 6, 2021.

[21] K. Takagi, T. Akiba, and H. Tsukada, "Semi-supervised asr based on iterative joint training with discrete speech synthesis," in *APSIPA ASC*, 2022, pp. 923–928.

[22] C. Yuan and Y.-C. Huang, "Personalized end-to-end mandarin speech synthesis using small-sized corpus," in *APSIPA ASC*, 2020, pp. 837–840.

[23] Z. Borsos, R. Marinier, D. Vincent, *et al.*, "Audiolm: A language modeling approach to audio generation," *ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2023.

[24] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2022.

[25] Anonymous, "CLam-TTS: Improving neural codec language model for zero-shot text-to-speech," in *ICLR*, 2023.

[26] Z. Wang, Y. Chen, L. Xie, Q. Tian, and Y. Wang, "Lm-vc: Zero-shot voice conversion via speech generation based on language models," in *IEEE Signal Processing Letters*, 2023, pp. 1157–1161.

[27] E. Kharitonov, D. Vincent, Z. Borsos, *et al.*, "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1703–1718, 2023.

[28] V. Iashin and E. Rahtu, "Taming visually guided sound generation," in *BMVC*, 2021.

[29] J. Kim and H.-G. Kang, "Contrastive learning based deep latent masking for music source separation," in *INTERSPEECH 2023*, 2023, pp. 3709–3713.

[30] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[31] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *INTERSPEECH*, 2020, pp. 3291–3295.

[32] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *APSIPA ASC*, 2017, pp. 006–012.

[33] S.-W. Fu, C. Yu, T.-A. Hsieh, *et al.*, "Metricgan+: An improved version of metricgan for speech enhancement," in *Interspeech 2021*, 2021, pp. 201–205.

[34] D. Kim, S.-W. Chung, H. Han, Y. Ji, and H.-G. Kang, "Hd-demucs: General speech restoration with heterogeneous decoders," in *INTERSPEECH 2023*, 2023, pp. 3829–3833.

[35] D. Kim, H. Han, H.-K. Shin, S.-W. Chung, and H.-G. Kang, "Phase continuity: Learning derivatives of phase spectrum for speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6942–6946.

[36] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *The Centre for Speech Technology Research*, vol. 6, p. 15, 2017.

[37] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," in *INTERSPEECH*, 2020, pp. 1748–1752.

[38] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *INTERSPEECH*, 2021, pp. 2127–2131.

MTFNet: Multi-Scale Transformer Framework for Robust Emotion Monitoring in Group Learning Settings

Yi Zhang[1†], FangYuan Liu[1†], JiaJia Song[1], Qi Zeng[2] and Hui He[1*]
1 Beijing Normal University, Zhuhai, ZhuHai, China
2 Beijing Normal University, Beijing, China

E-mail: hhdpc@outlook.com Tel: +86-17328420173

***Abstract*—Identifying students' learning states in authentic classroom settings is a prominent topic in educational technology. This study addresses the challenges posed by complex facial environments and the scarcity of data in such settings. we propose a Multi-Scale Transformer with Frame Shuffled Order Predict Network (MTFNet), based on a spatial-temporal feature extraction structure, to perform effective learning-related facial expression recognition in a primary school classroom. Specifically, we combine a Multi-Scale Facial Feature Fusion Module (MFFF) based on Grouped Spatial Convolution(GS Conv) to effectively capture multi-level facial features and improve the model's robustness in complex environments. Additionally, a Frame-wise Shuffle Order Prediction Module (FSOP) is introduced to enhance the model's ability to understand the dynamic changes of emotional intensity by predicting the emotion expression sequence. Experiments on both the DFEW dataset and our dataset demonstrate excellent performance and generalization in real-world applications.**

I. INTRODUCTION

Facial expressions are the most powerful and direct means for humans to convey emotions. In recent years, facial expression recognition (FER) has garnered significant research interest due to its potential applications in many fields [1][2]. For instance, students' facial expressions can reflect their feelings about the course content and indicate their engagement in group discussion, which allows teachers to verify whether students understand the material effectively and promptly, enabling timely adjustments to the course schedule [3][4]. However, facial expression detection and recognition in real-world settings faces challenges such as occlusion and low resolution, limiting the effectiveness of artificial intelligence in practical applications. To address this, some studies focus on optimizing model architecture or improving data quality to enhance inference efficiency and accuracy [5][6].

From the perspective of the model, it is worth noting that the presentation of facial expressions is often a dynamic process, so relying solely on a static single image as input will confuse the model on similar expressions. Compared with traditional hand-crafted feature extractions[7][8], deep learning methods have got more excellent grades in a large number of experiments. Specifically, CNN-based models such as ResNet [9] can easily extract spatial features from each frame while RNN-based models such as LSTM [10] have the ability to find temporal dependencies between frames. Combining the two methods is an excellent way to effectively model the dynamic facial expression process. However, a more effective approach today involves using transformer-based models, which consistently achieve state-of-the-art performance due to their robust global information perception capability [5][11][12]. These models can capture crucial information in the input without depending on sequence order. Therefore, it is possible to train the transformer-based model by shuffling the order of frames.

From the perspective of the dataset, while previous studies have demonstrated the potential benefits of deep learning models in complex situations [13], there is still insufficient evidence to support their effectiveness in classroom contexts. In fact, the large-scale classrooms in China create an urgent demand for a model that can track students' learning-related emotions during class, assisting teachers in better monitoring each student's emotional engagement. Although in the field of education, the types of emotions generated by students when they are learning are very complicated (e.g. bored, pride, and confused)[14], most of the literature on affective computing in education is still based on Ekman's discrete emotion theory, which is a general and convenient way to categorize emotions but does not always provide valuable and practical insights for users. Therefore, it is essential to develop a learning emotion dataset aligned with learning engagement theory, grounded in both theoretical and empirical research in the field of education.

All in all, the main contributions of this paper can be summarized as follows:

- Proposing a MFFF module based on GS convolution to

†: Equal contribution.
*: Corresponding author.

effectively capture features from video clips with occlusions and uneven lighting, which improves the performance of the model in practical applications.

- Adapting a FSOP module to guide the model to focus on changes between frames, thereby enhancing the transformer's ability to model effect on dynamic changes in facial expressions.
- Providing a small-scale video dataset, which records student's performance in a primary school classroom and includes four discrete learning-related emotions: interesting, happy, bored and confused.
- Identifying that proposed MTFNet achieves state-of-the-art performance on the evaluated datasets (i.e., DFEW and our datasets). We will release our source code upon acceptance.

II. Related Work

In this section, we review some related works of our proposed method for facial expression recognition in educational scenarios.

A. Emotion recognition based on facial expressions

In the 1970s, Ekman and Friesen found that the way humans perceive certain emotions is the same, regardless of cultural and racial differences, and defined six basic emotions: happiness, anger, disgust, fear, sadness and surprise[15]. Subsequently, in 1978, Ekman and colleagues further refined their researches on facial expressions and proposed the Facial Action Coding System (FACS) based on modular modeling of facial expression recognition muscles [16]. This system has a broader and more diverse range of expression portrayals, which has greatly promoted the research progress of facial expression recognition.

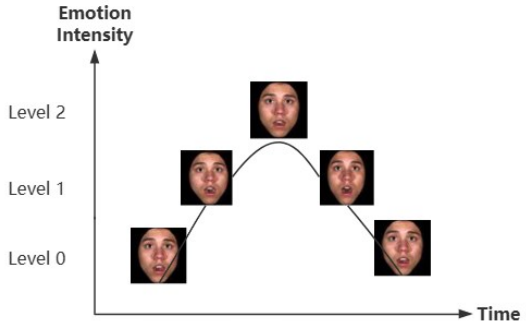


Fig. 1 Dynamic evolution of facial expressions chart

Dynamic expression recognition (DER) is to recognize the change process of a complete facial expression (e.g. neutral -> start -> apex -> end -> neutral). The process is shown in the Fig. 1 . In 1978, Suwa and colleagues began to recognize facial expressions from facial video in order to make the recognition results more representative, thus opening up the field of DER. Early researchers used manual operators (Gabor, LBP, Haar, etc.) to extract specific features for recognition[16][17]. In 2011, Gedeon proposed a large-scale datasets AFEW, significantly advancing the application of deep learning in DER [18]. Following this, researchers widely applied deep learning methods to the extraction of facial expression features such as CNN-based methods and RNN-based methods [19][20]. In 2017, with the success of the attention mechanism in the field of translation[11], which then was applied to spatial and temporal feature modeling. For instance, Zengqun designed a spatial and a temporal self-attention module at the level of model structure to fully model the expression features from the video [27]. In addition, some researchers have also combined temporal features with spatial features for processing, such as C3D and MAE-DEFR, which also have achieved great results on multiple dataset [21][22].

B. Identification of students' learning in classroom contexts

Currently, the recording and evaluation of students' learning status mainly adopts self-evaluation and classroom reports, which refers to students' evaluation of their own learning status after the course ends. With the rise of research on physiological signals in recent years, biofeedback technology has been used to obtain physiological indicators to analyze the learning status of learners. For instance, Wang Pengli and colleagues monitored learners' EEG signals to monitor their attention in class through brain-computer interface devices [23]. Moreover, the development of camera equipment and computer vision technology in recent years allows researchers focusing on identifying students' emotions through images, that is, identifying the emotional category based on changes in the learner's face, gestures and eye movements over a period of time. For instance, Sun Bo and colleagues [24] moved the learning scenario from online to offline and proposed a teacher-student emotional interaction system, which realized the emotion recognition and intervention functions of learners. Zhan Zehui et al. tracked learners' eye movement status and integrated it with emotional characteristics to build an intelligent agent learner emotion and cognitive model [25]. In summary, these systems enable teachers to monitor students' learning status in real time, allowing for timely intervention when issues arise.

III. Methodology

In this section, we introduce the proposed framework in detail, including three modules: Multi-Scale Facial Feature Fusion module (MFFF), Temporal-Encoder with FSOP and Design of a Joint Loss Function.

A. Overview of the Proposed Framework

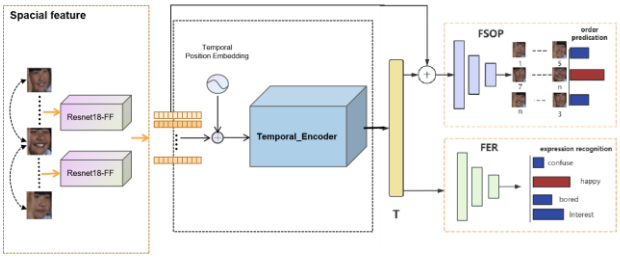


Fig. 2 Overview of the proposed framework for MTFNet from videos.

An overview of the proposed framework is presented in Fig. 2, which consists of three modules. First, For the input video V, the frame sequence$\{F_1, \cdots, F_n\}$ is extracted and then subsequently shuffled into $\{F_m, \cdots, F_t\}$. This shuffled sequence is fed into a MFFF module to extract frame-wise facial features $\{f_m, \cdots, f_t\}$. Following this, temporal feature extraction is performed by adding temporal positional encodings $\{em_m, \cdots, em_t\}$ to derive the video features. Finally, these

features are used for emotion classification module and frame order prediction module.

B. *Multi-Scale Facial Feature Fusion Module*

In real-world environments, students' facial emotion is often affected by lighting variations, occlusions, and poses changes. When the prominent features of a subject's face are obstructed, relying solely on deep, high-semantic features can severely impact the model's recognition performance. To mitigate the recognition challenges posed by occlusion, we employed a simple and efficient feature fusion strategy. Unlike the commonly used feature pyramid methods in object detection tasks, this approach focuses more on integrating features at different semantic levels rather than on positional information transmission. The model structure is shown in the Fig. 3. Given ResNet18's excellent generalization capabilities, it is used as the backbone network. The model includes two sets of feature fusion modules (FF). The first set integrates the low-semantic features from stages 1 to 3 of ResNet18,while the second set merges the fused features of the first set with the high-semantic features from stages 3 to 4, ultimately obtaining multi-level facial features :

$$f_{FF1} = FF1\big(Block1(F_i), Block2(F_i), Block3(F_i)\big) \quad (1)$$

$$f_i = FF2\big(f_{FF1}, Block3(F_i), Block4(F_i)\big) \quad (2)$$

As shown in the Fig. 4, each fusion unit comprises four functional channels, each utilizing two GS convolution modules. This module combines the original convolution kernels with depth-wise separable convolutions, ensuring feature extraction capability while reducing computational complexity. The GS convolution structure is shown in the Fig. 5.

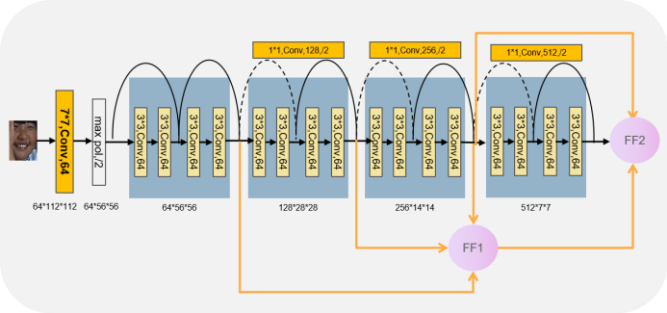


Fig. 3 ResNet18-FF Model Architecture Diagram

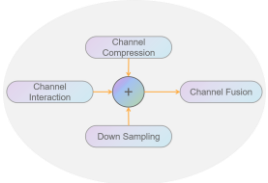


Fig. 4 Feature Fusion Module Structure Diagram

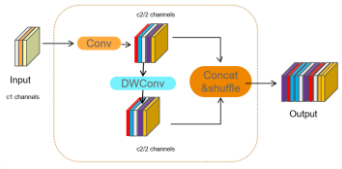


Fig. 5 GS Convolution Diagram

C. *Temporal-Encoder with FSOP*

In this module, leveraging the global modeling capabilities of the self-attention mechanism, we introduced the Temporal-Encoder for dynamic modeling of emotion sequences. The model structure is shown in the Fig. 6. Based on the random frame features $\{f_m, \cdots, f_t\}$ extracted by the Spatial Feature module, we added cosine positional encoding, Then mapped them to Q, K and V for the Temporal-Encoder to perform dynamic emotion feature modeling, thereby obtaining emotion classification features for the frame segments. The process show as:

$$\{f_{mm}, \cdots, f_{tm}\} = \{f_m, \cdots, f_t\} + \{em_m, \cdots, em_t\} \quad (3)$$

$$H = A(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (4)$$

Where $A(\cdot)$ represents the self-attention.

However, in practical applications, the multi-head attention mechanism is easily influenced by the intensity variations in facial features, leading to neglecting the overall motion features. For instance, when peak frames exhibit strong noise interference, the extraction of dynamic features is severely affected. To address this issue, Liu et al. proposed the Snippet-wise Shuffle Order Prediction module (SSOP), which divides emotion video sequences into multiple frame segments and performs order prediction at the segment level. While this method alleviates the problem, it is computationally too complex for real-time emotion prediction tasks. In this paper, we propose a Frame-wise Shuffle Order Prediction (FSOP) module, which performs order prediction at the frame level.

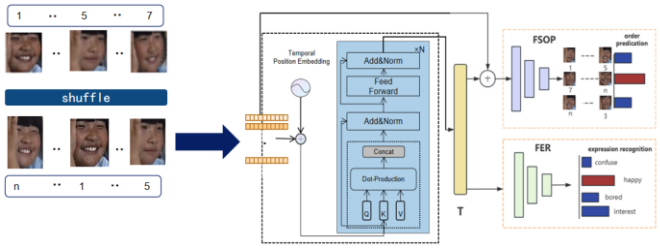


Fig. 6 Temporal-Encoder with FSOP

D. *Learning Criteria*

To supervise the learning of the proposed framework for expression recognition, we firstly consider cross entropy loss for Facial Expression Classification Module and Prediction of Frame Sequence Order, respectively. The cross entropy loss is defined as:

$$L_{cls} = -\sum_{i=1}^{N} y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i) \quad (5)$$

$$L_{fsp} = -\sum_{i=1}^{N} S_i * \log(\hat{S}_i) + (1 - S_i) * \log(1 - \hat{S}_i) \quad (6)$$

Where N denotes the number of samples, y_i denotes the true label, $\hat{y}_i$ denotes the predicted class probability, S_i denotes the true ordinal category, $\hat{S}_i$ denotes the predicted ordinal category probability.

Therefore, the overall loss function is defined as follows:

$$Loss = L_{cls} + aL_{fsp} \quad (7)$$

Where α denotes the weight of the proportion of category prediction results, Based on experimental results, the value of a was chosen to be 0.1428.

IV. Experiment

In this section, we present the experiments for evaluating the proposed framework for facial expression recognition. Specifically, i) we introduce the adopted databases in our experiments; ii) we present the implementation details of the proposed framework; iii) we show and analyze the comparisons of our method with other state-of-the-art methods, as well as the ablation studies for investigating the effects from the different designs in our proposed framework.

A. Datasets

In our experiments, we adopted two datasets to evaluate the performance of the proposed framework, including DFEW[28] and our dataset. Examples of the expressive faces from these databases are presented in Fig. 7.

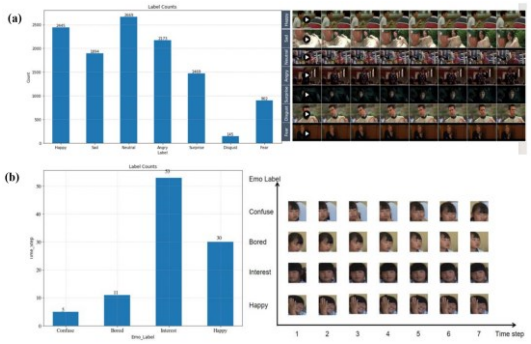


Fig. 7 Examples from (a) DFEW and (b) our dataset

The DFEW database is a large-scale, unconstrained dynamic facial expression database with a total of 12,059 single-label video clips. The dataset uses seven basic labels: Fear, Disgust, Surprise, Angry, Neutral, Sad, and Happy. Notably, the dataset encompasses various extreme scenarios such as lighting variations, occlusions, and pose changes, highlighting the significant challenges associated with accurate facial expression recognition in such complex environments.

Our dataset focuses on emotion recognition in group learning environments, addressing gaps in both domestic and international datasets.in collaboration with education experts, we defined four emotional categories: interested, happy, bored, and recorded five sets of classroom group discussion videos in primary school. (This program was approved by the Ethics Committee of the authors' university (No. BNU202201100014). Informed consent was obtained from all individual participants included in the study and their parents.) The videos were processed into image frame sequences, with frames showing clear emotional features labeled by experts. Seven frames before and after the labeled frame were extracted to form video clips. Further screening was conducted to ensure quality, removing unclear or blurred facial images. Fig. 7 shows examples of the retained frames.

B. Implementation Details

The DFEW dataset is divided into five groups, with an 8:2 split between the training and testing sets in each group. For the self-constructed dataset, due to the limited data volume, 20% of the data is used to fine-tune the classification layer of the model pre-trained on the DFEW dataset. The hyperparameters are shown in the Table 1.

Table 1 Training Hyperparameter Configuration

Hyperparameter	**Parmeter (DFEW)**	**Parmeter (our dataset)**
GPU Acceleration platform	CUDA11.1	
Framework	PyTorch 1.8 + python3.8	
Optimizer	ADAM	ADAM
Learning Rate	1e-3	1e-5
Random seed	3047	3047
Batch_size	32	32

C. Comparison With State-of-The-Art Methods

The Table 2 below shows that our model structure has significant improvements in recognizing multiple individual emotion categories as well as overall performance compared to other models. It can also be seen that the use of class-weighted optimization strategies effectively alleviates the long-tail problem caused by the uneven distribution of data samples, leading to a substantial improvement in the recognition of the Fear category. However, due to the limited amount of data in the Disgust category, the recognition performance for this category is very poor.

Table 2 Experimental Results on DFEW Dataset

Methods	**Accuracy of Each Emotion(%)**							**Metrics(%)**	
	Happy	Sadness	Neutral	Anger	Surprise	Disgust	Fear	**UAR**	**WAR**
VGG11+LSTM[24]	76.89	37.65	58.04	60.70	43.70	0.00	19.73	42.39	53.70
EC-STFL[25]	79.18	49.05	57.85	60.98	46.15	2.76	21.51	45.35	56.51
Resnet18+LSTM[23]	83.56	61.56	68.27	65.29	51.26	0.00	29.34	51.32	63.85
Resnet18+ GRU[23]	82.87	63.83	65.06	68.51	52.00	0.86	30.14	51.68	64.02
Former-DFER[27]	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70
EST[12]	86.87	66.58	67.18	71.84	47.52	**5.52**	28.49	53.43	65.85
STT[26]	87.36	**67.90**	64.97	71.24	53.1	3.49	34.04	**54.58**	66.45
MTFNet (w/o)	**91.12**	63.23	68.33	**72.13**	**68.10**	0.00	0.00	51.83	66.80
MTFNet(Ours)	89.78	58.73	**69.79**	65.90	52.90	0.00	**37.78**	53.55	**67.59**

The Table below compares the recognition performance of our model with that of some other models on the self-constructed dataset. It can be seen that on the self-constructed group learning dataset, our model outperforms the EST and STT models in terms of accuracy for multiple individual categories, as well as weighted and unweighted accuracy rates. The accuracy of the model for the interest category is relatively low, mainly because some data in the interest category are similar to the confuse category, making them more challenging to recognize.

Table 3 Comparison of Custom Group Learning Datasets

Methods	**Accuracy of Each Emotion(%)**				**Metrics(%)**	
	Confuse	Bored	Interest	Happy	**UAR**	**WAR**
EST[12]	40.00	63.63	37.78	86.67	57.00	55.56
STT[26]	40.00	72.72	**49.06**	86.67	62.13	62.62
MTFNet(Ours)	40.00	**90.91**	47.72	**93.33**	**67.85**	**64.65**

D. *Ablation study*

The results of the module ablation experiments on the DFEW dataset are shown in the Table below. The experimental results indicate that both the MFFF module and the FSOP module significantly contribute to this task, and their combination can further improve the model's performance.

Table 4 Results of Ablation Experiments on Modules

MFFF	FSOP	DFEW	
		WAR	UAR
		65.02	51.00
	√	65.31	52.73
√		66.35	51.22
√	√	**67.60**	**53.55**

We used t-SNE to visualize the learned feature distributions with and without the MFFF module, as shown in Figure 8. The feature distribution in the left figure, obtained after adding the MFFF module, shows greater inter-class differences and tighter intra-class cohesion, with clear boundary lines. This also verifies that the introduction of the MFFF module enriches the spatial semantic features of the images, thereby enhancing the model's accuracy.

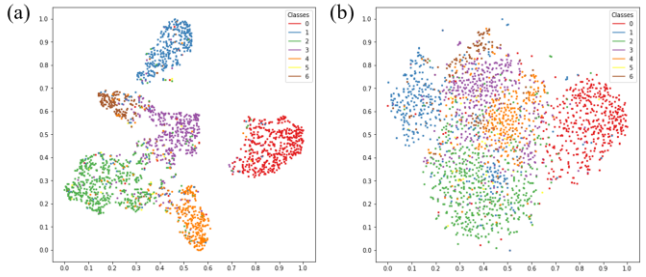


Figure 8 Visualisation of feature distribution (t-SNE)

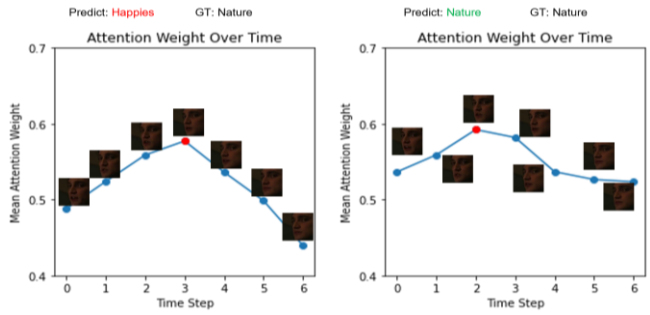


Figure 9 Temporal Attention Visualisation

We visualized the attention weights of the last layer features in the Temporal Transformer in Figure 9. The comparison of the samples shows that our model focuses more on the overall facial changes in the sequence, resulting in a more balanced distribution of weights. The introduction of the FSOP module reduces the impact of the multi-head attention mechanism's excessive focus on peak variation frames, leading to better temporal modeling of emotional information.

E. *Limitations and Future Works*

Currently, the challenges in facial expression recognition within group learning scenarios mainly stem from data scarcity and the loss of facial features in complex environments. To address this, we introduced the MFFF module in the spatial feature extraction part to capture multi-scale features and collected a small evaluation dataset. In the future, the performance of recognition in this scenario can be improved through the following approaches:

- Dataset Expansion: High-quality data will help the model better capture high-quality features.
- Semi-Supervised Learning: By combining general emotion datasets for learning, this approach can alleviate the issue of data scarcity in specific scenarios and expand the existing small dataset.

V. CONCLUSION

Given the complexity of group learning scenarios and the lack of relevant datasets, this paper proposes the MTFNet model, which is based on the efficiency of the spatial-temporal framework for recognition. Our model can be deployed on terminals and achieve real-time monitoring. In the spatial feature extraction part, we introduced a Multi-scale Facial Feature Fusion Module (MFFF) based on GS convolution, which integrates features from different semantic spaces while controlling the number of parameters, mitigating the decline in performance caused by complex scenarios. In the temporal feature extraction phase, we improved the Frame-wise Shuffle Order Prediction Module (FSOP) to guide the model's attention to changes between frames, enhancing the model's ability to capture dynamic emotional changes. Additionally, we collaborated with educational experts to collect data and construct a small evaluation dataset suitable for group learning scenarios in real settings. Finally, our model achieved excellent performance on both datasets, and the ablation study results on the DFEW dataset also demonstrate the model's applicability across different scenarios.

VI. ACKNOWLEDGMENT

We sincerely appreciate the generous support from several funding agencies for this research. First and foremost, this study was funded by the National Natural Science Foundation of China under the project titled "Multi-modal Assessment Models and Interventions for Student Engagement in Cooperative Learning" (No. 62277007) and the Guangdong Province University Characteristic Innovation Project (Natural Science) (No. 2022KTSCX205) . Additionally, we are grateful for the support from Industry-University-Research Innovation Fund for Chinese University(No. 2024XF014), as well as the aid provided by the Guangdong Province Undergraduate Course Teaching and Research Office Construction Project (No. jx2022303).

The funding from these institutions was crucial in facilitating the successful execution of our research plan and achieving the anticipated outcomes. References.

[1] R. Adyapady R and A. Basava, "A comprehensive review of facial expression recognition techniques," Multimedia Systems, vol. 29, Jul. 2022, doi: 10.1007/s00530-022-00984-w.

[2] Ö. Sümer, P. Goldberg, S. D’Mello, P. Gerjets, U. Trautwein, and E. Kasneci, ‘Multimodal engagement analysis from facial videos in the classroom’, IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1012 – 1027, 2021.

[3] S. Gupta, P. Kumar, and R. K. Tekchandani, “Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models,” Multimed Tools Appl, vol. 82, no. 8, pp. 11365 – 11394, Mar. 2023, doi: 10.1007/s11042-022-13558-9.

[4] R. Pekrun and L. Linnenbrink-Garcia, “Academic Emotions and Student Engagement,” in Handbook of Research on Student Engagement, S. L. Christenson, A. L. Reschly, and C. Wylie, Eds., Boston, MA: Springer US, 2012, pp. 259 – 282. doi: 10.1007/978-1-4614-2018-7_12.

[5] H. Li, H. Niu, Z. Zhu, and F. Zhao, “Intensity-Aware Loss for Dynamic Facial Expression Recognition in the Wild,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 1, Art. no. 1, Jun. 2023, doi: 10.1609/aaai.v37i1.25077.

[6] H. Kim, J.-H. Lee, and B. C. Ko, “Facial Expression Recognition in the Wild Using Face Graph and Attention,” IEEE Access, vol. 11, pp. 59774 – 59787, 2023, doi: 10.1109/ACCESS.2023.3286547.

[7] R. Zhi, M. Flierl, Q. Ruan, and W. Kleijn, “Graph-Preserving Sparse Nonnegative Matrix Factorization With Application to Facial Expression Recognition,” IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, vol. 41, pp. 38 – 52, Feb. 2011, doi: 10.1109/TSMCB.2010.2044788.

[8] T. Ojala, M. Pietikainen, and T. Maenpaa, ‘Multiresolution gray-scale and rotation invariant texture classification with local binary patterns’, IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 7, pp. 971 – 987, 2002.

[9] M. Shafiq and Z. Gu, ‘Deep residual learning for image recognition: A survey’, Applied Sciences, vol. 12, no. 18, p. 8972, 2022.

[10] S. Hochreiter and J. Schmidhuber, ‘Long Short-Term Memory’, Neural Computation, vol. 9, no. 8, pp. 1735 – 1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. of Advances in Neural Information Processing Systems (NIPS), pp. 5998-6008, Dec. 2017.

[12] Y. Liu, W. Wang, C. Feng, H. Zhang, Z. Chen, and Y. Zhan, ‘Expression snippet transformer for robust video-based facial expression recognition’, Pattern Recognition, vol. 138, p. 109368, Jun. 2023, doi: 10.1016/j.patcog.2023.109368.

[13] G. Zhao, Y. Zhang, and J. Chu, ‘A multimodal teacher speech emotion recognition method in the smart classroom’, Internet of Things, vol. 25, p. 101069, 2024.

[14] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, ‘Affective computing in education: A systematic review and future research’, Computers & Education, vol. 142, p. 103649, Dec. 2019, doi: 10.1016/j.compedu.2019.103649.

[15] P. Ekman and W. V. Friesen, Facial action coding system: investigator ’ s guide. Palo Alto, Calif.: Consulting Psychologists Press, 1978.

[16] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, ‘Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron ’ , in Proceedings Third IEEE International Conference on Automatic face and gesture recognition, 1998, pp. 454 – 459. Accessed: Jul. 07, 2024.

[17] C. Shan, S. Gong, and P. W. McOwan, ‘Robust facial expression recognition using local binary patterns’, in IEEE International Conference on Image Processing 2005, 2005, vol. 2, p. II – 370. Accessed: Jul. 07, 2024.

[18] M. F. Valstar et al., ‘Fera 2015-second facial expression recognition and analysis challenge’, in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, vol. 6, pp. 1 – 8. Accessed: Jul. 07, 2024.

[19] V. Vielzeuf, S. Pateux, and F. Jurie, ‘Temporal multimodal fusion for video emotion classification in the wild’, in Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow UK, Nov. 2017, pp. 569 – 576. doi: 10.1145/3136755.3143011.

[20] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, ‘Empirical evaluation of gated recurrent neural networks on sequence modeling’, CoRR, vol. abs/1412.3555, 2014

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, ‘Learning spatiotemporal features with 3d convolutional networks’, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489 – 4497. Accessed: Jul. 07, 2024.

[22] L. Sun, Z. Lian, B. Liu, and J. Tao, ‘MAE-DFER: Efficient Masked Autoencoder for Self-supervised Dynamic Facial Expression Recognition’, in Proceedings of the 31st ACM International Conference on Multimedia, Ottawa ON Canada, Oct. 2023, pp. 6110 – 6121. doi: 10.1145/3581783.3612365.

[23] P. L. Wang, Q. C. Ke, and J. Q. Zhang, ‘Research on application of brain-computer interface in smart classroom’, Open Education Research, vol. 26, pp. 72 – 81, 2020.

[24] S. Bo, L. Yongna, C. Jiubing, L. Jihong, and Z. Di, ‘Emotion analysis based on facial expression recognition in smart learning environment’, Modern Distance Education Research, vol. 2, pp. 96 – 103, 2015.

[25] Z. H. Zhan, ‘An emotional and cognitive recognition model for distance learners based on intelligent agent-the coupling of eye tracking and expression recognition techniques’, Mod. Dist. Educ. Res, vol. 5, pp. 100 – 105, 2013.

[26] F. Ma, B. Sun, and S. Li, ‘Spatio-Temporal Transformer for Dynamic Facial Expression Recognition in the Wild’. arXiv, May 10, 2022. doi: 10.48550/arXiv.2205.04749.

[27] Z. Zhao and Q. Liu, ‘Former-DFER: Dynamic Facial Expression Recognition Transformer’, in Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA, Oct. 2021, pp. 1553 – 1561. doi: 10.1145/3474085.3475292.

[28] X. Jiang et al., ‘DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild’, in Proceedings of the 28th ACM International Conference on Multimedia, New York, NY, USA, Oct. 2020, pp. 2881 – 2889. doi: 10.1145/3394171.3413620.

Target Speaker Extraction Method by Emphasizing the Active Speech with an Additional Enhancer

Xue Yang, Changchun Bao*, Xu Zhang and Xianhong Chen
Institute of Speech and Audio Information Processing, School of Information Science and Technology,
Beijing University of Technology, Beijing 100124, China
E-mail: yangx11@emails.bjut.edu.cn, baochch@bjut.edu.cn, xuzhang223@emails.bjut.edu.cn, chenxianhong@bjut.edu.cn

***Abstract*—Target speaker extraction (TSE) is a practical solution to the cocktail party problem. Recently, a novel embedding-free TSE method was proposed. In this method, the enrollment and the mixed signal are directly interacted to exploit the contextual information within the enrollment. In the absence of noise, the derived guidance exhibits the onset, offset and voice activity similar to the mixed signal. However, in the presence of noise, such similarity may be destroyed since the enrollment is interacted with both speech and noise signals in the mixture. If the noise (e.g., babble noise) contains components that resemble the enrollment to some extent, the misleading guidance may be generated after the direct interaction. To tackle this issue, an additional enhancer is designed in this paper to derive an auxiliary guidance that emphasizes the active speech. Specifically, this enhancer consists of a processing block and an interaction block. The processing block mainly utilizes the recurrent layers to model the temporal dynamics of the enrollment and mixed signal. In this block, the speech and noise signals are modeled in different manners and the similarity between the enrollment and noise can be reduced. Afterwards, the processed representations of the enrollment and mixed signal are utilized to derive an enhanced representation in the interaction block. This enhanced representation emphasizes the active speech and is employed as an auxiliary guidance for the extraction. Experimental results demonstrate the effectiveness of our proposed method in complex acoustic environments.**

I. INTRODUCTION

In complex acoustic environments, the mixed signal captured by the microphone may comprise speech signals from multiple speakers, noise and reverberation. While human beings can effectively focus on the speaker of interest in such environment, it is hard for the machine to possess this capability. This challenge is known as the cocktail party problem [1]. To tackle this problem, the speech separation (SS) and target speaker extraction (TSE) are primarily studied. The SS aims to estimate all speech signals within the mixed signal [2] and has recently achieved outstanding performance [3-6]. However, the SS assumes to know the number of speakers and faces the global permutation ambiguity [7-8]. These issues make the SS less practical in real-world scenarios. In contrast, the TSE focuses on isolating the target speaker's speech, guided by the auxiliary information of the target speaker [9]. Hence, the TSE is a more practical solution since it does not require to know the number of speakers and avoids the global permutation ambiguity. In this paper, the reference utterance of the target speaker, known as the enrollment, is employed as the auxiliary information due to its effectiveness and easy accessibility.

Depending on how enrollment is leveraged, the TSE can be divided into two categories: embedding-based TSE method and embedding-free TSE method. Typically, the embedding-based TSE method derives a speaker embedding from the enrollment and employs this embedding as guidance. In [10-11], the speaker embedder pretrained on the speaker verification or recognition task was employed to derive the embedding. In [12-13], the speaker embedder was jointly trained with the extraction network to address the sub-optimization issue. Besides, techniques such as speaker representation loss [14] or self-supervised disentangled representation [15] were adopted to improve the embedding discriminability. Following the derivation of the speaker embedding, the extraction process involves fusing the embedding with the features of the mixed signal and various fusion methods have been proposed [16-19]. Additionally, the multi-stage framework [20], onset/offset information [21] or iterative refined adaptation [22] have been used to boost the extraction performance. Nevertheless, being a compact vector, speaker embedding mainly summarizes the speaker characteristics and discards the content details. Thus, these embedding-based TSE methods may not fully leverage the contextual information within the enrollment.

Recently, the embedding-free TSE methods have gained increasing attention since both the speaker characteristics and the content details are leveraged. In [23-25], high-dimensional feature sequences derived from the enrollment and the mixed signal were fused through various attention mechanisms. In [26], the hidden states and cell states of the recurrent neural network (RNN) were adopted to summarize the speaker information. In [27], two pooling approaches were introduced to generate speaker representations. In these methods, the processed feature sequences or states were utilized as guidance and the contextual information is partially exploited. In [28], the enrollment and mixed signal were directly interacted through an attention mechanism. In the absence of noise, the derived guidance of this contextual information exploitation

network (CIENet) exhibits the onset, offset and voice activity similar to the mixed signal. However, in the presence of noise, such similarity may be destroyed since the enrollment is interacted with both speech and noise signals in the mixture. If the noise (e.g., babble noise) contains components that resemble the enrollment to some extent, the misleading guidance may be generated after the direct interaction. Therefore, it is necessary to explore more appropriate guidance in complex acoustic environments.

In this paper, we extend the CIENet with an additional enhancer to derive an auxiliary guidance that emphasizes the active speech. Specifically, this additional enhancer consists of a processing block and an interaction block. The processing block mainly utilizes the recurrent layers to model the temporal dynamics of the enrollment and mixed signal. In this block, the speech and noise signals are modeled in different manners and the similarity between the enrollment and noise can be reduced. Afterwards, the processed representations of the enrollment and the mixed signal are interacted to derive an enhanced representation in the interaction block. This enhanced representation emphasizes the active speech and is used as an auxiliary guidance for the extraction. This enhanced contextual information exploitation network (CIENet-Enh) achieves better performance in complex acoustic environments.

The rest of this paper is organized as follows. The proposed TSE method is first detailed in Section 2. Afterwards, the experimental setup is outlined in Section 3. The results and discussions are presented in Section 4. Finally, the conclusions are drawn in Section 5.

II. Proposed Method

A. Problem Formulation

Guided by the target speaker's enrollment $\boldsymbol{e}\in R^{1\times L_e}$, the TSE aims to isolate the target speaker's speech from the mixed signal $\boldsymbol{y}\in R^{1\times L_y}$, namely:

$$\hat{\boldsymbol{x}}_{tgt}=\mathcal{M}(\boldsymbol{y}|\boldsymbol{e})=\mathcal{M}\left(\left(\sum_{i=1}^{S}\boldsymbol{z}_i+\boldsymbol{n}\right)\middle|\boldsymbol{e}\right) \tag{1}$$

where L_e is the length of the enrollment, L_y is the length of the mixed signal, $\hat{\boldsymbol{x}}_{tgt}$ is the estimated signal of the target speaker, $\mathcal{M}(\boldsymbol{a}|\boldsymbol{b})$ is the mapping function of $\boldsymbol{a}$ given $\boldsymbol{b}$, S is the number of speakers, the vector $\boldsymbol{z}_i$ denotes the signal coming from the ith speaker and the vector $\boldsymbol{n}$ represents the additive noise.

B. Proposed Network Architecture

As illustrated in Fig. 1, our proposed CIENet-Enh extends the original CIENet with an additional enhancer.

1) Original CIENet

The original CIENet takes the mixed signal $\boldsymbol{y}$ and the enrollment $\boldsymbol{e}$ as inputs to isolate the target speaker's speech through mask estimation. As given in the lower part of Fig. 1, the mixed signal $\boldsymbol{y}$ and the enrollment $\boldsymbol{e}$ are first transformed into the time-frequency (T-F) domain using short-time Fourier transform (STFT). Their T-F representations $\boldsymbol{Y}\in C^{T_Y\times F}$ and $\boldsymbol{E}\in C^{T_E\times F}$ are compressed with the dynamic range compression (DRC) [29] to reduce the dynamic range and emphasize the regions with small values. The compressed T-F representations $\boldsymbol{Y}_c\in R^{2\times T_Y\times F}$ and $\boldsymbol{E}_c\in R^{2\times T_E\times F}$ are expressed as

$$\boldsymbol{Y}_c=\text{Concat}\left(|\boldsymbol{Y}|^{\gamma}\cos\boldsymbol{\theta}_Y,|\boldsymbol{Y}|^{\gamma}\sin\boldsymbol{\theta}_Y\right) \tag{2}$$

$$\boldsymbol{E}_c=\text{Concat}\left(|\boldsymbol{E}|^{\gamma}\cos\boldsymbol{\theta}_E,|\boldsymbol{E}|^{\gamma}\sin\boldsymbol{\theta}_E\right) \tag{3}$$

where T_Y and T_E are the frame numbers of the mixed signal and the enrollment, F is the number of frequencies, $|\boldsymbol{Y}|$ and $|\boldsymbol{E}|$ are the magnitude spectra, $\boldsymbol{\theta}_Y$ and $\boldsymbol{\theta}_E$ are the phase spectra, Concat(·, ·) is the concatenation operation, γ denotes the compression factor ranging from 0 to 1.

Afterward, the T-F representations $\boldsymbol{Y}_c$ and $\boldsymbol{E}_c$ are interacted in the interaction block to derive the guidance $\boldsymbol{F}_c\in R^{2\times T_Y\times F}$. For clarity, the magnitude spectrum comparison is depicted in Fig. 2. In the absence of noise, the direct interaction between the enrollment in Fig. 2(a) and the clean mixture in Fig. 2(b) leads to the guidance in Fig. 2(c). This derived guidance not only retains target speaker's characteristics to some extent but also exhibits the onset, offset and voice activity similar to the clean mixture. In the presence of noise, the enrollment is interacted with both speech and noise signals in the mixture. If the noise contains components that resemble the enrollment to some extent, the misleading guidance may be generated. For instance, the direct interaction between the enrollment in Fig. 2(d) and the noise in Fig. 2(e) is depicted in Fig. 2(f). Besides, the direct interaction between the enrollment in Fig. 2(g) and the noisy mixture in Fig. 2(h) is shown in Fig. 2(i). Compared with the guidance in Fig. 2(c), misleading guidance in Fig. 2(i) no longer exhibits the onset, offset and voice activity similar to the clean mixture. Thus, more appropriate guidance needs to be explored.

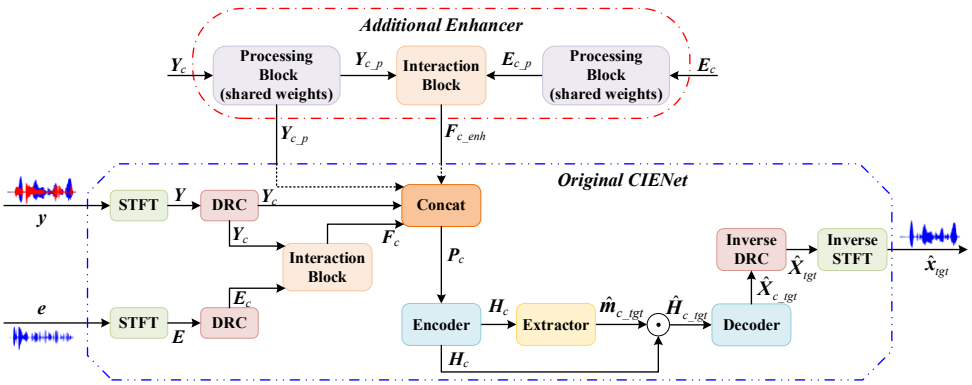


Fig.1 The general framework of CIENet-Enh.

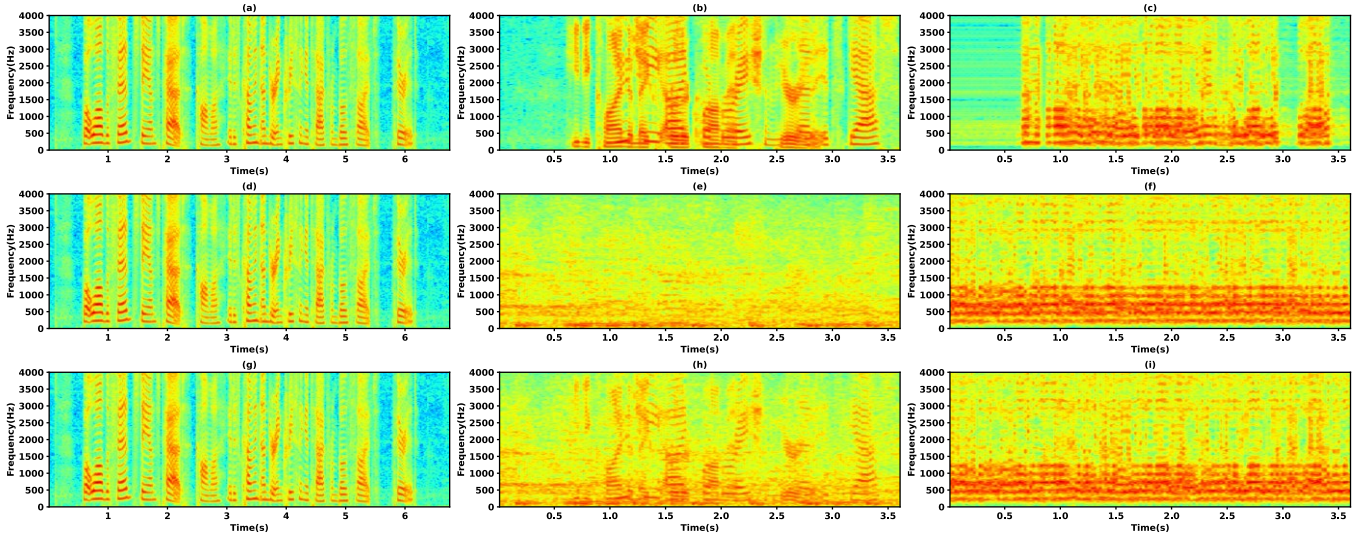


Fig. 2 The magnitude spectrum comparison in the original CIENet. (a) Enrollment; (b) Clean mixture; (c) Interaction between the enrollment and clean mixture; (d) Enrollment; (e) Noise; (f) Interaction between the enrollment and noise; (g) Enrollment; (h) Noisy mixture; (i) Interaction between the enrollment and noisy mixture.

In the original CIENet, the guidance $\boldsymbol{F}_c$ is concatenated with the T-F representation $\boldsymbol{Y}_c$ of the mixed signal. The stacked feature tensor of shape $4\times T_Y\times F$ is further processed by the encoder and the extractor to estimate the mask $\hat{\boldsymbol{m}}_{c_tgt}\in R^{K\times T_Y\times F}$. As illustrated in Fig. 1, this mask is element-wise multiplied with the output $\boldsymbol{H}_c$ of the encoder to derive the feature tensor $\hat{\boldsymbol{H}}_{c_tgt}$ of the target speaker, namely:

$$\hat{\boldsymbol{H}}_{c_tgt}=\boldsymbol{H}_c\odot\hat{\boldsymbol{m}}_{c_tgt} \tag{4}$$

where K is the channel dimension and the symbol "$\odot$" denotes the element-wise multiplication. The derived feature tensor $\hat{\boldsymbol{H}}_{c_tgt}$ is processed by the decoder to obtain the estimated T-F representation $\hat{\boldsymbol{X}}_{c_tgt}\in R^{2\times T_Y\times F}$. Finally, the inverse DRC and inverse STFT are applied to recover the estimated signal $\hat{\boldsymbol{x}}_{tgt}$. In practice, the STFT and inverse STFT are implemented with convolutional layer and transposed convolutional layer.

2) *Additional Enhancer*

To derive more appropriate guidance for complex acoustic environments, an additional enhancer is introduced. As shown in the upper part of Fig. 1, this enhancer consists of a processing block and an interaction block.

The T-F representation $\boldsymbol{Y}_c$ and $\boldsymbol{E}_c$ are individually processed by the processing block to derive the processed representations $\boldsymbol{Y}_{c_p}$ and $\boldsymbol{E}_{c_p}$. As depicted in Fig. 3(a), the channel dimension of the representation $\boldsymbol{Y}_c$ is first expanded to D using a two-dimensional convolutional layer. Then, the expanded feature sequence undergoes reshaping and modeling by two bi-directional long short-term memory (BLSTM) [30] layers. The resulting feature is reshaped and further transformed with a convolutional layer to obtain the processed representation $\boldsymbol{Y}_{c_p}\in R^{2\times T_Y\times F}$. Similarly, the T-F representation $\boldsymbol{E}_c$ of the enrollment is processed to derive the processed representation $\boldsymbol{E}_{c_p}\in R^{2\times T_E\times F}$. Note that the BLSTM layers are utilized to model the temporal dynamics of the enrollment and mixed signal. Since the temporal dynamics of speech and noise signals are distinct, they can be modeled in different manners. In this way, the similarity between the enrollment and noise can be reduced.

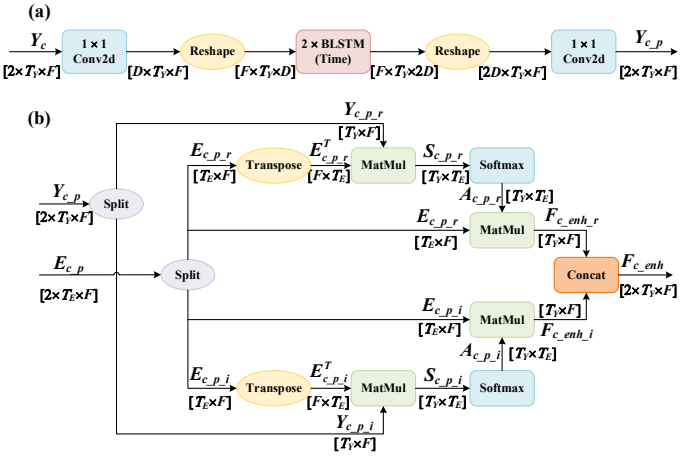


Fig.3 (a) Processing block; (b) Interaction block.

Afterwards, these processed representations $\boldsymbol{Y}_{c_p}$ and $\boldsymbol{E}_{c_p}$ are interacted in the interaction block to derive the enhanced representation $\boldsymbol{F}_{c_enh}\in R^{2\times T_Y\times F}$. As illustrated in Fig. 3(b), these representations are split into the real-part features ($\boldsymbol{Y}_{c_p_r}$ and $\boldsymbol{E}_{c_p_r}$) and imaginary-part features ($\boldsymbol{Y}_{c_p_i}$ and $\boldsymbol{E}_{c_p_i}$). The real-part feature $\boldsymbol{Y}_{c_p_r}$ is matrix-multiplied with the transposed real-part feature $\boldsymbol{E}^T_{c_p_r}$ to calculate similarity matrix $\boldsymbol{S}_{c_p_r}$ for different frames. Subsequently, the Softmax function is applied on the last dimension of this similarity matrix to derive the attention matrix $\boldsymbol{A}_{c_p_r}$. This attention matrix is further matrix-multiplied with the real-part feature $\boldsymbol{E}_{c_p_r}$ to obtain the enhanced real-part feature $\boldsymbol{F}_{c_enh_r}$. Similarly, the enhanced imaginary-part feature $\boldsymbol{F}_{c_enh_i}$ can be derived, namely:

$$\boldsymbol{F}_{c_enh_r}=\boldsymbol{A}_{c_p_r}\boldsymbol{E}_{c_p_r}=\mathrm{Softmax}\left(\boldsymbol{Y}_{c_p_r}\boldsymbol{E}^T_{c_p_r}\right)\boldsymbol{E}_{c_p_r} \tag{5}$$

$$\boldsymbol{F}_{c_enh_i}=\boldsymbol{A}_{c_p_i}\boldsymbol{E}_{c_p_i}=\mathrm{Softmax}\left(\boldsymbol{Y}_{c_p_i}\boldsymbol{E}^T_{c_p_i}\right)\boldsymbol{E}_{c_p_i} \tag{6}$$

where the superscript "T" is the transpose operation. These enhanced features are concatenated to produce the enhanced representation $\boldsymbol{F}_{c_enh}\in R^{2\times T_Y\times F}$, which is further employed as an auxiliary guidance. As shown in Fig. 1, the representations $\boldsymbol{Y}_c$,

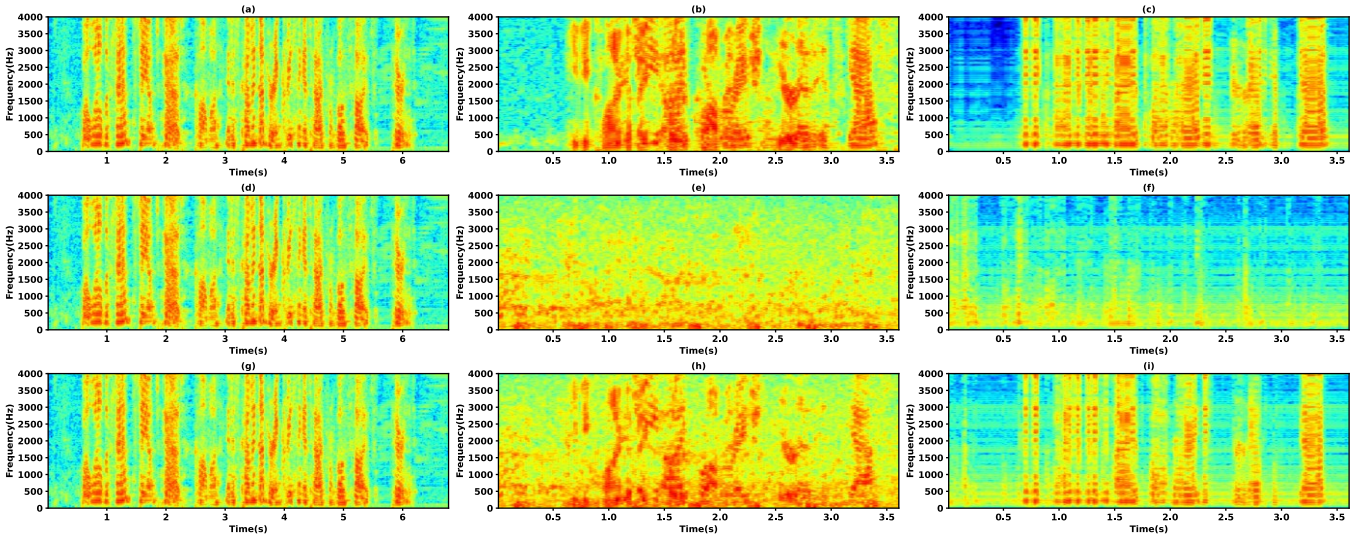


Fig. 4 The magnitude spectrum comparison in the additional enhancer.
(a) Processed enrollment; (b) Processed clean mixture; (c) Interaction between the processed enrollment and processed clean mixture;
(d) Processed enrollment; (e) Processed noise; (f) Interaction between the processed enrollment and processed noise;
(g) Processed enrollment; (h) Processed noisy mixture; (i) Interaction between the processed enrollment and processed noisy mixture.

$\boldsymbol{F}_c$, $\boldsymbol{Y}_{c_p}$ and $\boldsymbol{F}_{c_enh}$ are concatenated to form the stacked feature tensor $\boldsymbol{P}_c \in R^{8\times T_1\times F}$. Note that the same interaction block is used in the original CIENet.

For clarity, the magnitude spectrum comparison in this additional enhancer is depicted in Fig. 4. After modeling in the processing block, the harmonic structures of the processed enrollment in Fig. 4(a) and the processed clean mixture in Fig. 4(b) are somewhat distorted. However, their interaction in Fig. 4(c) still exhibits certain voice activity of the clean mixture and target speaker's characteristics. Additionally, the interaction between the processed enrollment in Fig. 4(d) and processed noise in Fig. 4(e) is given in Fig. 4(f). Compared with the direct interaction between the enrollment and noise in Fig. 2(f), the similarity between the processed enrollment and the processed noise is effectively reduced. This may be due to the different modeling manners for the speech and noise signals in the processing block. Furthermore, the interaction between the processed enrollment in Fig. 4(g) and the processed noisy mixture in Fig. 4(h) is shown in Fig. 4(i). It can be observed that this enhanced representation emphasizes the active speech in complex acoustics environments. Therefore, this enhanced representation is employed as an auxiliary guidance, although the over-suppression may occur.

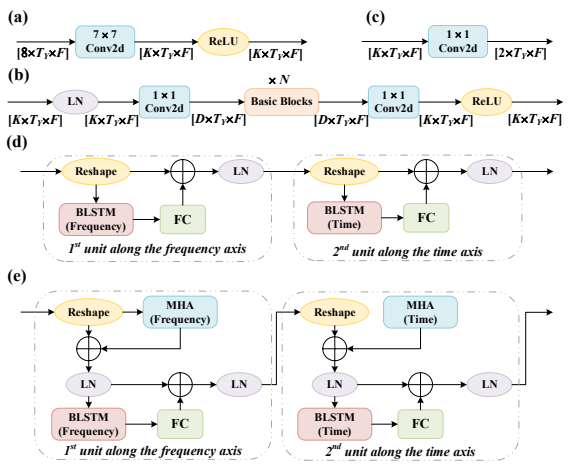


Fig. 5 (a) Encoder; (b) Extractor; (c) Decoder;
(d) mDPRNN block; (e) mDPTNet block.

3) *Details of Different Modules*

The details of the encoder, extractor and decoder are illustrated in Fig. 5. The encoder consists of a convolutional layer and a rectified linear unit (ReLU). The extractor comprises layer normalization (LN), N basic blocks positioned between two convolutional layers and ReLU. The decoder is a convolutional layer. For fair comparison with the original CIENet, the same types of basic blocks are used. The first type is the modified dual-path RNN (mDPRNN) block derived from [3]. This block includes two units to model dependencies along both the frequency and time axes. Each unit consists of BLSTM, fully connected (FC) layer, skip connection and LN. The second type is the modified dual-path transformer (mDPTNet) block derived from [4], which uses multi-head attention (MHA).

III. Experimental Setup

Three benchmark datasets derived from the Wall Street Journal (WSJ0) corpus are utilized in this paper. The first one is WSJ0 Hipster Ambient Mixtures (WHAM!) [31], which includes noises collected from the real-world scenarios. The second dataset WHAMR! [32] is the reverberant version of the WHAM! dataset. The third dataset is WSJ0-2Mix [33], where neither noise nor reverberation is introduced. Note that each speaker in the mixed signal is considered as the target speaker in turn. The sampling rate is 8kHz for all three datasets.

To perform TSE in the T-F domain, the Hanning window is utilized to split waveforms. The window length and the hop size are set to 32ms and 16ms, respectively. Besides, the number of frequencies F is 129. The compression factor γ in DRC is set to 0.5. The hyper-parameters K and D are 256 and 64, respectively. In the basic blocks, the number of hidden units for each direction of BLSTM is 128 and the number of attention heads

is 4. Besides, the number of basic blocks N is set to 6.

All models are trained with 4s long speech signals for 120 epochs. The Adam [34] optimizer is utilized and the initial learning rate is set to 0.0003. In the first 100 epochs, the learning rate is multiplied with a factor of 0.99 for ever two epochs. In the last 20 epochs, this factor is reduced to 0.9. The gradient clipping is adopted to limit the maximum L_2 norm to 1. The training objective is to maximize the scale-invariant signal-to-distortion ratio (SI-SDR) [35] between the estimated signal and the ground-truth signal of the target speaker.

IV. Results and Discussions

The SI-SDR improvement (SI-SDRi) and signal-to-distortion ratio [36] improvement (SDRi) are employed to evaluate the extraction accuracy.

A. Performance Comparison on the Complex Datasets

In Tab. 1, the CIENet-Enh is compared with several methods on the WHAM! and WHAMR! datasets. The embedding-based methods and embedding-free methods are separated with a double line. The first four methods conduct the TSE in the time-domain, while others perform the TSE in the T-F domain.

Tab. 1 Comparison on the complex datasets.

Methods	WHAM!		WHAMR!	
	SI-SDRi (dB)	SDRi (dB)	SI-SDRi (dB)	SDRi (dB)
SpEx [8]	12.2†	13.0†	10.3†	9.5†
SpEx+ [13]	13.1†	13.6†	10.9†	10.0†
DPRNN-Spe-IRA [22]	14.2	14.6	–	–
SpEx++ [20]	14.3	14.7	11.7	10.7
X-TF-GridNet [19]	15.7	16.1	15.3	14.2
Enhance-CIENet-mDPTNet	3.5	4.8	3.8	3.8
CIENet-mDPRNN [28]	15.7	16.1	15.5	14.1
CIENet-mDPTNet [28]	16.6	17.0	15.7	14.3
CIENet-Enh-mDPRNN	16.1	16.4	16.4	15.0
CIENet-Enh-mDPTNet	**17.2**	**17.5**	**17.2**	**15.8**

-Results with superscript "†" are given in [20].

On both datasets, the X-TF-GridNet [19] outperforms the time-domain approach using advanced network architecture [5] and fusion method. All five methods utilize the embedding as guidance and do not fully leverage the contextual information within the enrollment. In contrast, the CIENet achieves higher performance by fully leveraging the contextual information. However, in the presence of noise, misleading guidance may be generated. By introducing an additional enhancer, an auxiliary guidance emphasizing the active speech is leveraged and the CIENet-Enh can achieve better performance. Additionally, the Enhance-CIENet-mDPTNet is designed to show that CIENet-Enh is more effective than the enhance-then-extract approach. In this reference method, the mixed signal is first denoised with a pretrained model [37] and then the extraction is conducted with the pretrained CIENet-mDPTNet model. As shown in Tab. 1, this reference method performs the worst on both datasets. This is due to the sub-optimization issue of the cascaded system and the distortions of speech signals introduced during the enhancement process. In constrast, the CIENet-Enh can be optimized end-to-end, utilizing an enhanced representaion that emphasizes the active speech as an auxiliary guidance.

B. Performance Comparison on the WSJ0-2Mix Dataset

In Tab. 2, the CIENet-Enh is compared with a baseline method and CIENet on the WSJ0-2Mix dataset. The baseline method is similar to the CIENet-Enh, except that the direct interaction $\boldsymbol{F}_c$ is not used as guidance. Compared with CIENet-mDPRNN, the performance of the baseline method degrades, indicating that the direct interaction $\boldsymbol{F}_c$ is more effective on the clean dataset. Therefore, both the direct interaction $\boldsymbol{F}_c$ and the enhanced representation $\boldsymbol{F}_{c_enh}$ are utilized as guidance in the CIENet-Enh. Additionally, the CIENet-Enh does not exhibit a significant improvement over CIENet. This is reasonable since the direct interaction $\boldsymbol{F}_c$ can already provide effective guidance on this clean dataset. Furthermore, only a marginal increase is observed in the number of parameters (N_p) for CIENet-Enh.

Tab. 2 Comparison on the WSJ0-2Mix dataset.

Methods	N_p ($\times 10^6$)	WSJ0-2Mix	
		SI-SDRi (dB)	SDRi (dB)
Baseline-mDPRNN	2.9	20.4	20.6
CIENet-mDPRNN [28]	2.7	20.7	21.0
CIENet-mDPTNet [28]	2.9	21.4	21.6
CIENet-Enh-mDPRNN	2.9	20.9	21.1
CIENet-Enh-mDPTNet	3.1	**21.5**	**21.8**

V. Conclusions

In this paper, an additional enhancer was introduced to extend CIENet for better accommodation of complex acoustic environments. Specifically, this enhancer includes a processed block and an interaction block. In the processing block, the temporal dynamics of the enrollment and mixed signal were modeled. By modeling the speech and noise signals in different manners, the similarity between the enrollment and noise was reduced. As a results, an enhanced representation emphasizing the active speech was derived in the interaction block. This enhanced representation was employed as an auxilary guidance for the extraction. Our proposed CIENet-Enh achieved superior performance on both WHAM! and WHAMR! datasets.

VI. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants 61831019 and 62006010), R&D Program of Beijing Municipal Education Commission (KM202210005029).

References

[1] E. C. Cherry, On Human Communication, Cambridge, MA, USA: MIT Press, 1957.

[2] JD. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702-1726, 2018.

[3] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE ICASSP*, 2020, pp. 46-50.

[4] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, 2020, pp. 2642-2646.

[5] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221-3236, 2023.

[6] X. Yang, C. Bao, and X. Chen, "Coarse-to-fine speech separation in the time-frequency domain," in *Speech Communication*, vol. 155, 103003, 2023.

[7] M. Delcroix *et al.*, "Improving speaker discrimination of target speech extraction with time-domain Speakerbeam," in *IEEE ICASSP*, 2020, pp. 691-695.

[8] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370-1384, 2020.

[9] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Cernocky, and D. Yu, "Neural target speech extraction: An overview," in *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8-29, 2023.

[10] Q. Wang *et al.*, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, 2019, pp. 2728-2732.

[11] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *IEEE ASRU*, 2019, pp. 327-334.

[12] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE ICASSP*, 2018, pp. 5554-5558.

[13] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *Proc. Interspeech*, 2020, pp. 1406-1410.

[14] S. Mun, S. Choe, J. Huh, and J. S. Chung, "The sound of my voice: Speaker representation loss for target voice separation," in *IEEE ICASSP*, 2020, pp. 7289-7293.

[15] Z. Mu, X. Yang, S. Sun, and Q. Yang, "Self-supervised disentangled representation learning for robust target speech extraction," in *Proc. AAAI*, 2024, pp. 18815-18823.

[16] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for Speakerbeam target speaker extraction," in *IEEE ICASSP*, 2019, pp. 6965-6969.

[17] W. Wang, C. Xu, M. Ge, and H. Li, "Neural speaker extraction with speaker-speech cross-attention network," in *Proc. Interspeech*, 2021, pp. 3535-3539.

[18] K. Liu, Z. Du, X. Wan, and H. Zhou, "X-SEPFORMER: End-to-end speaker extraction network with explicit optimization on speaker confusion," in *IEEE ICASSP*, 2023, pp. 1-5.

[19] F. Hao, X. Li, and C. Zheng, "X-TF-GridNet: A time-frequency domain target speaker extraction network with adaptive speaker embedding fusion," in *Information Fusion*, 102550, 2024.

[20] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Multi-stage speaker extraction with utterance and frame-level reference signals," in *IEEE ICASSP*, 2021, pp. 6109-6113.

[21] Y. Hao, J. Xu, P. Zhang, and B. Xu, "Wase: Learning when to attend for speaker extraction in cocktail party environments," in *IEEE ICASSP*, 2021, pp. 6104-6108.

[22] C. Deng *et al.*, "Robust speaker extraction network based on iterative refined adaptation," in *Proc. Interspeech*, 2021, pp. 3530-3534.

[23] X. Xiao *et al.*, "Single-channel speech extraction using speaker inventory and attention network," in *IEEE ICASSP*, 2019, pp. 86-90.

[24] B. Zeng, S. Hongbin, Y. Wan, and M. Li, "SEF-Net: Speaker embedding free target speaker extraction network," in *Proc. Interspeech*, 2023, pp. 3452-3456.

[25] Y. Hu, H. Xu, Z. Guo, H. Huang, and L. He, "SMMA-Net: An audio clue-based target speaker extraction network with spectrogram matching and mutual attention," in *IEEE ICASSP*, 2024, pp. 1496-1500.

[26] L. Yang, W. Liu, L. Tan, J. Yang, and H.-G. Moon, "Target speaker extraction with ultra-short reference speech by VE-VE framework," in *IEEE ICASSP*, 2023, pp. 1-5.

[27] W. Zhang, L. Yang, and Y. Qian, "Exploring time-frequency domain target speaker extraction for causal and non-causal processing," in *IEEE ASRU*, 2023, pp. 1-6.

[28] X. Yang, C. Bao, J. Zhou, and X. Chen, "Target speaker extraction by directly exploiting contextual information in the time-frequency domain," in *IEEE ICASSP*, 2024, pp. 10476-10480.

[29] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," in *JASA Express Letters*, vol. 1, no. 1, pp. 014802, 2021.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[31] G. Wichern *et al.*, "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019, pp. 1368-1372.

[32] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *IEEE ICASSP*, 2020, pp. 696-700.

[33] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE ICASSP*, 2016, pp. 31-35.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[35] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?," in *IEEE ICASSP*, 2019, pp. 626-630.

[36] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462-1469, 2006.

[37] Y.-X. Lu, Y. Ai, and Z.-H, Ling, "MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *Proc. Interspeech*, 2023, pp. 3834-3838.

GE2E-AC: Generalized End-to-End Loss Training for Accent Classification

Chihiro Watanabe* and Hirokazu Kameoka†
* NTT Communication Science Laboratories, Japan
E-mail: ch.watanabe@ntt.com Tel: +81-7049347331
† NTT Communication Science Laboratories, Japan
E-mail: hirokazu.kameoka@ntt.com Tel: +81-462403645

***Abstract*—Accent classification or AC is a task to predict the accent type of an input utterance, and it can be used as a preliminary step toward accented speech recognition and accent conversion. Existing studies have often achieved such classification by training a neural network model to minimize the classification error of the predicted accent label, which can be obtained as a model output. Since we optimize the entire model only from the perspective of classification loss during training time in this approach, the model might learn to predict the accent type from irrelevant features, such as individual speaker identity, which are not informative during test time. To address this problem, we propose a GE2E-AC, in which we train a model to extract accent embedding or AE of an input utterance such that the AEs of the same accent class get closer, instead of directly minimizing the classification loss. We experimentally show the effectiveness of the proposed GE2E-AC, compared to the baseline model trained with the conventional cross-entropy-based loss[1].**

I. Introduction

Accent is one of the major non-linguistic factors of speech. To alleviate the performance degradation in speech recognition caused by different accents (e.g., English, American, and Scottish), a potential solution is to first classify the accent type of an input utterance and then use an accent-specific speech recognition model that has been trained for the predicted accent [1]–[3]. Such a task to classify an accent type of a given utterance is called accent classification (AC), and various data sets have been publicly available for AC which contain the accent labels of the utterances as well as the speaker ones [4]–[6].

Existing AC models are typically trained to minimize a loss function (e.g., cross-entropy or CE loss) between the target and predicted accent labels [7]–[9]. We will call this approach **CE-AC** (Fig. 1 (a)). While this approach directly optimizes the model's classification accuracy during training, it may lead to predictions that capture irrelevant information (e.g., speaker identity) from the input utterance to reduce the training loss, ultimately degrading test performance. This has also been pointed out in [10]. Moreover, CE-based approach is based on the implicit assumption that the set of potential accent types of input utterances is fixed. To utilize an additional data set with the accent types that have not been in the previous data set, a different model with a new output size must be retrained from scratch.

To address these limitations, we propose adapting the concept of generalized end-to-end (GE2E) loss training [11], which has proven highly effective in speaker verification tasks, for use in the AC task. Specifically, instead of training a model to directly minimize the error in the predicted accent class, we train it to extract accent embeddings (AEs), i.e., feature vectors representing accent information, from input utterances so that the AEs of the same accent class get closer, while those of different accent classes move farther apart. We call this approach **GE2E-AC** (Fig. 1 (b)).

This approach has the advantage that even when a test sample with an unknown accent comes in, it can determine which training sample is most similar. Additionally, it is easy to retrain without needing to modify the model structure. Our experiments show that this method is more effective for the AC task compared to the conventional CE-based baseline method. In the experiment, we also compared the basic GE2E-AC with **GE2E-AC-A**, which incorporates the additional adversarial speaker classification loss to further eliminate the speaker identity information from extracted AEs. Furthermore, since the optimal input for the AC task is not obvious, we also evaluated the performance of using phoneme-related bottleneck features (BNFs) [12] extracted from an encoder-decoder type automatic speech recognition (ASR) system and Hidden-Unit BERT (HuBERT) [13] features as inputs.

II. Baseline method: CE-based accent classification

Before we explain the proposed GE2E-based AC, we first describe a baseline CE-AC. In this approach, we use accent classifier $f_{\mathrm{CE}} : \mathbb{R}^{T\times F} \mapsto \mathbb{R}^{C}$ which maps input feature $X \in \mathbb{R}^{T\times F}$ (e.g., BNF [12] and HuBERT [13]) of an utterance to vector $\hat{\boldsymbol{y}} = f_{\mathrm{CE}}(X) \in [0,1]^{C}$, where T, F, and C indicate the numbers of time and frequency bins of an input feature and accent types, respectively. Throughout this paper, we assume that classifier f_{CE} is modeled as a neural network. By assuming that output vector $\hat{\boldsymbol{y}}$ represents the probabilities of categorical distribution on the accent types, we train the entire network f_{CE} by minimizing the CE loss $\mathcal{L}_{\mathrm{CE}}$ between the model output $\hat{\boldsymbol{y}}$ and the target accent label in one-hot

[1]This work was supported by JST CREST Grant Number JPMJCR19A3, Japan.

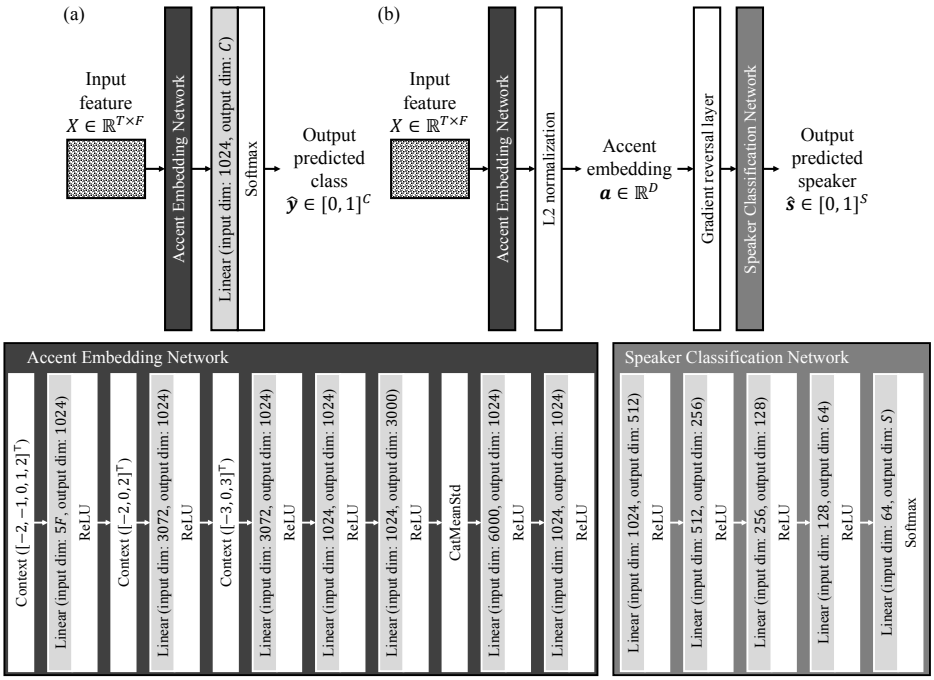


Fig. 1. The network architectures of (a) the baseline CE-based and (b) the proposed GE2E-based accent classification. For arbitrary vector $\boldsymbol{\phi} \in \mathbb{R}^K$, $\mathrm{Context}(\boldsymbol{\phi})$ indicates a function $f_{\boldsymbol{\phi}} : \mathbb{R}^{T\times F'} \mapsto \mathbb{R}^{T\times KF'}$, which maps matrix $X \in \mathbb{R}^{T\times F'}$ to $f_{\boldsymbol{\phi}}(X) = \left[M^{(1)}X, \cdots, M^{(K)}X\right]$, where $M^{(k)} \in \mathbb{R}^{T\times T}$ is a matrix each of whose entries $M^{(k)}_{ij}$ is one if $j = i + \phi_k$ and zero otherwise for all $k \in \{1, \ldots, K\}$. $\mathrm{CatMeanStd}$ is a function $f : \mathbb{R}^{T\times F'} \mapsto \mathbb{R}^{2F'}$, which maps matrix $X \in \mathbb{R}^{T\times F'}$ to $\left[\boldsymbol{\mu}^\mathsf{T}, \boldsymbol{\sigma}^\mathsf{T}\right]^\mathsf{T}$, where $\boldsymbol{\mu} \in \mathbb{R}^{F'}$ and $\boldsymbol{\sigma} \in \mathbb{R}^{F'}$ are the vectors each of whose entries μ_k and σ_k are the mean and the standard deviation of vector $[X_{1k}, \ldots, X_{Tk}]^\mathsf{T}$, respectively.

representation $\boldsymbol{y} \in \left\{\boldsymbol{y}' \in \{0,1\}^C \,|\, \sum_{j=1}^C y'_j = 1\right\}$. For a mini-batch with size B, the CE loss is given by

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{B}\sum_{l=1}^{B}\sum_{j=1}^{C} y_{lj} \log \hat{y}_{lj}, \tag{1}$$

where $\boldsymbol{y}_l = (y_{lj})_{1\le j\le C}$ and $\hat{\boldsymbol{y}}_l = (\hat{y}_{lj})_{1\le j\le C}$ are the target and predicted accent labels of the lth sample in a mini-batch, respectively.

In the experiment in Section IV, we used a classifier based on the architecture in [14], as shown in Fig. 1 (a). This architecture is also used as a sub-model in [15], which achieved the highest accuracy in track 1 (i.e., English accent recognition task) of the accented English speech recognition challenge 2020 (AESRC2020) [10].

III. Proposed method: GE2E-based accent classification

A. GE2E-AC

During training, the baseline method described in Section II focuses on directly minimizing the error in the predicted accent label. However, this training approach may cause the model to rely on irrelevant information, such as speaker identity, to reduce training loss, leading to poor test performance. To prevent this, we propose using a metric learning approach that trains the model to extract AEs from input utterances, ensuring that the AEs of utterances with the same accent class are closer together while those of different accent classes are farther apart. This approach also offers significant advantages in potential applications beyond improving the model's generalization ability. Specifically, the ability to extract AEs allows it to be used for tasks like zero-shot accent conversion, where input speech is converted to match the accent of a reference speech sample, even if the reference accent is unknown.

GE2E loss training is a type of metric learning approach originally used for speaker verification [11]. It has proven highly effective in this task by extracting speaker embeddings from input utterances so that embeddings from the same speaker are closer together, while those from different speakers are farther apart. In our proposed method, we adapt this approach for the AC task, treating accent classes similar to how speaker identities are handled in speaker verification. In the proposed GE2E-AC, by pulling the AEs from the utterances of the same accent type closer together and pushing those from differently accented utterances away, the model would learn to extract an abstract representation of accent, thereby avoiding overfitting to individual speakers. Although there are other possible methods of deep metric learning such as Siamese network [16], we employ GE2E loss here, since its effectiveness has already been shown in another speech processing task [11].

The actual flow of GE2E-AC is as follows (Fig. 1 (b)). We first extract a D-dimensional AE $\boldsymbol{a} = f_{\mathrm{AE}}(X) \in \mathbb{R}^D$ from

input feature X by using an accent embedding network $f_{\mathrm{AE}} : \mathbb{R}^{T\times F} \mapsto \mathbb{R}^D$. During training, we define a mini-batch of input features as a set of M utterances for each of C accent types. Let $\boldsymbol{a}_{ji} \in \mathbb{R}^D$ be the AE of the ith utterance of the jth accent for $i = 1, \ldots, M$, and $j = 1, \ldots, C$. As shown in Fig. 1 (b), we define an AE such that it has a unit L2 norm (i.e., $\|\boldsymbol{a}_{ji}\|_2 = 1$). Based on these AEs, we can compute two types of centroids for each accent class j, which are given by

$$\boldsymbol{c}_j = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{a}_{ji}, \quad \boldsymbol{c}_j^{(-i)} = \frac{1}{M-1}\sum_{\substack{m=1\\m\neq i}}^{M} \boldsymbol{a}_{ji}. \tag{2}$$

We also define the L2 normalization of the centroids by

$$\bar{\boldsymbol{c}}_j = \frac{\boldsymbol{c}_j}{\|\boldsymbol{c}_j\|_2}, \quad \bar{\boldsymbol{c}}_j^{(-i)} = \frac{\boldsymbol{c}_j^{(-i)}}{\|\boldsymbol{c}_j^{(-i)}\|_2}. \tag{3}$$

By using these L2 normalized centroids, we define $MC \times C$ similarity matrix $S = (S_{lk})_{1\leq l\leq MC, 1\leq k\leq C}$, whose (l,k)th entry is given by

$$S_{lk} = \begin{cases} w\boldsymbol{a}_{ji}\cdot\bar{\boldsymbol{c}}_k^{(-i)} + b & j = k, \\ w\boldsymbol{a}_{ji}\cdot\bar{\boldsymbol{c}}_k + b & \text{otherwise}, \end{cases} \tag{4}$$

where w and b are learnable weight and bias parameters for the similarity metric and $l = (j-1)M + i$ for $i = 1, \ldots, M$ and $j = 1, \ldots, C$. As in [11], we define each entry S_{lk} differently in the cases of $j = k$ and $j \neq k$ for training stability.

Based on similarity matrix S in (4), we train the entire network to minimize the GE2E loss, which is given by

$$\mathcal{L}_{\mathrm{GE2E}} = -\frac{1}{MC}\sum_{l=1}^{MC}\sum_{k=1}^{C} y_{lk}\log\hat{y}_{lk}, \tag{5}$$

where $\hat{y}_{lk} = \exp(S_{lk})/\{\sum_{k'=1}^{C}\exp(S_{lk'})\}$. In regard to the final formulation, the GE2E loss in (5) looks similar to the CE loss in (1). However, unlike in the CE loss, the predicted accent labels $\{\hat{y}_{lk}\}$ in the GE2E loss are computed based on the similarity $\{S_{lk}\}$ between the AEs and their centroids. To minimize the GE2E loss, the accent embedding network should learn to pull the AEs closer to the centroid of its accent type and to push them away from the other centroids. This facilitates the AEs of the same accent type getting closer together, even if they are uttered by mutually different speakers, which would improve the generalization performance in AC.

In the experiment in Section IV, to evaluate the effect of introducing the GE2E loss instead of the CE one, we set the shallower part of the model (i.e., "Accent Embedding Network" in Fig. 1) to be the same as that of CE-AC. In the standard GE2E-AC, we used only the shallower part than the gradient reverse layer.

B. GE2E-AC-A

Although introduction of GE2E loss alone would improve the performance of AC during test time, AEs of GE2E-AC might still contain some information related to speaker identity, since there are no samples of different accents from the same speaker. To further eliminate speaker-related information from AEs and improve the generalization performance, we also propose GE2E-AC-A, in which we incorporate an adversarial speaker classification or SC loss $\mathcal{L}_{\mathrm{SC}}$ into the training criterion. We consider such a loss to avoid mixing speaker information in AEs. It must be noted that a similar technique has also been used in [17], although our main contribution lies in introduction of GE2E loss for AEs, rather than in that of SC loss.

Specifically, the SC loss can be computed by using another network $f_{\mathrm{SC}} : \mathbb{R}^D \mapsto [0,1]^S$ (i.e., "Speaker Classification Network" in Fig. 1), which predicts a speaker index from an AE. Let $\boldsymbol{s} \in \left\{\boldsymbol{s}' \in \{0,1\}^S \middle| \sum_{j=1}^{S} s'_j = 1\right\}$ be the target speaker label in one-hot representation and let $\hat{\boldsymbol{s}} = f_{\mathrm{SC}}(\boldsymbol{a}) \in [0,1]^S$ be the predicted one. The SC loss is given by

$$\mathcal{L}_{\mathrm{SC}} = -\frac{1}{B}\sum_{l=1}^{B}\sum_{j=1}^{S} s_{lj}\log\hat{s}_{lj}, \tag{6}$$

where $\boldsymbol{s}_l = (s_{lj})_{1\leq j\leq S}$ and $\hat{\boldsymbol{s}}_l = (\hat{s}_{lj})_{1\leq j\leq S}$ are the target and predicted speaker labels of the lth sample in a mini-batch, respectively.

The training strategy of GE2E-AC-A is two-fold. The speaker classification network should be trained to minimize SC loss $\mathcal{L}_{\mathrm{SC}}$, while the accent embedding network should be trained to maximize it. We can implement such an adversarial learning by adding a gradient reversal layer (GRL) [18] before the speaker classification network. The GRL is an identity function in forward propagation, while it multiplies the gradient from the deeper layers by -1 in backpropagation. By introducing GRL, we can formulate the overall training criterion of GE2E-AC-A to be minimized by $\mathcal{L}_{\mathrm{GE2E}} + \lambda_{\mathrm{SC}}\mathcal{L}_{\mathrm{SC}}$, where λ_{SC} is a hyperparameter. As GE2E-AC, we used a network architecture in Fig. 1 (b) in the experiment in Section IV.

IV. Experiment

To confirm the effectiveness of the proposed GE2E-AC, we conducted an experiment using CSTR VCTK Corpus [19]. Among this data set, we only used the samples of accent types of "English," "American," "Scottish," "Irish," and "Canadian," which consist of 33, 22, 19, 9, and 8 speakers, respectively. For each accent type, we randomly chose five test speakers and used the other speakers for training. We trimmed leading and trailing silence of each sample and resampled it to 16 KHz.

As input features, we adopted BNF [12] and HuBERT [13]. We extracted these features from WAV samples by using the codes provided by [20], [21] and trained each model of CE-AC (baseline), GE2E-AC, and GE2E-AC-A with them.

The hyperparameters for training the models were as follows. For all the models, we used Adam [22] with the learning rate of 1.0×10^{-5} for optimization and set the minibatch size to 16, the number of epochs to 100, λ_{SC} to 1.0×10^{-5}, and the number M of utterances per accent type to 10. It must be noted that, since a mini-batch of GE2E-AC and GE2E-AC-A consists of the same number M of utterances for all the

accents with generally different sample sizes $\{n_1, \ldots, n_C\}$, we could not determine the meaning of one epoch in the usual sense. Therefore, for these methods, we defined the nominal "sample size" for a sample which includes one input feature for each accent as the maximum sample size $\max_{j=1,\ldots,C} n_j$ in the experiment. We clipped the max L2 norm of the gradients of the model parameters to 1.0 for CE-AC and 3.0 for GE2E-AC and GE2E-AC-A. Finally, we computed confusion matrices by using the trained models. The (i, j)th entry of a confusion matrix indicates the number of training or test samples whose target accent index is i and whose predicted counterpart is j. Since outputs of the GE2E-based models are accent embeddings $\boldsymbol{a}$, not predicted accent classes $\hat{\boldsymbol{y}}$, we need to determine $\hat{\boldsymbol{y}}$ from $\boldsymbol{a}$ to compute a confusion matrix. Based on the accent embeddings of all the training samples, we first computed the centroid (i.e., mean of embeddings) of each accent class. We defined the predicted accent class of a given sample with accent embedding $\boldsymbol{a}'$ as the class whose centroid had the highest cosine similarity with $\boldsymbol{a}'$.

Tables I, II, and III show the classification accuracy of each method in the cases of three, four, and five accent types, respectively. Aside from the case of using HuBERT for classifying four accent types, the proposed GE2E-AC or GE2E-AC-A achieved better test performance than that of the CE-based baseline method. In some cases (e.g., using HuBERT for classifying four accent types), the adversarial speaker classification loss in combination with the GE2E loss was effective, while in other cases it was not. In regard to the input feature type, HuBERT yielded better test performance than BNF did in most cases. As more detailed results, Figs. 2, 3, and 4 show the confusion matrices in the cases of three, four, and five accent types, respectively. Although there were slight differences in the values of the confusion matrices, they showed similar classification tendency to each other among all the settings. For instance, when the number of accent types was five, it was hard for all the models to correctly classify the Canadian speakers, and most of the utterances by Canadian speakers were classified as "American." This would be probably due to the bias in sample size, and to mitigate such a problem would be an important direction of a future study.

TABLE I
ACCURACY OF ACCENT RECOGNITION FOR EACH SETTING WITH **THREE** ACCENT TYPES.

Method (Feature type)	Training accuracy	Test accuracy
CE-AC (BNF)	99.9%	78.8%
GE2E-AC (BNF)	100.0%	**81.2%**
GE2E-AC-A (BNF)	99.7%	77.8%
CE-AC (HuBERT)	98.7%	81.9%
GE2E-AC (HuBERT)	99.9%	85.0%
GE2E-AC-A (HuBERT)	100.0%	**85.2%**

TABLE II
ACCURACY OF ACCENT RECOGNITION FOR EACH SETTING WITH **FOUR** ACCENT TYPES.

Method (Feature type)	Training accuracy	Test accuracy
CE-AC (BNF)	99.8%	59.2%
GE2E-AC (BNF)	100.0%	**63.8%**
GE2E-AC-A (BNF)	100.0%	62.1%
CE-AC (HuBERT)	99.7%	**65.9%**
GE2E-AC (HuBERT)	100.0%	63.5%
GE2E-AC-A (HuBERT)	99.6%	64.9%

TABLE III
ACCURACY OF ACCENT RECOGNITION FOR EACH SETTING WITH **FIVE** ACCENT TYPES.

Method (Feature type)	Training accuracy	Test accuracy
CE-AC (BNF)	99.2%	46.6%
GE2E-AC (BNF)	100.0%	**49.7%**
GE2E-AC-A (BNF)	100.0%	49.6%
CE-AC (HuBERT)	99.4%	50.9%
GE2E-AC (HuBERT)	100.0%	52.8%
GE2E-AC-A (HuBERT)	100.0%	**53.4%**

V. CONCLUSION

In this study, we proposed a new training method for AC based on the similarity between a pair of accent embeddings of the utterances so that the model can avoid overfitting to specific speakers. The effectiveness of the proposed GE2E-AC has been shown experimentally, compared to the existing CE-based method in which we directly minimize the classification loss for individual utterances. A possible future direction of this study would be to utilize the proposed method as a preliminary step toward accent conversion, where we aim to convert the accent of an input utterance to some different one. By extracting an AE from a target sample with a trained GE2E-AC model and using it for test-time accent conversion, it would be possible to convert accents in zero-shot manner.

REFERENCES

[1] V. Gupta and P. Mermelstein, "Effects of speaker accent on the performance of a speaker-independent, isolated-word recognizer," *The Journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1581–1587, 1982.

[2] L. M. Arslan and J. H. L. Hansen, "Language accent classification in american english," *Speech Communication*, vol. 18, no. 4, pp. 353–367, 1996.

[3] A. Faria, "Accent classification for speech recognition," in *International Workshop on Machine Learning for Multimodal Interaction*, 2006, pp. 285–293.

[4] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source multi-speaker corpora of the English accents in the British isles," in *Language Resources and Evaluation Conference (LREC)*, 2020, pp. 6532–6541.

[5] R. Sanabria, N. Bogoychev, N. Markl, A. Carmantini, and P. B. Ondrej Klejch and, "The Edinburgh international accents of English corpus: Towards the democratization of English ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] W. Wang, Y. Song, and S. Jha, "GLOBE: A high-quality English corpus with global accents for zero-shot speaker adaptive text-to-speech," arXiv:2406.14875, 2024.

[7] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning," in *Interspeech*, 2018, pp. 2454–2458.

[8] M. Lesnichaia, V. Mikhailava, N. Bogach, I. Lezhenin, J. Blake, and E. Pyshkin, "Classification of accented English using CNN model trained on amplitude mel-spectrograms," in *Interspeech*, 2022, pp. 3669–3673.

[9] V. Nechaev and S. Kosyakov, "Non-autoregressive real-time accent conversion model with voice cloning," arXiv:2405.13162, 2024.

[10] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, "The accented English speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6918–6922.

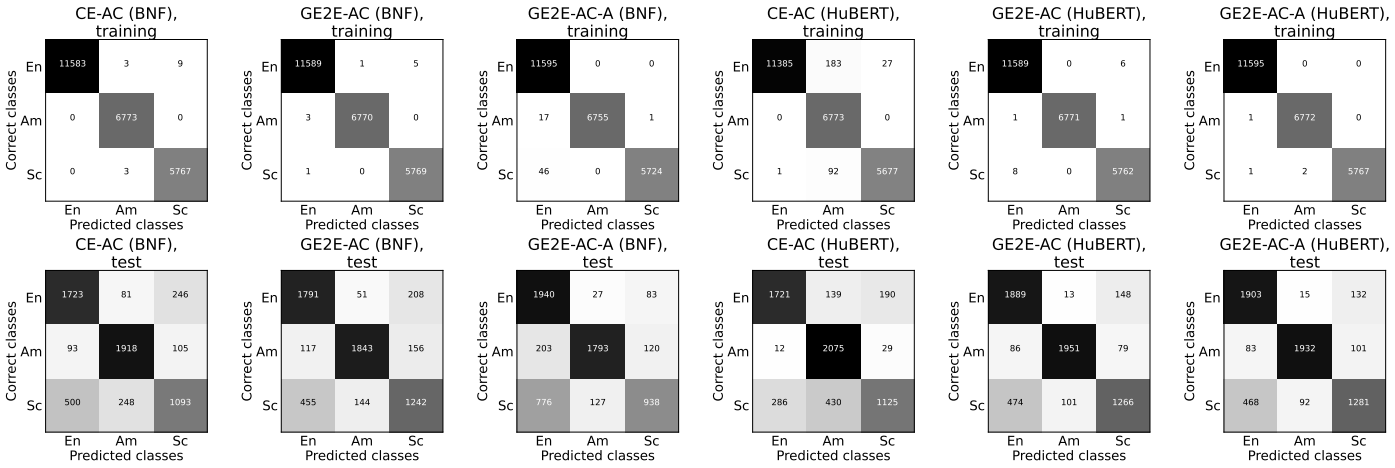


Fig. 2. Confusion matrices with **three** accent types. The value plotted in each entry indicates the number of corresponding samples. "En," "Am," and "Sc" stand for English, American, and Scottish, respectively.

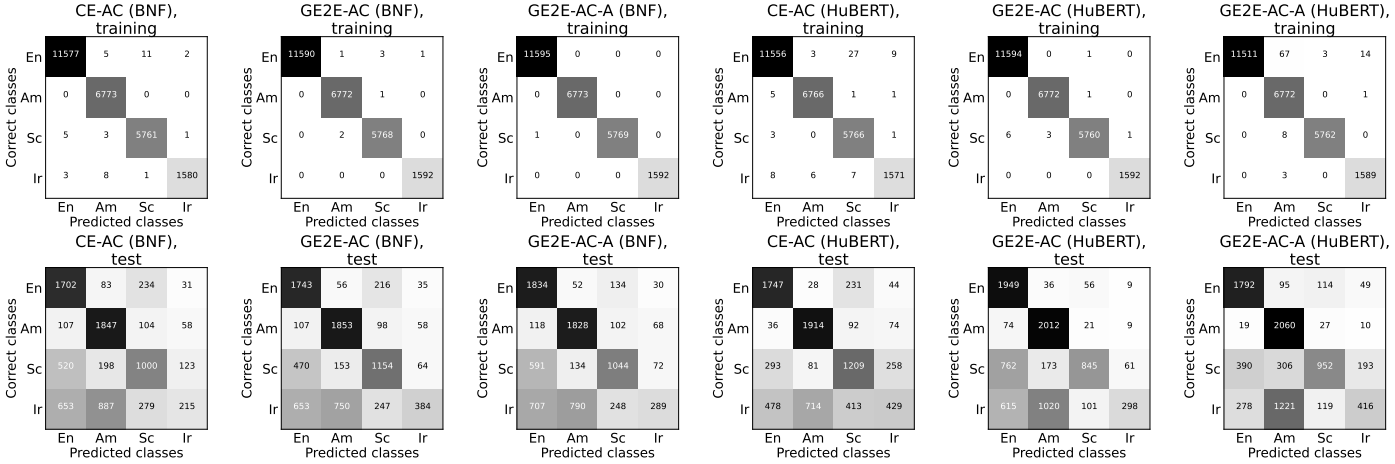


Fig. 3. Confusion matrices with **four** accent types. "Ir" stands for Irish.

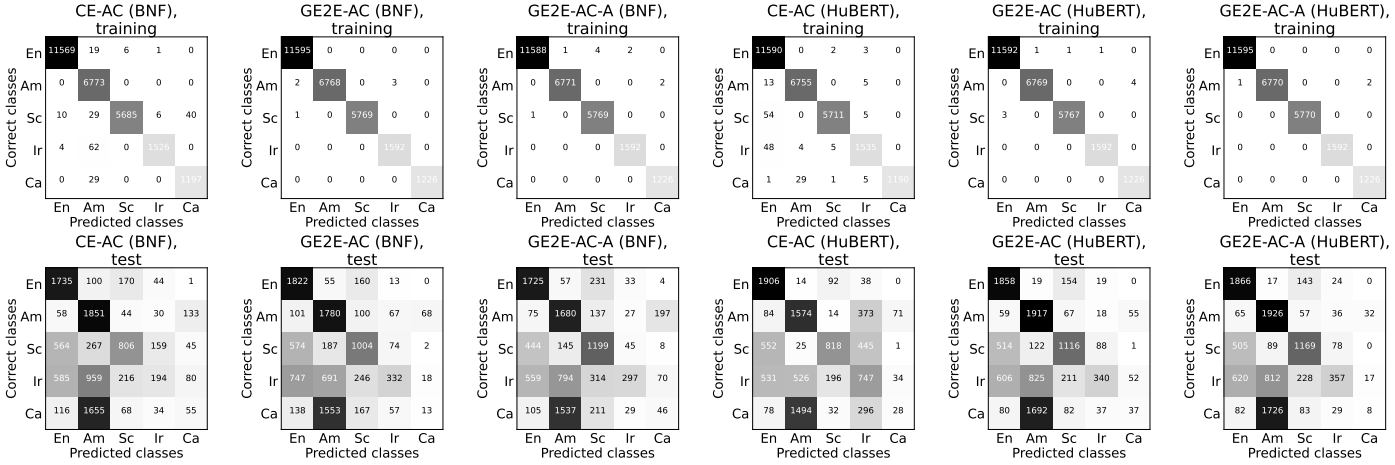


Fig. 4. Confusion matrices with **five** accent types. "Ca" stands for Canadian.

[11] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.

[12] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[14] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[15] H. Huang, X. Xiang, Y. Yang, R. Ma, and Y. Qian, "AISpeech-SJTU accent identification system for the accented English speech recognition challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6254–6258.

[16] A. Siddhant, P. Jyothi, and S. Ganapathy, "Leveraging native language speech for accent identification using deep Siamese networks," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 621–628.

[17] X. Zhou, M. Zhang, Y. Zhou, Z. Wu, and H. Li, "Multi-scale accent modeling with disentangling for multi-speaker multi-accent TTS synthesis," arXiv:2406.10844, 2024.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[19] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017.

[20] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021, https://github.com/liusongxiang/ppg-vc.

[21] "Hubert," Hugging Face, https://huggingface.co/docs/transformers/model_doc/hubert.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations*, 2015.

Efficient Feature Selection for Word Embedding Dimension Reduction

Jintang Xue, Yun-Cheng Wang, Chengwei Wei and C.-C. Jay Kuo
University of Southern California, Los Angeles, CA, USA

***Abstract*—Word embeddings transform each word into a vector space representation, making it possible for computers to process and analyze the human language. However, storing and processing these word vectors can become resource-intensive as the vector space dimension increases and the vocabulary expands. It poses a challenge, especially for mobile edge-device applications. This paper explores dimension reduction for word embedding to balance computational costs and performance. We propose an efficient weakly-supervised feature selection method called WordFS, which has two variants utilizing novel criteria for feature selection. Our experiments on various tasks, such as word similarity and binary and multi-class classification, show that WordFS outperforms other dimension reduction methods while requiring lower computational costs.**

I. Introduction

In recent years, large language models (LLMs) have made a breakthrough in natural language processing (NLP). These models have revolutionized the understanding and generation of human language tasks such as question-answering, machine translation, and text summarization [1], [2], etc. As a fundamental component in language models, word embedding represents words as vectors in a continuous, high-dimensional space [3]. Word vectors capture semantic and syntactic meanings of words, providing word relationships for various downstream tasks. Static word embedding methods assign a fixed vector to each word. They convert input words to corresponding vectors for further processing and model training. Contextual word embedding methods leverage deep-learning models to generate word vectors based on the context. A word can have a different contextual word embedding in a different sentence, allowing the model to capture subtle differences in the meaning of the same word in various contexts.

A common challenge for static and contextual word embedding methods is that the high dimension of a word vector leads to an enormous model size. Typically, a word vector has hundreds to thousands of dimensions. For instance, loading a 300-dimensional word embedding matrix with 2.5 million tokens would require up to 6 GB of memory on a 64-bit system [4]. On the one hand, high-dimensional word vectors provide a good representation of complex human language, which is crucial for performing downstream tasks. On the other hand, high-dimensional word vectors have higher demands on computational resources and memory requirements. Thus, dimension reduction is critical in the application of word vectors.

Existing dimension reduction methods mainly consist of traditional PCA-based and deep-learning-based models. Traditional unsupervised PCA-based methods are known for their efficiency and interpretability. In contrast, supervised deep-learning-based methods are inefficient and lack interpretability. Semi-supervised feature selection methods can also be used for dimension reduction. Since word vectors can be viewed as extracted features for each word, feature selection methods have the potential to provide a straightforward yet effective way to reduce the word dimension. It could help balance unsupervised PCA-based methods and resource-intensive deep-learning-based methods.

This paper investigates using a small subset of labeled word similarity pairs to develop a weakly supervised dimension reduction method called WordFS (i.e., word dimension reduction with **F**eature **S**election). We demonstrate that one can achieve dimension reduction effectively by supervising a limited number of word similarity pairs. Note that word similarity is typically used in evaluation tasks but is rare in word embedding dimension reduction, a key novelty of this work. The proposed WordFS method consists of three stages: 1) post-processing, 2) feature extraction, and 3) weakly-supervised feature selection. We apply WordFS to other downstream tasks to show its generalizability. Experimental results show that WordFS outperforms existing methods in word similarity and various tasks while achieving much lower computational costs.

This work has the following significant contributions.

- We propose a novel, effective, and efficient dimension reduction method called WordFS for word embeddings from the perspective of feature selection based on weakly-supervised learning.
- We demonstrate the potential of combining feature selection methods and word similarity for word embedding dimension reduction.
- We show the effectiveness and efficiency of our approach on various downstream tasks, including binary and multi-class classification tasks. Our method generally outperforms the existing techniques while being more straightforward.

The rest of the paper is organized as follows. Related work is reviewed in Sec. II. The proposed WordFS method is described in Sec. III. Experimental results are shown in Sec. IV. Finally, concluding remarks and future extensions are given in Sec. V.

II. Related work

Word embedding compression is an essential topic for storing and processing word embeddings, especially on computationally limited devices. Existing dimension reduction methods mainly consist of traditional and deep learning-based models.

A. Matrix Decomposition Techniques

Matrix decomposition techniques, such as singular value decomposition (SVD), principal components analysis (PCA), non-negative matrix factorization (NMF), and factor analysis (FA), have been applied to dimension reduction of word embeddings. While simple and efficient, these methods are not highly effective. Post-processing methods help enhance word embedding in dimension reduction. One can achieve lower dimensional word embeddings by combining post-processing algorithms (PPAs) to reduce the anisotropy [5] with PCA [4]. The dimension-reduced vectors can perform similarly or even better than the original pre-trained word embeddings, outperforming most unsupervised methods.

B. Deep Learning Methods

Deep learning-based models have recently become quite popular, and efforts have been made to explore their potential in reducing word embedding dimensions. EmbedTextNet [6] achieves better performance in low-dimensional embedding sizes by utilizing a VAE model with a correlation penalty added to the weighted reconstruction loss. However, deep learning models typically require more computational resources and longer training time, contradicting the original goal of dimension reduction.

C. Feature Selection Methods

Feature selection is another simple and intuitive method for removing redundant features. Feature selection methods can be applied to bag-of-words representations, where each unique word is treated as a distinct feature, for sentiment classification [7] tasks to reduce the number of terms and save storage and memory space. Also, feature selection methods can be applied for text classification to select the most representative word [8]. However, the role of feature selection for word embedding dimension reduction is under-explored. This paper shows that a straightforward and intuitive feature selection method can effectively reduce dimension. We believe applying the feature selection method for word embedding dimension reduction is novel.

III. Proposed Method

A. System Overview

Our WordFS method consists of three stages: 1) post-processing, 2) feature extraction, and 3) feature selection. We adopt word similarity datasets as the source of supervision to guide the feature selection in our model. As a widely-used benchmark of word embeddings, the word similarity task evaluates the quality of word embeddings in representing semantic and syntactic meanings by measuring how closely the representation of word vectors matches human perception of similarity [9], making them good supervision resources for word embedding dimension reduction. Also, since word similarity captures fundamental characteristics of word embeddings, which are crucial for many NLP tasks, its potential for generalizing to downstream tasks is noteworthy.

In our method, We begin by optionally applying a post-processing method to our pre-trained word embeddings. We then extract pair-wise features for each word pair by conducting element-wise production with normalization. Then, we adopt two different criteria for feature selection to evaluate each dimension and identify essential dimensions. The first one is based on a supervised feature selection method called RFT [10]. The second is based on Spearman's rank correlation coefficient between each feature dimension and word similarity scores. Finally, the selected dimensions of the word embeddings are kept, and other dimensions are discarded according to the requirements for dimension reduction. The details of each stage are elaborated below.

B. Post-processing

To improve the quality of word embedding in downstream applications, performing post-processing techniques on pre-trained word embeddings can be crucial. The post-processing algorithm (PPA) [5] is an effective PCA-based post-processing method to improve the isotropy of word representations by removing the top principal components of all words. Specifically, given a word vector matrix M with n word vectors and a threshold D, mean vectors are subtracted first:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} M_i, \quad M_i = M_i - \mu, \quad \forall i = 1, 2, \ldots, n. \tag{1}$$

Then, the principal components are computed:

$$U = \text{PCA}(M) \tag{2}$$

Finally, for each word vector $v \in M$, the top D principal components are removed:

$$v' = v - \sum_{i=1}^{D} (u_i^\top v) u_i, \quad \forall v \in M. \tag{3}$$

The output is the enhanced word vector matrix M'. The Algo method [4] applies PPA before and after dimension reduction to enhance both the original and the dimension-reduced word vectors. In our method, we only apply PPA before dimension reduction, which makes our method simpler. Applying PPA to the selected subset after feature selection may disrupt the previous results because PCA and the feature selection criteria can have different objectives. Specifically, PCA transforms features to maximize variance, while feature selection methods usually select features based on their correlation with labels. Also, the selected feature subset may lose crucial information for PCA to find valuable principal components. As an enhancement of original word embeddings before dimension reduction, we also make PPA an optional choice. The reason is that anisotropy may not always be harmful [11], and using PPA depends on the pre-trained word embeddings and specific application

scenarios. Specifically, word embeddings trained on less data may be vulnerable to noise. Applying PPA helps mitigate some of the noise and offers better performance. On the other hand, using PPA may result in the loss of crucial and intricate information related to word representation, which can harm performance. Word embeddings trained on relatively small data benefit from uniformly distributed word embeddings, while complex contextual tasks may require anisotropy to capture detailed information.

C. Feature Extraction

In this stage, we extract suitable features for feature selection based on the dataset that provides supervision. In this paper, we leverage word similarity data as the weak supervision of our model. Since the word similarity is based on the correlation (i.e., word similarity score) between every two words, we extract pair-wise features for each word similarity pair.

The pair-wise features are constructed in each dimension based on cosine similarity, an effective measure for evaluating the similarity between word embeddings. Cosine similarity between two-word vectors is derived by dividing their dot product by the product of their magnitudes:

$$\begin{aligned} \text{Sim}(\mathbf{a}, \mathbf{b}) &= \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \\ &= \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}}, \end{aligned} \tag{4}$$

where $\mathbf{a}$ and $\mathbf{b}$ are embeddings of the two words. The cosine similarity score can be viewed as the sum of the normalized results of element-wise products. Each pair of elements (i.e., each dimension) has a linear impact on the cosine similarity score of that pair of word vectors. Therefore, the normalized results of element-wise products can be viewed as features to predict the similarity score. The feature we use for each dimension is:

$$f_i = \frac{a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}}, \tag{5}$$

where f_i is the feature of the ith dimension extracted from the two word embeddings.

D. Feature Selection

The method for the feature selection process depends on the objective of downstream tasks. Specifically, we employ two feature selection methods for prediction and similarity tasks. These two feature selection methods are developed into our model's variants, WordFS-S and WordFS-P. We focus on filter feature selection techniques since they are simple and computationally efficient. Although different dimensions of distributed word embeddings may capture nuanced and interdependent semantic relationships, they allow us to identify dimensions that contribute most significantly to model performance. Besides, filter feature selection techniques select features based on their correlation with the labels and are independent of the classifier. As a result, the selected features are generalizable.

Prediction tasks: For prediction tasks, we leverage the methods from Relevant Feature Test (RFT) [10] to build our WordFS-P model. RFT is a filter feature selection technique that was proposed recently for regression tasks. It has been widely used in green learning architectures [13] for feature dimension reduction to reach a smaller model size. In our experiments, RFT is utilized for feature selection, and we adopt the word similarity scores as the labels. RFT partitions each feature dimension into two subintervals and calculates the overall mean square error (MSE) as the loss function. A smaller loss function means a better feature dimension. Specifically, given a feature of the ith dimension f_i, the feature space is partitioned into $B = 2^k, k = 1, 2, ...$ uniform segments and the optimal partition threshold is searched among $B - 1$ candidates in the range $[\min(f_i), \max(f_i)]$:

$$f_b^i = \min(f_i) + \frac{b}{B} \left[\max(f_i) - \min(f_i)\right], \tag{6}$$

where $b = 1, 2, ...B - 1$. The threshold t partitions the ith feature space into the left subset $S_{L,t}^i$ and the right subset $S_{R,t}^i$. The MSEs for regression $R_{L,t}^i$ and $R_{R,t}^i$ are separately calculated in the two subsets by using the mean of target values as the estimated regression value of all samples. Then, the RFT loss is defined as the weighted sum of the two MSEs:

$$R_t^i = \frac{N_{L,t}^i R_{L,t}^i + N_{R,t}^i R_{R,t}^i}{N}, \tag{7}$$

where $N_{L,t}^i$ and $N_{R,t}^i$ denote the number of samples in each subset and N is the total number of samples. The RFT loss of each feature dimension is defined as the minimal RFT loss among all the candidate thresholds:

$$R_i = \min_{t \in T} R_t^i. \tag{8}$$

Finally, all feature dimensions' RFT loss is ranked in ascending order, and the top K dimensions are selected, where K is the dimension of the word vectors after our dimension reduction approach.

Similarity tasks: Similarity tasks are usually done by calculating the cosine similarity between target vectors, which differs from prediction tasks. The cost function of DFT/RFT may not match well with the evaluation criteria of similarity tasks. Inspired by the evaluation of word similarity, we use Spearman's rank correlation coefficient to develop our WordFS-S model. It involves finding the correlation between the ranks generated by features extracted from each dimension and the target labels. First, the ranked values of each feature dimension and the labels are calculated. Second, Spearman's rank correlation coefficient for each feature dimension is calculated by the Pearson correlation coefficient between the ranked values of each dimension and the labels:

$$r_i = \rho_{R(f_i),R(y)} = \frac{\text{cov}(R(f_i), R(y))}{\sigma_{R(f_i)} \sigma_{R(y)}}, \tag{9}$$

TABLE I
COMPARISON OF SPEARMAN'S RANK CORRELATION COEFFICIENT ACROSS MULTIPLE WORD SIMILARITY DATASETS, WHERE **BOLDFACE** INDICATES THE BEST VALUE IN EACH COLUMN, AND UNDERLINE INDICATES THE SECOND BEST VALUE IN EACH COLUMN.

Dataset	MC -30	Men -TR -3k	M Turk -287	M Turk -771	RG -65	RW Stan ford	Sim Lex -999	VE RB -143	WS -353 -ALL	WS -353 -REL	WS -353 -SIM	YP -130	Avg
Glove-300D [12]	64.03	73.71	62.44	64.60	72.97	41.16	36.97	29.97	60.27	55.44	64.74	55.15	56.79
Algo-150D [4]	67.97	75.21	62.47	64.11	72.94	43.45	37.81	**38.06**	66.71	60.55	70.74	**58.09**	59.84
EmbedTextNet-150D [6]	**79.59**	75.43	58.36	62.91	**80.00**	41.31	37.49	27.33	63.06	55.35	67.50	55.80	58.68
WordFS-P-woP-150D (ours)	66.79	74.94	60.68	63.44	69.83	36.16	37.93	30.26	62.49	53.69	66.78	52.27	56.27
WordFS-P-wP-150D (ours)	73.63	75.38	64.38	64.24	70.37	45.01	42.47	34.88	**68.53**	**60.58**	70.25	50.54	60.02
WordFS-S-woP-150D (ours)	67.79	75.78	62.62	62.95	70.26	36.07	44.72	36.61	64.72	55.36	69.42	45.76	57.67
WordFS-S-wP-150D (ours)	72.27	**75.96**	**65.57**	**64.35**	70.69	**45.22**	**45.08**	36.03	67.16	59.18	**71.00**	52.90	**60.45**

TABLE II
PERFORMANCE COMPARISON OF SPEARMAN'S RANK CORRELATION COEFFICIENTS ON THE AGGREGATED WORD SIMILARITY DATASET.

Dimension	300	150	100	50
Glove [12]	45.74	–	38.45	36.24
Algo [4]	–	53.48	49.35	42.61
EmbedTextNet [6]	–	52.71	43.88	41.00
WordFS-P-woP (ours)	–	48.23	48.39	48.22
WordFS-P-wP (ours)	–	54.69	52.60	47.29
WordFS-S-woP (ours)	–	55.85	54.45	49.32
WordFS-S-wP (ours)	–	**56.76**	**54.73**	**50.11**

where $R(\cdot)$ represents the rank of given raw scores, $\text{cov}(\cdot)$ represents the covariance, and σ represents the standard deviation. The higher the correlation coefficient, the better the feature dimension is. Finally, the Spearman's rank correlation coefficients of all the feature dimensions are then ranked in descending order, and the top K dimensions are selected.

IV. EXPERIMENTS

We evaluate our method by applying it to pre-trained word embeddings. The word similarity datasets provide weak supervision to guide our feature selection module. The dimension-reduced word embeddings are used for word similarity tasks and downstream tasks. Then, we compare the results of our method with the original pre-trained word embeddings, the dimension-reduced word embeddings from the PCA-based method called Algo [4], and the results from the deep learning-based method called EmbedTextNet [6].

A. Pre-trained Word Embeddings

In our experiments, we use the Glove word embeddings [12] trained on Wikipedia 2014 and Gigaword 5 corpus (6B tokens, 400K vocabulary). They are available in multiple dimensions, precisely 50, 100, 200, and 300.

B. Word Similarity Datasets

We use the widely-used word similarity datasets [14] to evaluate our method. The word similarity datasets contain word pairs and their corresponding scores from human annotators based on perceived relatedness or similarity. The cosine similarity of each pair of words calculates the similarity score from the word embeddings. We use Spearman's rank correlation coefficient as our evaluation metric. It measures how closely the ranking derived from the cosine similarities of given word vectors matches those based on human judgments. A higher value of this metric indicates a better match to human-labeled similarity rankings.

We first evaluate our methods on twelve word similarity datasets. Since we choose word similarity as the guidance for feature selection, we apply 5-fold cross-validation to each word similarity dataset to train and test our method. The reported results are the average Spearman's rank correlation coefficients from the five folds. We evaluate the pre-trained word embeddings and all the methods on the same cross-validation fold and take the average for a fair comparison. All the results we report average over five different cross-validation trials in all the experiments. We set the number of bins in RFT to 4, the default and recommended value from the RFT model, and the threshold in the PPA to 7, the same as in the Algo model.

Table I shows the results across twelve word similarity datasets of different dimension reduction methods reduced from 300-dimensional Glove word embeddings to 150 dimensions. In the table, WordFS-P and WordFS-S represent our proposed methods based on RFT and similarity feature selection methods, respectively. P represents PPA, and wP and woP represent PPA and without PPA. The last column is the average of the previous columns. For the dimension-reduced vectors, our methods outperform the original word embeddings and other methods in most datasets. Our WordFS-S-wP method achieves the highest average score. Both feature selection-based approaches perform better than the existing methods. As expected, the WordFS-S-wP method performs better than the WordFS-P-wP method in this task. Our WordFS-S-wP method outperforms the best performance from existing methods with an average improvement of 0.61 in terms of Spearman's rank correlation coefficients when reducing to 150 dimensions. Our method can perform better than the original 300-dimensional word embeddings when reduced to 150 dimensions.

Then, we aggregated the twelve word similarity datasets. We first scaled them into the same range [0, 1] and then averaged the similarity scores of overlapped word pairs. There are two reasons to aggregate the datasets: 1) to provide an overall evaluation for the word similarity tasks and 2) to construct a comprehensive dataset to train our feature selection methods for various downstream tasks. We refer to our model as a weakly-supervised method for downstream tasks because there

TABLE III
COMPARISON OF THE RESULTS OF DOWNSTREAM PREDICTION TASKS. THE LAST COLUMN IS THE AVERAGE OF THE PREVIOUS COLUMNS.

Task	MR	CR	MPQA	SUBJ	STS-B	SST -FG	TREC	MRPC	SICK-E	Avg
Glove-300D [12]	74.99	75.81	86.75	91.39	78.20	40.77	66.6	75.28	77.19	73.81
Algo-150D [4]	73.70	75.32	85.31	89.59	**76.66**	40.95	60.7	**71.68**	76.49	72.27
EmbedTextNet-150D [6]	72.67	74.09	84.95	**90.08**	73.75	40.50	64.8	71.54	74.43	71.87
WordFS-P-woP-150D (ours)	73.09	74.86	**86.24**	89.86	76.22	39.55	**65.4**	71.01	**76.62**	72.54
WordFS-P-wP-150D (ours)	**74.03**	**76.00**	**86.24**	89.20	76.50	**41.27**	65.2	70.38	76.46	**72.81**
Algo-100D [4]	70.61	72.82	82.71	88.16	74.63	38.28	56.2	**71.83**	75.36	70.07
EmbedTextNet-100D [6]	71.55	72.87	84.44	**88.88**	70.18	38.46	62.2	71.48	75.14	70.58
WordFS-P-woP-100D (ours)	72.24	73.96	85.26	88.62	73.97	38.37	62.0	71.42	**76.60**	71.38
WordFS-P-wP-100D (ours)	**72.68**	**75.42**	**85.50**	88.25	**76.94**	**40.50**	**64.4**	71.59	74.87	**72.24**
Algo-50D [4]	66.24	70.23	76.35	84.32	69.36	35.88	48.1	71.39	72.98	66.09
EmbedTextNet-50D [6]	69.36	**72.34**	82.79	**87.65**	71.66	**38.64**	**60.2**	68.29	**74.20**	**69.45**
WordFS-P-woP-50D (ours)	68.54	70.54	81.35	85.51	69.85	35.61	56.2	**71.65**	73.19	68.05
WordFS-P-wP-50D (ours)	**69.96**	71.05	**83.29**	84.25	**72.10**	37.69	56.6	71.30	72.82	68.78

TABLE IV
COMPLEXITY COMPARISON OF THE ALGO METHOD, THE EMBEDTEXTNET METHOD, AND OUR METHODS WHEN REDUCING THE WORD EMBEDDING DIMENSIONS FROM 300 TO 150 IN HARDWARE CONFIGURATION AND TRAINING TIME.

Method	Hardware Configuration	Training Time (s)
EmbedTextNet [6]	GPU	2149.04 (1535.03X)
Algo [4]	CPU	21.50 (53.75X)
WordFS-S-wP (ours)		17.63 (44.08X)
WordFS-P-wP (ours)		11.84 (29.60X)
WordFS-S-woP (ours)		4.05 (10.13X)
WordFS-P-woP (ours)		0.40 (1X)

are only 7,705 human-labeled similarity scores for different word pairs in the aggregated dataset, which is much less than the total number of possible word pairs in the corpus to pre-train the word embeddings. For example, there are 400K word vectors in the pre-trained Glove word embeddings, which can generate approximately 80 billion unique word pairs. Compared with the possible word pairs, the guidance provided by the aggregated dataset is limited. We experimented with the aggregated dataset using the same settings as before.

Table II shows the results of various methods for dimension reduction on the aggregated dataset. Our WordFS-S methods perform best among all the experiments in different dimensions. And there is a large gap between our methods and the existing methods. Compared to other methods, our approaches significantly benefit this task, particularly in lower dimensions where the gap widens. It is worth noting that even when reduced to 50 dimensions, our WordFS-S-wP method achieves better performance compared to the original 300-dimensional word embeddings.

C. Downstream Tasks

It is insufficient to evaluate word embeddings only based on word similarity tasks. To demonstrate the power of our method, generalizability on downstream tasks is crucial. We utilize the SentEval toolkit [15] to evaluate the dimension-reduced vectors on various downstream tasks. We conducted experiments on nine prediction tasks, including binary and multi-class classification tasks (MR, CR, MPQA, SUBJ, STS-B, STS-FG, and TREC), paraphrase detection task (MRPC), and entailment and semantic relatedness task (SICK-E). These tasks directly take the word embeddings as input. We applied our WordFS-P method with the RFT feature selection based on weak guidance from the aggregated word similarity dataset to the pre-trained word embeddings in all the experiments to reduce the dimension. Then, we directly input the dimension-reduced embeddings to the downstream tasks and compare the test accuracy of different methods for each task.

Table III shows the results of various methods on the nine prediction tasks. Our experiments show that our method outperforms the Algo method on most prediction tasks. Also, our method achieves higher average scores than the Algo method in all the settings. Our method performs better than the EmbedTextNet model, although it performs slightly worse when reduced to 50 dimensions. However, as shown in the next section, our model is much more efficient and takes less time to train than the EmbedTextNet model. Our method retains more helpful information for most settings while reducing dimensions, achieving a much closer prediction performance to the original 300-dimensional word embeddings. It indicates the effectiveness of our method. Also, it proves that our method can generalize well on various prediction downstream tasks in a zero-shot manner.

D. Model Efficiency

We compare the model efficiency of the Algo method, the EmbedTextNet method, and our method when reducing the word embedding dimensions from 300 to 150 in Table IV. Different hardware is used for other methods because of their algorithmic nature. All the experiments were conducted on the same server, equipped with two AMD EPYC 7543 32-core CPUs and seven NVIDIA RTX A6000 GPUs. The deep learning-based model is trained using the GPUs, while the Algo model and our methods are trained using only the CPUs. The EmbedTextNet model takes approximately 30 minutes to train on the GloVe word embeddings. In contrast, the training time for our WordFS-P-wP and WordFS-S-wP models is less than 19 seconds and 13 seconds, respectively. Our methods without post-processing require even less time — the training of the WordFS-P-woP model is completed within one second,

making it thousands of times faster than the deep learning-based method.

Additionally, our methods are more efficient than the Algo method because they employ PPA before and after the PCA, while we only apply it once before feature selection. Our model, without the post-processing, can achieve even less time. The reason is the post-processing and PCA must be done on the entire vocabulary. However, our feature selection method only focuses on a subset of words that appear in the aggregated word similarity dataset, making the procedure very fast. In conclusion, Our methods take only a fraction of the time compared to deep learning methods and less than half the time compared to existing PCA-based methods. The experimental results demonstrate that our methods are much more efficient than existing methods.

In conclusion, our method finds a good balance between simplicity and performance compared to other methods. With much faster training times, our approach can outperform in most tasks and may only slightly underperform when the dimensionality decreases to a very small number.

V. Conclusion and Future Work

This paper proposes a general dimension reduction method called WordFS for pre-trained word embeddings with the post-processing algorithm (PPA) and simple yet effective feature selection methods based on weak supervision provided by a limited number of word similarity pairs. Our method is more straightforward, efficient, intuitive, and effective in most cases than the existing methods. Empirical results show that our method outperforms existing methods in word similarity tasks and generalizes well to various downstream tasks with much lower computational costs. In the future, we would like to explore the compression of word embeddings further. Since feature selection methods may result in information loss, there is still room for improvement in the performance of prediction tasks between the original word embeddings and the reduced word vectors. Additionally, suitable feature selection methods can be developed to compress pre-trained word vectors in specific domains.

References

[1] J. Xue, Y.-C. Wang, C. Wei, X. Liu, J. Woo, and C.-C. J. Kuo, "Bias and fairness in chatbots: An overview," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 2, 2024. DOI: 10.1561/116.00000064. [Online]. Available: http://dx.doi.org/10.1561/116.00000064.

[2] H. Naveed, A. U. Khan, S. Qiu, *et al.*, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[3] C. Wei, Y.-C. Wang, B. Wang, and C.-C. J. Kuo, "An overview of language models: Recent developments and outlook," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 2, 2024. DOI: 10.1561/116.00000010. [Online]. Available: http://dx.doi.org/10.1561/116.00000010.

[4] V. Raunak, V. Gupta, and F. Metze, "Effective dimensionality reduction for word embeddings," in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 2019, pp. 235–243.

[5] J. Mu and P. Viswanath, "All-but-the-top: Simple and effective postprocessing for word representations," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=HkuGJ3kCb.

[6] D. Y. Hwang, B. Taha, and Y. Nechaev, "Embedtextnet: Dimension reduction with weighted reconstruction and correlation losses for efficient text embedding," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 9863–9879.

[7] A. K. Uysal and Y. L. Murphey, "Sentiment classification: Feature selection based approaches versus deep learning," in *2017 IEEE International Conference on Computer and Information Technology (CIT)*, IEEE, 2017, pp. 23–30.

[8] W. Rui, J. Liu, and Y. Jia, "Unsupervised feature selection for text classification via word embedding," in *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, IEEE, 2016, pp. 1–5.

[9] B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C. J. Kuo, "Evaluating word embedding models: Methods and experimental results," *APSIPA transactions on signal and information processing*, vol. 8, e19, 2019.

[10] Y. Yang, W. Wang, H. Fu, and C.-C. J. Kuo, "On supervised feature selection from high dimensional feature spaces," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[11] M. Ait-Saada and M. Nadif, "Is anisotropy truly harmful? a case study on text clustering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023, pp. 1194–1203.

[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[13] C.-C. J. Kuo and A. M. Madni, "Green learning: Introduction, examples and outlook," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103 685, 2023.

[14] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 462–471.

[15] A. Conneau and D. Kiela, "Senteval: An evaluation toolkit for universal sentence representations," *arXiv preprint arXiv:1803.05449*, 2018.

Tiny Object Detection Enhancement for Large-Scale Remote Sensing Imagery

Tianwei Zhang* Lianru Gao† Xu Sun† and Lina Zhuang†
* Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
E-mail: zhangtianwei20@mails.ucas.ac.cn
† Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
E-mail: gaolr@aircas.ac.cn, sunxu@aircas.ac.cn, zhuangln@aircas.ac.cn

***Abstract*—The effective detection of targets of interest in large-scale remote sensing imagery has always been a crucial application task. Detecting tiny objects in these scenes poses particular challenges, as it can lead to the loss of robust features and difficulties in generating positive sample matches. To address this application scenario, this paper proposes a method that embeds the use of vague search (VGS) to match high-potential candidate regions within a coarse-to-fine detection framework. Additionally, a new centroid distance measurement (CDM) is proposed to distinguish between positive and negative proposals. The combination of VGS and CDM can effectively increases the number of positive samples that can be refined in the high-dimensional detection layer, thereby improving the precision of detecting tiny objects in large-scale scenes. We tested and validated our model on a high-resolution optical remote sensing scene dataset, where the overall mean Average Precision (mAP) for the target of interest, bridges, improved from 56.2% to 71.3%, achieving impressive results.**

I. Introduction

Detecting targets in large-scale optical remote sensing images has always been a challenging task, especially for tiny-sized targets. Generally, deep neural network-based object detectors [1]–[3] often struggle to capture robust feature information in tiny object detection (TOD) task, both semantically and spatially. Many researchers have proposed specific solutions for detecting tiny targets in ordinary and remote sensing images. For example, [4] uses a coarse-to-fine approach to improve the utilization of feature information within the receptive field by first screening high-potential proposals and then conducting fine classification in the second stage. [3], [5], [6]have proposed positive and negative sample learning strategies during the model training process to supply more high-quality positive samples that can be stably refined by the detector. [7] introduced the fountain fusion module to enhance the model's ability to fuse multi-scale features, promoting the extraction of spatial features from shallow feature layers. These methods have achieved good results in TOD tasks under different scenarios. However, their generalization performance in detecting small targets in large-scale remote sensing images has been unsatisfactory. Therefore, we deeply analyzed the issues within the network inference mechanisms for this detection scenario and proposed our solution, TOD-Net[1], to improve the efficiency and accuracy of the model in TOD scenarios.

In our application scenario, due to the large size of a single inference image, reaching 2000×2000 (to meet the inference speed requirements for large-scale remote sensing images), after a certain number of down-sampling layers in the feature extraction process, a single detection unit (DU) not only encompasses the target of interest but also cover a large area of the background, including the surrounding region of the target. The visualization results are shown in Fig 1.

We visualized the regions in the original input image corresponding to positive samples in a detection layer of size 256×256 (i.e., weight visualization). It can be seen that in the area where the target of interest—bridge appears, the effective receptive field (ERF) of this DU covers the target. However, smaller targets on the right side of the input image cannot generate effective positive samples (i.e., yellow points in Fig.1). This is due to the combined effect of the following two factors:

1) Foreground information of tiny remote sensing objects is easily submerged by the background during the down-sampling process of feature extraction..
2) Tiny objects are harder to highlight in the existing positive sample matching mechanisms, resulting in an inability to effectively and stably refine them.

Therefore, based on the coarse-to-fine feature extraction mechanism, we designed a vague search module (VGM) for filtering high-confidence proposal regions and combined it with centroid distance measurement (CDM) of polygonal areas to increase the number of positive detection units during model training. Based on the above, we formed the core algorithm mechanism of TOD-Net.

We tested TOD-Net on a self-collected remote sensing image dataset of an urban scene named GF2UBS (GF-2 urban bridge scene), where the images were captured by the GF-2 satellite and are single-band panchromatic images. Additionally, we validated the effectiveness and robustness of our algorithm on large public remote sensing datasets,

[1]This work was supported by the National Key Research and Development Program of China under Grant 2021YFB3900502.

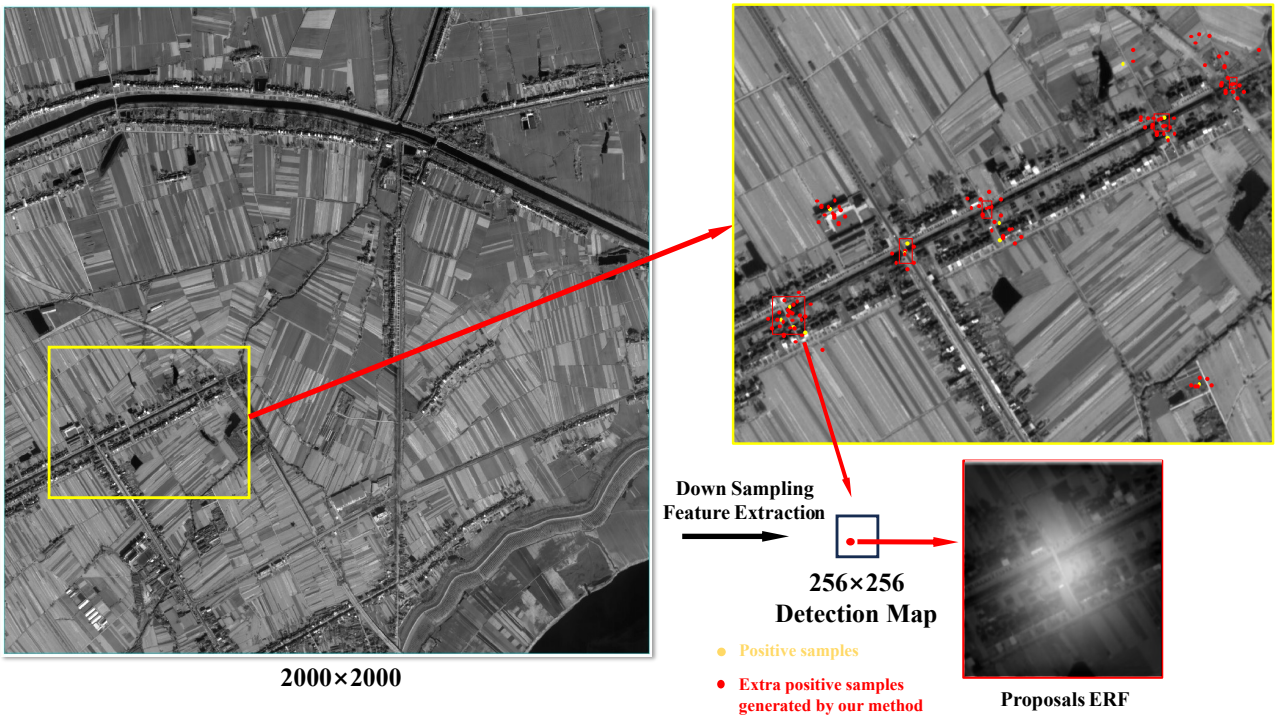


Fig. 1. Issues in detecting tiny objects of large-scale remote sensing imagery.

including the ship dataset HRSC2016 [8] and the multi-class remote sensing target dataset UCAS-AOD [9]. TOD-Net demonstrated superior performance in detecting small targets in large-scale remote sensing images, improving the overall detection accuracy (mAP) in the GF2UBS from 56.2% to an impressive 71.3%.

II. Methodology

A. Overall Pipeline of TOD-Net

TOD-Net directly uses large-sized remote sensing images (2000×2000) as network input during both training and testing phases. Consequently, under the same network configuration, the ERF of the DUs encompasses a larger spatial range of ground objects. The overall design process of TOD-Net is illustrated in Fig.2. After feature extraction through down-sampling by the deep neural network (DNN), multi-scale feature maps of varying sizes are obtained. TOD-Net adopts the classic ResNet [10] + FPN [11] combination as the backbone for feature extraction, followed by coarse-grained prediction box generation, resulting in the classification scores and regression results of the DUs. Then, two branches separately perform VGS and fine feature extraction, with the results of VGS serving as auxiliary information for the fine feature extraction step.

The multi-task overall loss in our experiments is defined as:

$$Loss_{Overall} = loss_{coarse} + loss_{fine}, \tag{1}$$

$loss_{coarse}$ and $loss_{fine}$ have the same format which is defined as:

$$loss_{coarse}, loss_{fine} = loss_{cls}(p, p^*) + loss_{reg}\left(\boldsymbol{t}, \boldsymbol{t}^*\right), \tag{2}$$

where the value p denotes the predicted classification score, p^* is the class label for all anchors ($p^* = 1$ for positive anchors and $p^* = 0$ for negative anchors). Symbol $\boldsymbol{t}$ is the position vector of predicted box offsets, and $\boldsymbol{t}^*$ denotes the GT box offsets. The position offsets vector $\boldsymbol{t}$ has the format (x, y, w, h, θ) where x and y denote the central coordinates, w the width, h the height, and θ the angle. For vector $\boldsymbol{t} = (t_x, t_y, t_w, t_h, t_\theta)$, we have:

$$\begin{aligned} t_x &= (x - x_a)/w_a, \quad t_y = (y - y_a)/h_a, \\ t_w &= \log(w/w_a), \quad t_h = \log(h/h_a), \\ t_\theta &= \tan(\theta - \theta_a), \end{aligned} \tag{3}$$

where x and x_a are the predicted box and anchor, respectively (likewise for y, w, h, θ). Focal Loss [12] is used in $loss_{cls}$ to reduce the contribution of easy-to-classify samples, and it alleviates the insufficient training caused by the class imbalance to a certain extent. The bounding box regression process (i.e., $loss_{reg}$) uses smooth L1 loss by default when there is no special statement.

B. VGS

For the obtained multi-scale feature maps $\mathbf{X}_i \in \mathbb{R}^{H\times W\times C}, i \in \{0, 1, 2, 3, 4\}$, i is the scale index in our TOD-Net. $\mathbf{X}_i(p)$ represents the feature vector corresponding to DU at position p in each feature map, which includes the confidence score of the predicted category generated in the coarse

detection stage and the regression parameters of the predicted box. MD_p is the positive and negative sample determination result at position p, $p \in \mathbb{P}\{0, 1, ..., W-1\} \times \{0, 1, ..., H-1\}$. For positive sample candidate regions, $MD_p = 1$, otherwise 0.

The core objective of VGS is to expand the search range within the coarse detection results where targets might appear, thereby maximizing the inclusion of valid targets within positive DUs that have larger receptive fields. This supports the feature extraction task in the subsequent fine-grained detection stage. As a result, searching expanding function $\tilde{\mathcal{F}}$ is defined as:

$$\begin{aligned} \tilde{\mathcal{F}} : MD_p = 1, \text{if } IoU(\mathbf{X}_i(p)) > 0.2, \\ \text{for } i \in [0, 5) \text{ and } p \in [p-1, p+1] \end{aligned} \tag{4}$$

where IoU is the traditional matching degree used in general object detectors. $\tilde{\mathcal{F}}$ is used to retrieve regions at any detection unit position that meet the criteria of having prediction confidence greater than 0.2 within all adjacent scales and surrounding 3×3 neighboring units. These proposals are then considered as positive samples, thereby generating potential candidate regions.

C. CDM

CDM is used to calculate the distance between the centroids of two polygonal regions as an evaluation metric. Its application is particularly suitable when the candidate region is large and the two regions have significant overlap characteristics. As shown in Fig. 4, the black dot $P_0(x_0, y_0)$ represents the centroid coordinates of the generated fuzzy candidate region, and the yellow dot $P_c(x_c, y_c)$ represents the centroid coordinates of the generated coupling region (CR). CR is the union of the predicted box (PB) generated during the coarse detection stage and the intersecting ground truth (GT). Therefore, the calculation method for Pc is:

$$x_c = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \textbf{and} \quad y_c = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{5}$$

After obtaining the centroid coordinates of CR, we can automatically determine whether the anchor box appearing in the refined feature extraction stage is a positive sample based on the inscribed circle radius of vague proposals, which is defined as:

$$CDM = Dis\left[P_0, P_c\right] <= \frac{\sqrt{2}}{2}\sqrt{\text{ratios}^{\,2}}, \tag{6}$$

where $ratios$ is the radius of the inscribed circle for vague proposals, Dis is the distance metric function, which can be measured using classical vector distance metrics such as Euclidean distance or Mahalanobis distance.

III. Experiments

A. Datasets

We first tested TOD-Net on a self-collected high-resolution optical remote sensing dataset, GF2UBS, which consists of GF-2 optical panchromatic images. The dataset contains 182

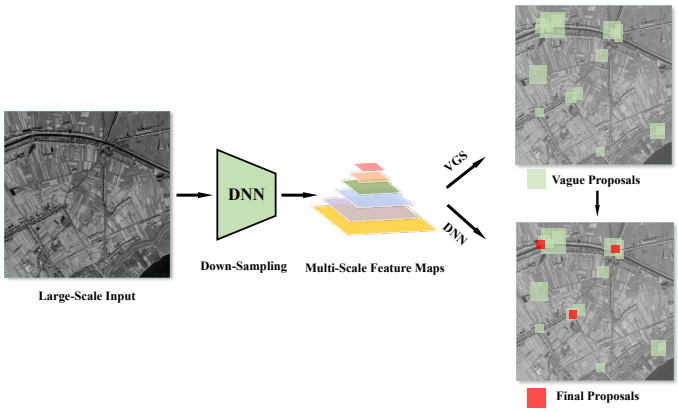


Fig. 2. Overall pipeline of TOD-Net. VGS denotes vague search strategy and we generate coarse detection results in the VGS process. CDM is used in the finer feature extraction process which cooperates with the output of VGS.

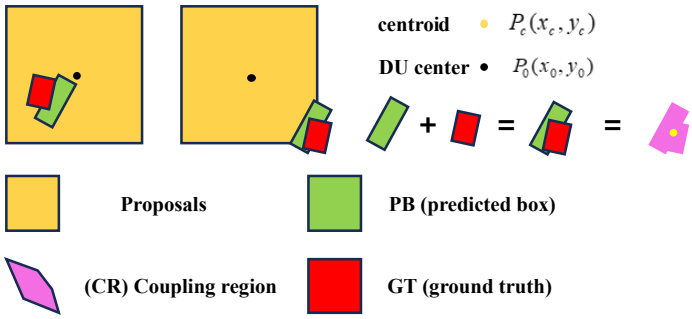


Fig. 3. Illustration of CDM.

panchromatic image slices, each sized 2000×2000 pixels. The dataset annotates 282 various types of bridges scattered throughout urban scenes, as shown in Fig. 4. Since most of the annotated bridges are small urban traffic bridges, their pixel sizes are relatively small, with a size distribution ranging from 7×10 to 121×36 pixels.

In addition to the remote sensing tiny object detection dataset in large-scale images, we also validated the generality of our method on two commonly used public remote sensing object detection datasets, which are introduced as follows:

1. HRSC2016 [8] is for ship detection that was collected from six harbors on Google Earth. It contains 1061 images (617 for training and 444 for testing). The image size ranges from 300 × 300 to 1500 × 900. We resize all images to 800 × 800.

2. UCAS-AOD [9] is an aerial image dataset for oriented aircraft and car detection which contains 1510 images including 1000 airplane images and 510 car images. We randomly divide it into the training set, validation set, and test set as 5:2:3.

B. Results on GF2UBS

Table I compares the detection accuracy results of TOD-Net with state-of-the-art object detection algorithms on the GF2UBS dataset. We used RetinaNet as the baseline for our model, with the baseline accuracy reaching only 56.2%. Under

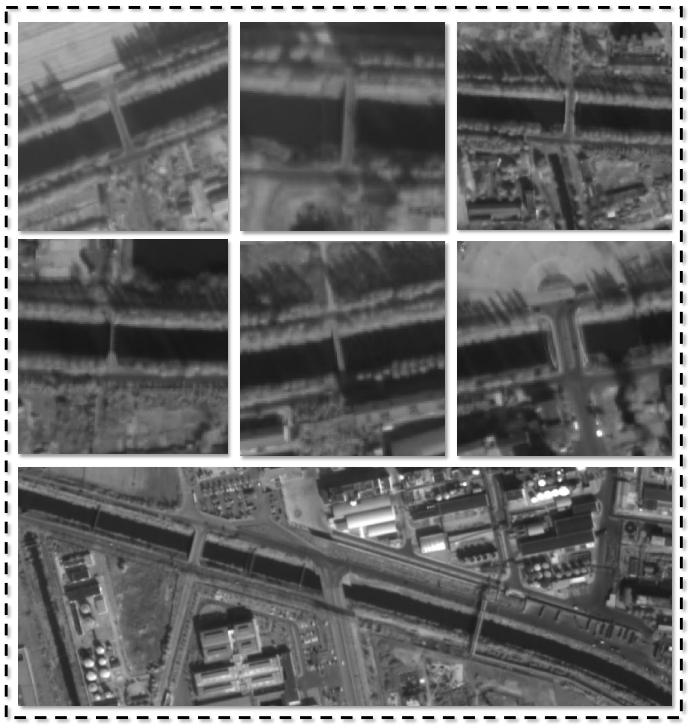


Fig. 4. Data examples of GF2UBS.

the same experimental configuration, TOD-Net increased the accuracy by 13.6%, which is quite impressive. This confirms the applicability and effectiveness of the proposed method in this paper.

When using more complex backbones for feature extraction, TOD-Net achieved a test accuracy of 71.3% on the GF2UBS, significantly outperforming other compared algorithms and achieving the fastest inference speed. Additionally, when using smaller image sizes as the benchmark for model training and testing, the mAP of TOD-Net further improved. This is because the model can capture more detailed spatial information despite reducing the network input size increases the overall inference time.

When focusing on the detection efficiency of tiny targets, the method proposed in this paper also achieved significantly better accuracy compared to other models. We evaluated the mAP-tiny of targets with pixel areas smaller than 30 across different algorithms. The performance of TOD-Net far surpassed that of the other methods, reaching a maximum of 40.1%. Although this result is not entirely satisfactory, it highlights the considerable difficulty in detecting small targets in the given scenes, suggesting there is still significant room for future improvement.

C. Comparison with State-of-Art Methods

To verify the generalizability of the method proposed in this paper, we also conducted validation on two additional public remote sensing datasets. On the HRSC2016 dataset, we compared our method with more state-of-the-art object detection algorithms, as shown in Table II and Table III. The

TABLE I
COMPARISONS WITH STATE-OF-THE-ART DETECTORS ON GF2UBS. BEST RESULTS ARE IN **BOLD**. RUNTIMES REPRESENTS THE TIME CONSUMED FOR MODEL INTERFERENCE OF TEST DATASET.

Methods	Backbone	Input Size	mAP-tiny	mAP-overall	runtimes (s)
S^2A-Net[13]	ResNet50	2000×2000	28.6	61.3	432.9
AOPG [14]	ResNet50	2000×2000	30.1	62.8	476.8
ReDet [6]	ResNet50	2000×2000	31.2	64.6	502.7
O-RCNN [15]	ResNet50	2000×2000	26.3	56.9	496.3
Baseline (RetinaNet)	ResNet50	2000×2000	24.6	56.2	465.4
TOD-Net (ours)	ResNet50	2000×2000	39.2	69.8	418.5
TOD-Net (ours)	ResNet50	800×800	44.3	76.5	538.5
TOD-Net (ours)	ResNet101	2000×2000	40.1	**71.3**	452.8

[1] mAp is calculated below VOC2012 rules.
[1] mAP-tiny refers to the detection precision of targets with a pixel area less than 30.

TABLE II
COMPARISONS WITH STATE-OF-THE-ART DETECTORS ON HRSC2016. NA DENOTES NUMBER OF ANCHORS. RUNTIMES REPRESENTS THE TIME CONSUMED FOR MODEL INTERFERENCE OF TEST DATASET.

Methods	Backbone	Input Size	NA	mAP (12)	runtimes (s)
*R^3Det-DCL [16]	ResNet101	800×800	-	96.41	134.6
*◇ DAL[1]	ResNet101	800×800	3	94.33	163.4
*◇ DRN[5]	ResNet101	800×800	-	95.65	154.1
*S^2A-Net[13]	ResNet50	800×800	1	95.01	128.6
*AOPG [14]	ResNet101	800×800	-	96.22	139.8
*ReDet [6]	-	800×800	-	97.63	167.1
*O-RCNN [15]	ResNet101	800×800	-	97.60	162.3
*CGC-Det[17]	ResNet101	800×800	-	97.86	168.3
TOD-Net (ours)	ResNet50	800×800	3	97.88	143.5
TOD-Net (ours)	ResNet101	800×800	3	**98.26**	152.8

[1] Best results are in **bold**.
[2] * represents the experimental runtimes are obtained from our local environment. ◇ indicates that the mAP is the result of re-realization.

results in these tables demonstrate that TOD-Net excels at enhancing the model's fine-grained feature extraction capabilities for targets in various scenes. This improvement effectively and consistently enhances the detector to extract robust features from target locations, thereby increasing the final accuracy.

On the HRSC2016 dataset, using ResNet101, we achieved a final mAP of 98.26%. On the UCAS-AOD dataset, we obtained the highest performances of 90.64% for car targets and 91.26% for airplane targets. These experimental results clearly demonstrate that TOD-Net not only performs excellently in large-scale remote sensing small target tasks but also adapts well to typical remote sensing target detection tasks in general application scenarios.

IV. CONCLUSIONS

In this paper, we propose a object detector for tiny objects in large-scale remote sensing images, named TOD-Net. TOD-Net employs a coarse-to-fine design and integrates a high-potential vague region screening strategy called VGS. VGS effectively expands the search range of detection units by searching for potential target areas across multiple scale levels and adjacent units. Additionally, we introduce a centroid distance metric to match small target positive samples, increasing the number of positive samples that can be refined during training.

TABLE III
DETECTION RESULTS ON UCAS-AOD DATASET.
BEST RESULTS ARE IN **BOLD**.

Methods	car	airplane	mAP
Baseline	0.8322	0.8643	0.8472
YOLOv3-O[18]	0.7463	0.8952	0.8208
Faster R-CNN-O [19]	0.8687	0.8986	0.8836
DAL[1]	0.8925	0.9049	0.8987
Oriented Reppoints [20]	0.8951	0.9070	0.9011
TOD-Net (ours)	**0.9064**	**0.9126**	**0.9106**

Experimental validation of TOD-Net on the large-scale scene remote sensing tiny target detection dataset GF2UBS shows an impressive 15.1% increase in the detection mAP for bridge targets, yielding remarkable results.

REFERENCES

[1] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," *arXiv preprint arXiv:2012.04150*, vol. 1, no. 2, p. 6, 2020.

[2] L. Ren, L. Gao, M. Wang, X. Sun, and J. Chanussot, "Hadgsm: A unified nonconvex framework for hyperspectral anomaly detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[3] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic r-cnn: Towards high quality object detection via dynamic training," *arXiv preprint arXiv:2004.06002*, 2020.

[4] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 821–830.

[5] X. Pan, Y. Ren, K. Sheng, *et al.*, "Dynamic refinement network for oriented and densely packed object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 207–11 216.

[6] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2786–2795.

[7] T. Zhang, X. Sun, L. Zhuang, *et al.*, "Ffn: Fountain fusion net for arbitrary-oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[8] W. LB *et al.*, "A high resolution optical satellite image dataset for ship recognition and some new baselines," 2017.

[9] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *2015 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2015, pp. 3735–3739.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[13] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *arXiv preprint arXiv:2008.09397*, 2020.

[14] G. Cheng, J. Wang, K. Li, *et al.*, "Anchor-free oriented proposal generator for object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[15] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3520–3529.

[16] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection. arxiv 2020," *arXiv preprint arXiv:2011.09670*, 2021.

[17] Y. Wang, Z. Zhang, W. Xu, *et al.*, "Learning oriented object detection via naive geometric computing," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2023. DOI: 10.1109/TNNLS.2023.3242323.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[19] G.-S. Xia, X. Bai, J. Ding, *et al.*, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.

[20] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1829–1838.

Scale-invariant Online Voice Activity Detection under Various Environments

Ryu Takeda* and Kazunori Komatani*
* SANKEN, Osaka University, Japan
E-mail: {rtakeda, komatani}@sanken.osaka-u.ac.jp Tel: +81-6-6879-8416

***Abstract*—Online voice activity detection (VAD) is an important front-end for spoken dialogue systems. However, different signal amplitudes and speech distortions under various environments cause performance degradation of neural VAD models due to the model mismatch. The amplitude and distortion problems were addressed during the feature extraction and training processes of neural networks, respectively. First, the signal amplitude was normalized block-wise to ensures the scale invariance mathematically. Such block-wise normalization was naturally introduced in our formulation of online VAD based on a recursive Bayesian estimation of speech activity. Second, over 1,000 hours of training data was augmented by simulating speech distortions, such as reverberations. Our VAD outperformed open VAD models, such as Silero, for a variety of datasets including a real spoken dialogue dataset in terms of speech and non-speech discrimination. Codes and models have been publicly released in a toolkit.**

I. Introduction

A. Background and Motivation

Online voice activity detection (VAD) [1] is an important front-end technology of spoken dialogue systems. The role of VAD is to detect the active-speech sections, i.e. pairs of start and end times of speech, from the (noisy) input signals captured by microphones. Additionally, online/real-time processing of VAD is an essential requirement because the latency of VAD obviously leads to the response time from the system, which affects the user experience [2]. Since spoken dialogue systems, such as guide robots, need to continue to operate and serve on demand from various speakers under various environments [3], [4] including remote spoken dialogue [5], the demand of *tuning-free and robust* VAD and its publicly-released models increases for real system developers. We assume a single speaker and a monaural microphone for the input signal hereafter.

Audio input always depends on target speakers, acoustic environments and recording conditions (Figure 1). Their influences are observed as the different signal amplitudes and speech distortions. The *signal amplitude* is affected by the *gain setting of audio devices* and the audio format used in recording and by the volume of each utterance by speakers. For example, there are both quiet and loud speakers, and the dynamic range of audio signal differs among them. Speech distortions can occur due to the various reverberations and background non-speech signals, as well as the lossy compression of speech data. For example, audio signals are compressed to reduce the amount of communication data in the case of dialogues on remote conference applications. These factors are usually not controllable by system developers.

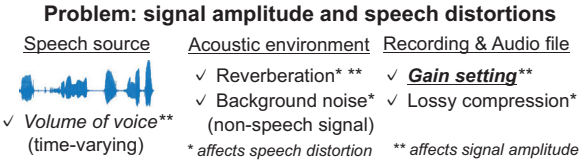


Fig. 1. Difficulties in VAD under various environments

Different signal amplitudes and speech distortions degrade the performance of current neural VAD models due to the model mismatch. Online VAD is more difficult than batch VAD because of the constraints of causal processing and low latency. Here, the online processing is different from batch processing that assumed in diarization tasks because the future inputs are not available (causality). For example, although Silero [6] is an open real-time VAD model that adopts the adaptive normalization of amplitude, its performance degrades for audio signals including long non-speech sections, background noises and reverberations with small amplitude. PyAnnote [7], [8] for diarization tasks (batch VAD) is not robust against background noises. Their models are based on a recurrent neural network (RNN) [9] (e.g., long short-term memory (LSTM) [10]), and a convolutional neural network (CNN) [11], [12]. Therefore, their actual performance indicates that the network structure or the training data of models were not designed for severe audio signals.

In light of this background, we propose a block-wise *scale-invariant* normalization and data augmentations for the amplitude and distortion problems, respectively. First, the signal amplitude was normalized in a block-wise manner to ensure the scale invariance of signal *mathematically*. Such block-wise normalization was naturally introduced in our formulation of online VAD based on a recursive Bayesian estimation of speech activity, i.e., a hybrid model of hidden Markov model (HMM) [13] and deep neural networks (DNN) using transformer-encoder. Second, over 1,000 hours of training data for VAD models was augmented by simulating speech distortions, such as reverberations. Since the process of DNN (transformer-encoder) is independent from recursive Bayesian filtering, its training becomes efficient for such large amount of data by mini-batch parallelization. The robustness of our VAD, Silero and PyAnnote was evaluated by using various open datasets including real spoken dialogue data.

The contributions of this paper are as follows.

- We focus on the problem of different signal amplitudes and speech distortions in terms of system development.
- We developed an online robust VAD based on scale-invariant transformation, and we also trained the model with data augmentation as solutions for the problems.
- A comprehensive evaluation of our VAD and two publically available VADs was conducted using various kinds of data in terms of scale invariance and noise robustness.
- Codes and models have been publicly released in the VAD toolkit[1].

B. Related Work

Although there has been extensive research on VAD, comprehensive research toward *tuning-free* and open-model VAD have not been tackled yet. While earlier works focused on clean speech signals, more recent works have examined additional functions such as online/real-time processing or noise robustness, as the performance of classifiers has been steadily improving thanks to the latest neural networks.

Standard VAD methods classify speech and non-speech sections by applying pattern recognition techniques. Traditionally, features and classifiers have been modeled separately. Examples of major features include energy [14], zero-crossing rate [15], and classifiers include logistic regression, support vector machine [16], and HMM [15]. The latest VADs are based on DNNs (RNN) which can optimize feature extraction and classifier simultaneously, e.g., feature learning with raw-waveform [17]. A variety of aspects of VAD has been also investigated: online VAD [18] with low-latency [19], noise-robust VAD using context information [20], [21], and integration of VAD with automatic speech recognition (ASR) [22].

II. Preliminaries

A. Problem Statement

Some variables are defined for our VAD formulation in the feature domain. The monaural signal is converted into a feature vector $\mathbf{x}_t$ at discrete time-frame index t, and the voice activity of speech at frame t is represented by its label $v_t \in \{0, 1\}$. If the speech exists at frame t, $v_t = 1$, otherwise, $v_t = 0$.

Online VAD is formulated as a sequential labelling problem, that is, the estimation of voice activity state z_t at frame t from the sequence of feature vectors $\mathbf{x}_{1:t} = [\mathbf{x}_1, ..., \mathbf{x}_t]$. The posterior probability $p(z_t|\mathbf{x}_{1:t})$ is usually utilized as its score, and the optimal $\hat{z}_t$ is estimated as

$$\hat{z}_t = \text{argmax}_{z_t} p(z_t|\mathbf{x}_{1:t}). \quad (1)$$

As an alternative to this criterion, $\hat{z}_t$ can be determined by thresholding the posterior probability. The definition of the voice activity state z_t may change depending on the design of VAD. For example, $z_t = v_t$ means the *filtering* setting, and $z_t = v_{t-m} (m > 0)$ means the *smoothing* setting in terms of sequential signal processing [23]. The setting of the allowed latency m depends on applications. We can also set the joint state of voice activity such as $z_t = (v_t, v_{t-m})$, which can model the transition of v_t explicitly.

Voice-active sections are estimated by using the $\hat{z}_t$ and its posterior probability. The detection is usually based on the transition of state $\hat{z}_t$, and the post-processing for the integration and the rejection of sections are usually applied. The detail of our implementation is explained in Section 3.

B. RNN Implementation

RNN series also used in Silero [6] can directly model the posterior probability $p(z_t|\mathbf{x}_{1:t})$ via hidden vector $\mathbf{g}_t$. The advantage of this approach is the end-to-end modelling whose parameters can be optimized in a supervised manner by using training data. Here, PyAnnote [7], [8] assumes a non-causal batch processing that is different from this sequential formulation.

The implementation of posterior probability *conceptually* consists of the function $\mathbf{G}$ for hidden vector extraction and the probability function $p(z_t|\mathbf{x}_t, \mathbf{g}_t)$, as

$$p(z_t|\mathbf{x}_{1:t}) := p(z_t|\mathbf{g}_t, \mathbf{x}_t), \text{and} \quad (2)$$

$$\mathbf{g}_t = \mathbf{G}(\mathbf{x}_t, \mathbf{g}_{t-1}) \quad (3)$$

where $\mathbf{g}_t$ represents all the hidden vectors among layers. Each function can be implemented by any networks, such as many-layered LSTM and CNN. The loss function is usually the corss-entropy between the probabilities of the ground truth label and the model's prediction.

C. DNN-HMM Implementation

A DNN-HMM is a state-space model [13] in which the likelihood is replaced by the point-wise posterior probability (snapshot estimation of state) using DNN [24]–[26]. The design of DNN posterior probability and the recursive Bayesian filtering are separated, that is, conditional independent each other in this model. Therefore, the training of the DNN becomes more efficient thanks to parallel processing using mini-batch.

DNN-HMM estimates the activity state z_t recursively by

$$p(z_t|\mathbf{x}_{1:t}) \propto \frac{p(z_t|\mathbf{x}_{1:t-1})p(z_t|\mathbf{x}_t; \mathbf{\Theta})}{p(z_t)}, \text{and} \quad (4)$$

$$p(z_t|\mathbf{x}_{1:t-1}) = \sum_{z_{t-1}} p(z_t|z_{t-1})p(z_{t-1}|\mathbf{x}_{1:t-1}), \quad (5)$$

where $p(z_t|\mathbf{x}_t; \mathbf{\Theta})$ is a posterior probability, $p(z_t)$ is a prior probability of voice state, and $p(z_t|z_{t-1})$ is a state transition probability. Here, the posterior $p(z_t|\mathbf{x}_t; \mathbf{\Theta})$ is modelled by a DNN with a parameter set $\mathbf{\Theta}$. The parameters of state transition and prior probabilities are trained or set manually if the size of them is small. The DNN parameters are also trained, and the cross-entropy is usually used as a loss function.

III. Proposed Online VAD

In this section, we first describe our strategy for the model, feature, and target state v_t for the proposed VAD. Then, we explain the network architecture including scale invariance

[1] pyadintool - https://github.com/ouktlab/pyadintool

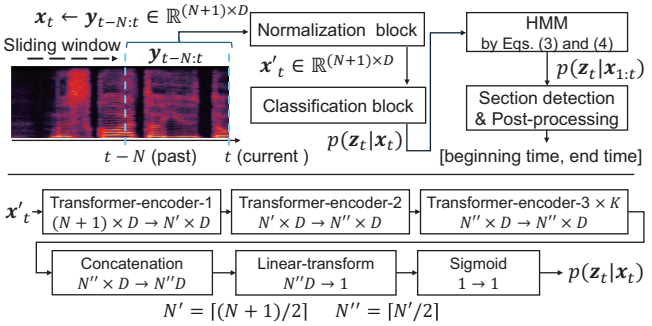


Fig. 2. Overview (top) and classification block (bottom)

transformation for the signal amplitude problem. Finally, we explain the training data augmentation for the speech distortion problems. The overview and relationships among each component are illustrated in Fig. 2.

A. Strategy for Model, Feature and Target State

We adopted a DNN-HMM because we can separate the design of the posterior from that of the sequential filtering. Since the latest neural architectures has a high discriminative ability, sequential filtering using HMM is enough for online VAD. The transformer encoder [27] is used to model the posterior probability $p(z_t|\mathbf{x}_t)$ because of its non-recurrent architecture and potential flexibility using prompts. The neural networks consist of normalization and classification blocks.

The input feature for DNN is a block of the amplitude spectrogram that is obtained by applying short-time Fourier transform (STFT) to the input signal. Given the D-dimensional amplitude spectrogram $\mathbf{y}_t \in \mathbb{R}^D$ in the STFT domain at frame t, our feature is the block-spectrogram $\mathbf{y}_{t-N:t} = [\mathbf{y}_{t-N}, ..., \mathbf{y}_t]$ from past $(t-N)$ to current t that corresponds to $\mathbf{x}_t$ in Eq. (1), that is, $\mathbf{x}_t \leftarrow \mathbf{y}_{t-N:t}$.

We applied the smoothing setting for Eq. (1) to estimate voice activities stably, in other words, our target state z_t for VAD is $v_{t-m} (m > 0)$. The parameter m controls the tradeoff between the latency and the estimation accuracy. We assumed about 200 ms for this delay that will be acceptable for ASR.

B. Scale-invariant Architecture and Post-processing

Normalization block: The scale of each block spectrogram $\mathbf{x}_t (\leftarrow \mathbf{y}_{t-N:t})$ is normalized to ensure the scale-invariant behavior of online VAD *mathematically*. We applied the *average-scale normalization* and layer normalization method [28]. While the latter normalization includes trainable parameters, the former normalization does not. The average-scale normalization also standardizes the parameter training behavior of the following networks including layer normalization.

Normalized block-spectrogram $\mathbf{x}'_t$ is obtained by normalizing each $\mathbf{x}_t$ by the following average-scale (avg.) and layer normalization (layer), respectively:

$$\tilde{\mathbf{x}}_t \xleftarrow[\text{avg.}]{} \frac{\mathbf{x}_t}{\text{avg}[\mathbf{x}_t] + \epsilon}, \tag{6}$$

$$\mathbf{x}'_t \xleftarrow[\text{layer}]{} \frac{\tilde{\mathbf{x}}_t - \text{avg}[\tilde{\mathbf{x}}_t]}{\sqrt{\text{var}[\tilde{\mathbf{x}}_t] + \epsilon}} \gamma + \beta, \tag{7}$$

where avg$[\cdot]$ and var$[\cdot]$ represent averaging and variance operator over all elements of $\mathbf{x}_t$, and ϵ is a small value for regularization. γ and β are trainable scale and offset parameters of layer normalization.

Classification block: Our classification block consists of transformer-encoder, linear transform and sigmoid layers. The input of this classification block is a normalized block-spectrogram $\mathbf{x}'_t$ with the dimension of $(N+1) \times D$, and the output is a posterior probability, i.e. $p(z_t|\mathbf{x}_t)$. Here, the transformer-encoder consists of multi-head attention and feed-forward networks. Positional encoding is applied to the input $\mathbf{x}'_t$ at first by considering the block-axis as a time-axis.

The three kinds of transformer-encoder networks were applied by reducing the block size $(N+1)$ to N''. This process can reduce the computational cost of these layers in proportion to the reduced block size. For example, the block sizes of each output from the first and second layers were halved as $N' = \lceil (N+1)/2 \rceil$ and $N'' = \lceil N'/2 \rceil$, respectively. The third transformer-encoder layer is repeated K times.

The block vectors from the three kinds of transformer-encoder layers were concatenated into a single vector, and linear transformation network was applied to it. The probability of voice activity state z_t was approximated by applying sigmoid function to the output from the linear transformation network.

Detection of Voice Active Section and Post Processing: The voice-active section is detected by using the estimated voice-activity state $\hat{z}_t$. In our case of $z_t = v_{t-m}$, the beginning frame t^s_k and the end frame t^e_k of the k-th voice-active section were detected by using the difference between $\hat{z}_t$ and $\hat{z}_{t-1}$. If $\hat{z}_t - \hat{z}_{t-1} = 1$, then it indicates the beginning frame: $t^s_k \leftarrow t - m$. If $\hat{z}_t - \hat{z}_{t-1} = -1$, then it indicates the end frame: $t^e_k \leftarrow t - m$.

The post processing is also applied to reject or merge the detected sections. The interval of the section is less than u_r, we discard it as a fluctuation. If the interval between the k-th and $(k+1)$-th sections is less than u_c, the two sections were merged into a new section t^s_k and t^e_{k+1}. Here, u_c should also be determined by the allowed latency of the system because it needs to wait for a while. Expansion of each section like $[t^s_k - u_m, t^e_{k+1} + u_m]$ with a margin parameter u_m is sometimes introduced to reduce the mismatch of voice-activity annotation criteria among datasets.

C. Augmentation Strategy for Speech Distortions

We augmented training data from the given seed audio signals. Three kinds of clean audio signals were used as the seed: clean speech signal, clean non-speech signal, and real impulse responses for reverberation simulation. Speech distortions were

simulated by combining with acoustic transformations excepts for the signal amplitude.

Our transformations for data augmentation are as follows.

1) Clean audio: Seed speech and non-speech signals.
2) Reverberant speech: Convolved speech signals by randomly selected impulse responses.
3) Reverberant and noisy speech: Non-speech signals added to the set 2) with randomly selected SNRs.
4) Collapsed speech: lossy compressions, such as mp3 and μ-low conversion, applied to the set 3).

After generating 2), 3) and 4), we applied oversampling [29] of the seed non-speech signals to relax the imbalance between active and non-active labels/data.

IV. Experiments

Our experiments investigate the *scale-invariance* and noise robustness of each VAD: two public VAD models (Silero with online processing and PyAnnote with batch processing) and our model (online processing). Note that the results of PyAnnote is reference because the processing style of VAD, such as online or batch (non-causal), usually affects VAD performance.

A. Data set

Seed corpora of speech: The corpora for speech signals mainly consist of nine *public* Japanese speech corpora: a core set of CSJ [30], S-JNAS, TWM, JEIDA-JCSD, ETL-WD copora[2], APP, APPDIC copora[3], SLC-3[4], and JVS [33]. These corpora cover over 4,000 speakers from 6 to 91 years old, utterances of words and sentences, and voices of whispers and falsetto. Extra utterances of single and double syllables were augmented by Google TTS and collected by our group, and the amount of its speech sections was about 50 hours. These total 677-hours speech data were downsampled to 16kHz.

Seed corpora of non-speech: The corpora for non-speech signals mainly consist of three *public* corpora: MUSAN [34], the training set of WHAM! [32], and the ProSoundEffects (PSE) corpus[5]. We also prepared 20 hours of different additional kinds of non-speech signals: pure tone signals, white/brown/pink noise signals, simulated babble noise signals, and environmental sound noise signals recorded by our group. These total 400-hours non-speech data were downsampled to 16kHz.

Training set (full and medium): The about 4,300 hours of the *full* training set consists of the seed non-speech corpora (four-times oversampled) and the augmented speech data. The transformations for augmentation were applied to the seed speech corpora, and we obtained the augmented speech data that was increased to 2,688 hours in total. The real impulse responses with 540 positions in a room of our group (RT_{20} 640 ms) were used to simulate *various real* reverberations. The SNR settings were randomly selected from $-10, -5, 0, 5, 10$, and 20 dB. The *medium* training set was randomly selected from the full set, and its amount was about 1,100 hours. The ratio between the speech and non-speech *sections* was $0.45 : 0.55$.

Test set: We prepared the three kinds of *open* test sets: non-speech, speech and dialogue sets. Non-speech set consist of *public* ESC-50 [31], the test set of WHAM! and the held-out set of PSE. Dialogue set consist of *public* real recorded spoken dialogue data called Hazumi 1911 and Hazumi2010 [5]. Speech set includes the clean test set of CSJ (*clean*), the simulated reverberant speech (*rev.*), and the simulated reverberant noisy speech (*rev.+bgn.*). Five kinds of impulse responses from the *public* RWCP-SSD corpus [35] were used to generate the reverberant speech from the CSJ test set. Then, signals in the non-speech set were added to them with randomly selected SNRs from $5, 10, 15$, and 20 dB. We changed their signal amplitude by multiplying $1.0, 0.5, 0.2$, and 0.05 and saved as 16bit-WAV format files to evaluate the robustness against amplitude changes.

Validation set: The validation set for monitoring the stationarity of F1-score change among epochs was constructed with different source signals by the same procedure as the test set: clean, rev. and rev.+bng sets. Source speech signals from the held-out set of CSJ, non-speech signals from the held-out set of PSE and the validation set of WHAM!. Different impulse responses were selected from the RWCP-SSD. The ratio between the speech and non-speech sections was almost equal.

B. Configurations

Our VAD model was built from scratch with the PyTorch library and its pre-defined classes [36]. The STFT parameters were the 512-point Hanning window and the 160-points shift (0.1 s). The label delay m was 20 (0.2 s), and the block size N was 50 (0.5 s). The transformer-encoder with eight-heads (default) and 512-hidden vectors was used, and the number of third transformer-encoders K was 2. The parameter initializations of each layer followed default setting of PyTorch. Gradient clipping [37] and Adam [38] were applied with clipping value 5 and learning rate 1.0×10^{-5}, respectively. The number of epochs was 30. The parameter sets of contiguous 10 epochs were **averaged** to reduce the influence of fluctuations at each epoch like [39], and then it was used for evaluations. The section (start and end epoch) for the parameter averaging was determined by the best corresponding averaged F1 score of *csj+rev+bgn.* set in our validation set. Here, the number of parameters of our models was five million while those of Silero (for 16k) and PyAnnote (for segmentation) public models were 0.15 and 1.5 million, respectively.

The parameters of post processing were the following: both u_c, and u_r were set to 10 (0.1 s), respectively. u_m was set to 10 (0.1 s) to reduce the influence of different annotation criteria among datasets. These parameters for PyAnnote and Silero were set to 0 because it performed best. Other hyper parameters of them were set to defaults. The prior probabilities

[2]https://research.nii.ac.jp/src/list.html
[3]https://www.atr-p.com/products/sdb.html
[4]https://alaginrc.nict.go.jp/slc-outline.html
[5]Pro Sound Effects Library. http://www.prosoundeffects.com

TABLE I

DETAILED RESULTS: F1-SCORE OF FRAME-WISE BINARY CLASSIFICATION IN PERCENTAGE. NOTE THAT THE RESULTS OF PYANNOTE ARE AS A REFERENCE BECAUSE OF ITS BATCH PROCESSING (NON-CAUSAL). THE LOWER AND UPPER BOUNDS OF ALMOST ALL CONFIDENCE INTERVALS WERE WITHIN ± 0.1. † AND ‡ INDICATE THAT THE CORRESPONDING BOUNDS WERE WITHIN ± 0.2 AND ± 0.3, RESPECTIVELY.

	Dataset	Non-speech set						Speech set (CSJ eval. [30])						Dialogue set [5]			
	Corpus/Condition	ESC-50 [31]		PSE		WHAM! [32]		clean		rev.		rev.+bgn.		Hazumi1911		Hazumi2010	
	Amplitude change	1.00	0.05	1.00	0.05	1.00	0.05	1.00	0.05	1.00	0.05	1.00	0.05	1.00	0.05	1.00	0.05
Baseline	Silero (online) [6]	†94.2	**99.2**	97.1	99.4	**94.8**	**100.0**	95.8	93.9	92.4	†72.5	†88.4	‡60.5	†85.9	†82.2	†92.6	†87.4
	PyAnnote (batch) [7]	93.6	†94.0	†95.2	95.9	87.4	91.5	**97.4**	**97.4**	91.5	91.5	90.8	90.4	†87.9	†87.1	†92.7	†92.4
Proposed (online)	medium training set	**97.3**	97.2	**98.6**	98.5	91.1	91.0	94.8	94.6	93.8	**94.0**	93.4	**93.4**	†89.9	†89.9	†92.9	†**92.9**
	full training set	**97.3**	97.3	98.3	98.3	93.4	93.4	94.6	94.5	93.0	93.5	92.9	92.9	†90.5	†**90.5**	†92.0	†92.8
Ablation	medium set w/o norm.	95.2	98.5	95.3	**99.6**	†84.0	99.8	94.4	†73.6	**94.4**	†69.0	**94.0**	†68.0	†**92.4**	‡60.5	†**94.0**	‡58.2

TABLE II

TOTAL PERFORMANCE AND REAL-TIME FACTOR OF PROPOSED METHOD

	Baseline		Proposed		Ablation
	Silero	PyAnnote	medium set	full set	medium w/o norm.
F1-score	90.51	92.48	**94.24**	94.14	89.80
Accuracy	89.45	90.75	**93.36**	93.20	89.64

	GPU: 1-board	CPU: 1-core	CPU: 2-core	CPU: 4-core
RTF	0.019	0.125	0.070	0.039

in HMM were set to 0.5 that means no prior knowledge, and the probability of self-transition was set to 0.99 manually.

C. Results

The evaluation metrics were frame-wise F1-score and accuracy, which were calculated as follows. First, the detected speech sections and ground-truth sections were converted into a sequence of binary voice-activities (0 or 1) at every 0.1-s frame for each test signals. The binary labels of the detection and the ground-truth compared to calculate the metrics by using the software [40] with 95% confidence-interval computation (without "condition" setting). The number of samples (labels) before applying amplitude changes was 1,469,155 for the non-speech set, 695,397 for the speech set, and 756,037 for the Hazumi set, respectively. **F1-score of non-active labels was used for non-speech set**. Here, the sections of system utterances in the Hazumi set were eliminated on the basis of their annotations.

Table I lists the F1-score results with the selected amplitudes 1.0 and 0.05. The baselines were Silero and PyAnnote, with PyAnnote listed as a *reference* due to its batch processing. The performance of our method without normalization processes was also investigated as an ablation study.

We found that the two baselines have some weaknesses. The change of signal amplitude and the speech distortion seriously degraded the performance of Silero by a maximum of about 20 points when we focus on the results of *rev.+bgn.* set in the Speech set. As for the Non-speech set, the performances under the small scale condition (0.05) were better than those under large scale condition (1.00). This indicates that the information of the absolute signal amplitude (not SNR) was implicitly used as a feature for classification in the Silero model. The performance of PyAnnote also degraded in the sets of WHAM! and *rev.+bgn.*, which may come from the training data set and the problem setting of PyAnnote. The performances of our methods with medium and full set training data were stable while our method *without normalization (ablation study)* did not work for different amplitude signals (over 20 points degradation for speech and dialogue set) even with the same number of parameters. These performance improvements related with the signal amplitude were brought mainly by our mathematical scale-invariant transformation, and slightly by the different number of parameters of each model.

Table II lists the macro-averaged F1-score and accuracy over non-speech, speech and dialogue sets as summary (data with four amplitude settings contained in each set). Our methods both with medium and full set outperformed others in both F1-sore and accuracy from 1 to 3 points. The performance of medium set model was slightly better than that of full set one just for data sets used in this experiment. Real-time factors (RTFs) with GPU (GeForce RTX 2080 Ti) and CPU (Intel Core i9-9980XE 3.0GHz) of our method are also shown. All RTFs were less than 0.15, and RTF with 2-core (2 threads) was less than 0.1. Thus, our VAD will not cause serious delays.

Limitations of our method are that 1) the balance between precision and recall depends on data set, 2) evaluations were limited to Japanese data, and 3) network structures were not optimized. As for the language dependency, our model may not cause serious performance degradations compared with natural language processing area because the speech feature used in ASR is usually similar among different languages. Of course, model training using multiple language data set will improve the VAD performance more thanks to the various of speech signals. Construction of lighter models is also a future work.

V. CONCLUSION

Online voice activity detection (VAD) is an important front-end for spoken dialogue systems. Different signal amplitudes and speech distortions under various environments cause performance degradation of neural VAD models due to the model mismatch. We addressed the two problems by scale-invariant normalization and data augmentation of various speech data, respectively. Our VAD outperformed Silero and PyAnnote for various datasets including a real spoken dialogue dataset. Future work includes the evaluation with other languages and spoken dialogue dataset.

ACKNOWLEDGMENT

This work was partly supported by JSPS KAKENHI Grant Numbers JP23H03457 and JP22H00536, and JST Moonshot R&D Grant Number JPMJPS2011, Japan.

REFERENCES

[1] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201–212, 1976. DOI: 10.1109/TASSP.1976.1162800.

[2] Y. Shangguan, R. Prabhavalkar, H. Su, J. Mahadeokar, Y. Shi, *et al.*, "Dissecting user-perceived latency of on-device E2E speech recognition," in *Proc. of Interspeech*, 2021, pp. 4553–4557. DOI: 10.21437/Interspeech.2021-1887.

[3] R. Higashinaka, T. Minato, K. Sakai, T. Funayama, H. Nishizaki, and T. Nagai, "Spoken dialogue system development at the dialogue robot competition," *The Journal of The Acoustic Society of Japan*, vol. 77, no. 8, pp. 512–520, 2021, ISSN: 03694232. DOI: 10.20697/jasj.77.8_512.

[4] T. Iio, Y. Yoshikawa, M. Chiba, T. Asami, Y. Isoda, and H. Ishiguro, "Twin-robot dialogue system with robustness against speech recognition failure in human-robot dialogue with elderly people," *Applied Sciences*, vol. 10, no. 4, 2020, ISSN: 2076-3417. DOI: 10.3390/app10041522.

[5] K. Komatani, R. Takeda, and S. Okada, "Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus," in *Proc. of Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2023, pp. 104–113. DOI: 10.18653/v1/2023.sigdial-1.9.

[6] Silero Team, *Silero VAD: Pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier*, https://github.com/snakers4/silero-vad, 2021.

[7] H. Bredin, R. Yin, J. M. Coria, G. Gelly, *et al.*, "Pyannote.audio: Neural building blocks for speaker diarization," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 7124–7128.

[8] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. of Interspeech*, 2021, pp. 3111–3115.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362, ISBN: 026268053X.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

[11] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.

[12] Y. LeCun, B. Boser, J. Denker, *et al.*, "Handwritten digit recognition with a back-propagation network," in *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 2, 1989, pp. 396–404.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

[14] L. Ye, W. Tong, C. Huijuan, and T. Kun, "Voice activity detection in non-stationary noise," in *Proc. of Multiconference on Computational Engineering in Systems Applications*, vol. 2, 2006, pp. 1573–1575. DOI: 10.1109/CESA.2006.4281886.

[15] J.-C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognizer," in *Proc. of Eurospeech*, 1991, pp. 1371–1374. DOI: 10.21437/Eurospeech.1991-313.

[16] M. Baig, S. Masud, and M. Awais, "Support vector machine based voice activity detection," in *Proc. of International Symposium on Intelligent Signal Processing and Communications*, 2006, pp. 319–322. DOI: 10.1109/ISPACS.2006.364896.

[17] R. Zazo, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform CLDNNs for voice activity detection," in *Proc. of Interspeech*, 2016, pp. 3668–3672. DOI: 10.21437/Interspeech.2016-268.

[18] P. R. Gudepu, J. M. Koroth, K. Sabu, and M. A. B. Shaik, "Dynamic encoder RNN for online voice activity detection in adverse noise conditions," in *Proc. of Interspeech*, 2023, pp. 5052–5056. DOI: 10.21437/Interspeech.2023-2466.

[19] M. Shi, Y. Shu, L. Zuo, *et al.*, "Semantic VAD: Low-latency voice activity detection for speech interaction," in *Proc. of Interspeech*, 2023, pp. 5047–5051. DOI: 10.21437/Interspeech.2023-598.

[20] X.-L. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2016. DOI: 10.1109/TASLP.2015.2505415.

[21] R. Masumura, K. Matsui, Y. Koizumi, T. Fukutomi, T. Oba, and Y. Aono, "Context-aware neural voice activity detection using auxiliary networks for phoneme recognition, speech enhancement and acoustic scene classification," in *Proc. of European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5. DOI: 10.23919/EUSIPCO.2019.8902703.

[22] Y. Sudo, S. Muhammad, K. Nakadai, J. Shi, and S. Watanabe, "Streaming Automatic Speech Recognition with Re-blocking Processing Based on Integrated Voice Activity Detection," in *Proc. Interspeech*, 2022, pp. 4641–4645. DOI: 10.21437/Interspeech.2022-11216.

[23] S. Haykin, *Adaptive Filter Theory*, 4th. Upper Saddle River, NJ 07458: Prentice-Hall, 1991.

[24] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 161–174, 1994. DOI: 10.1109/89.260359.

[25] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transaction," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 24–29.

[26] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, *et al.*, "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, *Layer normalization*, arXiv:1607.06450, 2016. arXiv: 1607.06450 [stat.ML].

[29] N. V. Chawla, K. W. Bowyer, *et al.*, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[30] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[31] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. of Annual ACM Conference on Multimedia*, Oct. 13, 2015, pp. 1015–1018, ISBN: 978-1-4503-3459-4. DOI: 10.1145/2733373.2806390.

[32] G. Wichern, J. Antognini, M. Flynn, *et al.*, "WHAM!: Extending speech separation to noisy environments," in *Proc. of Interspeech*, 2019, pp. 1368–1372.

[33] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, *JVS corpus: Free Japanese multi-speaker voice corpus*, arXiv:1908.06248, 2019. arXiv: 1908.06248 [cs.SD].

[34] D. Snyder, G. Chen, and D. Povey, *MUSAN: A Music, Speech, and Noise Corpus*, arXiv:1510.08484v1, 2015. eprint: 1510.08484.

[35] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. of the Second International Conference on Language Resources and Evaluation (LREC)*, 2000.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. of Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019, pp. 8024–8035.

[37] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. of International Conference on Machine Learning*, vol. 28, 2013, pp. 1310–1318.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of International Conference on Learning Representations*, 2015.

[39] P. Izmailov, D. Podoprikhin, *et al.*, "Averaging weights leads to wider optima and better generalization," in *Proc. of Conference on Uncertainty in Artificial Intelligence*, 2018.

[40] L. Ferrer and P. Riera, *Confidence intervals for evaluation in machine learning*, https://github.com/luferrer/ConfidenceIntervals.

Fine-Grained Quantitative Emotion Editing for Speech Generation

Sho Inoue*†, Kun Zhou‡, Shuai Wang†§, and Haizhou Li*†,
*School of Data Science †Shenzhen Research Institute of Big Data
The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China
‡Alibaba Group, Singapore
E-mail: shoinoue@link.cuhk.edu.cn, kun.z@alibaba-inc.com, wangshuai@cuhk.edu.cn, and haizhouli@cuhk.edu.cn

***Abstract*—It remains a significant challenge how to quantitatively control the expressiveness of speech emotion in speech generation. In this work, we propose an approach for quantitative manipulation of the emotion rendering for emotion editing in speech generation. We apply a hierarchical emotion distribution extractor, i.e. Hierarchical ED, that quantifies the intensity of emotions at different levels of granularity. Hierarchical ED is subsequently integrated into the FastSpeech2 framework, guiding the model to learn emotion intensity at phoneme, word, and utterance levels. During synthesis, users can manually edit the emotional intensity of the generated voices. Both objective and subjective evaluations demonstrate the effectiveness of the proposed network in terms of fine-grained quantitative emotion editing.**

I. Introduction

Speech emotion, characterized by prosodic patterns at phoneme, word, and utterance level [1]–[3], is effectively represented through a hierarchical structure [4], [5]. These patterns, including pitch, tempo, rhythm, stress, lexical emphasis, and speaker individuality, are crucial for accurate emotion rendering [6]–[9]. Existing methods, like style tokens, represent prosody patterns but often lack interpretability, quantitativeness, and fine-grained controllability [10], [11], highlighting the need for more effective models. In neural text-to-speech, speech editing techniques empower models to alter audio segments based on user instructions, enabling fine-grained prosody prediction in synthesized speech [12], [13]. However, this approach is deficient in quantitative control.

The complexity of speech emotions, as evidenced by various studies [14]–[17], indicates that merely adjusting prosodic features is inadequate. Additionally, the hierarchical nature of speech emotion [6], with unique characteristics at utterance [18], word [19], and phoneme [6] levels, calls for a multi-level approach for accurate and human-like emotional expression in speech synthesis. Previous studies on text-to-speech and emotional voice conversion [17], [20], [21] also highlight the importance of multi-level modeling for speech emotions. Thus, a quantifiable method capable of modeling the hierarchical structure of emotions is imperative for effective and nuanced emotion rendering in speech synthesis.

In our previous study [22], we investigated a fine-grained emotion control approach for text-to-speech (TTS) systems. In this paper, we further extend the approach to enable quantitative emotion editing. This provides an intuitive and quantitative method for manipulating emotion rendering at the phoneme, word, and utterance levels. Compared to our previous work [22], which integrated fine-grained embedding solely into the variance adaptor of FastSpeech2, our proposed emotion editing approach allows for fine-grained and flexible human editing across any TTS model.

Our contributions are summarized as follows:

- We propose a novel fine-grained emotion editing framework that facilitates the emotion rendering in previously unseen audios and allows for the adjustment of emotion intensity for a speech generation framework;
- We thoroughly compare our approach with the baseline method of phoneme-level emotion intensity control, demonstrating our model's superior emotion expressiveness and controllability.
- At run-time, users have the flexibility to analyze and manipulate the emotion distribution from the audio sample ("emotion editing"). This approach empowers users with a quantitative method to modify emotion rendering of a spoken utterance.

The rest of this paper is organized as follows: In Section 2, we discuss the related works. Section 3 describes our proposed methodology. In Section 4, we report our experiments and results. Section 5 concludes our study.

II. Related Works

A. Control of Speech Emotion

There has been a growing interest in enabling the control of synthesized emotions [23]–[25]. Emotion-enhanced GST [23] incorporated an emotion recognition task to facilitate the modeling of emotion-related prosody. Some studies explored quantitative methods for controlling emotions through relative attributes [26]–[29]. Another study [30] delved into inter-

§Corresponding Author
The research is supported by National Natural Science Foundation of China (Grant No. 62271432); Internal Project of Shenzhen Research Institute of Big Data (Grant No. T00120220002); Shenzhen Science and Technology Program ZDSYS20230626091302006; Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002); and CCF-NetEase ThunderFire Innovation Research Funding (No. CCF-Netease 202302).

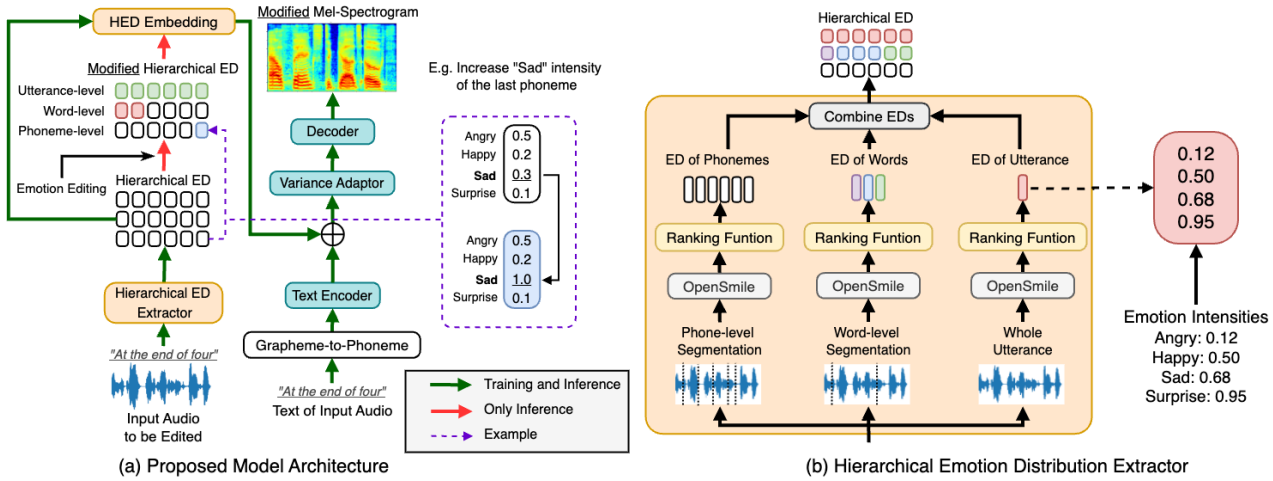


Fig. 1: (a) Model architecture with hierarchical emotion distribution (HED) mechanism in TTS. During inference, the framework extracts hierarchical ED from input audio ("emotion editing"). Users can manually modify hierarchical ED to control emotion intensities at phoneme, word, and utterance levels; (b) Hierarchical ED extraction workflow for emotion distributions at phoneme, word, and utterance levels.

and intra-class distances to achieve fine-grained control over recognizable intensity differences. EmoQ-TTS [31] adopted a distance-based intensity quantization approach to capture emotion intensity.

Unlike MsEmoTTS [32], which covers only utterance-level emotional change with fine-grained intensity changes, this work aims to offer quantitative and hierarchical control over emotions at various levels of speech units, e.g. phoneme, word, and utterance. It provides a versatile tool for editing emotion rendering in any speech segment.

B. Speech Editing

Speech editing, particularly semantic editing, focuses on altering the textual information of speech while preserving the naturalness of the synthesized output [12], [13], [33]. EditSpeech [33] employed partial inference and bidirectional fusion techniques to effectively modify speech semantics. Another research, EdiTTS [12], utilized perturbations in the Gaussian space to allow for detailed audio edits, adjusting context and pitch, and enabling user-guided prosody editing. Yet another study [13] presented a method for phoneme-level pitch adjustment and time-stretching, providing users with greater control over prosody.

This work is motivated to provide a systematic approach for precisely controlling high-level prosodic patterns, which differs from the prior studies which primarily focused on modifying the physical speech attributes, e.g. pitch and duration.

In this section, we delve into the details of the proposed fine-grained emotion editing mechanism. We begin by formulating the problems and introducing a quantitative and hierarchical emotion control mechanism. We then detail the design of the hierarchical emotion distribution (ED) extractor, along with its associated training scheme. Lastly, we explain how to render the desired emotions by editing.

III. Fine-Grained Quantitative Control

A. Problem Formulation

Given an audio input and its text transcript, we would like to develop a model to control the emotion intensity of the speech segments, as illustrated in Fig.1(a). This approach is expected to work for the rendering of a single emotion or a mixture of emotions and to work seamlessly with various text-to-speech frameworks. Typically, we label emotions at the utterance level in most speech databases without paying attention to the nuanced variations in emotion intensity. It remains a significant challenge as to how to achieve effective and quantifiable control over speech emotion. To address this, we aim to automatically generate fine-grained and quantitative intensity labels that act as 'soft labels' for speech generation models. In this way, we eliminate the need for manual labeling of emotional intensity. Our approach is effective in emotion intensity control and mixed-emotion rendering, which can be easily adapted to any speech generation frameworks including text-to-speech and voice conversion.

B. Hierarchical Emotion Distribution Extractor

Following [22], we apply a hierarchical emotion distribution extraction module consisting of an OpenSmile feature extractor and a pre-trained ranking function for each segmental level, which automatically quantifies the intensities of each emotion type in an utterance. This approach, grounded in the concept of relative attributes [34], allows us to measure the prominence of specific emotions in speech. By treating emotion style as a speech attribute, we model and rank the presence of various emotions relative to the other emotions.

In particular, we define the ranking function as $f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i + b$, where $\mathbf{x}_i$, $\mathbf{w}$, and b denote the acoustic features of the i-th training sample, the weight vector, and bias, respectively. The parameters ($\mathbf{w}$ and b) are optimized using

the support vector machine's objective function for binary classification (e.g. Angry and Non-angry classification) [35]. To provide a quantifiable measure, we further normalize the values obtained from the ranking function to a range of $[0, 1]$, where a larger value represents a stronger emotion intensity. Our methodology facilitates the labeling of training data with continuous emotion intensities across various emotions. Additionally, it allows for the quantification and regulation of emotion intensity in unseen utterances during run-time.

The hierarchical ED extractor consists of pre-trained ranking functions as illustrated in Fig.1(b). The extractor first segments the input audio signal into phoneme, word, and utterance levels via Montreal Forced Aligner [36]. Then, it extracts an 88-dimensional audio feature set for each temporal segment with OpenSMILE [37]. The pre-trained ranking function automatically estimates an ED vector for each audio segment. Each value in this vector represents the intensity of a specific emotion encoded in the audio segment. In practice, the utterance-level emotion distribution is duplicated across all phonemes, and the word-level emotion distribution is replicated across the corresponding phonemes. Once formulated, the hierarchical ED vectors are integrated with linguistic embeddings in the variance adaptor during training.

C. Hierarchical Emotion Intensity Modeling

We integrate the aforementioned hierarchical ED extractor into the FastSpeech2 framework, which comprises a text encoder, variance adaptor, and decoder, as depicted in Fig.1(a). During training, given an input pair <audio, text>, this extractor predicts the emotion intensity at phoneme, word, and utterance levels from the audio. These predictions form a hierarchical ED as illustrated in Fig.1 (b). The text encoder converts the phoneme sequence into text embeddings. The variance adaptor, utilizing both text and ED embeddings, predicts pitch, duration, and energy. Subsequently, the decoder reconstructs the Mel-spectrogram, guided by an L1 loss function. This process enables the framework to learn the hierarchical nature of emotion intensity and to establish a link between linguistic and hierarchical ED information.

D. Emotion Editing

During run-time, as depicted in Fig. 1(a), our framework enables emotion editing. For unseen <audio, text> inputs, it derives the hierarchical Emotion Distribution (ED) from the audio signal. Users can then edit the emotion rendering by adjusting emotion distributions at three different levels (Hierarchical ED).

IV. Experiments and Results

A. Experimental Setup

We conduct our experiments on two datasets: Blizzard Challenge 2013 dataset, i.e. Blizzard [38], and Emotion Speech Dataset, i.e. ESD [39], [40]. Blizzard, audiobooks read by a single female speaker, offers approximately 42 hours of expressive speech with varied prosody but lacks emotion labels. We randomly select 200 samples from four audiobooks for testing. On the other hand, ESD comprises over 29 hours of emotional speech in five categories—Neutral, Angry, Happy, Sad, and Surprise—from 20 speakers, including 10 native English and 10 Mandarin speakers. Our study exclusively employs English recordings. For training the hierarchical ED extractor's ranking functions, we select 20 random samples per speaker and emotion from ESD.

TABLE I: MOS with 95% confidence interval

Ground Truth	MsEmoTTS	**Hierarchical ED (ours)**
4.211±0.142	3.544±0.138	**3.596±0.141**

B. Model Architecture

We use FastSpeech2 as the backbone framework [41], which consists of a text encoder, variance adaptor, and decoder. A transformer [42] based text encoder converts an input phoneme sequence into a linguistic embedding. Variance adaptors use feed-forward networks for the duration, pitch, and energy prediction. A transformer-based decoder synthesizes a mel-spectrogram from the variance-adapted features. The loss function combines the L1 loss between the predicted and target mel-spectrogram and the mean squared error loss from predicted prosodic features. We opt for the Adam optimizer [43], setting a batch size of 32 and conducting 800,000 iterations of training over 48 hours on a single GPU. The ED embedding layers comprise fully connected layers with a Tanh activation function. We adopt the text-based emotion intensity predictor from MsEmoTTS [32], which uses two 1D convolution layers, layer normalization, and dropout. HiFiGAN [44] serves as the vocoder, pre-trained on the Blizzard dataset.

C. Results and Analysis

We conduct both objective and subjective evaluations, focusing on speech quality, emotional expressiveness, and emotion controllability. We integrate MsEmoTTS [32] into the FastSpeech2 framework as the baseline. Our subjective evaluation involves listening experiments with 18 participants, each listening to 120 synthesized samples under specific guidelines. We suggest readers to refer to our demo page[1].

1) Speech Quality: We conduct a Mean Opinion Score (MOS) test to evaluate the overall speech quality, where a higher MOS represents better speech quality. As shown in Table I, our model demonstrates superior performance over the baseline, as indicated by its consistently higher MOS scores.

2) Emotion Expressiveness: We conduct A/B preference tests where participants select the audio that more closely matches the reference in terms of emotional expressiveness. Both our model's hierarchical ED and the baseline's emotion intensity are derived from the same reference audio. As indicated in Table II, our model outperforms the baseline with a preference rate of 43.51%.

[1]**Speech Demos**: https://shinshoji01.github.io/Hierarchical-ED-Demo/

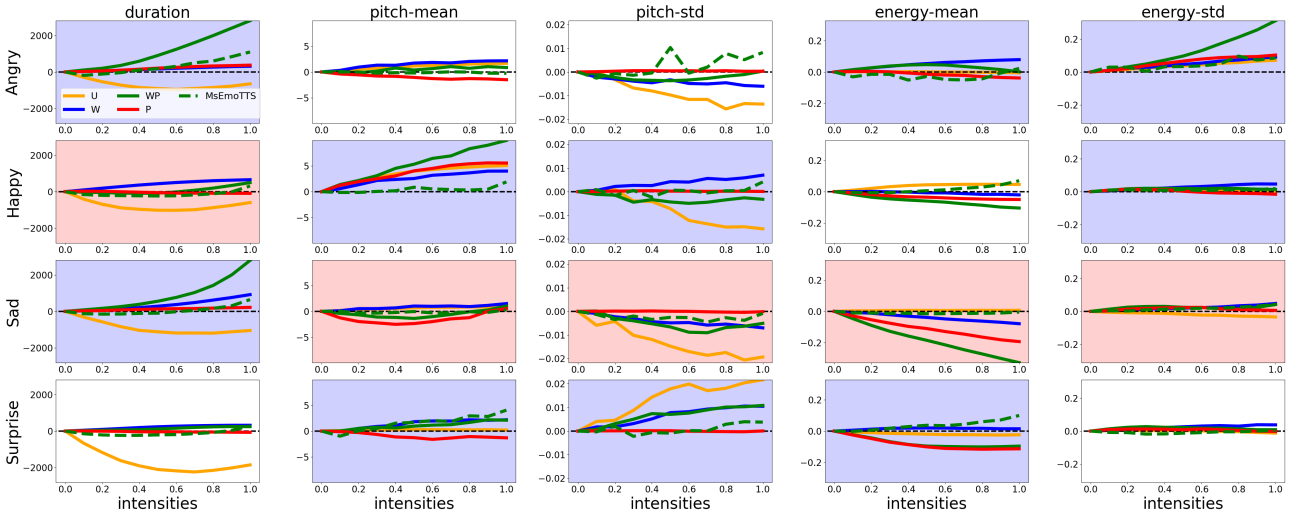


Fig. 2: The illustration of prosodic variants with intensity changes. The red background represents the expected negative trend, the blue indicates the expected positive trend, both summarized from the ESD dataset. 'U', 'W', 'P', and 'WP' indicate the controlling segments in Hierarchical ED (our model): Utterance, Word, Phoneme, and Word-and-Phoneme, respectively.

TABLE II: A/B preference test for emotion expressiveness with 95 % confidence interval.

MsEmoTTS	Neutral	**Hierarchical ED (ours)**
28.77±5.256	27.72±5.197	**43.51±5.756**

TABLE III: Objective evaluation for emotion expressiveness

	MCD	Pitch (10^1)	Energy (10^{-2})	FD (10^1)
MsEmoTTS	4.783±0.135	1.215±0.064	4.884±0.225	7.841±0.842
HED (ours)	**4.348±0.109**	**1.151±0.052**	**4.018±0.162**	**6.881±0.710**

To further objectively evaluate emotion expressiveness, we calculate various metrics: (1) Mel-Cepstral Distortion (MCD)[45] to measure spectral similarity, (2) Pitch/Energy Distortion for assessing prosody alignment, and (3) Frame Disturbance (FD)[46] to evaluate duration deviation. The results summarized in Table III confirms our framework's superiority over the baseline MsEmoTTS in replicating target emotions, thereby validating its effectiveness in emotion modeling.

3) Emotion Controllability: We conducted best-worst scaling (BWS) tests [47] to compare word-level emotion controllability between our model and the baseline. In these tests, we modified the emotion intensity of three words in each utterance to 0.0, 0.5, and 1.0. Evaluators listened to the audio samples and selected the least and most expressive ones. As shown in Table IV, our model demonstrated a distinct preference for the least expressive sample at the lowest intensity and the most expressive sample at the highest intensity, particularly for Sad and Surprise emotions. This result highlights our model's capability to effectively differentiate intensity variations.

We further validated the controllability of our model across fine-grained emotional variances, examining them at utterance, word, phoneme, and word-and-phoneme combination levels. We increased the emotion intensity values from 0.0 to 1.0 and subsequently computed various prosody features, such as duration, mean/standard deviation of pitch, and mean/standard deviation of energy, shown in Fig.2. These prosodic characteristics are known to correlate significantly with emotion intensity, as indicated in the literature [5] For example, sadness often manifests in a slower speaking rate and lower values for pitch and energy mean/standard deviation. We performed an analysis of ESD to explore the relationship between acoustic features and various emotions, using different colors in Figure 2 to indicate positive and negative correlations. A red background represents an expected negative trend, while a blue background signifies a positive trend as intensity increases. We have observed that our proposed model closely aligns with the expected trends. For example, ESD reveals a positive correlation between happiness and mean pitch, as well as a negative correlation between sadness and mean energy. Our models can effectively adjust these features as emotion intensity varies.

Furthermore, we observe significant prosodic changes when editing both word and phoneme-level emotions and the

TABLE IV: BWS Test Result: The value represents evaluator preferences (%), with red and blue indicating the heatmap for the least expressive and most expressive audio, respectively.

		Hierarchical ED (ours)				MsEmoTTS			
		Ang	Hap	Sad	Sur	Ang	Hap	Sad	Sur
Least	0.0	79	63	67	81	42	32	21	33
	0.5	0	28	14	14	47	47	54	42
	1.0	21	9	19	5	11	21	25	25
Most	0.0	11	18	16	7	11	12	30	18
	0.5	16	7	9	7	26	16	23	30
	1.0	74	75	75	86	63	72	47	53

changes in the standard deviation of pitch at the utterance level align with our expectations. The figure also illustrates that Hierarchical ED outperforms the baseline model, especially when controlling word and phoneme levels simultaneously. Our model exhibits a significantly closer alignment with the expected trend.

V. Conclusion

We introduce a fine-grained emotion editing approach for speech generation tasks. At run-time, users can quantitatively control emotion rendering at phoneme, word, and utterance levels. Both objective and subjective evaluations validate the effectiveness of our proposed idea. As a future work, we aim to enhance paragraph-level information processing and develop emotional text-to-speech models featuring robust emotion consistency. Additionally, the Hierarchical ED framework holds potential for application in diverse speech generation frameworks and contexts, including voice conversion tasks.

References

[1] A. Triantafyllopoulos, B. W. Schuller, G. İymen, *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proceedings of the IEEE*, 2023.

[2] Z. KUN, "Emotion modelling for speech generation," 2022.

[3] J. Hirschberg, "Pragmatics and intonation," *The handbook of pragmatics*, pp. 515–537, 2006.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[5] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[6] S. R. Krothapalli and S. G. Koolagudi, "Emotion recognition using prosodic information," in *Emotion Recognition using Speech Features*. New York, NY: Springer New York, 2013, pp. 79–91, ISBN: 978-1-4614-5143-3. DOI: 10.1007/978-1-4614-5143-3_5. [Online]. Available: https://doi.org/10.1007/978-1-4614-5143-3_5.

[7] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," *Proc. Interspeech 2020*, pp. 3416–3420, 2020.

[8] K. Zhou, B. Sisman, and H. Li, "Vaw-gan for disentanglement and recomposition of emotional elements in speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 415–422.

[9] Z. Du, B. Sisman, K. Zhou, and H. Li, "Disentanglement of emotional style and speaker identity for expressive voice conversion," *arXiv preprint arXiv:2110.10326*, 2021.

[10] C. Du and K. Yu, *Rich prosody diversity modelling with phone-level mixture density network*, 2021. arXiv: 2102.00851 [cs.SD].

[11] N. Tits, *Controlling the emotional expressiveness of synthetic speech: A deep learning approach*, 2022.

[12] J. Tae, H. Kim, and T. Kim, *Editts: Score-based editing for controllable text-to-speech*, 2022. arXiv: 2110.02584 [cs.SD].

[13] M. Morrison, L. Rencker, Z. Jin, N. J. Bryan, J.-P. Caceres, and B. Pardo, *Context-aware prosody correction for text-based speech editing*, 2021. arXiv: 2102.08328 [eess.AS].

[14] K. Zhou, B. Sisman, C. Busso, B. Ma, and H. Li, *Mixed-evc: Mixed emotion synthesis and control in voice conversion*, 2023. arXiv: 2210.13756 [eess.AS].

[15] Y. Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, 2011.

[16] J. Latorre and M. Akamine, "Multilevel parametric-base f0 model for speech synthesis," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[17] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 230–237.

[18] E. Rodero, "Intonation and emotion: Influence of pitch levels and contour type on creating emotions," *Journal of Voice*, vol. 25, no. 1, e25–e34, 2011, ISSN: 0892-1997. DOI: https://doi.org/10.1016/j.jvoice.2010.02.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199710000378.

[19] J. Hirschberg and G. Ward, "The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in english," *Journal of Phonetics*, vol. 20, no. 2, pp. 241–251, 1992, ISSN: 0095-4470. DOI: https://doi.org/10.1016/S0095-4470(19)30625-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447019306254.

[20] H. Ming, D.-Y. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion.," in *Interspeech*, 2016, pp. 2453–2457.

[21] S. Lei, Y. Zhou, L. Chen, *et al.*, *Towards multi-scale speaking style modelling with hierarchical context information for mandarin speech synthesis*, 2022. arXiv: 2204.02743 [cs.SD].

[22] S. Inoue, K. Zhou, S. Wang, and H. Li, "Hierarchical emotion prediction and control in text-to-speech synthesis," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 601–10 605. DOI: 10.1109/ICASSP48485.2024.10445996.

[23] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, *Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition*, 2021. arXiv: 2010.13350 [eess.AS].

[24] T. Li, S. Yang, L. Xue, and L. Xie, *Controllable emotion transfer for end-to-end speech synthesis*, 2020. arXiv: 2011.08679 [cs.SD].

[25] A. Rabiee, T.-H. Kim, and S.-Y. Lee, "Adjusting pleasure-arousal-dominance for continuous emotional text-to-speech synthesizer," in *Interspeech*, 2019. [Online]. Available: https : / / api . semanticscholar . org / CorpusID:189762268.

[26] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 192–199. DOI: 10.1109/ASRU46091.2019.9003829.

[27] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 31–48, 2023. DOI: 10.1109/TAFFC.2022.3175578.

[28] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3120–3134, 2023. DOI: 10.1109/TAFFC.2022.3233324.

[29] Y. Lei, S. Yang, and L. Xie, *Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis*, 2020. arXiv: 2011.08477 [cs.SD].

[30] S. Wang, J. Guonason, and D. Borth, "Fine-grained emotional control of text-to-speech: Learning to rank inter- and intra-class emotion intensities," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10097118.

[31] C.-B. Im, S.-H. Lee, S.-B. Kim, and S.-W. Lee, "Emoq-tts: Emotion intensity quantization for fine-grained controllable emotional text-to-speech," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6317–6321. DOI: 10.1109/ICASSP43922.2022.9747098.

[32] Y. Lei, S. Yang, X. Wang, and L. Xie, *Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis*, 2022. arXiv: 2201.06460 [cs.SD].

[33] D. Tan, L. Deng, Y. T. Yeung, X. Jiang, X. Chen, and T. Lee, *Editspeech: A text based speech editing system using partial inference and bidirectional fusion*, 2021. arXiv: 2107.01554 [eess.AS].

[34] D. Parikh and K. Grauman, "Relative attributes," in *2011 International Conference on Computer Vision*, IEEE, 2011, pp. 503–510.

[35] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[36] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502. DOI: 10.21437/Interspeech.2017-1386.

[37] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile – the munich versatile and fast open-source audio feature extractor," Jan. 2010, pp. 1459–1462. DOI: 10.1145/1873951.1874246.

[38] S. King and V. Karaiskos, "The blizzard challenge 2013," 2014.

[39] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[40] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 920–924.

[41] Y. Ren, C. Hu, X. Tan, *et al.*, *Fastspeech 2: Fast and high-quality end-to-end text to speech*, 2022. arXiv: 2006.04558 [eess.AS].

[42] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2017. arXiv: 1706.03762 [cs.CL].

[43] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG].

[44] J. Kong, J. Kim, and J. Bae, *Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis*, 2020. arXiv: 2010.05646 [cs.SD].

[45] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, 125–128 vol.1. DOI: 10.1109/PACRIM.1993.407206.

[46] B. Sisman, G. Lee, H. Li, and K. C. Tan, "On the analysis and evaluation of prosody conversion techniques," in *2017 International Conference on Asian Language Processing (IALP)*, 2017, pp. 44–47. DOI: 10.1109/IALP.2017.8300542.

[47] S. Kiritchenko and S. Mohammad, "Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, R. Barzilay and M.-Y. Kan, Eds., Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 465–470. DOI: 10.18653/v1/P17-2074. [Online]. Available: https://aclanthology.org/P17-2074.

YOLO-DC: Enhancing object detection with deformable convolutions and contextual mechanism

Dengyong Zhang*, Chuanzhen Xu*, Jiaxin Chen*, Xin Liao†
* School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China
E-mail: Dengyong Zhang, zhdy@csust.edu.cn; Chuanzhen Xu, chuanzhen@stu.csust.edu.cn; Jiaxin Chen, jxchen@csust.edu.cn
† College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
E-mail: xinliao@hnu.edu.cn

***Abstract*—Existing object detection methods often optimize model structure, loss functions, and data preprocessing, but enhancements via convolutional techniques are often overlooked. Additionally, increasing network depth leads to the loss of key features. This paper introduces YOLO-DC, which enhances object detection with deformable convolution and contextual mechanisms. YOLO-DC includes a Deformable Convolutional Module (DCM) and a Contextual Fusion Downsampling Module (CFD). The DCM module uses a modified deformable convolution of Multiscale Spatial Channel Attention (MSCA) to extend the receptive field and improve feature extraction. The CFD integrates contextual and local features post-downsampling, reducing information loss. Compared to YOLOv8-N, YOLO-DC-N improves average precision (AP) by 3.5%, reaching 40.8% on the Microsoft COCO 2017 dataset, with similar FPS and inference time. The model demonstrates superior performance compared to other state-of-the-art detection algorithms. The source code is available at: https://github.com/Object-Detection-01/YOLO-DC.git.**

I. Introduction

Current object detectors rely on Convolutional Neural Networks (CNNs) and transformer-based models. Transformer models like DETR, DINO, and RT-DETR [1]–[3] captured long-term dependencies and achieved high accuracy but needed speed improvements compared to CNNs. The YOLO family, based on CNNs, balanced performance and efficiency. Since YOLOv1, enhancements like DarkNet [4], CSPNet [5], and ELAN [6] improved feature extraction. Enhancements like SPP [7], PANet [8], RepGFPN [9], and GD [10] enhanced multi-scale feature fusion. YOLOX [11] and YOLOv10 [12] improved performance by optimizing label allocation strategy. YOLO-MS improved accuracy with deep convolution [13].

Researchers proposed various convolution techniques to balance performance and efficiency, such as group convolution [14], dilated convolution [15], depthwise separable convolution [16], and deformable convolution [17]–[20]. Deformable convolution (DCNv1) adjusts the receptive field by learning offsets for each sample point [17]. Subsequent versions, DCNv2 [18], DCNv3 [19], and DCNv4 [20], introduced improvements like weight coefficient masking, weight sharing, and removal of softmax normalization. Additionally, attention and contextual mechanisms are crucial for improving model performance. Attention mechanisms such as CA [21], EMA [22], and EPSANet [23] encode semantic information at different scales, while GCNet [24], CGNet [25], and F-SSD [26] enhance performance by combining multiscale semantic information with contextual mechanisms.

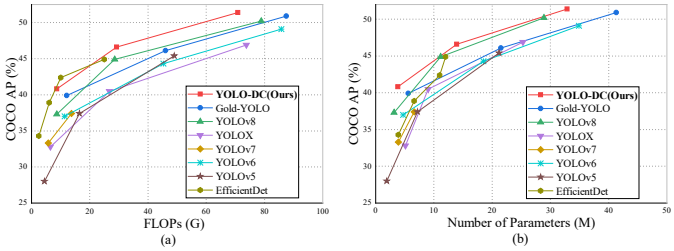


Fig. 1. Comparison with other SOTA object detection models on the COCO dataset: (a) AP vs. FLOPs; (b) AP vs. parameters. Notably, Faster R-CNN, RetinaNet, and DETR-DC models are not shown due to their high computational and parameter requirements. However, YOLO-DC outperforms these models, as detailed in Table I.

Despite significant progress by these techniques, the optimization space for convolutional enhancement was neglected in YOLOv8. In addition, increasing network depth led to information loss during feature extraction, causing biased gradient flow and affecting accuracy [27]. Consequently, YOLOv8 achieved sub-optimal performance. To address these issues, this paper proposes YOLO-DC, an object detection algorithm based on YOLOv8 [28]. As shown in Fig. 1, YOLO-DC achieves an optimal balance between computational load and parameter count, with superior performance compared to other state-of-the-art (SOTA) models. The main contributions of this paper include:

1) Improving DCNv2 by introducing a Multi-scale Spatial Channel Attention (MSCA), leading to the proposed DCN-MSCA convolution, which accurately extends the receptive field and constitutes a deformable convolutional module (DCM) for improved feature extraction.
2) Proposing the Context Fusion Downsampling (CFD) module to integrate contextual information and reduce redundancy after downsampling to improve information utilization.
3) Presenting YOLO-DC using DCM and CFD modules, which significantly outperforms existing SOTA models in comprehensive performance.

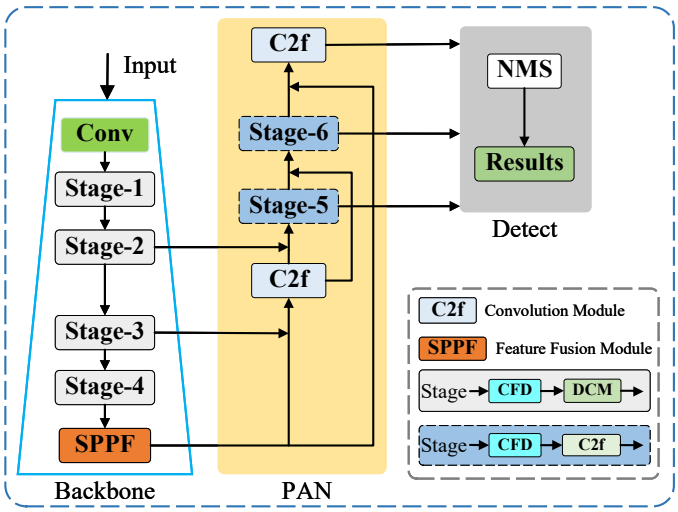


Fig. 2. YOLO-DC network structure diagram.

II. Proposed Method

A. Model Architecture

This paper presents YOLO-DC, an object detector based on YOLOv8, designed to enhance detection accuracy and address limitations in YOLOv8 convolution and information loss. The key components are the DCM and CFD modules. The CFD module replaces the downsampling operation in YOLOv8 [28], and the DCM module replaces the C2f module in the backbone. In each stage of the backbone network, the DCM provides robust feature extraction and an improved receptive field, while the CFD module reduces information loss during downsampling by enhancing features before feeding them into the DCM. In the PAN structure, each stage is fed into the C2f module after passing through the CFD module. This integrated approach significantly improves detection accuracy. The YOLO-DC network structure is shown in Fig. 2.

B. Contextual Fusion Downsampling Module (CFD)

The CFD module plays a crucial role in downsampling and information integration. It combines local characteristics with contextual semantic information and refines them with global contextual information, enhancing utilization efficiency. Positioned before the DCM module in the backbone and the C2f module in the neck, the CFD spans the entire network from spatial to semantic levels, facilitating rapid and accurate feature transfer and preventing information loss.

The CFD module first performs a 3×3 convolution with a stride of 2 for downsampling, followed by a regular convolution to collect local features and a dilated convolution to capture contextual features. It then passes through a joint feature extractor, including a concatenation layer, batch normalization, and SiLU activation to fuse local and contextual features. Finally, a global feature extractor, consisting of a global pooling layer and two fully connected layers, extracts features and generates a weight vector to guide joint feature fusion and produce the final output (See Fig. 3). Inspired by the Context Guided Block (CG Block) in CGNet [25], the CFD module combines surrounding contextual information with local features. While CG blocks are suitable for

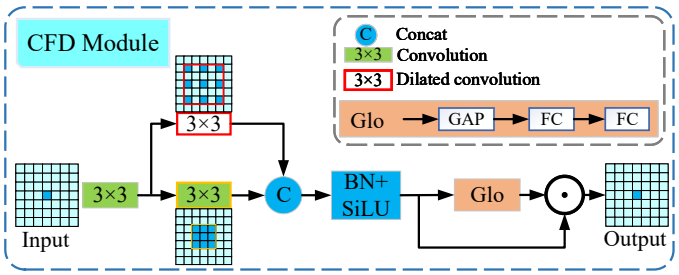


Fig. 3. CFD module schematic.

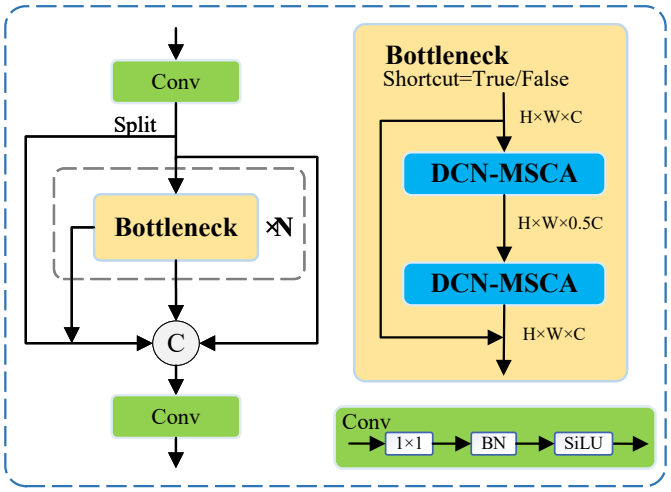


Fig. 4. DCM module schematic.

lightweight models, they may lack pronounced feature extraction capability. To maximize contextual information, the CFD module incorporates downsampling operations before the main convolutional module, facilitating comprehensive contextual information integration. This fused information enhances feature extraction and improves overall information extraction capability without significantly increasing computational load.

C. Deformable Convolution Module (DCM)

The DCM module integrates principles from the C2f module [28], preserving gradient flow. It combines residual connections with the CSPNet [5] structure to enhance learning. The DCM performs a convolution followed by a split into two branches. One branch passes through multiple Bottleneck modules with residual connections, then concatenates and fuses with the other branch and the output residuals of the Bottleneck module. The Bottleneck module in DCM employs DCN-MSCA to reduce irrelevant regions and improve feature extraction. The shortcut in the Bottleneck module connects at the backbone and disconnects at the neck (see Fig. 4).

Initially, we constructed the DCM with DCNv2, which slightly improved performance over the baseline model using normal convolution. However, the single convolution process in DCNv2 led to irrelevant regions. To address this, we added a Multiscale Spatial Channel Attention (MSCA) mechanism to DCNv2, enhancing offset attention in the X and Y directions. The improved DCN-MSCA method reconfigures offset generation in DCNv2, resulting in better accuracy and fewer

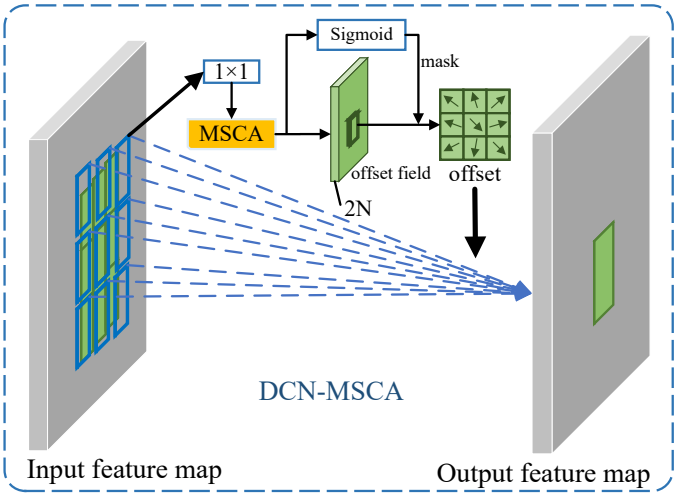


Fig. 5. DCN-MSCA convolution schematic.

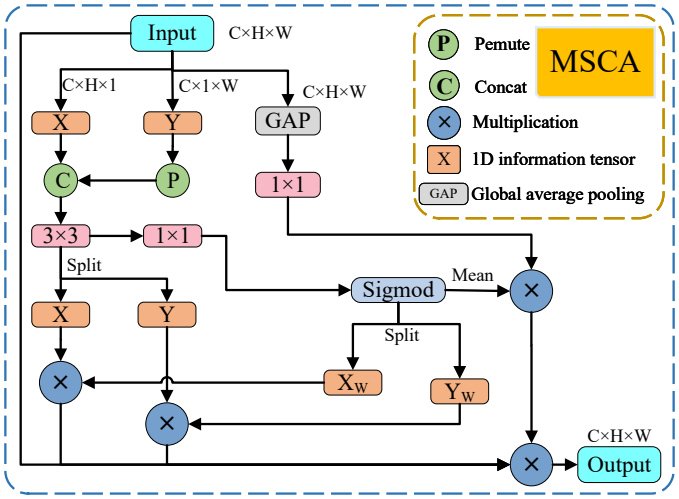


Fig. 6. MSCA mechanism schematic.

irrelevant regions (see Fig. 5).

The MSCA mechanism extends Coordinate Attention (CA) to tackle information loss by incorporating multiscale fusion. While CA decomposes the global pool into one-dimensional feature encoding for two spatial directions, its direct decomposition using two global pooling operations is somewhat crude. The subsequent splicing-then-convolution operation inadequately facilitates information exchange, contributing to suboptimal performance [21]. Our MSCA mechanism addresses this by integrating multiscale information to enhance accuracy by analyzing both spatial and channel information.

MSCA integrates multiscale information via CA, extracting weights from intermediate processes for bidirectional fusion. This enhances information interaction, with global information guiding both directions to mitigate loss. MSCA consists of three branches, each handling multiscale information at different gradients. The first branch transforms global pooling into one-dimensional codes for X and Y directions. The second branch convolves the first, extracting X_{weight} and Y_{weight} after Sigmoid activation to refine fusion accuracy. The third branch conducts global pooling on initial inputs, guiding encoded information from the second branch. The integrated output combines information from all three branches (see Fig. 6).

III. Experiment Evaluation

A. Experiment Setups

1) Datasets: We validated the performance of YOLO-DC using the Microsoft COCO 2017 dataset, the RUOD underwater object detection dataset, and the PASCAL VOC dataset (07+12). The COCO 2017 dataset includes 80 categories with 118,287 training images and 5,000 test images. The PASCAL VOC dataset (07+12) combines VOC 2007 and VOC 2012 training and validation sets, totaling 16,551 training images and 4,952 test images. The RUOD dataset, designed for underwater detection, comprises 9,800 training images and 4,200 test images.

2) Implementation Details: YOLOv8 was used as the baseline model. YOLO-DC was trained for 500 epochs, with mosaic enhancement disabled in the last 10 epochs to improve accuracy. The SGD optimizer with baseline hyperparameters was employed, using BCE loss for classification and DFL loss with CIoU loss for box regression. To ensure accurate evaluation, training was conducted from scratch without pre-training weights. Single-scale images (640×640) were used as input, following COCO evaluation metrics, reporting normalized AP at different IoU thresholds and AP for small, medium, and large targets (AP_s, AP_m, AP_l). Experiments were conducted on 2 NVIDIA RTX 3090 GPUs using PyTorch 2.0.

B. Comparisons

To evaluate YOLO-DC, we compared it against several SOTA models on the COCO 2017 dataset, including EfficientDet [31], RetinaNet [32], DETR-DC [1], Faster R-CNN [33], YOLOv7 [6], YOLOvX [11], Gold-YOLO [10], YOLO-MS [13], and the baseline YOLOv8 [28].

The experimental results in Table I showed that each version of YOLO-DC outperformed other models on the COCO 2017 dataset at the same computational overhead, demonstrating significant improvements in overall performance. The enhanced information utilization of the CFD module and the powerful feature extraction capability of the DCM module significantly improved the performance of YOLO-DC. Compared to YOLOv8-N, the AP of YOLO-DC-N increased by 3.5% to 40.8%, while maintaining similar computation and parameter counts, with almost the same inference time. The AP of YOLO-DC increased by 1.7% and 0.2% compared to YOLOv8-S and YOLOv8-M, respectively. Additionally, all versions of YOLO-DC outperformed EfficientDet, RetinaNet, and Faster R-CNN. Compared to YOLOv7-Tiny, YOLOX-N, and Gold-YOLO-N, YOLO-DC achieved AP improvements of 3.4%, 8%, and 0.9%, respectively. Both the S and M versions of YOLO-DC showed significant AP improvements over corresponding versions of these SOTA models. Compared with DETR-DC using ResNet-50 and ResNet-101, YOLO-DC-L outperformed them by 6.9% and 5.1% AP, respectively, with much lower computational overhead and inference time. Only YOLO-MS models had higher APs than YOLO-DC among the compared models. However, each version of YOLO-DC was

TABLE I
COMPARISON OF YOLO-DC WITH OTHER SOTA MODELS ON THE COCO 2017 DATASET. INFERENCE TIME ARE MEASURED WITH A BATCH SIZE OF 32.

Method	Input Size	$AP^{val}(\%)$	$AP^{val}_{50}(\%)$	$AP_s(\%)$	$AP_m(\%)$	$AP_l(\%)$	FPS	Latency	Params	FLOPs
EfficientDet-D0	512	34.3	54.2	12.0	37.3	50.2	95	10.5 ms	3.9 M	2.5 G
EfficientDet-D1	640	38.9	57.6	16.9	43.3	55.0	69	14.5 ms	6.6 M	6.1 G
EfficientDet-D2	768	42.4	61.3	20.5	46.0	57.4	51	19.6 ms	8.1 M	11 G
EfficientDet-D3	896	44.9	64.0	25.6	48.4	58.8	32	31.5 ms	12 M	25 G
RetinaNet-50	800	35.7	55.0	18.9	38.9	46.3	14	72.6 ms	29 M	165 G
RetinaNet-101	800	37.8	57.3	20.2	41.1	49.2	10	105.5 ms	38 M	205 G
Faster R-CNN	800	27.2	48.4	6.6	28.6	45.0	8	120 ms	42 M	180 G
Faster R-CNN +++	800	34.9	55.7	15.6	38.7	50.9	2	460 ms	60 M	246 G
DETR-DC5-R50	800	43.3	63.1	22.5	47.3	61.1	12	82.4 ms	41 M	187 G
DETR-DC5-R101	800	44.9	64.7	23.7	49.5	62.3	10	96.7 ms	60 M	253 G
YOLOv7-Tiny	416	33.3	49.9	-	-	-	1196	**0.8 ms**	6.2 M	5.8 G
YOLOv7-Tiny	640	37.4	55.2	19.9	41.1	50.8	519	1.9 ms	6.2 M	13.7 G
YOLOvX-N	416	32.8	50.3	14.0	35.5	48.3	1143	**0.9 ms**	5.1 M	6.5 G
YOLOvX-S	640	40.5	59.3	23.9	45.2	53.8	396	2.5 ms	9.0 M	26.8 G
YOLOvX-M	640	46.9	65.6	29.0	51.2	60.9	179	5.6 ms	25.3 M	73.8 G
Gold-YOLO-N	640	39.9	55.9	19.7	44.1	57.0	1030	1.0 ms	5.6 M	12.1 G
Gold-YOLO-S	640	46.1	63.3	25.3	50.2	62.6	446	2.2 ms	21.5 M	46.0 G
Gold-YOLO-M	640	**50.9**	68.2	32.3	55.3	66.3	220	4.5 ms	41.3 M	87.5 G
YOLO-MS-XS	640	**43.4**	**60.4**	23.7	48.3	60.3	131	7.6 ms	4.5 M	8.7 G
YOLO-MS-S	640	46.2	**63.7**	26.9	50.5	63.0	110	9.0 ms	8.1 M	15.6 G
YOLO-MS	640	**51.0**	**68.6**	33.1	56.1	66.5	80	12.3 ms	22.2 M	40.1 G
YOLOv8-N	640	37.3	52.6	15.3	35.6	54.7	734	1.4 ms	3.2 M	8.7 G
YOLOv8-S	640	44.9	60.8	23.6	47.1	65.7	387	2.6 ms	11.2 M	28.6 G
YOLOv8-M	640	50.2	67.2	28.9	53.6	69.6	176	5.7 ms	28.9 M	78.9 G
YOLO-DC-N (Ours)	640	**40.8**	56.9	16.6	39.6	60.0	676	1.5 ms	3.9 M	8.9 G
YOLO-DC-S (Ours)	640	**46.6**	63.5	24.6	48.5	65.8	334	3.0 ms	13.9 M	29.2 G
YOLO-DC-M (Ours)	640	**50.4**	67.3	27.9	53.9	69.7	147	6.8 ms	32.9 M	70.9 G

TABLE II
COMPARATIVE EVALUATION OF YOLO-DC WITH OTHER SOTA MODELS ON THE PASCAL VOC (07+12) AND RUOD DATASETS.

Mothod	Params	FLOPs	PASCAL VOC(07+12)		RUOD	
			$AP^{val}(\%)$	$AP^{val}_{50}(\%)$	$AP^{val}(\%)$	$AP^{val}_{50}(\%)$
YOLOv5-N [29]	1.9 M	4.5 G	45.1	72.5	53.6	72.1
YOLOv5-S	7.2 M	16.5 G	53.0	76.2	58.8	79.4
YOLOv6-N [30]	4.7 M	11.4 G	60.5	82.0	59.7	84.2
YOLOv7-Tiny [6]	6.2 M	13.7 G	53.8	79.3	57.2	85.3
YOLOv8-N [28]	3.2 M	8.7 G	59.1	79.9	61.9	85.3
YOLO-DC-N (Ours)	3.9 M	8.9 G	**62.6**	**82.4**	**62.8**	**85.6**

faster; for example, the inference time for YOLO-DC-N was only 20% of that for YOLO-MS-XS.

For a comprehensive evaluation, we conducted comparative experiments on the PASCAL VOC dataset (07+12) and the RUOD dataset, including models such as YOLOv5 (N, S), YOLOv6-N, YOLOv7-Tiny, and YOLOv8-N.

Table II presents the comparison results on the PASCAL VOC (07+12) and RUOD datasets. YOLO-DC-N achieved a significantly higher AP of 62.6% on PASCAL VOC, an improvement of 3.5% over YOLOv8-N, outperforming all selected SOTA models. On the RUOD dataset, YOLO-DC-N achieved an AP of 62.8%, a 1.1% increase over YOLOv8-N. Fig. 7 shows detection results comparison between YOLO-DC and YOLOv8, highlighting the superior performance of YOLO-DC in challenging environments.

Experiments show that YOLO-DC outperforms several SOTA algorithms across multiple datasets in parameters, computation, inference time, and detection accuracy.

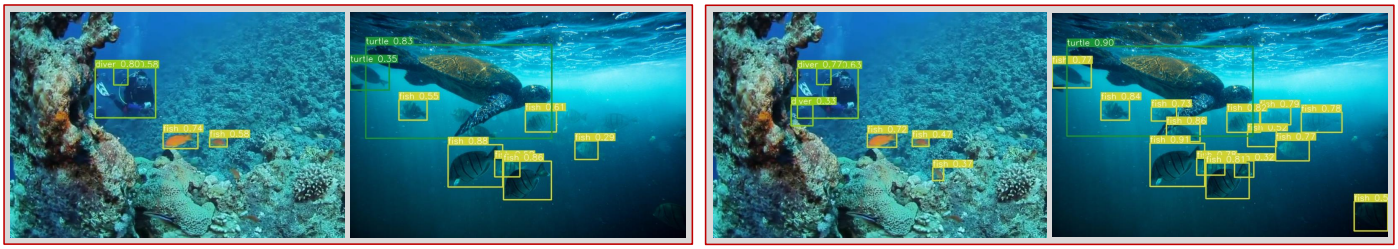
C. Ablation Study

To verify the effectiveness of our proposed improvements, we conducted ablation studies on DCNv2, DCM, CFD, and MSCA, using YOLOv8-N as the baseline model, on the Microsoft COCO 2017 dataset (Table III).

Starting with YOLOv8-N, which had an AP of 37.3%, replacing normal convolution in the C2f module with DCNv2 increased AP by 1.33% to 38.63%. Further enhancing DCNv2 with MSCA led to an additional 0.4% increase, reaching 39.03%. The CFD module alone improved AP by 1.14% to 38.44%. Combining DCM and CFD in YOLO-DC resulted in a significant AP improvement of 3.5%, reaching 40.8%. Despite a slight increase in parameters and computational load, inference time remained almost unchanged, indicating a favorable performance tradeoff. Fig. 8 illustrates the effectiveness of feature extraction for ordinary convolution, DCNv2, and DCN-MSCA in the model, highlighting the superior feature extraction capability of DCN-MSCA after applying MSCA.

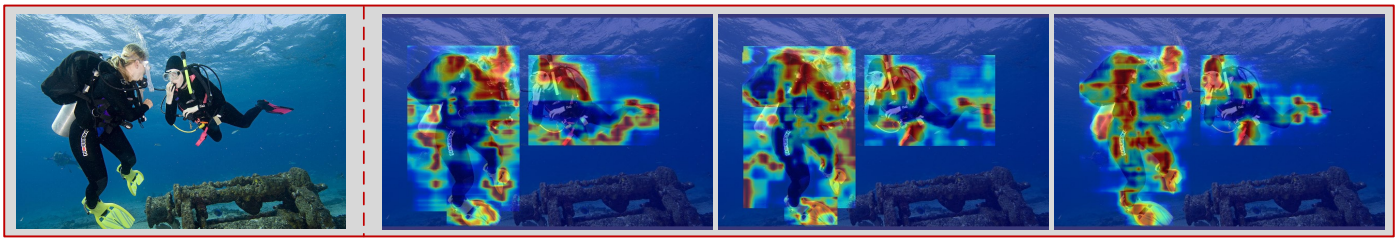
TABLE III
ABLATION STUDY ON DIFFERENT MODULES USING YOLOV8-N AS THE BASELINE MODEL: EVALUATING PERFORMANCE ON THE COCO 2017 DATASET.

Method	$AP^{val}(\%)$	$AP^{val}_{50}(\%)$	$AP_s(\%)$	$AP_m(\%)$	$AP_l(\%)$	FPS	Latency	Params	FLOPs
YOLOv8-N(Baseline)	37.30	52.61	15.28	35.59	54.66	734	1.4 ms	3.2 M	8.7 G
+DCNv2	38.63	54.44	15.62	37.63	55.91	756	1.3 ms	3.2 M	7.4 G
+DCN-MSCA(DCM)	39.03	54.97	15.72	37.93	56.32	724	1.4 ms	3.2 M	7.9 G
+CFD	38.44	53.83	15.24	36.81	55.68	789	1.3 ms	3.8 M	9.8 G
YOLO-DC(DCM+CFD)	**40.80**	**56.89**	**16.62**	**39.61**	**60.04**	676	1.5 ms	3.9 M	8.9 G



(a) YOLOv8 (b) YOLO-DC

Fig. 7. Comparison of image detection results selected from the RUOD dataset. (a) is the detection result of YOLOv8. (b) is the detection result of YOLO-DC.



(a) Input image (b) Ordinary convolution (c) DCNv2 (d) DCN-MSCA

Fig. 8. Comparison plots of feature extraction results with different convolutions using YOLOv8 as the base model. (a) shows the input image, while (b) to (d) display the visualized heatmaps (GradCAM) extracted by the model at the convolution stage of the backbone network after applying normal convolution, DCNv2, and DCN-MSCA, respectively.

IV. CONCLUSION

This paper introduces YOLO-DC, an object detection model with deformable convolution. YOLO-DC features a Deformable Convolution Module (DCM) and a Context Fusion Module (CFD). The DCM enhances feature extraction with multi-scale spatial channel attention, while the CFD reduces information loss after downsampling. Experiments show that YOLO-DC outperforms other state-of-the-art models in both general and underwater object detection.

REFERENCES

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[2] H. Zhang, F. Li, S. Liu, *et al.*, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2022.

[3] Y. Zhao, W. Lv, S. Xu, *et al.*, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16 965–16 974.

[4] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[5] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[6] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7464–7475.

[7] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-yolov4: Scaling cross stage partial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 024–13 033.

[8] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8759–8768.

[9] X. Xu, Y. Jiang, W. Chen, Y. Huang, Y. Zhang, and X. Sun, "Damo-yolo: A report on real-time object detection design," *arXiv preprint arXiv:2211.15444*, 2022.

[10] C. Wang, W. He, Y. Nie, *et al.*, "Gold-yolo: Efficient object detector via gather-and-distribute mechanism," in *Neural Information Processing Systems (NIPS)*, vol. 36, 2023, pp. 51 094–51 112.

[11] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[12] A. Wang, H. Chen, L. Liu, *et al.*, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[13] Y. Chen, X. Yuan, R. Wu, J. Wang, Q. Hou, and M.-M. Cheng, "Yolo-ms: Rethinking multi-scale representation learning for real-time object detection," *IEEE Transactions on Cognitive and Developmental Systems (TCDS)*, 2023.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, 2012.

[15] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2016.

[16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.

[17] J. Dai, H. Qi, Y. Xiong, *et al.*, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 764–773.

[18] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9300–9308.

[19] W. Wang, J. Dai, Z. Chen, *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14 408–14 419.

[20] Y. Xiong, Z. Li, Y. Chen, *et al.*, "Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications," *arXiv preprint arXiv:2401.06197*, 2024.

[21] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 713–13 722.

[22] D. Ouyang, S. He, G. Zhang, *et al.*, "Efficient multi-scale attention module with cross-spatial learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[23] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "Epsanet: An efficient pyramid squeeze attention block on convolutional neural network," in *Asian Conference on Computer Vision (ACCV)*, 2023, pp. 541–557.

[24] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," *International Conference on Computer Vision Workshop (ICCVW)*, pp. 1971–1980, 2019.

[25] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "Cgnet: A light-weight context guided network for semantic segmentation," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 1169–1179, 2020.

[26] J.-S. Lim, M. Astrid, H.-J. Yoon, and S.-I. Lee, "Small object detection using context and attention," in *International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, IEEE, 2021, pp. 181–186.

[27] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, *Yolov9: Learning what you want to learn using programmable gradient information*, 2024. arXiv: 2402.13616.

[28] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics YOLOv8*, https : / / github . com / ultralytics / ultralytics, version 8.0.0, Jan. 2023.

[29] G. Jocher, *YOLOv5 by Ultralytics*, https://github.com/ultralytics/yolov5, version 7.0, May 2020.

[30] C. Li, L. Li, H. Jiang, *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[31] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 781–10 790.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, pp. 1137–1149, 2016.

One-step Spectral Estimation for Euclidean Distance Matrix Approximation

Yicheng Li* and Xinghua Sun*
* School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 518107, China
E-mail: liych75@mail2.sysu.edu.cn; sunxinghua@mail.sysu.edu.cn

***Abstract*—This paper proposes and analyzes a new spectral estimator for Euclidean Distance Matrix Completion (EDMC) problem under Bernoulli sample model. This new estimator can be viewed as a simplified version of the well-known SVD (Singular Value Decomposition)-MDS (Multidimensional Scaling) approach by Drineas et al. as it reduces the computational cost while maintaining almost the same performance. Its approximation accuracy, when measured by the spectral norm, is analyzed under the presence of zero-mean, entry-wise independent, sub-Gaussian additive noise. Theoretical results show that the approximation error is well-controlled up to the noise amplification term as soon as the sample complexity reaches the order of $\mathcal{O}(\log^2 n/n)$. Numerical simulations verify the above assertions.**

I. Introduction

We are interested in the reconstruction of a point set configuration from its sampled pair-wise distance measurements, known as the Euclidean Distance Matrix Completion (EDMC) problem [1]. The EDMC serves as a fundamental problem in a variety of fields, including but not limited to data visualization and dimensionality reduction in machine learning [2, Sec. 3.1.1] [3], wireless channel estimation [4], phase retrieval with unknown sensing basis [5], computational biochemistry [6] [7] [8], operating wireless sensor networks [9] [10] and articulated robots [11]. The EDMC is known to be NP-hard without prior restrictions [12], yet practical solutions with either deterministic type guarantees [9] [13] or performance bound under random point set model [10] exist. Recently, [14] proved that if the underlying point set satisfies the incoherent assumption, i.e., a prevalent measure of the degeneracy of the problem in low-rank matrix recovery context, and the distance samples are taken uniformly at random, then the EDM can be exactly recovered with near-optimal sample complexity via the nuclear norm minimization paradigm [15].

A. Mathematical Setup

Consider a ground truth point set consists of n points, and embedded in the $\mathbb{R}^d$ space, $d = 2, 3 \ll n$. Denote its configuration by $\mathbf{Y}^\star = [\mathbf{p}_1^\star, \ldots, \mathbf{p}_n^\star]^T \in \mathbb{R}^{n\times d}$, and assume the Gram matrix $\mathbf{G}^\star := \mathbf{Y}^\star\mathbf{Y}^{\star T}$ has rank d. Let $d_{ij}^\star$ be the ground truth distance between the (i, j) point, then the symmetric matrix $\mathbf{D}^\star \in \mathbb{R}^{n\times n}$, which (i, j)-th entry is $(d_{ij}^\star)^2$ is the so-called EDM, i.e., $\mathbf{D}^\star \in \mathbb{EDM}^n$. Mathematically, the observed entries in the EDM satisfies

$$\mathbf{D}_{ij}^\star = (d_{ij}^\star)^2 = \mathrm{tr}(\mathbf{G}^\star\boldsymbol{\omega}_{\boldsymbol{\alpha}}), (i, j) \in \Omega, \tag{1}$$

where tr is the matrices inner product in $\mathbb{R}^{n\times n}$, i.e, $\mathrm{tr}(\mathbf{A}\mathbf{B}^T) = \langle \mathbf{A}, \mathbf{B}\rangle$. $\boldsymbol{\alpha} = (i, j)$ is a tuple, and $\boldsymbol{\omega}_{\boldsymbol{\alpha}} := (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T$, where $\mathbf{e}_i \in \mathbb{R}^n$ denotes the i-th canonical Euclidean basis. Ω is the sample set and we assume that it is generated by the widely used symmetric Bernoulli rule, that is $\Omega := \{(i, j)|\delta_{\boldsymbol{\alpha}} = 1, 1 \le i < j \le n\}$, where $\delta_{\boldsymbol{\alpha}}$ is a sequence of i.i.d. 0/1 Bernoulli random variables with $\mathbb{P}(\delta_{\boldsymbol{\alpha}} = 1) = p$. And we let the operator $\mathcal{P}_\Omega$ denote the projection onto the sample set

$$\mathcal{P}_\Omega(\cdot) := \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \delta_{\boldsymbol{\alpha}}(\langle \cdot, \mathbf{e}_i\mathbf{e}_j^T\rangle\mathbf{e}_i\mathbf{e}_j^T + \langle \cdot, \mathbf{e}_j\mathbf{e}_i^T\rangle\mathbf{e}_j\mathbf{e}_i^T), \tag{2}$$

where $\mathbb{I} := \{(i, j)|1 \le i < j \le n\}$[1]. This paper details a specific spectral method to approximately solve the EDMC under the setup of (1) (2) and provides a theoretical characterization of the reconstruction error of the point set. We will also demonstrate that such coarse guess can be used to initialize those non-convex successive refinements discussed in our previous work [16] and to achieve high-accuracy recovery of the point set numerically, which might shed some light on the guaranteed non-convex EDMC algorithms design.

Notations: **Bold** face lower letters and capital letters represent vectors or matrices, respectively. $\mathbf{1}$, $\mathbf{0}$, $\mathbf{I}$ is the vector of all ones, all zeros, and the identity matrix, respectively. $\mathbf{A}_{ij}$, $\mathrm{diag}(\mathbf{A})$ is the (i, j)-th entity, and the column vector formed by the diagonal elements of $\mathbf{A}$, respectively. The adjoint of a linear transformation $\mathcal{T}$ is denoted by $\mathcal{T}^*$ while $\mathcal{I}$ denotes the identity operator. $\|\mathbf{A}\|$, $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_\infty$ stand for the spectral, Frobenius, and entrywise l_∞ norm of $\mathbf{A}$. $\|\mathbf{a}\|_2$ is the vector l_2 norm. C, c are positive universal constants and may differ from line to line. $a \lesssim b$ and $a \gtrsim b$ stands for $a < Cb$ and $a > Cb$, respectively. $\mathbf{A} \succcurlyeq \mathbf{0}$ means that $\mathbf{A}$ is positive semidefinite.

Organization: In Section I-B we state our main findings while in Section I-C we briefly review the most related literature and provide a few remarks on the main result. Numerical tests that support our claims and a detailed proof are referred to Section II and Section III, respectively. This paper is concluded in Section IV.

B. Main Result

We propose the following one-step spectral estimator to approximate the EDM, i.e., to recover the point set configuration

[1]We note that since any EDM is a hollow diagonal matrix, the diagonal samples can be automatically dropped out without introducing perturbation.

up to translation and rotation ambiguity

$$\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T = \hat{\mathbf{G}}_\natural = \mathcal{T}_d\left[-\frac{1}{2p}\mathbf{J}(\mathcal{P}_\Omega\hat{\mathbf{D}})\mathbf{J}\right], \tag{3}$$

where $\mathcal{T}_d$ is the rank-d singular value reconstruction and $\mathbf{J} := \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$. $\hat{\mathbf{D}} = \mathbf{D}^\star + \mathbf{E}$ stands for the noisy observation and $\mathbf{E}$ is a symmetric noise matrix. We assume that $\{\mathbf{E}_{ij}\}_{i>j}$ are independent, zero-mean random variables with sub-Gaussian norm [17, Sec. 2.5] bounded by σ_n. Our main result controls the distortion of $\hat{\mathbf{G}}_\natural$ from $\mathbf{G}^\star$, which depends on the standard coherence parameter [15, Def. 1.2] of the ground truth.

Definition I.1 (Coherence)**.** *Let $\mathbf{G}^\star = \mathbf{U}^\star\mathbf{\Sigma}^\star\mathbf{U}^{\star T}$ denotes the rank-d thin Singular Value Decomposition (SVD) of $\mathbf{G}^\star$. Let the coherence parameter be defined as*

$$\mu^2 := \max_{i\in[n]} \frac{n}{d}\|\mathbf{e}_i^T\mathbf{U}^\star\|_2^2.$$

It is straightforward to check that $1 \le \mu^2 \le \frac{n}{d}$.

Theorem I.1 (Main)**.** *For a fixed, self-centered[2] point set configuration $\mathbf{Y}^\star$, under the setup of* (1) (2) (3) *and Definition I.1, for some $0 < \epsilon \le 1$, we have*

$$\|\hat{\mathbf{G}}_\natural - \mathbf{G}^\star\| \lesssim \epsilon\|\mathbf{G}^\star\| + \sigma_n\sqrt{\frac{n}{p}}, \tag{4}$$

holds with probability at least $1 - n^c$ as soon as $p \gtrsim \max\{\log^2 n, \mu^4 d^2/\epsilon^2\}/n$. The proof is referred to Section III.

C. Related Works

A detailed review of prior arts for the EDMC problem is far beyond the scope of this article, thus we choose only to discuss those that shared the most resemblance with us and refer interested readers to [1] [18] for the whole picture. The low-rank property of the EDM has been known for decades and might induce the initial interest in studying the low-rank recovery models [19]. By substituting any point set $\mathbf{Y} \in \mathbb{R}^{n\times d}$ into (1), and set $\Omega = \mathbb{I}$, i.e., a full observation, one finds

$$\mathbf{D} = g(\mathbf{G}) := \mathrm{diag}(\mathbf{G})\mathbf{1}^T + \mathbf{1}\mathrm{diag}(\mathbf{G})^T - 2\mathbf{G}, \tag{5}$$

where $\mathbf{G} = \mathbf{Y}\mathbf{Y}^T$. This yields the so-called EDM forward mapping g. From (5), it is straightforward to check that any EDM has rank no larger than $d + 2$ by the utilization of rank inequality. This fact gives rise to the SVD-MDS algorithm, proposed by Drineas et al. [20], which directly combines the idea from standard spectral method for Low-rank Matrix Completion [21, Sec. 3.8] (SVD, the first step) and the Multidimensional Scaling [2, Sec. 3.1] (MDS, the second step). It reads

$$\tilde{\mathbf{D}} = \mathcal{T}_{(d+2)}\left[\frac{1}{p}\mathcal{P}_\Omega\hat{\mathbf{D}}\right], \hat{\mathbf{G}} = \mathcal{T}_d\left[g^+(\tilde{\mathbf{D}})\right], \tag{6}$$

where g^+ is the pseudo-inverse of g (which will be introduced shortly). The performance guarantee of SVD-MDS was later refined in [22] by leveraging the moment-control Matrix Bernstein inequality [23, Thm. 6.2] and a perturbation bound on MDS. They also proved that the first step in (6) for EDMC is minimax suboptimal when the noise is small. While our approach removes the rank-$(d + 2)$ SVD thus reduces the computation cost approximately by half when n is small[3]. Moreover, when compared with [22, Thm. 2], Theorem I.1 directly establishes the reconstruction error of the point set with near-optimal sample complexity[4], and it is more general since we do not require the noise to be i.i.d. Gaussian. Beyond Theorem I.1, we have also observed that the performance gap between (3) and (6) is small in the numerical test, i.e., a reasonable compromise for the reduced computational complexity.

(3) is actually inspired by [14] [24], in which they introduced the dual-basis representation method to reformulate the forward and inverse EDM mapping. To proceed, we need the explicit expression of g^+

$$\mathbf{G} = g^+(\mathbf{D}) := -\frac{1}{2}\mathbf{J}\mathbf{D}\mathbf{J}, \mathbf{D} \in \mathbb{EDM}^n, \mathbf{G} \succcurlyeq \mathbf{0}. \tag{7}$$

This map is well-defined on $\{\mathbf{1}\}^\perp$ [1, Thm. 2], and we call $\mathbf{J}$ the geometry centering matrix as it forces $\mathbf{G}\mathbf{1} = \mathbf{0}$. [14] and the following up works showed that (5) (7) have alternative forms by noticing that

$$\begin{aligned}\mathcal{I}_{\{\mathbf{1}\}^\perp}(\cdot) = g^+g(\cdot) &= -\frac{1}{2}\mathbf{J}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot,\boldsymbol{\omega}_{\boldsymbol{\alpha}}\rangle(\mathbf{e}_i\mathbf{e}_j^T + \mathbf{e}_j\mathbf{e}_i^T)\right]\mathbf{J}\\ &= \sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot,\boldsymbol{\omega}_{\boldsymbol{\alpha}}\rangle\boldsymbol{\nu}_{\boldsymbol{\alpha}},\end{aligned} \tag{8}$$

where $\mathcal{I}_{\{\mathbf{1}\}^\perp}$ is the identity operator but restricted to $\{\mathbf{1}\}^\perp$, and $\boldsymbol{\nu}_{\boldsymbol{\alpha}} := -\frac{1}{2}\mathbf{J}(\mathbf{e}_i\mathbf{e}_j^T + \mathbf{e}_j\mathbf{e}_i^T)\mathbf{J}$ is the dual-basis of $\boldsymbol{\omega}_{\boldsymbol{\alpha}}$. Some linear algebra verifies this claim by checking that

$$\mathrm{tr}(\boldsymbol{\omega}_{\boldsymbol{\alpha}}\boldsymbol{\nu}_{\boldsymbol{\beta}}) = \begin{cases}1, & \boldsymbol{\alpha} = (i,j) = (k,l) = \boldsymbol{\beta}\\ 0, & \text{otherwise}\end{cases}.$$

In other words, the forward mapping can be established by the basis $\boldsymbol{\omega}_{\boldsymbol{\alpha}}$, and the inverse mapping is encoded implicitly in $\boldsymbol{\nu}_{\boldsymbol{\alpha}}$. One interesting operator originated from (8) is its sampled version, defined as follows

$$\mathcal{R}_\Omega := g^+\mathcal{P}_\Omega g = \sum_{\boldsymbol{\alpha}\in\mathbb{I}}\delta_{\boldsymbol{\alpha}}\langle\cdot,\boldsymbol{\omega}_{\boldsymbol{\alpha}}\rangle\boldsymbol{\nu}_{\boldsymbol{\alpha}}.$$

It is straightforward to check $\mathbb{E}\frac{1}{p}\mathcal{R}_\Omega(\mathbf{G}) = \mathbf{G}$ holds for self-centered $\mathbf{G}$ due to (8). Thus, one may then conjecture that $\frac{1}{p}\mathcal{R}_\Omega(\mathbf{G})$ concentrate around its mean nicely, i.e., the rank-d SVD of $\frac{1}{p}\mathcal{R}_\Omega(\mathbf{G})$ is close to the unperturbed matrix $\mathbf{G}$. After simple linear algebra, this idea gives rise to (3).

[24] proposed another spectral estimator for the same purpose. They use SVD to truncate the matrix after applying

[2]Without loss of generality, we assume that point sets in this work are always self-centered, i.e., $\mathbf{Y}^T\mathbf{1} = \mathbf{0}$. Otherwise, one can perform the projection $\mathbf{Y}' = \mathbf{J}\mathbf{Y}$ to force this property.

[3]We note that the main computation cost in either (3) or (6) comes from the thin SVD when n is small.

[4]The $\log^2 n$ term in Theorem I.1 comes from controlling the spectral norm of $\frac{1}{p}\mathcal{P}_\Omega\mathbf{E}$. Under noiseless situation, the bound can be improved to $p \gtrsim \max\{\log n, \mu^4 d^2/\epsilon^2\}/n$, which is only sub-optimal in d.

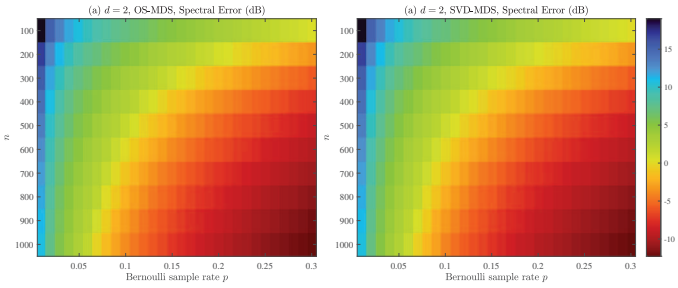


Fig. 1. The spectral error (in dB) of OS-MDS and SVD-MDS when varying Bernoulli sample rate p and the number of points n. For each pair (n, p), the result is the median value of 1000 independent trials due to the limited iteration cost when evaluating the spectral norm, i.e., to exclude outliers.

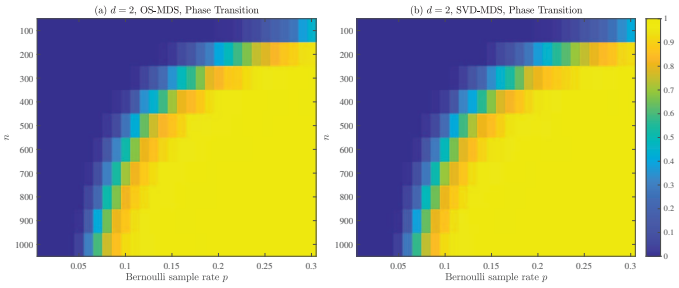


Fig. 2. The phase transition of OS-MDS and SVD-MDS when varying Bernoulli sample rate p and the number of points n. To verify Theorem I.1, we claim a success if the spectral error falls below 1. For each pair (n, p), the result is obtained by 1000 independent trials.

a self-dual operator $\mathcal{R}_\Omega^*\mathcal{R}_\Omega = g^*\mathcal{P}_\Omega g^+ g^+ \mathcal{P}_\Omega g$. The explicit expression of this scheme is given as follows

$$\hat{\mathbf{G}} = \mathcal{T}_d\left[\frac{1}{p^2} g^*\mathcal{P}_\Omega g^+ g^+ \mathcal{P}_\Omega(\hat{\mathbf{D}})\right]. \tag{9}$$

Even though (9) has almost the same computation cost as (3), the composition of $\mathcal{R}_\Omega^*\mathcal{R}_\Omega$ makes theoretical analysis challenging since g^+g^+ is not a diagonal operator. The non-orthonormality of $\boldsymbol{\nu}_{\boldsymbol{\alpha}}$ brings crossing term into $\mathcal{R}_\Omega^*\mathcal{R}_\Omega(\mathbf{G})$, and, the reuse of the sample set raises complicated dependency structure. This statement can be verified by expanding the terms, we have

$$\begin{aligned}\mathcal{R}_\Omega^*\mathcal{R}_\Omega(\mathbf{G}) = &\sum_{\boldsymbol{\alpha}=\boldsymbol{\beta}} \delta_{\boldsymbol{\alpha}}\langle \mathbf{G}, \boldsymbol{\omega}_{\boldsymbol{\alpha}}\rangle \|\boldsymbol{\nu}_{\boldsymbol{\alpha}}\|_F^2 \boldsymbol{\omega}_{\boldsymbol{\alpha}} \\ &+ \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \delta_{\boldsymbol{\alpha}}\delta_{\boldsymbol{\beta}}\langle \mathbf{G}, \boldsymbol{\omega}_{\boldsymbol{\alpha}}\rangle \langle \boldsymbol{\nu}_{\boldsymbol{\alpha}}, \boldsymbol{\nu}_{\boldsymbol{\beta}}\rangle \boldsymbol{\omega}_{\boldsymbol{\beta}},\end{aligned} \tag{10}$$

where we use the fact that $\delta_{\boldsymbol{\alpha}}^2$ has the same distribution as $\delta_{\boldsymbol{\alpha}}$. The first term in (10) introduces bias when measured on the population level, while the second term indicates that $\mathcal{R}_\Omega^*\mathcal{R}_\Omega$ has limited randomness structure and we leave this analysis to our future work.

II. Numerical Simulations

We present numerical results to verify the claimed performance of the proposed new estimator (3), which is referred to as OS-MDS (One Step-Multidimensional Scaling) in the following sections. The comparison between OS-MDS and SVD-MDS [20] is conducted in both noiseless and noisy tests over synthetic point sets, while OS-MDS initialized gradient refinement is also tested over four protein data downloaded from the Protein Data Bank [25].

A. Implementation Details

The synthetic ground truth point set $\mathbf{Y}^\star$ used in this study is generated by standard Gaussian distribution, while the sample mask Ω is forced to be symmetric, i.e., to satisfy (2). The noise $\mathbf{E}$ is a symmetric, hollow diagonal, Gaussian matrix, i.e., $\mathbf{E}_{ij} \sim \mathcal{N}(0, \sigma_n^2)$ with σ_n^2 varied for different noise intensities. The SVD truncation step is implemented by MATLAB `svds` function. If not specified otherwise, under each different scenario setup, we run these two spectral estimators for 1000 independent trials. The non-convex refinement discussed in [16], which takes the approximation result obtained by OS-MDS and SVD-MDS as initialization and perform Gradient Descent (GD) to optimize the s-stress function ((11) below) is also tests here. The GD solver comes from Manopt [26] default `steepestdescent` function, whose step size is chosen according to the widely used Armijo rule. Except the default stopping criteria, the gradient refinement is terminated if either it reaches the maximum iteration (800 times), or the Frobenius norm of the gradient is less than 10^{-6}.

$$f(\mathbf{Y}) = \frac{1}{2}\|\mathcal{P}_\Omega(g(\mathbf{Y}\mathbf{Y}^T) - \hat{\mathbf{D}})\|_F^2. \tag{11}$$

Three metrics are considered to evaluate the performance, i.e., the spectral error (12a), EDM recovery rate (12b), and the Frobenius error (12c), defined below.

$$\mathrm{SE} := \frac{\|\hat{\mathbf{G}} - \mathbf{G}^\star\|}{\|\mathbf{G}^\star\|}, \tag{12a}$$

$$\mathrm{RE} := \frac{\|\bar{\mathbf{D}} - \mathbf{D}^\star\|_F}{\|\mathbf{D}^\star\|_F}, \tag{12b}$$

$$\mathrm{FE} := \frac{\|\hat{\mathbf{G}} - \mathbf{G}^\star\|_F}{\|\mathbf{G}^\star\|_F}, \tag{12c}$$

where $\hat{\mathbf{G}}$ is the approximation generated by either OS-MDS or SVD-MDS, and $\bar{\mathbf{D}}$ denotes the recovered EDM after GD refinement. The SNR in noisy tests is defined by

$$\mathrm{SNR} := \frac{\|\mathbf{G}^\star\|_F^2}{n^2\sigma_n^2}.$$

We say "in dB" means taking $20\log_{10}$ to the residual in (12), or $10\log_{10}$ to the SNR. All experiments are run on a dual-socket Intel Xeon Gold 6226r server with 256 GB of RAM, NVIDIA A5000, Ubuntu 20.04.4, MATLAB 2022a.

B. Noiseless Test

Fig. 1 plots the spectral error of OS-MDS and SVD-MDS when varying Bernoulli sample rate p and the number of points n. We use the median value instead of the mean of 1000 independent trials mainly due to the observation of artificial outliers when evaluating (12a). This is caused by the limited iteration budget when using power method to calculate the spectral norm. Fig. 2 shows the phase transition of the test

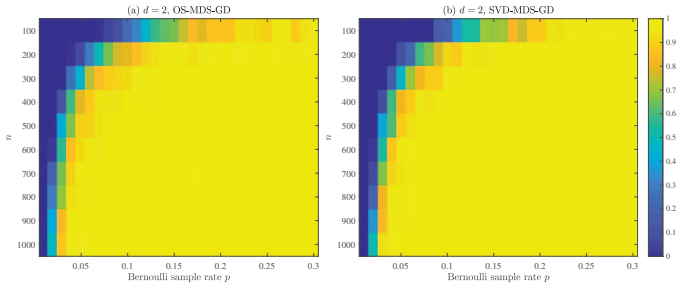


Fig. 3. The phase transition of OS-MDS and SVD-MDS initialized GD with Armijo step size, when varying Bernoulli sample rate p and the number of points n. We claim a success if the EDM recovery rate falls below 10^{-3}. For each pair (n, p), the result is obtained by 100 independent trials.

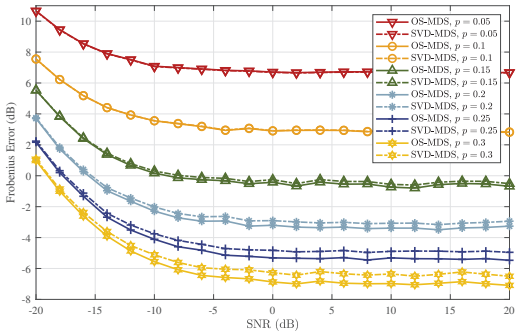


Fig. 4. The Frobenius error (in dB) of OS-MDS and SVD-MDS v.s. SNR (in dB) when varying Bernoulli sample rate p. We fix $n = 500$, $d = 2$. The results are the mean value of 1000 independent trials.

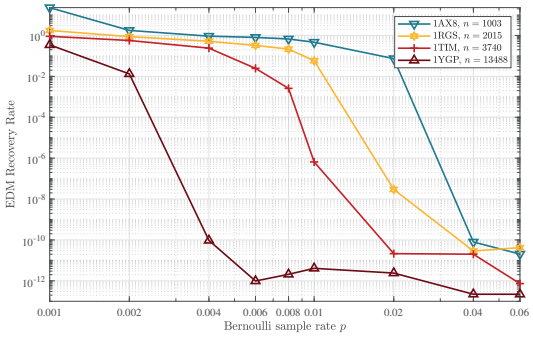


Fig. 5. The EDM recovery rate of OS-MDS initialized GD with Armijo step size when varying Bernoulli sample rate p, tested on four proteins. The proteins have different atom numbers (n) and fixed embedding dimension $d = 3$. 1AX8 and 1RGS are single-chain while the other are double-chain. The results are the mean value of five independent trials.

estimators. To verify Theorem I.1 with $\epsilon < 1$, we claim a success if (12a) falls below 1. The transition edge clearly shows the $p = \mathcal{O}(\log n/n)$ dependency, which matches the theoretical result[5]. By comparing Fig. 1(a)/Fig. 2(a) with Fig. 1(b)/Fig. 2(b), one finds that the performance gap between OS-MDS and SVD-MDS is ignorable.

The phase transition of GD refinement, when initialized by either OS-MDS or SVD-MDS is plotted in Fig. 3. Again, the two strategies show minor differences. It is interesting to see that GD can converge in a very less restrictive region around the ground truth. Yet a non-trivial explanation of this phenomenon, when specified to the EDMC problem, is still a mystery, please see [16] for more discussion.

C. Noisy Test

Fig. 4 plots the noise-resistance behavior, evaluated by the Frobenius error, of OS-MDS and SVD-MDS when varying Bernoulli sample rate p and the SNR. The range of p is chosen to be the period where phase transition of these two estimators happens. The result clearly shows the existence of an SNR threshold around -5 dB, i.e., below this, the noise dominates the approximation error. When the noise is ignorable, the first term in (4) starts to matter. It is surprising to see OS-MDS performs slightly better than SVD-MDS as the problem becomes well-posed, since it only uses one-step SVD reconstruction.

D. Protein Test

We perform a similar test over real protein data as in [14, Sec. V-A]. Specifically, we choose four of the tested proteins in [27, Sec. 5.3] and randomly delete inter-atom distance measurements according to the symmetric Bernoulli rule. Then the OS-MDS initialized GD is applied to solve this EDMC problem. The algorithm is run on GPU using MATLAB `gpuArray` to speed up computation. Fig. 5 presents the EDM recovery rate of this process when varying Bernoulli sample rate p and the protein type. It suggests that all four protein structures can be exactly recovered from less than 5% random distance samples by a simple, scalable, memory-friendly strategy. Yet this numerical witness does not mean that the protein configuration problem [8] is ready to be handled via delicate initialized GD, since the practical distance sampling scheme, caused by the behavior of Nuclear Magnetic Resonance (NMR) techniques[6], do not follow the Bernoulli rule. Please see [16, Sec. III-C] for more discussion regarding the impact of this practical scheme for non-convex EDMC algorithm design.

III. Proof

Our proof of Theorem I.1 relies on a celebrated sharp bound by Bandeira and van Handel [28], restated here as the lemma below. It is tighter than the matrix Bernstein inequality [23, Thm. 6.1] when applied to matrix with independent random entries.

Lemma III.1 ([28], Coro. 3.12, Remark 3.13)**.** *Let* $\mathbf{X}$ *be a symmetric matrix in* $\mathbb{R}^{n\times n}$ *whose entries* $\mathbf{X}_{ij}$ *are independent, zero-mean random variables. Let*

$$\sigma_x^2 := \max_i \sum_{j=1}^{n} \mathbb{E}[\mathbf{X}_{ij}^2],\ \max_{ij} |\mathbf{X}_{ij}| \le B.$$

[5]Since $d = 2$ and $n \ge 100$ in all tests and we use Gaussian point set ($\nu = \mathcal{O}(1)$), the $\log n$ term in Theorem I.1 dominates.

[6]NMR can estimate the distance between two atoms if their pair-wise distance is less than $6\mathring{A}$ [27]. A similar sample rule also appears in the sensor network localization problem [10].

Then there exist some universal constant $c > 0$, such that for every $t > 0$

$$\mathbb{P}\{\|\mathbf{X}\| > 3\sqrt{2}\sigma_x + t\} \leq n\exp(-\frac{t^2}{cB^2}).$$

One immediate corollary is that [21, Sec. 3.1], if $\mathbf{X}$ satisfies $\max_{ij}\mathbb{E}[\mathbf{X}_{ij}^2] \leq \sigma^2$, we have

$$\|\mathbf{X}\| \leq 3\sqrt{2}\sqrt{n\sigma^2} + c'B\sqrt{\beta\log n} \tag{13}$$

holds for some universal constant c' and $\beta > 1$ with probability at least $1 - n^{\beta-1}$.

Proof: We are now ready to proceed with the proof. Notice that

$$\begin{aligned}
&\|-\frac{1}{2p}\mathbf{J}(\mathcal{P}_\Omega\hat{\mathbf{D}})\mathbf{J} - \mathbf{G}^\star\| \overset{(i)}{=} \|\frac{1}{p}\mathcal{R}_\Omega(\mathbf{G}^\star) - \mathbf{G}^\star + g^+(\frac{1}{p}\mathcal{P}_\Omega\mathbf{E})\| \\
&\leq \|\frac{1}{p}\mathcal{R}_\Omega(\mathbf{G}^\star) - \mathbf{G}^\star\| + \|\frac{1}{2p}\mathbf{J}(\mathcal{P}_\Omega\mathbf{E})\mathbf{J}\| \\
&\overset{(ii)}{\leq} \|g^+(\frac{1}{p}\mathcal{P}_\Omega\mathbf{D}^\star - \mathbf{D}^\star)\| + \frac{1}{2}\|\frac{1}{p}\mathcal{P}_\Omega\mathbf{E}\| \\
&\overset{(iii)}{\leq} \frac{1}{2}(\underbrace{\|\frac{1}{p}\mathcal{P}_\Omega\mathbf{D}^\star - \mathbf{D}^\star\|}_{I_1} + \underbrace{\|\frac{1}{p}\mathcal{P}_\Omega\mathbf{E}\|}_{I_2}).
\end{aligned} \tag{14}$$

Where (i) from our model setup in Section I-B, (ii), (iii) by noticing that $\mathbf{J}$ is an orthogonal projection thus $\|\mathbf{J}\| \leq 1$. What remains is to handle I_1 and I_2.

Bound on I_1: It is straightforward to check that the matrix $\mathbf{X} := \frac{1}{p}\mathcal{P}_\Omega\mathbf{D}^\star - \mathbf{D}^\star$ satisfies all prerequisites of Lemma III.1. Specifically, $B = \|\mathbf{X}\|_\infty \leq \frac{1}{p}\|\mathbf{D}^\star\|_\infty$ and $\sigma^2 = \max_{ij}\mathbb{E}[\mathbf{X}_{ij}^2] \leq \mathbb{E}(\frac{1}{p}\delta_{\boldsymbol{\alpha}} - 1)^2\|\mathbf{D}^\star\|_\infty^2 \leq \frac{1}{p}\|\mathbf{D}^\star\|_\infty^2$. Using (13), we have

$$\begin{aligned}
I_1 &\leq c_1\sqrt{\frac{n}{p}\|\mathbf{D}^\star\|_\infty^2} + c_2\sqrt{\beta\log n}\frac{1}{p}\|\mathbf{D}^\star\|_\infty \\
&\overset{(i)}{\leq} 2c_1\sqrt{\frac{n}{p}\|\mathbf{D}^\star\|_\infty^2} \overset{(ii)}{\leq} 2c_1\sqrt{\frac{4\mu^4d^2}{np}}\|\mathbf{G}^\star\| \overset{(iii)}{\leq} \varepsilon\|\mathbf{G}^\star\|,
\end{aligned} \tag{15}$$

where c_1, c_2 are some universal constants. (i) from setting $p > (\frac{c_2}{c_1})^2\beta\log n/n$, and (iii) by forcing $p > 16c_1^2\mu^4d^2/(n\varepsilon^2)$ for some $0 < \varepsilon \leq 1/2$. (ii) from noticing that

$$\|\mathbf{D}^\star\|_\infty^2 \leq \max_{\boldsymbol{\alpha}}|\langle\boldsymbol{\omega}_{\boldsymbol{\alpha}}, \mathbf{G}^\star\rangle|^2 \overset{(a)}{\leq} 4\|\mathbf{G}^\star\|_\infty^2 \overset{(b)}{\leq} \frac{4\mu^4d^2}{n^2}\|\mathbf{G}^\star\|^2,$$

where (a) from expanding $\boldsymbol{\omega}_{\boldsymbol{\alpha}}$, i.e.,

$$\langle\boldsymbol{\omega}_{\boldsymbol{\alpha}}, \mathbf{G}^\star\rangle = \mathbf{G}^\star_{ii} + \mathbf{G}^\star_{jj} - \mathbf{G}^\star_{ij} - \mathbf{G}^\star_{ji},$$

then followed by Cauchy-Schwarz, and (b) by Definition I.1. Thus we conclude that (15) holds with probability at least $1 - n^{\beta-1}$ if $p \gtrsim \max\{\beta\log n, \mu^4d^2/\varepsilon^2\}/n$, $\beta > 1$.

Bound on I_2: This is a meanwhile standard result proved by [29, Lemma 11] using truncation argument. We briefly restate it here. By standard sub-Gaussian tail one finds $\max_{ij}|\frac{1}{p}\delta_{\boldsymbol{\alpha}}\mathbf{E}_{ij}| \leq C\sqrt{\sigma_n^2\log n/p}$ hold with probability $1 - n^{-c}$. Also $\mathbb{E}|\frac{1}{p}\delta_{\boldsymbol{\alpha}}\mathbf{E}_{ij}|^2 \leq C\sigma_n^2/p$. Using (13) and setting $p \gtrsim \frac{\beta\log^2 n}{n}$, $\beta > 1$, we have $I_2 \lesssim \sigma_n\sqrt{\frac{n}{p}}$ holds with probability at least $1 - n^{-c} - n^{\beta-1}$.

By substituting the above two estimates into (14) and using Weyl's theorem, we find

$$\begin{aligned}
\|\hat{\mathbf{G}}_\natural - \mathbf{G}^\star\| &\leq \|\mathcal{T}_d\left[-\frac{1}{2p}\mathbf{J}(\mathcal{P}_\Omega\hat{\mathbf{D}})\mathbf{J}\right] + \frac{1}{2p}\mathbf{J}(\mathcal{P}_\Omega\hat{\mathbf{D}})\mathbf{J}\| \\
&\quad + \|-\frac{1}{2p}\mathbf{J}(\mathcal{P}_\Omega\hat{\mathbf{D}})\mathbf{J} - \mathbf{G}^\star\| \\
&\leq 2\|-\frac{1}{2p}\mathbf{J}(\mathcal{P}_\Omega\hat{\mathbf{D}})\mathbf{J} - \mathbf{G}^\star\| \lesssim \varepsilon\|\mathbf{G}^\star\| + \sigma_n\sqrt{\frac{n}{p}},
\end{aligned}$$

where we conclude the proof after some trivial rescale of the constants. ∎

IV. Conclusion

This paper proposes a new spectral estimator, namely, the OS-MDS, for EDM approximation under Bernoulli sample model. Theoretical characterization of its behavior is conducted under the presence of additive, sub-Gaussian, and symmetric noise. Numerical results verify the claimed performance. It is of future interest to explore the theoretical guarantee of the algorithm proposed in [24], and to uncover the convergence behavior of the gradient method when applied to (11), starting from the initial point provided by OS-MDS.

Acknowledgment

This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B0101120003, and in part by Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515012015).

References

[1] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean distance matrices: Essential theory, algorithms, and applications," *IEEE Signal Process Mag.*, vol. 32, no. 6, pp. 12–30, 2015. DOI: 10.1109/MSP.2015.2398954.

[2] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative review," *J. Mach. Learn. Res.*, vol. 10, no. 66-71, p. 13, 2009.

[3] E. Arias-Castro, A. Javanmard, and B. Pelletier, "Perturbation bounds for procrustes, classical scaling, and trilateration, with applications to manifold learning," *J. Mach. Learn. Res.*, vol. 21, no. 1, Article 15, 2020.

[4] P. Agostini, Z. Utkovski, and S. Stańczak, "Channel Charting: An Euclidean distance matrix completion perspective," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 5010–5014. DOI: 10.1109/ICASSP40776.2020.9053639.

[5] S. Gupta, R. Gribonval, L. Daudet, and I. Dokmanić, "Don't take it lightly: Phasing optical random projections with unknown operators," *Adv. Neural Inf. Proces. Syst.*, vol. 32, 2019.

[6] G. M. Crippen, "Conformational analysis by energy embedding," *J. Comput. Chem.*, vol. 3, no. 4, pp. 471–476, 1982. DOI: 10.1002/jcc.540030404.

[7] H. Wang, J. Yang, Y. Zhang, J. Qian, and J. Wang, "Reconstruct high-resolution 3D genome structures for diverse cell-types using FLAMINGO," *Nat. Commun.*, vol. 13, no. 1, p. 2645, 2022. DOI: 10.1038/s41467-022-30270-2.

[8] B. Hendrickson, "The molecule problem: Exploiting structure in global optimization," *SIAM J. Optim.*, vol. 5, no. 4, pp. 835–857, 1995. DOI: 10.1137/0805040.

[9] A. M.-C. So and Y. Ye, "Theory of semidefinite programming for sensor network localization," *Math. Program.*, vol. 109, no. 2-3, pp. 367–384, 2006, ISSN: 0025-5610 1436-4646. DOI: 10.1007/s10107-006-0040-1.

[10] A. Javanmard and A. Montanari, "Localization from incomplete noisy distance measurements," *Found. Comput. Math.*, vol. 13, no. 3, pp. 297–345, 2013. DOI: 10.1007/s10208-012-9129-5.

[11] F. Marić, M. Giamou, A. W. Hall, S. Khoubyarian, I. Petrović, and J. Kelly, "Riemannian optimization for distance-geometric inverse kinematics," *IEEE Trans. Robot.*, vol. 38, no. 3, pp. 1703–1722, 2022. DOI: 10.1109/TRO.2021.3123841.

[12] J. B. Saxe, "Embeddability of weighted graphs in k-space is strongly np-hard," in *17th Allerton Conf. Commun. Control Comput., 1979*, 1979, pp. 480–489.

[13] A. M.-C. So and Y. Ye, "A semidefinite programming approach to tensegrity theory and realizability of graphs," in *Proc. Annu. ACM SIAM Symp. Discrete Algorithms*, 2006, pp. 766–775.

[14] A. Tasissa and R. Lai, "Exact reconstruction of euclidean distance geometry problem using low-rank matrix completion," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3124–3144, 2019. DOI: 10.1109/TIT.2018.2881749.

[15] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009. DOI: 10.1007/s10208-009-9045-5.

[16] Y. Li and X. Sun, "Sensor network localization via riemannian conjugate gradient and rank reduction," *IEEE Trans. Signal Process.*, vol. 72, pp. 1910–1927, 2024. DOI: 10.1109/TSP.2024.3378378.

[17] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. DOI: 10.1017/9781108231596.

[18] L. Liberti, C. Lavor, N. Maculan, and A. Mucherino, "Euclidean distance geometry and applications," *SIAM Rev.*, vol. 56, no. 1, pp. 3–69, 2014.

[19] M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices," in *Proc. Am. Control Conf.*, vol. 3, 2003, pp. 2156–2162. DOI: 10.1109/ACC.2003.1243393.

[20] P. Drineas, A. Javed, M. Magdon-Ismail, G. Pandurangan, R. Virrankoski, and A. Savvides, "Distance matrix reconstruction from incomplete distance information for sensor network localization," in *Ann. IEEE Commun. Soc. Ad hoc Commun. Netw. Secon*, vol. 2, 2006, pp. 536–544. DOI: 10.1109/SAHCN.2006.288510.

[21] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Spectral methods for data science: A statistical perspective," *Found. Trends Mach. Learn.*, vol. 14, no. 5, pp. 566–806, 2021. DOI: 10.1561/2200000079.

[22] H. Zhang, Y. Liu, and H. Lei, "Localization from incomplete euclidean distance matrix: Performance analysis for the svd–mds approach," *IEEE Trans. Signal Process.*, vol. 67, no. 8, pp. 2196–2209, 2019. DOI: 10.1109/TSP.2019.2904022.

[23] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, 2012. DOI: 10.1007/s10208-011-9099-z.

[24] C. M. Smith, S. P. Lichtenberg, H. Cai, and A. Tasissa, "Riemannian optimization for euclidean distance geometry," in *OPT 2023: Optimization for Machine Learning*, 2023.

[25] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–42, 2000. DOI: 10.1093/nar/28.1.235.

[26] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a Matlab toolbox for optimization on manifolds," *J. Mach. Learn. Res.*, vol. 15, no. 42, pp. 1455–1459, 2014. [Online]. Available: https://www.manopt.org.

[27] N.-H. Z. Leung and K.-C. Toh, "An sdp-based divide-and-conquer algorithm for large-scale noisy anchor-free graph realization," *SIAM J. Sci. Comput.*, vol. 31, no. 6, pp. 4351–4372, 2010. DOI: 10.1137/080733103.

[28] A. S. Bandeira and R. van Handel, "Sharp nonasymptotic bounds on the norm of random matrices with independent entries," *Ann. Probab.*, vol. 44, no. 4, pp. 2479–2506, 2016.

[29] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," *ArXiv*, vol. abs/1509.03025, 2015.

Handling Missing Data in Limited-View Photoacoustic Tomography Using Compressive Sensing Algorithm-Based Deep Learning

Mary John and Imad Barhumi
College of Engineering, EECE Department
United Arab Emirates University, Al Ain, Abu Dhabi, 15551, UAE
E-mail: {201990198, imad.barhumi}@uaeu.ac.ae

***Abstract*—In photoacoustic tomography, missing sensor data poses a significant challenge for accurate image reconstruction. This study aims to evaluate various imputation methods to address different missing data scenarios, including random missing data, communication loss, and environmental interference. We applied mean imputation, median imputation, k-nearest neighbors (KNN), and multivariate imputation by chained equations (MICE) to these scenarios and assessed their performance using metrics such as MS-SSIM, SSIM, PSNR, R-squared, NMSE, and Entropy. Our results demonstrate that the presence of missing data significantly degrades image quality, with MICE consistently providing the best reconstruction performance across all scenarios. In random missing data cases, MICE achieved the highest SSIM and PSNR values, closely approximating the no missing data scenario. Similar trends were observed in communication loss and environmental interference cases, where MICE outperformed other imputation methods, followed by mean and median imputation, with KNN generally showing the lowest performance. These findings suggest that MICE is a robust method for handling missing data in PAT, enhancing image reconstruction quality.**

Keywords: Photoacoustic Tomography, Compressive Sensing, Deep Learning, Missing Data, Image Reconstruction, Multivariate Imputation by Chained Equations.

I. Introduction

Photoacoustic tomography (PAT) [1] is a hybrid imaging technique that combines the high spatial resolution of optical imaging with the high contrast of ultrasound imaging. It works by illuminating tissue with pulsed laser light, which induces ultrasonic waves due to thermoelastic expansion. These ultrasonic waves are then detected by sensors, and the data is used to reconstruct images of the tissue's optical absorption properties. PAT has shown promise in various biomedical applications, including cancer detection, brain imaging, and functional imaging of blood oxygenation [2].

However, the quality of PAT images heavily depends on the completeness and accuracy of the data acquired by the sensors. In practical scenarios, missing data is a common issue due to various factors such as sensor failures, communication losses, environmental interferences, and scheduled maintenance. These missing data points can degrade the quality of the reconstructed images, making it challenging to achieve accurate diagnostics. To address this, researchers have explored various imputation methods to fill in the missing data before performing image reconstruction.

A. Motivation and Objectives

The primary motivation for this research is to enhance the robustness of PAT image reconstruction in the presence of missing data. Given the critical role of accurate imaging in medical diagnostics, it is essential to develop methods that can effectively handle missing data scenarios and still produce reliable images. This study aims to investigate the effectiveness of various imputation methods—mean, median, KNN, and MICE—in the context of PAT imaging. By simulating practical missing data scenarios, such as random sensor failures, communication losses, and environmental interferences, we seek to evaluate and compare the performance of these imputation methods. Furthermore, we aim to leverage compressive sensing algorithm-based deep learning network, specifically relaxed basis pursuit based on ADMM (Alternating Direction Method of Multipliers) [3], to reconstruct images from the imputed data. This network was initially trained on full-view sensor data and subsequently adapted through transfer learning to handle limited-view semicircular arrangements.

The objectives of this research are:

- To simulate various practical missing data scenarios in PAT.
- To apply and compare different imputation methods for handling missing data.
- To evaluate the performance of these imputation methods using metrics such as MS-SSIM, SSIM, PSNR, R-squared, NMSE, and entropy.
- To demonstrate the effectiveness of deep learning-based image reconstruction method in producing high-quality PAT images from imputed data.
- To provide insights into the strengths and limitations of each imputation method in the context of PAT.

By achieving these objectives, this research aims to contribute to the development of more reliable and robust PAT imaging techniques that can be used in clinical practice, ultimately improving patient outcomes through better diagnostics.

II. Related Works

Handling missing sensor data is a critical challenge in various imaging applications, including PAT. Numerous methods have been developed to address this issue, each with its unique approach and strengths. In [4] a comprehensive review of various imputation methods for missing sensor data is provided, categorizing them based on missingness patterns and properties. This foundational review highlights the diversity of imputation strategies and sets the stage for exploring more advanced techniques. Iterative Imputing Network is introduced in [5], which captures latent temporal dynamics to recover missing sensor data. This method leverages complex temporal patterns, showing significant improvements in applications like air quality and meteorological datasets. In [6] the authors proposed a framework to enhance MICE specifically for large gaps of missing data. Their approach involves reshaping the sensor data to exploit correlations between missing and observed values, splitting the data into extreme and normal values, and building separate MICE models for each. Several machine learning methods for imputing missing sensor data in cardiac imaging datasets are evaluated in [7], including mean imputation, KNN, C4.5 decision trees, and self-organizing maps. This study provides valuable insights into the effectiveness of various imputation techniques in medical imaging contexts. In [8] a novel deep learning-based framework for imputing missing values in sensor data is developed, specifically addressing the challenges of large segment imputation. Their multistage imputation strategy incorporates adaptive evaluation metrics and dynamic data length adjustment. A probabilistic method for imputing missing sensor data in multi-sensory activity recognition, using a conditional Gaussian distribution and exploiting correlation among classifier outputs is proposed in [9]. In [10] an efficient imputation technique based on k-means clustering for IoT sensor networks is presented. Their Best Fit Missing Value Imputation (BFMVI) method partitions incomplete data into clusters, estimates missing values using a random forest model trained on observed data within each cluster, and selects the most suitable imputation technique based on an error score function. These studies highlight the diverse approaches to imputation, from simple statistical methods to sophisticated machine learning techniques, providing a solid foundation for exploring the application of these imputation techniques to PAT imaging to enhance robustness and accuracy of image reconstruction in the presence of missing data.

III. Methodology

This section details the methodology used in our study, encompassing the compressive sensing algorithm, the deep learning network, the simulation of missing data scenarios, and the evaluation metrics employed to assess the performance of the imputation methods and image reconstruction.

A. Compressive Sensing Algorithm and Deep Learning Network

In this study, we employ a compressive sensing (CS) [11] algorithm-based deep learning network to reconstruct PAT images. Specifically, we use a relaxed basis pursuit algorithm based on the ADMM. Compressive sensing leverages the sparsity of signals to reconstruct high-quality images from under-sampled data, making it particularly effective for PAT where complete data acquisition is often challenging.

The significance of the rBP-ADMM (relaxed Basis Pursuit - ADMM) algorithm [12], [13] lies in its ability to address the complex inverse problem in PAT with high efficiency and robustness. The algorithm offers several advantages over traditional methods:

1) Robustness to Noise: rBP-ADMM algorithm integrates regularization terms that mitigate the impact of noise, resulting in more accurate and stable image reconstructions.
2) Iterative Refinement: It iteratively updates the solution, progressively refining the image, which is beneficial for dealing with the ill-posed nature of the PAT inverse problem.
3) Efficient Convergence: By splitting the optimization problem into smaller sub-problems, ADMM framework ensures faster convergence compared to traditional optimization methods, making it suitable for real-time applications.
4) Handling Incomplete Data: It effectively deals with missing or incomplete data, a common issue in practical PAT scenarios due to sensor limitations or failures.

The deep learning network architecture incorporates multiple layers, including convolutional layers, activation functions, and fully connected layers, to facilitate feature extraction and accurate image reconstruction. The network is initially trained on full-view sensor data to learn comprehensive patterns and features. Transfer learning is then applied to adapt the network for limited-view semicircular arrangements, enabling the model to leverage prior knowledge and fine-tuned for scenarios with partial data. The rBP-ADMM algorithm is depicted as in Algorithm 1, where $\mathbf{x}$ is the image to be reconstructed in vectorized form, $\mathbf{y}$ is the pressure sensor data, and $\mathbf{M}$ is the sensing matrix.

Algorithm 1 rBP-ADMM

Require: : $\rho, \alpha, \gamma, r, \tau, \mathbf{y}, \mathbf{M}, k_{\max}$, tol.
  Initialize: $k = 0$, $\mathbf{x}^0 = \mathbf{z}^0 = \mathbf{u}^0 = \mathbf{0}$, NMSE = ∞,
  while $k < k_{\max}$ or NMSE $>$ tol **do**
    $\mathbf{x}^{k+1} \leftarrow \left(\rho\mathbf{I} + \alpha\mathbf{M}^T\mathbf{M}\right)^{-1}\left(\rho\mathbf{z}^k - r\mathbf{u}^k + \alpha\mathbf{M}^T\mathbf{y}\right)$
    $\mathbf{z}^{k+1} \leftarrow S_{\frac{\gamma}{\rho}}\left(\frac{\mathbf{u}^k}{\rho} + \mathbf{x}^{k+1}\right)$
    $\mathbf{u}^{k+1} \leftarrow \mathbf{u}^k + \tau\left(\mathbf{x}^{k+1} - \mathbf{z}^{k+1}\right)$
    NMSE $\leftarrow \frac{\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2}{\|\mathbf{x}^{k+1}\|_2^2}$
    $k \leftarrow k + 1$
  end while
    $\mathbf{x}^* = \mathbf{x}^k$

The unrolled deep learning algorithm, termed U-rBP-ADMM, is designed based on the architecture of the rBP-ADMM algorithm, with iterations transformed into iteration

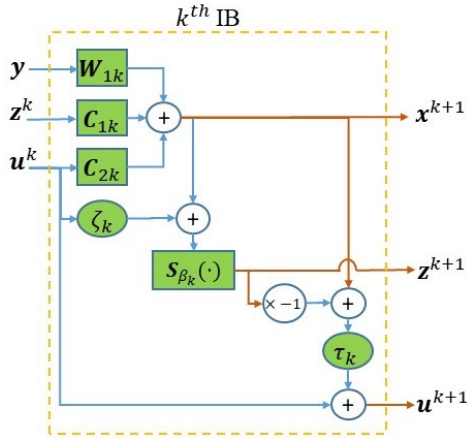


Fig. 1: k^{th} IB in U-rBP-ADMM network diagram.

blocks that perform similar operations to the iterative version using deep network layers. The updates at the $(k+1)$th intermediate block (IB) of the U-rBP-ADMM deep learning network [14] are designed as follows:

$$\mathbf{x}^{k+1} = \mathbf{C}_{1k}\mathbf{z}^k + \mathbf{C}_{2k}\mathbf{u}^k + \mathbf{W}_{1k}\mathbf{y}, \quad (1)$$

$$\mathbf{z}^{k+1} = S_{\beta_k}(\zeta_k\mathbf{u}^k + \mathbf{x}^{k+1}), \quad (2)$$

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \tau_k(\mathbf{x}^{k+1} - \mathbf{z}^{k+1}). \quad (3)$$

Here, $\mathbf{C}_{1k}$ resembles $(\mathbf{I} + \frac{\alpha}{\rho}\mathbf{M}^T\mathbf{M})^{-1}$, $\mathbf{C}_{2k}$ resembles $(\rho\mathbf{I}+\alpha\mathbf{M}^T\mathbf{M})^{-1}r$, and $\mathbf{W}_{1k}$ resembles $(\frac{\rho}{\alpha}\mathbf{I}+\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$. The matrices $\mathbf{C}_{1k}$ and $\mathbf{C}_{2k}$ are learnable weight matrices. The constant $\frac{1}{\rho}$ is implemented using ζ_k, a learnable parameter. The shrinkage operator with threshold value $\frac{\gamma}{\rho}$ is implemented using the layer $S_{\beta_k}(\cdot)$, with β_k as the learnable threshold value, specific to each IB.

The architecture and iterative updates are visually represented in Figure. 1

B. Simulation of Missing Data Scenarios

To evaluate the performance of the imputation methods, we simulate practical missing data scenarios commonly encountered in PAT and other imaging applications. The following scenarios are considered:

- **Random Sensor Failures**: Sensors can fail randomly due to hardware malfunctions or temporary issues, leading to sporadic missing data across the sensor array. To simulate this, we randomly set a certain percentage of data points to zero. For instance, if 10% of the data is considered missing, we randomly select 10% of the total data points and set their values to zero.
- **Temporary Communication Loss**: Intermittent communication issues can cause temporary loss of data from certain sensors. It simulated by setting blocks of 10 consecutive samples to zero across all sensors to mimic temporary communication interruptions. This method resulted in approximately 14.29% of the data being set to zero, helping to analyze the impact of systematic data loss on image reconstruction quality.
- **Environmental Interference**: Environmental factors such as obstructions or external noise can interfere with certain sensors, causing localized data loss. Environmental interference was simulated by setting data for specific sensors (5, 15, and 25) to zero for samples 30 to 45, mimicking interference in localized regions. This resulted in approximately 8% of the sensor data being lost.

These scenarios help in understanding the impact of different types of missing data on the reconstruction quality and the effectiveness of various imputation methods.

C. Imputation Methods

To address the missing data in the simulated scenarios, we employ and compare several imputation methods:

- **Mean Imputation**: This simple technique involves replacing missing data points with the mean value of the observed data. Although it is easy to implement, it does not account for the relationships between data points and can introduce bias.
- **Median Imputation**: Similar to mean imputation, but missing values are replaced with the median of the observed data. Median imputation is more robust to outliers compared to mean imputation.
- **KNN Imputation**: KNN imputation replaces missing values based on the k-nearest neighbors' observed data points. This method considers the relationships between data points and can provide more accurate imputations than mean or median imputation.
- **MICE**: MICE is an advanced imputation technique that performs multiple imputations by creating several plausible imputed datasets and then averaging the results. It accounts for the uncertainty around missing values and leverages the correlations among data points.

These methods are applied to the simulated datasets to fill in the missing values, and their performance is compared in terms of reconstruction quality.

D. Evaluation Metrics

To assess the performance of the imputation methods and the overall image reconstruction quality, we use the following evaluation metrics:

- **Multi-Scale Structural Similarity Index (MS-SSIM)**: MS-SSIM evaluates the structural similarity between the reconstructed image and the original image at multiple scales. It provides a comprehensive measure of image quality by considering luminance, contrast, and structural information.
- **Structural Similarity Index (SSIM)**: SSIM measures the similarity between two images based on luminance, contrast, and structure. It is a widely used metric for assessing image quality and is sensitive to changes in image structure.

- **Peak Signal-to-Noise Ratio (PSNR)**: PSNR measures the ratio between the maximum possible value of a signal and the power of corrupting noise. It is commonly used to evaluate the quality of reconstructed images, with higher values indicating better quality.
- **R-squared (R^2)**: R^2 quantifies the proportion of variance in the observed data that is predictable from the reconstructed data. It is a measure of goodness-of-fit, with values closer to 1 indicating better fit.
- **Normalized Mean Squared Error (NMSE)**: NMSE measures the mean squared error between the reconstructed and original images, normalized by the variance of the original image. Lower values indicate better reconstruction accuracy.
- **Entropy**: Entropy measures the randomness or complexity of the image. In the context of image reconstruction, it provides insights into the information content and quality of the reconstructed image.

These metrics provide a comprehensive evaluation of the imputation methods and the overall performance of the image reconstruction process. By comparing the results across different imputation techniques and missing data scenarios, we aim to identify the most effective methods for handling missing data in limited-view circle sensor arrangements in PAT imaging.

IV. Experimental Results

This section presents the experimental results obtained from the proposed methodology. We evaluate the performance of the imputation methods under various missing data scenarios and assess the reconstruction quality of PAT images using the proposed deep learning network.

A. Experimental Setup

The experiments were conducted using a limited-view semicircle sensor arrangement, as shown in Figure 2. The setup consists of 71 sensors arranged in a semicircular pattern, capturing 75 samples each. This arrangement simulates practical scenarios [15] where complete circular data acquisition is not feasible.

The data used for the experiments was subjected to different missing data scenarios, including random sensor failures, temporary communication loss, and environmental interference. The missing data was then imputed using various imputation methods such as mean, median, KNN, and MICE. The imputed data was subsequently fed into the trained deep learning network to reconstruct the PAT images. The performance analysis were conducted to reconstruct three test images as shown in Figure 3.

B. Results and Discussion

Analysis of the three cases of practical missing data scenarios are discussed in this section. Table I presents the results for Images 1, 2, and 3 with random data missing, communication loss, and environmental interference. The "No missing" case, where there is no missing data, serves as the reference for optimal image reconstruction quality.

TABLE I: U-rBP-ADMM performance with missing data.

Case	MS-SSIM	SSIM	PSNR	R^2	NMSE	Entropy
Random Data Missing						
Image 1						
No missing	0.9730	0.7768	31.99	0.9652	0.01968	5.664
Missing	0.9455	0.6529	28.27	0.9213	0.04454	5.725
Mean Imputed	0.9599	0.6943	29.91	0.9469	0.03007	5.797
Median Imputed	0.9598	0.6968	29.88	0.9455	0.03086	5.726
KNN Imputed	0.9415	0.6379	28.61	0.9297	0.03981	5.961
MICE Imputed	0.9664	0.7411	31.13	0.9573	0.02420	5.668
Image 2						
No missing	0.9627	0.7130	30.01	0.9654	0.02020	6.128
Missing	0.9313	0.5917	26.54	0.9262	0.04305	6.134
Mean Imputed	0.9487	0.6312	28.24	0.9501	0.02910	6.246
Median Imputed	0.9492	0.6387	28.31	0.9496	0.02936	6.129
KNN Imputed	0.9088	0.5315	25.78	0.9140	0.05016	6.425
MICE Imputed	0.9570	0.6857	29.27	0.9596	0.02355	6.199
Image 3						
No missing	0.9505	0.6433	28.42	0.9369	0.04489	5.320
Missing	0.9146	0.5313	26.25	0.8984	0.07235	5.395
Mean Imputed	0.9295	0.5545	27.08	0.9182	0.05826	5.510
Median Imputed	0.9275	0.5558	27.01	0.9155	0.06018	5.448
KNN Imputed	0.9132	0.5243	26.51	0.9058	0.06707	5.532
MICE Imputed	0.9409	0.6017	27.84	0.9281	0.05116	5.343
Communication Loss						
Image 1						
Missing	0.8973	0.5496	25.34	0.8498	0.08505	5.797
Mean Imputed	0.9279	0.6186	27.66	0.9140	0.04867	5.991
Median Imputed	0.9317	0.6264	27.85	0.9160	0.04755	5.882
KNN Imputed	0.8846	0.5334	25.70	0.8655	0.07614	6.099
MICE Imputed	0.9161	0.5601	27.77	0.9051	0.05374	5.847
Image 2						
Missing	0.8822	0.4844	23.78	0.8607	0.08121	6.146
Mean Imputed	0.9151	0.5397	25.58	0.9106	0.05213	6.328
Median Imputed	0.9197	0.5511	25.65	0.9104	0.05225	6.224
KNN Imputed	0.8443	0.4482	23.03	0.8407	0.09289	6.445
MICE Imputed	0.8757	0.4701	24.54	0.8821	0.06873	6.334
Image 3						
Missing	0.8906	0.4714	25.01	0.8616	0.09850	5.323
Mean Imputed	0.9133	0.5092	26.42	0.9057	0.06715	5.624
Median Imputed	0.9110	0.5156	26.21	0.9000	0.07122	5.548
KNN Imputed	0.8859	0.4660	24.95	0.8679	0.09405	5.664
MICE Imputed	0.8909	0.4614	25.54	0.8801	0.08537	5.719
Environmental Interference						
Image 1						
Missing	0.9665	0.7540	31.13	0.9580	0.02376	5.645
Mean Imputed	0.9685	0.7533	31.66	0.9627	0.02111	5.710
Median Imputed	0.9697	0.7582	31.72	0.9632	0.02084	5.699
KNN Imputed	0.9575	0.7225	31.02	0.9572	0.02422	5.764
MICE Imputed	0.9523	0.7066	30.67	0.9515	0.02745	5.744
Image 2						
Missing	0.9570	0.6920	29.56	0.9614	0.02252	6.084
Mean Imputed	0.9564	0.6833	29.64	0.9625	0.02189	6.132
Median Imputed	0.9590	0.6936	29.77	0.9634	0.02136	6.120
KNN Imputed	0.9181	0.6073	27.31	0.9372	0.03662	6.222
MICE Imputed	0.9093	0.5921	26.89	0.9285	0.04167	6.227
Image 3						
Missing	0.9489	0.6353	28.37	0.9360	0.04559	5.304
Mean Imputed	0.9468	0.6221	28.25	0.9348	0.04640	5.380
Median Imputed	0.9458	0.6175	28.23	0.9343	0.04677	5.376
KNN Imputed	0.9469	0.6228	28.27	0.9349	0.04634	5.358
MICE Imputed	0.9469	0.6251	28.35	0.9349	0.04634	5.303

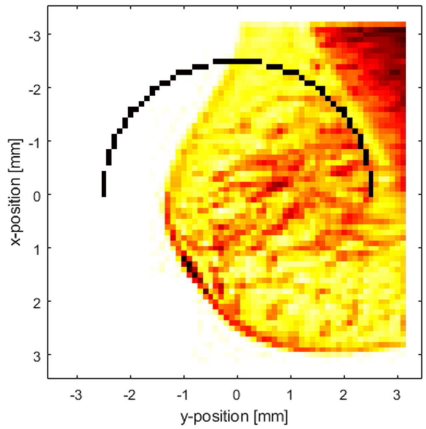
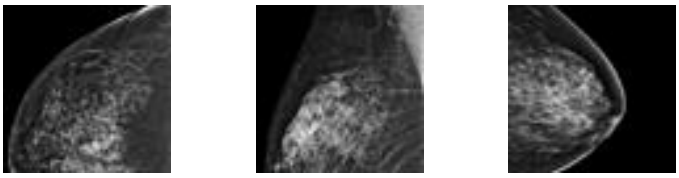


Fig. 2: Limited-View semicircle sensor arrangement.



(a) Image 1 (b) Image 2 (c) Image 3

Fig. 3: Test images.

1) Random Data Missing: The "No missing" case consistently shows the highest values across all metrics, indicating the best quality reconstruction. In the presence of missing data, there is a significant drop in image quality. Among the imputation methods, the "MICE Imputed" case provides the best performance overall, closely related to the no missing data scenario. For instance, in Image 1, the "No missing" case has the highest MS-SSIM (0.9730), SSIM (0.7768), PSNR (31.99), and R-squared (0.9652), and the lowest NMSE (0.01968) and Entropy (5.664). The "MICE Imputed" case shows the best performance among the imputed cases, with MS-SSIM of 0.9664, SSIM of 0.7411, and PSNR of 31.13. The "Mean Imputed" and "Median Imputed" cases show comparable performance, demonstrating noticeable improvement over the missing data case but not reaching the level of MICE imputation. The "KNN Imputed" case generally shows the lowest performance among the imputation methods tested.

2) Communication Loss: Among the imputation methods evaluated, the "MICE Imputed" method consistently yielded the best results, closely resembling the no missing data scenario. For example, in Image 1, the "MICE Imputed" case achieved the highest MS-SSIM (0.9161), SSIM (0.5601), PSNR (27.77), R-squared (0.9051), and maintained a lower Entropy (5.847) compared to other imputation methods. The "Mean Imputed" and "Median Imputed" methods showed comparable performance, typically being the second-best and third-best methods, respectively. They improved image quality over the "Missing" case but did not perform as well as MICE. The "KNN Imputed" method consistently showed the lowest performance among the imputation methods tested, indicating its relative ineffectiveness in handling communication loss scenarios.

3) Environmental Interference: From the analysis of all images, it is clear that the "MICE Imputed" method consistently provides the best results, closely approximating the no missing data scenario. For example, in **Image 1**, the "MICE Imputed" case achieves the highest performance with MS-SSIM of 0.9523, SSIM of 0.7066, PSNR of 30.67, R-squared of 0.9515, and entropy of 5.744. The "Mean Imputed" and "Median Imputed" cases also show substantial improvements over the missing data scenario, with "Median Imputed" slightly outperforming "Mean Imputed". The "KNN Imputed" case generally shows the least effective performance, often close to or slightly better than the "Missing" data case.

Overall, from the analysis, it is evident that missing data, whether from random data loss, communication loss, or environmental interference, significantly impacts the quality of image reconstruction in photoacoustic tomography. In all scenarios, the "No missing" data case provides the best image reconstruction quality across all images. When data is missing, there is a substantial drop in image quality. Among the imputation methods, "MICE Imputed" consistently yields the best results, closely approximating the no missing data case in all types of missing data scenarios. The "Mean Imputed" and "Median Imputed" cases show comparable results, better than the missing data scenario but not as good as MICE. The "KNN Imputed" case generally shows the lowest performance among the imputation methods tested. While environmental interference impacts image quality, its effect is less severe compared to random data loss and communication loss. Overall, "MICE Imputed" emerges as the most effective imputation method, followed by "Mean Imputed" and "Median Imputed," with "KNN Imputed" being the least effective.

V. Conclusion

In this study, we evaluated the effectiveness of various imputation methods—mean imputation, median imputation, KNN, and MICE for handling missing sensor data in PAT under different scenarios: random missing data, communication loss, and environmental interference. Our findings indicate that missing data significantly degrades image reconstruction quality in PAT. Among the tested imputation methods, MICE consistently outperformed others, yielding the best image reconstruction metrics across all scenarios. Mean and median imputations provided moderate improvements, while KNN generally exhibited the lowest performance. These results highlight the importance of selecting robust imputation strategies, such as MICE, to mitigate the impact of missing data on PAT imaging. Future research should explore further optimization of imputation techniques and their application to various imaging contexts.

References

[1] Wang, Lihong V. "Prospects of photoacoustic tomography." *Medical Physics*, vol. 35, no. 12, 2008, pp. 5758–5767. Wiley Online Library.

[2] Gu, Yanru, Sun, Yuanyuan, Wang, Xiao, Li, Hongyu, Qiu, Jianfeng, and Lu, Weizhao. "Application of photoacoustic computed tomography in biomedical imaging: A literature review." *Bioengineering & Translational Medicine*, vol. 8, no. 2, 2023, p. e10419. Wiley Online Library.

[3] Nishihara, Robert, Lessard, Laurent, Recht, Ben, Packard, Andrew, and Jordan, Michael. "A general analysis of the convergence of ADMM." In *International Conference on Machine Learning*, 2015, pp. 343–352. PMLR.

[4] R. N. Faizin, M. Riasetiawan, and A. Ashari, "A review of missing sensor data imputation methods," in *2019 5th International Conference on Science and Technology (ICST)*, vol. 1, pp. 1–6, 2019.

[5] J. Zhou and Z. Huang, "Recover Missing Sensor Data with Iterative Imputing Network," *ArXiv*, vol. abs/1711.07878, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:30222955

[6] R. Wu, S. D. Hamshaw, L. Yang, D. W. Kincaid, R. Etheridge, and A. Ghasemkhani, "Data imputation for multivariate time series sensor data with large gaps of missing data," *IEEE Sensors Journal*, vol. 22, no. 11, pp. 10671–10683, 2022.

[7] Y. Liu and V. Gopalakrishnan, "An overview and evaluation of recent machine learning imputation methods using cardiac imaging data," *Data*, vol. 2, no. 1, p. 8, 2017.

[8] J. Yang, Y. Shao, C. Li, and W. Wang, "A multistage deep imputation framework for missing values large segment imputation with statistical metrics," *Applied Soft Computing*, vol. 146, p. 110654, 2023.

[9] H. Sagha, J. R. Millán, and R. Chavarriaga, "A Probabilistic Approach to Handle Missing Data for Multi-Sensory Activity Recognition," in *Ubiquitous Computing*, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:12966009

[10] B. Agbo, H. Al-Aqrabi, R. Hill, and T. Alsboui, "Missing data imputation in the Internet of Things sensor networks," *Future Internet*, vol. 14, no. 5, p. 143, 2022.

[11] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[12] M. J. John and I. Barhumi, "Fast and efficient pat image reconstruction algorithms: A comparative performance analysis," *Signal Processing*, vol. 201, p. 108691, 2022.

[13] Chen, Shaobing, and Donoho, David. "Basis pursuit." In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, vol. 1, 1994, pp. 41–44. IEEE.

[14] M. John and I. Barhumi, "Advancing Sensor-Data Based PAT Image Reconstruction Through Efficient and Intelligible Unrolled Networks," *IEEE Access*, vol. 11, pp. 117053–117066, 2023.

[15] John, Mary Josy, and Barhumi, Imad. "Compressive Sensing Based Algorithms for Limited-View PAT Image Reconstruction." In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 1317–1322. IEEE.

Green Video Camouflaged Object Detection

Xinyu Wang*, Hong-Shuo Chen*, Zhiruo Zhou*, Suya You*, Azad M. Madni* and C.-C. Jay Kuo*
* University of Southern California, USA
{xwang350, hongshuo, zhiruozh, suya, azad.madni, jckuo}@usc.edu

***Abstract*—Camouflaged object detection (COD) aims to distinguish hidden objects embedded in an environment highly similar to the object. Conventional video-based COD (VCOD) methods explicitly extract motion cues or employ complex deep learning networks to handle the temporal information, which is limited by high complexity and unstable performance. In this work, we propose a green VCOD method named GreenVCOD. Built upon a green ICOD method, GreenVCOD uses long- and short-term temporal neighborhoods (TN) to capture joint spatial/temporal context information for decision refinement. Experimental results show that GreenVCOD offers competitive performance compared to state-of-the-art VCOD benchmarks.**

I. Introduction

Video Camouflaged Object Detection (VCOD) identifies objects that blend into their surroundings within video footage. Unlike image camouflaged object detection, VCOD needs to address more challenges posed by object movement, camera shaking, constantly changing environments, and the inherent variability and complexity of video scenes. In addition, the hidden object has a color or texture pattern highly similar to the background, further increasing the difficulty compared to traditional target detection and tracking tasks. Despite its challenges, VCOD is essential for a wide range of applications, such as surveillance and security [1], medical imaging [2]–[4], and autonomous driving.

Existing video detection approaches have demonstrated the importance of incorporating temporal motion cues of target objects. Traditional techniques, such as frame difference and optical flow, explicitly capture motion information through homo-graphic calculation and angle estimation. These motion patterns suggest that background noise easily influences the position and motion trajectory of the camouflaged object and may lead to the accumulation of errors. Implicit methods utilizing feature or prediction correlations have higher self-correction capabilities, but the learning results are highly dependent on the training samples and lack generalization ability. Also, the implicit motion cues are not explicitly evaluated, which makes the learned motion cues less reliable [5]. Consequently, extracting and incorporating frame-level spatial information and temporal motion cues improves detection accuracy and efficiency.

Additionally, given the applications in wildlife species protection and security operations, developing compact models for video object detection is essential to facilitate edge computing and enhance flexibility in mobile devices. Current models, which typically employ encoder-decoder architectures or deep neural networks, often require additional motion estimation modules. This requirement can significantly increase the computational cost, highlighting the need for more streamlined solutions that maintain performance while reducing resource demands.

In this work, we propose GreenVCOD, a novel green learning framework to deal with the VCOD problem feedforwardly. The framework comprises two essential components: the initial-stage prediction module and the temporal refinement module. The problem is first reduced into a series of image-based detection tasks to obtain the initial prediction maps. A cascaded multi-scale decision-boosting network is utilized to refine the prediction graph as the resolution increases gradually. Next, we incorporate temporal information to refine the initial predictions. Instead of explicit motion estimation, our temporal neighborhood prediction cube fixes the focus on a certain spatial location relative to the whole frame. It captures the scene change among adjacent frames, providing implicit spatial-temporal context information without additional computational cost. With features extracted at the initial stage, an additional XGBoost classifier is trained for temporal decision refinement. We build a parallel structure to balance the variety of object movements, including short-term and long-term refinement modules. In the concluding phase, a decision ensemble module is employed to integrate the outcomes derived from both the short-term and long-term refinement modules, thereby ensuring robustness and accuracy in the final result.

To summarize, this work has three significant contributions. First, we propose a novel framework, GreenVCOD, that combines frame-based camouflage detection with temporal refinement, realizing effective video camouflage detection in a feedforward manner. Second, we propose a temporal neighborhood prediction method. It extracts spatial information and motion cues while requiring no extra computational cost. Third, we build a parallel short - and long-term decision refinement structure that makes a good trade-off between different object motion patterns.

The rest of this paper is organized as follows. The related work is reviewed in Sec. II. The GreenVCOD method is described in Sec. III. Experimental results are shown in Sec.IV. Finally, the concluding remarks and future extensions are given in Sec.V.

II. Review of Related Work

A. Image Camouflaged Object Detection

Camouflaged object detection (COD) has become crucial in computer vision, focusing on identifying objects hidden in an

indistinguishable background. Due to the intrinsic similar patterns and unclear boundaries between objects and backgrounds, it presents a challenging task even for human eyes. A coarse-to-fine design is commonly adopted to tackle this problem. Some methods, such as SINet [2], ANet [6], and PFNet [7], employ a joint image classification and segmentation framework. By incorporating multi-tasking analysis, these methods first identify potential regions where camouflaged objects might exist and then apply segmentation tasks to refine the initial prediction maps. DGNet [8] decouples the task into context and texture encoders and links two branches by a gradient-induced transition to achieve joint feature learning. MGL [9] and BSANet [10] use edge details to refine segmentation decisions with boundary-aware modules. Inspired by U-Net [11], GreenCOD [12] constructs a cascaded multi-resolution boosting network that progressively refines the prediction map as the resolution increases.

B. Video Camouflaged Object Detection

Video Camouflaged Object Detection (VCOD) poses more challenges than image-based camouflaged object detection problems. It requires distinguishing objects in static images and dealing with dynamic scenes and potential noise introduced by the moving cameras. Combining motion cues with frame-by-frame predictions is the key to solving this problem. Conventional object detection methods explicitly calculate object motions by assessing differences between consecutive frames. Lamdouar et al. [13] proposed a model that integrates a differentiable registration module and a motion segmentation module, utilizing optical flow and difference images obtained from the differentiable module as input to produce camouflage segmentation results. The effectiveness of this model heavily relies on the performance of the explicit motion extraction process. Instead, SLTNet [14] uses a dense correlation pyramid to capture motions between neighboring frames implicitly. Our approach extracts motion cues via temporal neighborhood prediction cube reconstruction without requiring extra motion estimation modules and additional computational costs.

III. GreenVCOD Method

The green video camouflaged object detection method (GreenVCOD) is proposed in this section. First, we develop an initial-stage prediction module that extracts features and generates pixel-wise initial predictions using cascaded multi-resolution stages, as detailed in Sec.III-A. Next, we incorporate temporal information to refine the initial predictions. This temporal refinement consists of the short-term and long-term refinement modules described in Sec.III-B. In the concluding phase, a decision ensemble module, detailed in Sec.III-C, integrates the outcomes derived from both the short-term and long-term refinement modules, ensuring robustness and accuracy in the outcome. The overall framework is depicted in Fig.1.

A. Initial-stage Prediction Module

Videos are composed of successive image sequences and most video frames contain the target camouflaged object, hence we can realize a high-quality detection on the entire video, if we achieve a high quality image-level detection. Based upon this, we first simplify the video camouflaged object detection problem into a sequence of image-based or frame-based object detection problems and develop an initial-stage prediction. The initial-stage prediction module is built based on GreenCOD [12]. It consists of a feature extraction module and a multi-resolution decision module.

1) Feature Extraction Module: The feature extraction module employs the ImageNet pre-trained EfficientNetB4 backbone to efficiently extract discriminant features for each frame. The input frames are resized into $672 \times 672 \times 3$ and fed into the EfficientNetB4 backbone for feature extraction. Next, the features from different blocks are concatenated and resized to 168×168 standard size.

2) Multi-resolution Decision Module: The multi-resolution decision module adopts U-Net fashion to build a multi-resolution XGBoost boosting framework. As shown in Fig.1, this boosting framework utilizes four cascaded classifiers, where each classifier is trained using features of a certain resolution and the prediction result from the previous classifier. The standard size features in the feature extraction module are adjusted to a certain resolution to feed into each classifier. As the cascade progresses, the resolution of the features gradually increases. Specifically, the last classifier takes features of a standard size, while the previous classifier has one-fourth the resolution of the latter. Since the predictions of each layer classifier are pixel-wise results, to enrich the prediction map fed to deeper classifiers, the predictions of the 19×19 neighborhood centered on that pixel are considered the prediction information passed to the next classifier. The prediction map gradually enlarges throughout the boosting process, enhancing the overall accuracy.

B. Temporal Refinement Module

Video data often contains dynamic changes of the position and appearance of objects over time, hence static image information alone may not fully capture the behavior of objects. In challenging tasks like camouflaged video object detection, temporal information extraction is especially crucial. Here, we build a temporal refinement module to overcome the limitations of single-frame detection and provide more reliable and robust detection in complex scenes.

1) Temporal Neighborhood (TN) Prediction Cube Reconstruction: Traditional methods incorporate individual motion detection modules, such as optical flow, to capture the motion information of successive frames. Those explicit methods consume extra computation costs, and the inaccurate motion result may lead to unreliable predictions in subsequent frames. In addition, the optical flow and the pixel-wise classifier lack necessary contextual information, while the extracted motion information cannot be directly fed into the pixel-wise classifier.

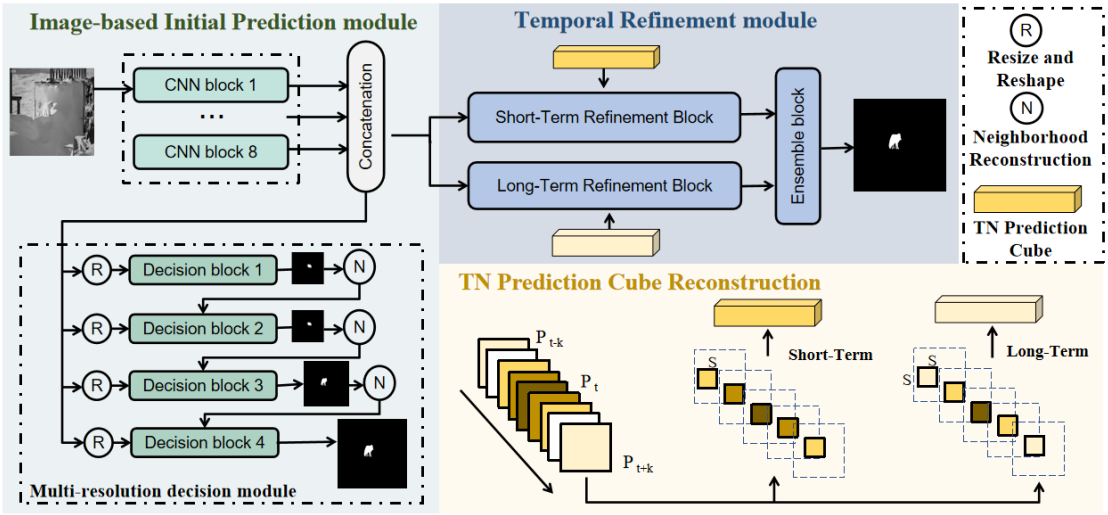


Fig. 1. The system diagram of the proposed green video camouflaged object detection method (GreenVCOD).

To provide adequate temporal information, we have to consider the extraction of features and the estimation of the motion position, which usually requires more computing resources and running time.

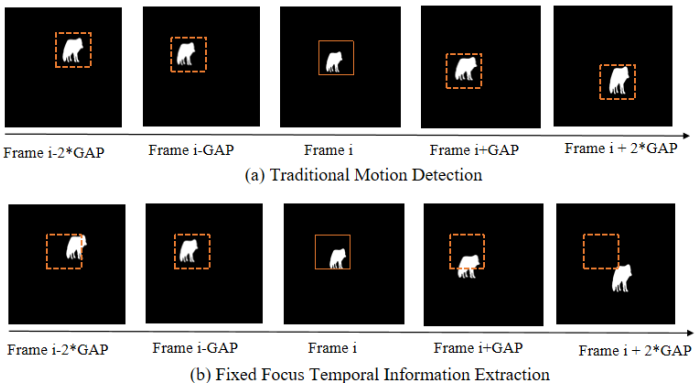


Fig. 2. Comparison between traditional explicit motion detection and our proposed fixed focus temporal information extraction.

To solve this problem, we propose the temporal neighborhood (TN) prediction cube to capture the temporal information implicitly. Prior works prefer to fix the background and move the focus along the movement track of the object relative to the background. Our method assumes the focus is fixed on a specific spatial location relative to the whole frame and captures the scene change within a particular time range. Since the object's movement is relatively slow between adjacent frames, fixing the focus can capture the object from different angles or focus on other parts, thereby providing adequate temporal information.

Compared with feature maps, prediction maps have only one value at each pixel location and are more efficient in providing spatial-context information. In the initial-stage XGBoost boosting framework, we adopt neighborhood predictions to enrich the feature set for decision refinement. If we consider the prediction map of each frame as an $H \times W \times 1$ cube, the spatial neighborhood is an $S \times S \times 1$ cube cropped from the original prediction cube, where the H, W, S represent the image height, the image width, and the side length of the square neighborhood region. Since the depth equals one, the prediction cube only contains spatial-context information. As described in Fig.3, when we consider all frames in a certain video, it will construct an $H \times W \times F$ prediction cube, where F is the number of frames. By extracting the neighborhood area at the same spatial location for K successive frames, an $S \times S \times K$ prediction cube is constructed containing spatial-context and temporal-context information.

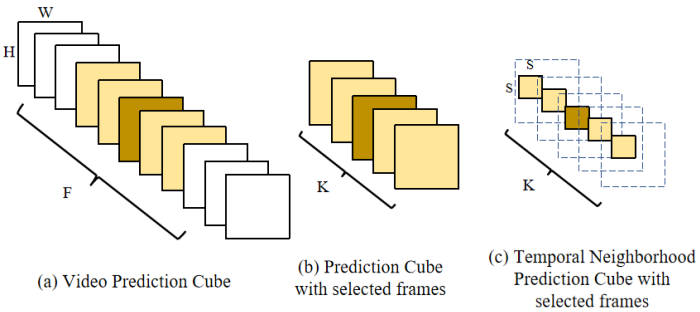


Fig. 3. TN Prediction Cube Reconstruction: The brown frame represents the current frame, while the yellow frames represent the selected adjacent frames.

For each single frame, we consider the current frame, $(K-1)/2$ previous frames, and $(K-1)/2$ subsequent frames to build the TN prediction cube for the frames at the beginning and the end of the video, we will retrieve the remaining frames in reverse direction. For instance, if $K = 5$ and we are at the video's final frame, we will mirror the two frames before the final frame to serve as the previous and subsequent frames of the final frame.

2) Long-Term and Short-Term Decision Refinement: After obtaining the TN prediction cube, we reshape it into a pre-

diction feature vector and concatenate that with the initial-stage features of the target pixel. As described in Fig.4, the enriched feature set trains an additional XGBoost to obtain refined decisions. By introducing motion detection or utilizing temporary information, more prosperous and more coherent contexts, and clues are captured to enable decision correction of complex frames.

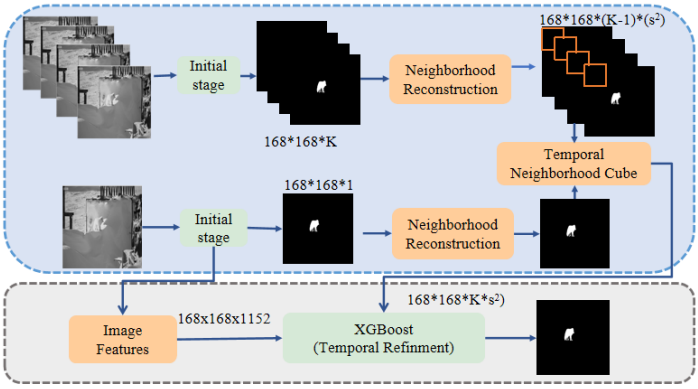


Fig. 4. Temporal Refinement Module Diagram.

Furthermore, when constructing a prediction cube, we select frames using a specific gap, referred to as GAP, rather than simply choosing K successive frames. For instance, if we are analyzing the ith frame with $GAP = 2$, the frames used to construct the prediction cube would be the $i-4^{th}, i-2^{th}, i^{th}, i+2^{th}, i+4^{th}$ frames. The conventional strategy of considering successive K frames is a case where $GAP = 1$. Building upon this, we refer to this approach as short-term when $GAP = 1$, and we call it long-term when $GAP > 1$. Since the long-term and short-term refinement modules only require different parameterized construction in the temporal prediction cubes, extracting temporal information under different term settings will not cause additional computation.

C. Decision Ensemble Module

The time range required to calculate adequate temporal information varies significantly in video object detection. Fast-moving targets necessitate a shorter time range to effectively capture meaningful information, while slow-moving targets benefit from a longer detection time. Our approach integrates short-term and long-term refinement strategies, where both refinement modules are built in a parallel architecture. Furthermore, we enhance our method by integrating prediction maps derived from these refinement modules and employing adaptive thresholds to balance the tradeoffs between short-term and long-term decisions, ensuring optimal performance across varying motion patterns.

IV. Experiments

We demonstrate the performance evaluation of our proposed framework on by MoCA-Mask dataset [14] in this section.

A. Datasets

MoCA-MASK is generated based on the original Moving Camouflaged Animals (MoCA) Dataset. The original ground truth bounding box is extended to dense segmentation masks for more comprehensive performance evaluation. The training set includes 71 videos with 19,313 frames, and the testing set contains 16 video sequences with 3,626 frames.

COD10K is one of the largest camouflage object detection datasets. It consists of 5066 camouflaged images and 1934 non-camouflaged images. The frame-based models within our initial-stage prediction modules are pre-trained on the COD10K [2] dataset.

B. Experimental Setup

We conduct the experiments on MoCA-Mask dataset. Since this dataset has insufficient frame and scene diversity and small camouflage objects, we choose to pre-train the cascaded multi-resolution decision part in the initial-stage prediction module is pre-trained on COD10K and CMAO [6] dataset. At the same time, the temporal refinement module is trained and fine-tuned only with MoCA-Mask training data. The final prediction map of each frame is of standard size 168 $\times$ 168, which will be resized back to its original frame size for performance evaluation. For a fair comparison with state-of-the-art, we utilize five evaluation metrics: enhanced-alignment measure (E_ϕ), weighted F-measure (F_w^β), mean absolute error (MAE), mean Dice (mDice) and the mean intersection of union (mIoU).

C. Performance Evaluation

1) Comparisons with State-of-The-Arts: As shown in Table I, we compare our proposed method with 6 state-of-the-art methods under MoCA-Mask dataset [14]. Our proposed method outperforms all methods on MAE (0.0079) and F_w^β (0.311), while we rank second in mean dice and third in average IOU. Notably, our approach is comparable in performance to recently proposed implicit motion-extraction-based methods, such as SLT-Net [14]. Fig.5 provides several visual examples to assess prediction quality. Compared to the most advanced detection methods, our method has more accurate detection results and higher visual quality, especially in the details of object shapes and more precise detection boundaries.

TABLE I
QUANTITATIVE RESULTS ON THE MOCA-MASK DATASET. THE BEST-PERFORMING METHOD OF EACH CATEGORY IS HIGHLIGHTED IN BOLD, WHILE THE SECOND PERFORMING METHOD IS HIGHLIGHTED WITH UNDERLINES.

	MoCA-Mask w/o pseudo labels #				
	$F_w^\beta \uparrow$	$E_\phi \uparrow$	$MAE \downarrow$	$mDice \uparrow$	$mIoU \uparrow$
SINet [2]	0.231	<u>0.699</u>	0.028	0.276	0.202
SINet-v2 [15]	0.204	0.642	0.031	0.245	0.180
PNS-Net [16]	0.097	0.510	0.033	0.121	0.101
RCRNet [17]	0.138	0.527	0.033	0.171	0.116
MG [18]	0.168	0.561	0.067	0.181	0.127
SLT-Net [14]	**0.311**	**0.759**	<u>0.027</u>	**0.360**	**0.272**
Ours	**0.311**	0.635	**0.00786**	<u>0.318</u>	<u>0.243</u>

We also show some sample video frames in Fig.6 to demonstrate performance on continuous video frames. For a more comprehensive evaluation, we provide two sample videos with different difficulty levels and focus, the Arctic Fox and Sand Cat videos. In the first video, the Arctic fox's shape and shooting angle change during rapid movement, while in the second video, the shading of bushes and intersecting light conditions make even ground truth unable to provide a clear outline of the sand cat. The experimental results demonstrate that our model can detect the camouflaged object stably and get clear moving animal boundaries under different challenging modes.

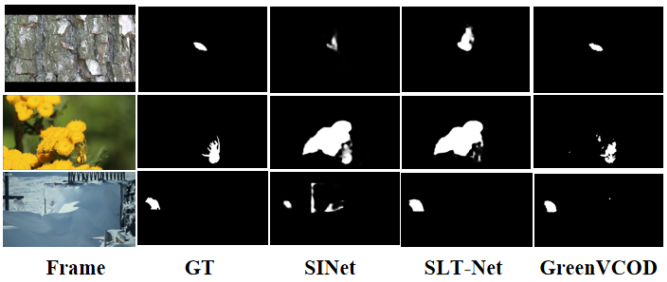


Fig. 5. Visual comparisons on example video frames with state-of-the-art camouflage detection methods.

2) Ablation Study: Short-term and Long-term Modules: For decision refinement, we use short - and long-term modules with parallel structures to balance our attention to the speed of motion of different objects. As shown in Table II, long-term optimization results outperform short-term optimization results in almost all evaluations. Because the different temporal modules extract complementary context information, the final ensemble result from the fusion of both optimizations is better than any single one in all evaluations.

TABLE II
ABLATION STUDIES OF SHORT-TERM AND LONG-TERM ARCHITECTURES ON MOCA-MASK TEST SET.

	MoCA-Mask w/o pseudo labels #				
	$F_w^\beta \uparrow$	$E_\phi \uparrow$	$MAE \downarrow$	$mDice \uparrow$	$mIoU \uparrow$
Ours (Short-Term)	0.284	0.630	0.00828	0.289	0.216
Ours (Long-Term)	0.304	0.629	0.00763	0.314	0.240
Ours (Ensemble)	0.311	0.635	0.00786	0.318	0.243

3) Computational Complexity in Inference: We calculate the model size and computational complexity of the proposed GreenVCOD framework, where the complexity is quantified through Multiply-Accumulate Operations (MACs) for evaluation. As shown in the table. III, the model size of GreenVCOD is approximately 19.9M, which is comparable to the smaller-sized image detection method SINet-v2 (26.98M). In addition, our model requires 17.42G MACs in the inference phase, of which 77.5% of the computational cost is due to the EfficientNet backbone. In contrast, the TN prediction cube is generated based on a prediction graph of successive frames and does not generate additional calculations, making the decision refinement module highly efficient.

TABLE III
THE NUMBER OF PARAMETERS AND COMPUTATIONAL COMPLEXITY OF GREENVCOD.

Module	Sub-Module	# Trees	Depth	Parameters (%)
Stage-1	EfficientNetB4	_	_	16,742,216 (84.1%)
Stage-1	XGBoost1	10000	3	220,000 (1.1%)
Stage-1	XGBoost2	10000	3	220,000 (1.1%)
Stage-1	XGBoost3	10000	3	220,000 (1.1%)
Stage-1	XGBoost4	10000	3	220,000 (1.1%)
Stage-2	XGBoost(long-term)	6000	6	1,140,000 (5.7%)
Stage-2	XGBoost(short-term)	6000	6	1,140,000 (5.7%)
Total	_	_	_	19,902,216

Module	Sub-Module	Map size	MACs (%)
Stage-1	EfficientNetB4	_	13,503,446,880 (77.5%)
Stage-1	XGBoost1	42×42	70,560,000 (0.4%)
Stage-1	XGBoost2	42×42	70,560,000 (0.4%)
Stage-1	XGBoost3	84×84	282,240,000 (1.6%)
Stage-1	XGBoost4	168×168	1,128,960,000 (6.4%)
Stage-2	XGBoost(long-term)	168×168	1,185,408,000 (6.8%)
Stage-2	XGBoost(short-term)	168×168	1,185,408,000 (6.8%)
Total	_	_	17,426,582,880

V. CONCLUSION AND FUTURE WORK

GreenVCOD was proposed for camouflage object detection in video footage in this work. It utilized an image-based COD method to obtain the initial prediction map and optimized predictions with short- and long-term ensembles. GreenVCOD offers competitive predictive performance and impressive efficiency compared to state-of-the-art VCOD methods. Due to the limitations of the MoCA-Mask dataset, most existing methods often exploited partial pre-training with more diverse datasets. As future extensions, we will address the few-shot camouflage detection problem.

REFERENCES

[1] T. Liu, Y. Zhao, Y. Wei, Y. Zhao, and S. Wei, "Concealed object detection for activate millimeter wave image," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9909–9917, 2019. DOI: 10.1109/TIE.2019.2893843.

[2] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *IEEE CVPR*, 2020, pp. 2777–2787.

[3] D.-P. Fan, T. Zhou, G.-P. Ji, *et al.*, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020. DOI: 10.1109/TMI.2020.2996645.

[4] D.-P. Fan, G.-P. Ji, T. Zhou, *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *International conference on medical image computing and computer-assisted intervention*, Springer, 2020, pp. 263–273.

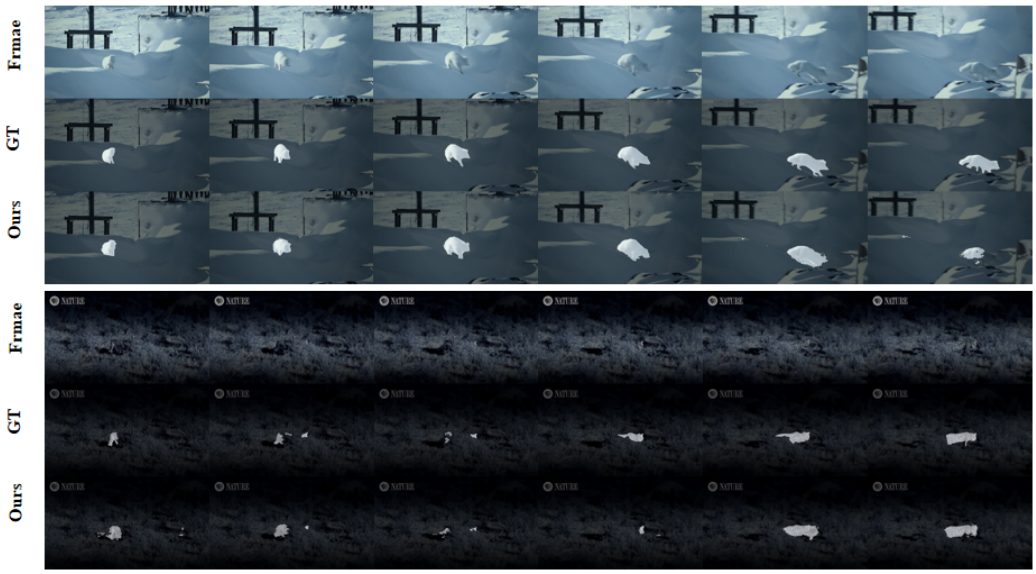


Fig. 6. Visual comparisons on consecutive video frames.

[5] X. Zhang, T. Xiao, G. Ji, X. Wu, K. Fu, and Q. Zhao, *Explicit motion handling and interactive prompting for video camouflaged object detection*, 2024. arXiv: 2403.01968 [cs.CV].

[6] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Computer Vision and Image Understanding*, vol. 184, pp. 45–56, 2019, ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2019.04.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314219300608.

[7] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8768–8777. DOI: 10.1109/CVPR46437.2021.00866.

[8] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, and L. Van Gool, "Deep gradient learning for efficient camouflaged object detection," *Machine Intelligence Research*, vol. 20, pp. 92–108, 1 2023.

[9] Q. Zhai, X. Li, F. Yang, *et al.*, "Mgl: Mutual graph learning for camouflaged object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 1897–1910, 2023. DOI: 10.1109/TIP.2022.3223216.

[10] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7471–7481. DOI: 10.1109/CVPR.2019.00766.

[11] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: 1505.04597 [cs.CV].

[12] H.-S. Chen, Y. Zhu, S. You, A. M. Madni, and C.-C. J. Kuo, "Greencod: A green camouflaged object detection method," *arXiv preprint arXiv:2405.16144*, 2024.

[13] H. Lamdouar, C. Yang, W. Xie, and A. Zisserman, *Betrayed by motion: Camouflaged object discovery via motion segmentation*, 2020. arXiv: 2011.11630 [cs.CV].

[14] X. Cheng, H. Xiong, D.-P. Fan, *et al.*, "Implicit motion handling for video camouflaged object detection," in *CVPR*, 2022.

[15] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE TPAMI*, vol. 44, no. 10, pp. 6024–6042, 2022. DOI: 10.1109/TPAMI.2021.3085766.

[16] G.-P. Ji, Y.-C. Chou, D.-P. Fan, *et al.*, "Progressively normalized self-attention network for video polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 142–152.

[17] P. Yan, G. Li, Y. Xie, *et al.*, "Semi-supervised video salient object detection using pseudo-labels," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7284–7293.

[18] C. Yang, H. Lamdouar, E. Lu, A. Zisserman, and W. Xie, "Self-supervised video object segmentation by motion grouping," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7157–7168. DOI: 10.1109/ICCV48922.2021.00709.

Improving Speaker Consistency in Speech-to-Speech Translation Using Speaker Retention Unit-to-Mel Techniques

Rui Zhou* and Akinori Ito† and Takashi Nose‡
* Graduate School of Engineering, Tohoku University, Sendai
E-mail: zhou.rui.p1@dc.tohoku.ac.jp
† Graduate School of Engineering, Tohoku University, Sendai
E-mail: aito.spcom@tohoku.ac.jp
‡ Graduate School of Engineering, Tohoku University,
E-mail: takashi.nose.b7@tohoku.ac.jp

***Abstract*—We propose a Speaker-Consistent Speech-to-Speech Translation (SC-S2ST) system that effectively retains speaker-specific information. While the paradigm of Speech-to-Unit Translation (S2UT) followed by Unit-to-Waveform Vocoder has become a mainstream for End-to-End S2ST systems, due to the substantial semantic content carried by discrete units, this approach primarily captures semantic information and often results in synthesized speech that lacks speaker-specific characteristics such as accent and individual voice qualities. Existing S2UT systems with style transfer face the issue of high inference latency. To address this limitation, we introduced a Speaker-Retention Unit-to-Mel (SR-UTM) framework designed to capture and preserve speaker-specific information. We conducted experiments on the CVSS-C and CVSS-T corpora for Spanish-English and French-English translation tasks. Our approach achieved BLEU scores of 16.10 and 21.68, which are comparable to those of the baseline S2UT system. Furthermore, our SC-S2UT system excelled in preserving speaker similarity. The speaker similarity experiments showed that our method effectively retains speaker-specific information without significantly increasing inference time. These results confirm that our primary approach successfully achieve speaker-consistent speech-to-speech translation.**

I. Introduction

Speech-to-speech translation (S2ST) refers to the process of converting spoken language in one language into spoken language with the same meaning in another language, facilitating communication between speakers of different languages. Traditional cascade S2ST systems operate by first using automatic speech recognition (ASR) to transcribe the source speech, followed by machine translation (MT) to translate the source text into the target text, and finally using text-to-speech (TTS) to synthesize the target speech. More recently, end-to-end speech-to-text (S2T) models[1] have been employed to replace the ASR and MT components, reducing the accumulation of errors.

However, the cascade systems mentioned above face significant challenges, such as long inference times and the large parameter size due to multiple models, making them impractical for industrial applications. Consequently, researchers have shifted their focus to E2E S2ST systems. Translatotron was the first proposed E2E S2ST, utilizing auxiliary recognition and attention-based sequence-to-sequence multitask learning to transform spectrograms from the source language to the target speech spectrograms[2]. Subsequently, Translatotron2, which utilizes a two-pass approach, was proposed[3]. This method establishes attention between the source language acoustic features and the target language phonemes, thereby reducing the complexity of directly predicting the target spectrogram. Additionally, phoneme-level content features are employed during inference, enhancing performance compared to the original Translatotron.

The Translatotron series of S2ST primarily focuses on transforming the spectrogram of the source language into the spectrogram of the target language. However, directly building a translation model between two spectrograms is exceptionally challenging due to the high dimension and complexity of spectrogram data. In recent years, with the advent of self-supervised learning, a new approach has emerged, which involves extracting speech representations using models like HuBERT[4] to capture rich and detailed features of the speech signal. These representations are then clustered to construct discrete units, which serve as a more manageable and semantically meaningful intermediate representation. A combination of unit-to-waveform conversion and advanced vocoders [5] has gradually become mainstream [6], [7].

Some previous works have applied style transfer in direct S2UT[8], [9]. These methods utilize an acoustic language model to generate units containing speaker information, which are then combined with semantic units and finally converted to speech using a vocoder. However, this approach used an auto-regressive method to generate units greatly increases inference time. Additionally, the large parameter size of the Residual Vector Quantization model results in reduced efficiency. To address these issues, we propose our Speaker-Retention Unit-to-Mel framework(SR-U2M), which enhances efficiency.

[0]This work was supported by JST SPRING, Grant Number JPMJSP2114 and JSPS KAKENHI JP 23K20725 and JP24H00085

There are two main motivations for our work. Firstly, we found that mel-spectrograms, compared to units, have higher dimensionality and contain richer information, making it easier to retain speaker information. Secondly, we were inspired by FastSpeech[10], a non-auto-regressive text-to-speech method that quickly synthesizes mel-spectrograms from discrete units.

Therefore, we introduce a novel structure SR-U2M, which integrates the speaker's reference utterance along with the speech units into the model. The SR-U2M framework first extracts speaker information using a speaker adapter, and then embeds this speaker information into the speech units. The combined information is subsequently converted into a mel-spectrogram through a unit-to-mel structure, and finally transformed into speech. This approach not only effectively preserves the speaker's unique characteristics but also significantly enhances inference efficiency.

The rest of this paper is organized as follows. First, we introduce related work in the next section. Then, we describe our SC-S2ST system and SR-U2M in Section III. Following this, we present and analyze our experimental results in Section IV. Section V summarizes with the experimental results and analysis. Finally, Section VI is the conclusion.

II. Related Works

A. Direct S2UT

Lee et al. were the first to propose a discrete unit-based S2ST system, utilizing a Transformer Encoder-Decoder structure to directly establish the relationship between the source mel spectrogram and the target discrete units[6]. This structure is very similar to that of S2T and employs the intermediate inputs of the encoder to conduct a multitask auxiliary recognition. They used HuBert to extract speech representations of the target language, followed by k-means clustering to derive the target language's discrete units. Leveraging the self-supervised learning's characteristic of not requiring paired text, they also proposed S2ST for unwritten languages[11], [12], which similarly only required the target language's discrete units rather than text. To address the issue of varying units for different speakers, Lee et al. further introduced norm-unit-based S2ST, which was applied to textless real data, thereby enhancing the robustness and applicability of the system[7]. Popuri et al. used pre-trained wav2vec instead of a transformer encoder to extract acoustic features, improving the model's effect[13]. Therefore, in our research, we also apply the S2UT method using pretrained wav2vec as the encoder. We will introduce the structure of our S2UT in the next section.

B. Speaker Information Retention

Preserving speaker-specific information during speech translation has been a long-standing challenge. Recent advancements have leveraged techniques from speaker verification to enhance multi-speaker speech synthesis. Jia et al. integrated a speaker encoder to extract speaker features then combined with the synthesizer encoder, facilitating the retention of speaker identity in the synthesized speech[14]. Translatotron also employed a similar approach by adding a speaker encoder to extract speaker information[2]. Therefore, we adopted this structure as well, with the difference being that we replaced the construction of the speaker encoder. Our method utilizes an ECAPA-TDNN-based speaker encoder, which has demonstrated superior performance in speaker verification tasks[15]. The detailed architecture of the ECAPA-TDNN will be elaborated in the next section.

III. Speaker-Consistent Speech-to-Speech Translation

A. Model Structure

Fig.1 illustrates the overall architecture of our proposed SC-S2ST. From left to right, it shows the overall framework, the SR-U2M structure, and the speaker adapter structure within SR-U2M. The source language speech is input into the S2UT module, which generates a sequence of speech units in the target language. These units are then processed by the Length Predictor to match the actual sequence. Simultaneously, the source speech is input into the Speaker Adapter to extract speaker-specific information. The SR-U2M module integrates the speech units and the extracted speaker information to generate a mel-spectrogram of the target speech, retaining the original speaker's characteristics. Finally, a mel vocoder converts the mel-spectrogram into the waveform of the target speech.

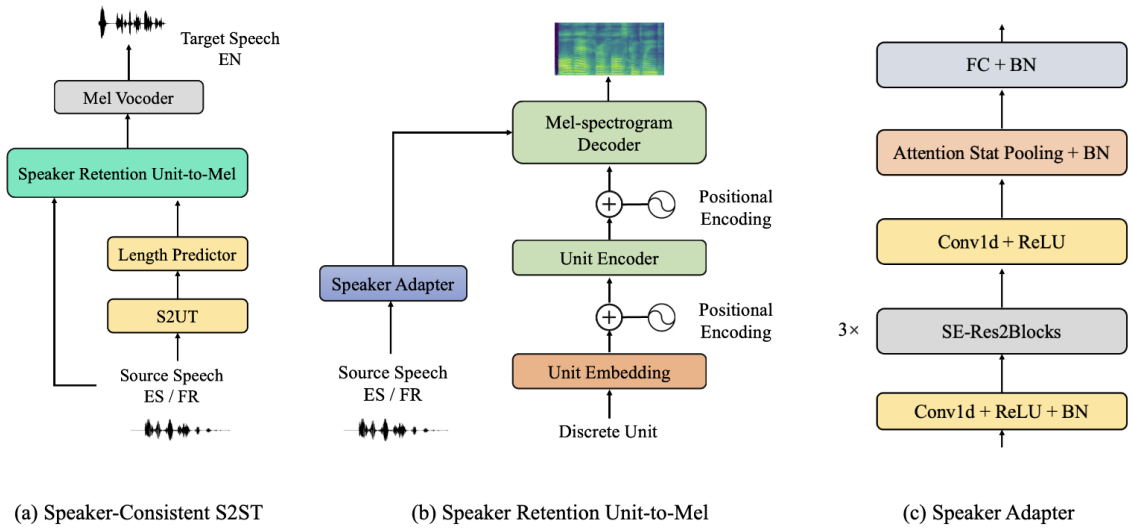
B. S2UT Model

Different from previous S2UT literature that used X-former (e.g., Transformer, Conformer)[6], [7], [11], [12], we utilize a pre-trained wav2vec2 model as the encoder. Wav2vec2 can extract richer speech representations from the original audio. For the decoder, we use the same Transformer decoder as others. Before training, we use the pre-trained HuBERT model to extract speech representations of the target language and then apply k-means clustering to obtain discrete units. Following Lee et al. we use a stacking method to convert adjacent identical units into a single unit, significantly speeding up both training and inference times[6]. Therefore, after S2UT, we need to add a length predictor to restore the units. Our length predictor is also derived from FastSpeech2, which consists of two 1D-convolutional layers, each with ReLU activation, followed by layer normalization and dropout, and a linear layer[16].

C. Speaker-Retention U2M

Our unit-to-mel approach is inspired by text-to-speech (TTS) systems, which convert discrete representations such as phonemes and text into mel-spectrograms. Examples of such systems include Tacotron and its variants[17], [18]. However, Tacotron uses an auto-regressive method, which is time-consuming during inference. To address this issue, we chose a FastSpeech-like structure that uses a non-auto-regressive approach to convert units into mel-spectrograms.

For the unit encoder, we employ the Feed-Forward Transformer (FFT) structure from FastSpeech[10]. This structure is similar to the encoder in the Transformer[19] model but



(a) Speaker-Consistent S2ST (b) Speaker Retention Unit-to-Mel (c) Speaker Adapter

Fig. 1. The overall architecture for Speaker-Consistent Speech-to-Speech Translation.

uses a 2-layer 1D convolutional network with ReLU activation instead of a 2-layer dense network, which has been shown to be more effective. Our mel-spectrogram decoder employs the traditional Transformer decoder structure. The specific parameters and configurations of our model will be detailed in the next section.

D. Speaker Adapter

The speaker adapter in our system is responsible for extracting speaker-specific information while disregarding other content information. Achieving this task primarily involves speaker identification[20]. The speaker adapter is designed to isolate and capture speaker-specific characteristics from the input speech signal. Our speaker adapter uses the ECAPA-TDNN architecture in its entirety to extract high-level speaker embeddings[15]. The structure is shown in subfigure(c) in Fig.1, which begins with a 1-dimensional convolutional layer followed by ReLU activation and batch normalization, capturing temporal features from the input speech signal. This is followed by three SE-Res2Blocks, which integrate Squeeze-and-Excitation mechanisms within residual blocks to emphasize important speaker-specific features. Another 1-dimensional convolutional layer with ReLU activation further refines these features. Attention-based statistical pooling then aggregates the most relevant frames into a fixed-dimensional speaker embedding, normalized by batch normalization. The final fully connected layer, also followed by batch normalization, transforms these aggregated features into the final speaker embedding, effectively encapsulating the unique vocal characteristics of the speaker.

IV. Experiments

A. Dataset

We performed our experiments using Spanish to English and France to English in CVSS multilingual speech-to-speech translation corpus[21]. The CVSS corpus includes two datasets: CVSS-C, which contains multi-speaker recordings in various languages translated into the voice of a single female English speaker, and CVSS-T, which contains multi-speaker recordings translated into multi-speaker English speech, maintaining similar pronunciation characteristics across both datasets. The number of samples and their respective durations are detailed in TABLE I.

We utilize the CVSS-T corpus to train our SR-U2M model because it provides the necessary multi-speaker target speech mel-spectrogram required for our approach. Conversely, the CVSS-C corpus is employed to train the S2UT model. Since discrete units only encapsulate semantic information, extracting these units from the same speaker's voice, as provided in the CVSS-C corpus, yields better performance.

B. System Implementation

1) SC-S2UT Model: we used the pre-trained HuBERT model[1] trained on LibriSpeech[22] and follow illustrate k-means with $K = 100$ to extract sixth layers representation to discrete unit. We utilized the pre-trained wav2vec to extract acoustic feature from source waveform, which trained on the es unlabeled subset of VoxPopuli corpus[23] for Spanish[2], and trained on 14K hours of French[3][24]. We did not employ multi-task learning in our experiments as focus of our study is to compare the extent to which our method can retain

[1]https://github.com/facebookresearch/fairseq/tree/main/examples/hubert
[2]https://github.com/facebookresearch/voxpopuli/
[3]https://huggingface.co/LeBenchmark/wav2vec2-FR-14K-large

TABLE I
STATISTICS (NUMBER OF SAMPLES AND DURATION) OF THE CVSS SPANISH-ENGLISH AND FRANCE-ENGLISH DATASET

	CVSS-C								CVSS-T		
	Es-En				Fr-En				Es-En		
	train	dev	dev-small	test	train	dev	dev-small	test	train	dev	test
# samples	79k	13.2k	0.99k	13.2k	20.7k	14.7k	0.49k	14.7k	79k	13.2k	13.2k
source (hrs)	69.5	12.4	1.31	12.4	174	13	0.64	13.3	73.7	13.2	13.3
target (hrs)	113.1	21.8	/	22.7	264.3	21.7	/	23.3	113.1	21.8	22.7

speaker information during the translation process. For SR-U2M, we compute 80-dimensional mel-spectrogram features at every 20-ms, making it matching the length of discrete unit. Hyperparameters of our model is shown in TABLE II.

TABLE II
HYPERPARAMETERS FOR SC-S2ST

Hyperparameter		SC-S2ST
S2UT Decoder	Unit neurons	103
	Decoder Block	6
	Decoder Hidden	512
	Encoder Attention Heads	8
	Encoder Dropout	0.1
SR-U2M Unit-Encoder Mel-Decoder	Encoder Block	6
	Decoder Block	6
	Encoder Kernel	31
	Hidden	512
	Attention Heads	8
	Dropout	0.1
Speaker Adapter	Channels	[1024, 1024, 1024, 1024, 3072]
	Kernel Sizes	[5, 3, 3, 3, 1]
	Dilations	[1, 2, 3, 4, 1]
	Groups	[1, 1, 1, 1, 1]
	Attention Channels	128

For waveform generation, we used the same unit-based vocoder as [13] serving as a baseline for comparison. In our SC-S2UT system, since the final output is a mel-spectrogram, we employ HiFi-GAN[5] as the vocoder[4] and trained with multispeaker LibriTTS corpus[25].

2) Baselines: We developed two cascaded baselines, ASR+MT+TTS and S2T+TTS, and one direct S2ST baseline, which also utilized a pre-trained wav2vec2 model as the encoder. For the Automatic Speech Recognition (ASR) component, we employed wav2vec2 pre-trained on XLSR[26] and fine-tuned it with Connectionist Temporal Classification (CTC) for French[5] and Spanish[6] speech recognition. For Machine Translation (MT), we used a pre-trained transformer model for both French and Spanish[7]. For Text-to-Speech (TTS), we leveraged the Massively Multilingual Speech (MMS) model for English text-to-speech[8][27]. Additionally, we used a transformer-based sequence-to-sequence (seq2seq) model for Spanish and French speech-to-text translation[28]. Furthermore, we implemented a wav2vec-based S2UT[13] and used a Unit-HiFi-GAN vocoder to convert the units into waveforms.

3) Evaluation: We evaluated the system's performance in three aspects: translation quality, speech quality, and speaker similarity. For translation quality, we followed previous literature by applying ASR to the speech output and computing BLEU scores. We used a transformer-based end-to-end model pre-trained on LibriSpeech, which achieved a WER of 2.27 on the test-clean set[9]. For translation quality, we used all development and test data in Spanish and French. For speech quality evaluation, we used only test data and utilized Mean Opinion Scores (MOS) to assess the naturalness of the output speech, with a scale from 1 (worst) to 5 (best). We used UTMOS based on the methodology described in [29]. We also conducted speed tests using the baseline for comparison. Our experiments were performed on an RTX 3090 GPU. Lastly, for speaker similarity evaluation, we used the CVSS-T corpus, which features source and target speakers with similar voices. Additionally, because our SR-UTM structure was trained using units and target speech, we compared the similarity between the translated speech and target speech as well as source speech. We extracted speaker features from the output speech using the Speaker Adapter structure within our system shown in subfigure(c) in Fig.1. We then computed normalized Mean Squared Error (NMSE), Standard Euclidean Distance (SED), and Cosine Similarity (COS-SIM) between different features to quantify speaker similarity. Since we trained our SR-UTM using only Spanish data, for similarity evaluation, we used a Spanish-English dataset. All our experiments were conducted using the SpeechBrain framework[30].

V. RESULTS

In this section, we present the results of our experiments on translation quality and speech quality. The primary goal of our approach is to maintain the source speaker's voice while achieving competitive translation and speech quality.

A. Translation Quality and Speech Quality

TABLE III compares the BLEU scores and MOS scores and inference time and speedup of our proposed method against the

[4]https://huggingface.co/speechbrain/tts-hifigan-libritts-16kHz
[5]https://huggingface.co/facebook/wav2vec2-large-xlsr-53-french
[6]https://huggingface.co/facebook/wav2vec2-large-xlsr-53-spanish
[7]https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/models
[8]https://huggingface.co/facebook/mms-tts-eng
[9]https://huggingface.co/speechbrain/asr-transformer-transformerlm-librispeech

TABLE III
BLEU SCORE AND MOS SCORE AND INFERENCE TIME AND SPEEDUP FOR OUR EXPERIMENT

	BLEU Score ↑				MOS Score ↑		Inference Time(s) ↓		Speedup ↑	
	ES-EN		FR-EN		ES-EN	FR-EN	ES-EN	FR-EN	ES-EN	FR-EN
	dev	test	dev	test	test	test	dev-small	dev-small	test	test
Direct System										
S2UT[13]	15.56	16.54	22.26	22.67	3.31 ± 0.11	3.29 ± 0.12	0.77	0.82	1.00×	1.00×
SC-S2UT	15.03	16.10	21.52	21.68	3.26 ± 0.13	3.20 ± 0.11	1.00	1.09	0.77×	0.75×
Cascade System										
ASR+MT+TTS	22.16	24.36	23.36	23.22	4.07 ± 0.06	4.10 ± 0.08	/	/	/	/
ST+TTS	13.82	14.13	21.05	20.72	4.11 ± 0.08	4.10 ± 0.07	/	/	/	/
Ground Truth	84.51	88.64	80.51	80.29	4.45 ± 0.11	4.44 ± 0.11	4.74	4.75	/	/

baseline. The MOS score is presented with the standard deviation. From the results, we observed that the ASR+MT+TTS cascade system still achieved the highest BLEU scores in both language pairs and datasets. Our proposed SC-S2UT method showed slightly lower BLEU scores compared to the baseline S2UT, indicating a small trade-off in translation quality when focusing on preserving the speaker's voice. However, the gap between our method and S2UT is very small, demonstrating that while our approach might lose a bit of content information due to its emphasis on retaining speaker characteristics, the loss is minimal and does not significantly impact content comprehension. In TableIII, inference time represents the average duration of inference, while ground truth denotes the average duration of the source speech. The results show that our method only slightly increases the required inference time, reducing the speed to approximately 0.76 times that of the original. Our proposed SC-S2UT method, while having a slightly lower Mean Opinion Score (MOS) compared to the cascade systems, is comparable to the baseline S2UT, achieving near state-of-the-art speech quality without significantly increasing inference time, thereby maintaining efficiency. Overall, our method, with the goal of preserving speaker information, achieved results very close to the baseline S2UT system.

B. Speaker Similarity

For speaker similarity evaluation, we compared the similarity between the speech obtained from the translation of source speech with the source speech and target speech, as shown in TABLE IV.

Based on the experimental results, we can find that, first, it is expected that the "*tal-tgt*" scenario shows the highest similarity, reflected by the COS-SIM score of 0.84 on the test set, and the lowest NMSE (0.051) and SED (0.51) values. These results indicate that our model effectively captures the characteristics of the target speech, as it was trained using target speech data. More importantly, the comparison between the translation and source speech demonstrates higher similarity and better performance in terms of NMSE and SED compared to the direct comparison between source and target speeches. Specifically, the COS-SIM score for "*tal-src*" (0.79) is higher than that for "*src-tgt*" (0.71). Additionally, the NMSE and SED values for "*tal-src*" are lower than those for "*src-tgt*" suggesting that our translation process preserves the speaker-specific information from the source speech more effectively than the direct reference between source and target. These findings suggest that our model excels in maintaining speaker similarity and preserving crucial speaker-specific attributes from the source speech.

TABLE IV
SPEAKER SIMILARITY EXPERIMENT RESULTS

	NMSE↓		SED↓		COS-SIM↑	
Compared Item	dev	test	dev	test	dev	test
tal-src	**0.070**	0.071	**0.60**	0.60	0.79	0.79
tal-tgt	0.075	**0.051**	0.64	**0.51**	**0.84**	**0.82**
src-tgt	0.089	0.072	0.71	0.62	0.77	0.71

"*tal-src*" refers to the comparison between translation and source speech, "*tal-tgt*" denotes the comparison between translation and target speech, and "*src-tgt*" represents the comparison between source speech and target speech. (↑) indicates that higher values are better, while (↓) mens that lower values are better.

VI. CONCLUSION

In this paper, we proposed a SR-UTM architecture designed to preserve the voice characteristics of the source speaker during speech-to-speech translation with low inference latency. We conducted experiments on the CVSS-C and CVSS-T corpora. The results demonstrated that our SC-S2UT system achieved translation quality comparable to the baseline S2UT system. Moreover, the speaker similarity experiments confirmed that our method effectively retains the speaker-specific information, without significantly increasing inference time. This highlights the robustness and efficacy of our approach in maintaining speaker identity during translation.

REFERENCES

[1] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744*, 2016.

[2] Y. Jia, R. J. Weiss, F. Biadsy, *et al.*, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint arXiv:1904.06037*, 2019.

[3] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: Robust direct speech-to-speech translation," 2021.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[5] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[6] A. Lee, P.-J. Chen, C. Wang, *et al.*, "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021.

[7] A. Lee, H. Gong, P.-A. Duquenne, *et al.*, "Textless speech-to-speech translation on real data," *arXiv preprint arXiv:2112.08352*, 2021.

[8] K. Song, Y. Ren, Y. Lei, *et al.*, "Styles2st: Zero-shot style transfer for direct speech-to-speech translation," *arXiv preprint arXiv:2305.17732*, 2023.

[9] Y. Wang, J. Bai, R. Huang, R. Li, Z. Hong, and Z. Zhao, "Speech-to-speech translation with discrete-unit-based style transfer," *arXiv preprint arXiv:2309.07566*, 2023.

[10] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[11] C. Zhang, X. Tan, Y. Ren, T. Qin, K. Zhang, and T.-Y. Liu, "Uwspeech: Speech to speech translation for unwritten languages," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 14 319–14 327.

[12] P.-J. Chen, K. Tran, Y. Yang, *et al.*, "Speech-to-speech translation for a real-world unwritten language," *arXiv preprint arXiv:2211.06474*, 2022.

[13] S. Popuri, P.-J. Chen, C. Wang, *et al.*, "Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation," *arXiv preprint arXiv:2204.02967*, 2022.

[14] Y. Jia, Y. Zhang, R. Weiss, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[15] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, 2020, pp. 3830–3834.

[16] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[17] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[18] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.

[19] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Systems with Applications*, vol. 90, pp. 250–271, 2017.

[21] Y. Jia, M. T. Ramanovich, Q. Wang, and H. Zen, "Cvss corpus and massively multilingual speech-to-speech translation," *arXiv preprint arXiv:2201.03713*, 2022.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[23] C. Wang, M. Riviere, A. Lee, *et al.*, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[24] T. Parcollet, H. Nguyen, S. Evain, *et al.*, "Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech," *Computer Speech & Language*, vol. 86, p. 101 622, 2024.

[25] H. Zen, V. Dang, R. Clark, *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[26] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[27] V. Pratap, A. Tjandra, B. Shi, *et al.*, "Scaling speech technology to 1,000+ languages," *arXiv*, 2023.

[28] C. Wang, Y. Tang, X. Ma, *et al.*, "Fairseq s2t: Fast speech-to-text modeling with fairseq," *arXiv preprint arXiv:2010.05171*, 2020.

[29] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.

[30] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, *SpeechBrain: A general-purpose speech toolkit*, arXiv:2106.04624, 2021. arXiv: 2106 . 04624 [eess.AS].

Keyword spotting for dialectal speech and Introduction of wav2vec2.0

Tomohiro Ariga*, Reo Minakawa†, Yoshiaki Itoh*, Kazunori Kojima*, Shi-wook Lee‡

* Iwate Prefectural University, Japan, E-mail: y-itoh@iwate-pu.ac.jp

† Graduated from Iwate Prefectural University

‡ National Institute of Advanced Industrial Science and Technology, Japan, E-mail: s.lee@aist.go.jp

Abstract— **Northern towns of the northeastern Tohoku region of Japan have folktales unique to each region. They are told by storytellers in local dialects which make them difficult for speakers of standard Japanese to understand. We have proposed a system to detect keywords when they are spoken, and to display an explanation so that anyone can understand the tale contents. Keyword spotting (KWS) can be used with spoken keywords, but it requires the creation of dialectal spoken keyword. Therefore, we propose a method by which a keyword is given as monophone or as a syllable series of text. Then a frame-level subword series is constructed and matched with input dialectal speech. Additionally, we introduce wav2vec2.0 to obtain the posterior probability vectors of monophone and syllable to improve the detection accuracy. First, we compare the proposed method with the Query-by-Example Spoken Term Detection (QbE-STD) method using spoken queries. Next, we compare the proposed method with our conventional method for text queries, which reveals that the proposed method achieves a 4.37 percentage point (%pt) improvement in detection accuracy, and a further 2.64%pt improvement using frame reduction. Furthermore, the introduction of wav2vec2.0 improved the detection accuracy by 5.14%pt.**

Index Terms: query-by-example, spoken term detection, blank-comp, Keyword Spotting, dialectal speech

I. INTRODUCTION

In the towns of Japan's northern Tohoku region, folktales unique to each region are passed down to younger generations by storytellers and performed in front of audiences. The stories are told with peculiar accents and dialects that are charming, but they are difficult for speakers of standard Japanese to understand. In fact, the dialects of Japan's northern Tohoku region have phonemes and word systems that are completely different from standard Japanese, rendering them almost incomprehensible to speakers other than those from the northern Tohoku region. Therefore, we have proposed a system to detect keywords that are important for understanding a story and for displaying them in real time so that anyone can understand the story contents. Using a speech recognition system is impossible because an automatic speech recognizer (ASR) for standard speech cannot be used for dialectal speech. Moreover, no transcription data exist for dialectal speech. It is also difficult to train and perform fine-tuning for deep neural networks and similar methods. We regard the problem as similar to the processing of a zero-resource language [1]. A search method that compares spoken keywords with speech data directly, called Query by Example-Spoken Term Detection (QbE-STD), can be applied without using ASR, making it independent of the availability of speech and linguistic resources for the target language [2]. The method is therefore useful for KWS of dialectal speech. In fact, KWS using spoken keywords requires the creation of spoken keywords, which are difficult to create for zero-resource speech and dialectal speech. We therefore construct a monophone or syllable series of the assumed keywords in text-based on keywords heard by speakers of standard Japanese. As described herein, we propose a frame-level matching method between text keyword and posteriorgram of speech data. Because the text keyword is at the subword level such as a syllable, monophone, and because the posteriorgram is at the frame level, the text keyword is stretched multiple times to correspond to the frame level. The frame-level keywords are matched with the posteriorgram of the speech data by continuous application of Dynamic Time Warping (DTW). The posteriorgram is transformed to local distances in advance for DTW matching. In KWS for speech commands, a sufficient amount of training is collected for classifying classes, regarding the keyword itself as a class; the deep neural model is trained as a classification task [3]. However, learning and relearning dialectal speech is difficult because there is no keyword speech for learning. Therefore, we use subwords of standard Japanese and perform fine-tuning of them using a standard Japanese corpus.

This study uses the Self-Supervised Learning (SSL) model, wav2vec2.0. Because it has been trained in English and many other languages in the pre-training process SSL model, wav2vec2.0 is assumed to have learned acoustic and linguistic features which are not learned only in standard Japanese speech. The dialectal speech used for this study is similar to zero-resource speech because, unlike the research on low-resource languages, it uses no dialectal speech for fine-tuning or re-training of wav2vec2.0. Some studies of zero-resource speech

have used intermediate features of models based on the bottleneck feature (BNF), clustering [3][4], and others. The method proposed herein extracts posterior probability vectors of monophones and syllables as intermediate features, and then uses them as a posteriorgram.

We first compare KWS methods using spoken keywords and our proposed method using text keywords. Then the effectiveness of the proposed method and introduction of wav2vec2.0 are verified using evaluation experiments.

II. Dialect Speech Comprehension Support Systems and Approaches

An overview of the dialectal speech understanding support system is given in II.A. Subwords such as monophones and syllables used for the system are explained in II.B Posteriorgram matching, a standard matching method for spoken keyword and interstate matching method [5], matches a given text keyword using a sequence of monophones or other subwords. They are explained respectively in II.C and II.D. Lastly, the frame reduction method of a posteriorgram used for this study is explained in II.E.

A. Dialect Speech Comprehension Support System

Fig. 1 presents an image of a system that supports understanding of a storyteller's narration speech. Keywords have been set in advance. While the storyteller is telling a folktale, a monitor displays an explanation of the keywords at the times when the keywords are spoken. This system helps the listener to understand the folktale without interrupting the storyteller's narration.

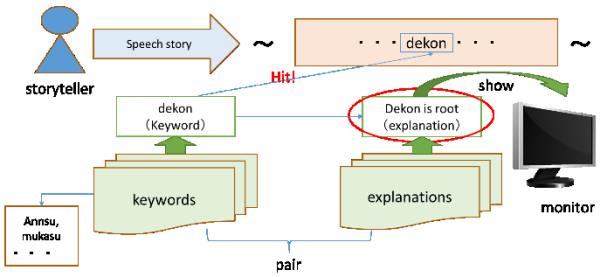


Figure. 1. The storyteller speech comprehension support system.

B. Subword

A subword is a unit smaller than a word: it is a character, a syllable, a monophone, or other utterance. Subwords are subdivided as shown in Table 1. For this study, monophones and syllables are used as subwords.

Table1. Subword Examples

word	Tokyo				
character	T	o	k	y	o
syllable	To		Kyo		
monophone	T	o	ky		o
triphone	#-t-o	t-o+ky	o-ky+o	ky-o+#	

C. Posteriorgram matching for QbE[6]

A posterior probability vector of triphone states, monophones, and others is obtainable by inputting the features of each frame of speech data into a deep learning model. The matrix of these posterior probability vectors in time order is called a posteriorgram. The posteriorgram of a query is searched among the posteriorgrams of the speech data by continual application of DTW method. We use as Continuous Dynamic Programming (CDP) for this purpose. The local distance used in DTW is obtained by finding the inner product of the posterior probability vectors of both the spoken query and the speech data and taking its negative logarithm.

D. Interstate matching method[5]

Speech data are recognized in subword series. Then the results are converted into a triphone state series beforehand. A text keyword is converted to a subword series. Then, after it is converted to a state series, CDP matching is performed using the acoustic distance between states [7] as the local distance.

E. Frame reduction method for posteriorgram

The same subword number sometimes continues several times in posteriorgrams using Connectionist Temporal Classification (CTC). Such duplicate information is often unnecessary. Therefore, we have proposed methods such as blank cut, blank compression, and Frame De-Duplication (FDD) to reduce them [8]. Because we show the results of blank compression herein, we explain the blank compression method.

In the blank compression method, the posterior probability values of all consecutive blank frames are averaged and compressed into a single frame in the case of output of a sequence of blank frames. This method is applied to all frames of speech data. The resulting posteriorgram is reconstructed. The frame number is reduced by the blank compression method.

III. Proposed Method

This section describes our proposed frame-level sequence matching method. First, a text keyword is converted into a subword number series such as monopone. When inputting speech, a deep learning model is composed to obtain the posterior probability of subwords. The posteriogram is localized by referring to the posterior probability values which match the number sequence of extracted keywords. CDP matching is then applied to perform this local distance and DTW continuously. Unlike inter-state matching, acoustic distance is not used. Using a distance based on the posteriogram, matching can be performed without loss of speech features. As shown in Fig2, assuming a subword number series of a keyword to be 4, 2, 3, only rows 4, 2, and 3 of the posteriorgram are extracted and rearranged. A distance is obtained by taking the negative logarithm of the posterior probability value. For example, the posterior probability subword number of 4 and

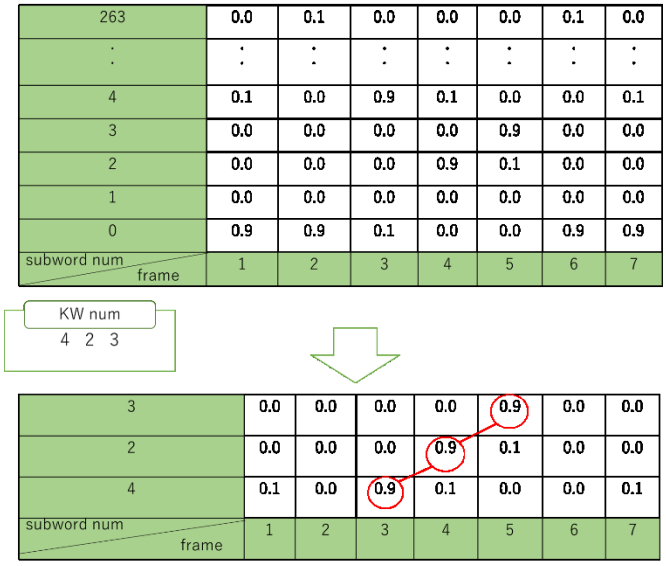


Figure 2: Schematic diagram of the proposed method

frame 3 is 0.9 and the local distance at that point is $-log_{10}0.9$ at the bottom figure in Fig. 2.

A keyword is transformed to a pseudo-frame-level subword series. We consider that the number of keyword subword series should be the same as the number of subword series in the speech segment where keywords are spoken. Therefore, each subword of a keyword is repeated to be the same as the number of frame level subword series in the speech data. This method improved the accuracy for standard Japanese test sets [10]. Figure 3 presents an example of a keyword using pseudo-frame-level subword series.

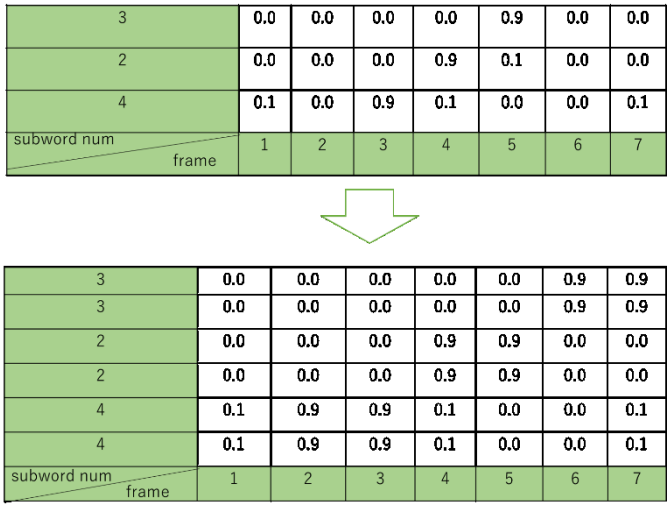


Figure 3. Example of keyword Stretching (keyword : 4 2 3)

Let the i-th subword of a keyword be S_i. The local distance $d(i, j)$ from the j frame of the speech data is calculated using the following equation (1). $(P_j(S_i))$ in equation (1) is the posterior probability corresponding to frame j and subword S_i in posteriorgram the speech data.

$$d(i, j) = -log_{10}(P_j(S_i)) \ . \qquad (1)$$

Consequently, frame level detailed matching can be performed using these local distances which use different posterior probability values for each frame.

IV. EVALUATION EXPERIMENT

A. *Experimental conditions*

For this experiment, we used three deep learning models: BLSTM; ESPnet [11], which has been used in our laboratory; and wav2vec2.0 [12], which is a self-supervised learning model with high recognition accuracy. Table 1 presents the fine-tuning conditions for BLSTM, ESPnet, and wav2vec2.0. We used approximately 600 hours of the Corpus of Spontaneous Japanese (CSJ) from 2702 lectures as training data for BLSTM and ESPnet, and as fine-tuning data for wav2vec2.0.

Table 2. learning conditions

model	BLSTM	ESPnet	wav2vec2
input	FBANK 120dim	FBANK 83 dim	wav data
train data	CSJ2702 speech (600hours)		
library	keras	Pytorch	
output	triphone state	syllable	syllable
outputs dim	3,009	264	263
epochs	30	26	50
window size	25ms		
frame shift	10ms		20ms

The base model, large model, and large-xlsr model described in [12] were used as pre-trained models for fine-tuning in wav2vec2.0. For the posteriorgrams used for matching, ESPnet and wav2vec2.0 used the output of CTC converted to probability values using the softmax function as the posterior probability vector.

B. *TEST DATA*

Seven dialectal tale speeches were used for evaluation. The keyword of each tale was found after confirming with the storyteller.

Table3. details of the tale

story		A	B	C	D	E	F	G
speech num		91	115	86	37	40	36	69
keyword	kind	11	19	10	9	9	10	11
	correct num	41	48	36	29	26	42	57
	speaker num	3 speakers						

Details of each tale and keyword are presented in Table 3. Mean Average Precision (MAP) was used as the evaluation index. The detection time is not considered in this experiment because can detect keywords in real time.

Table 4: Matching results for each model.

metric		MAP				
model		BLSTM	ESPnet	wav2vec2.0		
				base	large	large-xlsr
method	Poste	41.94	72.02	74.17	65.58	**81.55**
	Inter		77.24	**78.91**	73.67	77.61
	proposal	81.62	78.14	**83.28**	82.13	82.62

Acoustic features learned in the pre-training phase would be lost according to the progress of fine-tuning in wav2vec2.0. The detection accuracy is evaluated using a model with 1–50 epochs and the number of epochs was determined using cross-validation.

C. Results

Fig.4 shows the experimentally obtained results for the pre-trained model and the model with fine-tuning to demonstrate the effects of fine-tuning. The results of posteriorgram matching using fine-tuning and pre-trained models are also presented in Fig. 4. The output of the fine-tuning alone was unable to extract suitable features. It showed no high detection accuracy. We infer that fine-tuning for subwords is necessary to obtain high detection accuracy.

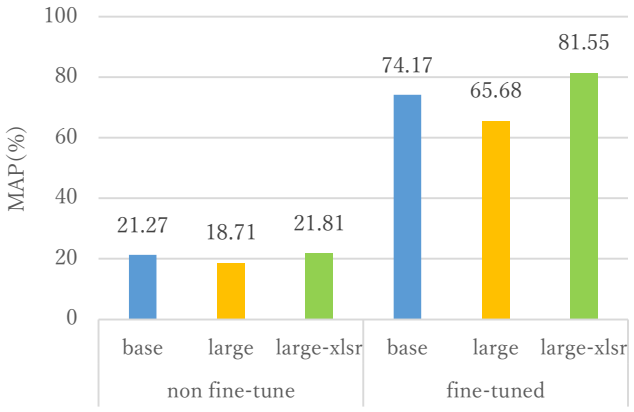


Figure 4: Detection accuracy of pre-trained and fine-tunig models

Next, Table 4 presents results of the comparison between spoken keywords and text keywords. The large-xlsr model of wav2vec2.0 showed the highest accuracy (81.55% in Table 4 Poste) for posteriorgram matching using spoken keywords. In the case of using spoken keywords, use of the multilingual large-xlsr model obtained various features which are thought to improve detection accuracy. The base model showed the highest detection accuracy using text keywords. Although the large-xlsr model showed high detection accuracy using spoken keywords, the proposed method using the base model showed the highest detection accuracy using text keywords. The reasons for this outcome were investigated. For posteriorgram matching, interstate matching, and proposed methods, wav2vec2.0 showed higher accuracy than that of either BLSTM or ESPnet. The results also indicate that our proposed method using text keywords is superior to the posteriorgram

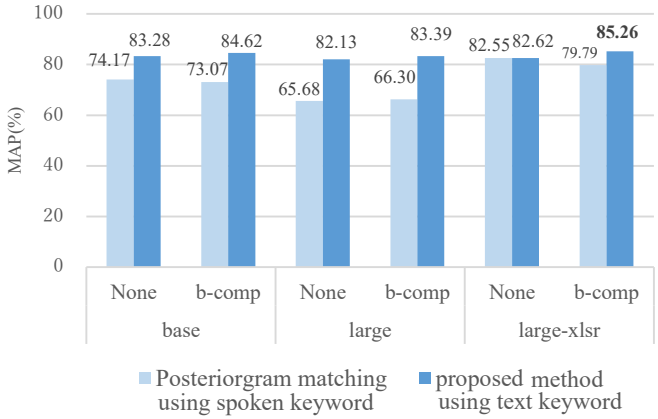


Figure5. Effects of Blank comp.

matching method using spoken keywords and interstate matching using text keywords. Therefore, the frame-level matching method using posteriorgrams and text keywords was more effective than using spoken keywords.

The results obtained using the frame reduction method “blank comp” are also shown in Fig. 5. Although “blank comp” does not always produce superior results for posteriorgram matching, it improved the detection accuracy of the proposed method and achieved 85.26% of the highest detection accuracy. The higher accuracy achieved when using the multilingual model suggests that multilingual speech data are effective for dialectal speech.

V. CONCLUSIONS

As described herein, we attempted to improve the detection accuracy of Japanese dialectal speech using a frame-level matching method on text keywords. The proposed method using text keywords achieved detection accuracy (MAP) that was more than 85% higher than that achieved using spoken keywords. Compared to the conventional method, the proposed method improved the detection accuracy. Moreover, the accuracy was improved further using the frame reduction. These results confirm the effectiveness of the proposed method for Japanese dialect speech. Compared to the conventional model, the proposed method with wav2vec2.0 improved the accuracy by approximately five percentage points. We plan to conduct experiments using other SSL models and low resource speech in future studies.

ACKNOWLEDGMENT

This work was supported by JSPS (C), KAKENHI Grant Number 21K12611.

REFERENCES

[1] J. Martin, et al. ‘‘Utterance Classification in Speech-to-speech Translation for Zero-resource Languages in the Hospital Administration Domain,’’ 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 303–309 (2015).

[2] S. Mizuochi, et al. “Spoken Term Detection of Zero-Resource Language Using Posteriorgram of Multiple Languages”, nterdisciplinary Information Sciences, Vol. 28, No. 1 (2022) 1–13

[3] D. Seo, et al. "Wav2KWS: Transfer Learning From Speech Representations for Keyword Spotting," in IEEE Access, vol. 9, pp. 80682-80691, 2021, doi: 10.1109/ACCESS.2021.3078715.

[4] Sung, Man-Ling and Tan Lee. “Unsupervised Spoken Term Discovery Based on Re-clustering of Hypothesized Speech Segments with Siamese and Triplet Networks.” ArXiv abs/2011.14062 (2020): n. pag.

[5] R. Konno, et al. “A Spoken term detection method matching at a frame level”, IEICE, Vol. 115, no. 146, SP2015-37, pp 7-12 (2015)

[6] M. Obara, et al. “Improvement of Search Accuracy of Spoken Term Detection Using Posteriorgram of Deep Neural Network”, Iwate Prefectural University（2016）

[7] K. Iwata, et al. “An Investigation of New Subword Models and Subword Phonetic Distance for Vocabulary-free Spoken Document Retrieval System”, JIP, vol.48, no.5, pp.1990-2000 (2007)

[8] H. Yamaga, et. Al.”Accurate and Practical Query-by-Example Using Multiple Deep Learning Models and Frame Compression Methods” APSIPA ASC 2023: 862-867

[9] A. Graves and N. Jaitly, “Towards End-to-End speech recognition with recurrent neural networks,” Proc. ICML (2014).

[10] R. Minakawa, et al. “Improving of Retrieval Accuracy for Spoken Term Detection in Frame Level Matching Method”, CiNii, Vol.2022-1R-07, No.2, pp.295-296, (2022)

[11] S. Watanabe, et. al , “ESPnet : End-to-end speech processing toolkit”, INTERSPEECH, 2018, pp. 2207–2211.

[12] A. Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” arXiv preprint arXiv:2006.11477, 2020.

LCMV-based Scan-and-Sum Beamforming for Region Source Extraction

Aoto Yasue*, Benjamin Yen*, Katsutoshi Itoyama† and Kazuhiro Nakadai*
* Tokyo Institute of Technology (currently, Institute of Science Tokyo)
E-mail: yasue@ra.sc.e.titech.ac.jp
† Honda Research Institute Japan Co., Ltd.

***Abstract*—This paper addresses region source extraction (RSE), which extracts the entire sound source originating from a specified target region. To solve this problem, an LCMV-based Scan-and-Sum Beamformer (LCMV-SS-BF) that can separate a wide range of continuous regions is proposed. This method is designed by incorporating the linearly constrained minimum variance Beamformer (LCMV-BF) as a filter for continuous regions into the Scan-and-Sum Beamformer (SS-BF) as a filter for wide-area sound source separation. In addition, an incremental estimation method using the Woodbury's formula for LCMV-BF filters is proposed to improve the processing speed of the filter computation of the LCMV-BF. Numerical simulation using convolved sound datasets showed 1) the effectiveness of the LCMV-SS-BF for surface source separation and 2) the computational cost advantage of the incremental estimation method.**

I. Introduction

Source enhancement techniques such as beamforming have been widely studied, and these signal models assume point sources [1], [2]. However, in reality, ideal point sources do not exist. Therefore, model errors are inherent in the extraction of sound sources that exist within a specific area, such as non-point sound sources like waterfalls, multiple instruments in an orchestra, and the utterances of a driver whose position is constantly changing, and the expected performance cannot be achieved. In addition, when a video camera or telepresence robot is used to track and direct attention to a specific target, both audio and visual information can be used, but in most cases, image and sound are used without linking them. For example, even if you zoom in on an object in terms of images, the process of zooming in is not performed in terms of audio, and only the background noise is processed. As a result, sounds outside the target region are also recorded. To solve this problem, it is necessary to link the image and audio and control the spatial recording range so that only acoustic signals from the same range as the camera's image range are recorded (called audio-visual zooming [3]). Therefore, a new technique is required to specify a target region and extract all acoustic signals coming from within the target region. In this paper, we call this technique "region source extraction" (RSE) and treat it as a method for isolating and extracting sound sources within a specified region.

Beamforming is a technique used to extract point source in a specific direction using a microphone array and utilizing the property that each microphone has a different time delay when the direction of arrival is different [2], [4]. Beamforming is a point source extraction method that basically specifies the target direction of the source to be extracted. However, RSE requires specifying a target region instead of a target direction and extracting all sound sources within the target region, which is beyond the capabilities of conventional beamforming.

In order to perform RSE, multi-modal methods using audiovisual information are proposed [5]–[9]. These methods use a camera and microphone to synchronously record images and sound. Then assuming a strong correlation between the recorded image and audio, the audio corresponding to the objects in the image is separated and extracted using deep learning. However, the system assumes that the target object is contained in the image, and cannot be used in situations where the object is hidden or far from the camera. In addition, a large data set is required to handle various situations.

On the other hand, as methods for RSE using only audio information, there are techniques such as acoustic zooming and a Scan-and-Sum beamforming (SS-BF).

Acoustic zooming detects all sound sources in the observed sound by sound source localization and extracts each detected source using conventional beamforming. RSE is then achieved by resynthesizing only the extracted sources within the target region [10]–[12]. This method requires the number and location of sound sources as a priori information, but it is difficult to provide this information in advance for a set of sound sources such as an orchestra or a non-point source. The same problem also occurs with blind source separation [13]–[15] approaches, which can separate multiple sound sources at once, to achieve RSE.

An SS-BF [16], [17] achieves RSE by the division of the target region (each divided small region is called a sub-region) and the sum of beamformers that extract sources within the sub-region. the SS-BF specifies a target region, divides it into sub-regions, then constructs sub-beamformers (sub-BF) for all sub-regions such that the sources in the center direction of the subregion are extracted. Finally, it forms a filter for RSE by the weighting sum of the sub-BFs. This method does not require the number or location of the sources, but the extraction performance is degraded for sources located off the center of the main lobe of each sub-BF. This is due to the fact that sub-BFs are computed for the extraction of discrete points in the target region, which means SS-BF is not a filter suitable for continuous region extraction.

As reflected in the existing studies, the problems of conven-

tional methods include 1) the requirement of image data, 2) the requirement of the number and location of sound sources, or 3) the inability to create filters for continuous regions. To solve these problems, we combine a linearly constrained minimum variance beamformer (LCMV-BF) with the SS-BF as an LCMV-based Scan-and-Sum Beamformer (LCMV-SS-BF) to achieve RSE without using image data, model training with data sets, or sound source localization as preprocessing.

The LCMV-BF constructs a filter by solving a minimization problem by placing no-distortion constraints on multiple directions [18], [19]. It is commonly used to separate multiple sources at once by simultaneously providing constraints on multiple source directions. On the other hand, it has been reported that by giving each constraint point in a neighborhood, stable extraction is possible for the continuous region between the constraint points. This property is generally used to deal with direction-of-arrival mismatch [20]. This property suggests that LCMV-BF could be used as a filter for RSE. However, since the constraints of the LCMV-BF must be placed in the neighborhood to acquire the property, it is not possible to perform wide region separation. Therefore, we propose an LCMV-SS-BF, which adopts the structure of the SS-BF to LCMV-BF. To cope with the huge computational time required for the SS-BF, we also propose a method to speed up the calculation of the LCMV-BF by using a recursive least square (RLS) method to estimate the filters incrementally.

We evaluate the proposed method by numerical simulation. For the evaluation, we created self-made datasets of mixed-source speech data by convolving transfer functions to each audio data from publicly available datasets and summing them over multiple audio data. Simulation results show that LCMV-SS-BF outperforms conventional methods for RSE, indicating the superiority of the proposed method. Furthermore, the incremental estimation method achieves faster computation.

The contributions presented in this paper are as follows:

- We propose an LCMV-SS-BF as a novel RSE method that can handle a wide range of continuous regions.
- We propose the use of an incremental RLS method for LCMV-BF estimation based on the Woodbury's formula and introduce it into LCMV-SS-BF to improve its processing speed.
- We show the effectiveness of the proposed LCMV-SS-BF, including its speed-up by the incremental estimation through simulations using the self-made sound datasets.

II. Preliminary

In this section, we formulate the problem setting of RSE and the LCMV beamformer on which the proposal is based.

A. problem setting of Region Source Extraction

Fig. 1 shows the coordinate system and the definition of target sources for RSE. Let the target region be a region cut off from the spherical coordinate system with azimuth θ and elevation ϕ, where $-\mathcal{R}/2 \leq \theta, \phi \leq \mathcal{R}/2$ (the orange region in Fig. 1), and the noise region be a region, where $-\mathcal{R}_{n_{\text{out}}}/2 \leq \theta, \phi \leq \mathcal{R}_{n_{\text{out}}}/2$ and $(-\mathcal{R}_{n_{\text{in}}}/2 \leq \theta, \phi \leq \mathcal{R}_{n_{\text{in}}}/2)^c$

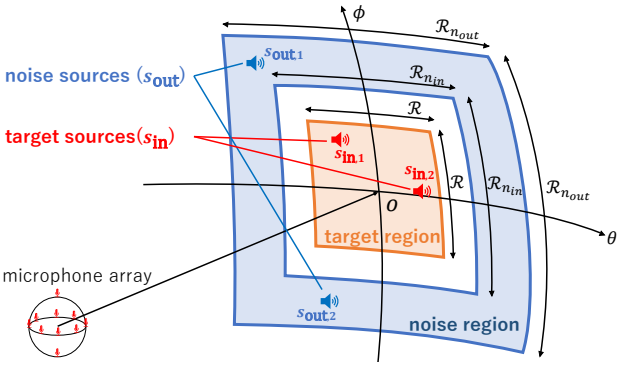


Fig. 1. A target surface region defined as a region that cuts off a portion of a spherical coordinate system.

(the blue region in Fig. 1). $(\cdot)^c$ means the residual phenomenon. These regions are parts of the spherical surface. Then target sources s_{in} are defined as the sources coming from within the target region and noise sources s_{out} are the sources coming from within the noise region (without the target region). When $z_{\omega,t,m}$ is the observed signal by the m-th microphone ($m = 1, \ldots, M$) at angular frequency ω, time frame t, the observed signal $\boldsymbol{z}_{\omega,t}$ is defined as,

$$\boldsymbol{z}_{\omega,t} = \sum_p \boldsymbol{a}_{\text{in},p,\omega,t} \cdot s_{\text{in},p,\omega,t} + \sum_q \boldsymbol{a}_{\text{out},q,\omega,t} \cdot s_{\text{out},q,\omega,t}, \quad (1)$$

$$\boldsymbol{z}_{\omega,t} = [z_{\omega,t,1}, \cdots, z_{\omega,t,M}]^T, \quad (2)$$

$$\boldsymbol{a}_{x,\omega,t} = [e^{-j\omega\tau_{1,x}}, \cdots, e^{-j\omega\tau_{M,x}}]^T, \quad (3)$$

where $\boldsymbol{a}_{x,\omega,t} \in \mathbb{C}^{M\times 1}$, $x \in [(\text{in}, p), (\text{out}, q)]$ denotes the transfer function corresponding to the sound source $s_{x,\omega,t}$, while $\cdot^T$ indicates the matrix transpose operator. $\tau_{m,x}$ indicates the time difference of arrival (TDOA) of the m-th microphone to the reference point in the x-th transfer function, $\boldsymbol{a}_{x,\omega,t}$. Now the objective of RSE can be written as extracting $\sum_p s_{\text{in},p,\omega,t}$.

B. LCMV Beamformer

When applying beamforming to the observed signal $\boldsymbol{z}_{\omega,t}$, the output signal $Y_{\omega,t}$ is defined as,

$$Y_{\omega,t} = \boldsymbol{w}_{\omega,t}^H \boldsymbol{z}_{\omega,t}, \quad (4)$$

$$\boldsymbol{w}_{\omega,t} = [w_{\omega,t,1}, \cdots, w_{\omega,t,M}]^T, \quad (5)$$

where $\boldsymbol{w}_{\omega,t} \in \mathbb{C}^{M\times 1}$ denotes the filter of a beamformer, while $\cdot^H$ indicates the matrix Hermitian transpose operator. ω and t will be omitted for simplicity thereafter.

The LCMV-BF is obtained as the solution to a constrained optimization problem that minimizes the output of the beamformer with all-pass characteristics for multiple directions as,

$$\tilde{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\text{argmin}}\, \boldsymbol{w}^H \boldsymbol{R} \boldsymbol{w} \quad (6)$$

$$\text{subject to } \boldsymbol{A}^H \boldsymbol{w} = \boldsymbol{f},$$

$$\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_k, \cdots, \boldsymbol{a}_K],$$

$$\boldsymbol{a}_k = [e^{-j\omega\tau_{1,k}}, \cdots, e^{-j\omega\tau_{M,k}}]^T,$$

$$\boldsymbol{f} = [f_1, \cdots, f_k, \cdots, f_K]^T,$$

where $\boldsymbol{R} = E[\boldsymbol{z}\boldsymbol{z}^H] \in \mathbb{C}^{M\times M}$ represents the spatial correlation matrix of the observed signals. $E[\boldsymbol{z}\boldsymbol{z}^H]$ denotes the

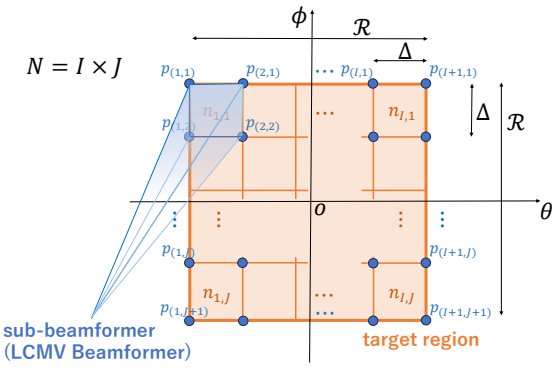


Fig. 2. LCMV-based Scan-and-Sum Beamformer (LCMV-SS-BF)

expected value of $\boldsymbol{z}\boldsymbol{z}^H$. $\boldsymbol{A} \in \mathbb{C}^{M\times K}$ is a matrix consisting of K transfer functions for constraint directions. f_k denotes the all-pass characteristics in the k-th direction.

By solving Eq. (6), the LCMV-BF is formulated as,

$$\tilde{\boldsymbol{w}} = \boldsymbol{R}^{-1}\boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\boldsymbol{f}. \tag{7}$$

III. Proposed Method

This section proposes and formulates the LCMV-SS-BF and the incremental RLS method for LCMV-SS-BF estimation.

A. LCMV-based Scan-and-Sum Beamformer

Fig. 2 shows the abstract of the proposed LCMV-SS-BF. The target region shown in Fig. 1 is considered as an $\mathcal{R}\times\mathcal{R}$ square area as shown in Fig. 2. To construct the LCMV-SS-BF, the target region is divided by small $\Delta\times\Delta$ square regions (sub-regions) in both θ and ϕ direction to obtain N sub-regions. The sub-region (i,j) is indicated as $n_{i,j}$, and the grid point (i,j) is indicated as $p_{(i,j)}$. After that, a filter for each sub-region is computed as a sub-BF, and all sub-BFs are summed up to form the SS-BF. For the sub-BF, the LCMV-BF is selected and it is designed to separate the entire sub-region with minimal distortion by constraining the all-pass characteristics at the four edges of each sub-region. Since the target region is divided into small sub-regions, the four edges are placed in a neighborhood, which maintains the robustness to the continuous direction between each point. Thus, the filter of the proposed LCMV-SS-BF, $\boldsymbol{w}_{\text{proposed}}$ is formulated as,

$$\boldsymbol{w}_{\text{proposed}} = \sum_{i=1}^{I}\sum_{j=1}^{J} b_{i,j}\cdot\boldsymbol{R}^{-1}\boldsymbol{A}_{i,j}(\boldsymbol{A}_{i,j}^H\boldsymbol{R}^{-1}\boldsymbol{A}_{i,j})^{-1}\boldsymbol{1}, \tag{8}$$

$$\boldsymbol{A}_{i,j} = [\boldsymbol{a}_{p_{(i,j)}}, \boldsymbol{a}_{p_{(i+1,j)}}, \boldsymbol{a}_{p_{(i,j+1)}}, \boldsymbol{a}_{p_{(i+1,j+1)}}], \tag{9}$$

where $\boldsymbol{A}_{i,j} \in \mathbb{C}^{M\times 4}$ includes the four corner transfer functions of $n_{i,j}$ (see Fig. 2) and $\boldsymbol{B} = [b_{1,1}, \cdots, b_{i,j}, \cdots, b_{I,J}]$ is an weight vector for each sub-region.

The LCMV-SS-BF compensates for the disadvantages of both the LCMV-BF and the SS-BF by combining them: by using the structure of the SS-BF, the constraint points of the LCMV-BFs can be placed in the neighborhood, allowing a wide range RSE. On the other hand, by using the LCMV-BF as a sub-BF, the SS-BF can construct a filter for continuous region extraction instead of the sum of filters for discrete points. So the LCMV-SS-BF can extract sound sources from a wide range and continuous region without using visual information and knowledge regarding the number and directions of sound sources included in the region.

B. Incremental RLS method for LCMV-BF estimation

The SS-BF requires repeated calculations of sub-BFs to estimate the filter, which increases the computation time. To solve this problem, we propose a method to reduce the computation time of the LCMV-BF, which is an adaptive filter and requires computationally expensive inverse matrices of $\boldsymbol{R}$ and $\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A}$. To avoid such inverse matrix computation, we propose an incremental estimation method based on recursive least square (RLS) method [21].

While the inverse of $\boldsymbol{R}_{\omega,t}$ needs to be computed for each time frame t and angular frequency ω, the RLS method can compute it incrementally using the result obtained at the previous time frame. First, $\boldsymbol{R}_{\omega,t}$ is defined in a recursive manner by omitting ω again for simplicity as,

$$\boldsymbol{R}_t = \alpha\boldsymbol{R}_{t-1} + (1-\alpha)\,\boldsymbol{z}_t\boldsymbol{z}_t^H, \tag{10}$$

where α is the forgetting coefficient. Using the Woodbury's formula [22], $\boldsymbol{R}_t^{-1}$ can be formulated as follows:

$$\boldsymbol{R}_t^{-1} = \frac{1}{\alpha}\left(\boldsymbol{R}_{t-1}^{-1} - \frac{1}{\gamma_t}\boldsymbol{R}_{t-1}^{-1}\boldsymbol{z}_t\boldsymbol{z}_t^H\boldsymbol{R}_{t-1}^{-1}\right) \tag{11}$$

$$\gamma_t = \frac{\alpha}{1-\alpha} + \boldsymbol{z}_t^H\boldsymbol{R}_{t-1}^{-1}\boldsymbol{z}_t. \tag{12}$$

Likewise, the incremental expression of $\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A}$ in Eq. (7) is formulated by using Eq. (11) as,

$$\boldsymbol{A}^H\boldsymbol{R}_t^{-1}\boldsymbol{A} = \frac{1}{\alpha}\left(\boldsymbol{A}^H\boldsymbol{R}_{t-1}^{-1}\boldsymbol{A} - \frac{1}{\gamma_t}\boldsymbol{l}_t\boldsymbol{l}_t^H\right) \tag{13}$$

$$\boldsymbol{l}_t = \boldsymbol{A}^H\boldsymbol{R}_{t-1}^{-1}\boldsymbol{z}_t. \tag{14}$$

Using the Woodbury's formula, $\boldsymbol{D}^{-1} = (\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}$ can be expressed as

$$\boldsymbol{D}_t^{-1} = \alpha\left(\boldsymbol{D}_{t-1}^{-1} + \frac{1}{\delta_t}\boldsymbol{D}_{t-1}^{-1}\boldsymbol{l}_t\boldsymbol{l}_t^H\boldsymbol{D}_{t-1}^{-1}\right) \tag{15}$$

$$\delta_t = \gamma_t - \boldsymbol{l}_t^H\boldsymbol{D}_{t-1}^{-1}\boldsymbol{l}_t. \tag{16}$$

By using Eqs. (11) and (15), the calculation of the LCMV-BF can be obtained as an incremental estimation without using inverse matrix calculations.

IV. Evaluation

The proposed methods are evaluated by simulations with convolved sound datasets with publically available audio datasets, in terms of extraction performance improvement and computational speed.

A. Experimental setting

As a recording device, a 10 ch spherical microphone array of radius 3 cm was selected, which has eight microphones at the equator and another two at the poles, as shown in Figure 1. The acoustic transfer functions were calculated geometrically assuming a free acoustic field and plane wave model. For the simulations, we prepared four datasets, Datasets 1 to 4. The

comparison of each dataset is shown in TABLE I. Datasets 1 and 2 were used to evaluate the extraction performance when different types of sound sources were mixed together as the target and noise sources. As the target sources, we used the train-clean-100 dataset from LibriSpeech [23], which is the dataset of human speech sounds. As the noise sources, we used the background noise sources from the CHiME3 dataset [24], which is the dataset of background noise sounds. Datasets 3 and 4, on the other wards, were used to evaluate the extraction performance when the same type of sound sources were mixed together, and the dev-clean dataset from LibriSpeech was used as both target and noise source. Datasets 1 and 3 were used to evaluate the extraction performance in the case of the mixture of one target source and one noise source, while Datasets 2 and 4 were used to evaluate the extraction performance in the case of the mixture of two target sources and two noise sources, which means the multiple sources extraction. An appropriate acoustic transfer function was applied to each source so that target sources were randomly placed in the target region and noise sources were randomly placed in the noise region. And by mixing them, we created 100 6-second-long synthetic sounds for each Dataset, each $\mathcal{R}$ size to be evaluated. The setting of $\mathcal{R}$ were conducted on 20°, 40°, and 60°. $\mathcal{R}_{n_{\text{in}}}$, $\mathcal{R}_{n_{\text{out}}}$ were set to $\mathcal{R} + 20°, \mathcal{R} + 100°$.

TABLE I
SETTING OF EACH DATASET

Dataset	the number of		the kind of	
	Target Sources	Noise Sources	Target Sources	Noise Sources
Dataset 1	1	1	Speech	Background
Dataset 2	2	2	Speech	Background
Dataset 3	1	1	Speech	Speech
Dataset 4	2	2	Speech	Speech

Six separation methods were compared, including the proposed LCMV-SS-BF (LCMV-SS), the Minimum Variance Distortionless Response beamformer (MVDR-BF, MVDR), the LCMV-BF (LCMV), the SS-BF (SS), Independent Low-Rank Matrix Analysis (ILRMA) [13], [14], and Neural Network based Generalized Eigenvalue Beamformer (NN-GEV) [1]. The first four methods, including the proposed method, are beamformer-based approaches. MVDR-BF is the generally used beamforming method for point source extraction. ILRMA is a kind of blind source separation method, which separates the mixture of sound without sound source or system information. It specifies the number of sound sources in the sound mixture and separates the sound mixture into the specified number of sound sources based on the sound source model. When used as RSE, the number of sound sources in the sound mixture must be known, and it is further necessary to determine whether the separated sound sources are contained within the target region. NN-GEV is a mask-based method with a neural network. It generates masks for the target and noise sources using neural network and calculates the GEV beamformer [25], [26] using the spatial correlation matrix computed from them.

For evaluating the computation time of LCMV-BF, three methods, that is, i) the proposed method (RLS2) that applies the RLS method to $\boldsymbol{R}^{-1}$ and $(\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}$, ii) the Block-wise method (Block) that calculates the inverse matrix, and iii) RLS1 that applies the RLS method only to $\boldsymbol{R}^{-1}$ were compared in terms of the processing speed for computing one LCMV sub-BF.

The signal-to-distortion ratio (SDR) [27] and the short-time objective intelligibility measure (STOI) [28] were used as the objective and subjective metric of sound source extraction performance, respectively. The values were calculated in the time domain with $\sum_p s_{\text{in},p,\omega,t}$ as the reference signal and the output of each method as the input, respectively. The real-time factor (RTF) was used to evaluate the performance of the processing speed, and it was measured for computing a single LCMV sub-BF. The measurements were performed using MATLAB on a PC equipped with an Intel(R) Core(TM) i7-1185G7 @ 3.00 GHz.

B. Parameter settings of each method

For LCMV-SS-BF, Δ, was set to $10°$, and a uniform weight vector was used for $\boldsymbol{B}$. For MVDR-BF, the target direction was set to the central direction of the target region, that is, $(\theta, \phi) = (0°, 0°)$, because only a single direction can be specified in the target region. For LCMV-BF, the constraint points are given at the four edges of the target region. For SS-BF, Δ was set to $4°$ and MVDR-BF oriented toward the center of each sub-region was used as the sub-BF. For ILRMA, the number of sound sources must be given in advance. Therefore, sound source separation was first performed with the number of sound sources set equal to or lower than the actual number of sound sources. Although it is also necessary to determine whether the sound sources after separation are included in the target region, this time we ignored this. And selected the one with the best separation performance among the combinations of the sums of the separated sound sources as the result. For NN-GEV, the neural network was trained to estimate a mask that separates speech from speech mixed with environmental noise with CHiME3 dataset. This condition is favorable for Datasets 1 and 2, while unfavorable for Datasets 3 and 4. This method is used for comparison but is not an method for RSE because it depends on the types of sound sources.

For the calculation methods, RLS1 was initialized with $\boldsymbol{R}_0^{-1} = \boldsymbol{I} \times 10^5$ and $\alpha = 0.995$. In RLS2, the first 100 frames are calculated with RLS1, and then $\boldsymbol{R}^{-1}$ and $(\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}$ are initialized using the values at that time, and start estimation.

C. Results and discussions

TABLE II and III show the average SDR and STOI for each method, respectively. TABLE IV shows the average RTF per LCMV for each calculation method. The results of the method with the best average performance in the same dataset are shown in bold. In addition, P-tests were performed between the method that produced the best average performance (the method shown in bold) and the other methods to verify the difference in superiority between methods. This was done

TABLE II
COMPARISON IN SDR [DB]

Dataset	$\mathcal{R}$	Method					
		LCMV-SS (proposed)	MVDR	LCMV	SS	ILRMA	NN-GEV
Dataset 1	20°	10.51 ± 3.91	2.73 ± 3.98	8.44 ± 3.91	2.94 ± 2.40	**12.45 ± 7.03**	9.01 ± 3.13
	40°	9.99 ± 3.37	−0.81 ± 4.52	4.62 ± 4.54	0.41 ± 2.93	**13.74 ± 6.53**	9.05 ± 3.15
	60°	8.96 ± 2.96	−1.81 ± 4.49	0.55 ± 4.19	−1.15 ± 3.63	**13.60 ± 7.15**	9.23 ± 3.20
Dataset 2	20°	**9.85 ± 3.69**	3.22 ± 3.02	8.03 ± 3.68	4.32 ± 2.06	−0.74 ± 4.61	6.24 ± 3.38
	40°	**10.04 ± 2.86**	0.44 ± 2.84	4.98 ± 3.77	2.36 ± 2.36	−0.75 ± 4.54	5.27 ± 3.42
	60°	**8.60 ± 2.39**	−0.57 ± 4.00	0.97 ± 3.90	1.09 ± 2.84	−0.80 ± 4.55	3.98 ± 3.47
Dataset 3	20°	**13.32 ± 3.84**	4.45 ± 3.76	10.31 ± 4.46	4.60 ± 1.69	•10.80 ± 10.61	−3.48 ± 6.78
	40°	•12.66 ± 2.86	0.92 ± 4.19	7.37 ± 4.34	1.57 ± 1.93	**13.30 ± 9.95**	−2.68 ± 6.76
	60°	•10.33 ± 3.37	−2.14 ± 4.14	3.54 ± 4.72	−0.83 ± 2.99	**11.42 ± 10.42**	−3.09 ± 6.77
Dataset 4	20°	**10.40 ± 3.69**	4.76 ± 3.39	8.70 ± 3.38	6.31 ± 2.41	2.44 ± 3.70	−5.34 ± 5.31
	40°	**10.67 ± 3.32**	1.54 ± 3.23	6.27 ± 3.42	4.59 ± 2.33	2.40 ± 4.18	−4.55 ± 4.84
	60°	**9.46 ± 3.06**	−0.16 ± 3.96	2.88 ± 3.34	3.21 ± 2.54	2.24 ± 4.13	−6.30 ± 5.88

TABLE III
COMPARISON IN STOI [%]

Dataset	$\mathcal{R}$	Method					
		LCMV-SS (proposed)	MVDR	LCMV	SS	ILRMA	NN-GEV
Dataset 1	20°	92.77 ± 7.98	87.30 ± 8.86	89.41 ± 9.08	89.91 ± 7.84	•93.23 ± 9.01	**94.11 ± 7.11**
	40°	•92.74 ± 6.90	82.54 ± 9.31	82.63 ± 10.96	88.38 ± 7.42	**95.06 ± 6.99**	•94.75 ± 6.58
	60°	91.57 ± 7.11	79.22 ± 9.94	73.36 ± 12.36	86.40 ± 8.13	•93.52 ± 9.63	**95.13 ± 5.72**
Dataset 2	20°	**95.14 ± 3.38**	82.67 ± 7.08	91.08 ± 5.29	90.47 ± 4.46	69.63 ± 12.35	•94.68 ± 4.20
	40°	**94.57 ± 2.93**	74.11 ± 7.87	83.71 ± 9.22	85.98 ± 6.29	68.60 ± 12.46	•94.19 ± 3.80
	60°	**92.45 ± 3.55**	70.90 ± 10.36	71.78 ± 12.77	84.36 ± 7.66	68.95 ± 12.00	•92.01 ± 5.16
Dataset 3	20°	**95.02 ± 4.57**	90.62 ± 4.86	91.92 ± 6.13	92.84 ± 3.48	86.18 ± 13.21	62.42 ± 19.39
	40°	**95.10 ± 2.88**	84.63 ± 8.14	88.32 ± 7.14	89.89 ± 4.61	89.77 ± 12.53	63.62 ± 19.37
	60°	**93.46 ± 3.90**	80.75 ± 8.08	81.40 ± 10.37	88.33 ± 5.20	86.98 ± 13.26	61.11 ± 19.76
Dataset 4	20°	**91.80 ± 5.79**	83.10 ± 8.90	89.64 ± 6.00	•90.25 ± 5.02	73.84 ± 10.80	55.22 ± 16.90
	40°	**91.91 ± 5.29**	73.60 ± 10.26	84.95 ± 7.50	85.90 ± 5.87	73.33 ± 11.11	54.77 ± 16.37
	60°	**90.92 ± 5.01**	69.06 ± 11.44	76.50 ± 9.48	82.37 ± 7.76	71.79 ± 12.45	50.10 ± 18.33

for each dataset and each $\mathcal{R}$. If the P value is greater than 0.01, there is no superiority difference. Results that were determined to have no superiority difference (equivalent to the best-performing method) are indicated by a black circle (•) in front of the number.

From TABLE II and III, In Dataset 2, Dataset 3, and Dataset 4, the proposed method is found to have the best performance with a superiority difference (although it was determined to be equivalent to ILRMA for SDR in Dataset 3). The result of the proposed method on Dataset 1 also showed performance that was second to the best method, although a difference in superiority occurred between it and the best method. This means that the proposed method has the best RSE performance among these methods. Dataset 1 is the most typical situation of one speaker and one background noise, where enhancement can be achieved with various source enhancement techniques that are not specific to RSE. In particular, IRLMA omits sound source localization, so experiments are conducted under favorable conditions.

In addition, what is noteworthy about the proposed method is its robustness to various situations. Observing the performance over varying $\mathcal{R}$, it is shown that the performance degradation of the proposed method from $\mathcal{R} = 20°$ to $\mathcal{R} = 60°$ is limited to 1.80 dB in SDR and 1.58 % in STOI. This is the smallest degradation among the beamforming-based methods. This indicates that the problem that RSE over a wide region was difficult with a single LCMV-BF has been greatly relaxed. For the variation in the number of sound sources in the target region, cases where one target source (Datasets 1 and 3) and two target sources (Datasets 2 and 4) are compared. The proposed method maintains consistent performance between Datasets 1, 3 and 2, 4, with an SDR difference of 1.12 dB and STOI difference of 0.65 %, regardless of the number of the target sources. This can be said to be a property of beamforming-based methods that are based on spatial models. On the other hand, IRLMA significantly deteriorated as the number of sources increased. This is because as the number of sound sources increases, more complex models for generating sound sources become necessary. To observe the performance over types of sound sources mixed, source separation of the mixture of different types (Datasets 1 and 2) and that of the same type (Datasets 3 and 4) are compared. The results showed that the performance of the proposed method was independent of the source type, with an SDR difference of 1.48 dB and STOI difference of 0.17 % between Datasets 1, 2 and 3, 4, while NN-GEV showed significant performance degradation under unfavorable conditions. These results suggest the robustness of the proposed method against the changes in the region size, number of sources, and types of sources. The proposed method can be said to be the most suitable method for RSE.

From TABLE IV, the proposed method showed the best RTF and achieved real-time processing. The proposed method was 2.85× and 1.18× faster than Block and RLS1, respectively. This demonstrates the effectiveness of the proposed method

TABLE IV
COMPARISON IN RTF

Method	Block	RLS1	RLS2
RTF	1.77	0.73	**0.62**

also in computational time.

V. CONCLUSION

This paper presented surface source separation for a wide range of continuous regions at high speed and proposed LCMV-SS-BF by integrating LCMV-BF and SS-BF with an incremental RLS method for LCMV-BF estimation. Simulation results showed that the proposed method outperformed other methods such as beamforming, blind separation, and mask-based separation using a neural network, and the robustness for the changes in region size, number of sources, and types of sources. In terms of processing speed, it worked in real time and was 2.85 times faster than a conventional method with inverse matrix calculations. Future work includes validation in a reverberation environment.

ACKNOWLEDGMENT

This work was supported by KAKENHIJP22F22769 and JP22KF0141, and F-REI (JPFR23010102).

REFERENCES

[1] J. Heymann *et al.*, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016, pp. 196–200. DOI: 10.1109/ICASSP.2016.7471664.

[2] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.

[3] A. A. Nair *et al.*, "Audiovisual zooming: What you see is what you hear," in *ACM Multimedia*, 2019, pp. 1107–1118.

[4] J. Benesty *et al.*, *Microphone array signal processing*. Springer Science & Business Media, 2008.

[5] H. Zhao *et al.*, "The sound of pixels," in *ECCV*, 2018, pp. 570–586.

[6] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, 2018, Article No. 112, ISSN: 0730-0301. DOI: 10.1145/3197517.3201357.

[7] Y. Ji *et al.*, "Self-supervised fine-grained cycle-separation network (fscn) for visual-audio separation," *IEEE Transactions on Multimedia*, vol. 25, pp. 5864–5876, 2023. DOI: 10.1109/TMM.2022.3200282.

[8] R. Lu *et al.*, "Listen and look: Audio–visual matching assisted speech source separation," *IEEE Signal Processing Letters*, vol. 25, no. 9, pp. 1315–1319, 2018. DOI: 10.1109/LSP.2018.2853566.

[9] Y. Gao *et al.*, "Audio compensation network: To improve the quality of low-energy audio in visual sound separation," in *ICCECE*, 2021, pp. 727–732. DOI: 10.1109/ICCECE51280.2021.9342383.

[10] O. Thiergart *et al.*, "An acoustical zoom based on informed spatial filtering," in *IWAENC*, 2014, pp. 109–113. DOI: 10.1109/IWAENC.2014.6953348.

[11] F. Rund *et al.*, "Objective quality assessment for the acoustic zoom," in *TSP*, 2015, pp. 392–396. DOI: 10.1109/TSP.2015.7296290.

[12] W. Ruochen *et al.*, "Acoustic zooming based on real-time metadata control," in *IC-NIDC*, 2014, pp. 338–342. DOI: 10.1109/ICNIDC.2014.7000321.

[13] D. Kitamura *et al.*, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016. DOI: 10.1109/TASLP.2016.2577880.

[14] D. Kitamura *et al.*, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, pp. 1–25, 2018.

[15] A. Ozerov *et al.*, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *ICASSP*, 2011, pp. 257–260. DOI: 10.1109/ICASSP.2011.5946389.

[16] Z. Zhong *et al.*, "Design and assessment of a scan-and-sum beamformer for surface sound source separation," in *IEEE/SICE SII*, 2020, pp. 808–813.

[17] Z. Zhong *et al.*, "Assessment of a beamforming implementation developed for surface sound source separation," in *2021 IEEE/SICE International Symposium on System Integration (SII)*, 2021, pp. 369–374.

[18] O. Thiergart *et al.*, "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates," in *ICASSP*, 2013, pp. 659–663. DOI: 10.1109/ICASSP.2013.6637730.

[19] O. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972. DOI: 10.1109/PROC.1972.8817.

[20] T. Chakrabarty *et al.*, "Performance investigation of robust linearly constrained minimum variance beamforming for direction of arrival mismatch," in *ICECTE*, 2019, pp. 149–152.

[21] T. Nakatani and K. Kinoshita, "Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer.," in *Interspeech*, 2019, pp. 111–115.

[22] H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *SIAM Review*, vol. 23, no. 1, pp. 53–60, 1981.

[23] V. Panayotov *et al.*, "LibriSpeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[24] J. Barker *et al.*, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *ASRU*, 2015, pp. 504–511. DOI: 10.1109/ASRU.2015.7404837.

[25] T. Kagoshima, N. Ding, and H. Fujimura, "Adaptive noise suppression for wake-word detection by temporal-difference generalized eigenvalue beamformer," in *APSIPA ASC*, 2020, pp. 805–809.

[26] C. Boeddeker, P. Hanebrink, L. Drude, J. Heymann, and R. Haeb-Umbach, "Optimizing neural-network supported acoustic beamforming by algorithmic differentiation," in *ICASSP*, 2017, pp. 171–175. DOI: 10.1109/ICASSP.2017.7952140.

[27] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006. DOI: 10.1109/TSA.2005.858005.

[28] C. H. Taal *et al.*, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2125–2136, 2011. DOI: 10.1109/TASL.2011.2114881.

Multi-Channel Fusion Human Activity Recognition Algorithm Based on Millimeter-Wave Radar

Junda Zhu*, Shisheng Guo*†‡, Longzhen Tang*, Cui Guolong*†
* School of Information and Communication Engineering,
University of Electronic Science and Technology of China, Chengdu 611731,China
† Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Quzhou 324000, China
‡ Corresponding author: ssguo@uestc.edu.cn

***Abstract*—Traditional datasets for human activity recognition (HAR) are typically obtained from single-channel radar, which does not effectively utilize the target feature information. This paper introduces a multi-channel fusion HAR algorithm based on millimeter-wave radar. The algorithm leverages the multi-channel data acquisition capability of multi-input multi-output (MIMO) radar and employs three independent MobileNetV3 lightweight network models for classification and recognition. Subsequently, a hard classification algorithm is utilized to fuse the multi-channel information, enabling the identification of human behaviors. Experimental results indicate that the proposed algorithm achieves an excellent classification performance of 97.05% on the experimental dataset. Moreover, it demonstrates a marked improvement in accuracy and stability compared to traditional single-channel classification methods. Furthermore, when compared with commonly used classification networks for human behavior, MobileNetV3 used in the algorithm highlights its ability to achieve commendable classification performance with minimal computational cost. This validates the proposed method as a viable lightweight hardware-transplantation approach.**

***Index Terms*—multi-input multi-output (MIMO), human activity recognition (HAR), multi-channel fusion, lightweight neural network**

I. Introduction

With the continuous advancement in the field of artificial intelligence and the maturation of Internet of Things (IoT) technologies, human activity recognition (HAR) has emerged as a subject of significant interest for further research in various domains such as public safety, smart elderly care, interactive gaming, and traffic monitoring [1]–[4]. Among the various methods of HAR, the use of millimeter-wave radar technology has gained favor due to its advantages of continuous monitoring, non-intrusive nature, and lack of privacy concerns. The implementation of millimeter-wave radar for HAR currently involves several approaches.

The first kind of approach involves the application of logical judgments and threshold detection, which offers simplicity and rapid response processing. J. Baik *et al.* proposed a method that increases the number of judgment threshold features by adding two additional detection features, which are centroid distance and distance width, to the conventional speed and acceleration, achieving certain classification results using these four features [5]. However, the primary limitation of this method is the poor classification performance due to false negatives and false positives, and the number and accuracy of thresholds also affect the performance of algorithm.

The second kind of approach utilizes traditional machine learning methods, relying on features extracted from sliding windows and classic machine learning algorithms. A. Shrestha *et al.* processed continuous data using a sliding window approach and introduced Sequential Forward Selection (SFS) as a feature selection tool to further optimize the classification performance of the Support Vector Machine (SVM) [6]. F. J. Abdu *et al.* employed Canonical Correlation Analysis (CCA) algorithm in conjunction with the SVM classifier to create a special discriminant vector for activity recognition [7]. Nevertheless, traditional machine learning methods are plagued by the complexities of manual feature extraction and performance constraints.

Lastly, there is the increasingly popular method of deep learning, with an increasing number of high-performance networks being applied to HAR. This classification approach has gradually become mainstream [8]–[12]. J. Maitre *et al.* proposed a deep neural network (DNN) model consisting of a convolutional neural network (CNN), a long-short-term memory (LSTM) network, and a fully connected neural network for HAR [8]. H. Li *et al.* proposed a framework based on a bidirectional LSTM network that integrates various methods for multi-modal sensor fusion in HAR [9]. Deep learning eliminates the tedious step of feature extraction in traditional machine learning, allowing raw datasets to be trained directly. However, datasets are typically obtained from single-channel radar data, which fails to effectively utilize the target feature information captured by MIMO radar.

Building upon this research, this paper presents a human activity recognition algorithm based on MIMO millimeter-wave radar. Specifically, after necessary preprocessing of the multi-channel echo data acquired by the MIMO radar, the data are classified through three independent, lightweight MobileNetV3 network models. Subsequently, a hard classification algorithm is utilized to integrate the multi-channel information, yielding the final recognition outcomes. To validate recognition performance of the algorithm, a dataset comprising 600 groups of each of the six common human activities is collected for verification. Experimental results indicate that the algorithm significantly improves recognition accuracy compared to tra-

This work was supported by the Municipal Government of Quzhou under Grant 2023D032.

ditional single-channel HAR algorithms. Simultaneously, our comparative experiments demonstrate that the proposed algorithm is a viable lightweight hardware-transplantation method.

II. Experimental Setup and Data Processing

A. Experimental Setup

The primary hardware components utilized in this experiment include the 60 GHz millimeter-wave radar sensor IWR6843, introduced by Texas Instruments (TI), and the DCA1000 data capture card, as illustrated in Fig. 1.

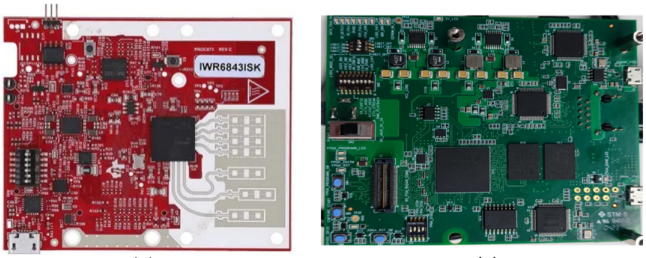


Fig. 1. Experimental instruments: (a) IWR6843 millimeter-wave radar and (b) DCA1000 data capture card

TABLE I presents the settings for the key parameters of the experimental radar system. Each frame of data lasts for 50 milliseconds, with a total of 100 frames collected in a single trial. The duration of data collection for each experiment is 5 seconds. The experiment employs a configuration of 3 transmitting and 4 receiving antennas, resulting in 12 data collection channels.

TABLE I
Experimental Radar System Parameter Settings

Start Frequency	60 GHz
Continuous Frequency Modulation Signal Bandwidth	1200 MHz
Continuous Frequency Modulation Signal Period	50 μs
Number of Transmitter and Receiver Antennas	3 transmitters 4 receivers
ADC Sampling Frequency	2 MHz
Radar Sampling Points	64
Continuous Frequency Modulation Slope	30.018 MHz/μs
Signal Rise Time	40 μs
Signal Sampling Gap	10 μs
Frame Sampling Periods	128 chirps per frame
Maximum Detectable Distance	10 m
Maximum Measurable Velocity	8.33 m/s

In this experiment, we focus on the multi-classification of six common human behaviors, which include waving hand, standing up, sitting down, falling, getting up and sleeping. Illustrations of these six behaviors are shown in Fig. 2. To enhance the generalization ability of the recognition results, all behavior data collection in this study is performed from different angles and distances. The specific experimental scenario model is depicted in Fig. 3, where the experiment is conducted in a relatively open space. All human behaviors are collected within a 5*5 meter square area, with the radar positioned at a height of 2.8 meters above the ground.

There are 10 participants, with heights ranging from 1.6 m to 1.85 m. Each participant performs behaviors facing the radar, but their orientation relative to the radar is random. The specific behavior dataset is introduced in TABLE II. While

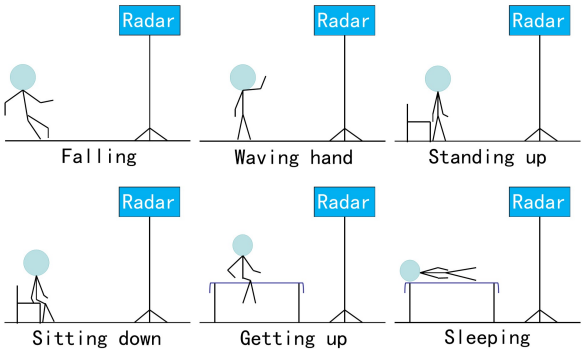


Fig. 2. Schematic diagram of six common human behaviors

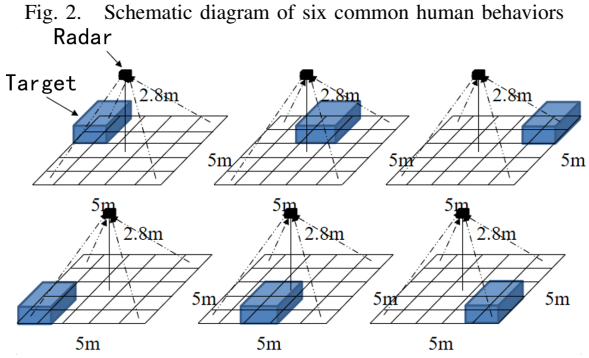


Fig. 3. Experimental scenario model

the radar detection range may have included interference from other static objects, the presence of rapidly moving dynamic objects is avoided. The actual experimental setup is depicted in Fig. 4.

TABLE II
Experimental Radar System Parameter Settings

Behavior Types Collected	6
Number of Targets	10 individuals
Data Volume per Single Target and Single Behavior	60
Target Activity Range	5 m*5 m
Radar Height from Ground	2.8 m
Total Dataset Data Volume	3600



Fig. 4. Actual experimental scenes: (a) empty scene and (b) HAR scene

B. Experimental Data Processing

In the 3-transmit 4-receive mode used in this experiment, the radar RF transmission process is altered compared to the single-transmit mode. The radar divides each chirp into 3 time blocks, and during different time blocks, the 3 TX antennas sequentially transmit their signals, while the 4 RX antennas simultaneously receive the echo signals. This allows for radar

data collection under the 3-transmit 4-receive mode. After the data obtained from each channel is stored as a row vector in a specific order, we obtain the initial single-pulse sampling data matrix.

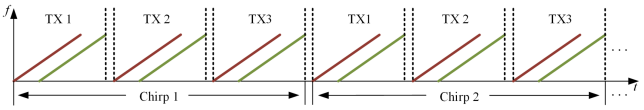


Fig. 5. Radar signal transmission model in 3-transmit 4-receive mode

After preprocessing the intermediate frequency signals and reshaping the data matrix, this paper obtains a radar three-dimensional data matrix. For the data from each channel, the Fast Fourier Transform (FFT) in the range dimension is performed, which is applied to each row of the data matrix. The specific expression for the FFT is as follows:

$$TR_{(m,k)} = \sum_{u=1}^{N_s} \|W_u S_{(n-u,m)} e^{-j2\pi ku/N}\|^2 \quad (1)$$

where $TR_{(m,k)}$ represents the amplitude of the mth chirp after discrete FFT at the digital frequency point $2\pi k/N$,where N is the number of points for the discrete FFT, W_u is a predefined Hamming window function, and $S_{(n-u,m)}$ denotes the data of the $n-u$ sample point of the mth chirp, where n represents the sample point, and u is the integration variable. The obtained Range-Time map is shown in Fig. 6.

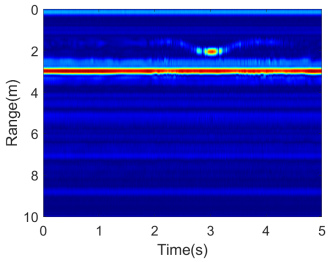


Fig. 6. Obtained Range-Time map

In practice, during the radar data collection process, in addition to human targets, many other objects can interfere with the echo signals, affecting the identification of the distance component. For these constant direct current (DC) components that are present in each collection, this paper employs intra-frame cancellation methods to remove them. The calculation formula is as follows:

$$MTI_{(m,k)} = TR_{(m,k)} - TR_{(m+j,k)} \quad (2)$$

where $MTI_{(m,k)}$ represents the Range-Time map after cancellation, and j denotes the span of continuous frequency modulation for each cancellation. After the aforementioned processing, the final Range-Time map can be obtained, as shown in Fig. 7.

After the application of range-FFT to the matrix data of each frame following clutter suppression, the sum of each row is computed. Subsequently, the maximum value is sought across each column, and the index of this maximum value is interpreted as the target position within the frame. A Doppler dimension FFT is then performed at the target position of each frame. Through a process analogous to the acquisition of the Range-Time map described previously, a Frequency-Time map is obtained.

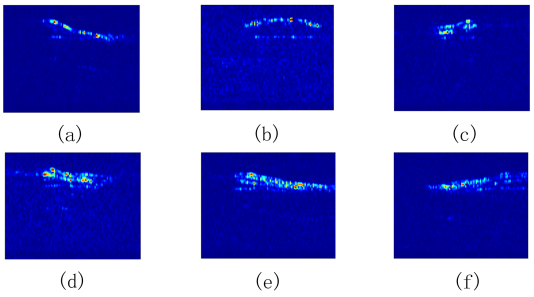


Fig. 7. Processed Range-Time map for 6 types of behaviors: (a) falling (b) waving hand (c) standing up (d) sitting down (e) sleeping (f) getting up

III. Multi-Channel Fusion Algorithm for Human Behavior Information Based on MobileNetV3

A. Network Classification Architecture Based on MobileNetV3

To enhance the generalization ability of the algorithm, this paper proposes to use the MobileNetV3 network, which has shown superior performance in the field of lightweight classification, as the basic architecture for the classification network.

The core idea of the MobileNetV3 is to use depthwise separable convolutions to replace the traditional pointwise convolution methods, thereby significantly reducing the computational complexity and achieving network lightweighting. Depthwise separable convolutions consist of two steps: depthwise convolution and pointwise convolution, as illustrated in Fig. 8. Depthwise convolution refers to the application of convolution to each input channel separately, resulting in multiple output channels. Pointwise convolution involves convolving each output channel, thus merging the output channels from multiple depthwise convolutions. This method of depthwise separable convolution reduces the amount of computation and the number of parameters while maintaining good accuracy and generalization capabilities.

The core structure of MobileNetV3 is the bottleneck (bneck) structure, as shown in Fig. 9. The structure begins with an expanded convolution, which increases the number of channels through a 1*1 convolution. This is followed by a depthwise separable convolution. MobileNetV3 then introduces a Squeeze-and-Excitation (SE) channel attention mechanism. After passing through the residual structure of the linear bottleneck layer, the h-swish activation function is used [13]. This maintains performance while improving computational efficiency.

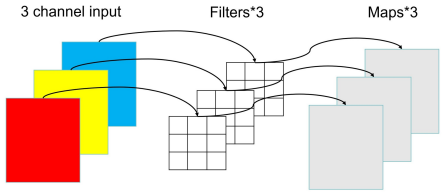


Fig. 8. Depthwise convolution in depthwise separable convolution

Compared to traditional CNN represented by ResNet, MobileNetV3, through techniques such as depthwise separable

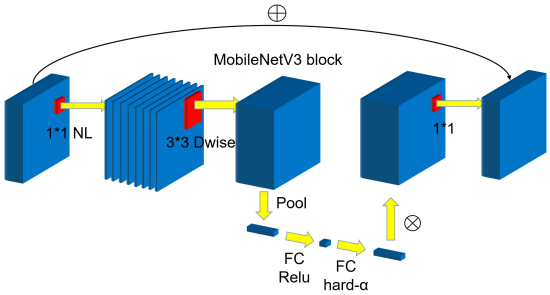


Fig. 9. MobileNetV3 bneck structure

convolutions and the Squeeze-and-Excitation (SE) channel attention mechanism, can significantly reduce computational complexity and the number of parameters while maintaining good accuracy and generalization capabilities. This has made it one of the important models for efficiently performing visual tasks such as image classification and object detection on mobile devices.

B. Multi-Channel Information Fusion Algorithm Based on Hard Classification

In typical HAR tasks, limited by radar hardware or other reasons, other algorithms often use single-channel radar data for HAR, which can lead to two issues. Firstly, single-channel radar systems can only provide information from a single angle or direction, which is insufficient for acquiring a comprehensive spatial understanding of target behaviors. Secondly, datasets acquired using a single radar are prone to environmental noise and interference, leading to significant fluctuations in training and testing results, particularly on more lightweight network architectures. By integrating multi-channel data for HAR, we can not only extract target behavior features from different orientations and at multiple levels but also enhance the generalization ability of the HAR task through certain channel information fusion algorithms, thereby improving task performance.

For the 3-transmit 4-receive MIMO radar used in this experiment, as shown in Fig. 10, in order to fully obtain information from different angles, the experiment plans to use the data from channels 1, 8, and 11 to synthesize target information. As illustrated, these three channels can fully capture target behavior information from multiple angles for HAR. After appropriate preprocessing, the information from the three channels is processed through three independent MobileNetV3 network classification architectures for classification training.

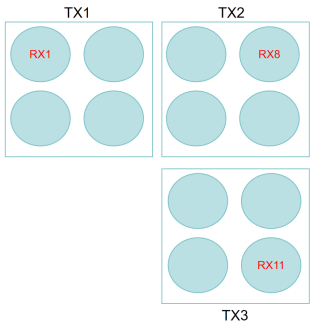


Fig. 10. Data channels used in 3-transmit 4-receive MIMO radar

Hard classification is a fundamental classification method in the fields of machine learning and pattern recognition, where each sample is assigned a clear category label. For a given input sample, a hard classifier outputs a category, usually the index or name of the category. Compared to the soft classification used in neural network training, hard classification has a stronger ability to resist overfitting in small datasets. Moreover, in the recognition of easily confused human behaviors, hard classification can better integrate the target feature differences from different channels, thus increasing the recognition accuracy of easily confused behaviors.

The experimental multi-channel information fusion algorithm based on hard classification decision is illustrated in Fig. 11. Initially, the range echo data from different channels are classified using independent network architectures to obtain decision vectors. Subsequently, a non-probabilistic model is applied to the decision vectors of each channel, selecting the decision vector that appears most frequently among those representing different target behaviors as the final decision vector, which is then used for the multi-class classification task of behavioral targets.

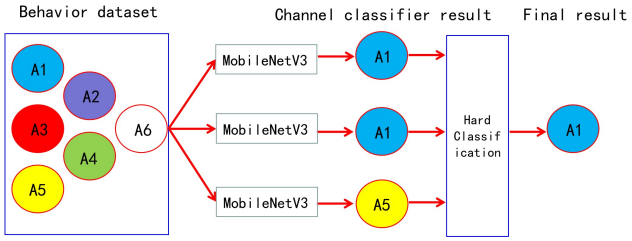


Fig. 11. Multi-channel information fusion model based on hard classification

In cases where discrimination is not possible, a further hard classification is performed on the frequency-time maps from different channels to obtain classification results. If discrimination still fails, a reclassification based on the independent lightweight network model for different channels is conducted for the easily confused behaviors recorded in the aforementioned two rounds of decision-making. Unlike the previous two rounds of discrimination, in this round, the test set categories are only labeled with the recorded easily confused behaviors, allowing for a targeted hard classification to obtain the final decision label. The specific multi-round hard classification process is depicted in Fig. 12. The aforementioned method not only comprehensively integrates multi-channel information for more accurate judgments but also synthesizes multi-domain information and addresses easily confused human behaviors, thereby effectively achieving the task of HAR.

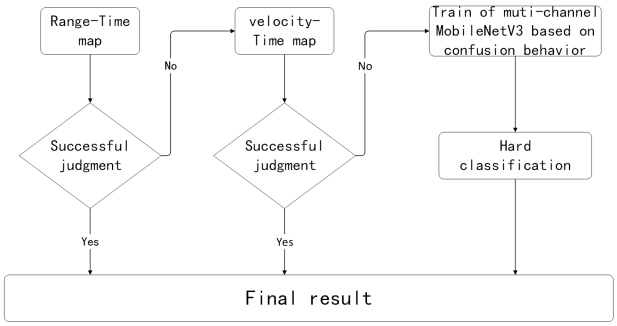


Fig. 12. Multi-round hard classification

IV. Experimental Results and Analysis

A. Network Training Details

Regarding the division of the training set, for the 6 types of behaviors, each behavior has a dataset of 600. The experiment uses a 7:3 ratio to divide the training and test sets, meaning each behavior has a training set of 420 and a test set of 180. The experiment samples using a uniform sampling scheme to ensure the generalization ability of the experiment.

As for the setting of experimental hyperparameters, the experiment uses the Adam optimizer and cross-entropy loss function for training. The training learning rate is set to 0.001, with a total of 100 epochs for training. The GPU device used for training is the RTX4070Ti.

B. Ablation Experiment Results of Single-Channel and Multi-Channel

In the ablation experiments, we conduct classification experiments using the MobileNetV3-s network architecture for a single channel. To mitigate the potential impact of insufficient data volume on the results of the single-channel ablation experiment, we integrate the data from three channels as the training and test sets for the single-channel experiment, thereby ensuring that the training data volume for the multi-channel and single-channel experiments is identical. The confusion matrices for multi-channel classification based on hard classification algorithm decisions and single-channel classification are depicted in Fig. 13(a) and (b) respectively. The average accuracy comparison is presented in TABLE III.

We can observe that, under the condition of equal training data volume, the multi-channel classification based on hard classification algorithm decisions not only comprehensively integrates target behavior information but also demonstrates a significant improvement in HAR accuracy. Moreover, it effectively enhances the generalization ability of recognition.

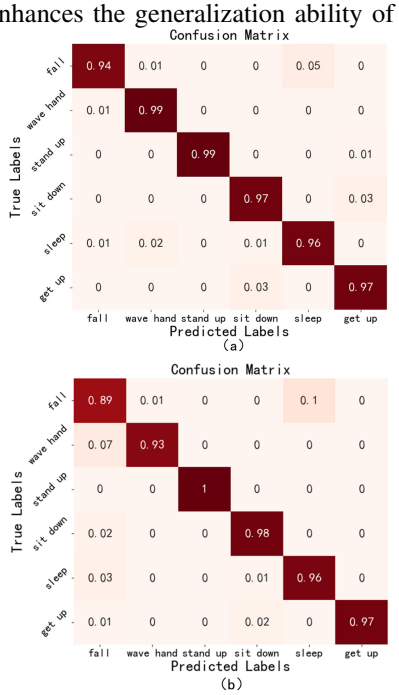


Fig. 13. Ablation experiment classification confusion matrices: (a) multi-channel classification and (b) single-channel classification

TABLE III
Ablation Experiment Average Training Accuracy Comparison

Different Channel MobileNetV3 Model	Average Accuracy
Three-Channel Fusion	97.05%
Channel 1	94.04%
Channel 8	94.52%
Channel 11	93.25%

C. Comparative Experiment Results with MobileNetV3 and Different Network

For the comparative experiment, we select several network structures that are popular in lightweight networks, including GhostNet and EfficientNetv2-s, as well as traditional CNN networks such as ResNet18, and Vision Transformer (ViT) networks with transformer structures. Fig. 14 display the classification confusion matrices for MobileNetV3-s, GhostNet, EfficientNetv2-s, ResNet18, and ViT networks. TABLE IV presents the comparison of average training accuracy, the number of parameters and floating point operations (FLOPs) for each network model. It can be seen that MobileNetV3 achieves a high classification accuracy with extremely small computational costs, demonstrating unique advantages in scenarios where radar device integration and small data processing are required.

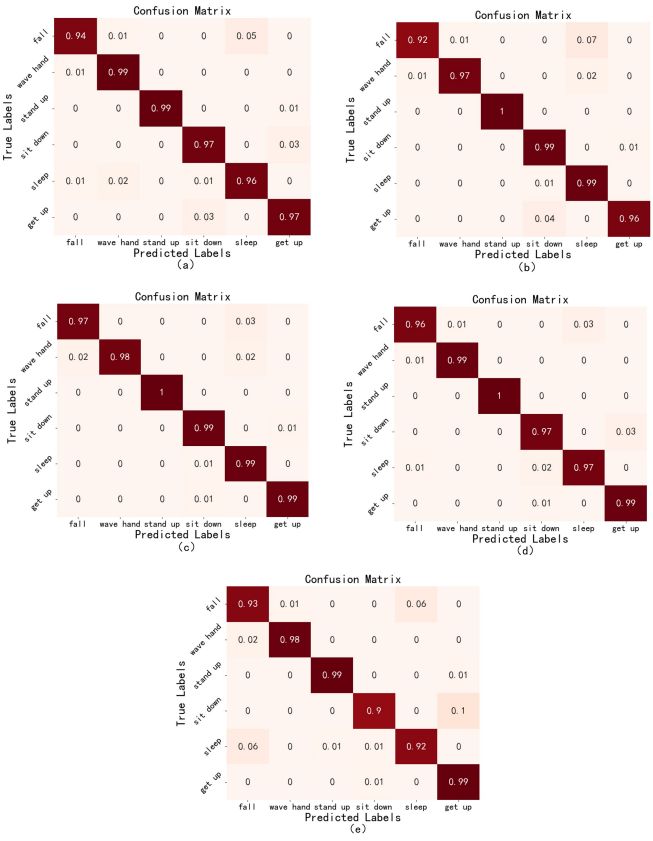


Fig. 14. Comparative experiment classification confusion matrices: (a) MobileNetV3-s (b) GhostNet (c) EfficientNetv2-s (d) ResNet18 (e) ViT

TABLE IV
COMPARISON OF AVERAGE TRAINING ACCURACY AND NUMBER OF PARAMETERS AND FLOPS FOR EACH NETWORK MODEL

Network Model	Average Accuracy	Parameters	FLOPs
MobileNetV3-s	97.05%	1.62M	0.061G
GhostNet	96.75%	4.40M	0.22G
EfficientNetv2-s	97.42%	20.19M	2.89G
ResNet18	97.13%	11.2M	1.82G
ViT	95.23%	86.5M	16.86G

In summary, through comparative and ablation experiments, it can be observed that the multi-channel classification algorithm based on MobileNetV3 proposed in this paper achieves superior HAR performance with a smaller computational cost in application scenarios involving embedded devices and small datasets. Compared to other traditional HAR algorithms, it exhibits strong advantages.

V. CONCLUSIONS

This paper proposes an algorithm for HAR using MIMO millimeter-wave radar, which is based on the MobileNetV3 network and multi-channel information fusion. Through experiments, it can be observed that the proposed multi-channel information fusion algorithm outperforms traditional single-channel algorithms in terms of recognition accuracy and stability and the recognition accuracy reaches 97.05%. For comparative experiments, the MobileNetV3 network achieves a relatively high recognition accuracy with smaller parameter count and FLOPs compared with other popular HAR classification networks, indicating that the MobileNetV3 satisfies the condition of achieving superior HAR classification with minimal computational cost, and represents a viable lightweight hardware-transplantation recognition approach.

REFERENCES

[1] Y.Qiao, J.Luo, T.Cui, *et al,* "Soft Electronics for Health Monitoring Assisted by Machine Learning,"*Nano-Micro Letters*,2023,15(5): 83-168.

[2] Yussof, Salleh, Tokhi, *et al,* "Towards understanding on the development of wearable fall detection an experimental approach,"*IEEE Sensors Journal*,2022,12(2):345-358.

[3] M.Vito, A.Federica, P.Luca, *et al,* "Biomechanical Measures for Fall Risk Assessment and Fall Detection in People with Transfemoral Amputations for the Next-Generation Prostheses: A Scoping Review,"*Journal of prosthetics and orthotics: JPO*,2022,34(3):144-162.

[4] A.B.H.Khaled, A.Khalfallah, M.S.Bouhlel, "Systematic review of indoor fall detection systems for the elderly using Kinect,"*International journal of telemedicine and clinical practices*, 2022,3(4):276-301.

[5] J.Baik, C.Jung, A.Nam, *et al, Fall detection and reducing detection error using FMCW radar*.2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC). IEEE, 2022, pp.553-555.

[6] H.Li, A.Shrestha, H.Heidari, *et al,Activities recognition and fall detection in continuous data streams using radar sensor*. 2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC). IEEE, 2019,1, pp.1-4.

[7] F.J.Abdu, Y.Zhang, Z.Deng, "Activity classification based on feature fusion of FMCW radar human motion micro-Doppler signatures,"*IEEE Sensors Journal*, 2022, 22(9): 8648-8662.

[8] J.Maitre, K.Bouchard, S.Gaboury, "Fall detection with UWB radars and CNN-LSTM architecture," *IEEE Journal of Biomedical and Health Informatics*, 2020, 25(4): 1273-1283.

[9] H.Li, A.Shrestha, H.Heidari, *et al,* "Bi-LSTM network for multimodal continuous human activity recognition and fall detection,"*IEEE Sensors Journal*,2019, 20(3): 1191-1201.

[10] T.Liang, H.Xu, *A posture recognition-based fall detection system using a 24GHz CMOS FMCW radar SoC*. 2021 IEEE MTT-S International Wireless Symposium (IWS). IEEE,2021, pp. 1-3.

[11] A.S.Shah, A.Tahir, J.Le Kernec, *et al,* "Data portability for activities of daily living and fall detection in different environments using radar micro-doppler,"*Neural Computing and Applications*, 2022, 34(10): 7933-7953.

[12] X.Yu, C.Feng, L.Yang, *et al,* "Human Fall Detection by FMCW Radar Based on Time-Varying Range-Doppler Features,"*International Journal of Computer and Systems Engineering*,2022, pp. 318-323.

[13] A.Howard, M.Sandler, G.Chu, *et al, Searching for mobilenetv3*. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1314-1324.

Run-Time Adaptation of Neural Beamforming for Robust Speech Dereverberation and Denoising

Yoto Fujita[1,2], Aditya Arie Nugraha[2], Diego Di Carlo[2],
Yoshiaki Bando[3,2], Mathieu Fontaine[4,2], and Kazuyoshi Yoshii[5,2]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]Center for Advanced Intelligence Project (AIP), RIKEN, Japan
[3]AIRC, National Institute of Advanced Industrial Science and Technology (AIST), Japan
[4]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[5]Graduate School of Engineering, Kyoto University, Japan

***Abstract*—This paper describes speech enhancement for real-time automatic speech recognition (ASR) in real environments. A standard approach to this task is to use neural beamforming that can work efficiently in an online manner. It estimates the masks of clean dry speech from a noisy echoic mixture spectrogram with a deep neural network (DNN) and then computes a enhancement filter used for beamforming. The performance of such a supervised approach, however, is drastically degraded under mismatched conditions. This calls for run-time adaptation of the DNN. Although the ground-truth speech spectrogram required for adaptation is not available at run time, blind dereverberation and separation methods such as weighted prediction error (WPE) and fast multichannel nonnegative matrix factorization (FastMNMF) can be used for generating pseudo ground-truth data from a mixture. Based on this idea, a prior work proposed a dual-process system based on a cascade of WPE and minimum variance distortionless response (MVDR) beamforming asynchronously fine-tuned by block-online FastMNMF. To integrate the dereverberation capability into neural beamforming and make it fine-tunable at run time, we propose to use weighted power minimization distortionless response (WPD) beamforming, a unified version of WPE and minimum power distortionless response (MPDR), whose joint dereverberation and denoising filter is estimated using a DNN. We evaluated the impact of run-time adaptation under various conditions with different numbers of speakers, reverberation times, and signal-to-noise ratios (SNRs).**

***Index Terms*—speech enhancement, dereverberation, neural beamforming, blind source separation,**

I. Introduction

Robust speech enhancement is a key technique in practical automatic speech recognition (ASR) systems that work in real time in real environments. For this purpose, one may use blind source separation (BSS) methods such as multichannel nonnegative matrix factorization (MNMF) [1], [2], independent low-rank matrix analysis (ILRMA) [3], and FastMNMF [4], [5]. Among these, FastMNMF is a state-of-the-art method that has been shown to outperform MNMF and ILRMA [5]. These methods are based on unsupervised learning (maximum likelihood estimation) of probabilistic models of mixture signals and are thus essentially free from the condition mismatch problem of supervised learning methods. However, these methods are hard to use for real-time systems due to the computationally demanding iterative optimization required at run time.

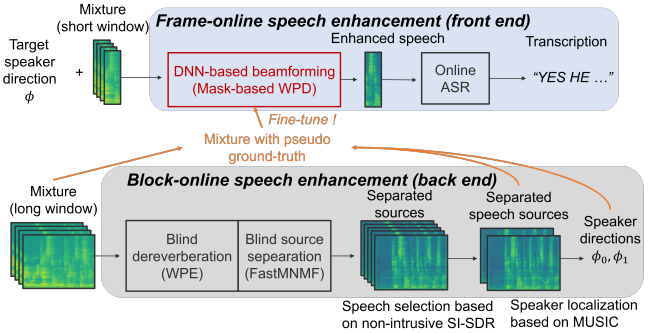


Fig. 1. The overview of our proposed adaptation of neural beamforming based on blind dereverberation by WPE and blind source separation by FastMNMF.

In recent years, deep neural networks (DNNs) have widely been used for speech enhancement. This approach, for example, performs direct mapping from a noisy mixture into multiple speech sources [6], mixture-conditioned deep speech generation [7], [8] for single-channel data, and neural beamforming with a DNN-based mask estimator [9] for multichannel data. In general, while supervised training of a DNN is computationally demanding, inference with the DNN works fast even on edge devices [10]. Considering the low-latency and high-performance nature and the potential contribution to the ASR [11], we focus on DNN-based beamforming (a.k.a. neural beamforming) as a front end of real-time distant ASR.

In neural beamforming, a DNN is used for estimating speech masks in the time-frequency domain. To separate clean speech from noisy speech mixture, the spatial covariance matrices (SCMs) of speech and noise are computed from the estimated masks and a enhancement filter used for beamforming is computed from the SCMs. The DNN is trained in a supervised manner by using pairs of noisy mixture and clean speech, often with the directions of the target speakers [10], [12]. However, it is not realistic to collect training data covering diverse acoustic environments that the model is potentially applied to. This makes the model less robust to unseen acoustic environments.

A promising solution to this problem is a run-time adaptation

of neural beamforming [10], [13]. The biggest challenge in this task is that no ground-truth data (clean speech) are available unlike in standard offline benchmarks. To solve this problem, one can fine-tune a DNN-based mask estimator using *pseudo* ground-truth data given by FastMNMF. In a dual-process system [10], a light-weight minimum variance distortionless response (MVDR) beamforming (front end) is used for streaming speech enhancement, where the mask estimator is fine-tuned with the target speech separated by asynchronously-running FastMNMF (back end), often with the direction of the target speaker. It has been shown that the ASR performance tends to improve along with the amount of fine-tuning data (e.g., multi-party conversation data) [10]. This system also uses weighted prediction error (WPE) [14], [15], a popular blind dereverberation method, before MVDR beamforming and FastMNMF for improved ASR. However, since the adaptation is only applied to the mask estimator for beamforming, the adaptation capability of this system is thus limited to MVDR beamforming only.

In this paper, we propose run-time adaptation of neural beamforming for joint speech dereverberation and denoising. Since WPE works stably in various environments thanks to the unsupervised nature [10], we aim to draw its full potential with its neural extension. Specifically, we use weighted power minimization distortionless response (WPD) beamforming [16], [17], a unified version of WPE and a minimum power distortionless response (MPDR), where a DNN-based mask estimator is used to estimate a unified dereverberation and denoising filter. Both the speech dereverberation and denoising functions of the system can be adapted to a test environment while considering the mutual dependency of both tasks. We comprehensively investigate acoustic conditions in which the adaptation effectively contributes to the improvement of speech enhancement and ASR.

II. Related Work

This section reviews speech dereverberation based on WPE [14], [15], speech enhancement based on MPDR beamforming, joint dereverberation and denoising based on WPD beamforming [16], [17], and BSS based on FastMNMF [5].

A. Dereverberation

WPE is a well-known blind dereverberation method based on an autoregressive model of late reverberation. Let $\mathbf{x}_{ft} \in \mathbb{C}^M$ be the short-time Fourier transform (STFT) spectrum of an observed mixture captured by an M-channel microphone array at frequency $f \in [1, F]$ and time frame $t \in [1, T]$, where F is the number of frequency bins and T is the number of frames. We assume $\mathbf{x}_{ft}$ can be decomposed as follows:

$$\mathbf{x}_{ft} = \mathbf{d}_{ft} + \mathbf{r}_{ft}, \tag{1}$$

where $\mathbf{d}_{ft} \in \mathbb{C}^M$ is direct signals with early reflections, $\mathbf{r}_{ft} \in \mathbb{C}^M$ is the spectrum of late reverberation. The late reverberation is assumed to be the weighted sum of past observations as follows:

$$\mathbf{r}_{ft} = \sum_{\tau=b}^{L} \mathbf{W}_{f\tau}^{\mathsf{H}} \mathbf{x}_{f,t-\tau}, \tag{2}$$

where $\mathbf{W}_{f\tau} \in \mathbb{C}^{M \times M}$ is a mixing filter for delay τ, L is a tap length, and b is a prediction delay representing the boundary between the early reflections and late reverberation. The target $\mathbf{d}_{ft}$ is thus given by

$$\mathbf{d}_{ft} = \mathbf{x}_{ft} - \sum_{\tau=b}^{L} \mathbf{W}_{f\tau}^{\mathsf{H}} \mathbf{x}_{f,t-\tau}. \tag{3}$$

The filter is estimated by minimizing the weighted power of the estimated direct signal as follows:

$$\widehat{\mathbf{W}}_f = \underset{\mathbf{W}_f}{\operatorname{argmin}}\, \mathbb{E}_t \left[\frac{|\mathbf{x}_{ft} - \sum_{\tau=b}^{L} \mathbf{W}_{f\tau}^{\mathsf{H}} \mathbf{x}_{f,t-\tau}|^2}{\sigma_{ft}^2} \right], \tag{4}$$

where σ_{ft}^2 represents the time-varying power spectral density (PSD) of the target speech. The PSD can be obtained through iterative updates of the estimated target speech and its power [18], or by source mask estimation using a DNN [19].

B. Speech Enhancement

The multichannel signal model in (1) can be rewritten by decomposing a target speech signal as the product of a steering vector $\mathbf{a}_f \in \mathbb{C}^M$ and a source $s_{ft} \in \mathbb{C}$, while considering as noise other components including non-target components, early reflections, and late reverberations as follows:

$$\mathbf{x}_{ft} = \mathbf{a}_f s_{ft} + \mathbf{n}_{ft}, \tag{5}$$

where $\mathbf{n}_{ft}$ is the spectrum of noise. The target speech $\widehat{d}_{ft}$ is estimated by applying an enhancement filter $\mathbf{w}_{f0} \in \mathbb{C}^M$ to the mixture $\mathbf{x}_{ft}$ as follows:

$$\widehat{d}_{ft} = \mathbf{w}_{f0}^{\mathsf{H}} \mathbf{x}_{ft}. \tag{6}$$

In MPDR beamforming [20], the filter is estimated by minimizing the power of the observed mixture $\mathbf{x}_{ft}$, while maintaining a distortionless response in the direction of the steering vector $\mathbf{a}_f$ as follows:

$$\widehat{\mathbf{w}}_{f0}^{\text{MPDR}} = \underset{\mathbf{w}_{f0}}{\operatorname{argmin}}\, \mathbb{E}_t \left[|\mathbf{w}_{f0}^{\mathsf{H}} \mathbf{x}_{ft}|^2 \right] \quad \text{s.t.} \quad \mathbf{w}_{f0}^{\mathsf{H}} \mathbf{a}_f = 1. \tag{7}$$

The closed-form solution of the optimal filter is given by

$$\widehat{\mathbf{w}}_{f0}^{\text{MPDR}} = \frac{\mathbf{K}_f^{-1} \mathbf{a}_f}{\mathbf{a}^{\mathsf{H}} \mathbf{K}_f^{-1} \mathbf{a}_f}, \tag{8}$$

where $\mathbf{K}_f = \mathbb{E}_t[\mathbf{x}_{ft} \mathbf{x}_{ft}^{\mathsf{H}}]$ is the SCM of the mixture.

C. Joint Speech Dereverberation and Denoising

WPD beamforming is formulated by integrating WPE and MPDR beamforming for jointly dereverberation and enhancement. Specifiacally, using (3) and (6), the signal obtained with WPD beamforming is given by

$$\widehat{d}_{ft} = \mathbf{w}_{f0}^{\mathsf{H}} \left(\mathbf{x}_t + \sum_{\tau=b}^{L} \mathbf{W}_{f\tau}^{\mathsf{H}} \mathbf{x}_{f,t-\tau} \right) = \overline{\mathbf{w}}_f^{\mathsf{H}} \overline{\mathbf{x}}_{ft}, \tag{9}$$

where $\overline{\mathbf{w}}_f \in \mathbb{C}^{(L-b+1)M}$ is an integrated filter consisting of $\{\mathbf{w}_{ft}\}_{t=0,b,\ldots,L}$ and $\overline{\mathbf{x}}_{ft} \in \mathbb{C}^{(L-b+1)M}$ is the concatenation of the current and past observations $\{\mathbf{x}_{f,t-\tau}\}_{\tau=0,b,\ldots,L}$.

Using (4) and (7), the filter $\overline{\mathbf{w}}_f$ is estimated as

$$\widehat{\mathbf{w}}_f^{\text{WPD}} = \underset{\overline{\mathbf{w}}_f}{\operatorname{argmin}}\, \mathbb{E}_t \left[\frac{|\overline{\mathbf{w}}_f^{\mathsf{H}} \overline{\mathbf{x}}_{ft}|^2}{\sigma_{ft}^2} \right] \text{ s.t. } \mathbf{w}_{f0}^{\mathsf{H}} \mathbf{a}_f = 1. \quad (10)$$

The closed-form solution of the optimal filter is given by

$$\widehat{\mathbf{w}}_f^{\text{WPD}} = \frac{\overline{\mathbf{K}}_f^{-1} \overline{\mathbf{a}}_f}{\overline{\mathbf{a}}_f^{\mathsf{H}} \overline{\mathbf{K}}_f^{-1} \overline{\mathbf{a}}_f}, \quad (11)$$

where $\overline{\mathbf{K}}_f = \mathbb{E}_t[\overline{\mathbf{x}}_{ft} \overline{\mathbf{x}}_{ft}^{\mathsf{H}} \sigma_{ft}^{-2}]$ is the SCM of the mixture compensated by the PSD of the target speech, and $\overline{\mathbf{a}}_f \in \mathbb{C}^{(L-b+1)M}$ is the concatenation of the steering vector $\mathbf{a}_f$ and a zero vector $\mathbf{0} \in \mathbb{R}^{(L-b)M}$.

Using the SCM of the target speech $\overline{\mathbf{R}}_f = \overline{\mathbf{a}}_f \overline{\mathbf{a}}_f^{\mathsf{H}} |s_{ft}|^2$, (11) can be reformulated as:

$$\widehat{\mathbf{w}}_f^{\text{WPD}} = \frac{\overline{\mathbf{K}}_f^{-1} \overline{\mathbf{R}}_f}{\operatorname{tr}(\overline{\mathbf{K}}_f^{-1} \overline{\mathbf{R}}_f)} \overline{\mathbf{u}}_q, \quad (12)$$

where $\overline{\mathbf{u}}_q \in \mathbb{R}^{(L-b+1)M}$ is a one-hot vector whose q-th element (reference channel) takes one and zero otherwise.

In mask-based WPD beamforming, the PSD σ_{ft}^2 and SCM $\overline{\mathbf{R}}_f$ of the target speech are computed with $\widehat{\mathbf{s}}_{ft} = \omega_{ft} \mathbf{x}_{ft}$, where a time-frequency (TF) speech mask $\omega_{ft} \in [0, 1]$ is estimated by a DNN [21], [22].

D. Blind Source Separation

The general goal of BSS is to separate a mixture spectrogram $\{\mathbf{x}_{ft}\}_{f,t=1}^{F,T}$ into N source spectrograms $\{\{\mathbf{x}_{nft}\}_{f,t=1}^{F,T}\}_{n=1}^{N}$, where $\mathbf{x}_{ft}, \mathbf{x}_{nft} \in \mathbb{C}^M$. In modern BSS methods, each source $\mathbf{x}_{nft}$ is typically assumed to follow an M-variate circularly-symmetric complex Gaussian distribution as follows:

$$\mathbf{x}_{nft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \lambda_{nft} \mathbf{G}_{nf}), \quad (13)$$

where λ_{nft} and $\mathbf{G}_{nft}$ are the PSD and SCM of the source n. Assuming the additivity of complex spectrograms and using the reproductive property of the Gaussian distribution, the mixture $\mathbf{x}_{ft}$ is given by

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft} \mathbf{G}_{nf} \right). \quad (14)$$

In MNMF [1], [2] and its variants including FastMNMF [5], the PSDs $\{\lambda_{nft}\}_{f,t=1}^{F,T}$ are factorized with NMF as follows:

$$\lambda_{nft} = \sum_{k=1}^{K} u_{nkf} v_{nkt}, \quad (15)$$

where $\mathbf{u}_{nk} \in \mathbb{R}_+^F$ is a basis vector, $\mathbf{v}_{nk} \in \mathbb{R}_+^T$ is an activation vector, and K is the number of bases. In FastMNMF, the SCM $\mathbf{G}_{nf}$ is also factorized as follows:

$$\mathbf{G}_{nf} = \mathbf{Q}_f^{-1} \operatorname{Diag}(\tilde{\mathbf{g}}_n) \mathbf{Q}_f^{-\mathsf{H}}, \quad (16)$$

where $\mathbf{Q}_f \in \mathbb{C}^{M \times M}$ is a time-invariant diagonalization matrix and $\tilde{\mathbf{g}}_n \in \mathbb{R}_+^M$ is a frequency-invariant nonnegative vector.

The model parameters $\{\mathbf{u}_{nk}\}_{n,k=1}^{N,K}$, $\{\mathbf{v}_{nk}\}_{n,k=1}^{N,K}$, $\{\mathbf{Q}_f\}_{f=1}^{F}$, and $\{\tilde{\mathbf{g}}_n\}_{n=1}^{N}$ are estimated with an iterative optimization algorithm such that the likelihood of the parameters for the mixture given by (14) is maximized [5]. Given the optimal parameters, the separation filter $\mathbf{w}_{nft}^{\text{BSS}}$ is given by

$$\widehat{\mathbf{w}}_{nft}^{\text{BSS}} = \mathbf{Q}_f^{\mathsf{H}} \operatorname{Diag} \left(\frac{\lambda_{nft} \tilde{\mathbf{g}}_n}{\sum_{n'} \lambda_{n'ft} \tilde{\mathbf{g}}_{n'}} \right) \mathbf{Q}_f^{-\mathsf{H}} \mathbf{u}_q, \quad (17)$$

where $\mathbf{u}_q \in \{0, 1\}^M$ is a one-hot vector whose q-th element (reference channel) takes one and zero otherwise.

III. Proposed method

This section describes the proposed adaptive joint dereverberation and denoising system based on a dual-process architecture consisting of mask-based WPD beamforming with FastMNMF-guided fine-tuning. This system uses the WPD beamforming to perform low-latency speech dereverberation and denoising, resulting in a single-channel speech signal useful for the ASR system. To be adaptive to dynamic environments, its DNN-based mask estimator is fine-tuned at run time using speech signals dereverberated and separated with high-latency yet environment-robust WPE and FastMNMF.

A. Joint Neural Speech Dereverberation and Denoising

Given a mixture spectrogram $\mathbf{X} \triangleq \{\mathbf{x}_{ft} \in \mathbb{C}^M\}_{f=1,t=1}^{F,T}$ with target speaker DOAs $\boldsymbol{\phi} \triangleq \{\phi_t \in [0, 2\pi]\}_{t=1}^{T}$, a DNN $\mathcal{F}_{\boldsymbol{\Theta}}$ parameterized by $\boldsymbol{\Theta}$ is used for estimating TF masks $\boldsymbol{\omega} \triangleq \{\omega_{ft}\}_{f=1,t=1}^{F,T}$ as follows:

$$\boldsymbol{\omega} = \mathcal{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \boldsymbol{\phi}). \quad (18)$$

The PSD $\widehat{\sigma}_{ft}^2$ and SCM $\widehat{\mathbf{R}}_f$ of the target speech can be computed using the mask estimate as follows:

$$\widehat{\sigma}_{ft}^2 = \frac{1}{M} \sum_{m=1}^{M} |\omega_{ft} x_{ftm}|^2, \quad (19)$$

$$\widehat{\mathbf{R}}_f = \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathbf{s}}_{ft} \widehat{\mathbf{s}}_{ft}, \quad (20)$$

where $\widehat{\mathbf{s}}_{ft} = [\omega_{ft} \mathbf{x}_{ft}^{\top}, \omega_{f,t-b} \mathbf{x}_{f,t-b}^{\top}, \ldots, \omega_{f,t-L} \mathbf{x}_{f,t-L}^{\top}]^{\top} \in \mathbb{C}^{(L-b+1)M}$. The WPD filter $\widehat{\mathbf{w}}_f^{\text{WPD}}$ is then computed from $\widehat{\mathbf{K}}_f = \sum_t \overline{\mathbf{x}}_{ft} \overline{\mathbf{x}}_{ft}^{\mathsf{H}} \widehat{\sigma}_{ft}^{-2}$ and $\widehat{\mathbf{R}}_f$ as in (12). Finally, the target speech signal $\widehat{d}_{ft}$ corresponding to the DOA ϕ_t is obtained by applying $\widehat{\mathbf{w}}_f^{\text{WPD}}$ to $\overline{\mathbf{x}}_{ft}$ as in (9).

B. Pretraining of Mask Estimator

The DNN-based mask estimator $\mathcal{F}_{\boldsymbol{\Theta}}$ is pretrained using triples consisting of an M-channel mixture, a reference speech, and a target DOA. It is optimized to minimize the negative signal-to-distortion ratio (SDR) between the estimated time-domain speech signal $\widehat{\mathbf{d}} \in \mathbb{R}^S$ and the reference time-domain speech signal $\mathbf{d}^{\text{ref}} \in \mathbb{R}^S$ given by

$$\mathcal{L} = -10 \log_{10} \frac{\widehat{\mathbf{d}}^{\top} \widehat{\mathbf{d}}}{(\widehat{\mathbf{d}} - \mathbf{d}^{\text{ref}})^{\top} (\widehat{\mathbf{d}} - \mathbf{d}^{\text{ref}})}, \quad (21)$$

where S denotes the number of samples. The estimated time-domain speech signal $\widehat{\mathbf{d}}$ is obtained by applying the inverse STFT to the estimated speech signal in the STFT domain $\{\widehat{d}_{ft}\}_{f=1,t=1}^{F,T}$.

C. Run-Time Adaptation of Mask Estimator

To make the mask estimator $\mathcal{F}_{\boldsymbol{\Theta}}$ adaptive to environmental changes, we fine-tune it at run time using triples of the observed mixture $\{\mathbf{x}_{ft}\}_{f=1,t=1}^{F',T'}$, the pseudo ground-truth speech signal $\mathbf{d}_p^{\text{ref}} \in \mathbb{R}^S$, and the pseudo ground-truth DOA ϕ_p to minimize the negative SDR loss in (21), where F' and T' are the number of frequency bins and that of frames of the mixture used for fine-tuning, respectively.

To obtain the pseudo ground-truth speech signal, we first dereverberate the mixture $\mathbf{x}_{ft}$ using WPE as follows:

$$\mathbf{x}_{ft}^{\text{dry}} = \mathbf{x}_{ft} - \sum_{\tau=b}^{L} \mathbf{W}_{f\tau}^{\mathsf{H}} \mathbf{x}_{f,t-\tau}. \tag{22}$$

where $\mathbf{x}_{ft}^{\text{dry}}$ is the dereverberated mixture and $\mathbf{W}_{f\tau}$ is a filter for delay τ. The filter is estimated as (4) with the PSD σ_{ft}^2 obtained through iterative updates of the estimated dereverberated signal $\mathbf{x}_{ft}^{\text{dry}}$ and its power σ_{ft}^2 [18]. Then, we separate the sources $\{\{x_{nft}\}_{f,t=1}^{F',T'}\}_{n=1}^{N}$ from the dereverberated mixture using FastMNMF as follows:

$$x_{nft} = \left(\widehat{\mathbf{w}}_{nft}^{\text{BSS}}\right)^{\mathsf{H}} \mathbf{x}_{ft}^{\text{dry}}, \tag{23}$$

where the separation filter $\mathbf{w}_{nft}^{\text{BSS}}$ is calculated as in (17). We measure the signal quality of each separated source signal using the reference-less non-intrusive scale-invariant SDR [23], [24]. We take $N' (\leq N)$ separated signals that satisfy a predefined threshold α and consider these as pseudo ground-truth speech signals $\{\mathbf{d}_{p,n}^{\text{ref}}\}_{n=1}^{N'}$. Finally, the corresponding N' pseudo ground-truth DOAs $\{\phi_{p,n}\}_{n=1}^{N'}$ are estimated based on multiple signal classification (MUSIC) [25].

IV. Evaluation

We report a comprehensive evaluation conducted for assessing the performance of our adaptive system in various simulated acoustic environments.

A. Dataset

For the mask estimator $\mathcal{F}_{\boldsymbol{\Theta}}$, we made a *training dataset* comprising 36,000 triples and a *validation dataset* comprising 3,600 triples. Each triple consisted of a 2-second 7-channel mixture signal, a 2-second 7-channel target speech signal, and the corresponding target DOA. Each mixture was composed of two speech signals by different speakers, who were stationary during the recording, and a diffuse noise signal. The speech signals were randomly taken from the training set (for the *training dataset*) and the development set (for the *validation dataset*) of Librispeech [26], while the diffuse or moving noise signals were randomly taken from the DEMAND dataset [27].

Both mixture and target speech signals were simulated using Pyroomacoustics [28] by considering a 7-channel circular microphone array, configured to match the geometry constraints of the DEMAND dataset, within a 2-dimensional room. The room length and width were randomly sampled from uniform distributions $\mathcal{U}(7.6\text{m}, 8.4\text{m})$ and $\mathcal{U}(5.6\text{m}, 6.4\text{m})$, respectively. The 2-dimensional coordinates of the array center were sampled from $\mathcal{U}(3.6\text{m}, 4.4\text{m})$ and $\mathcal{U}(2.6\text{m}, 3.4\text{m})$, respectively. The distance between the array center and each speaker was sampled from $\mathcal{U}(1\text{m}, 2\text{m})$. The reverberation time (RT60) for the mixtures was varied between 0.25 and 0.7 seconds. The target speech signals, which were supposed to include the direct path and early component, were simulated in the same rooms as the mixtures, but with the RT60 fixed at 0.25 seconds. Indoor noise signals from the DEMAND dataset, excluding the environment "PSTATION" that was used for the test set (see below), were randomly selected and added to the simulated mixtures, with an SNR between -5.0 and 5.0 dB.

To evaluate the effectiveness of our adaptation method in various acoustic environments, we prepared a *test dataset* consisting of seven distinct simulation settings, i.e., one "default" setting and six variations. The default setting considered mixtures of two speakers with an RT60 of 0.5 seconds, an SNR of 30.0 dB, and other parameters were randomly sampled as in the pretraining dataset. For the other six settings, we independently varied the number of stationary speakers (3 and 4), the RT60s (0.8 and 1.2 seconds), and the SNR (5.0 and -5.0 dB). Speech signals were taken from the test set of Librispeech, with diffuse or moving noise signals from the environment "PSTATION" of the DEMAND dataset. For each setting, we generated 30 triples, each consisting of an approximately 8-minute mixture, the target speech signal, the target DOA, and corresponding transcriptions. The first four minutes of each recording were used for fine-tuning the mask estimator, with the remaining duration reserved for evaluation.

B. Experimental Settings

Both the front and back ends operated in the STFT domain. The STFT coefficients were computed using a window size of 1024 ($F = 513$) with a hop length of 256.

For the mask-based WPD beamforming, the prediction delay and the tap size of the convolutional filter were set to $b = 3$ and $L = 8$, respectively. The TF mask was estimated using a DNN as in [10]. The DNN was composed of a preprocessing network, a direction attractor network (DAN), and a bidirectional long short-term memory (BLSTM) network. The preprocessing network took as input the concatenation of the log magnitude of the mixture, the inter-channel phase difference, and the beamforming output by delay-sum beamforming, while the DAN took the target DOA as input. Given the outputs of these two networks, the BLSTM then estimates a TF mask. This mask estimator was pretrained on the *training dataset* using the AdamW optimizer with a learning rate of 10^{-4} and a batch size of 4. The model was trained for 20 epochs, and the model with the lowest validation negative SDR loss was selected for evaluation.

For the joint adaptation of mask-based WPD beamforming, the prediction delay, the tap size, and the number of iterations for parameter updates in WPE were set to $b = 3$ and $L = 13$, and 3, respectively. The number of sources, the number of bases, and the number of iterations for parameter updates in FastMNMF were set to $N = 5$, $K = 16$, and 200, respectively. The threshold for non-intrusive SI-SDR was set to $\alpha = 10.0$. The window size of the mixture given as the input to FastMNMF was set to 30 seconds. The learning rate was set to 4×10^{-5},

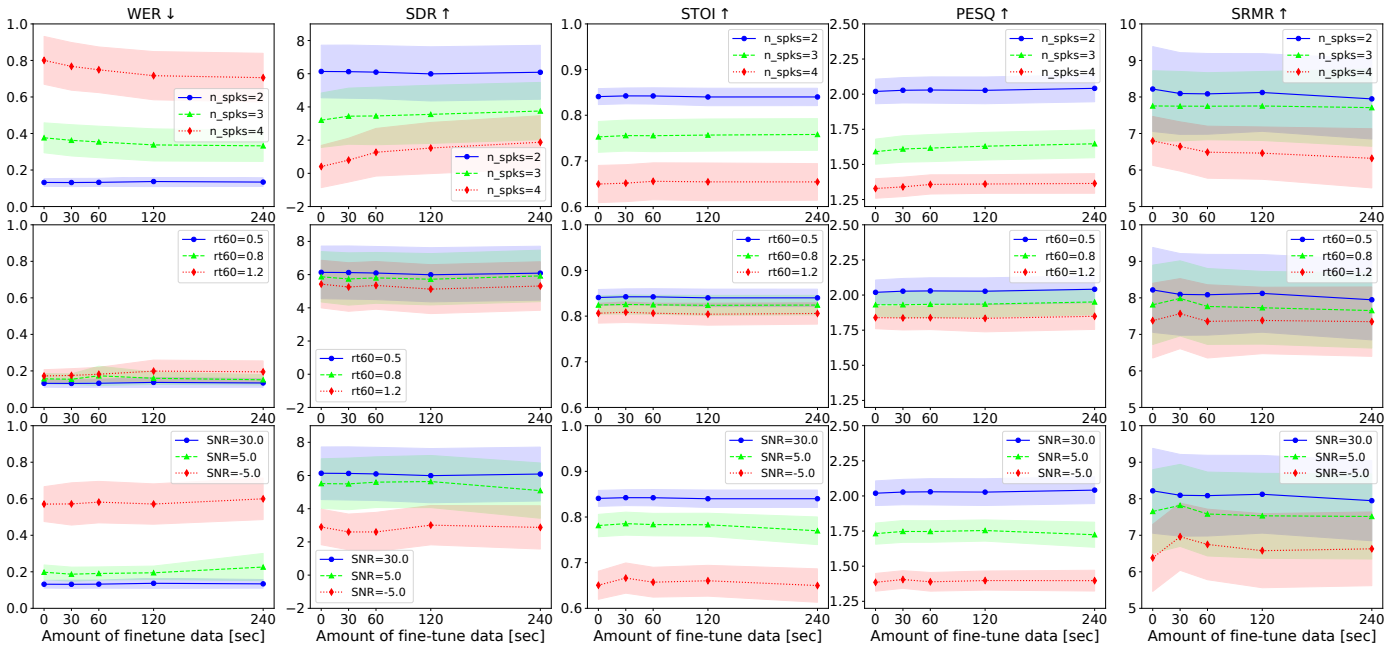


Fig. 2. The evaluation result of the joint adaptation of dereverberation and denoising by mask-based WPD beamforming based on blind dereverberation by WPE and BSS by FastMNMF. Blue lines refer to the "default" setting, constant across different plots. The shaded regions surrounding each line indicate the 95% confidence intervals.

and the batch size was 4 for fine-tuning. To stabilize the fine-tuning, we added the same amount of the pretraining data to the fine-tuning data so a batch may contain these two types of data.

The evaluation compared different amounts of fine-tuning data (30, 60, 120, and 240 seconds) against several key metrics. These metrics include word error rate (WER), signal-to-distortion ratio (SDR), short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ), and speech-to-reverberation modulation energy ratio (SRMR). We used a Transformer-based encoder-decoder ASR model from the SpeechBrain toolkit [29] to measure WER. This ASR model was trained on the Librispeech dataset. For all metrics except WER, higher values are better.

C. Experimental Results

Figure 2 shows the evaluation results of the joint adaptation of dereverberation and denoising using mask-based WPD beamforming with different durations of fine-tuning data and various simulation settings. The upper left plot in the figure indicates that our adaptation method improved WER, SDR, STOI, and PESQ across different numbers of stationary speakers, which proves the effectiveness of our adaptation. These improvements seem to benefit from the robust separation capability of FastMNMF for stationary sources.

However, when we used a large amount of fine-tuning data, the WER was slightly degraded as the RT60 increased or the SNR decreased. This would be because the pretrained mask-based WPD beamforming already has a strong capability of dereverberation, and FastMNMF is less robust to a mixture with moving sources. In contrast, when we used a small amount of fine-tuning data, the ASR performance hardly changed and the other metrics, STOI, PESQ, and SRMR improved in noisy or reverberant conditions. This suggests that 30 seconds of fine-tuning data is optimal for practical use.

V. Conclusion

This paper proposes a run-time adaptation method for the joint neural dereverberation and denoising with mask-based WPD beamforming using fine-tuning data obtained using WPE and FastMNMF. Evaluations showed robust improvements of the ASR performance across different numbers of stationary speakers, RT60s, and SNRs when we used a small amount of fine-tuning data. For future work, BSS methods that are capable of dealing with more various acoustic conditions should be investigated to improve the ASR performance even in noisy conditions with moving sources.

Acknowledgment

This work was partially supported by JST FOREST No. JPMJFR2270, JSPS KAKENHI Nos. 24H00742, 23K16912, and 23K16913, ANR Project SAROUMANE (ANR-22-CE23-0011), and Hi! Paris Project MASTER-AI.

References

[1] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[2] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 971–982, 2013.

[3] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[4] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 371–375.

[5] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2610–2625, 2020.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[8] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7402–7406.

[9] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Comput. Speech Lang.*, vol. 46, pp. 374–385, 2017.

[10] K. Sekiguchi, A. A. Nugraha, Y. Du, Y. Bando, M. Fontaine, and K. Yoshii, "Direction-aware adaptive online neural speech enhancement with an augmented reality headset in real noisy conversational environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 9266–9273.

[11] CHiME-4 Challenge Organizers, *CHiME-4 results*, [Online] Available: https://www.chimechallenge.org/challenges/chime4/results, [Accessed Jul. 6, 2024], 2016.

[12] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Proc. IEEE Spoken Language Technol. Workshop*, 2018, pp. 558–565.

[13] T. Aizawa, Y. Bando, K. Itoyama, K. Nishida, K. Nakadai, and M. Onishi, "Unsupervised domain adaptation of universal source separation based on neural full-rank spatial covariance analysis," in *Proc. IEEE Int. Workshop Mach. Learn. Signal. Process.*, 2023, pp. 1–6.

[14] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[15] T. Yoshioka and T. Nakatani, "Generalization of multichannel linear prediction methods for blind MIMO impulse response shortening," *IEEE Audio, Speech, Language Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.

[16] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, 2019.

[17] T. Nakatani and K. Kinoshita, "Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer," in *Proc. INTERSPEECH*, 2019, pp. 111–115.

[18] L. Drude, C. Boeddeker, J. Heymann, *et al.*, "Integrating neural network based beamforming and weighted prediction error dereverberation.," in *Proc. INTERSPEECH*, 2018, pp. 3043–3047.

[19] M. Togami, "Joint training of deep neural networks for multi-channel dereverberation and speech source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 3032–3036.

[20] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[21] T. Nakatani, R. Takahashi, T. Ochiai, *et al.*, "DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6399–6403.

[22] W. Zhang, A. S. Subramanian, X. Chang, S. Watanabe, and Y. Qian, "End-to-end far-field speech recognition with unified dereverberation and beamforming," *Proc. INTERSPEECH*, pp. 324–328, 2020.

[23] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[24] A. Kumar, K. Tan, Z. Ni, *et al.*, "Torchaudio-Squim: Reference-less speech quality and intelligibility measures in torchaudio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[25] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[27] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *Proc. Mtgs. Acoust.*, vol. 19, no. 1, pp. 1–6, 2013.

[28] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 351–355.

[29] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

Iterative Demographic Attentional Feature Fusion-based CNN and Transformer Network for Accurate Cuffless Blood Pressure Estimation

Liwen Tang [1], Dingchang Zheng [2] and Fei Chen [1]
[1] Southern University of Science and Technology, Shenzhen, China
E-mails: 12332129@mail.sustech.edu.cn, fchen@sustech.edu.cn
[2] Research Centre for Intelligent Healthcare, Coventry University, Coventry, United Kingdom
E-mail: ad4291@coventry.ac.uk

***Abstract*— Continuous cuffless blood pressure (BP) estimation is essential for cardiovascular disease monitoring. Recently, the use of deep learning models to automatically extract features and combine them with demographic features for continuous cuffless BP estimation has gained interest. Based on the observation that demographic features are highly correlated with BP estimation, this work proposes a new iterative demographic attentional feature fusion (AFF)-based CNN and Transformer network for better fusing the demographic features with the electrocardiogram (ECG) and photoplethysmography (PPG) features, as well as accurate BP estimation. This work tested model performance using a large open BP dataset, i.e., PulseDB. The BP estimation performance of the proposed model on PluseDB dataset meets the standards of the Association for the Advancement of Medical Instrumentation (AAMI) and achieves Grade A at the British Hypertension Society (BHS) standard in the estimate of systolic blood pressure (SBP) and diastolic blood pressure (DBP). The estimations of SBP and DBP have mean absolute errors (MAE) of 3.79 mmHg and 2.37 mmHg, respectively.**

I. INTRODUCTION

Cardiovascular disease (CVD) is the most common cause of death worldwide and is now responsible for nearly 32% of all deaths, according to the World Health Organization's statistics [1]. Systolic blood pressure (SBP) and diastolic blood pressure (DBP) are important indicators of cardiovascular health. Although the traditional cuff-based blood pressure (BP) measurement method can measure BP accurately, it is inconvenient and cannot be used to measure BP continuously. To achieve continuous and accurate BP monitor, cardiovascular signals like electrocardiogram (ECG) and photoplethysmogram (PPG) are used to estimate SBP and DBP by combining machine learning (ML) and deep learning methods.

Pulse wave analysis (PWA) is a traditional technology that estimates BP from cardiovascular signals, such as PPG. It manually extracts meaningful features from the clean signals and then finds a mapping from the extracted features and BP. Different regression ML algorithms, such as multiple linear regression [2], regression tree [3], gaussian process regression (GPR) [4], artificial neural network (ANN) [5], support vector machine (SVM), random forest (RF), and adaptive boosting (AdaBoost) [6], can be used to find the mapping between the features and BP. However, there are several drawbacks to the PWA method. First, extracting features is time-consuming [7], and another well-designed feature selection algorithm is needed for effective feature extraction [2, 4]. In addition, high-quality signals are essential during the feature extraction process.

Recently, deep learning has been used more and more in BP estimation. Unlike the ML-based PWA method, the deep learning model can automatically extract features from the row signals. Since the cardiovascular system is complex, many factors can influence BP estimation from cardiovascular signals [8]-[10]. Among these factors, demographic features (e.g., age, gender, height, and weight) are highly correlated with blood pressure, and a lot of works have tried to combine demographic features in their deep learning models [11]-[15]. Early studies only concatenate the demographic features with the features extracted by the deep learning model for later BP regression. This simple concatenation may not allow the deep learning model to effectively map characteristics of the cardiovascular signals to BP values for specific demographics. Therefore, the Attention pooling-based demographic feature fusion model is proposed to capture more essential feature frames in cardiovascular signals based on demographic features, and their works have shown that demographic features are important in BP estimation [14, 15].

On top of the research mentioned above, the aim of this study is to introduce an attentional feature fusion (AFF) module [16] to fuse features on the baseline CNN and Transformer model. Briefly, AFF will simultaneously consider the global and local characteristics of two features with inconsistent semantics and scales at the channel level and make a soft selection among these two features [16]. Meanwhile, considering a large dimensionality gap between the demographic features and cardiovascular signals features, the iterative demographic AFF-base CNN and Transformer network is proposed. In this work, the cardiovascular signal features are downsampled by

convolution and average pooling layers and then fused with demographic information several times. Also, we compare our new feature fusion module with other features fusion modules before. The results show that the model proposed in this work can achieve high accuracy of BP estimation in both SBP and DBP estimation.

II. Materials and Methods

A. Dataset

This study trains and tests the proposed model using a large dataset, PulseDB [17]. PulseDB is an open, cleaned dataset designed to benchmark BP estimation models. It uses MIMIC-III [18] and VitalDB [19] as its data sources. Synchronized ECG, PPG, and arterial blood pressure (ABP) signals were organized as 10-s segments at a 125 Hz sampling rate. The data processing details can be found in [17]. Reference SBP and DBP were calculated for each segment as the average of beat-to-beat SBP and DBP values retrieved from the ABP signal. Demographic information, including age (years), gender, height (cm), weight (kg), and body mass index (BMI) (kg/m^2), can be obtained from the VitalDB part of PulseDB. Still, the MIMIC-III part of PulseDB can only get age information. Therefore, we only use the VitalDB part of PulseDB to train and test our model to get more usable demographic information.

PulseDB randomly splits the training and calibration-based testing subsets. These two subsets contain 1,293 subjects' ECG, PPG segments, and BP labels from the VitalDB part of PulseDB. Each subject's segments are equal in the two subsets, 360 and 40, for each subject in the training and testing subsets, respectively. In this study, deep learning models are trained and tested using these two subsets, which means the testing is calibration-based. Table I summarizes the distribution of demographic characteristics among the 1,293 subjects involved in this study and the BP distribution of the training and testing dataset.

TABLE I

DISTRIBUTION OF SUBJECTS' DEMOGRAPHIC CHARACTERISTICS AND BP VALUES IN TRAINING AND TESTING SUBSETS

Item	Value	
Subjects	1293	
Age (years, mean ± SD)	59.0 ± 15.0	
Gender	746 Male, 547 Female	
Height (cm, mean ± SD)	162.5 ± 9.6	
Weight (kg, mean ± SD)	60.8 ± 11.7	
BMI (kg/m^2, mean ± SD)	22.9 ± 3.4	
	Training subset	Test subset
Segments	465480 (360 per subject)	51720 (40 per subject)
SBP (mmHg, mean ± SD)	115.48 ± 18.93	115.50 ± 18.85
DBP (mmHg, mean ± SD)	62.92 ± 12.08	62.94 ± 12.07

B. Iterative Demographic AFF-based CNN and Transformer Network

The iterative demographic AFF-based CNN and Transformer network uses CNN and Transformer as feature extraction models to extract features from PPG and ECG signals and combines the AFF module to fuse demographic information with extracted features for better BP estimation. Figure 1 depicts the structure of the proposed model, including several modules, i.e., CNN model, Transformer encoder, AFF, iterative demographic AFF, and regression.

The CNN module is built on the ResNet architectures [20]. The input signals pass through a basic convolution layer and three residual blocks. In residual blocks, the network directly adds the input to the convolved result by shortcut connection to minimize the problem of gradient explosion or disappearance of the model during the training process. After passing through the CNN module, the original input is changed from the original two-channel signals (PPG and ECG) with 1250 sequence length in time dimension to 256-channel feature vectors with 157 sequence length in time dimension. In this process, convolutional kernels extract local features of the input within the receptive field.

The sequence of local feature vectors extracted by the CNN module is used as input to the Transformer encoder. The Transformer encoder computes the global correlation between the feature vectors by using a multi-head attention (MHA) mechanism and extracts the important features using weighted averaging of feature vectors. The output of MHA passes through a feed-forward layer to obtain the final sequence of feature vectors. Since the attention mechanism cannot consider the temporal order between feature vectors, the positional encoding matrix is first added to the input to introduce temporal information about the sequence. The shortcut connection and layer norm methods are also used in the Transformer module. After going through the transformer encoder, the input sequence of feature vectors is further extracted to produce a sequence of global feature vectors with the same dimension.

As local and global physiological signal features, such as morphological and statistical features, are both important for BP estimation [5], making the model consider these two kinds of features is helpful. The AFF module is used to fuse features of inconsistent semantics and scales, like global and local features. Firstly, the points-wise summation of two features with different semantics is done. Then, AFF focuses on both local and global characters in each channel of the summation result and generates a weight mask of the same dimension. Point-wise soft selection between two features is done based on the mask. Then, the output features can include information about both features with different semantics. Here, local features extracted by the CNN model and global features

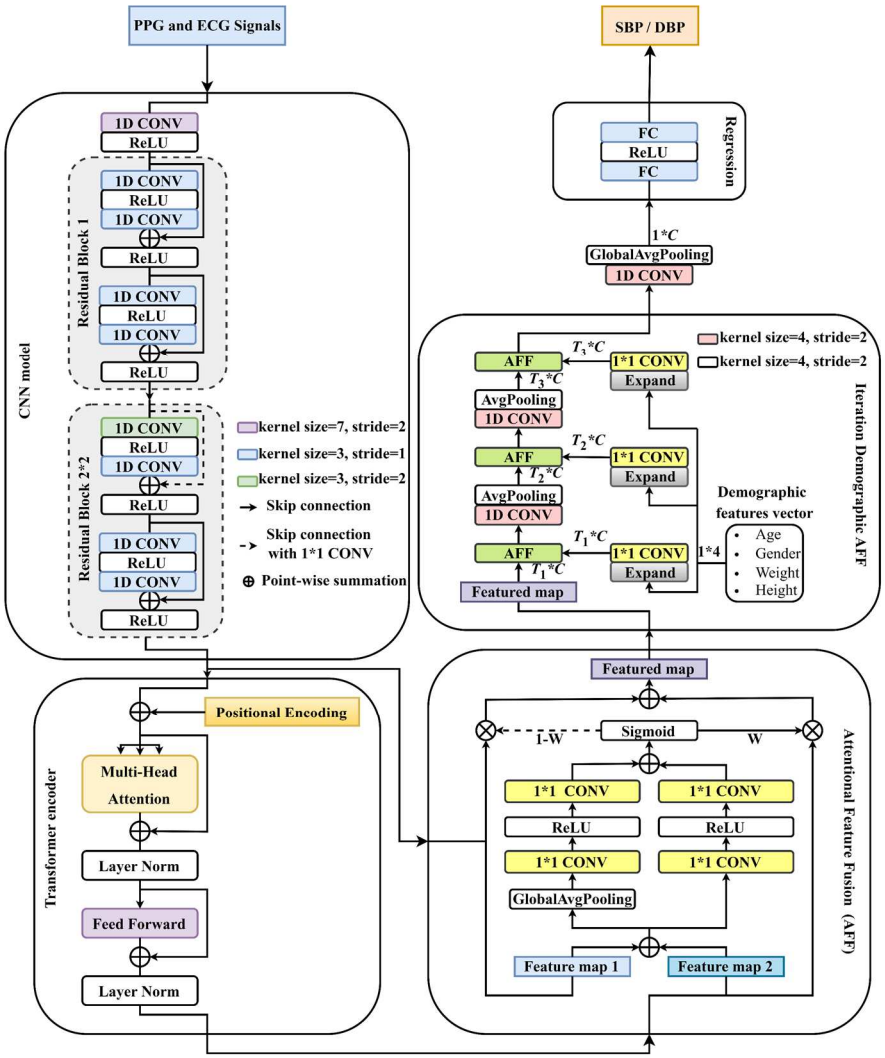


Fig. 1. Schematic diagram of iterative demographic AFF-based CNN and transformer network.

extracted by the Transformer encoder are fused by AFF.

Similarly, demographic information reflects the overall characteristics of a person, whereas features extracted from the PPG and ECG signals reflect the local characteristics of a person's cardiovascular system. Then, AFF is used to fuse demographic information with features extracted from PPG and ECG signals. Since AFF requires the inputs to be consistent in dimension, the demographic vector is expanded to vector sequence alone the time dimension by replicating. Then, the vector sequence goes through a 1×1 convolution to make it consistent in dimension with the PPG and ECG features. Considering that the demographic vector and the PPG and ECG features are greatly inconsistent in dimension, an iterative fusion method of the demographic vector is used. Specifically, the output of AFF is downsampled by convolution and pooling layer, and then the downsampled result is fused with demographic vectors again. The work uses three-step iterative AFF, as shown in Fig. 1.

Finally, the output, which fuses demographic information and cardiovascular signal features, goes through a regression model that includes two fully connected layers and ReLU to get the SBP or DBP estimation value.

III. EXPERIMENTS

A. Training and Testing Schemes

The training and calibration-based testing subsets in PulseDB were used as training and testing datasets, respectively. Each PPG and ECG segment was normalized using z-score normalization. The demographic feature vector includes age, height, and weight stored in their original numerical values and the gender encoded as either 0 or 1. PPG,

ECG segments, and demographic feature vectors were used as input for model training and testing.

The optimization of training was selected as AdamW optimizer with 0.0001 learning rate and 128 batch size. Mean squared error (MSE) loss was selected as the model's loss function. All the model's convolutional and dense layers were initialized using the Kaiming normal distribution. Early-stopping was applied after 10 epochs of less than 0.5 loss reduction on the testing set to avoid overfitting. The model weight at the last epoch yielding the lowest loss was saved for later performance evaluation.

B. Model Performance Evaluation Metrics and Standards

As the BP estimation problem is a classic regression problem, we used mean absolute error (MAE) and standard derivation of error (SDE) as the primary evaluation metrics. In addition, the British Hypertension Society (BHS) and Association for the Advancement of Medical Instrumentation (AAMI) standards were used to evaluate the performance of the proposed model. The BHS standard evaluates the performance of BP estimation based on the cumulative percentages of the estimated BP with absolute errors ≤ 5 mmHg, 10 mmHg, and 15 mmHg, and it grades BP devices into different levels based on thresholds [21]. The AAMI standard requires that the MAE and SDE of the BP measurement method be ≤ 5 mmHg and 8 mmHg, respectively, and the minimum number of participants in the study is 85 [22].

C. Ablation Experiments and Results

To validate the effectiveness of the AFF module and iterative demographic AFF module in our proposed model, we conducted four cases of ablation experiments as follows: (1) using only CNN and transformer encoder module, (2) using CNN, transformer encoder modules, and AFF module to fuse CNN and transformer encoder outputs, (3) using CNN, transformer encoder, AFF modules, and demographic AFF only once, and (4) using CNN, transformer encoder, AFF modules, and iterative demographic AFF, which is the model proposed in this work. To keep the input of the regression model consistent dimensionality, global average pooling was performed along the sequence of feature vectors in the first three models before making BP regression.

As shown in Table II, the results of ablation experiments are consistent in SBP and DBP estimation. Model 4, which includes AFF and iterative demographic modules, achieves the minimum MAEs and SDEs, 3.79±5.33 mmHg and 2.37±3.55 mmHg for SBP and DBP, respectively. Specifically, Model 2 uses the AFF module based on Model 1 to perform feature fusion between the CNN output and the Transformer encoder output. The results show significant improvement in SBP and less in DBP, indicating that with the AFF module, the model can catch features with different scales, thus achieving better

TABLE II

BP ESTIMATION PERFORMANCE WITH DIFFERENT MODULE COMBINATIONS

Model		MAE±SDE (mmHg)	
		SBP	**DBP**
1	CNN, Transformer	5.19±7.48	3.06±4.51
2	CNN, Transformer and AFF	4.60±6.70	3.03±4.53
3	CNN, Transformer, AFF, and demographic AFF once	4.48±6.35	2.62±3.51
4	CNN, Transformer, AFF, and iterative demographic AFF	**3.79±5.33**	**2.37±3.55**

BP estimation. Model 3 further incorporates demographic features AFF only once, based on Model 2. The results show an improvement in the estimation of both SBP and DBP, which suggests that incorporating demographic features through the AFF can better predict BP. Model 4 uses iterative demographic AFF with three times iteration, achieving a better performance than Model 3. This suggests a strong correlation between demographic information and BP, and the iterative fusion method enables the model to better focus on demographic information and more accurately estimate blood pressure.

D. BP Estimation Performance Evaluation Using the BHS and AAMI Standards

Table III displays the specific grade split by BHS and the performance of the proposed model in the estimation of SBP and DBP. The cumulative percentages with absolute errors ≤ 5 mmHg, 10 mmHg, and 15 mmHg are 74.2%, 94.3%, 98.4% in SPB estimation, 90.0%, 98.7%, and 99.7% in DBP estimation. The results show that the performance of the proposed model in SBP and DBP achieves Grade A at the BHS standard.

Table IV shows that the MAE and SDE of the proposed model in SBP estimation are 3.79 mmHg and 5.33 mmHg. The MAE and SDE in DBP estimation are 2.37 mmHg and 3.55 mmHg. Both SBP and DBP estimations meet the AAMI standard.

E. Performance Comparison with Other Feature Fusion Methods

To comprehensively demonstrate the advantages of our proposed method, we compared it with previous methods, and the results are shown in Table V.

In a recent study [14], the author summarized the commonly used feature fusion methods used in demographic feature fusion [11]-[13], and proposed an attentive pooling method, named as CNN-GRU$_{Emb}$ and CNN-GRU$_{Attn}$. As shown in Table V, under the same model framework, the CNN-GRU$_{Attn}$ outperforms CNN-GRU$_{Emb}$ in SBP estimation at PulseDB dataset. The proposed method in this work further improves the

TABLE III

EVALUATION OF BHS STANDARD

		Cumulative Error Percentage		
		≤5 mmHg	≤10 mmHg	≤15 mmHg
Proposed model	**SBP**	74.2%	94.3%	98.4%
	DBP	90.0%	98.7%	99.7%
BHS	**Grade A**	60%	85%	95%
	Grade B	50%	75%	90%
	Grade C	40%	60%	85%

TABLE IV

EVALUATION OF AAMI STANDARD

		MAE (mmHg)	SDE (mmHg)	Number of subjects
Proposed model	**SBP**	3.79	5.33	1293
	DBP	2.37	3.55	
AAMI standard		≤5	≤8	85

TABLE V

COMPARISON OF BP ESTIMATION WITH OTHER FEATURE FUSION METHODS

Method	Dataset	**MAE±SDE (mmHg)**	
		SBP	**DBP**
(Proposed) CNN-Transformer and AFF	PulseDB	**3.79±5.33**	**2.37±3.55**
$CNN\text{-}GRU_{Attn}$ ([14])	PulseDB	4.90±7.00	-
$CNN\text{-}GRU_{Emb}$ ([11]-[13])	PulseDB	5.82±8.26	-

BP estimation in PulseDB dataset. This means the CNN-Transformer architecture and our newly proposed demography feature fusion method based on AFF can better fuse demography features and achieve more accurate BP estimation.

IV. CONCLUSION

Based on the observation that demographic information is highly correlated with blood pressure, this work proposes an iterative demographic AFF-based CNN and transformer network. Through ablation experiments, we demonstrated that the AFF module can fuse features with different scales from the CNN module and transformer encoder module. The iterative demographic AFF module uses multi-times feature fusion to make the model fuse demographic features better. By combining both modules, the model can significantly improve BP estimation.

V. ACKNOWLEDGMENT

This work was supported by Shenzhen Key Technology Program Funding (JSGG20220831103803006).

REFERENCES

[1] N. Townsend, L. Wilson, P. Bhatnagar, K. Wickramasinghe, M. Rayner, and M. Nichols, "Cardiovascular disease in Europe: epidemiological update 2016," Eur. Heart J., vol. 37, no. 42, pp. 3232–3245, 2016.

[2] W.-H. Lin, F. Chen, Y. Geng, N. Ji, P. Fang, and G. Li, "Towards accurate estimation of cuffless and continuous blood pressure using multi-order derivative and multivariate photoplethysmogram features," Biomed. Signal Process. Control, vol. 63, p. 102198, Jan. 2021.

[3] S. G. Khalid, H. Liu, T. Zia, J. Zhang, F. Chen, and D. Zheng, "Cuffless blood pressure estimation using single channel photoplethysmography: a two-step method," IEEE Access, vol. 8, pp. 58146–58154, 2020.

[4] M. H. Chowdhury et al., "Estimating blood pressure from the photoplethysmogram signal and demographic features using machine learning techniques," Sensors, vol. 20, no. 11, p. 3127, Jun. 2020.

[5] P. Yao et al., "Multi-dimensional feature combination method for continuous blood pressure measurement based on wrist PPG sensor," IEEE J. Biomed. Health Inform., vol. 26, no. 8, pp. 3708–3719, Aug. 2022.

[6] G. Thambiraj, U. Gandhi, U. Mangalanathan, V. J. M. Jose, and M. Anand, "Investigation on the effect of Womersley number, ECG and PPG features for cuff less blood pressure estimation using machine learning," Biomed. Signal Process. Control, vol. 60, p. 101942, 2020.

[7] W.-H. Lin, D. Zheng, G. Li, H. Zhou, and F. Chen, "Investigation on pulse wave forward peak detection and its applications in cardiovascular health," IEEE Trans. Biomed. Eng., vol. 69, no. 2, pp. 700–709, Feb. 2022.

[8] M. E. Safar, J. Blacher, and P. Jankowski, "Arterial stiffness, pulse pressure, and cardiovascular disease—Is it possible to break the vicious circle?," Atherosclerosis, vol. 218, no. 2, pp. 263–271, Oct. 2011.

[9] W.-H. Lin, D. Zheng, G. Li, and F. Chen, "Age-related changes in blood volume pulse wave at fingers and ears," IEEE J. Biomed. Health Inform., pp. 5070- 5080, 2024.

[10] W.-H. Lin, X. Li, Y. Li, G. Li, and F. Chen, "Investigating the physiological mechanisms of the photoplethysmogram features for blood pressure estimation," Physiol. Meas., vol. 41, no. 4, p. 044003, May 2020.

[11] Y. Zhang, X. Ren, X. Liang, X. Ye, and C. Zhou, "A refined blood pressure estimation model based on single channel Photoplethysmography," IEEE J. Biomed. Health Inform., vol. 26, no. 12, pp. 5907–5917, Dec. 2022.

[12] D. Lee et al., "Beat-to-beat continuous blood pressure estimation using bidirectional long short-term memory network," Sensors, vol. 21, no. 1, p. 96, Dec. 202.

[13] S. Yang, Y. Zhang, S.-Y. Cho, R. Correia, and S. P. Morgan, "Non-invasive cuff-less blood pressure estimation using a hybrid deep learning model," Opt. Quantum Electron., vol. 53, no. 2, p. 93, Feb. 2021.

[14] W. Wang, P. Mohseni, K. L. Kilgore, and L. Najafizadeh, "Demographic information fusion using attentive pooling in CNN-GRU model for systolic blood pressure estimation," in 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia: IEEE, Jul. 2023, pp. 1–4.

[15] Z.-D. Liu et al., "HGCTNet: Handcrafted feature-guided CNN and Transformer network for wearable cuffless blood pressure measurement," IEEE J. Biomed. Health Inform., vol. 28, no. 7, pp. 3882–3894, Jul. 2024.

[16] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 3559–3568.

[17] W. Wang, P. Mohseni, K. L. Kilgore, and L. Najafizadeh, "PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods," Front. Digit. Health, vol. 4, p. 1090854, Feb. 2023.

[18] A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," Sci. Data, vol. 3, no. 1, p. 160035, May 2016.

[19] H.-C. Lee, Y. Park, S. B. Yoon, S. M. Yang, D. Park, and C.-W. Jung, "VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients," Sci. Data, vol. 9, no. 1, p. 279, Jun. 2022.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778. Accessed: Jul. 15, 2024.

[21] E. O'Brien et al., "The British Hypertension Society protocol for the evaluation of blood pressure measuring devices," J Hypertens, vol. 11, no. Suppl 2, pp. S43–S62, 1993.

[22] G. S. Stergiou et al., "A universal standard for the validation of blood pressure measuring devices: Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) collaboration statement," Hypertension, vol. 71, no. 3, pp. 368–374, Mar. 2018.

Performance Evaluation of Acoustic Echo and Noise Canceller with Variable-Step-Size Shared-Error NLMS Algorithm under Double-Talk Conditions

Kenta Iwai* and Takanobu Nishiura*
* Ritsumeikan University, Japan
E-mail: {iwai18sp@fc}{nishiura@is}.ritsumei.ac.jp

***Abstract*—In this paper, we present the evaluation of the acoustic echo and noise reduction performance of the variable-step-size shared-error normalized least mean square (VSS-SENLMS) algorithm under double-talk conditions. The VSS-SENLMS algorithm was developed for the acoustic echo and noise canceller (AENC) to simultaneously optimize both the acoustic echo canceller and the adaptive noise canceller by variable step size parameters. Our previous study showed that the VSS-SENLMS-based AENC achieved reductions of the acoustic echo and background noise under the single-talk condition with nonstationary background noise. Hence, we evaluate the acoustic echo and noise reduction performance of the VSS-SENLMS-based AENC under double-talk conditions. By computer simulation, this system is found to diverge under the double-talk condition owing to the poor estimation accuracy of variances for variable step size parameters. On the other hand, this system maintains the acoustic echo and noise reduction performance without the double-talk detector when the variances used for the variable step size parameters are accurately estimated.**

I. Introduction

In this paper, we present the evaluation of the acoustic echo and noise reduction performance of the variable-step-size shared-error normalized least mean square (VSS-SENLMS) algorithm [1] under double talk conditions [2]. This acoustic echo and noise canceller (AENC) consists of the acoustic echo canceller (AEC) with the time-domain adaptive digital filter and Widrow's adaptive noise canceller (ANC) [3]. Then, the AEC and ANC are updated by the same error signal. This adaptive algorithm is called a shared-error normalized least mean square (SENLMS) algorithm [4]. The SENLMS algorithm uses fixed step size parameters to update the AEC and ANC. Hence, the performance of acoustic echo and noise reduction degrades under nonstationary noise and double-talk conditions.

In our previous study [1], we proposed the VSS-SENLMS algorithm to improve the performance of acoustic echo and noise reduction under the nonstationary noise conditions. The VSS-SENLMS algorithm is derived from the *a posteriori* error, and variable step size parameters are updated at each time step. Simulation results showed that the VSS-SENLMS algorithm achieved acoustic echo and noise reduction under the nonstationary noise condition. However, we did not conduct the evaluation under double-talk conditions.

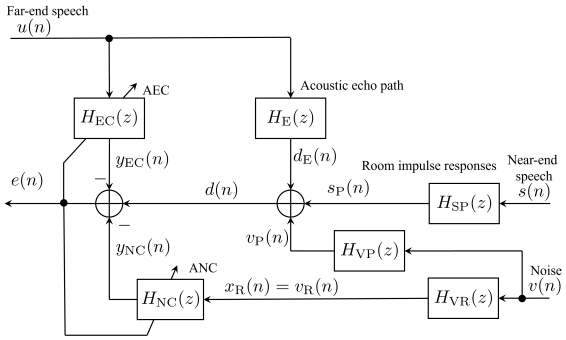


Fig. 1: Block diagram of acoustic echo and noise canceller with SENLMS algorithm.

An AEC generally suffers from the double-talk problem [2], [5] and is combined with a double-talk detector [6]–[9]. Hence, we evaluate the performance of the VSS-SENLMS algorithm under double-talk conditions. Here, the VSS-SENLMS algorithm uses the estimation of the variances of the error signal and noises. Therefore, we investigate the effectiveness of the estimation accuracy of each variance used to determine variable step size parameters under double-talk conditions.

II. VSS-SENLMS algorithm

A. SENLMS algorithm [4]

Fig. 1 shows the block diagram of the AENC based on the SENLMS algorithm. This AENC consists of the AEC with the time-domain adaptive digital filter and Widrow's ANC [3], and updates them by the SENLMS algorithm, which uses the shared-error signal $e(n)$. By using the SENLMS algorithm, the AEC and ANC are simultaneously optimized.

To derive the SENLMS algorithm, the error signal in the normalized least mean square (NLMS) algorithm [10]–[12] is replaced by the shared-error signal $e(n) = d(n) - y_{\mathrm{EC}}(n) - y_{\mathrm{NC}}(n)$, where $d(n)$ is the microphone input of the primary microphone, $y_{\mathrm{EC}}(n)$ is the output signal of the AEC, and $y_{\mathrm{NC}}(n)$ is the output signal of the ANC. Then, the shared-error signal $e(n)$ is normalized by the norm of the input signal vector. The update equation is obtained by the partial derivation of the normalized shared-error signal. The update equations of

the SENLMS algorithm are shown as

$$\mathbf{h}_{\mathrm{EC}}(n+1) = \mathbf{h}_{\mathrm{EC}}(n) + \alpha_{\mathrm{EC}} \frac{e(n)\mathbf{u}(n)}{\|\mathbf{u}(n)\|^2 + \beta_{\mathrm{EC}}}, \quad (1)$$

$$\mathbf{h}_{\mathrm{NC}}(n+1) = \mathbf{h}_{\mathrm{NC}}(n) + \alpha_{\mathrm{NC}} \frac{e(n)\mathbf{x}_{\mathrm{R}}(n)}{\|\mathbf{x}_{\mathrm{R}}(n)\|^2 + \beta_{\mathrm{NC}}}, \quad (2)$$

where $\mathbf{h}_{\mathrm{EC}}^{\mathrm{T}}(n)$ and $\mathbf{h}_{\mathrm{NC}}^{\mathrm{T}}(n)$ are the filter coefficient vectors of the AEC and ANC, respectively, $\mathbf{u}(n)$ and $\mathbf{x}_{\mathrm{R}}(n)$ are the input signal vectors, and $\|\cdot\|$ represents the norm of the vector. Also, α_{EC} and α_{NC} are the step size parameters, and β_{EC} and β_{NC} the regularization parameters of the AEC and ANC, respectively.

As seen in (1) and (2), the SENLMS algorithm utilizes the fixed step size parameters α_{EC} and α_{NC}. Hence, the step size parameters α_{EC} and α_{NC} should be manually adjusted depending on the type of background noise to avoid divergence. Since the shared-error signal $e(n)$ contains both the AEC output $y_{\mathrm{EC}}(n)$ and the ANC output $y_{\mathrm{NC}}(n)$, the entire system diverges when the AEC or ANC diverges.

B. VSS-SENLMS algorithm [1]

To solve the above problem, we proposed the VSS-SENLMS algorithm [1]. The VSS-SENLMS algorithm is based on the nonparametric VSS-NLMS (NP-VSS-NLMS) algorithm [13], and was derived by minimizing the mean square value of the *a posteriori* error signal.

The *a posteriori* error signal $\varepsilon(n)$ is represented as

$$\begin{aligned}\varepsilon(n) = d_{\mathrm{E}}(n) - \mathbf{h}_{\mathrm{EC}}^{\mathrm{T}}(n+1)\mathbf{u}(n) \\ + v_{\mathrm{P}}(n) - \mathbf{h}_{\mathrm{NC}}^{\mathrm{T}}(n+1)\mathbf{x}_{\mathrm{R}}(n) + s_{\mathrm{P}}(n),\end{aligned} \quad (3)$$

where $d_{\mathrm{E}}(n)$ is the acoustic echo, $v_{\mathrm{P}}(n)$ is the background noise, and $s_{\mathrm{P}}(n)$ is the near-end speech. Here, the update equations (1) and (2) are rewritten as

$$\mathbf{h}_{\mathrm{EC}}(n+1) = \mathbf{h}_{\mathrm{EC}}(n) + \alpha_{\mathrm{EC}}(n)e(n)\mathbf{u}(n), \quad (4)$$

$$\mathbf{h}_{\mathrm{NC}}(n+1) = \mathbf{h}_{\mathrm{NC}}(n) + \alpha_{\mathrm{NC}}(n)e(n)\mathbf{x}_{\mathrm{R}}(n), \quad (5)$$

where $\alpha_{\mathrm{EC}}(n)$ and $\alpha_{\mathrm{NC}}(n)$ are the variable step size parameters for the AEC and ANC, respectively. From (3)–(5), the mean square value of the *a posteriori* error signal is obtained as

$$\begin{aligned}\mathbb{E}\left[\varepsilon^2(n)\right] = \sigma_e^2(n) \\ \left[1 - \|\mathbf{u}(n)\|^2\alpha_{\mathrm{EC}}(n) - \|\mathbf{x}_{\mathrm{R}}(n)\|^2\alpha_{\mathrm{NC}}(n)\right]^2,\end{aligned} \quad (6)$$

where $\mathbb{E}[\cdot]$ is the expectation operator and $\sigma_e^2(n) = \mathbb{E}\left[e^2(n)\right]$. Here, $\|\mathbf{u}(n)\|^2$ and $\|\mathbf{x}_{\mathrm{R}}(n)\|^2$ are treated as the constants. Since the *a posteriori* error signal $\varepsilon(n)$ is obtained using $\mathbf{h}_{\mathrm{EC}}(n+1)$ and $\mathbf{h}_{\mathrm{NC}}(n+1)$,

$$\mathbb{E}\left[\varepsilon^2(n)\right] = \mathbb{E}\left[s_{\mathrm{P}}^2(n)\right] = \sigma_s^2(n). \quad (7)$$

From (7), the constraint of the step size parameters of the VSS-SENLMS algorithm is derived as

$$\|\mathbf{u}(n)\|^2\alpha_{\mathrm{EC}}(n) + \|\mathbf{x}_{\mathrm{R}}(n)\|^2\alpha_{\mathrm{NC}}(n) = 1 - \sqrt{\frac{\sigma_s^2(n)}{\sigma_e^2(n)}}. \quad (8)$$

Here, the left side of (8) is the function of $\alpha_{\mathrm{EC}}(n)$ and $\alpha_{\mathrm{NC}}(n)$. Since there is only (8) with the two variables $\alpha_{\mathrm{EC}}(n)$ and $\alpha_{\mathrm{NC}}(n)$, the desired variable step size parameters cannot be obtained using (8).

In [1], the NP-VSS-NLMS algorithm [13] is adopted to determine $\alpha_{\mathrm{EC}}(n)$ since it was developed for the AEC. The variable step size parameter of the AEC is determined as

$$\alpha_{\mathrm{EC}}(n) = \frac{\alpha_{\mathrm{VSS,EC}}(n)}{\|\mathbf{u}(n)\|^2}, \quad (9)$$

$$\alpha_{\mathrm{VSS,EC}}(n) = \frac{1}{\|\mathbf{u}(n)\|^2}\left[1 - \sqrt{\frac{\sigma_{v,\mathrm{EC}}^2(n)}{\sigma_e^2(n)}}\right], \quad (10)$$

$$\sigma_{v,\mathrm{EC}}^2(n) = \mathbb{E}\Big[\{v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n) + s_{\mathrm{P}}(n)\}^2\Big], \quad (11)$$

where $v_{\mathrm{P}}(n)$, $y_{\mathrm{NC}}(n)$, and $s_{\mathrm{P}}(n)$ are treated as the disturbances of the AEC. By using (9) and (8), $\alpha_{\mathrm{NC}}(n)$ can be determined as

$$\alpha_{\mathrm{NC}}(n) = \frac{\alpha_{\mathrm{VSS,NC}}(n)}{\|\mathbf{x}_{\mathrm{R}}(n)\|^2}, \quad (12)$$

$$\alpha_{\mathrm{VSS,NC}}(n) = \frac{1}{\|\mathbf{x}_{\mathrm{R}}(n)\|^2}\left[1 - \sqrt{\frac{\sigma_s^2(n)}{\sigma_e^2(n)}} - \alpha_{\mathrm{VSS,EC}}(n)\right]. \quad (13)$$

As in (9) and (12), $\sigma_e^2(n)$, $\sigma_s^2(n)$, $\sigma_{v,\mathrm{EC}}^2(n)$, and $\sigma_e^2(n)$ should be estimated to determine $\alpha_{\mathrm{EC}}(n)$ and $\alpha_{\mathrm{NC}}(n)$. According to [13], $\sigma_e^2(n)$ can be estimated as

$$\sigma_e^2(n) \approx \lambda\sigma_e^2(n-1) + (1-\lambda)\,e^2(n), \quad (14)$$

where $\lambda (0 << \lambda < 1)$ is the forgetting factor. Here, $\sigma_{v,\mathrm{EC}}^2(n)$ and $\sigma_s^2(n)$ cannot be estimated by a recursive equation like (14) since $v_{\mathrm{EC}}(n) (= v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n) + s_{\mathrm{P}}(n))$ and $s_{\mathrm{P}}(n)$ cannot be directly obtained. To solve this problem, $\sigma_{v,\mathrm{EC}}^2(n)$ and $\sigma_s^2(n)$ are estimated using the cross-correlation vector as [14]

$$\sigma_{v,\mathrm{EC}}^2(n) \approx \sigma_e^2(n) - \frac{\|\mathbf{r}_{eu}(n)\|^2}{\sigma_u^2(n)}, \quad (15)$$

$$\sigma_s^2(n) \approx \sigma_e^2(n) - \frac{\|\mathbf{r}_{eu}(n)\|^2}{\sigma_u^2(n)} - \frac{\|\mathbf{r}_{ex}(n)\|^2}{\sigma_x^2(n)}, \quad (16)$$

$$\sigma_u^2(n) \approx \lambda\sigma_u^2(n-1) + (1-\lambda)\,u^2(n), \quad (17)$$

$$\sigma_x^2(n) \approx \lambda\sigma_x^2(n-1) + (1-\lambda)\,x_{\mathrm{R}}^2(n), \quad (18)$$

$$\mathbf{r}_{eu}(n) \approx \lambda\mathbf{r}_{eu}(n-1) + (1-\lambda)\,e(n)\mathbf{u}(n), \quad (19)$$

$$\mathbf{r}_{ex}(n) \approx \lambda\mathbf{r}_{ex}(n-1) + (1-\lambda)\,e(n)\mathbf{x}_{\mathrm{R}}(n). \quad (20)$$

As seen in (9) and (12), the variable step size parameters of the VSS-SENLMS algorithm depend on each variance obtained using (14)–(20). In [1], the VSS-SENLMS algorithm was not evaluated under double-talk conditions but was evaluated in a nonstationary noise environment under the single-talk condition. In the next section, we will show the evaluation results under the double-talk condition.

TABLE I: Simulation conditions.

Far-end speech	Female speech (4507)[17]
Near-end speech	Male speech (1188)[17]
Background noise	White noise
Signal-to-noise ratio (SNR)	-5 dB
Sampling frequency	16 kHz
Tap length of each path	1024
Tap length of ADF of AEC	1024
Tap length of ADF of ANC	1024
Regularization parameter	1.0×10^{-6}

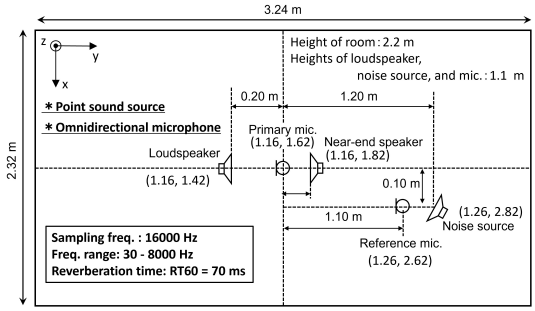
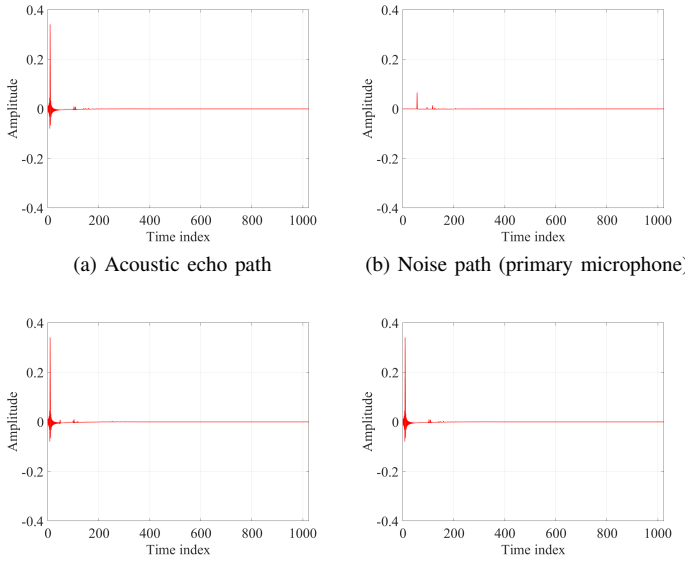


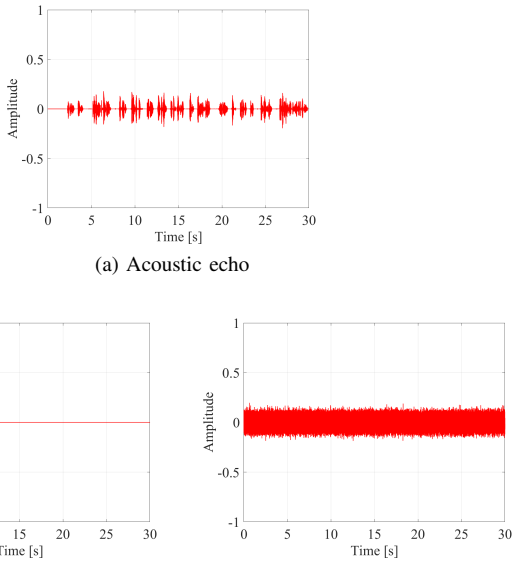
Fig. 2: Arrangement for RIR generation.



(a) Acoustic echo path (b) Noise path (primary microphone)

(c) Noise path (reference microphone) (d) Near-end speech path

Fig. 3: Impulse responses of each path.



(a) Acoustic echo

(b) Near-end speech (c) Background noise

Fig. 4: Time waveforms of each signal.

III. Computer Simulation

The computer simulation was conducted to evaluate the performance of the VSS-SENLMS algorithm under the double-talk condition. In this simulation, the variances $\sigma_s^2(n)$ and $\sigma_{v,\mathrm{EC}}^2(n)$ were estimated using not only (15) and (16) but also the following recursive equations:

$$\sigma_{v,\mathrm{EC}}^2(n) \approx \lambda\sigma_{v,\mathrm{EC}}^2(n-1) + (1-\lambda)v_{\mathrm{EC}}^2(n), \quad (21)$$

$$\sigma_s^2(n) \approx \lambda\sigma_s^2(n-1) + (1-\lambda)s_{\mathrm{P}}^2(n). \quad (22)$$

Hereafter, the results obtained using (15) and (16), and (21) and (22) are shown as "C-Corr" and "Rec," respectively. Equations (21) and (22) accurately estimate each variance since $v_{\mathrm{EC}}(n)$ and $s_{\mathrm{P}}(n)$ are directly used [1].

In this simulation, each impulse response was generated by a room impulse response generator [15]. The arrangement of the sound sources and microphones is shown in Fig. 2, and the generated impulse responses are shown in Fig. 3. Here, the direct-to-reverberation ratio (DRR) [16] was set to about 26 dB. Also, the near-end speech was 6 dB larger than the acoustic echo. Simulation conditions are shown in Table I. The forgetting factor λ was set to $\lambda = 1 - 1/2N$ with the tap length of the adaptive filter being N. The time waveforms of the acoustic echo, background noise, and near-end speech are shown in Fig. 4. Double-talk occurred from about 10.0 to 12.5 s.

The time waveforms of the error signals are shown in Fig. 5. From Fig. 5, the VSS-SENLMS algorithm "C-Corr" diverges after the double-talk period. On the other hand, the VSS-SENLMS algorithm "Rec" does not diverge after the double-talk period and maintains the performance of the acoustic echo and noise reduction. Here, the dynamics of the step size parameters of "C-Corr" and "Rec" are shown in Figs. 6 and 7, respectively. From Figs. 6 and 7, the step size parameter of the AEC for the VSS-SENLMS algorithm, "C-Corr," is seen to be almost one, particularly, after 10 s. In other words, the AEC is still updated during double-talk even though this is undesirable [9], [18]. On the other hand, the VSS-SENLMS algorithm "Rec" retains the step size parameters with zero and stops the update of the AEC and ANC. Therefore, the VSS-SENLMS algorithm does not require the double-talk detector if each variance can be estimated accurately.

The dynamics of each variance is shown in Fig. 8. Fig. 8

[1]This condition cannot be realized in the actual situation.

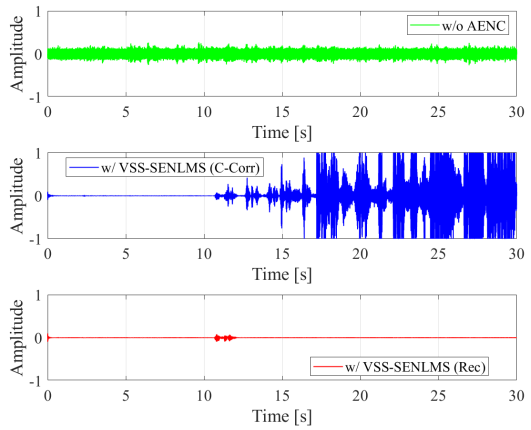
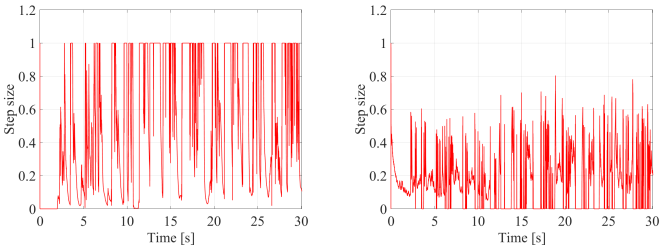
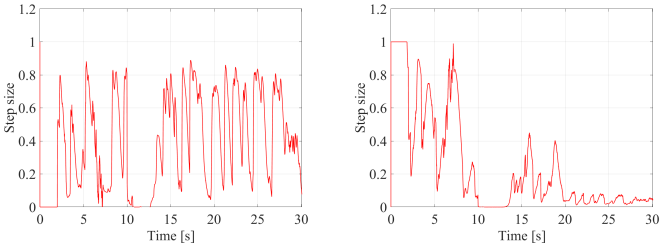


Fig. 5: Time waveform of error signal.



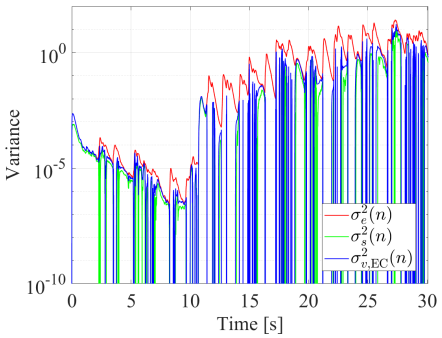
(a) AEC (b) ANC

Fig. 6: Dynamics of each step size (C-Corr).

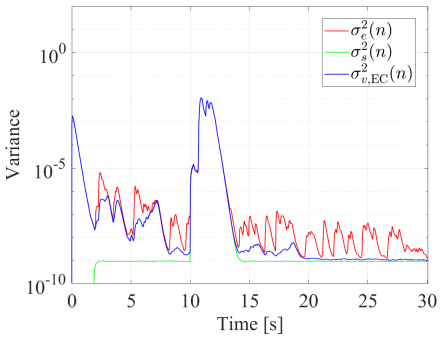


(a) AEC (b) ANC

Fig. 7: Dynamics of each step size (Rec).



(a) Using cross-correlation ($\sigma_e^2(n)$ is estimated using (14))



(b) Using recursive process

Fig. 8: Dynamics of each variance.

shows that each variance is different from the others depending on the estimation method for each variance. For example, $\sigma_{v,\mathrm{EC}}^2(n)$ is generally similar to $\sigma_e^2(n)$ during the nonspeech period. In other words, the error signal $e(n)$ and the disturbance of AEC $v_{\mathrm{EC}}(n)$ become

$$e(n) = v_{\mathrm{EC}}(n) = v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n). \tag{23}$$

Then, $\sigma_e^2(n) = \sigma_{v,\mathrm{EC}}^2(n) = \sigma_s^2(n)$. On the other hand, during the double-talk period, the error signal $e(n)$ and the disturbance of AEC $v_{\mathrm{EC}}(n)$ become

$$e(n) = d_{\mathrm{E}}(n) - y_{\mathrm{EC}}(n) + v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n) + s_{\mathrm{P}}(n), \tag{24}$$

$$v_{\mathrm{EC}}(n) = v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n) + s_{\mathrm{P}}(n). \tag{25}$$

If the AEC and ANC converge, i.e., $d_{\mathrm{E}}(n) - y_{\mathrm{EC}}(n) \approx 0$ and $v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n) \approx 0$, then $\sigma_e^2(n) \approx \sigma_{v,\mathrm{EC}}^2(n)$. As shown in Fig. 8 (b), the variances estimated using the recursive equation are similar. However, as shown in Fig. 8 (a), the variances estimated by cross-correlation differ from each other owing to the estimation error, which may be caused by the second term of (15) and the third term of (16). The estimation error affects the misadjustment of the step size parameters shown in Fig. 6.

From the above results, we observe that the algorithm can control the step size parameters during double-talk with accurately estimated variances without the double-talk detector, although the VSS-SENLMS algorithm diverges after the double-talk period owing to the estimation error of the variances.

IV. Conclusions

In this paper, we evaluated the performance of the VSS-SENLMS algorithm under double-talk conditions. From computer simulation results, the VSS-SENLMS algorithm was found to diverge when each variance was not accurately estimated. On the other hand, the VSS-SENLMS algorithm maintained the performance of acoustic echo and noise reduction when each variance was accurately estimated. In this case, the double-talk detector was not required. Hence, it is important to investigate the estimation method for each variance.

Acknowledgment

This work was supported by JSPS KAKENHI (JP21K17773).

REFERENCES

[1] K. Iwai and T. Nishiura, "Step size control of shared-error normalized least mean square algorithm for acoustic echo and noise canceller," in *Proc. 2023 Asia-Pacific Signal Inf. Process. Association Annual Summit Conf. (APSIPA ASC)*, 2023, pp. 1150–1154.

[2] M. Z. Ikram, "Blind source separation and acoustic echo cancellation: A unified framework," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2012, pp. 1701–1704.

[3] B. Widrow, J. Glover, J. McCool, *et al.*, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.

[4] K. Iwai and T. Nishiura, "Acoustic echo and noise canceller using shared-error normalized least mean square algorithm," in *Proc. 2022 Asia-Pacific Signal Inf. Process. Association Annual Summit Conf. (APSIPA ASC)*, 2022, pp. 281–285.

[5] S. Ohta, Y. Kajikawa, and Y. Nomura, "Acoustic echo cancellation using sub-adaptive filter," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2007, pp. I-85–I-88.

[6] D. L. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Trans. Commun.*, vol. 26, no. 5, pp. 647–653, 1978.

[7] J. Benesty, D. Morgan, and J. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech, Audio Process.*, vol. 8, no. 2, pp. 168–172, 2000.

[8] C. Paleologu, J. Benesty, T. Gaensler, and S. Ciochină, "Class of double-talk detectors based on the Holder inequality," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2011, pp. 425–428.

[9] M. Z. Ikram, "Double-talk detection in acoustic echo cancellers using zero-crossings rate," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 1121–1125.

[10] S. O. Haykin, *Adaptive Filter Theory*, Fourth. Upper Saddle River, NJ: Prentice Hall, 2001.

[11] A. Zerguine, "Convergence behavior of the normalized least mean fourth algorithm," in *Proc. Conf. Record Thirty-Fourth Asilomar Conf. Signals Syst. Comput.*, 2000, pp. 275–278 vol.1.

[12] E. Eweda and A. Zerguine, "A normalized least mean fourth algorithm with improved stability," in *Proc. Conf. Record Forty Fourth Asilomar Conf. Signals Syst. Comput.*, 2010, pp. 1002–1005.

[13] J. Benesty, H. Rey, L. Vega, and S. Tressens, "A nonparametric VSS NLMS algorithm," *IEEE Signal Process. Lett.*, vol. 13, no. 10, pp. 581–584, 2006.

[14] M. A. Iqbal and S. L. Grant, "Novel variable step size NLMS algorithms for echo cancellation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2008, pp. 241–244.

[15] E. Habets, *Room impulse response generator*, the International Audio Laboratories Erlangen, 2010.

[16] J. Eaton, A. H. Moore, P. A. Naylor, and J. Skoglund, "Direct-to-Reverberant Ratio estimation using a null-steered beamformer," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 46–50.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 5206–5210.

[18] S. Kim, H. Kwon, K. Bae, K. Byun, and K. Kim, "Performance improvement of double-talk detection algorithm in the acoustic echo canceller," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2001, 3249–3252 (vol.5).

Optimizing Computational Efficiency: In-Memory Computing with Dynamic Switching

Kun-Lin Tsai and Chao-Ting Huang
Dept. of Electrical Engineering, Tunghai University, Taichung, Taiwan (R.O.C.)
E-mail: kltsai@thu.edu.tw; kimihuang4509@gmail.com

Abstract — **Modern computing architectures mainly utilize the Von-Neumann architecture. However, the increasing demand for processing vast volumes of data worsens the Von-Neumann bottleneck, which results in significant energy consumption and computational latency. In-memory computing provides an effective solution for Von-Neumann bottleneck by performing fundamental logical or arithmetic operations inside the memory directly. In this paper, an in-memory computing architecture with dynamic switching function is proposed to provide multiple logical or arithmetic operations. The 8+T SRAM cell and pass transistor logic circuits are utilized in the proposed in-memory computing architecture. Specifically, we introduce the design of a dynamic adder-subtractor capable of switching between addition and subtraction, thereby enhancing the efficiency of in-memory computing.**

I. Introduction

The widespread application of modern artificial intelligence and deep learning underscores the inefficiencies of the traditional Von Neumann architecture [1], particularly in managing high-speed, large-volume data transfers. The separation of the Arithmetic Logic Unit (ALU) and Memory Unit necessitates repeated data transfers, resulting in the well-documented Von Neumann bottleneck [2], which reduces system throughput and increases power consumption and performance degradation.

In-Memory Computing (IMC) [3-5] addresses this bottleneck by performing logical or arithmetic operations within memory units using simple circuits, thus reducing the frequency of data transfers. This paper aims to efficiently execute the six basic logical operations (AND, NAND, OR, NOR, XOR, XNOR) and two arithmetic operations (full adder and full subtractor) using 8+T Static Random Access Memory (SRAM) [8-11]. This architecture leverages Read Bit Lines (RBL, RBLB) for AND and NOR results and uses inverters for NAND and NOR results. Additionally, XOR and XNOR operations are achieved using two transistors, as referenced in previous studies [17-19]. For arithmetic operations, we adopt the full adder architecture [13-14], utilizing two sets of four transistors for Sum (S) and Carry-out (Co) outputs. By comparing the truth tables of the full adder and full subtractor, the design is enhanced to produce Difference (D) and Borrow-out (Bo) outputs for the full subtractor. Dynamic input switching allows for quick retrieval of full adder and full subtractor values within the memory, enabling multiple operations within a single memory unit.

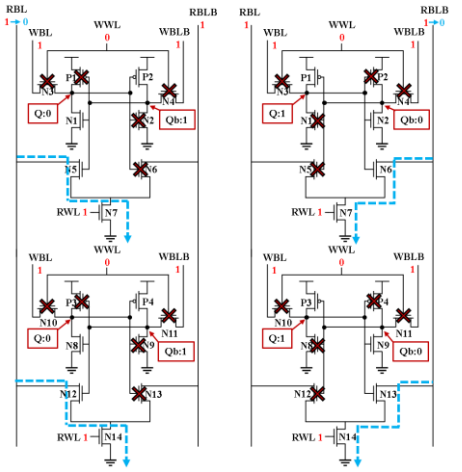


Fig. 1. The 8+T SRAM operations in case (0,0) and case (1,1)

II. 8+T SRAM and Pass Transistor Operations

Many researchers have proposed various ideas in the field of IMC. Agrawal *et al*. [10] have suggested using two cells of 8+T SRAM in series to obtain logical operation results. Fig. 1 shows the 8+T SRAM operations in case (0,0) and case (1,1). In the left part of Fig. 1, both cells are written with 0, and the 8+T SRAM begins read mode, activating RWL. Q1 and Q2 are 0, causing N6 and N13 to be cut off, keeping RBLB at a high potential. Qb1 and Qb2 are 1, causing N5 and N12 to conduct and discharge RBL to GND. Therefore, in case (0,0), RBL drops to low voltage, i.e., logic 0, and RBLB remains at high voltage, i.e., logic 1. The right diagram in Fig. 1 operates on the same principle as the left diagram described above. The conclusion is RBL remains at high voltage, i.e., logic 1, and RBLB also remains at high voltage.

In Fig. 2, the operating principles described above also apply. Therefore, for case (0,1) and case (1,0), the conclusion is RBL will drop to low voltage, and RBLB will also drop to low voltage. Table 1 lists the truth table of AND and NOR functions, it can be observed that RBL exhibits AND logic while RBLB exhibits NOR logic.

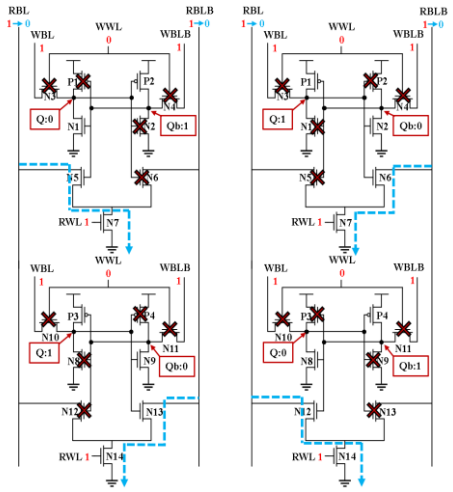


Fig. 2 The 8+T SRAM operations in case (0,1) and case (1,0)

Table 1 Logical Behavior Table of 8+T SRAM

Q1	Q2	RBL (AND)	RBLB (NOR)
0	0	0	1
0	1	0	0
1	0	0	0
1	1	1	0

This means that if an inverter is connected the resulting output will be NAND logic. Conversely, if an inverter is connected next to RBLB, the resulting output will be OR logic. In this manner, all four basic logical operations can be quickly obtained within the memory using RWL control.

Singh *et al.* [13] used Pass Transistor Logic (PTL) to generate the Sum and Cout of a full adder. This architecture first outputs the input values as XOR logic and connects them to an inverter to generate XNOR. These results are then used to produce the output of the full adder. Compared to CMOS, PTL uses fewer transistors to create various logic gates. In PTL, transistors act as switches between circuit nodes to determine logic levels without directly connecting to a voltage source, often resulting in lower output voltage. Multiple transistors in series usually require a stabilized power supply to restore the signal voltage.

For instance, Fig. 3 compares CMOS and PTL architectures for an AND gate. The PTL significantly reduces the number of required transistors. However, the XOR generation architecture in the referenced paper has flaws. Consequently, this study adopts only the full adder architecture from that reference, as illustrated in Fig. 4. Utilizing PTL, this design completes the full adder operation and significantly reduces transistor count compared to traditional implementations. Furthermore, PTL's design inherently lowers power consumption due to its lack of reliance on a stable voltage source.

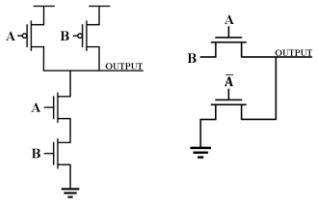


Fig. 3 AND Logic Gate in CMOS and PTL Architectures

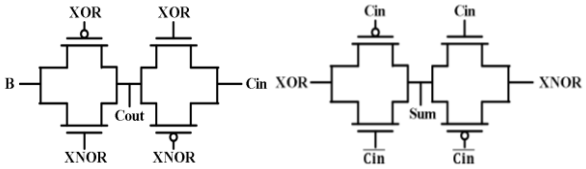


Fig. 4 Illustration of Full Adder Implementation Using PTL

III. Dynamic Switch in Memory Computing

This paper selects 8+T SRAM for its unique characteristics, enabling the generation of multiple logic gates with a few inverters. Separating the read and write lines addresses the data inversion issue found in 6T SRAM. While 8+T SRAM inherently produces four logic gates (AND, NAND, OR, NOR), most arithmetic operations necessitate XOR and XNOR gates. Referencing [7], this study generates XOR and XNOR gates by connecting two transistors to the Q1 and Q2 points in two cells.

These XOR and XNOR gates are then integrated into the PTL adder architecture to yield a two-bit adder output. By comparing the truth tables of addition and subtraction, the architecture can be slightly modified to produce subtraction outputs. Finally, incorporating input switching functionality allows for dynamic switching between addition, subtraction, and logic operations within the memory.

A. Logical Operations

According to one of the papers discussed in the Related Studies, we can understand that through the discharge characteristics of 8+T SRAM, an AND logic gate result can be obtained at the RBL end, and a NOR logic gate result can be obtained at the RBLB end. However, if an inverter is added to each end, NAND and OR results can be achieved, as shown in Fig. 5. Additionally, the extra inverter acts as a buffer, making the output logic more stable.

Next, we will generate XOR and XNOR based on the methodology described in reference [7]. As shown in Fig. 6, connecting the Q1, Q2, and Qb1 points from Fig. 5 using the CMOS inverter architecture yields the XNOR result. The circuit logic is as follows: when Q2 is 1, the NMOS transistor conducts and the PMOS transistor is cut off, making Q1 represent the XNOR outcome. Conversely, when Q2 is 0, the PMOS transistor conducts and the NMOS transistor is cut off, making Qb1 represent the XNOR outcome. An additional

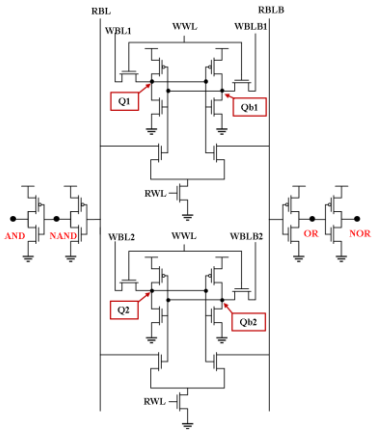


Fig.5 8+T SRAM Generating AND, NAND, OR, NOR

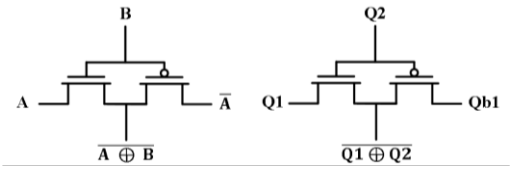


Fig.6 The XNOR integrated with the 8+T SRAM circuit.

inverter is then used to produce the XOR output. The derived truth table for XOR and XNOR logic with respect to Q1 and Q2 follows from this circuit logic.

B. Addition and Subtraction Operation

Upon completing the six basic logical operations, we utilize the previously generated XOR and XNOR results to perform addition and subtraction operations. We first examine the circuit logic for Cout, as illustrated in Fig. 7. When the XOR result of Q1 and Q2 is 1 and the XNOR result is 0, Cout takes the Cin value. Conversely, when the XOR result is 0 and the XNOR result is 1, Cout takes the Q2 value. The corresponding truth table is presented in

Table 2 references in alphabetical order or in order of appearance in the paper [6]. Next, let's discuss the circuit logic for the Sum. As shown in Fig.8, when Cin is 1 and $\overline{\text{Cin}}$ is 0, Sum directly takes the XNOR value of Q1 and Q2. Conversely, when Cin is 0 and $\overline{\text{Cin}}$ is 1, Sum directly takes the XOR value of Q1 and Q2. The corresponding truth table is presented in Table 3.

Table.2 Truth Table of Cout Using PTL Architecture

Q2	Cin	XOR	XNOR	Cout
0	0	1	0	0(Cin)
0	0	0	1	0(Q2)
0	1	1	0	1(Cin)
0	1	0	1	0(Q2)
1	0	1	0	0(Cin)
1	0	0	1	1(Q2)
1	1	1	0	1(Cin)
1	1	0	1	1(Q2)

Table. 3 Truth Table of Cout Using PTL Architecture

XOR	XNOR	Cin	$\overline{Cin}$	Sum
0	0	1	0	0(XNOR)
0	0	0	1	0(XOR)
0	1	1	0	1(XNOR)
0	1	0	1	0(XOR)
1	0	1	0	0(XNOR)
1	0	0	1	1(XOR)
1	1	1	0	1(XNOR)
1	1	0	1	1(XOR)

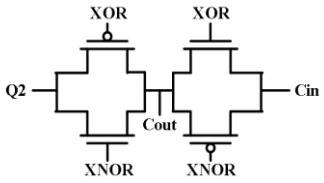


Fig.7 Circuit Diagram of Cout in PTL Architecture

Comparing the truth tables of the full adder and full subtractor reveals that the Difference in the full subtractor and the Sum in the full adder are identical. This suggests that the PTL full adder architecture generating the Sum for addition can also produce the Difference for subtraction.

Next, we consider integrating Cout and Bout. Based on the previously introduced PTL adder circuit logic, when Q1 and Q2 yield XNOR = 1, Bout equals Q2. Conversely, when Q1 and Q2 yield XOR = 1, Bout equals Cin, the inverse logic of addition. Thus, by interchanging Q2 and Cin in the PTL architecture generating Cout, as shown in Fig. 9, Bout can be obtained.

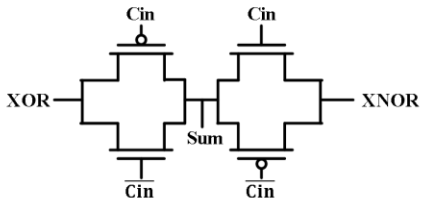


Fig.8 Circuit Diagram of Sum in PTL Architecture

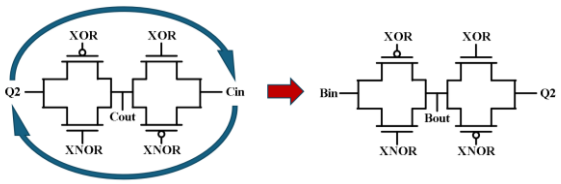


Fig.9 Generating Bout in PTL Architecture

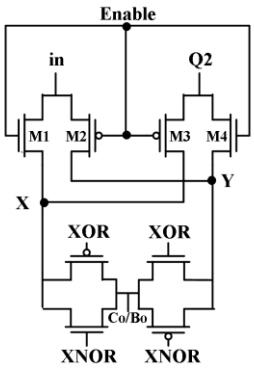


Fig.10 Dynamic Switching with PTL Architecture for Generating Co/Bo

Given that both addition and subtraction can be integrated within the PTL architecture, we aim to enable dynamic switching between the Q2 and Cin/Bin points in the Cout/Bout generation. Therefore, we have designed the following architecture. For readability, both Cin and Bin are referred to as "in," as shown in Fig. 10. When the Enable input is 0, M2 and M3 conduct, while M1 and M4 are cut off, allowing the data from the "in" terminal to reach point Y through M2, and the data from Q2 to reach point X through M3. This means that "when Enable = 0, the PTL architecture performs addition to generate Cout." Conversely, when the Enable input is 1, M1 and M4 conduct, while M2 and M3 are cut off, allowing the data from the "in" terminal to reach point X through M1, and the data from Q2 to reach point Y through M4. This means that when Enable = 1, the PTL architecture performs subtraction to generate Bout.

The entire design and operation of the circuit have been introduced above. Utilizing the 8+T SRAM in combination with the PTL adder-subtractor allows for efficient and rapid logical and arithmetic operations within the memory. With this, the complete in-memory computing architecture has been thoroughly explained.

IV. Experimental Results

A. *Logical Operation Results*

The preceding sections thoroughly introduce the architecture and operation of this study. Next, we will present the results and validate the functionality. All simulations and measurements were conducted using HSPICE, utilizing the 180-nanometer (nm) virtual process from TSRI at 25°C and 1.8V.

And then, we will explain the results of the logical operations. As seen in the various figures, before 5 microseconds (μs), the WWL is activated to allow data to be written into the memory. Subsequently, the WWL is deactivated, and the RWL is activated to enable data to be read and correctly generate the logical output. Therefore, the logical output results before 5 microseconds (μs) are not the final results.

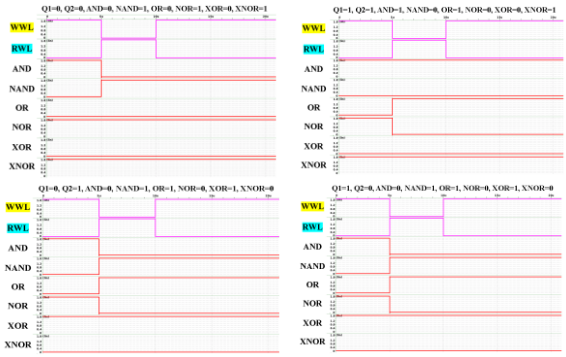


Fig.11 Logical Outputs Generated by 8+T SRAM in Various States

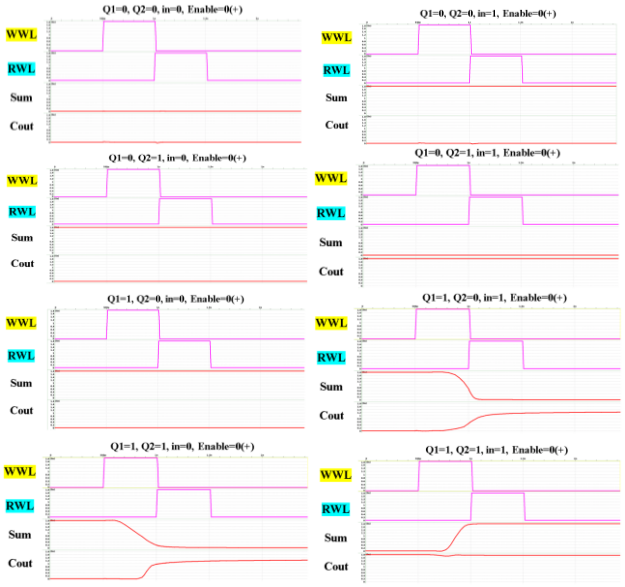


Fig.12 Addition Results Generated by the Proposed Architecture in Various States

The results from the Fig.11 match those of the basic logic gates.

B. *Addition and Subtraction Operation Results*

Next, we will explain the results of the addition operations in the Fig. 12. By comparing the experimental results with the truth table of the full adder, it can be observed that the two results match each other.

Next, we will explain the results of the subtraction operations in the Fig.13. By comparing the experimental results with the truth table of the full subtractor, it can be observed that the two results match each other.

C. *Delay Time Results*

The propagation delay, measured in picoseconds (ps), is calculated from the activation of the WWL or RWL until the voltage reaches VDD/2 and from the initiation of operations (XOR, XNOR, Sum/Diff, Cout/Bout, AND, NAND, OR, NOR) until their output reaches VDD/2.

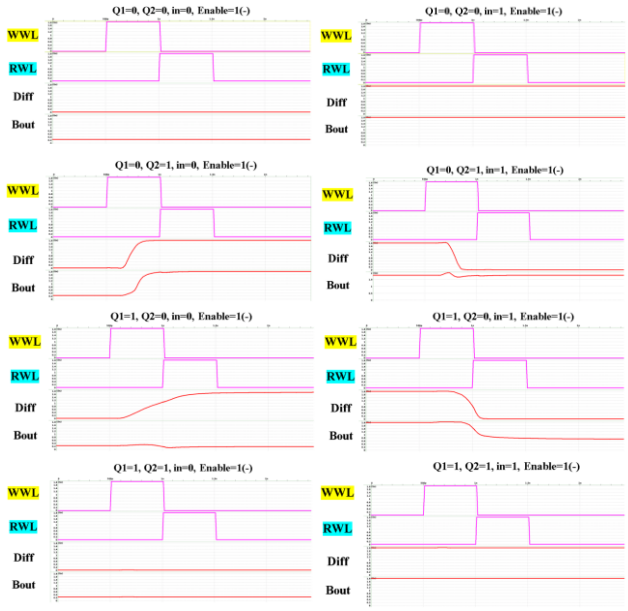


Fig. 13 Subtraction Results Generated by the Proposed Architecture in Various States

Table 4 Delay Time of the Proposed Architecture in This Paper (Unit: ps)

	Addition	**Subtraction**
AND	84.1875	82.9674
NAND	61.6961	60.2848
OR	63.1859	64.0917
NOR	85.8013	86.7333
XOR	249.7036	301.3825
XNOR	165.0656	238.0201
Sum/Diff	477.7527	492.1859
Cout/Bout	466.3265	630.9141

Table 5 Overall Power Consumption (Unit: μW)

Cin/ Bin	ADD/ SUB	Q1/Q2 Individual Write Values			
		(0,0)	(0,1)	(1,0)	(1,1)
in=0	ADD	48.6280	80.9560	84.4149	304.558
	SUB	49.5960	286.847	459.489	320.609
in=1	ADD	49.5421	80.9561	489.497	286.461
	SUB	49.5962	280.869	469.211	317.727
AVG		49.3405	182.407	375.653	307.338

Table 4 shows the maximum delay times for the proposed architecture in each state. In some states, there is no delay time. For Sum/Diff, Cout/Bout, XOR, and XNOR outputs, delays occur only in cases (1,0) and (1,1). In other states, the output values before and after WWL activation to VDD/2 are the same, resulting in no delay. For AND and NAND, there is no delay in case (1,1); for OR and NOR, there is no delay in case (0,0).

D. Power Consumption Calculation

This paper measures the average power consumption for each state when Q1, Q2, and Cin/Bin are individually written. The unit of measurement is fixed at microwatts (μW). Table 5 shows that overall power consumption peaks in the (1,0) state. This is likely due to the PTL architecture's primary drawback of lacking full swing at the output, leading to static power consumption. This issue can be mitigated using dual-edge-triggered flip-flops [20]. However, the architecture's advantage lies in performing both addition and subtraction through the exchange of two transistors, significantly reducing the required area. Future research will focus on reducing power consumption

V. Conclusion

The remarkable growth of artificial intelligence and deep learning has led to unprecedented data volumes, straining existing hardware. In-memory computing has thus become a critical research focus. While most research targets logical operations, the increasing demand for arithmetic operations is evident. This study explores using 8+T SRAM to enable both logical and arithmetic operations. Using Pass Transistor Logic (PTL) circuits, we designed an adder generating Sum and Carry-out outputs. Additionally, by examining adder and subtractor truth tables, we developed a dynamically switchable adder-subtractor.

Integrating computational functions within memory mitigates data transfer and processing delays, enhancing overall system performance. The dynamically switchable adder-subtractor offers high flexibility and scalability, addressing diverse computational needs and aligning with the performance demands of modern computing systems.

This research aims to provide valuable insights into advancing in-memory computing technologies and offers new perspectives on enhancing computer system performance and efficiency. Future research will optimize and apply this technology for broader applications and significant breakthroughs.

References

[1] Backus, J. (1978). Can programming be liberated from the von Neumann style? A functional style and its algebra of programs. *Communications of the ACM*, 21(8), 613-641.

[2] Zou, X., Xu, S., Chen, X., Yan, L., & Han, Y. (2021). Breaking the von Neumann bottleneck: architecture-level processing-in-memory technology. *Science China Information Sciences*, 64(6), 160404.

[3] Elliott, D. G., Stumm, M., Snelgrove, W. M., Cojocaru, C., & McKenzie, R. (1999). Computational RAM: Implementing processors in memory. *IEEE Design & Test of Computers*, 16(1), 32-41.

[4] Gauchi, R., Kooli, M., Vivet, P., Noel, J. P., Beigné, E., Mitra, S., & Charles, H. P. (2019, October). Memory sizing of a scalable SRAM in-memory computing tile based architecture. In *2019*

IFIP/IEEE 27th International Conference on Very Large Scale Integration (VLSI-SoC) (pp. 166-171). IEEE.

[5] Wang, J., Nalam, S., & Calhoun, B. H. (2008, August). Analyzing static and dynamic write margin for nanometer SRAMs. In Proceedings of the 2008 international symposium on Low Power Electronics & Design (pp. 129-134).

[6] Zhang, J., Wang, Z., & Verma, N. (2017). In-memory computation of a machine-learning classifier in a standard 6T SRAM array. *IEEE Journal of Solid-State Circuits*, 52(4), 915-924.

[7] Athe, P., & Dasgupta, S. (2009, October). A Comparative Study of 6T, 8T and 9T Decanano SRAM cell. *In 2009 IEEE Symposium on Industrial Electronics & Applications* (Vol. 2, pp. 889-894). IEEE.

[8] Ramani, A. R., & Choi, K. (2011, May). A novel 9T SRAM design in sub-threshold region. In 2011 *IEEE INTERNATIONAL CONFERENCE ON ELECTRO/INFORMATION TECHNOLOGY* (pp. 1-6). IEEE.

[9] Kulkarni, J. P., Goel, A., Ndai, P., & Roy, K. (2010). A read-disturb-free, differential sensing 1R/1W port, 8T bitcell array. *IEEE transactions on very large scale integration (VLSI) systems*, 19(9), 1727-1730.

[10] Agrawal, A., Jaiswal, A., Lee, C., & Roy, K. (2018). X-SRAM: Enabling in-memory Boolean computations in CMOS static random access memories. *IEEE Transactions on Circuits and Systems I: Regular Papers, 65(12)*, 4219-4232.

[11] Song, S., & Kim, Y. (2021, October). Novel in-memory computing circuit using muller C-element. In *2021 18th International SoC Design Conference (ISOCC)* (pp. 81-82). IEEE.

[12] B. Holdsworth . (2002) Digital Logic Design.

[13] Singh, S. S., Leishangthem, D., Shah, M. N., & Shougaijam, B. (2020, November). A Unique design of hybrid full adder for the application of low power VLSI circuits. In *2020 4th international conference on electronics, communication and aerospace technology (ICECA)* (pp. 260-264). IEEE.

[14] Mahendran, G. (2024). CMOS full adder cells based on modified full swing restored complementary pass transistor logic for energy efficient high speed arithmetic applications. *Integration, 95*, 102132.

[15] Devi, S., Suhas, N. S., & Vijay, K. A. (2017, February). Design of full subtractor using DPL logic and MTCMOS technique to reduce the leakage current and area. In *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-4). IEEE.

[16] Ding, L., Zhang, Z., Pei, T., Liang, S., Wang, S., Zhou, W., ... & Peng, L. M. (2012). Carbon nanotube field-effect transistors for use as pass transistors in integrated logic gates and full subtractor circuits. *ACS nano, 6(5)*, 4013-4019.

[17] Sharma, T., & Kumre, L. (2017, October). A comparative performance analysis of CMOS XOR XNOR circuits. In *2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE)* (pp. 473-478). IEEE.

[18] Frontini, L., Shojaii, S., Stabile, A., & Liberali, V. (2012, December). A new XOR-based Content Addressable Memory architecture. In *2012 19th IEEE International Conference on Electronics, Circuits, and Systems (ICECS 2012)* (pp. 701-704). IEEE.

[19] Babu, H., KV, R. R., & Dhanabal, R. (2014, February). Comparative performance analysis of XOR-XNOR function based high-speed CMOS full adder circuits. In *2014 International Conference on Reliability Optimization and Information Technology (ICROIT)* (pp. 432-436). IEEE.

[20] Hossain, R., Wronski, L. D., & Albicki, A. (1994). Low power design using double edge triggered flip-flops. *IEEE transactions on very large scale integration (VLSI) systems*, *2*(2), 261-265.

Speech Separation using Neural Audio Codecs with Embedding Loss

Jia Qi Yip*† Chin Yuen Kwok* Bin Ma† and Eng Siong Chng*
* Nanyang Technological University, Singapore
† Alibaba Group
E-mail: jiaqi006@e.ntu.edu.sg

***Abstract*—Neural audio codecs have revolutionized audio processing by enabling speech tasks to be performed on highly compressed representations. Recent work has shown that speech separation can be achieved within these compressed domains, offering faster training and reduced inference costs. However, current approaches still rely on waveform-based loss functions, necessitating unnecessary decoding steps during training. We propose a novel embedding loss for neural audio codec-based speech separation that operates directly on compressed audio representations, eliminating the need for decoding during training. To validate our approach, we conduct comprehensive evaluations using both objective metrics and perceptual assessment techniques, including intrusive and non-intrusive methods. Our results demonstrate that embedding loss can be used to train codec-based speech separation models with a 2x improvement in training speed and computational cost while achieving better DNSMOS and STOI performance on the WSJ0-2mix dataset across 3 different pre-trained codecs.**

I. Introduction

Speech separation, also known as the cocktail party problem [1], is the task of isolating individual speakers from overlapping audio, has seen remarkable progress in recent years [2]. While current models achieve impressive signal-to-noise ratios on clean datasets [3] [4], the field is now advancing towards more generalized capabilities in diverse environments. This shift towards broader applicability necessitates even larger training datasets [5], a requirement that is currently hindered by the substantial computational costs associated with training speech separation models.

The computational intensity of speech separation stems from two primary factors. First, traditional models typically employ minimal time-compression in their encoders, often using simple 1D convolutions that achieve only 8x downsampling in the time dimension [6]. Recent approaches have revisited hybrid frequency and time domain methods [7] [8], incorporating greater downsampling to enhance computational efficiency without sacrificing performance.

Second, the loss function and evaluation metrics used in speech separation are inherently expensive. The Permutation Invariant Training (PIT) [9] Loss, essential for handling the speaker order ambiguity, requires computing $O(N!)$ combinations of outputs and ground truth, where N is the number of speakers in the mixture. Furthermore, the commonly used Scale-Invariant Signal to Distortion Ratio (SI-SDR) [10] is a waveform comparison loss function, which is computationally demanding for the long sequences typical in audio waveforms.

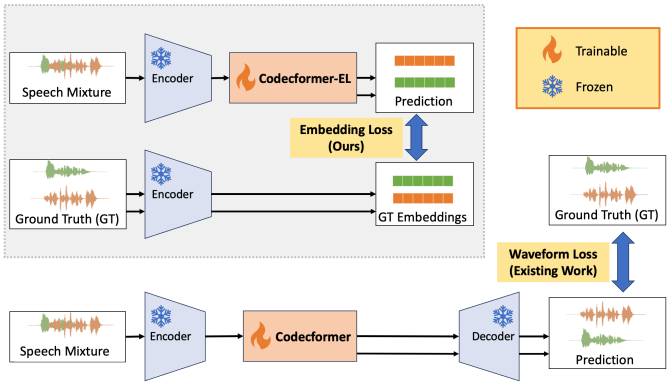


Fig. 1. Comparison between our proposed embedding loss (Top Left) with the conventional waveform loss used in a previous codec-based speech separation approach (Bottom). The key advantage of the embedding loss training method is that the decoder of the neural audio codec is not necessary for training, significantly reducing the computational cost during training.

Neural Audio Codecs [11] [12] [13], self-supervised models trained on vast amounts of general audio data, offer a promising solution to these challenges. These codecs, structured as autoencoders with a quantizer [14] between the encoder and decoder, can achieve time-domain compression ratios of approximately 100x, depending on the specific codec architecture. Due to the strong representations learned by neural audio codecs, they have been found to have useful applications beyond compression. They have been utilized in various downstream tasks, including automatic speech recognition [15] [16], text-to-speech [17] [18], speaker verification [19], and singing voice synthesis [20]. Recently, speech separation has also been added to this list of applications [21].

The recently proposed Codecformer [21], a neural audio codec-based speech separation model, has demonstrated the potential of integrating neural audio codecs with speech separation models to reduce computational requirements. However, as shown in Figure 1, this approach still relies on waveform-comparison losses, leaving room for further optimization.

In this paper, we propose a novel embedding loss for speech separation using neural audio codecs as shown in Figure 1. Our approach operates directly on the encoder representations of the codec, eliminating the need for expensive waveform comparisons during training. While PIT is still necessary, the shorter sequence length of the embeddings substantially decreases the computational cost of loss calculation.

Our contributions are threefold:

- We demonstrate that neural audio codec-based speech separation models can be trained from a variety of codecs using only embedding level loss, resulting in 2x improvement in training speech and computational cost with better performance on perceptual metrics, despite lower objective scores.
- While Codecformer [21] only supported the Descript Audio Codec (DAC) [13] and the PESQ [22] perceptual quality metric, we expand support to include EnCodec and SoundStream, as well as include additional perceptual quality metrics, STOI [23] and DNSMOS [24].
- We investigate the impact of pre-training data quantity and diversity on separation performance, comparing separation using the DAC model trained on open source datasets, LibriTTS, AMUSE, and the original DAC [13] dataset.

II. Methodology

In this work we propose Codecformer-EL, based on the recently proposed Codecformer [21], which modifies the training method for Codecformer by employing embedding loss as shown in Figure 1. While Codecformer utilizes both the encoder and decoder of the neural audio codec and performs waveform comparison loss, Codecformer-EL only requires the encoder. The loss computed is mean squared error (MSE) loss wrapped within the PIT [9] algorithm.

As shown in Figure 1, we perform speech separation by training a separator model, Codecformer [21], that makes use of a frozen encoder and decoder from a pre-trained neural audio codec. During inference and evaluation, the decoder is used to generate the necessary waveform for comparison.

A. Embedding Loss Function

Our proposed Embedding loss function can be written as follows:

$$e = \text{Encoder}(s_{gt})$$
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{e}_i - e_i)^2 \quad (1)$$

where s_{gt} is the ground truth speech and $\hat{e}$ is the separated embeddings produced by the separator in Codecformer-EL and n is the number of samples in each batch. This loss function is only used during training. Since the encoder used here to produce the groundtruth embeddings, e, is frozen, the embeddings can either be generated during training to save memory, or can be pre-computed before training to save on compute.

The choice of mean squared error loss in this study is inspired by early speech separation work that used this loss over spectrograms. However, because the neural audio codecs are learned representations, the embeddings obtained by the encoder are not spectrograms. Nevertheless, these embeddings are representations of audio and can be decoded into audio. Thus they may be thought of as spectrogram-like, which makes MSE loss more appropriate as opposed to Cosine Similarity loss used for semantic representations.

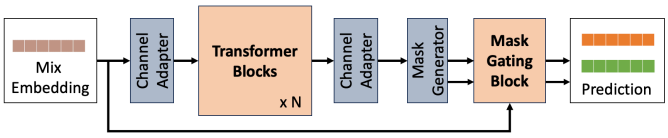


Fig. 2. Overview of the Codecformer model based on [21]. The bulk of the model consists of N transformer blocks along with channel adapter layers, a mask generator and the mask gating block.

B. Neural Audio Codecs

To verify that the embedding loss proposed in this paper is applicable to neural audio codec-based speech separation more generally, we make use of three popular neural audio codecs in this study, SoundStream [11], EnCodec [12] and Descript Audio Codec (DAC) [13] as described in Table I. We also compare the differences in performance across different training datasets, LibriTTS [25], AMUSE [26], and the original DAC dataset for the DAC model. All the codecs with embedding size of 512 in Table I were pre-trained using ESPnet-Codec [26].

TABLE I
COMPARISON OF THE DIFFERENT CODECS USED IN THIS WORK

Model	Pre-training Dataset	Embedding Size	Params (M)
SoundStream	AMUSE	512	19.7
EnCodec	AMUSE	512	19.7
DAC	LibriTTS, AMUSE	512	19.7
DAC	Original	1024	74.2

SoundStream [11] pioneered the use of Residual Vector Quantization (RVQ) in neural audio codec development. Its architecture combined a SEANet-based encoder-decoder with an RVQ quantizer. For discrimination, it employed both waveform and spectral discriminators. The model's loss functions included adversarial and feature-matching losses from these discriminators, along with a multi-resolution mel-spectrogram loss.

EnCodec [12] built upon SoundStream's design, substituting the STFT discriminator with a multi-scale version. It introduced a probabilistic approach to discriminator updates and implemented a loss balancing mechanism to automatically adjust loss scales during training.

DAC [13] further refined EnCodec by incorporating several enhancements. These included the adoption of the snake activation function, advanced RVQ featuring factorized and L2-normalized codes, increased use of quantizer dropout, and implementation of an MSMPMB discriminator. These improvements aimed to tackle codebook collapse, the phenomenon where a fraction of the codes are unused.

The LibriTTS dataset [25] is a commonly used for training neural audio codecs. The Audio, Music, and Speech Ensemble (AMUSE) dataset developed for ESPnet-Codec [26], is a fused corpus consisting of multiple open source high quality audio data, which also includes the original DAC dataset. Thus, the models trained on the AMUSE dataset represents the models with the largest-scale pre-training.

C. Codecformer-EL

The Codecformer with embedding loss (Codecformer-EL) model proposed in this study makes use of the same basic architecture as Codecformer [21] as shown in Figure 2. The bulk of the model consists of a stack of N transformer blocks followed by a mask gating block that is responsible for modifying the mix embedding received from the encoder. The Channel Adapters are linear layers that adapt the codec embedding size to the required dimensions of the transformer blocks. The mask generator is a linear layer that increase the number of channels to match the number of speakers in the mixture.

One weakness of the original Codecformer [21] it that the model only supported the DAC encoders and decoders. In this work we adapted Codecformer to support EnCodec and SoundStream as well. This required modifying the activation function in the Mask Gating Block from the Snake activation function used in DAC, to the ELU [27] activation function used in EnCodec and SoundStream. By matching the activation function in Codecformer with that of the neural audio codecs that they are embedded in, the generated outputs of the separation model are more likely to adopt a distribution more similar to the distribution expected by the decoder, potentially resulting in better generation by the decoder.

III. Experiments

A. Dataset

Our experiments utilize the widely-used WSJ0-2mix benchmark dataset [28]. This dataset is derived from the WSJ0 corpus, which consists of English read speech from the Wall Street Journal. The dataset comprises speech mixtures created from clean audio recordings. All audio, including both the mixtures and their corresponding ground truth, is sampled at 8kHz and recorded in acoustically controlled settings. The synthetic mixtures are generated using the 'min' condition, where the length of each mixture is cropped to match the duration of the shorter of the two input speeches. The WSJ0-2mix dataset is structured as follows: A training set consisting of 20,000 mixtures (30 hours) from the WSJ0 si_tr_s set, validation set consisting of 5,000 mixtures (10 hours) from the WSJ0 si_dt_05 set and a test set consisting of 3,000 mixtures (5 hours) from the si_et_05 set.

B. Codec Implementation

To ensure a consistent implementation, the model code and pre-trained weights for the codecs used were obtained from the ESPnet-Codec repository [26] except for the DAC Original implementation where the code and weights were obtained from the official release of the original authors[1]. Meanwhile, the training and evaluation of the models was implemented on the speechbrain toolkit [29] with the models from ESPnet handled by a wrapper written for speechbrain. For all neural audio codec models, the 16kHz version of the model is used, which is the lowest sampling rate available across all models. To manage the difference in sampling rate of the dataset and the neural audio codecs, the input audio is resampled using the torchaudio resampling method. Although this means that the codec receives audio that has missing frequencies, it is not expected that this results in significant performance impact.

C. Evaluation Methods

In addition to the standard objective metrics used for speech separation, such as SI-SDR, SI-SDRi, SDR, SDRi, we make use of perceptual evaluation metrics to measure the performance of the models. The perceptual evaluation metrics are necessary when performing neural audio codec-based speech separation because distortions introduced by the codecs result in distortions that are heavily penalized by objective metrics but are do not affect perceptual quality.

The perceptual quality metrics used in this study are deep noise suppression mean opinion score (DNSMOS) [24], perceptual evaluation of speech quality (PESQ) [22] and short-time objective intelligibility (STOI) [23]. Among these metrics, DNSMOS is a non-intrusive metric that does not rely on reference speech, while PESQ and STOI are intrusive metrics that do rely on some reference speech. Due to the ambiguity in the output order of the model, the ideal permutation for each of these metrics is computed based on maximizing the SI-SDR over all possible permutations. The permutation of outputs against the ground-truth that produces the highest average SI-SDR is used to compute all of the perceptual quality metrics. This also ensures that a consistent permutation is chosen across all objective and perceptual evaluation metrics.

D. Training Procedure

All models were trained for 20 epochs on a V100 with 32GB RAM. The Adam optimizer was used with an initial learning rate of $1.5e^{-4}$ and the LR scheduler was set to halve the learning rate with a patience of 2 after epoch 5. The models trained using the embedding loss were trained with a batch size of 20, while the those trained using the waveform loss were trained with a batch size of 3 due to memory limitations.

Following the settings of [21], the Codecformer separator consists of 16 transformer blocks with an input embedding size of 256. This is regardless of the embedding size of the neural audio codec used, which is either 1024 in the case of the original DAC and 512 for the ESPnet models, ensuring a fair comparison. The adapter layer in Codecformer was used to map the representations to the correct number of channels.

IV. Results

A. Improvements to Training Speed and Computation

One of the key features of Codecformer-EL is the improvement in training speed and computational efficiently. We measure this in Table IV using Multiple and Accumulate operations (MACs) computed using the PyTorch-OpCounter[2] as well as the number of hours required to train each epoch on the WSJ0-2mix dataset. We can see that Codecformer-EL trained 2.5x

[1] https://github.com/descriptinc/descript-audio-codec

[2] https://github.com/Lyken17/pytorch-OpCounter

TABLE II
COMPARISON OF SEPARATION PERFORMANCE ON DIFFERENT NEURAL AUDIO CODEC MODELS TRAINED ON EMBEDDING AND WAVEFORM LOSSES. ALL CODEC MODELS WERE PRE-TRAINED ON THE AMUSE DATASET

Model	**Loss Type**	**Objective Metrics**				**DNSMOS**				**PESQ**	**STOI**
		SI-SDR	**SI-SDRi**	**SDR**	**SDRi**	**OVRL**	**SIG**	**BAK**	**p808**		
DAC	Embedding	-1.2	-1.3	1.1	0.9	1.80	**2.12**	2.99	**2.58**	1.80	**0.81**
DAC	Waveform	2.8	2.8	4.9	4.7	**1.81**	2.04	**3.38**	2.55	**2.06**	0.79
EnCodec	Embedding	-29.1	-29.1	-8.5	-8.6	**1.89**	**2.20**	**3.18**	**2.65**	**1.92**	**0.80**
EnCodec	Waveform	-9.0	-9.0	-2.6	-2.8	1.24	1.26	2.75	2.29	1.28	0.51
SoundStream	Embedding	-14.2	-14.2	-4.9	-5.00	**1.80**	**2.07**	2.98	**2.63**	**1.94**	**0.83**
SoundStream	Waveform	-2.5	-2.4	1.2	1.0	1.73	1.95	**3.35**	2.45	1.77	0.74

TABLE III
COMPARISON OF CODECFORMER SEPARATION PERFORMANCE ON THE DAC MODEL WITH DIFFERENT PRE-TRAINING DATASETS

Pre-training Dataset	**Loss Type**	**Objective Metrics**				**DNSMOS**				**PESQ**	**STOI**
		SI-SDR	**SI-SDRi**	**SDR**	**SDRi**	**OVRL**	**SIG**	**BAK**	**p808**		
AMUSE	Embedding	-1.16	-1.16	1.10	0.94	1.80	2.12	2.99	2.58	1.80	0.81
	Waveform	2.8	2.7	4.9	4.7	1.81	2.04	3.38	2.55	2.06	0.79
Original	Embedding	-21.9	-22.0	-2.4	-2.6	1.57	1.91	2.53	2.38	1.57	0.71
	Waveform	6.7	6.7	7.9	7.7	2.19	2.53	3.40	2.80	2.20	0.85
LibriTTS	Embedding	-3.4	-3.3	-0.1	-0.2	1.89	2.26	2.98	2.62	1.75	0.80
	Waveform	1.0	1.1	3.0	2.9	1.82	2.03	3.43	2.60	1.76	0.75

faster and with 1.9x fewer MACs compared to Codecformer. Additionally, it trained 6.8x faster than Sepformer and requires 97x fewer MACs. The MACs were calculated on 2 seconds of 8kHz audio data and a V100 GPU with 16GB of RAM utilizing the maximum possible batch size was used to calculate the training speed.

TABLE IV
COMPUTATION AND TRAINING SPEED OF CODECFORMER-EL COMPARED AGAINST CODECFORMER AND SEPFORMER

Model	GMACs	Training Time (h/epoch)
Sepformer [2]	77.3	2.7
Codecformer [21]	1.5	1.0
Codecformer-EL (Ours)	**0.8**	**0.4**

B. Comparison between Codecformer-EL and Codecformer

Table II presents the separation performance of three neural audio codecs pre-trained on the AMUSE dataset, comparing our Codecformer-EL (embedding loss) method with the original Codecformer (waveform loss) approach.

A consistent pattern emerges: Codecformer models achieve higher scores on objective metrics (SI-SDR, SI-SDRi, SDR, SDRi), while Codecformer-EL models perform better on perceptual metrics, despite lower objective scores. This discrepancy highlights a known limitation of objective waveform-matching metrics for evaluating resynthesized speech [30].

The neural audio codec's decoder, trained with GAN-based loss, prioritizes generating perceptually natural speech over exact waveform reconstruction. This approach explains the disparity between objective and perceptual metric scores. Across the models, we observe varying degrees of performance difference. DAC models show the smallest gap between methods, with embedding loss outperforming on some perceptual metrics. EnCodec models display the most dramatic difference, with embedding loss achieving the highest perceptual scores despite poor objective metrics. SoundStream models also demonstrate a clear advantage for embedding loss in perceptual metrics.

These results underscore the potential of Codecformer-EL for real-world applications where perceived quality is crucial. They also emphasize the importance of using perceptual quality metrics for evaluating resynthesized speech separation performance. The success of our method across different codec architectures demonstrates its versatility and potential for wide applicability in speech separation tasks.

C. Comparison of speech separation performance across different codec pre-training datasets

In Table III we show the performance of the DAC model on speech separation after pre-training on different sets of data. In this analysis we see that regardless of the pre-training data, the DAC model performs similarly on the AMUSE and LibriTTS dataset, with embedding loss having a similar or better separation performance than waveform loss. However, in the model trained on the Original DAC dataset, the embedding method performs worse than the waveform loss and the Original DAC model with waveform loss performs the best out of all the models. This could potentially be due to the larger embedding size of 1024 compared to 512 in the other models. The embedding loss training method, due to its heavy reliance on the compressed representations, could be penalised more for the mismatch in dimension sizes between the transformer blocks and the DAC encoder. This could be further investigated in future work with ablation studies over different embedding sizes for the transformer in Codecformer-EL.

D. Practical Considerations

As discussed in Section IV-A, Codecformer-EL offers significant advantages in scenarios where neural audio codecs are already deployed or being considered. In these scenarios, because neural audio codecs are regenerating audio in a non-linear manner during decoding, it can cause distortions that result in a domain mismatch for signal-level speech separation tasks [21].

One solution to this problem is to retrain or fine-tune a speech separation model on the outputs of neural audio codecs. However, this process can be slow and costly. On the other hand, the efficiency of Codecformer-EL makes it particularly valuable for rapid adaptation to new codecs, and is a more convenient approach compared to re-training separation models on new codec outputs at the signal level.

By using compressed representations, Codecformer-EL reduces computational requirements for both training and inference, which could lead to faster deployment cycles and lower operational costs. The reduced computational demands also enhance scalability, potentially enabling deployment on a wider range of devices, including those with limited processing power.

E. Limitations

Despite its advantages, Codecformer-EL faces several limitations. Firstly, the upper bound of separation performance is inherently constrained by the quality and capabilities of the underlying neural audio codec. Since the model is trained on embeddings, the performance of Codecformer-EL heavily relies on the quality and generalizability of the pre-trained neural audio codec for both extracting mixture embeddings and providing the ground-truth labels. Therefore, suboptimal performance may result if the training data of the neural audio codec significantly differs from the data used to train Codecformer-EL.

Secondly, due to the generative nature of neural audio codecs, it is difficult to properly evaluate the performance of the separation output, limiting competitiveness with prior state-of-the-art results reported using signal-level loss. The variability introduced by the neural audio codec generation means that the output does not always align perfectly with original signal characteristics, despite having good perceptual quality. As observed in our results in Table II and III, this creates a discrepancy between objective and perceptual metrics in the results, and poses challenges in directly comparing Codecformer-EL's performance with methods evaluated primarily on objective metrics.

Finally, working with learned representations in the compressed domain may reduce the interpretability of the separation process compared to methods operating directly on waveforms or spectrograms. Additionally, performance may vary across different neural audio codec architectures, potentially requiring codec-specific optimizations to achieve optimal results.

V. Conclusions

In this paper, we introduced Codecformer-EL, a novel approach to speech separation that leverages neural audio codecs and employs an embedding-level loss function. Our method demonstrates significant improvements in computational efficiency and training speed compared to previous approaches, while maintaining comparable or superior perceptual quality in separated speech. We showed that Codecformer-EL is effective across multiple neural audio codecs and pre-training datasets, highlighting its versatility and potential for widespread application.

Our findings reveal an important dichotomy between objective and perceptual metrics in evaluating resynthesized speech. While objective metrics sometimes indicated poorer performance, perceptual quality metrics often showed comparable or superior results, underscoring the importance of using appropriate evaluation methods for this type of speech processing task.

Codecformer-EL's improved efficiency paves the way for more scalable speech separation models capable of handling real-world audio challenges, particularly in applications where audio compression is necessary for transmission, such as smartphones offloading computationally intensive tasks to cloud servers [21]. However, it's crucial to acknowledge potential limitations, including the trade-off between compression and information preservation in codec embeddings. Highly compressed embeddings may lead to loss of fine-grained audio details, potentially affecting separation quality in complex acoustic environments [19]. Moreover, Codecformer-EL's performance is inherently tied to the quality and generalizability of the pre-trained neural audio codec, which may lead to suboptimal results when the codec's pre-training data significantly differs from the target separation task.

Future work could explore techniques to mitigate these limitations, such as fine-tuning the codec encoder for specific separation tasks or investigating adaptive compression levels based on input complexity. Additionally, research into the optimal balance between compression and separation quality could further enhance the practical applicability of this approach.

In conclusion, Codecformer-EL represents a significant step forward in efficient, high-quality speech separation, opening new avenues for both research and real-world applications in audio processing and communication technologies.

Acknowledgment

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore. We would like to acknowledge Alibaba-NTU Joint Research Institute, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 21–25.

[3] J. Q. Yip, S. Zhao, Y. Ma, *et al.*, "Spgm: Prioritizing local features for enhanced speech separation performance," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[4] S. Zhao, Y. Ma, C. Ni, *et al.*, "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[5] J. Pons, X. Liu, S. Pascual, and J. Serrà, "Gass: Generalizing audio source separation with large-scale data," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 546–550.

[6] C. Subakan, M. Ravanelli, S. Cornell, F. Grondin, and M. Bronzi, "Exploring self-attention mechanisms for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2169–2180, 2022.

[7] Z. Wang, S. Cornell, S. Choi, Y. Lee, B. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *TASLP*, vol. 31, pp. 3221–3236, 2022.

[8] C. Chen, C.-H. H. Yang, K. Li, Y. Hu, P.-J. Ku, and E. S. Chng, "A neural state-space model approach to efficient speech separation," in *Proc. Interspeech*, 2023.

[9] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245. DOI: 10.1109/ICASSP.2017.7952154.

[10] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 626–630.

[11] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[12] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[13] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] R. M. Gray, "Vector quantization," *IEEE ASSP Magazine*, vol. 1, pp. 4–29, 1984.

[15] T. Wang, L. Zhou, Z. Zhang, *et al.*, "Viola: Unified codec language models for speech recognition, synthesis, and translation," *arXiv:2305.16107*, 2023.

[16] A. Gupta, G. Saon, and B. Kingsbury, "Exploring the limits of decoder-only models trained on public speech recognition corpora," *arXiv:2402.00235*, 2024.

[17] Z. Ju, Y. Wang, K. Shen, *et al.*, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in *ICML*, 2024.

[18] D. Yang, D. Wang, H. Guo, X. Chen, X. Wu, and H. Meng, "Simplespeech: Towards simple and efficient text-to-speech with scalar latent transformer diffusion models," *arXiv:2406.02328*, 2024.

[19] K. C. Puvvada, N. R. Koluguri, K. Dhawan, J. Balam, and B. Ginsburg, "Discrete audio representation as an alternative to mel-spectrograms for speaker and speech recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 111–12 115.

[20] X. Chang, J. Shi, J. Tian, *et al.*, "The Interspeech 2024 challenge on speech processing using discrete units," in *Proc. Interspeech*, 2024.

[21] J. Q. Yip, S. Zhao, D. Ng, E. S. Chng, and B. Ma, "Towards audio codec-based speech separation," in *Proc. Interspeech*, 2024.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, IEEE, vol. 2, 2001, pp. 749–752.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *TASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[24] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6493–6497.

[25] H. Zen, V. Dang, R. Clark, *et al.*, "LibriTTS: A corpus derived from librispeech for text-to-speech," in *Proc. Interspeech*, 2019.

[26] J. Shi, J. Tian, Y. Wu, *et al.*, "Espnet-codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech," in *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2024.

[27] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential

linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[28] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 31–35.

[29] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, *SpeechBrain: A general-purpose speech toolkit*, arXiv:2106.04624, 2021. arXiv: 2106 . 04624 [eess.AS].

[30] J. Shi, X. Chang, T. Hayashi, Y.-J. Lu, S. Watanabe, and B. Xu, *Discretization and re-synthesis: An alternative method to solve the cocktail party problem*, 2022. arXiv: 2112 . 09382 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2112.09382.

Sound Quality Improvement in Visual Microphone by Emphasizing Focused Area Based on Focal Rate

Hayata Nakano*, Yuting Geng*, Kenta Iwai* and Takanobu Nishiura*
* Ritsumeikan University, Osaka, Japan
E-mail: {is0516vr@ed, geng@fc, iwai18sp@fc, nishiura@is}.ritsumei.ac.jp

***Abstract*—This paper presents a method of improving the sound quality of visual microphone by emphasizing the in-focus area in a captured video. Visual microphone extracts sound by measuring the displacement of a vibrating object from the video of an object vibrating due to sound. In the captured video, there may arise blurred (out-of-focus) areas may arise due to depth of field. In these out-of-focus areas of the captured video, it is difficult to accurately measure the displacement of the vibrating object, which may reduce the sound quality of the extracted sound. However, out-of-focus areas are not taken into account in conventional sound extraction methods. We propose a method to improve the sound quality of the extracted sound in the visual microphone by emphasizing the in-focus area in the captured video. The proposed method emphasizes the measured displacement in the in-focus area of the captured video by using the focal rate that represents the degree of focus. Experimental results show that the proposed method improves the quality of the extracted sound compared to the conventional methods.**

I. Introduction

In sound recording, an air-conduction microphone is typically used to convert a sound wave into an electrical signal when a sound wave arrives at the diaphragm of the microphone. Consequently, the microphone acquires noise around itself in addition to the target sound. The visual microphone [1], [2] was proposed to address this problem. The visual microphone extracts a sound signal from a video of the object vibrating due to the target sound captured by a consumer camera. The sound signal extracted by the visual microphone contains little noise that occurs near the camera.

Many consumer-level cameras are equipped with rolling shutter image sensors [3]. These cameras expose and read pixels from the top row of an image to the bottom. Therefore, when capturing an object vibrating due to sound, rolling shutter distortion [3] arises in the captured images as a result of differences in exposure start times. The visual microphone extracts the sound signal by measuring the displacement of the object in each row of the captured images using this distortion. Here, we consider the case where out-of-focus (i.e., blurred) areas arise in the captured video. Therefore, when extracting sound including out-of-focus areas in the captured video, it is difficult to accurately measure the displacement of the vibrating object, leading to degradation in the quality of the extracted sound signal. However, out-of-focus areas are not taken into account in conventional sound extraction methods [1], [2].

Taking into account the out-of-focus areas in the captured video, this paper aims to improve the sound quality of the extracted sound in the visual microphone by emphasizing the in-focus area in the captured video. In our previous research [4], we proposed two methods: removing out-of-focus areas and weighting phase variation. By improving the sound quality of the extracted sound using our previous methods [4], we further emphasize the measured displacement in the in-focus area based on focal rate, which represents the degree of focus, to improve the sound quality of the extracted sound in the visual microphone. We conducted evaluation experiments to verify the effectiveness of the proposed method for sound extraction.

II. Conventional sound extraction method in visual microphone

The conventional method for sound extraction in the visual microphone [1] is illustrated in Fig. 1. The conventional method extracts sound signal $u(t)$ by calculating the displacement of the object using a complex steerable pyramid [5] on a video $I(x, y, n)$ of the object vibrating due to sound, where x and y are the row and column indices, n is the frame index, and t is the time index. Here, the vibration of the object appears as phase variation [6], [7], [8] enabling us to extract sound signal based on the phase variation of each row in each frame. The displacement of the object is obtained based on the phase variation calculated using the complex steerable pyramid.

The complex steerable pyramid consists of sub-band images for each scale r and orientation θ. Amplitude $A(r, \theta, x, y, n)$ and phase $\phi(r, \theta, x, y, n)$ are calculated by applying the complex steerable pyramid to the video $I(x, y, n)$. $A(r, \theta, x, y, n)$ and $\phi(r, \theta, x, y, n)$ are downsampled to 2^{-r} times the original resolution per scale r. As the conventional method uses the phase variation of each row in each video frame, the orientation θ is fixed to 0. Therefore, $A(r, \theta, x, y, n)$ and $\phi(r, \theta, x, y, n)$ are replaced by $A(r, x, y, n)$ and $\phi(r, x, y, n)$.

The phase variation $\phi_{\mathrm{v}}(r, x, y, n)$ is calculated from the phase difference with the reference frame $\phi(r, x, y, n_0)$, where n_0 denotes the index representing the reference frame. We calculate the row-wise weighted average of $\phi_{\mathrm{v}}(r, x, y, n)$ to obtain $\Phi(r, y, n)$. Here, $\Phi(r, y, n)$ for each r is a two-dimensional signal. Therefore, the averaged phase variation is transformed to a one-dimensional signal $\tilde{\Phi}(r, t)$ for each r.

The frame gaps arise due to the characteristics of rolling shutter image sensors [3]. Therefore, an autoregressive model [9] is utilized to interpolate the frame gaps in the extracted signal $\tilde{\Phi}(r, t)$. Finally, the extracted sound signal $u(t)$ is

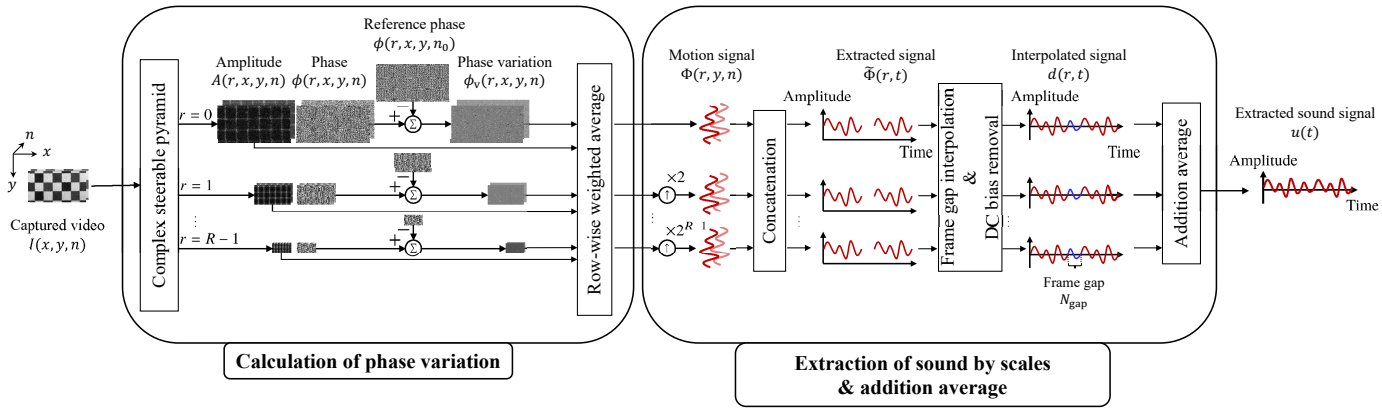


Fig. 1. Overview of conventional sound extraction method of visual microphone [1].

obtained by calculating the average of interpolated signal $d(r,t)$ at each scale r.

Out-of-focus areas are not taken into account in conventional methods [1], [2]. Hence, it is difficult to accurately calculate the phase variation when there are out-of-focus areas in the captured video. This causes the sound quality of the extracted sound to deteriorate. In this paper, we propose to improve the sound quality of the extracted sound in the visual microphone by emphasizing the in-focus area in the captured video.

III. Proposed Sound Extraction Method for Visual Microphone

Our proposed sound extraction method emphasizes the in-focus area in the captured video to improve the sound quality of the extracted sound. The proposed method is based on a combination of two methods, removing out-of-focus areas (Removed) and weighting phase variation (Weighted), proposed in [4]. The procedure of the proposed method is illustrated in Fig. 2.

Step 1: Calculation of Focal Rate and Out-of-focus Areas Removal

As the phase variation is not measured accurately in the out-of-focus areas of the captured video, we only utilize the in-focus area to extract the sound. Here, in the out-of-focus areas, the edges of the captured video are blurred, whereas the edges are relatively clear in the in-focus area. Therefore, the focal rate is calculated using a Sobel filter, which is commonly used as an edge enhancement filter [10]. This method only uses the displacement in the horizontal direction; thus, a horizontal Sobel filter is used. In out-of-focus areas, the change in pixel values is smooth, resulting in small absolute values of horizontal gradients. Therefore, we define the absolute values of horizontal gradients as the focal rate. The focal rate $M(x,y,n)$ for each frame of the captured video is calculated as

$$M(x,y,n) = |K(x,y) * I(x,y,n)|, \tag{1}$$

where $*$ denotes convolution operator and $K(x,y)$ denotes a 3×3 horizontal Sobel filter.

In this paper, we assume that the in-focus area does not change significantly between frames of the captured video. Therefore, we remove the out-of-focus areas by using the focal rate $M(x,y,n_0)$ calculated from the n_0-th frame.

First, the column-wise mean of $M(x,y,n_0)$ is calculated as

$$\overline{M}(x,n_0) = \frac{1}{H}\sum_{y=0}^{H-1} |M(x,y,n_0)|, \tag{2}$$

where H denotes the number of rows in $I(x,y,n)$. Next, we define $\boldsymbol{F}$ as the set of column indices x such that $\overline{M}(x,n_0)$ is larger than the threshold value $\overline{M}_{\mathrm{th}}$:

$$\boldsymbol{F} := \{x \mid \overline{M}(x,n_0) \geq \overline{M}_{\mathrm{th}}\}, \tag{3}$$

$$\overline{M}_{\mathrm{th}} = \frac{\alpha}{W}\sum_{x=0}^{W-1} \overline{M}(x,n_0), \tag{4}$$

where $\overline{M}_{\mathrm{th}}$ is calculated as α times the row-wise means $\overline{M}(x,n_0)$. The effect of removing out-of-focus areas can be amplified by increasing this α. The column indices indicating the start point F_{start} and endpoint F_{end} of the in-focus area are calculated as

$$(F_{\mathrm{start}}, F_{\mathrm{end}}) = \left(\min_{x\in\boldsymbol{F}} x, \max_{x\in\boldsymbol{F}} x\right). \tag{5}$$

Finally, we remove the out-of-focus areas from the captured video $I(x,y,n)$ and calculate $I_{\mathrm{focal}}(\tilde{x},y,n)$ which is the video of the in-focus area as

$$I_{\mathrm{focal}}(\tilde{x},y,n) = I(\tilde{x}+F_{\mathrm{start}},y,n), 0 \leq \tilde{x} \leq \tilde{W}-1, \tag{6}$$

$$\tilde{W} = F_{\mathrm{end}} - F_{\mathrm{start}} + 1, \tag{7}$$

where $\tilde{W}$ denotes the number of columns in $I_{\mathrm{focal}}(\tilde{x},y,n)$.

Step 2: Calculation of Phase Variation

As in the conventional method [1], amplitude $A(r,\tilde{x},y,n)$ and phase $\phi(r,\tilde{x},y,n)$ are calculated by applying the complex steerable pyramid to $I_{\mathrm{focal}}(\tilde{x},y,n)$. $A(r,\tilde{x},y,n)$ and

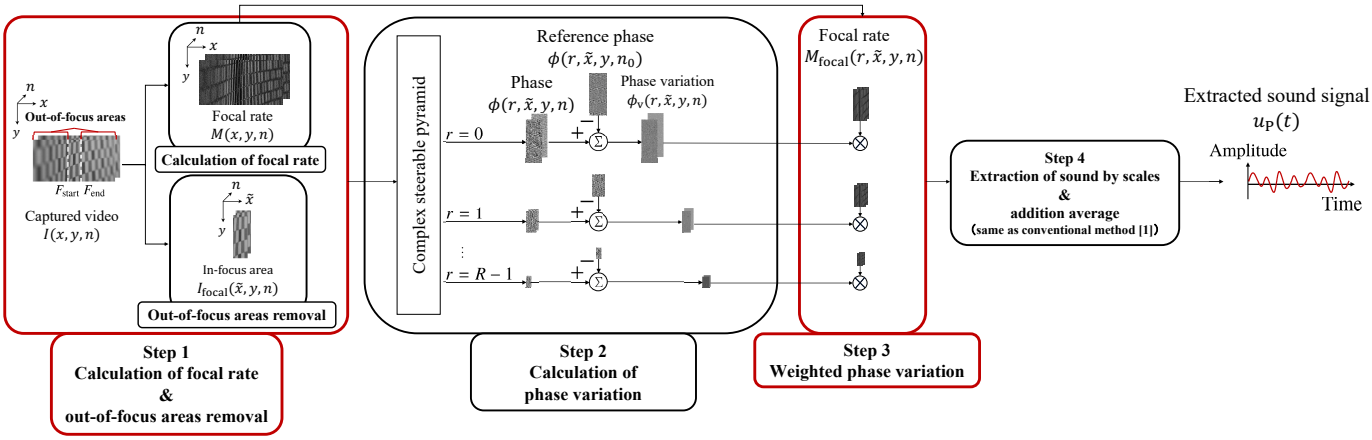


Fig. 2. Overview of proposed sound extraction method for visual microphone.

$\phi(r,\tilde{x},y,n)$ are downsampled to 2^{-r} times the original resolution per scale r. Then, the numbers of rows and columns in $A(r,\tilde{x},y,n)$ and $\phi(r,\tilde{x},y,n)$ for each scale r are given by

$$\tilde{W}_r = 2^{-r}\tilde{W}, \tag{8}$$

$$H_r = 2^{-r}H, \tag{9}$$

where $\tilde{W}$ and H denote the numbers of columns and rows in $I_{\mathrm{focal}}(\tilde{x},y,n)$. The phase variation $\phi_{\mathrm{v}}(r,\tilde{x},y,n)$ is calculated from the phase difference with the reference frame $\phi(r,\tilde{x},y,n_0)$ as

$$\phi_{\mathrm{v}}(r,\tilde{x},y,n) = \phi(r,\tilde{x},y,n) - \phi(r,\tilde{x},y,n_0). \tag{10}$$

Step 3: Weighted Phase Variation

The phase variation $\phi_{\mathrm{v}}(r,\tilde{x},y,n)$ obtained in Step 2 is weighted by the focal rate $M(x,y,n)$. The focal rate $M_{\mathrm{focal}}(\tilde{x},y,n)$ of the in-focus area is calculated as

$$M_{\mathrm{focal}}(\tilde{x},y,n) = M(\tilde{x}+F_{\mathrm{start}},y,n),\ 0 \le \tilde{x} \le \tilde{W}-1. \tag{11}$$

The focal rate $M_{\mathrm{focal}}(r,\tilde{x},y,n)$ for each scale r is calculated. The focal rate $M_{\mathrm{focal}}(r,\tilde{x},y,n)$ for each scale r is calculated by downsampling the $M_{\mathrm{focal}}(\tilde{x},y,n)$ to 2^{-r} times the resolution. In the conventional method, the phase variation $\phi_{\mathrm{v}}(r,x,y,n)$ is weighted by the square of the amplitude $A(r,x,y,n)$ since the phase can be accurately estimated for pixels with a large amplitude. The row-wise weighted average signal (the motion signal) $\Phi(r,y,n)$ is calculated as

$$\Phi(r,y,n) = \sum_{x=0}^{W_r-1} A^2(r,x,y,n)\phi_{\mathrm{v}}(r,x,y,n), \tag{12}$$

where W_r denotes the number of columns in $I(x,y,n)$ for each scale r.

In the proposed method, (12) is replaced by (13), weighting $\phi_{\mathrm{v}}(r,\tilde{x},y,n)$ by $M_{\mathrm{focal}}(r,\tilde{x},y,n)$ to emphasize the phase variation at the edges in $I_{\mathrm{focal}}(\tilde{x},y,n)$.

$$\Phi_{\mathrm{P}}(r,y,n) = \frac{1}{\tilde{W}_r}\sum_{\tilde{x}=0}^{\tilde{W}_r-1} M^2_{\mathrm{focal}}(r,\tilde{x},y,n)\phi_{\mathrm{v}}(r,\tilde{x},y,n). \tag{13}$$

As the length of $\Phi_{\mathrm{P}}(r,y,n)$ is H_r for each r, the signal length H_r is aligned to H through upsampling by a factor of 2^r for each r. Here, $\Phi_{\mathrm{P}}(r,y,n)$ for each r is a two-dimensional signal. Therefore, the averaged phase variation is transformed to a one-dimensional signal $\tilde{\Phi}_{\mathrm{P}}(r,t)$ for each r as

$$\tilde{\Phi}_{\mathrm{P}}(r,t) = \Phi_{\mathrm{P}}(r,y,n), \tag{14}$$

$$t = y + (H + N_{\mathrm{gap}})n, \tag{15}$$

where N_{gap} is the number of samples in the frame gap. As noted in Section 2, the frame gaps arise due to the characteristics of rolling shutter image sensors [3]. Therefore, an autoregressive model [9] is utilized to interpolate the frame gaps in the extracted signal $\tilde{\Phi}(r,t)$ to obtain an interpolated signal $d_{\mathrm{P}}(r,t)$.

Step 4: Extraction of Sound by Scales

The extracted sound signal $u_{\mathrm{P}}(t)$ is obtained by calculating the average of interpolated signal $d_{\mathrm{P}}(r,t)$ at each scale via

$$u_{\mathrm{P}}(t) = \frac{1}{R}\sum_{r=0}^{R-1} d_{\mathrm{P}}(r,t), \tag{16}$$

where R denotes the scale number of the complex steerable pyramid.

IV. Evaluation Experiments

We conducted evaluation experiments to investigate the effectiveness of the proposed method by comparing the sound extraction using the proposed method (Prop.) with the conventional methods (Conv. [1], Removed [4], and Weighted [4]).

A. Experimental Conditions for Sound Extraction

Figs. 3 shows the experimental setup and equipment arrangement, and 4 shows an A4 paper with a printed pattern used as the vibrating object. Moreover, Tables I, II and III show the experimental conditions, equipment and number of rows of video in each α in (4). As shown in Fig. 3, the A4 paper placed in front of a loudspeaker was captured with a rolling shutter camera, and the sound signal was extracted from the

captured video. To obtain the video with out-of-focus areas, the A4 paper was captured at 30°. This takes into account that capturing the subject at this angle causes out-of-focus areas in the video resulting from the relationship between focal length and depth of field. The pattern shown in Fig. 4 was printed on the A4 paper from the Salzburg Texture Image Database (STex) [11]. In addition, we used a floodlight for illumination so that we could accurately capture the A4 paper's vibration when the shutter speed was high. Fig. 5 shows one frame of the captured video.

We used sine waves as the sound source. Additionally, each sine wave was radiated from a loudspeaker for 5 s. In this experiment, we set $n_0 = 0$ as the reference frame and $r_0 = 0$ as a reference scale.

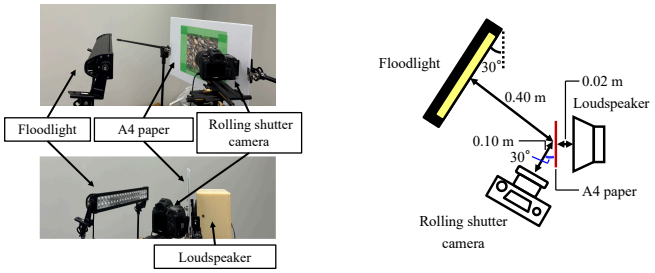


Fig. 3. Experimental setup and arrangement of experimental equipment.

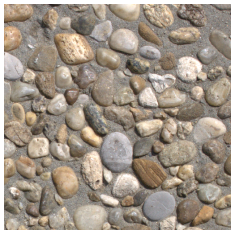


Fig. 4. Printed pattern.

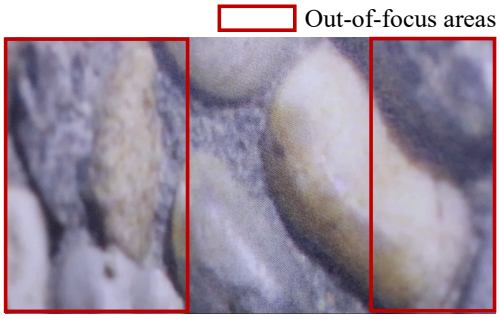


Fig. 5. One frame of captured video.

TABLE I
EXPERIMENTAL CONDITION

Ambient noise level	L_A=41.8 dB
Temperature / Humidity	20.8°C/51.7%
Sound source	Sine wave (300, 500, ···, 1,500 Hz)
Sound pressure level	85 dB at 0 m from A4 paper
Sampling frequency / Quantization	8,000 Hz / 16 bits
Resolution of captured video (width × height)	1,920 × 1,080 px
Frame rate of captured video	60 fps
Exposure time of camera	1/4,000 s

TABLE II
EXPERIMENTAL EQUIPMENT

Camera	Canon EOS 5D MarkIV
Camera lens	Canon MP-E 65 mm f/2.8 1-5x
Loudspeaker	FOSTEX FE83En
Loudspeaker amplifier	BOSE 1705II
Floodlight	GOODGOODS GDGDS-WL02 (10,000 lm)

B. Experimental Results of Sound Extraction

We evaluated the sound quality of extracted sound in terms of signal-to-distortion ratio (SDR). The SDR is calculated as

$$\mathrm{SDR} = 10\log 10 \left[\frac{\sum_{t=0}^{T-1} u_{\mathrm{c}}^2(t)}{\sum_{t=0}^{T-1} \{u_{\mathrm{c}}(t) - \lambda u(t)\}^2} \right], \quad (17)$$

$$\lambda = \sqrt{\frac{\sum_{t=0}^{T-1} u_{\mathrm{c}}^2(t)}{\sum_{t=0}^{T-1} u^2(t)}}, \quad (18)$$

where $u_{\mathrm{c}}(t)$ denotes the sine wave to be used as the sound source. Additionally, we experimented with $T = 80$ samples (0.01 s under sampling frequency of 8,000 Hz). The phase was adjusted by using cross-correlation so that $u(t)$ and $u_{\mathrm{c}}(t)$ were in phase.

Fig. 6 presents the evaluation results of the quality of the extracted sound in terms of SDR. The values in Fig. 6 indicate the average SDR. Fifty signals are randomly selected from the sound signal extracted by each method and the average SDR of the obtained Fifty signals is calculated.

Fig. 6 shows that the average SDR of Prop. is the highest at 500 – 1,500 Hz. This can be explained by the weighting based on the focal rate and the removal of out-of-focus areas, which can increase the contribution rate of the in-focus area in sound extraction, thereby enabling accurate measurement of the displacement. In particular, the accuracy of the sound quality improvement of the extracted sound of Prop. is highest at 500 Hz. Meanwhile, the average SDR of Conv. is highest at 300 Hz. However, it should be noted that the average SDR at 300 Hz is considerably lower than at other frequencies for all methods. Therefore, these results demonstrate that emphasizing the in-focus area is effective in improving the sound quality of the extracted sound, provided the sound quality is maintained at a certain level.

Fig. 7 presents the time waveforms of the 500 Hz sine waves extracted by the Conv., Weighted, Removed, and Prop. methods. Specifically, each figure shows the time waveform of 0.01 s signals from the extracted sound and the sound source, depicted as red lines and blue dotted lines respectively. The threshold values α for the Removed and Prop. methods are

Condition	Frequency [Hz] 300	500	700	900	1,100	1,300	1,500
Conv.	**-1.39**	**7.08**	**3.82**	**3.51**	**2.04**	**1.42**	**1.97**
Weighted	**-2.22**	**7.25**	**4.02**	**3.84**	**2.36**	**1.77**	**2.20**
Removed ($\alpha = 0.5$)	**-1.44**	**7.03**	**3.89**	**3.53**	**2.05**	**1.50**	**2.15**
Removed ($\alpha = 1.0$)	**-1.46**	**7.75**	**4.03**	**3.93**	**2.38**	**1.66**	**2.51**
Removed ($\alpha = 1.5$)	**-1.51**	**7.81**	**4.07**	**3.96**	**2.50**	**1.68**	**2.31**
Removed ($\alpha = 2.0$)	**-1.56**	**7.88**	**4.01**	**3.86**	**2.48**	**1.69**	**2.25**
Removed ($\alpha = 2.5$)	**-1.6**	**8.12**	**4.08**	**3.86**	**2.39**	**1.67**	**2.17**
Removed ($\alpha = 3.0$)	**-1.65**	**8.20**	**3.92**	**3.81**	**2.16**	**1.57**	**2.06**
Removed ($\alpha = 3.5$)	**-1.97**	**6.81**	**3.33**	**3.05**	**1.94**	**1.66**	**1.98**
Prop. ($\alpha = 0.5$)	**-2.22**	**7.25**	**4.02**	**3.84**	**2.36**	**1.77**	**2.2**
Prop. ($\alpha = 1.0$)	**-2.24**	**8.57**	**4.12**	**3.96**	**2.48**	**1.85**	**2.32**
Prop. ($\alpha = 1.5$)	**-2.29**	**8.85**	**4.21**	**4.40**	**2.82**	**2.01**	**2.75**
Prop. ($\alpha = 2.0$)	**-2.31**	**8.88**	**4.25**	**4.33**	**2.61**	**2.04**	**2.64**
Prop. ($\alpha = 2.5$)	**-2.34**	**8.98**	**4.28**	**4.27**	**2.73**	**2.00**	**2.67**
Prop. ($\alpha = 3.0$)	**-2.36**	**9.09**	**4.27**	**4.27**	**2.69**	**2.00**	**2.57**
Prop. ($\alpha = 3.5$)	**-2.36**	**9.22**	**4.16**	**4.24**	**2.49**	**1.90**	**2.54**

Fig. 6. Comparison of sound extraction accuracy in terms of signal-to-distortion ratio (SDR).

TABLE III
NUMBER OF ROWS OF VIDEO IN EACH α

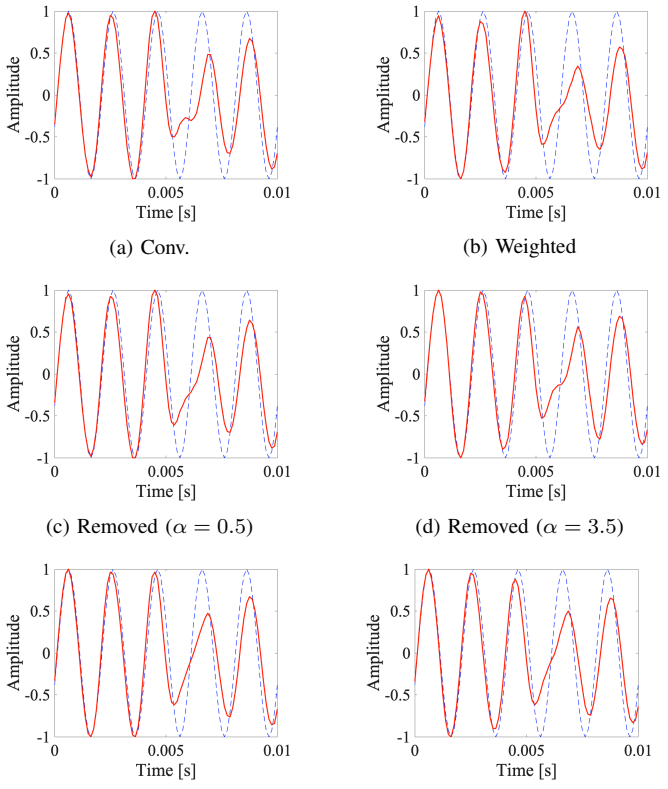
α in Eq. (4)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
In-focus area $\tilde{W}$	1,871	1,046	292	254	220	192	160

set to $\alpha = 0.5, 3.5$ in (4). The results show that the highest accuracy of sound extraction is achieved using Prop. method. This is because the value used to weight the phase variation was changed from amplitude to focal rate, which enabled the weighting of pixels with a higher degree of confidence.

These results show that the proposed method improves the sound quality of the extracted sound for sine waves of several frequencies.

V. CONCLUSION

In this paper, we have presented that the sound quality of the visual microphone can be improved by emphasizing the in-focus area of the captured video. Utilizing the focal rate that represents the degree of focus, we proposed a sound extraction method that emphasizes the displacement measured in the in-focus area of the captured video. Experimental results demonstrate that the proposed method is effective for improving the sound quality of the visual microphone. In the future, we will conduct experiments under different conditions, changing the pattern and material of the vibrating object, as well as the distance between the vibrating object and the sound source.



(a) Conv. (b) Weighted

(c) Removed ($\alpha = 0.5$) (d) Removed ($\alpha = 3.5$)

(e) Prop. ($\alpha = 0.5$) (f) Prop. ($\alpha = 3.5$)

Fig. 7. Time waveform (500 Hz) of sound extracted by each method (red line: extracted sound, blue dotted line: sound source) .

ACKNOWLEDGMENT

This work was partly supported by Ritsumeikan University R-GIRO and RARA, and JSPS KAKENHI Grant Numbers JP21H03488, JP23H03425, JP23K21691, JP23K28115, and JP24K20803.

REFERENCES

[1] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The Visual Microphone: Passive Recovery of Sound from Video," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 1–10, 2014.

[2] K. Terano, H. Shindo, K. Iwai, T. Fukumori, and T. Nishiura, "Sound capture from rolling-shuttered visual camera based on edge detection," In *Proceedings of 23rd International Congress on Acoustics*, pp. 2878–2884, 2019.

[3] O. Ait-Aider, N. Andreff, J. M. Lavest, and P. Martinet, "Exploiting Rolling Shutter Distortions for Simultaneous Object Pose and Velocity Computation Using a Single View," In *Proceedings of Fourth IEEE International Conference on Computer Vision Systems (ICVS'06)*, pp. 35–41, 2006.

[4] H. Nakano, T. Yoshizawa, Y. Geng, K. Iwai, and T. Nishiura, "Speech Quality Improvement Utilizing Out-of-Focus Areas in Rolling-Shutter

Video on Speech Extraction," In *Proceedings of 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2320–2325, 2023.

[5] J. Portilla and E. P. Simoncelli, "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients," *International Journal of Computer Vision*, vol. 40, pp. 49–70, 2000.

[6] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Phase-Based Video Motion Processing," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 1–10, 2013.

[7] T. Gautama and M. A. Van Hulle, "A phase-based approach to the estimation of the optical flow field using spatial filtering," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1127–1136, 2002.

[8] H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 188–200, 2002.

[9] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 2, pp. 317–330, 1986.

[10] O. R. Vincent, O. Folorunso, et al., , "A descriptive algorithm for sobel image edge detection," In *Proceedings of Information Science & IT Education Conference (InSITE)*, vol. 40, pp. 97–107, 2009.

[11] Multimedia Signal Processing and Security Lab, University of Salzburg, Salzburg, Austria, "Salzburg Texture Image Database (STex)," https://wavelab.at/sources/STex/, 2023.

Augmented sound-image perception using pre-virtual-leading ultrasounds based on precedence effect

Ryota Imanaka*, Yuting Geng*, Masato Nakayama† and Takanobu Nishiura*
* Ritsumeikan University, Osaka, Japan
E-mail: {is0545rs@ed, geng@fc, nishiura@is}.ritsumei.ac.jp
† Osaka Sangyo University, Osaka, Japan
E-mail: nakayama@ise.osaka-sandai.ac.jp

***Abstract*—In recent decades, stereo technologies have been proposed to control sound-images using binaural effects. Among these, some are based on the precedence effect, which is one of binaural effects. This phenomenon can lead the listener to localize the sound-image to a direction which is difficult to perceive with other binaural effects, so it can be utilized to augment sound perception in stereo audio. Moreover, emitted ultrasounds have been reportezd to change the listener's sensation of the sound, but how ultrasounds influence sound-image localization has not been fully considered yet. In this study, we focus on these two facts and propose a sound-image augmentation method with the precedence effect caused by ultrasounds. In this method, the sound-image is constructed with stereo audible sounds, while the precedence effect is provided by stereo ultrasounds. By utilizing the precedence effect, the perceived direction of a sound-image can be augmented, which is difficult with common stereo technologies. In this paper, we define these ultrasounds as pre-virtual-lead signals, as they are emitted before the main audible signals and provide a virtual perception of direction. Moreover, the ultrasounds are inaudible, so they will not affect listening to the contents, such as music and speech. In addition, ultrasounds with low-frequency envelopes are utilized as the precedence effect caused by them is expected to be perceived easily. We conducted subjective evaluation experiments on augmented sound perception by the proposed method and found the proposed method to be effective.**

I. Introduction

In recent decades, there has been a growing interest in technologies for reproducing a highly realistic sound field that can represent sound to listeners with a high sense of presence [1]. Highly realistic sound reproduction technologies can be broadly classified into two types: those based on psychoacoustic models, such as 5.1 ch. surround systems [2] and 22.2 ch. surround systems [3] and those based on physical acoustic models, such as wavefront synthesis [4] and ambisonics [5]. There are the methods for reproducing the direction of sound on psychoacoustic models [6] $\sim$ [8]. These are based on an extension of sound-image control using stereo technology and have been proposed to control sound-images using binaural effects [9].

Some of these methods are based on the precedence effect, which is one of the binaural effects. The precedence effect is a phenomenon in which a sound-image localizes to the direction of the first direct sound when similar sounds arrive to the listener from different directions with a slight time difference [10], [11]. This phenomenon can lead the listener to localize the sound-image to a direction that is difficult to perceive with other binaural effects, so it can be utilized to augment sound perception in stereo audio. Moreover, sound-image control methods using the precedence effect have been proposed. In these methods, sound emitted from loudspeakers is delayed appropriately, and the localization of sound-images can be controlled due to the precedence effect. These methods have been used for evacuation guidance [12], [13]. In these methods, the sound-image can be presented only in the loudspeaker direction. In this study, we utilize this phenomenon to augment sound perception between the stereo loudspeakers. Moreover, it has been reported that emitted ultrasounds change the listener's sensation of the sound, and this phenomenon is called the hypersonic effect [14]. Based on this report, we consider the possibility that the ultrasounds may influence sound-image localization.

In this study, we focus on these two facts and propose a sound-image augmentation method with the precedence effect caused by ultrasounds. In this method, the sound-image is constructed with stereo audible sounds, while the precedence effect is provided by stereo ultrasounds. The sound-image constructed by the conventional time panning method is augmented by adding signals that provide the precedence effect. By utilizing the precedence effect, the perceived direction of the sound-image can be augmented, which is difficult with common stereo technologies. We add signals before the main signals are emitted, and these additional signals provide a leading perception of direction. Moreover, we use ultrasound signals as pre-lead signals. The ultrasounds are inaudible, so they will not affect listening to the contents, such as music and speech. Therefore, we define these ultrasound signals as pre-virtual-lead signals. In addition, ultrasounds with low-frequency envelopes are utilized as the precedence effect caused by them is expected to be perceived easily. In this paper, we investigate the influence of the perception of sound-image by ultrasounds, and we also investigate the perception of the sound-image changes when pre-virtual-lead signals are

added to the main signals that construct sound-images. We conducted subjective evaluation experiments for augmented sound localization by the proposed method and found the proposed method to be effective.

II. Sound-image Perception

In the psychoacoustic models, sound-images are presented by reproducing the sensations at the listening point caused by binaural effects, such as interaural time difference (ITD) and interaural level difference (ILD) [15]. The binaural effect refers to the perception of sound direction, location, and spatial impression based on the difference in arrival time and sound pressure perceived by each ear when listening to a sound with both ears. In the horizontal direction, the direction of sound is mainly perceived the basis of the ITD and ILD [15].

A. Interaural Time Difference (ITD)

The ITD is the arrival time difference in sound emitted from a certain position due to the distance between the ears. In the horizontal direction, it is one of the cues to perceive the direction and position of sound along with the ILD, and its sensitivity maximize in the range from 500 to 1,000 Hz. In the frequency range above about 1,600 Hz, human beings cannot sense the ITD and perceive sound-images in the horizontal direction due to the ILD [16].

In sound-image construction using stereo technology, the direction and position of a sound-image perceived by the listener can be controlled by adding a time difference to the signals emitted from each loudspeaker in accordance with the sound-image construction direction. This time difference can be obtained by the sound-image construction direction, interaural distance, and sound speed, and its value is about 0.5 ms at the maximum. Up to this value, as the time difference between the signals emitted from each channel increases, the sound-image moves from the center between the two loudspeakers to the direction of the loudspeaker that emits the preceding sound. When the time difference between the sound emitted from each channel exceeds this maximum value, the sound-image localizes to the loudspeaker position due to the precedence effect [17].

B. Interaural Level Difference (ILD)

The ILD is the level difference in sound emitted from a certain position due to the distance between the ears. In the horizontal direction, it is one of the cues to perceive the direction and position of sound along with the ITD, and it functions as a cue to perceive the horizontal directions in the entire audible range [11]. In sound-image construction using stereo technology, the direction and position of a sound-image perceived by the listener can be controlled by adding a sound pressure difference to the sound emitted from each loudspeaker in accordance with the sound-image construction direction. This method is called amplitude panning [6] $\sim$ [8].

C. Precedence Effect

The precedence effect is one of the perceptions related to the direction and position of sound. When similar sounds are emitted from different directions with a slight time difference, the sound-image localizes to the direction of the first direct sound [10], [11]. This effect can lead the listener to localize the sound-image to a direction that is difficult to perceive with other binaural effects, so it can be utilized to augment sound perception. As the conditions for the precedence effect, the time difference between the preceding sound and the following sound should be about 2 to 100 ms [11], which is much larger than the maximum ITD. Therefore, this effect is a phychological perception that does not correspond with physical phenomena. When the time difference is large, listeners perceive the sound-images as separating into two sounds. When the time difference is larger than about 0.5 ms and smaller than about 2 ms, listeners perceive the sound-image in the middle of the two sound sources.

D. Discussions for Perception of Ultrasounds

It has been reported that emitted ultrasounds change the listener's sensation of the sound [14], but how ultrasounds influence sound-image localization has not been fully considered yet. Ultrasounds are inaudible, but human beings perceive these sounds other than by hearing. For example, tactile feedback methods have been proposed in which pressure is exerted on the skin by emitting amplitude-modulated ultrasonic signals from loudspeakers [18]. Since the frequency response of this vibrotactile sensation maxim at approximately 250 Hz [19], ultrasounds with low-frequency envelopes are utilized in this study as the precedence effect caused by them is expected to be perceived easily.

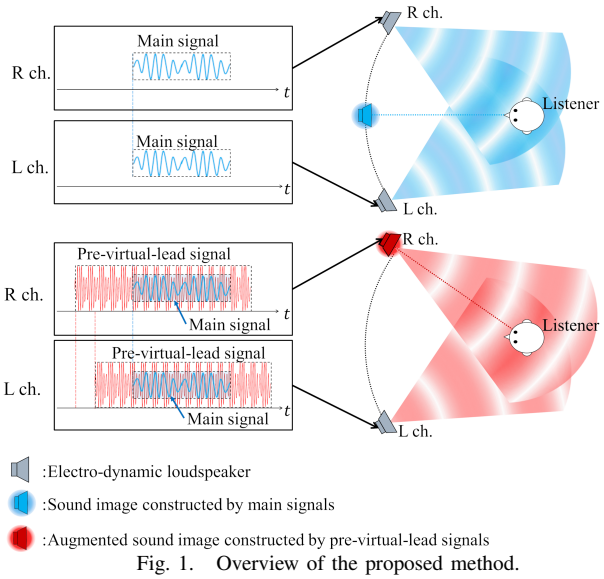


Fig. 1. Overview of the proposed method.

III. Proposed Augmented Sound Perception Using Pre-virtual-leading Ultrasounds Based on Precedence Effect

A. Overview of the Proposed Method

Fig. 1 shows an overview of the proposed method. In this paper, we focus on the precedence effect, which is one of the binaural effects. We propose a method of augmented sound perception in a sound-image control method using the stereo technology by adding pre-virtual-lead signals due to the precedence effect. First, we add delay to main signals, so that main signals are emitted after pre-virtual-lead signals. Next, we design the pre-virtual-lead signals. In this method, we use ultrasound signals for the lead signals. The ultrasounds are inaudible, so they will not affect listening to the main signals. On the other hand, ultrasound signals are difficult to perceive. Therefore, because the sensitivity of the human sense of vibrotactile is highest at around 250 Hz [19], we use signals with upper side band (USB) amplitude modulation [20] of a low-frequency signal around 250 Hz as pre-virtual-lead signals. In this way, pre-virtual-lead signals in the ultrasonic band can be designed while maintaining the low-frequency envelope. As a result, these signals stimulate the listener's vibrotactile sensations, and they can perceive the pre-virtual-lead signals. Finally, we add time delays to the designed pre-virtual-leading signals. These signals provide the precedence effect and are added to the delayed main signals. In this way, augmented sound perception using ultrasound pre-virtual-leading signals is achieved.

B. Design of Signals in the Proposed Method

Fig. 2 shows the signal design flow. Where $s_{\mathrm{m}}(t)$ is the main signal, $\tau_{\mathrm{m_L}}$ and $\tau_{\mathrm{m_R}}$ are the delay time added to the main signals emitted from each channel, $s_{\mathrm{pvl}}(t)$ is the pre-virtual-lead signal, $s_{\mathrm{c}}(t)$ is the carrier wave, $s'_{\mathrm{pvl}}(t)$ is the USB modulated pre-virtual-lead signal, $\tau_{\mathrm{pvl_L}}$ and $\tau_{\mathrm{pvl_R}}$ are the delay time added to pre-virtual-lead signals emitted from each channel, and $s_{\mathrm{L}}(t)$ and $s_{\mathrm{R}}(t)$ are the signals emitted from each channel. Fig. 3 shows the concept of the signals in the proposed method, where $s'_{\mathrm{m_L}}(t)$ and $s'_{\mathrm{m_R}}(t)$ are the main signals emitted from each channel after adding delay, f_{pvl} is the envelope frequency of the pre-virtual-lead signal, f_{c} is the carrier frequency of the USB amplitude modulation, f_{main} is the highest frequency of main signal, $\tau_{\mathrm{len_{pvl}}}$ is the length of the pre-virtual-lead signal, and $\tau_{\mathrm{len_m}}$ is the length of the main signal. First, in Step 1 in Fig. 2, we add delay to the main signals, so that the main signal is emitted after the pre-virtual-lead signal. $\tau_{\mathrm{m_L}}$ and $\tau_{\mathrm{m_R}}$ are calculated based on the time panning method [6]. Equations (1) and (2) show the signals to which delay is added.

$$s'_{\mathrm{m_L}}(t) = s_{\mathrm{m}}(t-\tau_{\mathrm{m_L}}), \quad (1)$$

$$s'_{\mathrm{m_R}}(t) = s_{\mathrm{m}}(t-\tau_{\mathrm{m_R}}), \quad (2)$$

Next, in Step 2, we design the pre-virtual-lead signals. A pre-virtual-lead signal is an ultrasound signal, so we add a low-frequency envelope to it. The USB amplitude modulation

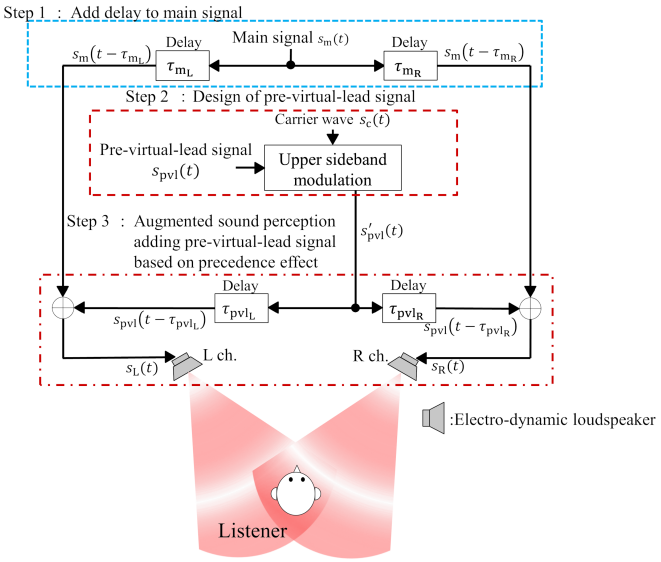


Fig. 2. Block diagram of the proposed method.

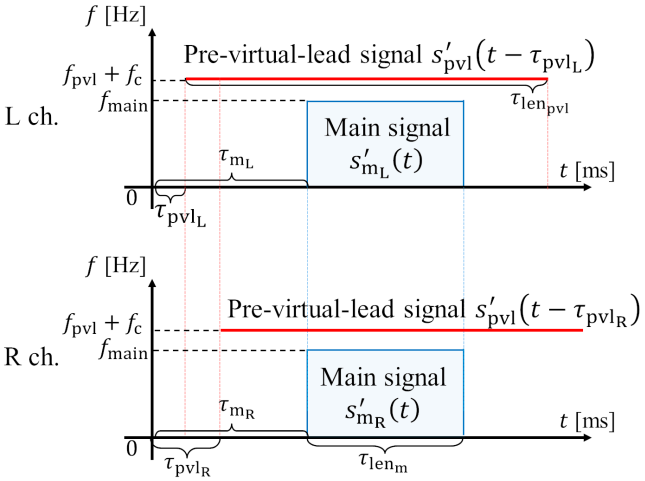


Fig. 3. Concept of the proposed method in time-frequency domain.

wave and its modulation factor are expressed by Equations (3) and (4), respectively [20]. Since the sensitivity of the human sense of vibrotactile is highest at around 250 Hz [19], we use a low-frequency signal that is modulated by ultrasound signal. In this way, a pre-virtual-lead signal can be designed with a low-frequency envelope while keeping it in the ultrasonic band. Equations (3) and (4) show the equations for the USB modulated wave.

$$s'_{\mathrm{pvl}}(t) = \{1+ms_{\mathrm{pvl}}(t)A_{\mathrm{c}}\cos(2\pi f_{\mathrm{c}}t)\} - m\hat{s}_{\mathrm{pvl}}(t)A_{\mathrm{c}}\sin(2\pi f_{\mathrm{c}}t), \quad (3)$$

$$m = \frac{A_{\mathrm{pvl}}}{A_{\mathrm{c}}}, \quad (4)$$

where A_{C} is the maximum amplitude of the carrier wave, m is the modulation factor $(0 < m \leq 1)$, A_{pvl} is the maximum amplitude of the pre-virtual-lead signal, and $\hat{s}_{\mathrm{pvl}}(t)$ is the pre-virtual-lead signal with a phase shift of $\pi/2$. In addition, the USB modulated pre-virtual-lead signal can be expanded by

using the trigonometric addition formulas as in Equation (5).

$$s'_{\mathrm{pvl}}(t) \quad = \quad A^2_{\mathrm{pvl}}\cos\{2\pi(f_{\mathrm{c}} + f_{\mathrm{pvl}})t\}. \tag{5}$$

In this case, if the frequency of the pre-virtual-lead signal $f_{\mathrm{c}} + f_{\mathrm{pvl}}$ is audible or below f_{main}, the signal affects listening to the contents of the main signals. Therefore, the constraint conditions of the pre-virtual-lead signals are shown in Equations (6) and (7).

$$f_{\mathrm{pvl}} \quad > \quad f_{\mathrm{main}}, \tag{6}$$

$$f_{\mathrm{pvl}} \quad \geq \quad 20000 \text{ Hz}. \tag{7}$$

Finally, in Step 3, we add time delays to the designed pre-virtual-leading signals. These signals provide the precedence effect, and are added to the delayed main signals. In this way, augmented sound perception using ultrasound pre-virtual-leading signals is achieved. $s'_{\mathrm{pvl}}(t)$ modulated by USB in Step 2 is delayed on each channel and added to the main signal $s'(t)$ generated in Step 1. Equations (8) and (9) show the equations for the signals emitted from each channel.

$$s_{\mathrm{L}}(t) \quad = \quad s'_{\mathrm{m_L}}(t) + s'_{\mathrm{pvl_L}}(t - \tau_{\mathrm{pvl_L}}), \tag{8}$$

$$s_{\mathrm{R}}(t) \quad = \quad s'_{\mathrm{m_R}}(t) + s'_{\mathrm{pvl_R}}(t - \tau_{\mathrm{pvl_R}}). \tag{9}$$

As a result, the precedence effect occurs in the pre-virtual-lead signals, and an augmented sound perception method using ultrasound pre-virtual-lead signals is achieved.

IV. Subjective Evaluation Experiment

A. Experimental Conditions

To evaluate the effectiveness of the proposed method, we conducted subjective evaluation experiments. In our experiments, we used $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\text{ Hz}, 20,000\text{ Hz})$, $(250\text{ Hz}, 20,000\text{ Hz}), (600\text{ Hz}, 20,000\text{ Hz})$ as the pre-virtual-lead signals. In each condition, we used two modulation factors, $m = 1.00, 0.05$. In addition, experiments were conducted using a sine wave at 20,000 Hz as a pre-virtual-lead signal. The delay times of the pre-virtual-lead signals in each channel $\tau_{\mathrm{pvl_L}}$, $\tau_{\mathrm{pvl_R}}$ are $(\tau_{\mathrm{pvl_L}}, \tau_{\mathrm{pvl_R}}) = (2\text{ ms}, 0\text{ ms}), (0\text{ ms}, 2\text{ ms})$, $(0\text{ ms}, 0\text{ ms})$, and either the left or right pre-virtual-lead signal preceded the other or the left and right pre-virtual-lead signals were emitted simultaneously. This delay time is equivalent to the direction of each loudspeaker in these experimental conditions. The delay times of main signals $\tau_{\mathrm{m_L}}$ and $\tau_{\mathrm{m_R}}$ were set in $\tau_{\mathrm{m_L}} = \tau_{\mathrm{m_R}}$.This makes the sound-image localized in front of the subject. The time difference between the preceding pre-virtual-lead signal and main signals was 1,000 ms. The length of the pre-virtual-lead signals $\tau_{\mathrm{len_{pvl}}}$ was set to $s'_{\mathrm{m_L}}(t) = 5{,}000$ ms, and the length of the main signal $\tau_{\mathrm{len_m}}$ was set to 20 ms. For the main signal, we used white noise with a bandwidth of 0 to 16,000 Hz. The equipment and the subject arrangement in the experiments are shown in Fig. 4. In the experiments, we had subjects sit in the position shown in Fig. 4. The subjects were asked to answer whether the perceived sound-image was "shifted to the left," "shifted to the right," or "stationary" compared to when the signal was emitted without the pre-virtual-lead signal. We define the correct answer as being when the direction of the sound-image perceived by the subject matched the direction of the

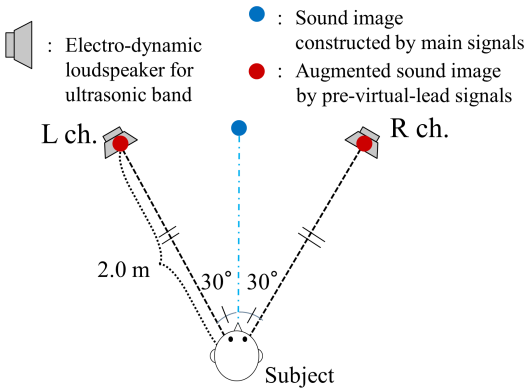


Fig. 4. Arrangement in the experiment.

preceding pre-virtual-lead signal, and we calculated the correct answer rate. In the case when the left and right pre-virtual-lead signals were emitted simultaneously, we specify that the correct answer is "stationary". The number of subjects was 10 (9 males and 1 female), aged 21 $\sim$ 23 years. Experimental conditions and equipment used are listed in Tables I and II.

TABLE I
Experimental conditions.

Environment	Office room
Reverberation time	T_{60} = 1.5 s
Ambient noise level	L_{A} = 32.4 dB
Sampling frequency	96,000 Hz
Quantization	24 bits

TABLE II
Experimental equipment.

Electro-dynamic loudspeaker	Technics, SB-G90M2
Power amplifier	Technics, SU-G700M2
Audio interface	RME, Fireface UFX

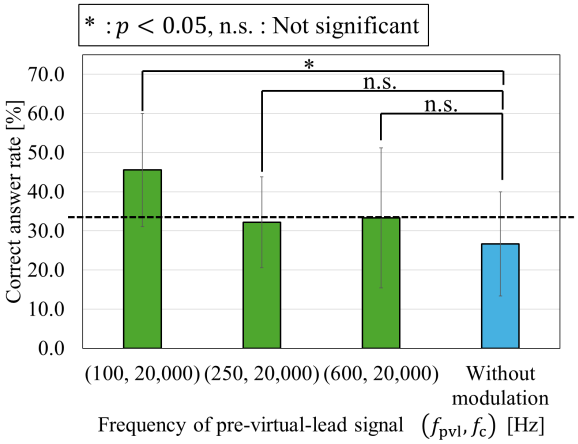


Fig. 5. Correct answer rate for each $(f_{\mathrm{pvl}}, f_{\mathrm{c}})$, $(m = 1.00)$.

B. Experimental Results

Figs. 5 and 6 show the correct answer rate for each condition at each modulation factor m. The horizontal axis is

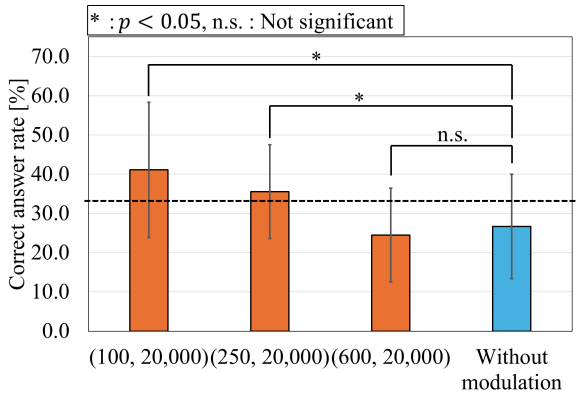


Fig. 6. Correct answer rate for each $(f_{\mathrm{pvl}}, f_{\mathrm{c}})$, $(m = 0.05)$.

TABLE III
PERCENTAGES OF RESPONSES AND PERCENTAGES OF CORRECT ANSWERS
$(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$, $m = 1.00$.

Response \ Ground truth	Shift to left	Shift to right	Stationary
Shift to left	**40.0%**	16.7%	36.7%
Shift to right	36.7%	**63.3%**	30.0%
Stationary	23.3%	20.0%	**33.3%**

$*$ Correct answer rate : 45.6%

TABLE IV
PERCENTAGES OF RESPONSES AND PERCENTAGES OF CORRECT ANSWERS
$(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$, $m = 0.05$.

Response \ Ground truth	Shift to left	Shift to right	Stationary
Shift to left	**43.3%**	36.7%	33.3%
Shift to right	36.7%	**50.0%**	36.7%
Stationary	20.0%	13.3%	**30.0%**

$*$ Correct answer rate : 41.1%

TABLE V
PERCENTAGES OF RESPONSES AND PERCENTAGES OF CORRECT ANSWERS
$(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (250\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$, $m = 0.05$.

Response \ Ground truth	Shift to left	Shift to right	Stationary
Shift to left	**40.0%**	33.3%	30.0%
Shift to right	46.7%	**43.3%**	46.7%
Stationary	13.3%	23.4%	**23.3%**

$*$ Correct answer rate : 35.6%

the frequency of the pre-virtual-lead signals, and the vertical axis is the correct answer rate. Error bars indicate the range of standard deviations added positively or negatively to the percentage correct for each condition. The dashed line indicates a reference level of 33.3%, which is the expected value for the correct answer rate when answering randomly. Also, Tables III, IV, and V list that the percentages of responses and percentages of correct answers. Table III lists the condition in $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$, modulation factor $m = 1.00$, Table IV lists the condition in $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$, modulation factor $m = 0.05$, and Table V lists the condition in $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (250\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$, modulation factor $m = 0.05$.

From Figs. 5 and 6, $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$ at modulation factor $m = 1.00$, and $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$, $(250\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$ at modulation factor $m = 0.05$, the correct answer rate obtained exceeded the expected value (33.3%). Also, from Tables III, IV, and V, we see that $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$ at modulation factor $m = 1.00$ and $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$ at modulation factor $m = 0.05$, the percentages of responses that matched the correct answers were higher than the percentage of the other options chosen. Therefore, we found the proposed method to be effective when $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$ at modulation factor $m = 1.00, 0.05$. In addition, we conducted paired t-tests on the correct answer rates for each frequency and modulation factor condition, as well as the correct answer rate for the no modulation condition. In this case, there are siginificant differences at the significance level $\alpha = 0.05$ in the condition of $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$ in modulation factor $m = 1.00$, and $(f_{\mathrm{pvl}}, f_{\mathrm{c}}) = (100\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$, $(250\ \mathrm{Hz}, 20,000\ \mathrm{Hz})$ in the modulation factor $m = 0.05$. Therefore, we demonstrated the effectiveness of modulating the pre-virtual-lead signals and enveloping their amplitude.

V. CONCLUSIONS

In this paper, we proposed a method of augmented sound perception by adding pre-virtual-lead signals to the signal that is presented in front of the listener. The results of subjective evaluation experiments showed that the proposed method can augment perception of the presented sound-image, but it has a problem in terms of localization accuracy in the desired direction. In the future, we will evaluate the effectiveness of the proposed method when the main signals are presented from directions other than in front of the listener. In addition, the pre-virtual-lead signals have an attack part and a sustain part. We intend to evaluate the contribution rate of these parts. Moreover, we will investigate the phase difference between the binaural ears when the pre-virtual-lead signals arrive at the listener. Specifically, by changing the modulation frequency of the pre-virtual-lead signals and using a signal whose frequency fluctuates over time, we will consider signals to augment left-right localization using the pre-virtual-lead signals for the sound-image presented to the listener.

ACKNOWLEDGMENTS

This work was partly supported by Ritsumeikan University R-GIRO and RARA, and JSPS KAKENHI Grant Numbers JP21H03488, JP21H04427, JP23H03425, JP23K21691, JP23K28115 and JP24K20803.

REFERENCES

[1] A. Ando, "Theory of Three-Dimensional Sound Field Reproduction," IEICE, SP, Fundamentals Review, vol. 3, no. 4, pp. 33-46, 2010.
[2] ITU-R Recommendation BS. 775-4, "Multichannel stereophonic sound system with and without accompanying picture," International Telecommunication Union, 2022.
[3] K. Hamasaki, "A 22.2 multichannel sound system for ultrahigh-definition TV (UHDTV)," SMPTE Motion Imaging Journal, vol. 117, no. 3, pp. 40-49, 2008.

[4] J. Ahrens, “Analytic Methods of Sound Field Synthesis,” Springer Berlin Heidelberg, 2012.
[5] R. Nishimura, “5. Ambisonics”, The Journal of The Institute of Image Information and Television Engineers, vol. 68, no. 8, pp. 616-620, 2014.
[6] V. Plukki, “Virtual sound source positioning using vector base amplitude panning,” Journal of the Audio Engineering Society, vol. 45, no. 6, pp. 456-466, 1997.
[7] J. Blauert, “Spatial Hearing: The Psychophysics of Human Sound Localization,” The MIT Press, 1996.
[8] D. M. Leakey, “Some Measurements on the Effects of Interchannel Intensity and Time Differences in Two Channel Sound Systems,” The Journal of the Acoustical Society of America, vol. 31, no. 7, pp. 977-986, 2005.
[9] K. Hamasaki, “2-3. Sound Technology for Reproducing Spatial Sound with Presence and Reality,” The Journal of The Institute of Image Information and Television Engineers, vol. 65, no. 5, pp. 604-609, 2011.
[10] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, “The precedence effect,” The Journal of the Acoustical Society of America, vol. 106, no. 4, pp. 1633-1654, 1999.
[11] H. Wallach, E. B. Newman, and M. R. Rosenzweig, “The Precedence Effect in Sound Localization,” The American Journal of Psychology, vol. 62, no. 3, pp. 315-336, 1949.
[12] T. Miyoshi, “Evaluation of Recognition Performance for Direction Detection of Sequential Sound in Evacuation Guidance System,” Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, vol. 32, no. 5, pp. 891-896, 2020.
[13] T. Miyoshi, ”Characteristic of Sound Localization for Emergency Guide System using Sound Control,” Proceedings of the Fuzzy System Symposium, vol. 34, pp. 24-26, 2018
[14] T. Oohashi, E. Nishina, M. Honda, Y. Yonekura, Y. Fuwamoto, N. Kawai, T. Maekawa, S. Nakamura, H. Fukuyama, and H. Shibasaki. ”Inaudible high-frequency sounds affect brain activity: hypersonic effect,” Journal of Neurophysiology, vol.83, no. 6, pp. 3548-3558, 2000.
[15] A. Roginska and P. Geluso, ”Immersive Sound The Art and Science of Binaural and Multi-Channel Audio,” Routledge, 2017.
[16] A. Brughera, L. Dunai, and W. M. Hartmann, “Human interaural time difference thresholds for sine tones: The high-frequency limit,” The Journal of the Acoustical Society of America, vol. 133, no. 5, pp. 2839-2855, 2013.
[17] Y. Nobe and Y. Kajikawa, “A Study on Reproduction of Varied Sound Image Localization for Real Source in Stereo Audio System,” IEICE technical report, vol. 109, no. 436, pp. 125-130, 2010.
[18] A. Ueda, T. Takahashi, and M. Nakayama, ”Texture Reproduction of Ultrasonic Mid-Air Haptics Based on Amplitude Modulation Signal Generation Using Fricative Sounds Feature Extraction and Hand Tracking,” IEICE Technical Report, vol. 122, no. 387, pp.171-176, 2023.
[19] E.R. Kandel, J.H. Schwartz, and T. Jessell, ”Principles of Neural Science,” McGraw-Hill Companies, 2000
[20] Y. Wang, X. Li, L. Xu, and L. Xu, ”SSB Modulation of the Ultrasonic Carrier for a Parametric Loudspeaker,” International Conference on Electronic Computer Technology 2009, pp. 669-673, 2009

Virtual multi-boosted amplitude modulation toward high-pressure audible sound with parametric array loudspeakers

Yoto IKEZAKI*, Yuting GENG*, Masato NAKAYAMA†, and Takanobu NISHIURA*
* Ritsumeikan University, Osaka, Japan
E-mail: {is0492ss@ed, geng@fc, nishiura@is}.ritsumei.ac.jp
† Osaka Sangyo University, Osaka, Japan
E-mail: nakayama@ise.osaka-sandai.ac.jp

***Abstract*—A parametric array loudspeaker (PAL) can achieve sharp directivity in the audible band by using ultrasounds. Ultrasounds are emitted intensely from a PAL, and the difference frequency component self-demodulates because of nonlinear interactions in the air. As a result, the sound pressure of the audible sound reproduced by a PAL is lower than that reproduced by a conventional electro-dynamic loudspeaker. This makes it difficult to use a PAL in a noisy environment. To solve this problem, we modify conventional modulations to improve the demodulated sound pressure. When the amplitude-modulated wave is emitted as a primary wave, distortion of the primary wave generates various components, including the audible sound, as higher-order components. Conventional modulation methods have difficulty using these components of the ultrasound band effectively. Hence, we propose virtual multi-boosted amplitude modulation toward high-pressure audible sound with PALs. In the proposed method, the modulated wave is designed so that frequency components of the carrier waves and the demodulated sound are generated as the higher-order components, which are the distortion components of the primary waves. These components can boost the carrier waves and the demodulated sound, thus increasing the demodulated sound pressure. The effectiveness of the proposed method is demonstrated by comparing the audible sound pressure of the proposed and conventional methods.**

I. Introduction

Loudspeakers are commonly used to reproduce audible sounds like speech and music [1], [2]. Because a typical electro-dynamic loudspeaker has a wide directivity and propagates the sound over a wide area, non-target listeners may regard the spreading sound as noise. Recently, a parametric array loudspeaker (PAL) has attracted attention because of its sharp directivity [3], [4].

A PAL reproduces the audible sound by using nonlinear interactions in the air. When two ultrasounds of frequencies f_1 and f_2 are emitted intensely as primary waves, sounds at the difference frequency $|f_1 - f_2|$, the sum frequency $f_1 + f_2$, and the harmonic frequencies $2f_1$ and $2f_2$ are generated as secondary waves due to the nonlinear interactions in the air [5], [6]. Secondary waves as well as higher-order components are generated. Through the application of this principle, an audible sound f_s can be reproduced by emitting a modulated wave. The modulated wave typically comprises a carrier wave with frequency f_c and sideband waves with frequencies $f_\mathrm{s} + f_\mathrm{c}$ and $f_\mathrm{s} - f_\mathrm{c}$[7], [8]. When the modulated wave is emitted from the PAL, the difference frequency between the carrier wave and the sideband waves f_s is self-demodulated. The generated virtual secondary sources form an end-fire array within the primary beam. Thus, the audible sound can achieve very narrow beams even at low frequencies [9].

However, as the demodulated sound is generated by nonlinear interactions, it has much lower sound pressure than the emitted ultrasounds [3], [9]. Therefore, the signal-to-noise ratio becomes small in noisy environments, making it difficult to hear the reproduced sound. A solution to this problem is to use multiple PALs to improve the demodulated sound pressure [10], [11]. However, the multiple PALs must be large, limiting the locations where they can be installed. Hence, we aim to improve the demodulated sound pressure of small PALs by modifying a modulation method.

In particular, one such method is amplitude modulation, which modulates the ultrasound carrier wave's amplitude in accordance with the target sound's amplitude [12]$\sim$ [14]. Other modulation methods have been proposed to improve demodulated sound pressure, such as a frequency modulation method [15], modulation methods combining frequency and amplitude-modulated waves [16], and modulation methods using square-wave carriers [17]. However, although these methods can increase the demodulated sound pressure, the sound quality is degraded because many frequency components other than the audible sound are generated in the audible band. In addition, there are multiplexed double sideband modulation [18], [19] for improved sound quality and modulation for reduced harmonic distortion [20]. However, with the above modulation methods, these higher-order components contribute little to the demodulated sound. To overcome this issue, a method was proposed that uses the modulated wave's higher-order components [21], [22]. However, this method is less effective for target sounds without harmonic structure. Hence, we aim to improve the demodulated sound pressure by effectively applying the modulated wave's higher-order components, regardless of the type of target sound.

In this paper, we propose a virtual multi-boosted amplitude modulation toward high-pressure audible sound with PALs. In the proposed method, the modulated wave is designed so that its higher-order components, which are the distortion components of the primary waves, match the frequencies of the carrier waves and the target sound.

The modulated wave of the proposed method is generated by combining two modulated waves with different carrier

frequencies. When this modulated wave is emitted intensely from the PALs, the audible sound self-demodulates in the same way as conventional methods. On the other hand, a feature of the proposed method is that the interactions of the two carrier waves also generate frequency components that match the frequencies of the carrier waves. As the carrier waves are boosted by the generation of higher-order components, the demodulated sound is consequently also boosted, regardless of the characteristics of the target sound. Furthermore, the interactions of the secondary waves also generate frequency components that match the frequency of the target sound. These frequency components are continuously distributed in the direction of propagation of the primary waves as virtual sources without substance, similar to the demodulated sound. Therefore, higher-order components other than the demodulated sound boost the carrier waves and demodulated sound while maintaining sharp directivity. Thus, the effective use of the distortion components of the modulated wave improves the demodulation efficiency and achieves a high-pressure audible sound with PALs. In a previous study, we showed the effectiveness of the proposed method in numerical simulations of demodulated sound pressure [23]. In this study, we conducted an evaluation experiment to evaluate the proposed method's effectiveness in a real environment.

II. Proposed virtual multi-boosted amplitude modulation

A. Overview of the proposed method

We propose virtual multi-boosted amplitude modulation toward high-pressure audible sound with PALs. Fig.1 shows a block diagram of the proposed method. In the proposed method, the modulated wave is designed so that the higher-order components, which are the distortion components of the primary waves, match the frequency of the carrier waves and target sound. The modulated wave is generated by combining two modulated waves with different carrier frequencies. Specifically, a low-frequency carrier wave $v_{\mathrm{c_l}}(t)$ and a high-frequency carrier wave $v_{\mathrm{c_h}}(t)$ are generated with a carrier frequency f_{c} and a frequency ratio γ. The low-frequency carrier wave $v_{\mathrm{c_l}}(t)$ is modulated by the target sound $v_{\mathrm{s}}(t)$ via lower sideband (LSB) modulation [13], while the high-frequency carrier wave $v_{\mathrm{c_{high}}}(t)$ is modulated by $v_{\mathrm{s}}(t)$ via upper sideband (USB) modulation [13]. The LSB-modulated wave $v_{\mathrm{l}}(t)$ and the USB-modulated wave $v_{\mathrm{h}}(t)$ are weighted by w_{l} and w_{h} and then added to generate the modulated wave $v(t)$ used in the proposed method. Details of the design method for the modulated waves used in the proposed method are given in Section II-B.

When this modulated wave $v(t)$ is emitted intensely from the PALs, the audible sound $v'_{\mathrm{s}}(t)$ self-demodulates in the same way as in conventional methods. On the other hand, a feature of the proposed method is that the interactions of the two carrier waves also generate frequency components that match the frequencies of the carrier waves. Furthermore, the interactions of the secondary waves also generate frequency components that match the frequency of the target sound. Details of the demodulation process are given in Section II-C. As a result, the higher-order components of the modulated wave boost the carrier waves and demodulated sound. Therefore, the distortion components of the modulated wave are effectively used to improve demodulation efficiency and to achieve high-pressure audible sound with PALs.

B. Design of virtual multi-boosted amplitude-modulated wave

Fig.2 shows the design flow of the modulated wave for virtual multi-boosting in the frequency domain. The frequencies of $v_{\mathrm{c_l}}(t)$ and $v_{\mathrm{c_h}}(t)$ are denoted by $f_{\mathrm{c_l}}$ and $f_{\mathrm{c_h}}$, respectively. These carrier wave frequencies can be expressed by the following equation:

$$f_{\mathrm{c_l}} = f_{\mathrm{c}}, \tag{1}$$
$$f_{\mathrm{c_h}} = \gamma f_{\mathrm{c}}, \tag{2}$$

where the setting of the value of γ is described in Section II-C. In addition, we should guarantee that no secondary waves other than the demodulated sound wave are generated in the hearing range. This constraint can be expressed as follows:

$$f_{\mathrm{c_h}} - f_{\mathrm{c_l}} \geq \text{Upper bound of hearing range.} \tag{3}$$

By using two carrier frequencies, the low-frequency carrier wave $v_{\mathrm{c_l}}(t)$ and the high-frequency carrier wave $v_{\mathrm{c_h}}(t)$ can be expressed as follows:

$$v_{\mathrm{c_l}}(t) = A_{\mathrm{c}} \cos(2\pi f_{\mathrm{c_l}} t), \tag{4}$$
$$v_{\mathrm{c_h}}(t) = A_{\mathrm{c}} \cos(2\pi f_{\mathrm{c_h}} t), \tag{5}$$

where A_{c} is the respective maximum amplitudes of the carrier wave and t is the time index.

Next, the target sound $v_{\mathrm{s}}(t)$ is amplitude-modulated with the determined carrier frequency. The LSB-modulated wave $v_{\mathrm{l}}(t)$ and the USB-modulated wave $v_{\mathrm{h}}(t)$ can be represented by Equations (6) and (7), respectively:

$$\begin{aligned} v_{\mathrm{l}}(t) = {} & \{1 + m_{\mathrm{ssb}} v_{\mathrm{s}}(t)\} A_{\mathrm{c}} \cos(2\pi f_{\mathrm{c_l}} t) \\ & + m_{\mathrm{ssb}} \widehat{v}_{\mathrm{s}}(t) A_{\mathrm{c}} \sin(2\pi f_{\mathrm{c_l}} t), \end{aligned} \tag{6}$$

$$\begin{aligned} v_{\mathrm{h}}(t) = {} & \{1 + m_{\mathrm{ssb}} v_{\mathrm{s}}(t)\} A_{\mathrm{c}} \cos(2\pi f_{\mathrm{c_h}} t) \\ & - m_{\mathrm{ssb}} \widehat{v}_{\mathrm{s}}(t) A_{\mathrm{c}} \sin(2\pi f_{\mathrm{c_h}} t), \end{aligned} \tag{7}$$

where m_{ssb} is modulation factor ($0 < m_{\mathrm{ssb}} \leq 1$) and $\widehat{v}_{\mathrm{s}}(t)$ is a signal with a phase shift of $\pi/2$ added to $v_{\mathrm{s}}(t)$.

The LSB-modulated wave $v_{\mathrm{l}}(t)$ and the USB-modulated wave $v_{\mathrm{h}}(t)$ are weighted by w_{l} and w_{h} and then added to generate the modulated wave for virtual multi-boosting $v(t)$. The modulated wave for virtual multi-boosting $v(t)$ is represented by the following equation:

$$v(t) = w_{\mathrm{l}} v_{\mathrm{l}}(t) + w_{\mathrm{h}} v_{\mathrm{h}}(t). \tag{8}$$

C. Demodulation process of virtual multi-boosted amplitude-modulated wave

In the proposed method, the value of the frequency ratio γ is set so that the interaction between the LSB-modulated wave $v_{\mathrm{l}}(t)$ and the USB-modulated wave $v_{\mathrm{h}}(t)$ generates frequency components that match the frequencies of carrier waves or target sound as the higher-order components. When the two carrier waves are emitted, sounds at the difference frequency $(\gamma-1) f_{\mathrm{c_l}}$, the sum frequency $(\gamma+1) f_{\mathrm{c_l}}$, and doubled

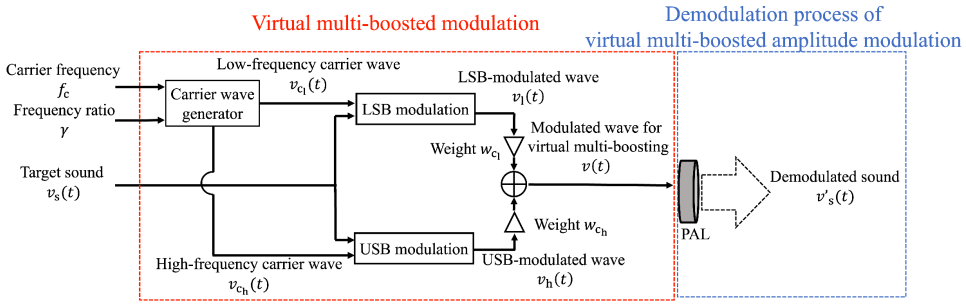


Fig. 1: Block diagram of the proposed method

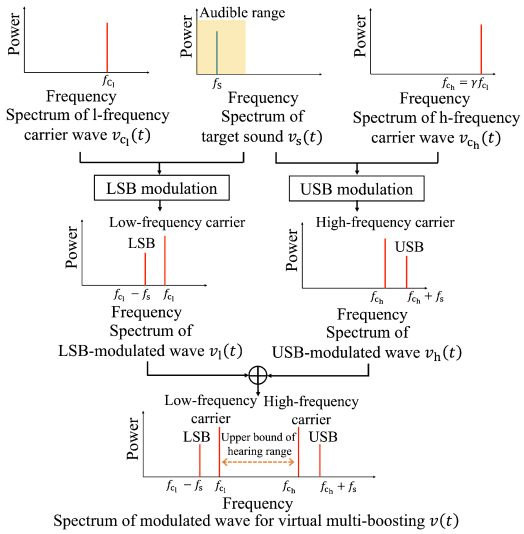


Fig. 2: Modulation flow in the frequency domain

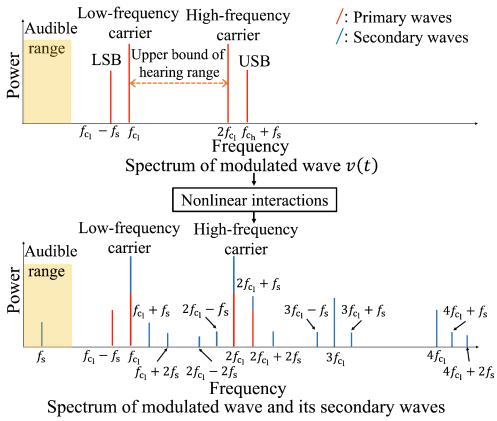


Fig. 3: Frequency relationship between the primary and secondary waves of the modulated wave

frequency components $2f_{c_l}$ and $2\gamma f_{c_l}$ are generated. Fig. 3 shows the frequency relationship between the carrier waves and the secondary waves resulting from their interactions when $\gamma = 2$. Here, the relationship between the weights for the LSB-modulated wave and the USB-modulated wave is $w_l = w_h$. When γ =2, the higher carrier frequency f_{c_h} becomes twice the lower carrier frequency f_{c_l}, and f_{c_l} matches the difference frequency between the two carriers. Therefore, the secondary waves boost the carrier waves $v_{c_l}(t)$ and $v_{c_h}(t)$ of the primary waves. Hence, the modulated wave of the proposed method can be expressed:

$$v(t) = A_{c_l}\cos(2\pi f_{c_l}t + \theta_{c_l}) + A_{b_l}\cos(2\pi f_{b_l}t + \theta_{b_l}) + A_{c_h}\cos(2\pi f_{c_h}t + \theta_{c_h}) + A_{b_h}\cos(2\pi f_{b_h}t + \theta_{b_h}), \quad (9)$$

where A_{c_l} is the maximum amplitude of the low-frequency carrier wave, A_{b_l} is the maximum amplitude of the LSB, A_{c_h} is the maximum amplitude of the high-frequency carrier wave, A_{b_h} is the maximum amplitude of the USB, f_{b_l} is the LSB frequency, f_{b_h} is the frequency of the USB, and θ_{c_l}, θ_{b_l}, θ_{c_h}, and θ_{b_h} are the phase terms of each frequency component, respectively.

When the modulated wave is emitted as a primary wave, higher-order components such as demodulated sound are generated because of nonlinear interactions in the air. Assuming that the generated frequency components are in phase, the upper limit of the power of the demodulated sound can be expressed as a theoretical value by the following equation:

$$P_s = \sum_{n=2}^{N}\left(\sum_{I_n^s}\beta_{I_n^s}\prod_{k=1}^{n}A_{I_k^s}\right), \quad (10)$$

where n is the order of the generated higher-order components, $I_n^s(I_1^s, \cdots, I_N^s \in \{c_l, c_h, b_l, b_h\})$ is the frequency of the primary wave responsible for generating n-th order components whose frequency match the target sound, N is the maximum order of higher-order components, c_l is the subscript corresponding to a low-frequency carrier component, c_h is the subscript corresponding to a high-frequency carrier components b_l is the subscript corresponding to the LSB component, b_h is the subscript corresponding to the USB component, k is the index of the primary wave responsible for generating n-th order components, $A_{I_k^s}$ is the amplitude of the primary wave responsible for generating n-th order wave components, and $\beta_{I_n^s}$ is the energy conversion efficiency from the primary wave to the n-th order components. Here, G_n^s represents the combinations of $I_1^s, \cdots, I_n^s$ for which n-th order components match the frequencies of the target sound. When n =2, G_2^s represents the combinations of I_1^s, I_2^s for which secondary components match the frequencies of the target sound ($|f_{I_1^s} - f_{I_2^s}| = f_s$), and the upper limit of the power of the target sound generated in the secondary waves can be calculated using Equation (10). By changing the value of n accordingly, the upper limit of the power of the higher-order components can be calculated in the same way.

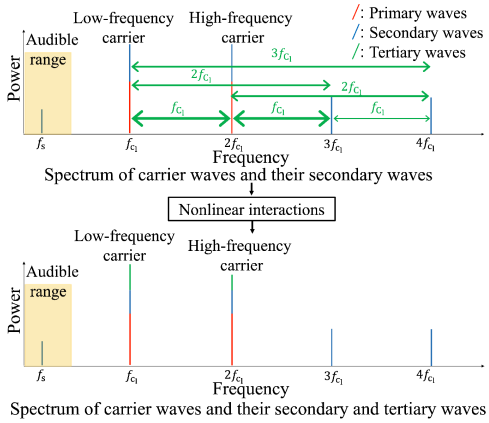


Fig. 4: Frequency relationship between the primary and secondary waves of the carrier waves, and their interaction

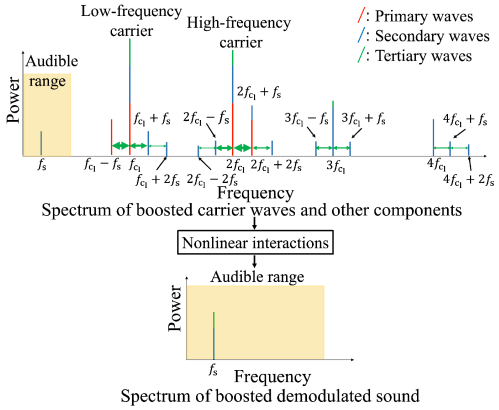


Fig. 5: Frequency relationship between the boosted carrier waves and other components, and their interaction

On the other hand, a feature of the proposed method is that the interactions of the two carrier waves also generate frequency components that match the frequencies of the carrier waves. Fig. 4 shows the frequency relationship between the carrier waves and their secondary and tertiary waves. Here, the two-way arrow line represents the frequency of the generated tertiary wave component: the thicker the line, the higher the power of the generated tertiary wave component. As shown in Fig. 4, the frequency components f_{c_l}, $2f_{c_l}$, $3f_{c_l}$ and $4f_{c_l}$ are generated as secondary waves. The frequencies f_{c_l} and $2f_{c_l}$ of the secondary waves match the frequencies of the carrier waves. The secondary waves boost the carrier waves $v_{c_l}(t)$ and $v_{c_h}(t)$. The interactions between the carrier waves and their secondary waves also generate tertiary waves. As a result, the boost from the secondary and tertiary waves increases the sound pressure of the carrier waves. The upper limits of the power of the low- and high-frequency carrier waves boosted by the higher-order components can be expressed by Equations (11) and (12), respectively:

$$P_{c_l} = \rho_l A_{c_l}^2 + \sum_{n=2}^{N} \left(\sum_{I_n^{c_l}} \beta_{I_n^{c_l}} \prod_{k=1}^{n} A_{I_k^{c_l}} \right), \quad (11)$$

$$P_{c_h} = \rho_l A_{c_h}^2 + \sum_{n=2}^{N} \left(\sum_{I_n^{c_h}} \beta_{I_n^{c_h}} \prod_{k=1}^{n} A_{I_k^{c_h}} \right), \quad (12)$$

where $I_n^{c_l}$ and $I_n^{c_h}$ are the frequencies of the primary waves responsible for generating n-th order components whose frequency matches the low-frequency and high-frequency carrier waves, ρ_l and ρ_h are power attenuation of low-frequency and high-frequency carrier waves, and $\beta_{I_n^{c_l}}$ and $\beta_{I_n^{c_l}}$ are the energy conversion efficiency from the primary wave to the n-th order components. The carrier components generated in the secondary and tertiary waves shown above can be calculated from Equations (11) and (12).

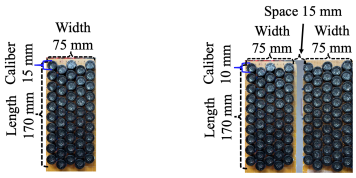
Fig. 5 shows the frequency relationship between the boosted carrier waves and other components and their interaction. Regarding the frequency components of the secondary waves, because the difference frequencies $f_{c_l} - (f_{c_l} - f_s)$ and $(f_{c_h} + f_s) - f_{c_h}$ match the demodulated sound frequency f_s, the demodulated sound is reproduced. As the sound pressure of the carrier waves increases, the demodulated sound pressure also increases. As shown in Fig. 5, there are multiple secondary wave components whose frequency interval is the frequency of the target sound. This means that there are multiple combinations of the numbers G_3^s and G_4^s in Equation (10). These components' interaction generates tertiary and fourth-order components that match the target frequency f_s. Thus, the higher-order components boost the demodulated sound. Therefore, we expect to boost the carrier waves and demodulated sounds with the higher-order components of the modulated wave, thereby increasing the demodulated sound pressure.

III. Evaluation Experiment

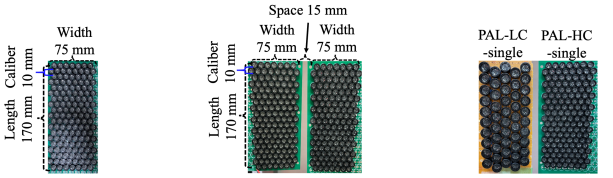
A. Experimental conditions

The modulated wave $v(t)$ of the proposed method consists of a LSB-modulated wave $v_l(t)$ and USB-modulated wave $v_h(t)$ with different carrier frequencies. In this experiment, the frequency ratio of the low-frequency carrier wave to the high-frequency carrier wave γ =2 and the carrier frequencies f_{c_l} = 25 kHz and f_{c_h} = 50 kHz were set. A PAL capable of emitting both 25 and 50 kHz should be used to realize the proposed method. However, as ultrasonic transducers capable of emitting both frequency bands could not be identified, PALs combining two ultrasound elements with different resonance frequencies are used. Fig. 6 shows PALs used in this experiment. An ultrasonic transducer (NICERA, T2516A1) with a resonance frequency of 25 kHz was used to emit a frequency band with a carrier frequency of f_{c_l} = 25 kHz. An ultrasonic transducer (SPL, UT1007-Z325R) with a resonance frequency of 40 kHz was used to emit a frequency band with a carrier frequency of f_{c_h} = 50 kHz. The PAL-LC-single and PAL-HC-single were made so that the areas of the emitting surfaces were equal. PAL-LC-double and PAL-HC-double are PALs with two PAL-LC-singles and two PAL-HC-singles in a row. PAL-LCHC is a PAL with a PAL-LC-single on the left and a PAL-HC-single on the right. The space between the two PALs is 15 mm, which is the diameter of the ultrasonic transducer (NICERA, T2516A1).

The following six conditions were compared.



(a) PAL-LC-single (b) PAL-LC-double



(c) PAL-HC-single (d) PAL-HC-double (e) PAL-LCHC

Fig. 6: PALs used in this experiment

- LSB single: Only the LSB-modulated wave $w_{\mathrm{l}}v_{\mathrm{l}}(t)$ is emitted from the PAL-LC-single.
- LSB double: Only the LSB-modulated wave $w_{\mathrm{l}}v_{\mathrm{l}}(t)$ is emitted from the PAL-LC-double.
- USB single: Only the USB-modulated wave $w_{\mathrm{h}}v_{\mathrm{h}}(t)$ is emitted from the PAL-HC-single.
- USB double: Only the USB-modulated wave $w_{\mathrm{h}}v_{\mathrm{h}}(t)$ is emitted from the PAL-HC-double.
- LSB+USB (w/o multi-boost): By using PAL-LCHC, the LSB-modulated wave $w_{\mathrm{l}}v_{\mathrm{l}}(t)$ is emitted from PAL-LC-single and the USB-modulated wave $w_{\mathrm{h}}v_{\mathrm{h}}(t)$ from PAL-HC-single. However, to remove the boosting effect of the proposed method, it is approximated by a superposition of the LSB single and USB single recorded sounds.
- Prop. (LSB+USB w/ multi-boost): By using PAL-LCHC, the LSB-modulated wave $w_{\mathrm{l}}v_{\mathrm{l}}(t)$ is emitted from PAL-LC-single and the USB-modulated wave $w_{\mathrm{h}}v_{\mathrm{h}}(t)$ from PAL-HC-single. The boosting effect of the proposed method occurs because the two modulated waves are emitted simultaneously.

In each condition, the impulse responses were calculated by the time stretched pulse (TSP) method [24] using the TSP signal as the target sound. The modulation factor $m_{\mathrm{ssb}} = 1.0$ was set so that the demodulated sound pressure is maximized under conditions where no over-modulation occurs.

Figs. 7 and 8 show the arrangement of the equipment and the photograph of the equipment. As shown in Fig. 7, in the LSB double, USB double, and Prop., the two PALs were set up to form a focal point at 0.5 m by shifting them by 10 degrees. In the LSB single, the PAL-LC-single is arranged in the PAL position on the left side when viewed from the microphone. In the USB single, the PAL-HC-single is arranged in the PAL position on the right side when viewed from the microphone. The w_{l} and w_{h} were adjusted so that the demodulated sound pressure of the LSB single and USB single was 60 dB at 0.5 m from the PAL.

Microphones were placed at $d = 0.5$, 1.0, and 2.0 m from the front of the PAL and at a height of 1.2 m from the floor. Tables I and II list the experimental equipment and conditions in the evaluation experiments. The impulse response was measured from the recorded sound and the sound pressure level of the demodulated sound was calculated. The sound pressure level of

TABLE I: Experimental conditions

Environment	Soundproof room (T_{60} =150 ms)
Ambient noise level	L_A = 21.9 dB
Sampling frequency	192 kHz
Quantization	32 bit
Target sound	TSP signal (0∼4.0 kHz)

TABLE II: Experimental equipment

Ultrasonic transducer	SPL, UT1007-Z325R NICERA, T2516A1
Power amplifier	VICTOR, PS-A2002
Microphone	SENNHEISER, MKH 8020
D/A, A/D converter	RME, Fireface UFX

the demodulated sound L [dB] was calculated by the following equation:

$$L = L_{\mathrm{pis}} - 10\log_{10}\left(\frac{\sigma_{\mathrm{rec}}^2}{\sigma_{\mathrm{pis}}^2}\right), \tag{13}$$

where L_{pis} is the sound pressure level of the acoustic calibrator measured by the sound level meter and σ_{rec} and σ_{pis} are the standard deviation of the recorded sound of the acoustic calibrator and the audible sound, respectively.

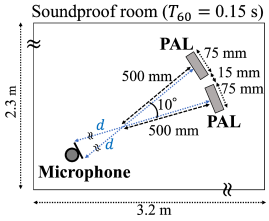


Fig. 7: Arrangement of the equipment

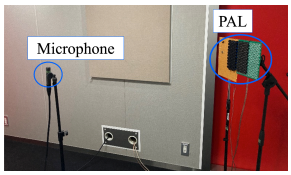


Fig. 8: Photograph of equipment

B. Results and discussion of experiment

In this experiment, ten recordings were made in each condition, and the average sound pressure was compared. Fig.9 shows the results for demodulated sound pressure. The horizontal axis represents the distance from the PAL, and the vertical axis represents the mean sound pressure level of the demodulated sound and the error bars represent the standard deviation. As shown in Fig. 9, the sound pressure level of Prop. was 0.81 dB higher at 0.5 m, 0.99 dB higher at 1.0 m, and 1.56 dB higher at 2.0 m than that of LSB double. In addition, the sound pressure level of Prop. was 0.01 dB higher at 0.5 m, 0.49 dB higher at 1.0 m, and 0.62 dB higher at 2.0 m than that of USB double. Therefore, the demodulated sound pressure of the proposed method was higher than those of the conventional methods at 1.0 and 2.0 m away from the focal point. As shown in Fig. 9, the sound pressure level of Prop. was 0.50 dB higher at 0.5 m, 0.46 dB higher at 1.0 m, and 0.44 dB higher at 2.0 m than that of LSB+USB(w/o multi-boost). Therefore, at all recording points, the demodulated sound pressure of the proposed method was higher than in the condition without boosting. Figs.10a and 10b show the spectrum of recorded sound in LSB+USB and Prop. at 1.0 m. As shown in Figs.

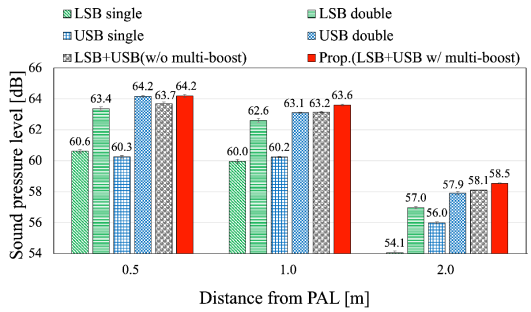
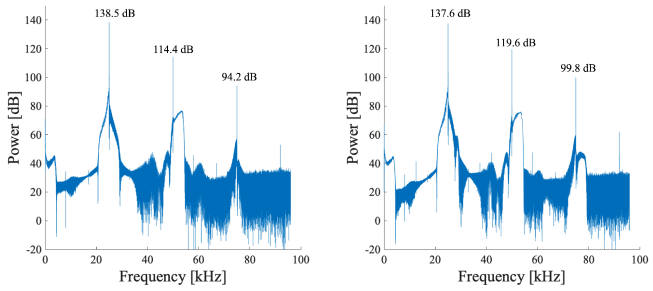


Fig. 9: Experimental results for demodulated sound pressure



(a) LSB+USB (w/o multi-boost) (b) Prop. (LSB+USB w/ multi-boost)

Fig. 10: Spectrum of recorded sound at 1.0 m

10(a) and 10(b), the sound pressure level of the low-frequency carrier wave ($f_{c_l} = 25$ kHz) in Prop. was comparable to that in LSB+USB. However, the sound pressure level of the high-frequency carrier wave ($f_{c_h} = 50$ kHz) in Prop. was 5.26 dB higher than that in LSB+USB. In addition, the sound pressure level of the frequency component $3f_{c_l}$ in Prop. was 5.61 dB higher than that in LSB+USB. Thus, the carrier waves and demodulated sound are boosted and the demodulated sound pressure is increased.

IV. Conclusion

We have proposed a virtual multi-boosted amplitude modulation toward high-pressure audible sound with PALs. In this method, higher-order components other than the demodulated sound boost the demodulated sound and carrier waves. Therefore, the distortion components of the modulated wave are effectively used to improve demodulation efficiency and to achieve high-pressure audible sound with PALs. To evaluate the proposed method's effectiveness, we conducted an objective evaluation experiment of the demodulated sound pressure. As a result, we revealed that the proposed method increases demodulated sound pressure by boosting the carrier wave and demodulated sound with higher-order components. However, in practical applications, the demodulated sound pressure needs to be increased even further. To address this, we plan to implement phase control of the modulated wave to maximize the boosting effect contributed by the higher-order components. This approach is expected to further increase the demodulated sound pressure of the proposed method. In addition, we plan to conduct experiments to evaluate the quality of the demodulated sound.

Acknowledgments

This work was partly supported by Ritsumeikan University R-GIRO and RARA, and JSPS KAKENHI Grant Numbers JP21H03488, JP21H04427, JP23H03425, JP23K21691, JP23K28115 and JP24K20803.

References

[1] R. H. Small, "Closed-box loudspeaker systems part 1: Analysis," Journal of the Audio Engineering Society, vol. 20, no. 10, pp. 799–808, 1972.
[2] R. H. Small, "Direct-radiator loudspeaker system analysis," IEEE Transactions on Audio and Electroacoustics, vol. 19, no. 4, pp. 269–281, 1971.
[3] C. Shi and W. Gan, "Development of parametric loudspeaker," IEEE Potentials, vol. 29, no. 6, pp. 20–24, 2010.
[4] K. Aoki, T. Kamakura, and Y. Kumamoto, "Parametric loudspeaker—characteristics of acoustic field and suitable modulation of carrier ultrasound," Electronics and Communications in Japan (Part III: Fundamental Electronic Science), vol. 74, no. 9, pp. 76–82, 1991.
[5] W. S. Gan, J. Yang, and T. Kamakura, "A review of parametric acoustic array in air," Applied Acoustics, vol. 73, no. 12, pp. 1211–1219, 2012.
[6] T. Fujii, M. Nakayama, and T. Nishiura, "Demodulation distance control based on analytic model between film gas-lens depth and demodulation distance for parametric array loudspeaker," Proceedings of the 23rd International Congress on Acoustics (ICA 2019), pp. 1061–1068, 2019.
[7] T. Matsui, D. Ikefuji, M. Nakayama, and T. Nishiura, "Multiple audio spots design based on separating emission of carrier and sideband waves," INTER-NOISE 2014, vol. 249, no. 5, pp. 2924–2928, 2014.
[8] Y. Geng, S. Sayama, M. Nakayama, and T. Nishiura, "Movable virtual sound source construction based on wave field synthesis using a linear parametric loudspeaker array," APSIPA Transactions on Signal and Information Processing, vol. 12, no. 3, e14, 2023.
[9] P. J. Westervelt, "Parametric acoustic array," The Journal of the Acoustical Society of America, vol. 35, no. 4, pp. 535–537, 1963.
[10] R. Konabe, T. Matsui, D. Ikefuji, M. Nakayama, and T. Nishiura, "A design of reflective audio spot with parabolic reflector for sound pressure improvement on separating emission of carrier and sideband waves," INTER-NOISE 2014, vol. 249, pp. 2890–2896, 2014.
[11] Y. Ogami, T. Fukumori, M. Nakayama, and T. Nishiura, "Enhanced audio spot based on multiple carrier waves with extra parametric loudspeaker," INTER-NOISE 2017, vol. 255, pp. 3233–3240, 2017.
[12] M. Yoneyama, J. Fujimoto, Y. Kawamo, and S. Sasabe, "The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design," The Journal of the Acoustical Society of America, vol. 73, no. 5, pp. 1532–1536, 1983.
[13] J. J. Croft, "Theory, history, and the advancement of parametric loudspeaker," White paper of American Technology Corporation, 2001.
[14] S. Sakai and T. Kamakura, "Dynamic single sideband modulation for realizing parametric loudspeaker," AIP Conference Proceedings, pp. 613–616, 2008.
[15] Y. Hatano, and C. Shi, and Y. Kajikawa, "Compensation for nonlinear distortion of the frequency modulation-based parametric array loudspeaker," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 8, pp. 1709–1717, 2017.
[16] M. Nakayama and T. Nishiura, "Synchronized amplitude-and-frequency modulation for a parametric loudspeaker," APSIPA-ASC 2017, pp. 130–135, 2017.
[17] K. Asahi, M. Nakayama, and T. Nishiura, "Audible sound enhancement based on carrier with periodic square-wave for parametric loudspeaker," 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), pp. 12–15, 2020.
[18] Y. Nakano, T. Fukumori, M. Nakayama, and T. Nishiura, "A study on audible low-frequency sound emphasis based on multiplexed double sideband modulation in parametric loudspeaker," INTER-NOISE 2018, vol. 258, pp. 2054–2065, 2018.
[19] Y. Geng, Y. Nakano, M. Nakayama, and T. Nishiura, "Development of multi-way parametric array loudspeaker using multiplexed double sideband modulation," Proceedings of the 23rd International Congress on Acoustics (ICA 2019), pp. 1069–1076, 2019.
[20] T. D. Kite, J. T. Post, and M. F. Hamilton, "Parametric array in air: Distortion reduction by preprocessing," The Journal of the Acoustical Society of America, vol. 103, no. 5, p. 2871 1998.
[21] Y. Geng, M. Nakayama, and T. Nishiura, "Virtual bass enhancement based on harmonics control using missing fundamental in parametric array loudspeaker," INTER-NOISE 2019, vol. 259, no. 7, pp. 2660–2671, 2019.
[22] Y. Geng, M. Nakayama, and T. Nishiura, "Demodulated sound quality improvement for harmonic sounds in over-boosted parametric array loudspeaker," Applied Acoustics, vol. 186, p. 108460, 2022.
[23] Y. Ikezaki, Y. Geng, M. Nakayama, and T. Nishiura, "Demodulated sound enhancement based on virtual multi-boosted amplitude modulation with parametric array loudspeaker," INTER-NOISE 2023, vol. 268, no. 6, pp. 2477–2487, 2023.
[24] Y. Suzuki, F. Asano, H. Y. Kim, and T. Sone, "An optimum computer - generated pulse signal suitable for the measurement of very long impulse responses," The Journal of the Acoustical Society of America, vol. 97, no. 2, pp. 1119–1123, 1998.

Deep-Learning-Based Speech Enhancement with Rough-Focused Optical Laser Microphone by Reconstructing Complex Spectrum

Yuki Nakano*, Yuting Geng*, Kenta Iwai*, Takanobu Nishiura*
* Ritsumeikan University, Osaka, Japan
E-mail: {is0570he@ed, geng@fc, iwai18sp@fc, nishiura@is}.ritsumei.ac.jp

***Abstract*—Rough-focused recording with an optical laser microphone allows for recording that is wide ranging and robust against changes in the position of a vibrating object. However, the recorded speech suffers from noise due to laser diffusion and missing signal components. To solve this problem, we propose a speech enhancement method for rough-focused recordings that reconstructs a complex spectrum. The conventional speech enhancement method for rough-focused optical laser microphones reconstructs only an amplitude spectrogram without considering phase components, resulting in lower quality speech enhancement. In contrast, the proposed method simultaneously reconstructs both amplitude and phase components of rough-focused recordings by reconstructing a complex spectrum. We compared the method with the conventional method. The results show that the proposed method performed equivalently to or better than the conventional method in objective evaluations.**

I. Introduction

The optical laser microphone [1] has attracted attention for use in acoustic systems capable of recording target speech from a distance [2], [3]. It measures speech-induced vibration by irradiating a laser beam onto the surface of a vibrating object and captures the speech. Generally, acoustic measurements with this microphone have been conducted with a focused laser beam to obtain sufficient reflected light. In contrast, a recording method that operates in an out-of-focus state, called rough-focused recording [4], has been studied. In rough-focused recording, the optical laser microphone irradiates the laser beam over a wide area, allowing for recording that is both wide-ranging and robust against changes in the position of a vibrating object. However, the wide irradiation area of the laser beam results in the reflection of the laser becoming diffused when measuring acoustic signals, causing noise and missing signal components in the recorded signal. Therefore, noise suppression and reconstruction of missing signal components are the main tasks of speech enhancement for the rough-focused optical laser microphone.

To address this issue, deep-learning-based speech enhancement for rough-focused recording has been proposed [4]. In [4], a single model enables speech enhancement for recorded speech in different rough-focus conditions. Then, the model reconstructs an amplitude spectrum of clean speech from one of recorded speech. Therefore, the speech enhancement performance in [4] is inadequate in that only the amplitude spectrum of the speech signal is reconstructed while the phase spectrum is not taken into account.

In this paper, we propose a deep-learning-based speech enhancement for the rough-focused optical laser microphone that reconstructs a complex spectrum and discuss the feasibility of speech enhancement for rough-focused recording with simultaneously reconstructed both amplitude and phase components. Since the complex spectrum contains both amplitude and phase components, the reconstruction of the spectrum allows for phase-aware speech enhancement. The complex-spectrum reconstruction model is designed on the basis of a previous study [5]. To improve the performance in reconstructing signal components in the high-frequency band, we also investigate the design of the loss function.

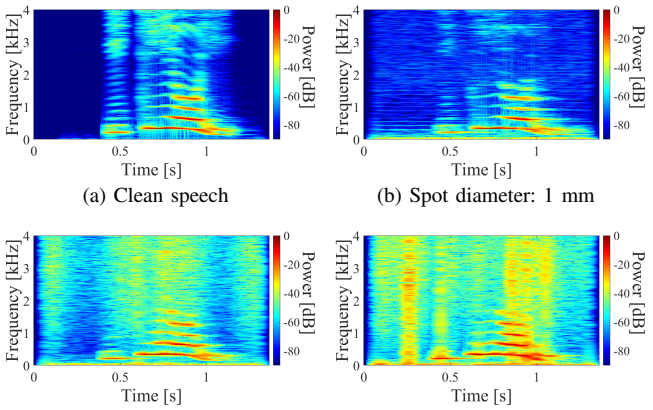
The remainder of the paper is organized as follows. In Section II, we describe speech-quality degradation with the rough-focused optical laser microphone and related work. In Section III, we present the proposed method. In Section IV we present the experiments and the results. Finally, we conclude the paper in Section V.

II. Speech-Quality Degradation in Rough-Focused Optical Laser Microphone

In a rough-focused optical laser microphone, recording robust against movements of vibrating objects comes at the cost of degradation in speech quality, as discussed below. In this paper, we use the spot diameter of the laser beam on a vibrating object as a metric to represent the state of rough focus.

Spectrograms of clean speech and recorded speech with the rough-focused optical laser microphone are shown in Fig. 1. Fig. 1 (b) shows that signal components in the high-frequency band (2 to 4 kHz) are missing. Also, Fig. 1 (d) shows that the recorded speech contains impact and stationary noises. This is attributed to the decrease in the intensity of reflected light due to rough-focus recording, resulting in measurement errors at the detector and subsequent noise contamination [6]. Then, Fig. 1 shows that the larger the spot diameter, the greater the degradation in sound quality.

Therefore, to obtain high-quality speech with the rough-



(a) Clean speech (b) Spot diameter: 1 mm

(c) Spot diameter: 60 mm (d) Spot diameter: 180 mm

Fig. 1: Spectrograms of recorded speech with rough-focused optical laser microphone.

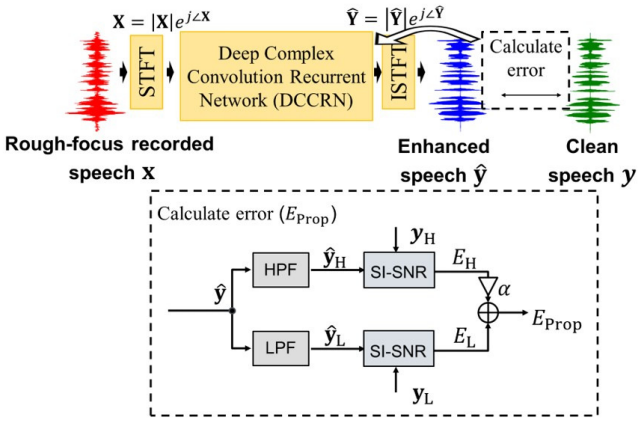


Fig. 2: Overview of training model in speech enhancement for recorded speech with rough-focused optical laser microphone.

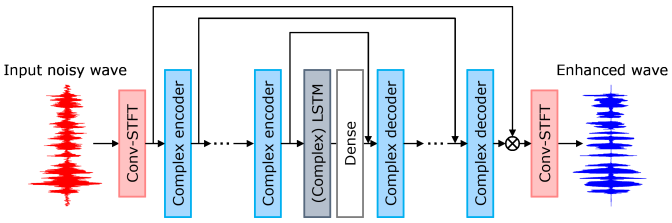


Fig. 3: DCCRN network.

focused optical laser microphone, it is necessary to suppress the noise and enhance the high-frequency components of the speech through speech enhancement.

To address the above problems, a previous study [4] proposed deep neural network (DNN)-based speech enhancement by reconstructing a log-power spectrum for rough-focus recorded speech. In the method [4], rough-focus recorded speech $\mathbf{x} \in \mathbb{R}^T$ is subjected to a short-time Fourier transform (STFT), and the complex spectrum $\mathbf{X} \in \mathbb{C}^{K \times L}$ is calculated as

$$\mathbf{X} = \mathrm{STFT}[\mathbf{x}], \tag{1}$$

where T is the number of samples, L is the number of frames, K is the number of frequency bins, and STFT$[\cdot]$ represents the STFT operator, respectively. Then, the amplitude spectrograms and phase spectrograms of $\mathbf{X}$ are denoted as $|\mathbf{X}|$ and $\angle\mathbf{X}$. Next, the log-power spectrogram $10\log_{10}|\mathbf{X}|^2$ is input to DNN to obtain $10\log_{10}|\hat{\mathbf{Y}}|$, where $10\log_{10}|\mathbf{X}|^2$ is the log-power spectrogram of $\mathbf{X}$, and $10\log_{10}|\hat{\mathbf{Y}}|^2$ is the log-power spectrogram of enhanced speech. Finally, the amplitude spectrogram $|\hat{\mathbf{Y}}|$ is calculated from $10\log_{10}|\hat{\mathbf{Y}}|^2$, and enhanced speech $\hat{\mathbf{y}} \in \mathbb{R}^T$ is obtained by performing an inverse STFT (ISTFT) using $|\hat{\mathbf{Y}}|$ and $\angle\mathbf{X}$.

$$\hat{\mathbf{y}} = \mathrm{ISTFT}[|\hat{\mathbf{Y}}|e^{j\angle\mathbf{X}}], \tag{2}$$

where ISTFT$[\cdot]$ represents the ISTFT operator. In (2), the degraded phase spectrum of the roughly recorded speech is used as is to calculate $\hat{\mathbf{y}}$. Therefore, the speech enhancement performance is insufficient as phase components are not considered.

III. Proposed Speech Enhancement for Rough-Focused Optical Laser Microphone

A. Overview of proposed speech enhancement

We focus on the deep complex convolution recurrent network (DCCRN) [5] for its network architecture, which is designed to handle the complex spectrum. Since the complex spectrum contains amplitude components and phase components, DCCRN enables phase-aware speech enhancement. An overview of the proposed method is shown in Fig. 2. First, $\mathbf{x}$ is first subjected to STFT, and $\mathbf{X}$ is calculated. Next, $\mathbf{X}$ is input to DCCRN, and the complex spectrum of enhanced speech $\hat{\mathbf{Y}}$ is obtained. In the stage of model training, the error E_{Prop} is calculated using the proposed loss function from the enhanced speech $\hat{\mathbf{y}}$ and the clean speech $\mathbf{y}$. A detailed definition of the loss function is given in Section III-C. In the proposed method, the speech enhancement model with DCCRN is trained by minimizing the error calculated from E_{Prop} for speech recorded with a rough-focused optical laser microphone.

B. Network architectures

The network architectures of DCCRN and the components of the complex encoder are shown in Figs. 3 and 4. DCCRN, originally described in [5], is an encoder-decoder architecture with two long short term memory (LSTM) layers. As shown in Fig. 4, the complex encoder consists of complex convolutional layers, complex batch normalization, and an activation function (PReLU). In the complex convolutional layers, complex-valued filters are convolved with the input on the basis of the rules of complex multiplication. The complex decoder is composed of the complex convolutional layers, complex batch normalization, and an activation function (PReLU) as in the complex encoder. Then, in the complex convolutional layers of the complex decoder, transposed convolution is performed.

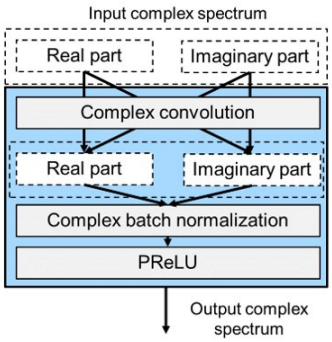


Fig. 4: Complex encoder.

C. Loss function

Generally, the loss function used in training of the DCCRN is the scale-invariant source-to-noise ratio(SI-SNR). SI-SNR is defined as

$$\begin{cases} \mathbf{y}_{\text{target}} & := \dfrac{\langle \hat{\mathbf{y}}, \mathbf{y} \rangle \cdot \mathbf{y}}{\|\mathbf{y}\|^2} \\ \mathbf{e}_{\text{noise}} & := \hat{\mathbf{y}} - \mathbf{y}_{\text{target}} \\ \text{SI-SNR} & := 10 \log_{10} \dfrac{\|\mathbf{y}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{noise}}\|^2}, \end{cases} \tag{3}$$

where $\mathbf{y}$ is clean speech, $\langle \cdot, \cdot \rangle$ denotes the inner product of vectors, and $\| \cdot \|$ denotes the Euclidean norm. Additionally, the loss $E_{\text{SI-SNR}}$ is expressed as

$$E_{\text{SI-SNR}} = -\,\text{SI-SNR}. \tag{4}$$

In this paper, we propose a method using loss function $E_{\text{SI-SNR}}$ and a method using the proposed loss function E_{Prop}. The method of calculating the proposed loss function E_{Prop} is explained as follows. First, $\hat{\mathbf{y}}$ is obtained from $\mathbf{x}$ through a series of processes based on the DCCRN. Next, high-pass and low-pass filters are applied to both $\hat{\mathbf{y}}$ and $\mathbf{y}$ in order to calculate $\hat{\mathbf{y}}_{\text{H}}$, $\hat{\mathbf{y}}_{\text{L}}$ and $\mathbf{y}_{\text{H}}$, $\mathbf{y}_{\text{L}}$. In this context, $\hat{\mathbf{y}}_{\text{H}}$, $\hat{\mathbf{y}}_{\text{L}}$ and $\mathbf{y}_{\text{H}}$, $\mathbf{y}_{\text{L}}$ represent the high-frequency and low-frequency components of $\hat{\mathbf{y}}$ and $\mathbf{y}$. Then, from $\hat{\mathbf{y}}_{\text{H}}$, $\mathbf{y}_{\text{H}}$ and $\hat{\mathbf{y}}_{\text{L}}$, $\mathbf{y}_{\text{L}}$, the loss in the high and low frequency components E_{H}, E_{L} are calculated. Finally, the total loss is calculated as

$$E_{\text{Prop}} = \alpha E_{\text{H}} + E_{\text{L}}, \tag{5}$$

where α is the weight of E_{H}. By adjusting α, it is possible to correct the influence of the low-frequency and high-frequency components on $E_{\text{SI-SNR}}$. Given the missing components in the high-frequency band in speech recorded with a rough-focused optical laser microphone, we believe that using E_{Prop} as the loss function can help the model better focus on reconstructing high-frequency components.

IV. Experiments and Results

We conducted experiments to evaluate the performance of the proposed method. In the experiments, we refer to the training method with the original loss function as Prop. A, and we refer to that with the proposed loss function as Prop. B.

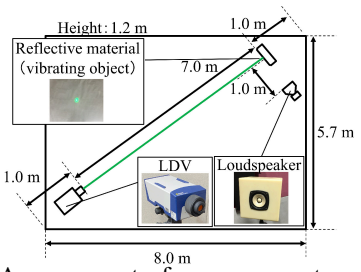


Fig. 5: Arrangement of measurement equipment.

TABLE I: Recording conditions.

Reverberation time T_{60}	300 ms
Ambient noise level L_{A}	29.7 dB
Sampling frequency	8 kHz
Quantization	16 bits
Temperature	23.8°C
Humidity	19.5%

TABLE II: Recording equipment.

Laser doppler vibrometer (LDV)	Polytec, VFX-F-110
A/D, D/A converter	RME, Fireface UFX
Loudspeaker	Fostex, FE83En
Loudspeaker amplifier	BOSE, 1705II
Reflective material	ONOSOKKI, LV-0012

A. Experimental setups

We recorded a speech dataset using a rough-focused optical laser microphone. The dataset was divided into two parts to train and evaluate the speech enhancement model. Specifically, 503 Advanced Telecommunications Research (ATR) phoneme-balanced sentences [7] (average duration: approximately 11.8 seconds per file) were used as the training data, and 216 ATR phoneme-balanced words [8] (average duration: 2.1 seconds per file) were used as the evaluation data. The arrangement of the equipment used for recording the training and evaluation data, the recording conditions, and the recording equipment are shown in Fig. 5 and Tables I and II, respectively. In this recording, 15,368 utterances from datasets [7], which include 300 speakers (150 female and 150 male speakers), were recorded with an optical laser microphone in various focus settings (spot diameter: 1, 60, 120, and 180 mm) as the training data. Also, 3,024 utterances from datasets [8], which include 14 speakers (7 female and 7 male speakers) were recorded as the evaluation data. The experimental settings for training of DNN are shown in Table III.

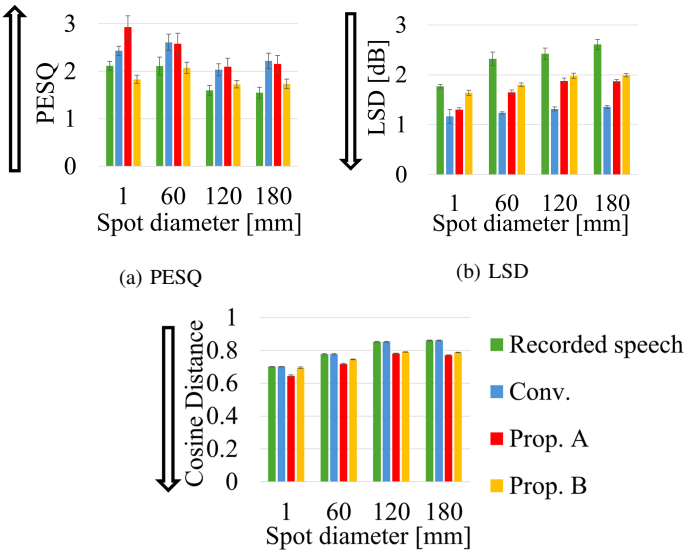
We used the perceptual evaluation of speech quality (PESQ) [9], log-spectral distance (LSD) [10], and cosine distance as the evaluation metrics. PESQ is used for speech in the range of 300–3400 Hz and is highly correlated with subjective evaluation of speech quality. The PESQ score ranges from -0.5 to 4.5, where higher scores indicate speech that is more intelligible to human auditory perception. LSD is a metric that shows magnitude distortion, where lower scores indicate higher quality of amplitude components in the evaluated speech. Cosine distance is a measure of the phase errors, defined as

$$\text{cosine distance} = 1 - \cos(\theta_{\text{Clean}} - \theta_{\text{Enh}}), \tag{6}$$

where θ_{Clean} is the phase spectrum of clean speech, and θ_{Enh} is the phase spectrum of enhanced and recorded speech. The

TABLE III: Experimental settings.

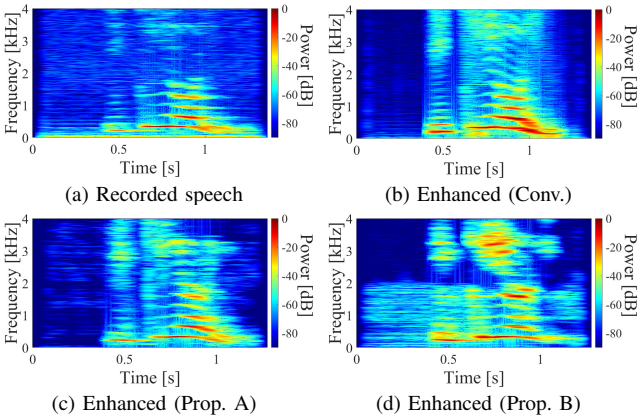
Sampling frequency	8 kHz
Quantization	16 bit
Window function	Hamming window
STFT length	512
Hidden layers	5 layers
LSTM layers	2 layers
Epochs	600
Batch size	32
Learning rate	0.0001
Optimizer	Adam
Loss function	$E_{\text{SI-SNR}}$ / E_{Prop}
Cutoff frequency in E_{Prop}	2 kHz
α in E_{Prop}	5



(a) PESQ (b) LSD

(c) Cosine distance

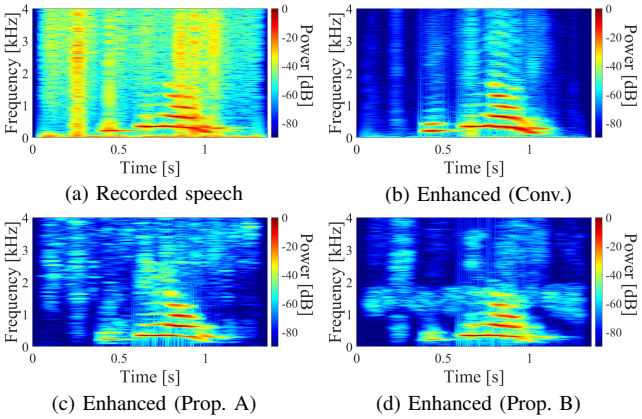
Fig. 6: Evaluation results.



(a) Recorded speech (b) Enhanced (Conv.)

(c) Enhanced (Prop. A) (d) Enhanced (Prop. B)

Fig. 7: Spectrograms of recorded and enhanced speech (spot diameter: 1 mm)



(a) Recorded speech (b) Enhanced (Conv.)

(c) Enhanced (Prop. A) (d) Enhanced (Prop. B)

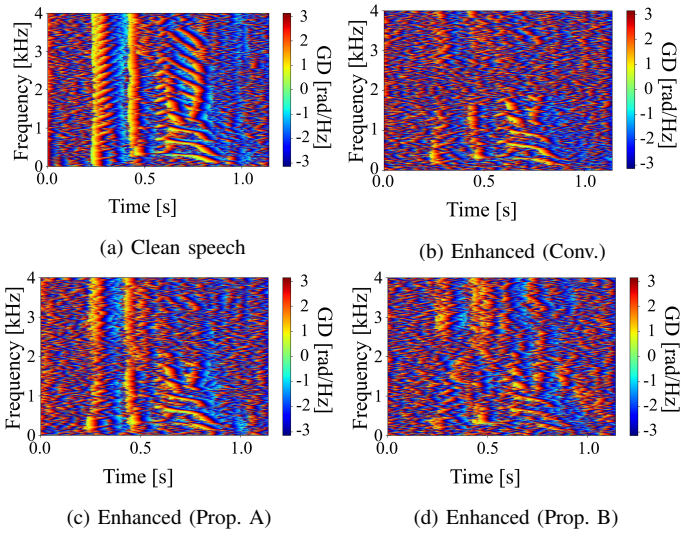
Fig. 8: Spectrograms of recorded and enhanced speech (spot diameter: 180 mm)

cosine distance score ranges from 0.0 to 2.0, where lower scores indicate higher quality of phase components in the evaluated speech.

B. Experimental results and discussion

The evaluation results in terms of PESQ, LSD, and cosine distance are shown in Fig. 6. From Fig. 6 (c), the cosine distance of recorded speech and the conventional method (Conv.) had the same score because enhanced speech with method [4] is calculated using STFT with the estimated amplitude and the phase of the recorded speech. Fig. 6 (a) shows that the PESQ score of the enhanced speech with Prop. A surpassed that of the enhanced speech with Conv. for the spot diameter of 1 mm. On the other hand, the PESQ score of the enhanced speech with Prop. B was lower than both the PESQ score of the enhanced speech with Prop. A and that with the conventional method. Therefore, the experimental results for PESQ show that Prop. A had the highest performance at a spot diameter of 1 mm, and Prop. A and Conv. performed equivalently at spot diameters other than 1 mm. Fig. 6 (b) shows that the LSD score of the enhanced speech with Prop. A was comparable to that with Conv. only for the spot diameter of 1 mm. The LSD score of the enhanced speech with Prop. B was higher than that with Conv. for all spot diameters. Therefore, Conv. and Prop. A had comparable amplitude reconstruction performance at the spot diameter of 1 mm. However, Conv. had the highest amplitude reconstruction performance at spot diameters other than 1 mm. Fig. 6 (c) shows that the cosine distance of the enhanced speech with Prop. A and that with Prop. B were lower than that with Conv. for all spot diameters. This indicates that Prop. A and Prop. B achieved phase reconstruction simultaneously with amplitude reconstruction at all spot diameters. Despite reconstructing both amplitude and phase simultaneously at all spot diameters, the PESQ score only improved with Prop. A at a spot diameter of 1 mm. This may be caused by the inferior amplitude reconstruction performance compared to Conv. at spot diameters other than 1 mm, as indicated in Fig. 6 (b) .

The spectrograms of the recorded speech and enhanced speech are shown in Figs. 7 and 8. Fig. 7 (b) and Fig. 8 (b)



(a) Clean speech (b) Enhanced (Conv.)

(c) Enhanced (Prop. A) (d) Enhanced (Prop. B)

Fig. 9: Group delay of clean speech and enhanced speech (spot diameter: 1 mm)

show that method [4] suppressed impact and stationary noises in the recorded speech and reconstructed the high-frequency speech components. This could be the reason for the good scores of method [4] in PESQ and LSD at all spot diameters. Fig. 7 (c) and Fig. 8 (c) show that Prop. A suppressed both impact and stationary noises and failed to reconstruct the speech components in the frequency band above 2 kHz. Given the low requirement for the reconstruction of the high-frequency components attributable to the slight missing components in speech recorded at a 1-mm-spot diameter, Prop. A was found to score well in PESQ and LSD at this diameter. Fig. 7 (d) and Fig. 8 (d) show that Prop. B suppressed noises in the frequency band above 2 kHz and reconstructed the high-frequency speech components. Given the higher performance in enhancing high-frequency components compared with Prop A, Prop. B shows the feasibility of improving the performance of reconstructing high-frequency components through E_{Prop}.

The group delay spectrograms [11] of clean and enhanced speech are shown in Fig. 9. It is known that the group delay and amplitude have strongly constrained relationships [12], and it can be confirmed that the group delay spectrogram has a similar structure to the log-power spectrogram in clean speech. Figs. 9 (a) and (b) show the distortion of phase components in the frequency band above 2 kHz. Fig. 9 (c) shows the reconstruction of the phase components in the frequency band above 2 kHz with Prop. A. It is clear that Prop. A achieves phase-aware speech enhancement. As shown above, the amplitude reconstruction of Conv. and Prop. A is accurate, while the phase reconstruction is done only by Prop. A. This could be the reason for the significant improvement in PESQ score of Prop. A.

In summary, the proposed method improves the quality of speech recorded with a rough-focused optical laser microphone. However, except for the case with a spot diameter of 1 mm, the performance of Prop. A is equivalent to or slightly lower than that of the conventional method. The reason for this is that the amplitude components could not be sufficiently reconstructed as a result of considering the phase components. Additionally, the enhanced speech with Prop. B, which modified the loss function to reconstruct high-frequency components, demonstrates the feasibility of improving the performance of reconstructing high-frequency components through E_{Prop}.

V. CONCLUSION

We proposed a DCCRN-based speech enhancement method for the rough-focused optical laser microphone. We evaluated our method by comparing PESQ, LSD, and cosine distance. Objective experiments showed that the proposed method, which considers phase components, is superior to the conventional method. In the future, we aim to achieve further high-quality speech for the rough-focused optical laser microphone by modifying the loss function and using features other than the complex spectrum.

VI. ACKNOWLEDGEMENTS

This work was partly supported by the Ritsumeikan Advanced Research Academy (RARA), Ritsumeikan Global Innovation Research Organization (R-GIRO), and by JSPS KAKENHI Grant Numbers JP21H03488, JP23H03425, JP23K21691, JP23K28115, and JP24K20803.

REFERENCES

[1] J. A. Bucaro and N. Lagakos, “Lightweight fiber optic microphones and accelerometers,” *Review of Scientific Instruments*, vol. 72, no. 6, pp. 2816–2821, 2001.

[2] T. Fukumori, C. Cai, Y. Zhang, L. E. Hafi, Y. Hagiwara, and T. Nishiura, “Optical laser microphone for human-robot interaction: Speech recognition in extremely noisy service environments,” *Advanced Robotics*, vol. 36, no. 5-6, pp. 304–317, 2022.

[3] Y.-M. Lin, J.-Y. Han, C.-H. Lin, and Y.-H. Lai, “Optical microphone-based speech reconstruction system with deep learning for individuals with hearing loss,” *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 12, pp. 3330–3341, 2023.

[4] K. Miyazato, W. Haonan, K. Iwai, and T. Nishiura, “Study on speech quality improvement of rough pint recorded by optical laser microphone (in Japanese),” *2022 Spring Meeting of Acoustical Society of Japan*, pp. 403–406, 2022.

[5] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, “DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement,” in *Proceedings Interspeech 2020*, 2020, pp. 2472–2476.

[6] M. Johansmann, G. Siegmund, and M. Pineda, “Targeting the limits of laser Doppler vibrometry,” *Proceedings IDEMA*, pp. 1–12, 2005.

[7] Y. Sagisaka, "A large-scale Japanese speech database," *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, pp. 1089–1092, 1990.

[8] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[9] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *International Telecommunication Union Telecommunication Standardization Sector Recommendation*, no. P.862, 2001.

[10] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition.* (Prentice Hall signal processing series). Prentice Hall, 1993.

[11] N. B. Thien, Y. Wakabayashi, K. Iwai, and T. Nishiura, "Inter-frequency phase difference for phase reconstruction using deep neural networks and maximum likelihood," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1667–1680, 2023.

[12] F. Auger, É. Chassande-Mottin, and P. Flandrin, "On phase-magnitude relationships in the short-time Fourier transform," *IEEE Signal Processing Letters*, vol. 19, no. 5, pp. 267–270, 2012.

A Study on Multimodal Fusion and Layer Adapter in Emotion Recognition

Xiaohan Shi*,Yuan Gao†, Jiajun He*, Jinyi Mi*, Xingfeng Li‡, Tomoki Toda*
* Nagoya University, Japan.
E-mail: {xiaohan.shi, jiajun.he, mi.jinyi}@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp
† Kyoto University, Japan.
E-mail: gao.yuan.75x@st.kyoto-u.ac.jp
‡ City University of Macau, Macau.
E-mail: xfli@cityu.edu.mo

***Abstract*—Multimodal emotion recognition (MER) is a rapidly evolving field aimed at integrating information from various modalities, such as speech and text, to deepen our understanding of emotions. However, challenges in feature extraction and fusion hinder further advances in MER performance. To address these challenges, we propose different modality representations to capture emotional information comprehensively. Additionally, we introduce a novel layer adapter and multimodal fusion method to explore modality-dependent and modality-invariant interactions in MER. Extensive experimental results show the effectiveness of our approach, achieving state-of-the-art results with an absolute improvement of 1.89% over the baseline. The significant improvements validate the effectiveness of our proposed method in enhancing the MER system.**

I. Introduction

Emotion recognition is a growing field driven by the advancements in human-computer interactions (HCI) [1]. Earlier research focused mainly on using speech signals for emotion recognition, but recent studies show that unimodal systems are inadequate for complex HCI contexts. To improve emotion recognition accuracy, recent studies advocate integrating diverse modalities, including textual and visual data, to foster a comprehensive understanding of human emotional states [2]–[4]. Multimodal emotion recognition (MER) has applications in various domains, including social media analytics [5], market research [6], and enhancing customer service experiences [7]. Through its interdisciplinary implications, MER holds promise in uncovering novel insights into human emotional dynamics, thereby paving the path for innovative solutions across a myriad of practical domains. Despite substantial progress, two key challenges remain: extracting effective acoustic and lexical features for distinguishing emotions and developing suitable fusion methods to integrate multiple modalities in emotion recognition. This study endeavors to address each of these challenges, aiming to model the emotion recognition process in a multimodal context.

In MER, most features are employed in the speech and text modalities as these modalities predominantly promote emotion recognition within existing datasets. For example, Majumder et al. [8] propose a multimodal sentiment analysis approach based on hierarchical fusion and context modeling, enhancing sentiment recognition accuracy and robustness. Chuang et al. [9] introduce a multi-modal deep learning model for integrating speech and text data, achieving significant performance improvements in emotion recognition tasks.

In recent years, with the emergence of self-supervised learning (SSL) models, more researchers have shifted towards using pre-trained SSL features instead of traditional deep learning features to enhance the performance [10]. For example, Zou et al. [11] incorporated multiple acoustic features, including Mel-frequency cepstral coefficients, spectrograms, and Wav2vec 2.0 embeddings, for categorical emotion recognition, achieving a 7.03% improvement compared to using Wav2vec 2.0 embeddings alone. Additionally, Padi et al. [12] present a MER framework using mel-spectrogram and fine-tuning pre-trained BERT models, providing complementary emotional insights from the speech and text modalities.

Motivated by these findings, this paper focuses on extracting multimodal emotion features by integrating general speech and text self-supervised representations and context representations from the automatic speech recognition (ASR) module to enhance the MER system. Fusion methods are another key aspect of this task. In the literature, feature-level and decision-level fusion are widely used for multimodal systems. For instance, Yoon et al. [13] proposed a deep dual recurrent neural network to encode audio-text sequences, then concatenated their outputs to predict the emotion. Similarly, Pepino et al. [14] designed multiple dual RNNs to represent audio-text sequences and compared feature-level fusion and decision-level fusion approaches, highlighting their comparable performance. In contrast, our study introduces a fusion method at both the feature level and decision level, integrating a multi-level fusion module to extract discriminative features for MER.

Our contributions can be summarized as follows:

- We propose a novel feature extraction approach for learning multimodal emotion information by leveraging multimodal self-supervised representations.
- We introduce a multimodal fusion module that comprehensively integrates modality-dependent and modality-invariant emotional information from both speech and text modalities.

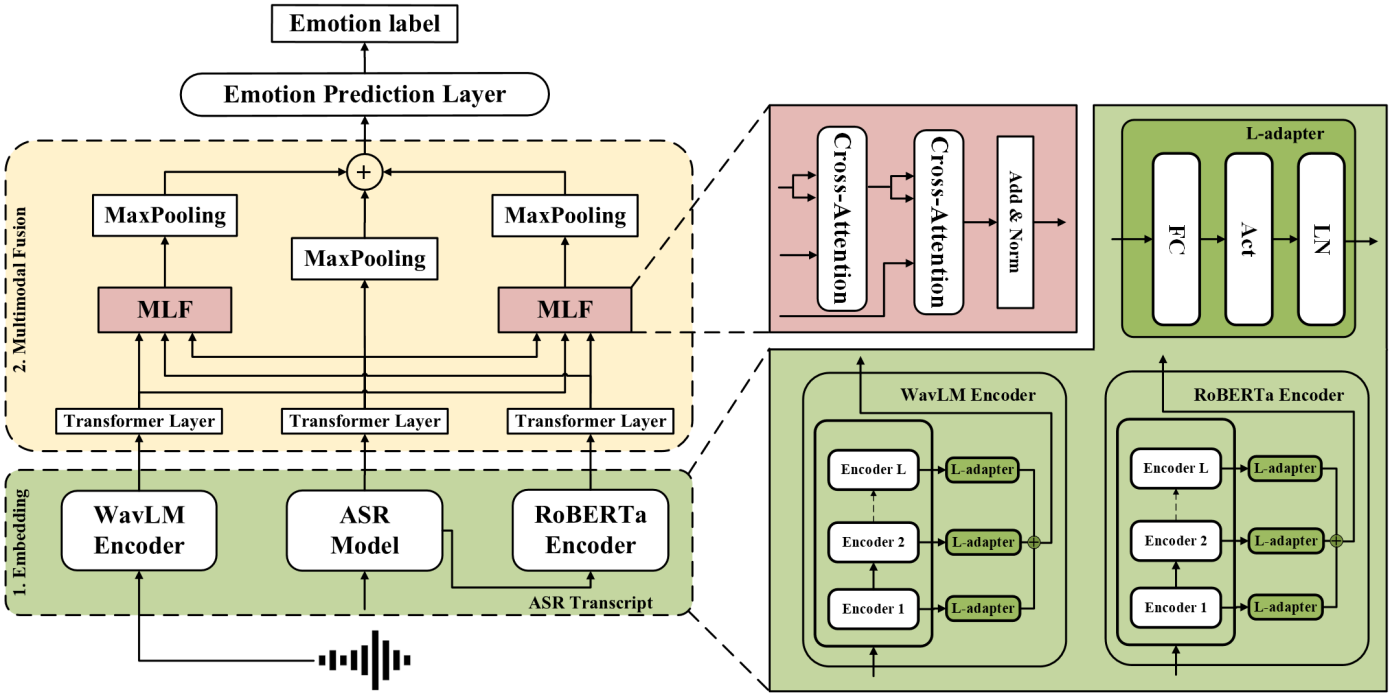


Fig. 1. The overall architecture of our proposed method.

- Our experimental results demonstrate that the proposed approach effectively addresses MER tasks.

II. Proposed Method

In this section, we outline our MER system and detail the multimodal fusion method. As depicted in Fig. 1, the network comprises four primary components: an ASR module for extracting textual and contextual information from raw speech, two embedding modules for encoding self-supervised representations, a multimodal fusion module for integrating modality-dependent and modality-invariant emotional information, and an emotion prediction module for predicting the emotion label.

A. Model Description

As depicted, raw audio utterances are directed into an acoustic feature encoder to extract self-supervised representations. Simultaneously, an ASR module is employed to obtain textual and contextual information. The transcripts then undergo a textual feature encoder to extract text self-supervised representations. These different modality representations are subsequently combined using the proposed multimodal fusion method, streamlining the final emotion recognition process.

B. Problem Formulation

The overall system can be expressed as the function $f(S,T,C) = L$, where each of the speech and context features $S = (s_1, s_2, \cdots, s_m)$ and $C = (c_1, c_2, \cdots, c_m)$ consists of m frames extracted from an utterance. The text modality $T = (t_1, t_2, \cdots, t_n)$ represents the original ASR hypotheses of an utterance, comprising n tokens. The model output $L \in \{l_1, l_2, \cdots, l_e\}$, where e represents the emotional categories.

C. Embedding Module

1) Speech Representations: To obtain a comprehensive understanding of acoustic features, we employ a pretrained SSL model, WavLM [15], as our speech self-supervised encoder.

WavLM employs a hybrid architecture consisting of convolutional neural network layers and a transformer encoder to capture speech features and contextual information effectively. We denote $H_S = (h_S^{(1)}, h_S^{(2)}, \cdots, h_S^{(m)})$ to represent the speech self-supervised representations, where m denotes the number of frames extracted from an utterance.

2) Context Representations: To acquire comprehensive information regarding contextual information from speech, we employ ASR representation as context features. In this study, we leverage Whisper [16] as our ASR module. Whisper is a supervised model specifically designed to transcribe spoken language into text. It has been trained on a diverse dataset comprising approximately 680,000 hours of speech collected from the web.

Whisper adopts an encoder-decoder transformer architecture; the encoder encodes the mel spectrogram information, and the decoder generates text as a sequence of words. In whisper, first, the raw audio inputs are converted to a log-mel spectrogram by the action of the feature extractor. Then, the Transformer encoder encodes the spectrogram to form a sequence of encoder-hidden states. Finally, the decoder

autoregressively predicts text tokens, conditional on both the previous tokens and the encoder's hidden states.

We denote $H_C = (h_C^{(1)}, h_C^{(2)}, \cdots, h_C^{(m)})$ to represent the context representations from whisper encoder layer, where *m* denotes the number of frames extracted from an utterance.

3) Text Representations: To acquire comprehensive information regarding lexical features, we leverage a pretrained SSL model, RoBERTa [17], as our text encoder. RoBERTa is an extension of the bidirectional encoder representations from the transformers model, which is specifically designed to address challenges related to long-range dependencies, and finely tuned for various natural language processing tasks. Pretrained on extensive corpora, including a dataset comprising 58 million tweets, RoBERTa exhibits exceptional contextual understanding, thereby enhancing text-related tasks. We denote $H_T = (h_T^{(1)}, h_T^{(2)}, \cdots, h_T^{(n)})$ to represent the text self-supervised representations, where *n* denotes the number of tokens extracted from an utterance.

4) L-adapter module (LA): To harness intermediate representations from the initial fine-tuning stages, layer adapters establish pathways from each encoder from the speech and text self-supervised model. Each layer adapter comprises a Fully connected (FC) layer, succeeded by a non-linear activation function and layer normalization, as illustrated in Fig. 1b. The application of the layer adapter results in adapted representations, computed as follows:

$$\alpha_S^l(\text{ or } \alpha_T^l) = \text{FC}((H_S^l) \text{ or } H_T^l) \tag{1}$$

for $l = 1, 2, \ldots, L$ in the encoder, and the weighted sum of the adapted representations is computed as:

$$H_S^*(\text{ or } H_T^*) = \sum_{l=1}^{L} w_l \alpha_S^l(\text{ or } \alpha_T^l) \tag{2}$$

This is fed into the emotion recognition classification, where w_l are learnable weights.

D. Multimodal Fusion (MF) Module

Inspired by [18], our MF is composed of two Multi-Level Fusion (MLF) blocks. The objective is to facilitate the learning of modality-dependent representations and modality-invariant representations.

In this section, we provide an in-depth explanation of the operation of each MLF block.

MLF Block adheres to the structure of a standard transformer layer, incorporating two cross-attention modules, and residual connections.

Initially, we employ three multi-head transformer layers to derive speech, text, and context representations.

$$H_S'(\text{ or } H_T', H_C') = \text{Transformer}(H_S^*(\text{ or } H_T^*, H_C)) \tag{3}$$

Then, two MLF blocks are used to derive speech, text self-supervised representations, and context representation. This is achieved by utilizing H_S' (or H_T') as queries and H_T' (or H_S') as keys and values within the first cross-attention block.

$$Q_1 = H_S'(\text{ or } H_T'), K_1 = H_T'(\text{ or } H_S'), V_1 = H_T'(\text{ or } H_S'). \tag{4}$$

$$H_S^{\text{T}}(\text{ or } H_T^{\text{S}}) = \text{Cross-Attention}(Q_1, K_1, V_1). \tag{5}$$

Next, the first cross-attention block output is used as queries and H_C' as keys and values within the second cross-attention block.

$$Q_2 = H_S^{\text{T}}(\text{ or } H_T^{\text{S}}), K_2 = H_C', V_2 = H_C'. \tag{6}$$

$$H_C^{\text{ST}}(\text{ or } H_C^{\text{TS}}) = \text{Cross-Attention}(Q_2, K_2, V_2). \tag{7}$$

Finally, we adopt Max pooling to obtain a 1-dimensional vector for each output. The final speech, context and text representations (H_C^{ST}, H_C^{TS}, H_C') are concatenated and written as follows:

$$H_{STC} = H_C^{\text{ST}} \oplus H_C^{\text{TS}} \oplus H_C'. \tag{8}$$

E. Emotion Classification Module

Emotion classification is conducted using the output representations H_{STC} of the MF module, which is subsequently passed through a FC layer and a SoftMax activation function.

$$P(y_{\text{emo}}|H_{STC}) = \text{SoftMax}(\text{FC}(H_{STC})). \tag{9}$$

where y_{emo} is the predicted emotion classification.

The loss function $\mathcal{L}_{\text{Emotion Classification}}$ is formulated using cross-entropy:

$$\mathcal{L}_{\text{Emotion Classification}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}) \tag{10}$$

where N denotes the total number of samples, C is the count of emotion classes, y_{ic} is the ground truth label for sample i and class c, and $\hat{y}_{ic}$ is the predicted probability of class c for sample i. This loss function penalizes the model based on the discrepancy between predicted probabilities and ground truth labels across all samples and classes, facilitating accurate classification.

III. Experimental Setup

A. Dataset

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is a widely utilized corpus in MER [19]. It contains 12 hours of audiovisual data, including audio, video, and textual transcriptions from 10 speakers. In each session, one male and one female performed a series of scripts or improvisational scenarios. For each utterance, three annotators assigned the categorical labels. Following common practice, we merged 'happy' and 'excited' into a single emotion class labeled 'happy'. We implemented the common practice of merging "happy" and "excited" into one emotion class "happy"; thus, the emotion labels in this dataset are happy (29%), sad (20%), angry (20%), and neutral (31%).

B. Experimental Procedure

We conduct two experiments in this study. Experiment 1 examines the impact of different modality representations on MER by comparing our approach with previous studies:

- **SAWC** [20]: This method modifies the importance weights based on confidence measures, thereby reducing the impact of ASR errors by focusing on relevant speech segments.
- **RMSER-AEA** [21]: This approach incorporates complementary semantic information, adjusts for ASR errors via an auxiliary task, and combines text and acoustic representations for SER.
- **SAMS** [22]: This approach utilizes high-level emotion representations as supervisory signals to establish a multi-spatial learning framework for each modality, facilitating cross-modal semantic learning and the exploration of fusion representations.

Meanwhile, we provide ablation experiments for each modality representation by visualizing the feature distributions using t-distributed stochastic neighbor embedding (t-SNE) [23], as shown in Fig. 2.

In Experiment 2, we explore the performance of the proposed fusion methods, namely Layer Adapter (LA) and Multimodal Fusion (MF), across different combinations of modality representations.

C. Implementation

Our deep learning models were developed using Python 3.7 and PyTorch 1.11.0. The speech self-supervised encoder was initialized using the WavLM-base model[1], producing speech self-supervised representations with a dimensionality of 768. For the text self-supervised encoder, we utilized the RoBERTa-base model[2], which has a hidden size of 768, 12 attention layers, and 12 attention heads. The context encoder was initialized using the Whisper-small model[3], resulting in context representations with a dimensionality of 768. The word error rate (WER) is 33.87% on the IEMOCAP dataset. the WavLM and RoBERTa models were fine-tuned, while the Whisper model was frozen during the training process. We used the Adam optimizer [24], with a dropout rate of 0.5 and a weight decay of 1e-5 to avoid overfitting and ensure model generalization. During training, we used a batch size of 32 and trained for 30 epochs.

D. Evaluation

To evaluate our results on the IEMOCAP dataset, which lacks a standard train/dev/test split, we adopt a leave-one-section-out cross-validation approach, consistent with prior studies [25]–[27]. We assess categorical MER performance using Unweighted Average Recall (UAR) and F1 scores, metrics commonly employed in experiments with imbalanced data to gauge performance [28]–[30] across the four discrete emotional labels.

[1]https://huggingface.co/microsoft/wavlm-base

[2]https://huggingface.co/FacebookAI/roberta-base

[3]https://huggingface.co/openai/whisper-small.en

IV. Results and Discussion

To analyze the different modality representations for MER, we compare the recognition performance of single modality and multimodal system, as shown in Table 1.

TABLE I
COMPARISON RESULTS OF OUR PROPOSED METHOD AND PREVIOUS MER STUDIES.

Modality	Model	UAR (%)	F1 (%)
Single modal	Speech (WavLM)	63.55	64.26
	Text (RoBERTa)	69.29	68.77
	Context (Whisper)	71.66	70.85
Multimodal	SAWC [20]	76.6	-
	RMSER-AEA [21]	76.4	-
	SAMS [22]	76.6	-
	Proposed	**77.73**	**77.24**
	w/o Speech	74.81	74.71
	w/o Text	67.21	67.01
	w/o Context	76.31	76.29

The effectiveness of incorporating features from multiple modalities is evaluated. The results demonstrate the effectiveness of multi-modal feature extraction. The improvements in UAR and F1 scores were 14.18% and 12.98% respectively, when compared to using only speech representation. When compared to using only text representation, the improvements are 8.44% and 8.47%, respectively, and compared to using only context representation, the improvements are 6.07% and 6.39%, respectively. Furthermore, compared with the MER previous studies, our proposed model achieves a 1.13% absolute improvement in UAR. Ablation results indicate that the most critical representation is the text representation, with its removal causing a 10.52% and 10.23% decrease in UAR and F1 scores, respectively. On the other hand, the context representation has the least impact on performance, leading to a 1.42% and a 0.95% decrease in UAR and F1 scores, respectively. These results highlight the significant role of multimodal representations in enhancing MER performance. The text modality is particularly influential. The minor impact of context representation suggests that its contribution is not as pivotal as that of speech and text. Furthermore, the feature distributions under the different modality representations and the proposed method are visualized using t-SNE, as depicted in Fig. 2. This visualization provides additional evidence supporting the effectiveness of our proposed method.

To analyze the effects of MF and LA on MER, we compared their performance across different modalities in detail, as shown in Table 2.

The results indicate that the combination of MF and LA yields superior performance compared to using only either MF or LA. Specifically, when utilizing speech and text representations, the observed improvements in UAR and F1 scores are 2.45% and 2.64%, respectively, compared to using only MF. With LA, these improvements are 1.75% and 3.08%,

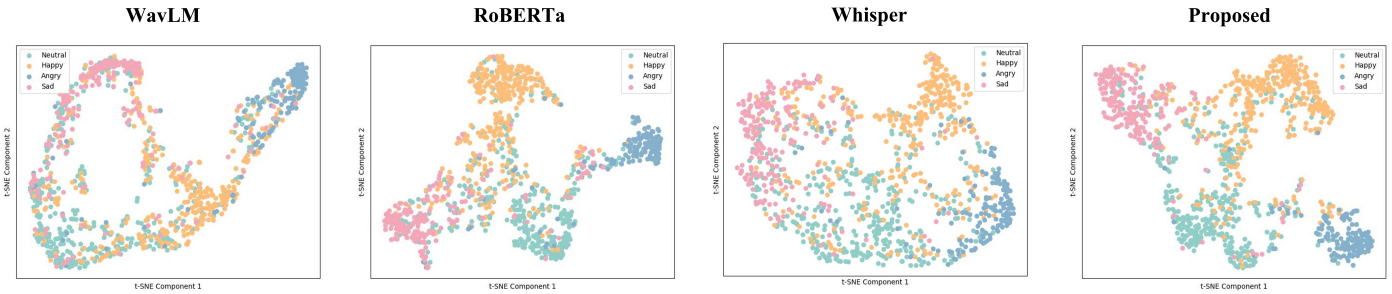


Fig. 2. The t-SNE visualization of different modality representations on IEMPOCAP session 2.

TABLE II
COMPARISON OF MER PERFORMANCE OBTAINED BY THE PROPOSED MF AND LA FUSION METHODS.

Model	Method	UAR (%)	F1 (%)
Speech + Text	-	72.27	72.06
	MF	73.86	73.65
	LA	74.56	73.21
	MF + LA	**76.31**	**76.29**
Speech + Context	-	65.04	64.39
	MF	66.89	66.59
	LA	66.74	66.94
	MF + LA	**67.21**	**67.01**
Text + Context	-	72.65	72.20
	MF	73.86	73.65
	LA	73.90	73.65
	MF + LA	**74.81**	**74.71**
Speech + Text + Context	-	75.84	75.59
	MF	76.83	76.24
	LA	77.61	77.05
	MF + LA	**77.73**	**77.24**

respectively. In the speech and context representations, the enhancements in UAR and F1 scores are 0.32% and 0.42%, respectively, when using only MF, and 0.47% and 0.07%, respectively, with LA. For text and context representations, the improvements in UAR and F1 scores are 0.95% and 1.06%, respectively, with MF, and 0.91% and 1.06%, respectively, with LA. Finally, when employing speech, text, and context representations, the observed enhancements in UAR and F1 scores are 0.9% and 1%, respectively, with MF, and 0.12% and 0.19%, respectively, with LA. In summary, our analysis highlights that the combination of MF and LA consistently improves MER performance across various modality combinations.. This finding highlights the significant potential of integrating multimodality representation to achieve robust and reliable emotion recognition.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a multimodal self-supervised representation extraction and fusion approach to capture emotional information comprehensively. In the proposed method, we introduce a novel layer adapter to explore modality-dependent and modality-invariant interactions in MER. Our findings reveal that the proposed representations across different modalities perform better than single-modality representations. Moreover, our novel layer adapter and modality fusion method for integrating modality-invariant emotional information consistently achieves higher accuracy in MER.

For future research, we advocate for further exploration of innovative modality fusion methods to further enhance the accuracy of MER.

VI. ACKNOWLEDGMENT

This work was financially supported by JST SPRING, Grant Number JPMJSP2125, and in part by JST CREST Grant Number JPMJCR19A3, Japan, and JSPS KAKENHI Grant Number 21H05054.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] Y. Gao, H. Shi, C. Chu, and T. Kawahara, "Speech emotion recognition with multi-level acoustic and semantic information extraction and interaction," in *Interspeech 2024*, 2024, pp. 1060–1064.

[3] H. Sun, S. Zhao, X. Wang, W. Zeng, Y. Chen, and Y. Qin, "Fine-grained disentangled representation learning for multimodal emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 051–11 055.

[4] J. Tian, D. Hu, X. Shi, *et al.*, "Semi-supervised multimodal emotion recognition with consensus decision-making and label correction," in *Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing*, 2023, pp. 67–73.

[5] N. Andalibi and J. Buss, "The human in emotion recognition on social media: Attitudes, outcomes, risks," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–16.

[6] S. Lugović, I. Duner, and M. Horvat, "Techniques and applications of emotion recognition in speech," in *2016 39th international convention on information and communication technology, electronics and microelectronics (mipro)*, IEEE, 2016, pp. 1278–1283.

[7] W. Han, T. Jiang, Y. Li, B. Schuller, and H. Ruan, "Ordinal learning for emotion recognition in customer service calls," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2020, pp. 6494–6498.

[8] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-based systems*, vol. 161, pp. 124–133, 2018.

[9] Z.-J. Chuang and C.-H. Wu, "Multi-modal emotion recognition from speech and text," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing*, 2004, pp. 45–62.

[10] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, "Using Semi-supervised Learning for Monaural Time-domain Speech Separation with a Self-supervised Learning-based SI-SNR Estimator," in *Proc. INTERSPEECH 2023*, 2023, pp. 3759–3763. DOI: 10.21437/Interspeech.2023-85.

[11] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7367–7371.

[12] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models," *arXiv preprint arXiv:2202.08974*, 2022.

[13] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE spoken language technology workshop (SLT)*, IEEE, 2018, pp. 112–118.

[14] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6484–6488.

[15] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers," *arXiv preprint arXiv:2307.03183*, 2023.

[17] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[18] J. He, X. Shi, X. Li, and T. Toda, "Mf-aed-aec: Speech emotion recognition by leveraging multimodal fusion, asr error detection, and asr error correction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 066–11 070.

[19] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[20] S. Dutta and S. Ganapathy, "Multimodal transformer with learnable frontend and self attention for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6917–6921.

[21] B. Lin and L. Wang, "Robust multi-modal speech emotion recognition with asr error adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[22] M. Hou, Z. Zhang, C. Liu, and G. Lu, "Semantic alignment network for multi-modal emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 5318–5329, 2023.

[23] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] X. Shi, X. Li, and T. Toda, "Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation," in *Proc. Interspeech*, 2023, pp. 765–769.

[26] H. Sun, S. Zhao, X. Kong, *et al.*, "Iterative prototype refinement for ambiguous speech emotion recognition," in *Interspeech 2024*, 2024, pp. 3200–3204.

[27] X. Shi, X. Li, and T. Toda, "Multimodal fusion of music theory-inspired and self-supervised representations for improved emotion recognition," pp. 2024–2350, 2024.

[28] Y. Gao, L. Wang, J. Liu, J. Dang, and S. Okada, "Adversarial domain generalized transformer for cross-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, 2023.

[29] X. Shi, S. Li, and J. Dang, "Dimensional emotion prediction based on interactive context in conversation.," in *INTERSPEECH*, 2020, pp. 4193–4197.

[30] X. Li, X. Shi, D. Hu, *et al.*, "Music theory-inspired acoustic representation for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2534–2547, 2023.

Analyzing House Music: Relations of Audio Features and Musical Structure

Justin Tomoya Wulf* and Tetsuro Kitahara*
* Nihon University, Tokyo, Japan
E-mail: {wulf, kitahara}@kthrlab.jp

***Abstract*—House music is composed differently from popular music; each part repeats the same pattern multiple times, and the structure is formed by inserting and/or deleting patterns. However, it is unclear what specific factors determine the structure of house music. Our previous study attempted to analyze the musical structure of some pieces of house music, but the number of pieces used in the experiment was insufficient. In this paper, we analyze 194 pieces of house music from different perspectives: frequency, sound pressure, and drum patterns. First, each piece is segmented and a label is assigned to each segment. Then, the spectral centroid, sound pressure, onsets for each drum, and other features are calculated for each section (intro, drop, break, and outro). The relationship between these features and sections is then investigated. Results show that sound pressure and other featurs of unseparated sources vary between sections, while the spectral centroid and similar features do not show significant differences.**

I. Introduction

House music is a unique genre that differs significantly from conventional popular music. In the latter, pitch-related features like melodies and chord progressions are the primary factors. However, house music typically involves each instrument part repeatedly playing the same patterns. The musical structure is shaped through the addition and removal of these instrument parts. For instance, the intro of a piece often starts just with drums. Following this, the bass begins to play. After the intro, a higher-pitch instruments come into the mix. This gradual layering of instruments leads to the unfolding of the musical (and often emotional) progression. Understanding these structural elements in crucial for house music composers.

The study of musical structure analysis has been pursued by numerous researchers. Levy et al. proposed a technique for music segmentation using constrained clustering [1]. Similarly, Paulus et al. introduced a method for musical structure analysis deploying probabilistic fitness measures and a greedy search [2]. Meanwhile, McFee et al. utilized spectral clustering for structure analysis [3]. Some studies have leveraged deep neural networks to enhance musical segmentation [4], [5], [6]. Other studies were more focused on visualizing musical structures instead of automatic detection [7], [8]. Nonetheless, no one has yet explored the correlation between musical features and musical structure specifically in house music.

In our previous paper [9], we examined the musical structure of several pieces of house music. We focused on the sound pressure of individual instrument parts, generating single-instrument sources through sound source separation and visualizing their sound pressures. From this, we inferred that the sound pressure of the bass and drums might be highly correlated to the musical structure. Nevertheless, our analysis was restricted by the limited number of pieces examined, and there was potential future work regarding features beyond sound pressure.

In this paper, we concentrate on frequency, sound pressure, and drum patterns. Once the musical structure of a given source is examined, factors like spectral centroid, sound pressure, and the onsets of drum instruments are analyzed for each section (i.e., intro, drop, break, and outro). Our goal is to establish relationships between such features and musical structure through the analysis of 194 pieces.

II. Experimental Methods

In this paper, we analyze audio features that are highly related to musical structures in house music. We define the following four sections:

- *Intro*: The introductory part of the track
- *Drop*: The climax part of the track
- *Break*: The part other than the climax part of the track
- *Outro*: The ending part of the track

A. Hypotheses

We formulated the following three hypotheses about house music:

1) Analysis of Frequency: Our preliminary experiment [9] suggests that the sound pressure of the drums and bass tends to increase in the drop. The drums and bass parts emphasize low-frequency sounds. Therefore, lower-frequency sounds would be strengthened in the drop compared to other sections.

2) Analysis of Sound Pressure: The presence of drums and bass increases the sound pressure, making it higher in the drop section compared to other sections.

3) Analysis of Drum Sounds: The sound pressure of the drums part depends on the drum pattern, as well as the intensity of each note. For instance, highly dense drum patterns usually have higher sound pressure on temporal average. Therefore, drum patterns would have specific characteristics in different sections.

B. Dataset

To ensure a diverse collection of pieces, seven notable compilation albums were selected from well-known house music labels, totaling 194 pieces. Pieces that were not house music in the albums were excluded. The selected albums are listed in Table I.

TABLE I
HOUSE MUSIC ALBUMS USED IN THE EXPERIMENT

Label Name	Album Name
De La Groove	3 Years On
Toolroom Records	Best Of Toolroom 2018
Defected Records	Defected Presents Most Rated 2017
Defected Records	Defected Presents Most Rated 2019
Defected Records	Defected Presents Most Rated 2020
Glasgow Underground	Glitterbox - Discotheque
Toolroom Records	This Is Toolroom 2020

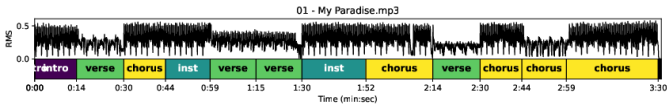


Fig. 1. Example of segmentation with All-In-One

C. Flow of Analysis

1) Obtaining Segment Labels: Each sound source is initially divided into sections. Following this, each section is assigned a descripter (label) using All-In-One [10]. Figure 1 showcases an example of segmentation utilizing All-In-One.

The All-In-One system provides ten types of segment labels. However, as only four types of sections are used in our experiment, the labels are converted to the sections that are most similar in terms of their roles within the musical structure. The conversion rules are illustrated in Table II.

While the segment labels obtained with All-In-One are generally reliable for the intro and outro, errors were frequently observed for drops and breaks. Thus, the segment labels are adjusted with simple rules to fit the following structure:

$$\text{Intro} \rightarrow [\text{Break}] \rightarrow (\text{Drop} \rightarrow \text{Break})^* \rightarrow [\text{Drop}] \rightarrow \text{Outro}$$

where [ ] indicates optional segments, and ( )* indicates zero or more repetitions.

2) Analysis of Frequency: The sound source in question is separated into four parts (drums, bass, vocals, and others) using Demucs [11]. Subsequently, the spectral centroid is computed for each of these parts. This feature signifies the frequency point at which energy is evenly balanced, both above and below the frequency. The spectral centroid for each instrument part is calculated using Librosa [12]. To prevent the calculation of the spectral centroid in silent sections, intervals where the root mean square (RMS) falls below a specified threshold are excluded from the calculation.

Next, utilizing the segment labels acquired with All-In-One, the temporal averages of the spectral centroid for each section are calculated from both separated (i.e., drums, bass, vocals, and others) and unseparated sources.

3) Analysis of Sound Pressure: The RMS of the waveform is calculated as a feature representing the sound pressure. Similarly to the analysis of frequency, the sound source is separated to four parts (drums, bass, vocals, and others) and the RMS is calculated from both separated and unseparated sources. The temporal averages of the RMS for each section are calculated.

TABLE II
CONVERSION OF SEGMENT LABELS

Original Section	Converted Section
start	intro
intro	intro
break	break
bridge	break
inst	break
verse	break
solo	drop
chorus	drop
outro	outro
end	outro

4) Analysis of Drum Patterns: The drum sound source obtained with Demucs is transcribed in the MIDI format using Omnizart [13]. We count note-on messages for the kick, snare, and hi-hat (referred to here as individual drums). In addition, to consider their temporal evolution, we calculate the counts of their note-on messages every eight measures. We also analyze the relationship between section changes and drum-pattern changes. First, we calculate the inter-onset intervals (IOIs). Then, we identify timings where their IOIs deviate significantly from the temporal average (using the standard deviation as a threshold in this analysis). These timings are considered *drum-pattern boundaries* (the points where the drum pattern changes). If the drum-pattern boundaries and section boundaries occur within one second of each other, they are considered to be synchronized.

III. RESULTS

We discuss the results of the analysis of frequency, sound pressure, and drum patterns.

A. Analysis of Frequency

Figure 2 illustrates the distributions of the temporal averages of the spectral centroid for each section calculated from unseparated sources. As this figure indicates, these distributions' medians approximately fall within the 4000 to 4500 Hz range for all sections. Our original hypothesis, presented in Section 2, that the drop section would emphasize lower frequencies and the break section would de-emphasize them, found no significant differences in frequencies between two sections. Furthormore, the drop and break sections reveal only slight dispersion in values, constrasting the intro and outro sections which demonstrate more considerable dispersion.

Figure 3 displays calculations derived from separated sources (drums, bass, vocal, and others). Various instrument parts present distinct temporal averages of the spectral centroid throughout all sections; however no significant differences are identified among the sections for all parts. In house music, each instrument tends to repeat the same phrase, and our result matches this tendency.

The bass possesses the lowest spectral centroid across all sections, while the drums consistently exhibit the highest. This could be attributed to the presence of high-frequency instruments within the drums, such as snares and hi-hats.

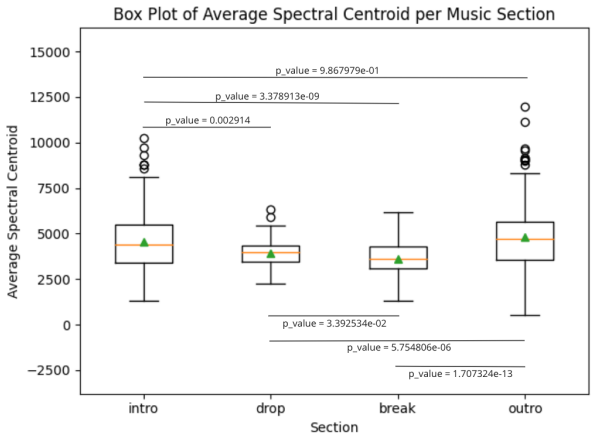


Fig. 2. The temporal averages of the spectral centroid for each section (calculated from unseparated sources). The white circles represent outliers that exceed 1.5 times the interquartile range. The green triangles represent the mean values. The orange lines represent the median values.

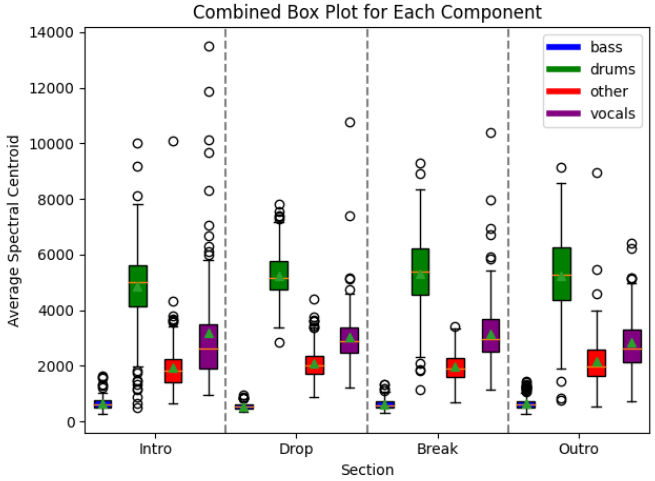


Fig. 3. The temporal averages of the spectral centroid for each section (calculated from instrument-separated sources). The white circles represent outliers that exceed 1.5 times the interquartile range. The green triangles represent the mean values. The orange lines represent the median values.

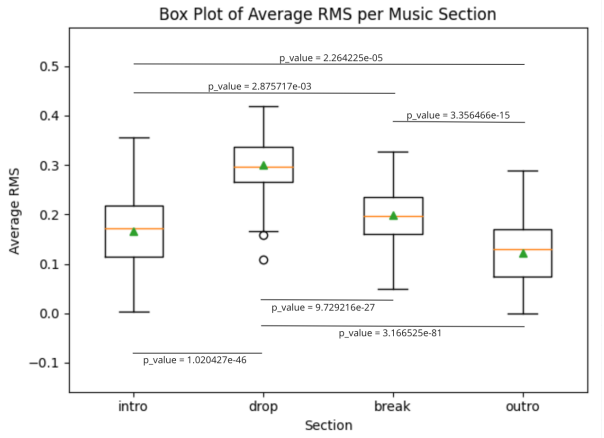


Fig. 4. The temporal averages of the sound pressure (the RMS of the waveform) for each section (calculated from unseparated sources). The white circles represent outliers that exceed 1.5 times the interquartile range. The green triangles represent the mean values. The orange lines represent the median values.

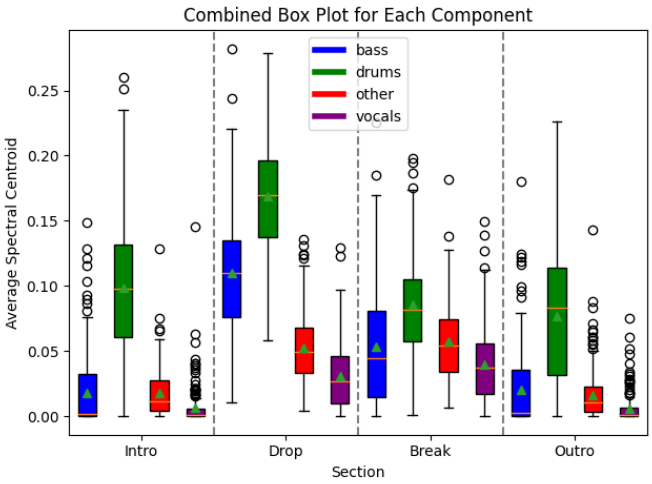


Fig. 5. The temporal avereages of the sound pressure (the RMS of the waveform) for each section (calculated from instrument-separated sources). The white circles represent outliers that exceed 1.5 times the interquartile range. The green triangles represent the mean values. The orange lines represent the median values

When calculating the temporal average of the spectral centroid, we excluded silent frames with an amplitude (RMS) lower than a specified threshold. Since this preprocessing is performed independently for each source, the number of frames differs among the instrument parts.

B. Analysis of Sound Pressure

Figure 4 illustrates the distributions of the temporal averages of the sound pressure (the RMS of the waveform) for each section, calculated from unseparated sources. This figure highlights significant differences in sound pressure across the sections. The drop section possesses the highest sound pressure, succeeded by the break, intro, and outro sections. The intro and outro sections exhibit a larger dispersion in values compared to the drop and break.

Figure 5 illustrates calculations derived from separated sources (bass, drums, vocals, and others). The sound pressure of the drums and bass is elevated in the drop section compared to other instrument parts, with a decrease noticeable in the break section. This suggests that the presence of these instruments plays a significant role in identifying the drop and break sections. Moreover, the sound pressure of the drums is markedly higher in the intro and outro sections compared to other instrument parts. The bass exhibits subdued sound pressure in sections other than the drop. A noticeable distinction in the sound pressue of the vocals and other instrument parts between the drop and break sections is absent, and they are almost silent in the intro and outro sections. The presence or absence of vocals is likely a key factor distinguishing the intro/outro sections from other sections.

Figure 6 presents the distributions of section-wise sound pressures in the bass, drums, and vocal+others parts. These diagrams reveal that the drum sound pressure in the drop

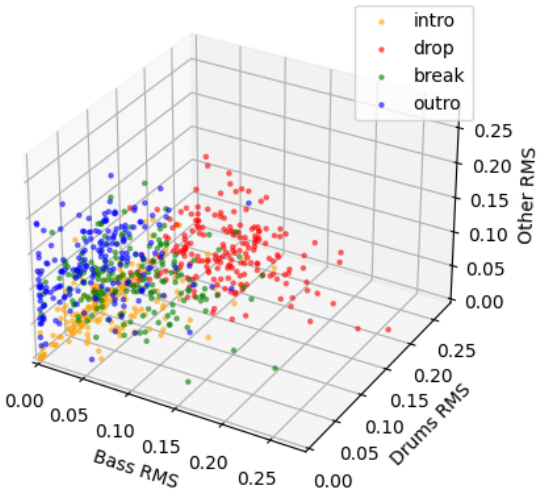


Fig. 6. Distributions of the section-wise sound pressures (classified by color)

section is the highest among all sections, with the bass and vocal+others parts also showing relatively high. In the break section, vocal+others parts sustain a sound pressure similar to that in the drop section, but the drums and bass display lower sound pressures than in comparison. During the intro section, the drum sound pressure exhibits variability, while the bass and vocal+others parts yield barely any sound. These trends align with those previoulsy identified in Figure 5.

On the other hand, the sound pressure of the drums and bass in the outro section remains unchanged from the intro section. However, the sound pressure of the vocal+others parts is higher compared to the intro section. This may be because it is common for the intro to start with only the drums and for the outro to conclude with the drums and other instrument parts in house music.

C. Analysis of Drum Patterns

Figure 7 illustrates the number of onsets for each individual drum in each section. Although the numbers of onsets vary among the individual drums, there are no significant differences in the onsets of each drums across the sections.

Figures 8 to 10 display the temporal development of the number of onsets for each individual drum, calculated every eight bars. These figures reveal that the kick drum occurs approximately 64 times per eight bars, the snare drum about 32 times, and the hi-hat around 128 times. This suggests that the kick drum is played in the eighth notes, the snare in the quarter notes, and the hi-hat in the sixteenth notes. Furthermore, greater variances are observed in the snare and hi-hat compared to the kick drum, indicating that they are more susceptible to variation. In contrast, the low variance exhibited by the kick drum suggests its rhythm remains consistent. No temporal trends were identificed in the onsets of each drum.

Figures 11 and 12 show that the rate of synchronization between the section and drum-pattern boundaries is around 30% to 40%.

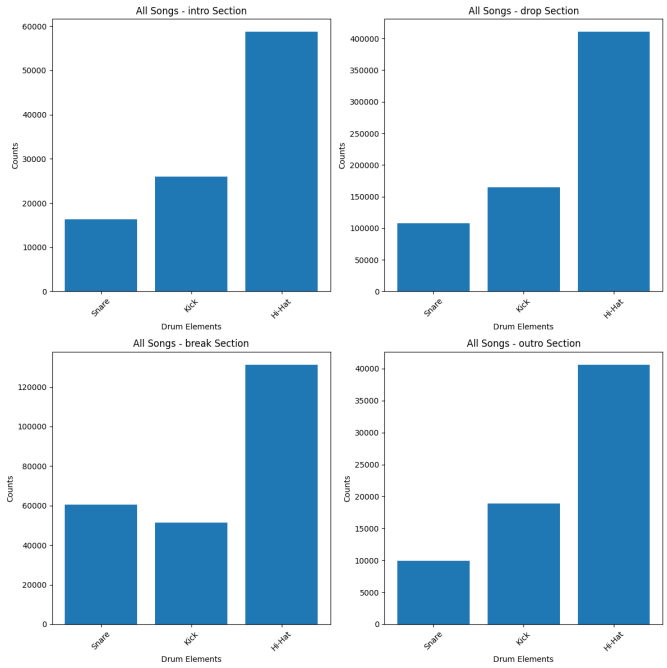


Fig. 7. The number of onsets in each section

IV. Conclusion

In this paper, we investigated audio features that relate to the musical structure of house music. First, we hypothesized that the spectral centroid for the drop section should be significantly higher than that for other sections and analyzed each section's spectral centroid with 194 pieces. The results, however, show no significant differences of the spectral centroid among sections. Second, we hypothesized that the sound pressure for the drop section should be significantly higher than that for other sections. The results show the sound pressure for the drop is higher as expected. Finally, we analyzed rhythm patterns played with the drums. The results imply each drum (i.e., the kick, snare, and hi-hat) has a specific periodicity.

To help musicians compose house music, we need more findings. For example, the tendency of spectral centroid and sound pressure may depend on composers. Also, we have to address how to generate specific advice from our results. Of course, more detailed experiments, such as investigating more audio features and more musical pieces, are needed. Comparative analysis with pieces from other genres is also slated to be conducted to further understand the unique characteristics of house music.

Acknowledgments

This work has been supported by JSPS Kakenhi No. 23K24966 and No. 24H00748.

References

[1] Mark Levy and Mark Sandler: Structural segmentation of musical audio by constrained clustering, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, Issue 2, pp. 318–326, 2008.

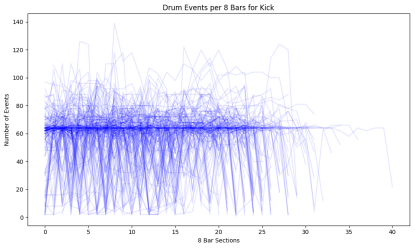


Fig. 8. The number of onsets in every eight bars (kick)

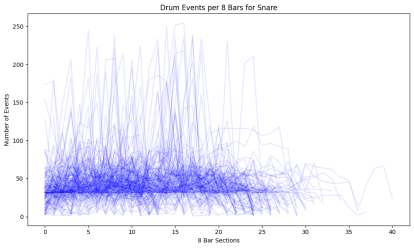


Fig. 9. The number of onsets in every eight bars (snare)

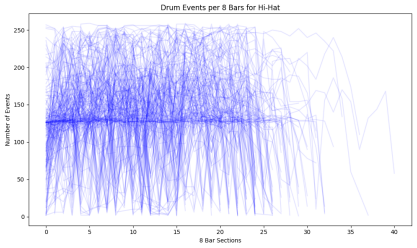


Fig. 10. The number of onsets in every eight bars (hi-hat)

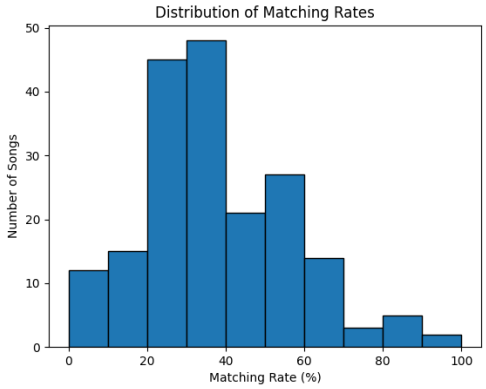


Fig. 11. The synchronization rate of section and drum-pattern boundaries (the number of synchronized boundaries divided by the total number of all drum-pattern boundaries)

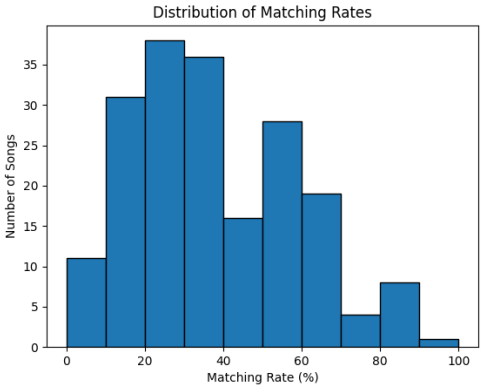


Fig. 12. The synchronization rate of section and drum-pattern boundaries (the number of synchronized boundaries divided by the total number of all section boundaries)

[2] Jouni Paulus and Anssi Klapuri: Music Structure Analysis Using a Probabilistic Fitness Measure and a Greedy Search Algorithm, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, Issue: 6, pp. 1159–1170, 2009.

[3] Brian McFee and Dnaiel P.W. Ellis: Analyzing Song Structure with Spectral Clustering, Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014), pp. 405–410, 2014.

[4] Justin Salamon, Oriol Nieto, and Nicholas J. Bryan: Deep Embeddings and Section Fusion Improve Music Segmentation, Proceedings of the International Society for Music Information Retrieval (ISMIR 2021), pp. 594–601, 2021.

[5] Carlos Hernandez-Olivan and Jose R. Beltran, David Diaz-Guerra: Music Boundary Detection using Convolutional Neural Networks: A comparative analysis of combined input features, International Journal of Interactive Multimedia & Artificial Intelligence, Vol. 7, No. 2, pp. 78–88, 2021.

[6] Morgan Buisson, Brian Mcfee, and Slim Essid, Helene-Camille Crayencour: Learning Multi-Level Representations for Hierarchical Music Structure Analysis, Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022), pp. 591–597, 2022.

[7] Jonathan Foote: Visualizing Music and Audio using Self-Similarity, Proceedings of the Seventh ACM International Conference on Multimedia (Part 1), pp. 77—80, 1999.

[8] Meinard Muller and Nanzhu Jiang: A Scape Plot Representation for Visualizing Repetitive Structures of Music Recordings, ISMIR, 2012.

[9] Justin Tomoya Wulf and Tetsuro Kitahara: Visualizing Musical Structure of House Music, Proceedings of the 16th International Symposium on Computer Music Multidisciplinary Research (CMMR 2023), pp. 784–787, 2023.

[10] Taejun Kim and Juhan Nam: All-In-One Metrical And Functional Structure Analysis With Neighborhood Attentions on Demixed Audio: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2023.

[11] Alexandre Défossez: Hybrid Spectrogram and Waveform Source Separation, ISMIR 2021 MDX Workshop, p. 11, 2021.

[12] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto: librosa: Audio and Music Signal Analysis in Python, SciPy 14th Python in Science Conference, 2015.

[13] Yu-Te Wu, Yin-Jyun Luo, Tsung-Ping Chen, I-Chieh Wei, Jui-Yang Hsu, Yi-Chin Chuang, and Li Su: Omnizart: A General Toolbox for Automatic Music Transcription, Journal of Open Source Software, Vol. 6, No. 68, 3391, 2021.

Band-Split Inter-SubNet: Band-Split with Subband Interaction for Monaural Speech Enhancement

Yen-Chou Pan*, Yih-Liang Shen*, Yuan-Fu Liao†, and Tai-Shih Chi*
* Institute of Communications Engineering, National Yang Ming Chiao Tung University, Taiwan
E-mail: yenchoupan@gmail.com; dennis831209@gmail.com; tschi@nycu.edu.tw
† Institute of Artificial Intelligence Innovation, Industry-Academia Innovation School,
National Yang Ming Chiao Tung University, Taiwan
E-mail: yfliao@nycu.edu.tw

***Abstract*—Speech enhancement models are developed to improve quality and intelligibility of speech for numerous daily applications. With the rapid development of technology, the neural network based speech enhancement models show significantly improved performance. The subband-based models focus on local spectral patterns and achieve outstanding results with fewer parameters. In this paper, we propose a subband-based composite model named Band-Split Inter-SubNet. It adopts the new constant-Q band-split setting to mimic human auditory perception. The proposed model demonstrates superior performance to other state-of-the-art models on the DNS Challenge - Interspeech 2021 dataset. Detailed analyses on experimental results demonstrate that the proposed band-split setting is effective, and the influence of neighboring frequency bins on the center-frequency bin across different frequency bands varies slightly.**

I. Introduction

In daily life, noise interference is one major factor hindering human speech communication and automatic speech recognition. It degrades both speech quality and intelligibility. The goal of speech enhancement models is to eliminate the interference of noise on speech. Conventional speech enhancement methods primarily adopt statistical theories to suppress stationary noise. With the advancement of technology, neural network (NN) based methods have shown significantly improved results in many complex and challenging situations, such as in low signal-to-noise ratios (SNRs) and reverberant conditions. At present, NN-based speech enhancement models are mainly implemented in the time domain or the frequency domain. The time-domain models [1], [2] directly generate a clean audio waveform from the noisy audio waveform. The frequency-domain models either directly predict the spectrogram of clean speech [3] or predict relevant masks, such as the ideal binary mask (IBM) [4], the ideal ratio mask (IRM) [5], and the complex ideal ratio mask (cIRM) [6][7], from the noisy spectrogram. Considering system robustness and computational complexity, frequency-domain models are more widely used than time-domain models.

Recently, the frequency-domain subband-based speech enhancement models [8], [9] have shown excellent performance in monaural speech enhancement. This approach is based on the idea that nearby frequency components are more influential and effective in distinguishing speech from noise than far-away frequency components. There are mainly two types of subband speech enhancement models, i.e., band-split [10], [11] and neighboring-frequency [12] models, as shown in Fig. 1. The band-split models divide the full frequency band into multiple subbands, with each subband processed independently. This allows for the capture of more local information. However, this type of model is limited by treating each band as a separate channel, preventing sharing information across bands. Consequently, various modules have been developed to enhance the interaction between channels to alleviate this limitation.

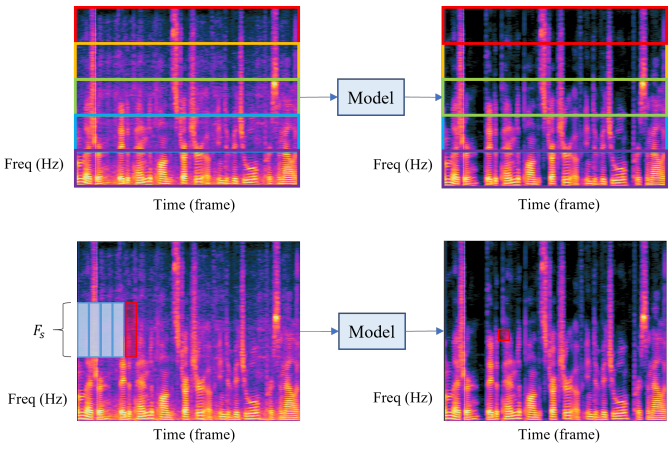


Fig. 1. Subband models: (Top) Band-split model; (bottom) neighboring-frequency model.

On the other hand, the neighboring-frequency models predict the center-frequency bin by considering its neighboring frequency bins as additional information. However, due to the lack of fullband information, this type of models can be less effective in conditions where local time-frequency features are not well-observed. To tackle this problem, neighboring-frequency models incorporating fullband information have been developed [13], [14], [15], enabling models to capture both local and global features. Not surprisingly, this combination results in increased number of model parameters and complexity. To reduce the complexity, Chen et al. [16] proposed a novel lightweight framework named Inter-SubNet, which comprises an interactive SubInter module to effectively captures global and local spectral patterns. Inspired by concepts of both types of subband models, we propose

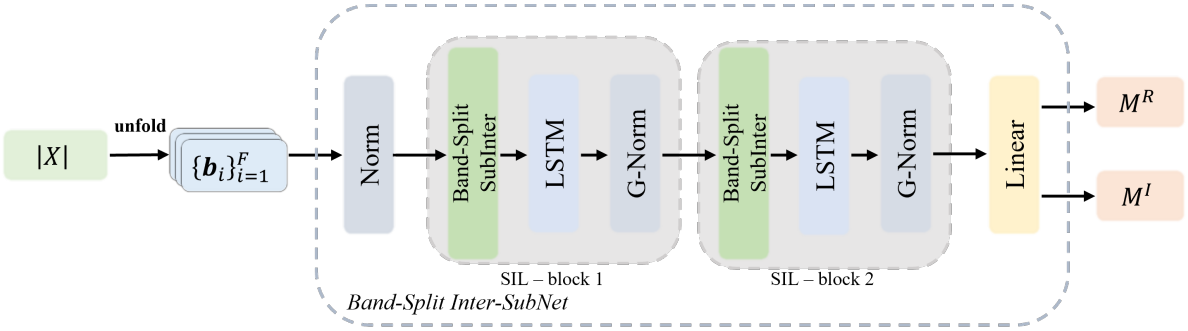


Fig. 2. The architecture of the proposed Band-Split Inter-SubNet model, which is extended from Inter-SubNet [16].

a combinational model, named Band-Split Inter-SubNet, to leverage the influence of neighboring frequency bins on the center-frequency bin across different frequency bands. The contributions of this paper can be summarized as follows:

1) We propose a combinational subband model from two types of models to achieve better performance.
2) Rather than the uniform splitting setting, we use the constant-Q band-split setting by mimicking human auditory perception to achieve better performance.

The rest of the paper is organized as follows. In Section 2, we introduce related work and the proposed Band-Split Inter-SubNet model. The experimental setup including datasets, training setup and baseline models are introduced in Section 3. In Section 4, we present results and discussions. Lastly, we conclude the paper in Section 5.

II. Proposed Model

A. Problem formulation

The signal model in the short-time Fourier transform (STFT) domain for monaural speech can be written as

$$X(t,f) = S(t,f) + N(t,f) \tag{1}$$

where t and f are the indexes of time frame and frequency bin, respectively. $X(t,f)$, $S(t,f)$ and $N(t,f)$ denote the complex spectrograms of noisy speech, clean speech, and interference noise.

The architecture of the proposed Band-Split Inter-SubNet model is depicted in Fig. 2. It is composed of two stacked SubInter-LSTM (SIL) blocks and one linear layer. The SIL block has been shown effective in Inter-SubNet [16]. It can capture global spectral information while maintaining the ability to focus on local spectral patterns. The "G-norm" indicates group normalization [17]. The magnitude spectrogram $|X| \in \mathbb{R}^{F\times T}$ and the real part and the imaginary part of cIRM (M^R and M^I) are the input and the output of the model, respectively, where F and T denote the numbers of overall frequency bins and time frames.

We concatenate the i-th frequency bin $|X_i| \in \mathbb{R}^{1\times T}$ with the $2\times n$ neighboring frequency bins to form a subband unit b_i by the "unfold" operation:

$$b_i = [|X_{i-n}|, \cdots, |X_i|, \cdots, |X_{i+n}|] \in \mathbb{R}^{F_s\times T} \tag{2}$$

where $F_s = 2n+1$.

B. Band-Split SubInter module

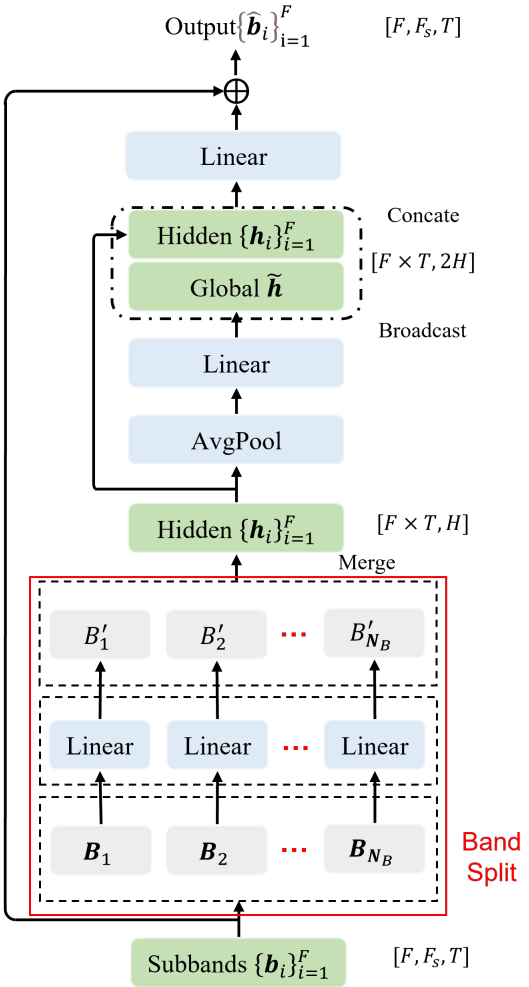


Fig. 3. Detailed structure of the Band-Split SubInter module.

The proposed Band-Split SubInter module shown in Fig. 2 is detailed in Fig. 3, which is slightly different from the original SubInter module in the Inter-SubNet framework [16]. Instead of using the same linear layer for all subband units,

the subband units of all frequencies are firstly divided into N_B bands and the subband units of each band are processed by each linear layer separately and merged together as the subband features $\{\boldsymbol{h}_i\}_{i=1}^{F}$, which contain local information from neighboring frequencies. After that, the subband features $\{\boldsymbol{h}_i\}_{i=1}^{F}$ are concatenated with global features $\tilde{\boldsymbol{h}}$, obtained by averaging local subband features followed by the second linear layer, and sent to the third linear layer. At the end, the final output subband features $\{\hat{\boldsymbol{b}}_i\}_{i=1}^{F}$ are obtained by applying a residual connection between the input subband features and the extracted subband features.

C. Setting for splitting bands

Typical band-split models split bands uniformly along the linear-frequency axis, i.e., all bands have the identical frequency bandwidth. In this paper, we propose a new band-split setting, as shown in Fig. 4 to mimic human auditory perception, which is more akin to a logarithmic frequency scale. Similar to the constant-Q transform, each bandwidth is half that of the previous band from high to low frequency. The first band is derived by splitting the fullband with the highest frequency of F_B Hz in half. And then, the second band is obtained by splitting the remaining frequencies ranging from 0 Hz to $\frac{1}{2}F_B$ Hz in half, and so on. The lowest two bands have the same bandwidth of $2^{-N_B+1} \times F_B$ Hz when we split the fullband into N_B bands. Compared with the uniform band-split setting, this new setting can effectively divide the low-frequency region into more bands without significantly increasing the number of model parameters.

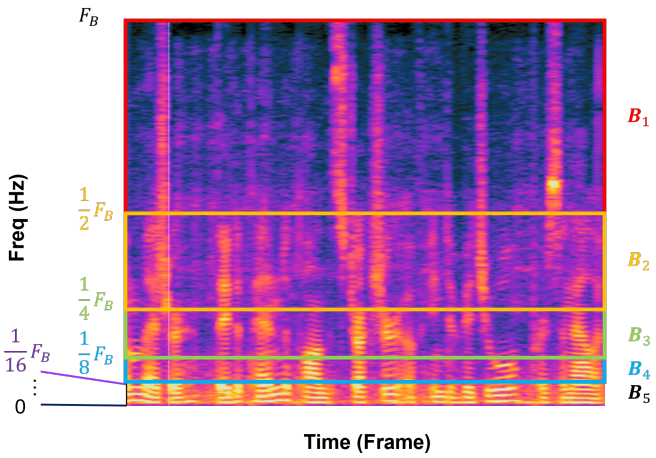


Fig. 4. Setting for splitting bands. The fullband is split in a way similar to the constant-Q transform.

III. Experimental Setup

A. Datasets

The proposed model was trained and evaluated on a subset of the Deep Noise Suppression Challenge (DNS Challenge) Interspeech 2021 dataset [18]. The clean speech dataset contains 562.72 hours of clips from 2150 speakers. The noise dataset includes 181 hours of 65302 clips from over 150 classes. Moreover, we randomly select and add room impulse responses from the openSLR26 and openSLR28 datasets [19] to 75% of the clean speech. Then, we mix clean or reverberant speech with noise at SNRs ranging from -5 to 20 dB to generate noisy speech. To evaluate models' performance, we used the public test set from DNS Challenge, which contains synthesized clips of two classes, namely with and without reverberations. It includes 150 noisy clips with SNRs ranging from 0 to 20 dB.

B. Training setup

We followed the settings of DNS Challenge for model training. The speech waveforms with 16 kHz sampling rate were converted into STFT spectrograms using the Hanning window with the window length of 32 ms and the frame shift of 16 ms. Adam optimizer was adopted with the learning rate of $1e^{-3}$ and the mean square error (MSE) and PReLU were used as the loss function and the activation function. The number of neighboring frequency bins on each side n was set to 15 and the number of bands N_B was set to 5. The number of frames of input and output sequences was set to 192 frames (approximately 3 s) in training. Compared models were evaluated by the common speech enhancement metrics WB-PESQ [20], NB-PESQ [21], STOI [22], and SI-SDR [23]. The following statements describe particular settings for the proposed model, inspired from the baseline Inter-SubNet model [16], and its variants.

Band-Split Inter-SubNet (prop): This is the basic version of the proposed model. The Band-Split SubInter module contains 93 and 307 hidden units in the first and the second SIL blocks, respectively. Both LSTMs in the first and the second SIL blocks are built with 384 hidden units.

Band-Split Inter-SubNet (lin): This is the proposed model with the uniform band-split setting in the linear-frequency axis. The performance of this variant is examined to see if the proposed constant-Q band-split setting offers benefits.

Band-Split Inter-SubNet (+): This is the plus version of the proposed model. In the basic version, the frequency band splitting is only applied in the first linear layer of the Band-Split SubInter module. In the plus version, the frequency band splitting is applied to all linear layers.

IV. Results and Discussions

As shown in Table I, the performance of the plus version is not significantly different from that of the basic version with $N_B = 5$. Therefore, we conclude that splitting the frequency band only in the first linear layer of the SubInter module is enough to achieve the optimal performance. Furthermore, we observe the proposed constant-Q band-split setting provides more benefits to performance than the uniform band-split setting under all conditions in terms of all metrics. The score differences under non-reverberant conditions are quite noticeable.

Comparison results between the baseline model and the proposed model indicate that band splitting is effective but not significantly so, suggesting that the influence of neighboring

TABLE I
DENOISING PERFORMANCE COMPARISON AMONG BAND-SPLIT MODELS

Model	Band-Split (Scale)	N_B	#Para (M)	With Reverb				Without Reverb			
				WB-PESQ	NB-PESQ	STOI	SI-SDR	WB-PESQ	NB-PESQ	STOI	SI-SDR
Noisy	-	-	-	1.822	2.753	86.62	9.03	1.582	2.454	91.52	9.07
Inter-SubNet [16]	-	-	2.29	3.207	3.659	93.98	16.76	2.997	3.504	96.61	18.05
Band-Split Inter-SubNet (prop)	log	2	2.41	3.216	3.656	93.90	16.64	2.995	3.497	96.53	17.98
Band-Split Inter-SubNet (prop)	log	3	2.54	3.223	3.651	93.97	16.66	3.005	3.501	96.63	18.10
Band-Split Inter-SubNet (prop)	log	4	2.66	3.221	**3.666**	93.92	16.77	2.987	3.495	96.54	17.99
Band-Split Inter-SubNet (prop)	log	5	2.78	**3.230**	3.665	94.03	16.80	**3.023**	**3.509**	**96.77**	18.23
Band-Split Inter-SubNet (prop)	log	6	2.90	3.229	3.659	93.95	16.71	3.005	3.502	96.71	18.19
Band-Split Inter-SubNet (lin)	linear	5	2.78	3.228	3.657	93.91	16.80	2.989	3.494	96.54	18.13
Band-Split Inter-SubNet (+)	log	5	3.75	3.222	3.662	**94.04**	**16.84**	2.995	3.506	96.72	**18.27**

frequencies on the center frequency bin does not vary greatly across different frequency ranges. In addition, the primary distinction between the uniform splitting and the proposed constant-Q splitting lies in the low-frequency range. Specifically, the uniform splitting produces the lowest frequency band from 0 to 1600 Hz, while the constant-Q splitting produces the band from 0 to 500 Hz. The superior performance from the constant-Q splitting suggests that lower-frequency bands have a more profound impact on model performance. The poorer performance observed with $N_B = 6$ compared to $N_B = 5$ may be attributed to overly meticulous division of the low-frequency range, leading to sub-optimal results.

Additionally, we conducted a detailed analysis on results of the proposed model with $N_B = 5$ across different SNRs. The box plots of four scores under reverberant and non-reverberant conditions are illustrated in Fig. 5 and Fig. 6, respectively. These box plots show that the proposed model makes an improvement from the baseline model in overall performance under both reverberant and non-reverberant conditions at low SNRs (< 5 dB). However, there is a trade-off, as scores of some metrics exhibit slight compromises at high SNRs. We speculate that by logarithmically splitting the frequency bands, the model can effectively learn the characteristics of each band, especially the low-frequency band, thereby enhancing its robustness in low SNR environments. However, this band-split approach also sacrifices some overall coherence across the full frequency band, which explains why the proposed model does not perform as well at high SNRs compared to the baseline

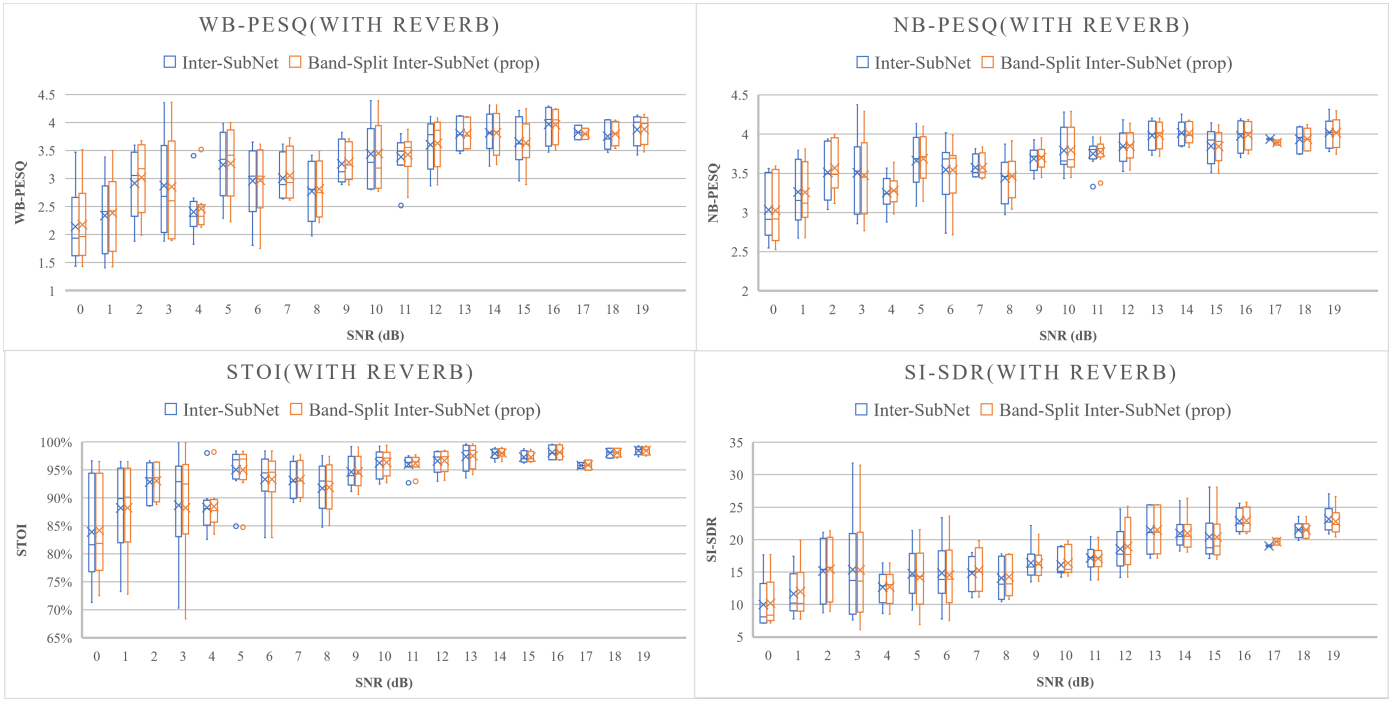


Fig. 5. Box plots of denoising scores of the proposed and baseline models across different SNRs under reverberant conditions.

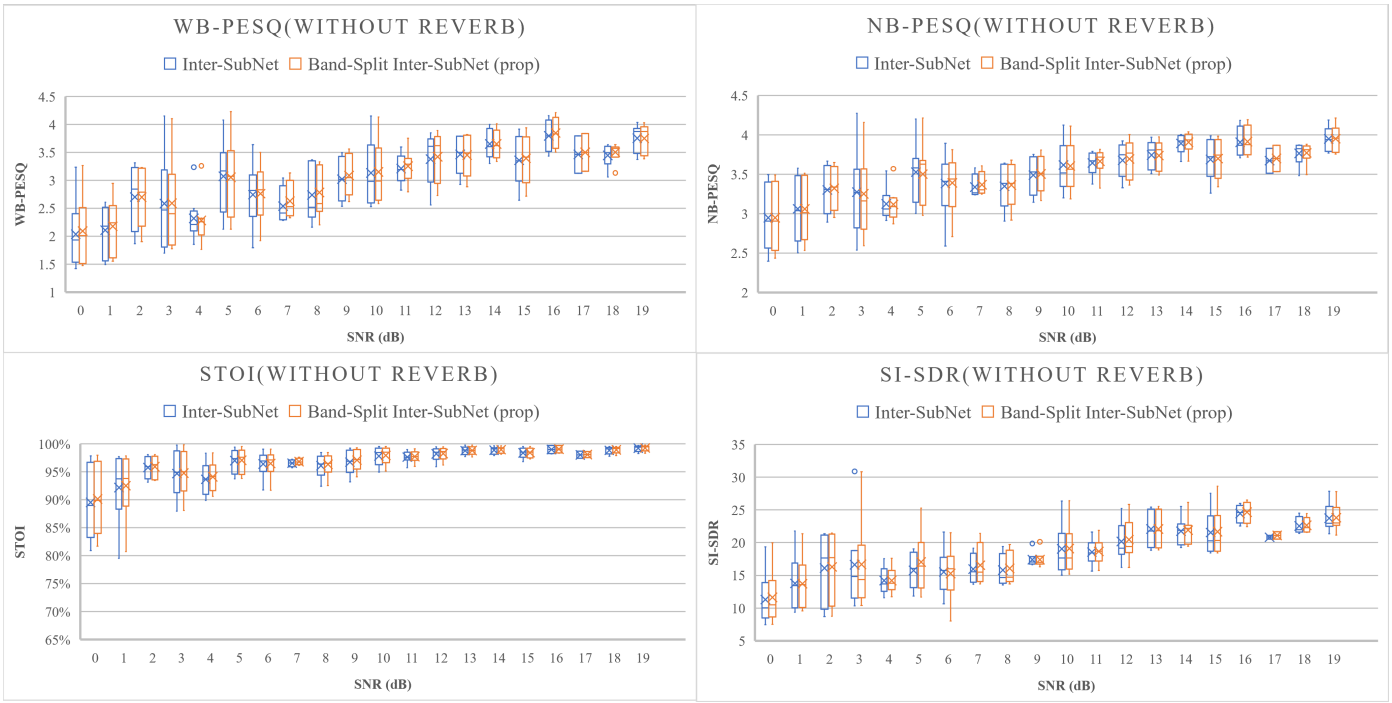


Fig. 6. Box plots of denoising scores of the proposed and baseline models across different SNRs under non-reverberant conditions.

model.

Finally, Table II shows the performance of the proposed model and other state-of-the-art models developed in recent years. Compared with other models, the proposed model shows superb performance in the denoising task with fewer model parameters. It is because the proposed model with the constant-Q band-split setting effectively highlights the distinctions between bands, resulting in superior model performance.

V. Conclusion and Future Work

In this paper, we propose a composite model named Band-Split Inter-SubNet. In addition, we adopt a new constant-Q band-split setting to mimic human auditory perception. Results demonstrate that the proposed model outperforms the original Inter-SubNet model. The new band-split setting enhances the robustness at low SNR conditions with slightly increased number of model parameters. The Band-Split SubInter module effectively learns the characteristics of each band and captures

TABLE II
Denoising performance comparison between the proposed model and other state-of-the-art models

Model	Year	#Para	With Reverb				Without Reverb			
		(M)	WB-PESQ	NB-PESQ	STOI	SI-SDR	WB-PESQ	NB-PESQ	STOI	SI-SDR
Noisy		-	1.822	2.753	86.62	9.03	1.582	2.454	91.52	9.07
DCCRN-E [24]	2020	3.7	-	3.077	-	-	-	3.27	-	-
Conv-TasNet [2]	2020	5.08	2.750	-	-	-	2.730	-	-	-
PoCoNet [25]	2021	50	2.832	-	-	-	2.748	-	-	-
DCCRN+ [8]	2021	3.3	-	3.300	-	-	-	3.330	-	-
TRU-Net [26]	2021	0.38	2.740	3.350	91.29	14.87	2.860	3.360	96.32	17.55
CTS-Net [27]	2021	4.99	3.020	3.470	92.70	15.58	2.940	3.420	96.66	17.99
FullSubNet [13]	2021	5.64	3.057	3.584	92.11	16.04	2.882	3.428	96.32	17.30
FullSubNet+ [14]	2022	8.67	3.177	3.648	93.64	16.44	3.002	3.503	96.67	18.00
FS-CANet [15]	2022	4.21	3.218	**3.665**	93.93	**16.82**	3.017	**3.513**	96.74	18.08
Inter-SubNet [16]	2023	2.29	3.207	3.659	93.98	16.76	2.997	3.504	96.61	18.05
Band-Split Inter-SubNet (prop)	2024	2.78	**3.230**	**3.665**	**94.03**	16.80	**3.023**	3.509	**96.77**	**18.23**

local patterns. Compared with other state-of-the-art models, the proposed model shows better performance on denoising under both reverberant and non-reverberant conditions. However, the proposed model demands high computations such that developing simpler modules and more efficient band-split settings still require further effort. In the future, we will explore new band-split settings and refine our modules to reduce model parameters and computations. Nevertheless, applying the model to real-time noise reduction remains a challenging task.

ACKNOWLEDGMENT

This work was supported in part by the Ministry of Science and Technology, Taiwan under Grant MOST 110-2221-E-A49-115-MY3, and in part by the Co-creation Platform of the Speech-AI Research Center, Industry-Academia Innovation School, NYCU, under the framework of the National Key Fields Industry-University Cooperation and Skilled Personnel Training Act, from the Ministry of Education (MOE), the National Development Fund (NDF), and industry partners in Taiwan.

REFERENCES

[1] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6629–6633.

[2] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[4] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[5] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 136–140.

[6] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[7] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.

[8] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-Wise Subband DCCRN with SNR Estimation for Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2816–2820.

[9] J. Li, D. Luo, Y. Liu, Y. Zhu, Z. Li, G. Cui, W. Tang, and W. Chen, "Densely connected multi-stage model with channel wise subband feature for real-time speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6638–6642.

[10] Y. Luo and J. Yu, "Music source separation with band-split rnn," *arXiv preprint arXiv:2209.15174*, 2022.

[11] J. Yu, H. Chen, Y. Luo, R. Gu, and C. Weng, "High Fidelity Speech Enhancement with Band-split RNN," in *Proc. Interspeech 2023*, 2023, pp. 2483–2487.

[12] X. Li and R. Horaud, "Online monaural speech enhancement using delayed subband lstm," *arXiv preprint arXiv:2005.05037*, 2020.

[13] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.

[14] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7857–7861.

[15] J. Chen, W. Rao, Z. Wang, Z. Wu, Y. Wang, T. Yu, S. Shang, and H. Meng, "Speech Enhancement with Fullband-Subband Cross-Attention Network," in *Proc. Interspeech 2022*, 2022, pp. 976–980.

[16] J. Chen, W. Rao, Z. Wang, J. Lin, Z. Wu, Y. Wang, S. Shang, and H. Meng, "Inter-subnet: Speech enhancement with subband interaction," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[17] Y. Wu and K. He, "Group normalization," 2018. [Online]. Available: https://arxiv.org/abs/1803.08494

[18] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *Proc. Interspeech 2021*, 2021, pp. 2796–2800.

[19] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[20] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[21] I. Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH–Geneva*, vol. 41, pp. 48–60, 2005.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[23] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 2472–2476.

[25] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy, "PoCoNet: Better Speech Enhancement with Frequency-Positional Embeddings, Semi-Supervised Conversational Data, and Biased Loss," in *Proc. Interspeech 2020*, 2020, pp. 2487–2491.

[26] H.-S. Choi, S. Park, J. H. Lee, H. Heo, D. Jeon, and K. Lee, "Real-time denoising and dereverberation wtih tiny recurrent u-net," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5789–5793.

[27] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.

Speech Synthesis from IPA Sequences through EMA Data

Koki Maruyama, Shun Sawada, Hidefumi Ohmura, Kouichi Katsurada
Tokyo University of Science, Noda, Chiba 278-8510, Japan
E-mail: 6323542@ed.tus.ac.jp, {sawada, katsurada}@rs.tus.ac.jp, ohmura@is.noda.tus.ac.jp

***Abstract*—Previous research on articulatory synthesis has partially implemented and verified the speech synthesis systems that mimic the human vocalization process. This study proposes a speech synthesis system that more faithfully replicates the human vocalization process by constructing a model that generates speeches from the international phonetic alphabet (IPA) sequence through articulatory movement data. For training and evaluating this model, we used the ATR phoneme-balanced 503 sentence electromagnetic articulography (EMA) database that contains pair data of speech and EMA data of a male Japanese speaker. The experimental results showed that using articulatory movement data as intermediate information improved the quality of synthesized speeches regarding objective scores such as mel-cepstral distortion (MCD), phoneme error rate (PER), and perceptual evaluation of speech quality (PESQ). In particular, the effectiveness of using articulatory movement data was confirmed in the case of speech where articulation points were included in the EMA database.**

1. Introduction

Human beings generate speech by converting linguistic information into movements of the articulatory organs such as the lip and tongue, thereby controlling the airflow. This fact shows that considering articulatory movements on speech synthesis has the potential to improve the quality of synthesized speeches. Although many text-to-speech (TTS) systems express the movements of articulatory organs with Mel-spectrogram, which illustrates the resonant frequency generated from the organs, Mel-spectrogram is the final result of articulation that can be directly converted to waveform with the vocoder. Modeling the movements of articulatory organs between text and Mel-spectrogram can more accurately simulate the human speech production process; thus, potentially improving the quality of the generated speeches.

Based on this background, there have been some studies regarding the relationships between articulatory movements, linguistic information, and speech. There are several types of articulatory movement data, with three notable examples: first, real-time magnetic resonance imaging (rtMRI), which measures the movements of the articulatory organs using an MRI device [1-3]. Although recording rtMRI data is costly, it is characterized by its ability to capture the movements of the entire articulatory organs. Second, electromagnetic articulography (EMA), which measures the movements of the articulatory organs using coils attached to them [4-7]. Although EMA data do not include data on vocal fold movements, it is characterized by its high sampling rate and spatial resolution. Third, ultrasound tongue imaging (UTI), which uses ultrasound to measure the movements of the tongue [8,9]. UTI captures only the movements of the tongue but it is characterized by its ease of measurement. These studies generated articulatory movement data from linguistic features or speech from articulatory movement data. Thus, these studies have only partially implemented and evaluated speech synthesis systems that mimic the human vocalization process.

In this study, we propose a model that synthesizes speech from linguistic features through articulatory movement data, using EMA data as the articulatory movement. We compared three models; (1) generating Mel-spectrogram from the international phonetic alphabet (IPA) sequence through estimated EMA, (2) directly generating Mel-spectrogram from the IPA sequence, and (3) generating Mel-spectrogram from both the IPA sequence and estimated EMA. We confirm the effectiveness of using articulatory movement data with some objective scores such as Mel-cepstral distortion (MCD) [10], phoneme error rate (PER), and perceptual evaluation of speech quality (PESQ) [11] to evaluate the quality of generated speeches. We also analyzed which position of articulation data was effective for speech generation and which phones benefited from the information on articulatory movement data.

2. Related Work

2.1 EMA data generation from the IPA

We have proposed a method for generating EMA data from the IPA [12]. IPA is a symbol that contains a wealth of articulatory information which is defined in terms of articulation method and position of articulation. Each phone in any language is represented as a symbol of the IPA. EMA [13,14] is an instrument that measures the articulatory movement using a coil attached to the articulatory organs, and is characterized by its ability to capture the movements of

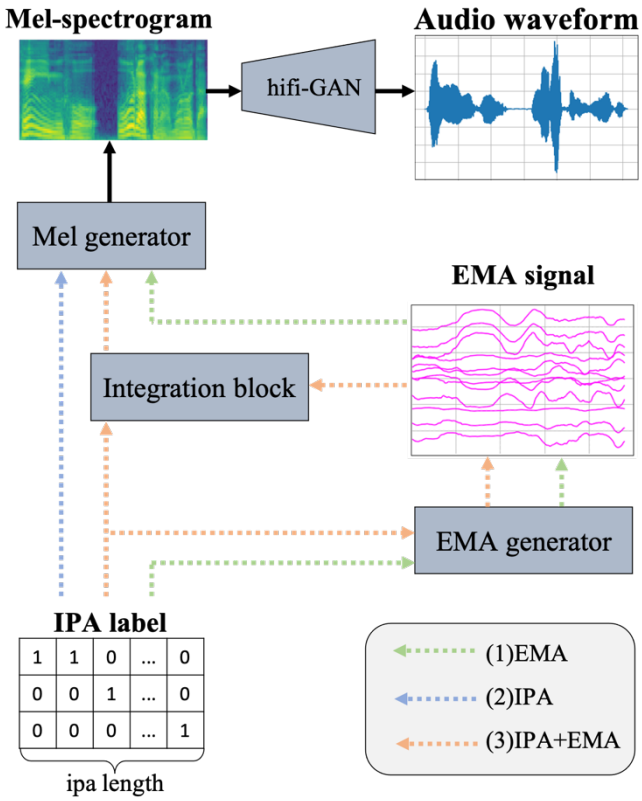


Fig. 1 Structure of speech synthesis model from the IPA sequence through EMA data.

articulatory organs at a high sampling rate and spatial resolution.

In [12], EMA data is generated from IPA using the 1DCNN-based SENet [15]. In this model, the receptive field, which is the input window width that affects the output, was optimized for generating EMA data from the IPA sequence. Additionally, dedicated models were constructed for each articulatory organ, and the final outputs were generated by combining the outputs of these multiple models as an ensemble model. This model enabled the generation of EMA data with high accuracy and speed.

2.2 Speech synthesis from EMA data

Kim et al. proposed a multi-speaker speech synthesis model from EMA data and speaker embedding through the Mel-spectrogram [16]. The proposed model consists of four modules; the first is an encoder consisting of the residual connections and CNN layers. The second is a generation model of F_0 and energy consisting of linear and CNN layers. The errors of F_0 and energy were used as the loss during training. The third is a model for generating the Mel-spectrogram, which consists of the Conformer [17]. The fourth is a vocoder that accepts Mel-spectrogram as input. This model uses the hifi-GAN [18] as a vocoder. The model showed high scores in both subjective and objective evaluations.

3. Proposed Models

The structure of the proposed model is shown in Fig. 1. To verify the effectiveness of using EMA data and the IPA information separately, we created three types of models: (1) EMA, which uses EMA data as intermediate information, (2) IPA, which generates speech directly from the IPA sequence, and (3) IPA+EMA, which uses both EMA and IPA sequence as the intermediate information and the input, respectively. For all models, hifi-GAN was used as the vocoder for generating speech from the Mel-spectrogram.

3.1 (1) EMA Model

The EMA Model first outputs EMA data from the IPA sequence and then generates a Mel-spectrogram from the output EMA data. IPA is considered as a suitable input for the EMA data generation model because it contains a wealth of articulatory information. In this model, the *EMA generator*, which generates EMA data from the IPA, consists of a 1DCNN-based SENet, and the *Mel generator*, which generates Mel-spectrogram from EMA data, is composed of the model proposed by Kim et al. [16], which learns losses including F_0.

3.2 (2) IPA Model

The IPA Model directly generates a Mel-spectrogram from the IPA sequence only. This model is the same as the *Mel generator* in the (1) EMA Model.

3.3 (3) IPA+EMA Model

The IPA+EMA Model first outputs EMA data from the IPA sequence and then generates a Mel-spectrogram from the output EMA data along with the IPA. The models that output EMA data from the IPA sequence and that generate Mel-spectrogram from the EMA data are the same as the *EMA generator* and *Mel generator* in the (1) EMA Model, respectively. The EMA data and IPA are combined in the *Integration block*, and this combined data is used as input to the *Mel generator*.

4. Experiment

In the experiment, we compared three models: (1) EMA, (2) IPA, and (3) IPA+EMA by using four evaluation metrics regarding the quality of the synthesized speeches.

4.1 Datasets

4.1.1 EMA database [19]

The EMA database contains IPA labeling, EMA, and speech data of the ATR phoneme-balanced 503 sentences [20] spoken by one male Japanese speaker. In the experiment, 450 sentences were used as training data, 27 sentences as validation data, and

Table 1 F_0RMSE, MCD, PER, and PESQ of synthesized speeches.

Model	F_0RMSE (Hz)	MCD (dB)	PER	PESQ
(1) EMA	39.34	13.36	0.26	1.09
(2) IPA	35.73	10.58	0.10	1.11
(3) IPA+EMA	**33.47**	**10.27**	**0.09**	**1.12**

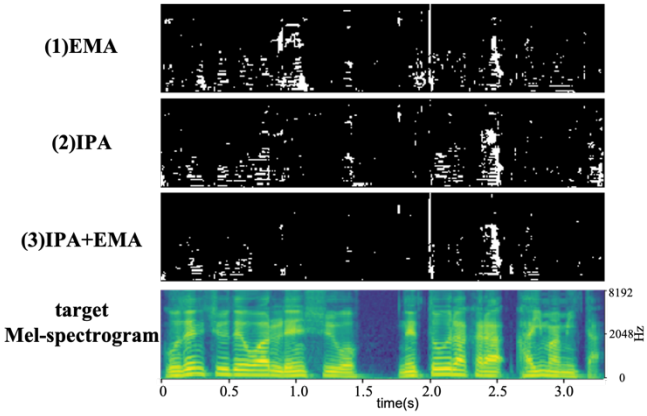


Fig. 2 Difference images of Mel-spectrograms between proposed models and ground truth.

26 sentences as test data, and were used for learning the *EMA generator* and *Mel generator* and fine-tuning the hifi-GAN.

Ninety-six IPA phones were included in the database and were preprocessed by converting them into one-hot vectors. EMA data consists of 12 dimensions of the back-forth and up-down movement data of six points on the mid-sagittal plane: the upper lip, lower lip, lower jaw, tongue tip, tongue body, and tongue dorsum. We preprocessed these data by normalizing them to range from -1 to 1. The speech data were transformed such that the Mel-spectrogram had 60 dimensions.

4.1.2 JVS corpus [21]

The JVS corpus contains speech data of ATR phoneme-balanced 503 sentences spoken by multiple Japanese speakers. In the experiment, only the data from male speakers were used for pre-training the hifi-GAN, with 5732 sentences as training data and 636 sentences as verification data. This is because the EMA database contains only the data from one male speaker. The speech data were transformed such that the Mel-spectrogram had 60 dimensions.

4.2 Evaluation metrics

As evaluation metrics, we used F_0 Root Mean Squared (RMSE) to measure the accuracy of F_0 estimation. Haverst [22] was used to extract F_0. Additionally, to evaluate the quality of the synthesized speech, we used MCD [10], PER and PESQ [11]. PER was calculated with the speech recognition

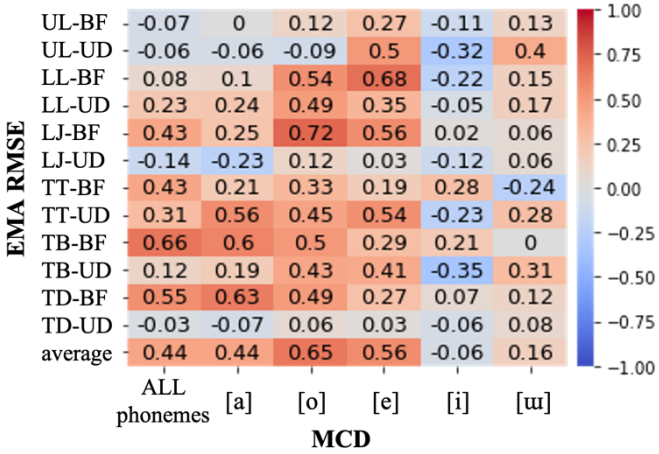


Fig. 3 Heatmap of Pearson correlation coefficients between the MCD and the RMSE of EMA data in the (3) IPA+EMA Model. "UL" means upper lip, "LL" means lower lip, "LJ" means lower jaw, "TT" means tongue tip, "TB" means tongue body, "TD" means tongue dorsum, "BF" means back-forth, and "UD" means up-down.

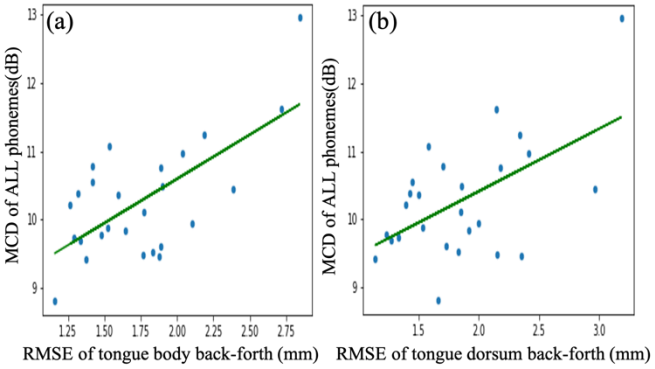


Fig. 4 Scatter plot (blue) and regression line (green).
(a) : "MCD of all phonemes" and "RMSE of tongue body back-forth position" in the EMA.
(b) : "MCD of all phonemes" and "RMSE of tongue dorsum back-forth position" in the EMA.

model Julius [23]. Lower values of F_0RMSE, MCD, and PER indicate better evaluation, whereas higher values of PESQ indicate better speech quality.

5. Results and Discussions

5.1 Overall results

Table 1 summarizes the overall experimental results. For all evaluation metrics, the (3) IPA+EMA Model showed the best results. The addition of EMA data as intermediate information enabled more accurate modeling of vocal tract characteristics, resulting in improved speech quality. However, F_0RMSE was 33.47 (Hz), indicating that synthesized speech cannot be considered natural. This is because EMA data do not contain information about the vocal folds, which are closely related to F_0.

Fig. 2 shows the difference images of Mel-spectrograms between the target speech and the synthesized speeches. White

Table 2 MCD for each phoneme category.

Model	MCD(dB)			
	vowel	consonant	semivowel	silence
(1) EMA	12.03	13.86	12.30	18.05
(2) IPA	9.85	10.68	**9.99**	11.63
(3) IPA+EMA	**9.69**	**10.37**	11.06	**11.07**

Table 3 MCD for each vowel.

Model	MCD(dB)				
	[a]	[o]	[e]	[i]	[ɯ]
(1) EMA	10.92	11.46	10.18	14.04	12.66
(2) IPA	9.97	9.79	8.81	**10.34**	**10.04**
(3) IPA+EMA	**9.55**	**9.50**	**8.52**	10.41	10.49

areas indicate where the Mel-spectrogram of the target speech was not accurately reproduced. Comparing the three models using the difference images, we can observe that the Mel-spectrogram of the speech synthesized from the (3) IPA+EMA Model exhibited the best reproducibility. However, none of the models accurately reproduced the target Mel-spectrogram around the 2.5 seconds. The sound at this point is [a], and the issue may be attributed to amount of articulatory movement; further analysis is required.

Fig. 3 shows the Pearson correlation coefficient between the RMSE of EMA data and the MCD in the (3) IPA+EMA Model. Focusing on the MCD of ALL phonemes, the correlation coefficient with the RMSE of the tongue body back-forth position (TB-BF) was the highest at 0.66, and the correlation coefficient with the RMSE of the tongue dorsum back-forth position (TD-BF) was the second highest at 0.55. Fig. 4 shows the scatter plot and regression line for the two pairs that recorded high correlation coefficients in Fig. 3. Fig. 4 (a) and (b) show the graphs which were depicted from the data of TB-BF and TD-BF, which showed the two best performances. The scatter plot and correlation coefficient indicate that the higher the accuracy of the estimation tongue body or dorsum back-forth position in the EMA, the better the MCD values for all phonemes. This is because MCD is an index of vocal tract characteristics, and the back-forward movement of the tongue has a significant effect on the shape of the vocal tract.

5.2 MCD for each phoneme category

Table 2 compares three models by the MCD of vowels, consonants, semivowels, and silence. The MCD of vowels, consonants, and silence was the best with the (3) IPA+EMA Model. However, that of semivowels was the best with the (2) IPA Model. The test data included only the semivowels [j] (voiced palatal approximant) and [w] (voiced labial–velar approximant), both of which are articulated at the palate as the point of articulation. However, because the EMA data do not contain information about the palate, the EMA data used as intermediate information was considered to be noise, leading to this result.

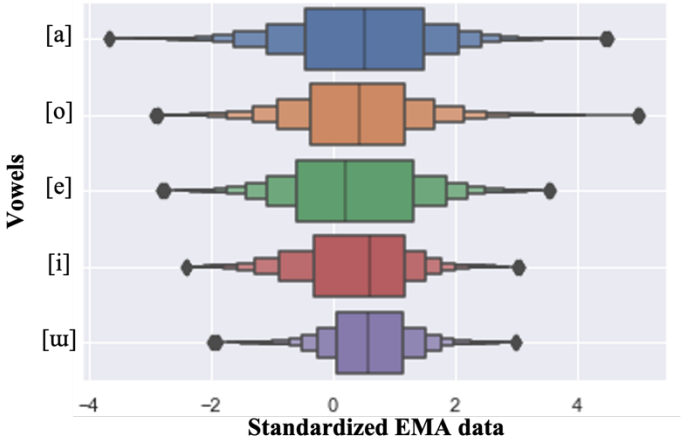


Fig. 5 Boxenplot of standardized EMA data for each vowel in test data.

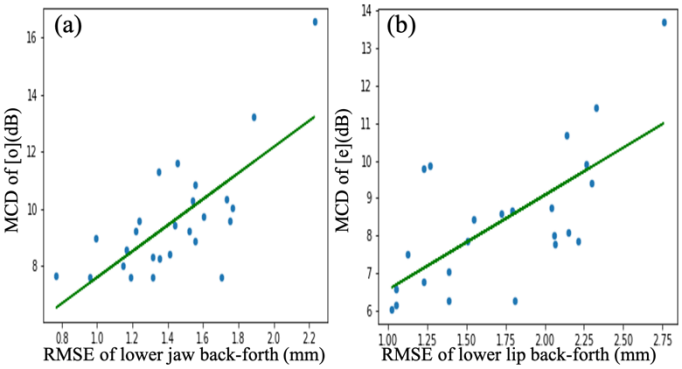


Fig. 6 Scatter plot (blue) and regression line (green).
(a): "MCD of [o]" and "RMSE of the lower jaw back-forth position" in the EMA.
(b) : "MCD of [e]" and "RMSE of the lower lip back-forth position" in the EMA

5.3 MCD for each vowel

Table 3 compares the three models using the MCD of single vowels. The MCD of [a], [o], and [e] was the best with the (3) IPA+EMA Model. These three vowel sounds were either open or close-mid vowels. Fig. 5 shows boxenplots that display the standardized EMA data categorized by vowels in the test data. The EMA data was standardized for each measurement point and converted from 12 dimensions to one dimension. From these plots, it can be observed that the data distribution is scattered in the order [a], [o], [e], [i], [ɯ]. That is, the articulatory movements are more intense in [a], [o], and [e] compared to [i] and [ɯ]. Therefore, three open or close-mid vowel sounds had a certain amount of articulatory movement, and the accuracy of MCD was improved by using the EMA data as intermediate information. However, the MCDs of [i] and [ɯ]

Table 4 MCD of consonants categorized by articulation points

Model	MCD(dB)						
	Labiodental	Palatal	Bilabial	Alveolar	Velar	Retroflex	Glottal
(1) EMA	17.59	13.78	14.11	12.60	14.85	15.37	17.58
(2) IPA	12.65	10.20	10.68	10.38	11.68	**11.10**	**10.83**
(3) IPA+EMA	**11.24**	**9.26**	**10.08**	**10.08**	**11.65**	11.31	11.17

were the best with the (2) IPA Model. These two vowel sounds were close vowels, and Fig. 5 indicates that close vowels have less articulatory movement. Thus, the MCDs in these two vowel sounds were not improved by using the EMA data as intermediate information.

Focusing on the MCD for each vowel in Fig. 3, the correlation coefficient between the MCD and RMSE of the lower jaw back-forth position of [o] ([o]-LJ-BF) was the highest at 0.72, and the correlation coefficient between the MCD and RMSE of the lower lip back-forth position of [e] ([e]-LL-BF) was the second highest at 0.68. Fig. 6 (a) and (b) show the scatter plots and the regression lines which were depicted from data [o]-LJ-BF and [e]-LL-BF, respectively. The results indicate that the higher the accuracy of the estimation lower jaw or lower lip back-forth position in the EMA, the better the MCD values of [o] or [e]. The vowel [o] is a close-mid back rounded vowel, pronounced with lip rounding. Because the lower jaw moves to create lip rounding, a high correlation was obtained. Fig. 7 shows boxenplots that display the standardized EMA data of [e] categorized by measurement points of EMA in the test data. Focusing on the mean values, LL-BF, UL-BF, and LJ-BF in this order, were the farthest from the mean value of zero. In other words, compared to other sounds except [e], the articulatory organs of these EMA points moved in unique positions. Therefore, it is considered that a high correlation coefficient between the MCD and [e]-LL-BF was observed. Additionally, [e]-UL-BF and [e]-LJ-BF also showed relatively high correlation coefficients when compared to other vowels.

However, the correlation coefficients between the MCD of close vowels [i][ɯ] and the RMSE of EMA data were lower compared to those of open or close-mid vowels. This indicates that no correlation between the estimation accuracy of the EMA data and MCD of close vowels existed. This may also be owing to the less articulatory movement of close vowels.

5.4 MCD for consonant categories

Table 4 compares three models using the MCD of each consonant articulation point. The MCD of labiodental, palatal, bilabial, alveolar, and velar sounds was the best with the (3) IPA+EMA Model. These sounds involved the lips or tongue as the place of articulation, which could be supplemented with EMA data, resulting in improved MCD.

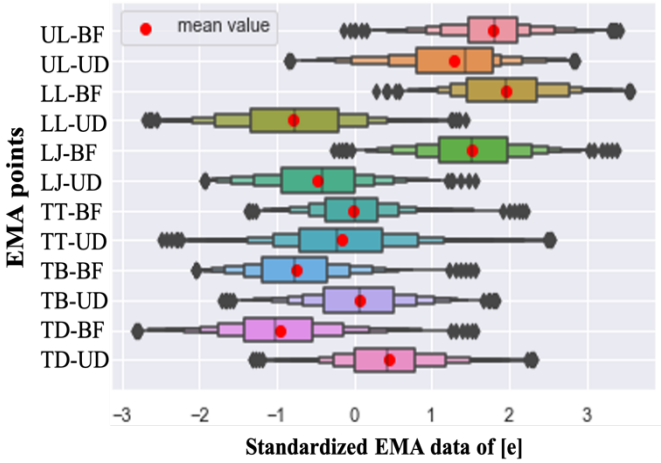


Fig. 7 Boxenplot of standardized [e]-EMA data for each EMA point of measurement in test data.

However, the MCD of the retroflex and glottal sounds was the best with the (2) IPA Model. Retroflex sounds are pronounced when the curled tongue approaches the hard palate. The EMA data do not contain information about the hard palate and the tongue movements are complex. Thus, the accuracy of MCD in the sounds was not improved by using the EMA data as intermediate information. Additionally, EMA data do not contain information about the glottis which is the articulation point of the glottal sound [h]; hence, the (2) IPA Model had the best results in the MCD of glottal sounds.

6. Conclusions and Future Work

In this study, we evaluated three models for generating speeches from IPA sequences through EMA data, which contains a wealth of information about articulatory movement. Experiments confirmed that using EMA data as intermediate information improves the quality of the synthesized speech. In particular, the effectiveness of using EMA data as intermediate information was confirmed in the case of speech where articulation points were included in the EMA data.

However, even when using EMA data, the estimation accuracy of F_0 was low. Therefore, in the future, we will investigate methods to accurately estimate F_0 using EMA data as input.

7. ACKNOWLEDGMENT

This work was supported by the JSPS KAKENHI Grant Numbers JP22K12100 (Kouichi Katsurada), and JP24K00071 (Kikuo Maekawa).

REFERENCES

[1] R. Tanji, H. Ohmura, and K. Katsurada, "Using Transposed Convolution for Articulatory-to-Acoustic Conversion from Real-Time MRI Data," *INTERSPEECH*, pp. 3176-3180, 2021.

[2] S. Udupa, and PK. Ghosh, "Real-Time MRI Video synthesis from time aligned phonemes with sequence-to-sequence networks," *ICASSP*, pp. 1-5, 2023.

[3] Y. Otani, S. Sawada, H. Ohmura, and K. Katsurada, "Speech Synthesis from Articulatory Movements Recorded by Real-time MRI," *INTERSPEECH*, pp. 127-131, 2023.

[4] Z. Wei, Z. Wu, and L. Xie, "Predicting articulatory movement from text using deep architecture with stacked bottleneck features," *ICASSP*, pp. 1-6, 2016.

[5] ZC. Liu, ZH. Ling, and LR. Dai, "Articulatory-to-Acoustic Conversion with Cascaded Prediction of Spectral and Excitation Features Using Neural Networks," *INTERSPEECH*, pp. 1502-1506, 2016.

[6] K. Katsurada, and K. Richmond, "Speaker-independent mel-cepstrum estimation from articulator movements using D-vector input," *INTERSPEECH*, ISCA, pp. 3176-3180, 2020.

[7] YW. Chen, KH. Hung, SY. Chuang, and J. Sherman, et al., "EMA2S: An End-to-End Multimodal Articulatory-to-Speech System," *ISCAS*, pp. 1-5, 2021.

[8] N. Kimura, M. Kono, and J. Rekimoto, "SottoVoce: An ultrasound imaging-based silent speech interaction using deep neural networks," *the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-11, 2019.

[9] G. Gosztolya, T. Grósz, T. Tóth, and L. Markó, et al., "Applying DNN adaptation to reduce the session dependency of ultrasound tongue imaging-based silent speech interfaces," *Acta Polytechnica Hungarica*, 17(7), 109-124, 2020.

[10] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *ICASSP*, 1993, pp. 125–128.

[11] AW. Rix, JG. Beerends, MP. Hollier, and AP. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *ICASSP*, pp. 749–752, 2001.

[12] K. Maruyama, S. Sawada, H. Ohmura, and K. Katsurada, "Generation of Articulatory Movement Data from IPA using 1DCNN," *ASJ*, 3-Q-38, September 2023.

[13] PW. Schönle, K. Gräbe, P. Wenig, and J. Höhne, et al., "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, pp. 26–35, 1987.

[14] H. Horn, G. Göz, M. Bacher, and M. Müllauer, et al., "Reliability of electromagnetic articulography recording during speaking sequences," *Eur. J. Orthod,* pp. 647–655, 1997.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CVPR*, pp. 7132-7141, 2018.

[16] M. Kim, Z. Piao, J. Lee, and HG. Kang, "Style Modeling for Multi-Speaker Articulation-to-Speech," *ICASSP*, pp. 1-5, 2023.

[17] A. Gulati, J. Qin, CC. Chiu, and N. Parmar, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv*:2005.08100, 2020.

[18] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *NeurIPS*, pp. 17022-17033, 2020.

[19] K. Wakamiya, F. Taguchi, R. Watanabe, and K. Katsurada, et al., "Collection of large-scale Japanese articulatory-acoustic parallel data," *IEICE Technical Report*, pp. 7-12, June 2019.

[20] A. Kurematsu, K. Takeda, Y. Sagisaka, and S. Katagiri, et al., "ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis," *Speech Commun*, vol. 9, no. 4, pp. 357–363, 1990.

[21] S. Takamichi, K. Mitsui, Y. Saito, and T. Koriyama, et al., "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint*, 1908.06248, Aug 2019.

[22] M. Morise, "Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals," *INTERSPEECH*, pp. 2321-2325, 2017.

[23] A. Lee, and T. Kawahara, "Recent development of open-source speech recognition engine Julius," *APSIPA*, pp. 131-137, 2009.

A Survey on Objective Quality Assessment of Omnidirectional Images

Xiangjie Sui†, Shiqi Wang*, and Yuming Fang‡
† City University of Macau, Macau
E-mail: xjsui@cityu.edu.mo
* City University of Hong Kong, Hong Kong
E-mail: shiqwang@cityu.edu.hk
‡ Jiangxi University of Finance and Economics, China
E-mail: fa0001ng@e.ntu.edu.sg

***Abstract*—Omnidirectional (360°) image quality assessment (OIQA) has become crucial with the rising popularity of virtual reality (VR). Despite significant efforts to develop objective OIQA methods, there is a lack of comprehensive reviews of these methods, which hinders in-depth understanding and analysis. This paper presents an extensive survey of objective OIQA methods, classifying them into three principal types: 2D-plane-based, sphere-based, and viewport-based methods. We then analyze the limitations of current methods, highlighting challenges such as insufficient consideration of viewing conditions and diverse viewing behaviors. Moreover, this paper suggests potential future research directions, such as multi-modality approaches and personalized assessments. By systematically reviewing existing methodologies and proposing potential advancements, this paper aims to guide future research towards more accurate and efficient OIQA solutions.**

I. Introduction

The advent of virtual reality (VR) technology, particularly propelled by advancements in smart wearable devices such as the Apple Vision Pro, has significantly increased attention on 360° images. These images, created by capturing and stitching scenes from multiple fisheye lenses, form a complete $360° \times 180°$ view. When exploring a scene via head-mounted displays (HMDs), 360° images offer an immersive experience that closely mimics physical presence in the scene. Compared to traditional 2D images, 360° images provide enhanced realism, interactivity, and immersion, steering digital media toward higher quality and richer information content.

The processing pipeline for 360° images encompasses several stages, including image capture, stitching and projection, encoding, transmission, decoding, and rendering [1]. Throughout these stages, 360° images may suffer significant quality degradation due to various factors, making it challenging to maintain high-quality outputs. For instance, the visual quality of 360° images is often affected by the accuracy of image registration, fusion, and projection algorithms. Thus, researching effective methods to automatically provide quantitative quality indicators for 360° images is essential for optimizing the processing pipeline and ensuring high-quality outputs.

Omnidirectional image quality assessment (OIQA) can be broadly divided into subjective and objective methods. Subjective quality assessment involves human subjects who evaluate the visual quality in controlled experimental environments. In contrast, objective quality assessment aims to develop computational models that automatically infer visual quality by simulating the human visual system [2]–[8], viewing behaviors [9]–[12], natural statistical features [13], [14]. Despite significant efforts to develop objective OIQA methods, there is a lack of comprehensive reviews categorizing these methods and analyzing their limitations and future research directions [15]. This paper addresses this gap by reviewing objective OIQA methods from the perspective of common formats of 360° images: 2D projection planes, spheres, and viewports. Moreover, we discuss the limitations of current OIQA methods and suggest potential future research directions to guide future research towards more accurate and efficient OIQA solutions.

II. Objective Quality Assessment of 360° Images

360° images can be represented in several ways, primarily as 2D projection planes, spheres, and viewports (see Fig. 1). Typically, 360° images are stored in a 2D plane format achieved via equirectangular projection. During viewing, these images are decoded and reprojected onto a spherical surface for rendering and display. The viewport, representing the visual content observed by the viewer at a given moment, can be extracted using rectilinear projection. Based on these representations, current OIQA methods are categorized into 2D-plane-based, sphere-based, and viewport-based methods. This section provides a detailed introduction to these categories.

A. 2D-Plane-Based Methods

In the 2D projection plane, 360° images can be processed similarly to traditional 2D images, which allows for the extension of 2D quality metrics to OIQA. However, different map projections introduce distinct problems. For instance, equirectangular projection causes significant shape distortions near the poles (see Fig. 1). Thus, the fundamental aim of 2D-plane-based methods is to address the non-uniform sampling

TABLE I
SUMMARY OF OIQA METHODS

Type	Model	Heuristic	Data-driven	Scanpath	Viewing conditions	Weighting allocation	Re-sampling
2D-Plane-Based	[1]CPP-PSNR [16]	✓					✓
	[2]WS-PSNR [17]	✓				✓	
	WS-SSIM [18]	✓				✓	
	DeepVR-IQA [19]		✓			✓	
	[3]SAP-Net [2]		✓				
	Liu23 [20]	✓					
Shpere-Based	[4]S-PSNR [21]	✓					✓
	S-SSIM [6]	✓				✓	✓
	Sendjasni23 [3]		✓			✓	✓
Viewport-Based	[5]MC360IQA [5]		✓				✓
	[6]VGCN [4]		✓				✓
	[7]Sui21 [9]	✓		✓	✓	✓	✓
	Zhou21 [8]		✓				✓
	MFILGN [13]	✓					✓
	MP-BOIQA [14]	✓					✓
	Fang22 [22]		✓		✓		✓
	Zhang22 [7]		✓				✓
	TVFormer [10]		✓	✓		✓	✓
	PW-360IQA [23]		✓	✓			✓
	[8]Assessor360 [11]		✓	✓	✓		✓
	[9]GSR-X [12]		✓	✓	✓		✓
	Liu24 [24]		✓	✓	✓		✓

Open source (retrieved in July 2024):

[1] https://github.com/Samsung/360tools [2] https://github.com/Rouen007/WS-PSNR [3] https://github.com/yanglixiaoshen/SAP-Net [4] https://github.com/Samsung/360tools
[5] https://github.com/sunwei925/MC360IQA [6] https://github.com/weizhou-geek/VGCN-PyTorch [7] https://github.com/xiangjieSui/img2video
[8] https://github.com/TianheWu/Assessor360 [9] https://github.com/xiangjieSui/GSR

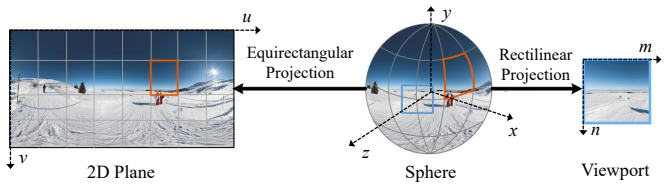


Fig. 1. An illustration of different projection of a 360° image.

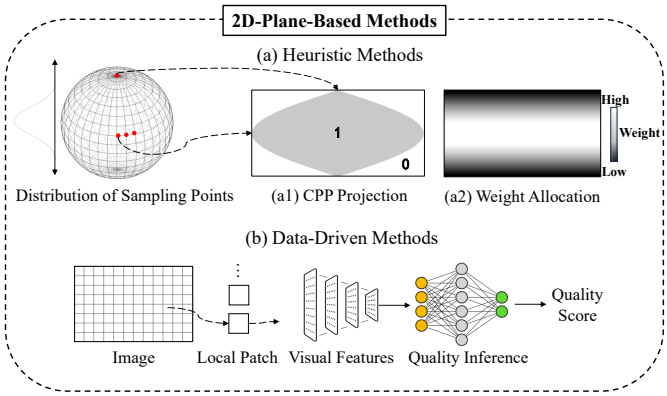


Fig. 2. Basic strategies of 2D-plane-based OIQA methods.

resulting from sphere-to-plane projection. These methods can be further divided into heuristic and data-driven approaches (see Fig. 2).

1) Heuristic Methods: Heuristic methods typically employ weighting matrices to model the density of sampling points on the spherical domain. Specifically, given an omnidirectional image $\mathbf{I}^{h\times w}$, a 2D weighting matrix $\mathbf{W}^{h\times w}$ is used to compensate for non-uniform sampling in quality inference: $\hat{Q} = \sum_{(x,y)\in\mathbf{I}} \hat{q}(x,y)\mathbf{W}(x,y)$, where $\hat{q}(x,y)$ represents the predicted quality score of the image pixel located at (x,y), and $\hat{Q}$ is the predicted overall score of the omnidirectional image. CPP-PSNR [16] employed Craster Parabolic Projection (CPP) to guarantee uniform sampling density. Given an initial weighting matrix $\mathbf{W} = 0$, the CPP projection maps uniformly distributed sample points (ϕ,θ) on the sphere to the corresponding 2D plane coordinates (x,y) by $x = \sqrt{\frac{3}{\pi}}\theta(2\cos\frac{2\phi}{3} - 1)$ and $y = \sqrt{3\pi}\sin\frac{\phi}{3}$. Then, by setting $\mathbf{W}(x,y) = 1$, the point (x,y) becomes a valid sample for Peak Signal-to-Noise Ratio (PSNR) calculation. Fig. 2 (a1) shows the resulting weighting matrix, where the number of sampling points decreases from the center to the sides, alleviating the oversampling problem near the poles. Similarly, WS-PSNR [17] and WS-SSIM [18] assigned weights to all pixels of the 360° image. Taking the equirectangular projection as an example, the value of $\mathbf{W}(x,y)$ can be determined by the corresponding stretch

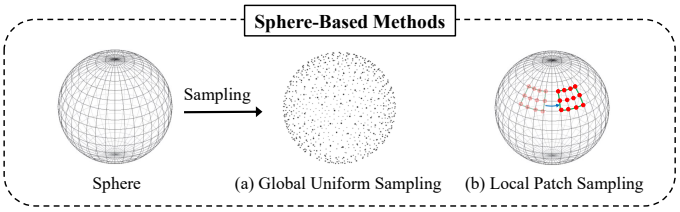


Fig. 3. Basic strategies of sphere-based OIQA methods.

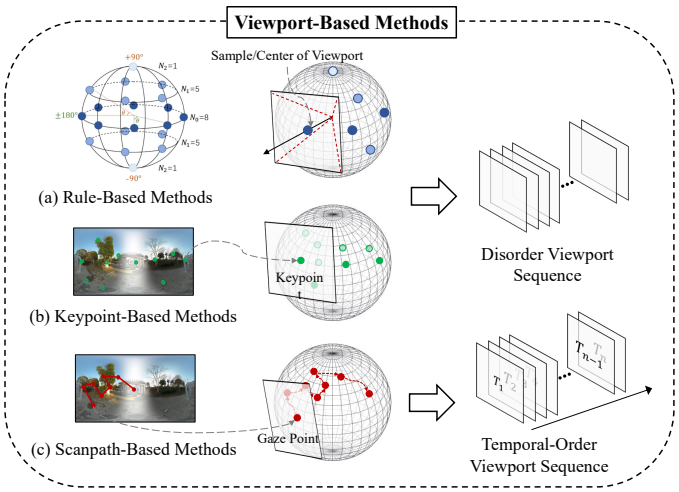


Fig. 4. Basic strategies of viewport-based OIQA methods.

ratio: $\cos\frac{y+0.5-\frac{h}{2}}{h}$, where h represents the height of the 360° image. Fig. 2 (a2) shows the resulting weighting matrix, where we can observe that the weight values decrease from the center to the sides. By considering that quality degradation leads to structural changes on the 2D projection plane, Liu *et al.* [20] proposed using histogram features to measure structural degradation and combining them with statistical and saliency features to compute the quality of 360° images. However, they did not consider the non-uniform sampling issue on the 2D projection plane.

2) Data-Driven Methods: Data-driven methods leverage deep learning to infer local image patch quality, which is then aggregated to predict the overall quality of the 360° image [2], [25]. More specifically, the 360° image is first divided into N image patches, which are then input into a pre-trained network to extract visual features $f_{1:N}$. Then, these features are fed into the quality inference module $\mathcal{Q}$ to obtain local image patch quality scores. Finally, a fusion strategy $\mathcal{W}$ is designed to integrate the local quality scores to obtain a global quality score. Overall, the goal of these methods is to maximize the prediction accuracy of the network, which could be expressed by: $\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} p(Q|\mathcal{W}(\mathcal{Q}(f_1), \mathcal{Q}(f_2), ..., \mathcal{Q}(f_N)); \boldsymbol{\alpha})$, where Q represents the ground-truth quality label, and $\boldsymbol{\alpha}$ denotes the learnable parameters of the network.

DeepVR-IQA [25] used ResNet-50 [26] as the visual feature extraction network and several fully connected layers to construct the quality inference module. Moreover, a fusion strategy was proposed to learn the weight of each image patch by encoding its coordinates. The final overall quality score of the 360° image was obtained by weighted averaging all local quality scores. SAPNet [2] included a self-supervised image enhancement module [27], which leveraged discrete wavelet transform features to enhance the quality of image patches. Then, the enhanced patches were regarded as "reference images" for quality inference. Finally, the overall quality of the 360° image was obtained by directly averaging the quality scores of all local image patches.

B. Sphere-Based Methods

Sphere-based methods calculate local quality estimates directly on the sphere domain, employing either global uniform sampling or local patch sampling strategies (see Fig. 3).

1) Global Uniform Sampling Methods: Global uniform sampling methods, such as S-PSNR [21], extracted a large number (*i.e.*, $655,362$) of predefined uniformly distributed sample points on the sphere. The corresponding 2D plane coordinates were computed to retrieve pixel values from distorted and reference images. Then, PSNR score was calculated to obtain the overall quality score of the 360° image.

2) Local Patch Sampling Methods: By considering that image patches contain structural information for quality inference, local patch sampling methods aggregate quality scores of local image patches (on sphere) to infer the overall quality of the 360° image. For instance, S-SSIM [6] calculated the Structural Similarity Index Measure (SSIM) [28] scores of image patches with a circularly symmetric Gaussian weighting function $g = \left\{g_i | \sum_{i=1}^{11\times11} g_i = 1\right\}$. These scores were weighted pooling based on coordinate stretch ratios to obtain an overall quality score. Sendjasni *et. al* [3] employed Convolutional Neural Networks (CNNs) to learn the quality-aware features of local image patches on the sphere and adaptively assigned weights to different local image patches based on the visual saliency map of the 360° image, giving higher weights to the salient regions in quality fusion.

C. Viewport-Based Methods

Given the limited field of view (FoV) in human perception, focusing only on the quality of the entire images might not be consistent with human perception. To address this problem, viewport-based methods assess the perceptual quality of 360° images by combining quality scores of viewport images (see Fig. 4). These methods are categorized into rule-based, keypoint-based, and scanpath-based approaches based on their viewport sampling strategies. Rule-based and keypoint-based methods commonly focus on learning spatial quality features of viewport images. While scanpath-based methods aim to model the spatio-temporal quality of 360° images by simulating the dynamic viewing experience of users.

1) Rule-Based Methods: Rule-based methods design sampling strategies based on characteristics of 360° images. For example, Fang22 [22] and MC360IQA [5] sampled viewports at equal intervals, with the number of samples decreasing

from the equator to the poles to maintain a relatively uniform sampling density (see Fig. 4 (a)). Additionally, MC360IQA rotated the 360° image around the y-axis at equal intervals and repeated the sampling process to augment training and testing data. In quality inference, both Fang22 [22] and MC360IQA [5] used ResNet [26] to extract pre-trained features of viewport images, and employed fully connected layers to regress these features to quality scores. Notably, Fang22 [22] encoded the coordinates of the viewport center as learnable weights to adaptively fuse the quality scores of different viewports. Several studies [8], [13], [14] used cubemap projection to obtain six minimally overlapping viewport of the 360° image. Zhou *et al.* [8] designed a multi-task learning-based OIQA quality inference network with quality score prediction as the main task and distortion type classification as an auxiliary task. MFILGN [13] and MP-BIQA [14] extracted statistical features of viewport images, and used support vector regression[13] or random forest [14] to map the statistical features to quality scores.

2) Keypoint-Based Methods: Keypoint-based methods commonly extract keypoints on the 2D projection plane, and then extract viewport centered on these points for quality inference, as illustrated in Fig. 4 (b). VGCN [4] used the Speeded Up Robust Features (SURF) algorithm [29] to extract 20 key points for viewport extraction. A graph convolutional network was constructed to learn the spatial relationships between viewports to adaptively fuse the quality scores of different viewport images. Similarly, Zhang *et al.* [7] used the Oriented FAST and Rotated BRIEF (ORB) algorithm [30] to extract keypoints for viewport extraction. By considering the potential prediction bias cased by relying on local viewports only, the studies [4], [7] measured the global quality of 360° images by measuring the quality of 2D projection image [4] or the quality of a set of viewport extracted along hypothetical scanpaths [7]. The final quality score of a 360° image was calculated by combining the local and global quality scores [4], [7].

3) Scanpath-Based Methods: Scanpath-based OIQA methods aim to predict the perceived quality 360° images by learning the spatio-temporal quality features of the viewport sequences extracted along scanpaths. The studies [9], [24] proposed OIQA methods based on human scanpath, the quality of viewport sequences were measured by mature 2D metrics [9] or the sequence model [24]. TVFormer [10] consisted of a scanpath prediction network and a quality inference network. The scanpath prediction network was based on the ViT [31] architecture, which included a memory unit modeling the memory mechanism of human. The quality inference network included the global and local branches, where the global branch learned quality features from the 2D projection plane, and the local branch learned spatio-temporal quality features of viewport sequences. Assessor360 [11] included a scanpath prediction strategy based on the entropy of viewport images. This was inspired by that human visual attention tends to focus on scenes with higher information content. In quality inference, a CNN-based module was proposed to learn the multi-scale spatial features of viewport images. Then, these features were fed into a temporal modeling module to adaptively fuse the spatial quality-aware features. The study [12] conducted a unique generative scanpath representation (GSR), which aggregates gaze-focused patches of different hypothesis users at each moment [32]. This representation provided a comprehensive global overview of dynamic perceptual experiences of multi-hypothesis users. For quality inference, the video backbone [33] was employed to learn the spatio-temporal features of GSR sequences for quality inference.

III. Limitations and Future Directions

A. Limitations

Despite notable advancements in the field of OIQA, several critical limitations persist. These limitations hinder the comprehensive application of OIQA methods, necessitating further research to address these challenges.

- **Overlook of viewing conditions**. Current OIQA methods commonly neglect the impact of varying viewing conditions, which significantly influence user scanpath patterns and perceptual quality [9], [34]. Although several studies [9], [11], [12], [22], [24] have considered factors such as the starting point of viewing and exploration time, other critical aspects remain under explored. For instance, the resolution, FoV, and display constraints of HMDs can significantly affect the perceptual quality of 360° images [9]. Therefore, a more comprehensive consideration of these viewing conditions is essential for developing robust OIQA models.
- **Neglect of diverse viewing behaviors**. Most OIQA methods rely on a single, fixed viewport sequence to predict the perceptual quality of 360° images [4], [5], [7], [8], [10], [13], [14]. Such a deterministic approach fails to account for the probabilistic nature of human viewing behaviors, which exhibit considerable variability and randomness. As a result, these methods may introduce prediction bias and fail to accurately reflect the perceptual quality experienced by users. Incorporating models that simulate diverse and dynamic viewing behaviors is crucial for enhancing the reliability of OIQA methods.
- **High computational complexity**. The high computational complexity of current OIQA methods poses significant challenges for real-time applications. One of the reasons is that the process of extracting viewports is often cumbersome and time-consuming. For example, methods such as MC360IQA [5] require the generation of a substantial number of viewports (*e.g.*, $1,080$) for a single 360° image before performing quality inference. Similarly, scanpath-based methods [9], [24] demand extensive viewport extraction. Streamlining these processes is vital for the practical deployment of OIQA methods in real-time scenarios.
- **Lack of large-scale datasets**. There is a notable scarcity of large-scale annotated datasets for training and evaluating OIQA models. To our best knowledge, the largest

available OIQA dataset [22] comprises only 258 reference images and 1032 distorted images. Insufficient training and testing data might limit the generalization ability of these methods. For example, the performance of OIQA models, especially those based on natural statistical features, is heavily dependent on the diversity and scale of the training data. Expanding the availability of large-scale and diverse datasets is critical for advancing the field.

B. Promising Directions

In addition to addressing the above limitations, several promising directions can be pursued:

- **Multi-modality OIQA methods**. Developing multi-modality OIQA methods that integrate various modalities, such as visual, text (*e.g.*, information of viewing conditions), auditory, and haptic feedback, can provide a more holistic evaluation of 360° image quality. By leveraging complementary information from different modalities, these methods can enhance the robustness and accuracy of quality assessments, offering a deeper insight of the user experience.
- **Personalized OIQA methods**. Personalized OIQA methods that account for individual user preferences and viewing behaviors can significantly improve the relevance and applicability of quality metrics. By tailoring the evaluation process to reflect the subjective quality experiences of different users, these methods can provide more accurate and user-specific assessments. This personalization can be achieved through adaptive models that learn from user interactions and feedback.
- **Quality assessment of generated 360° images**. The increasing use of AI techniques to generate 360° images presents unique challenges for quality assessment. Research in this direction might focus on developing specialized criteria and algorithms to evaluate the perceptual quality of synthetic content. These methods could consider multiple dimensions of quality (*e.g.*, authenticity and fidelity), ensuring that AI-generated 360° images meet high visual standards and provide satisfactory user experiences.

IV. Conclusion

In this survey, we have explored the current landscape of objective OIQA methods, categorizing them into 2D-plane-based, sphere-based, and viewport-based approaches. Despite significant advancements in the field, challenges such as high computational complexity, limited consideration of diverse viewing behaviors and viewing conditions, and the scarcity of large-scale datasets still persist. In addition to addressing these challenges, we point out several future research directions. By highlighting these limitations and potential research directions, we aim to guide future studies towards more precise, efficient, and comprehensive OIQA solutions. Ultimately, the evolution of OIQA methods will play a crucial role in advancing VR technology and improving the quality of experience in immersive environments.

References

[1] Y. Ye, E. Alshina, and J. Boyce, "JVET-G1003: Algorithm description of projection format conversion and video quality metrics in 360lib version 4," Joint Video Exploration Team, Turin, Italy, Rep. JVET-G1003, Tech. Rep., 2017.

[2] L. Yang, M. Xu, X. Deng, and B. Feng, "Spatial attention-based non-reference perceptual quality prediction network for omnidirectional images," in *IEEE International Conference on Multimedia and Expo*, 2021, pp. 1–6.

[3] A. Sendjasni and M.-C. Larabi, "Attention-aware patch-based CNN for blind 360-degree image quality assessment," *Sensors*, vol. 23, no. 21, 2023.

[4] J. Xu, W. Zhou, and Z. Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1724–1737, 2021.

[5] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma, "MC360IQA: A multi-channel CNN for blind 360-degree image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 64–77, 2020.

[6] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang, "Spherical structural similarity index for objective omnidirectional video quality assessment," in *IEEE International Conference on Multimedia and Expo*, 2018, pp. 1–6.

[7] C. Zhang and S. Liu, "No-reference omnidirectional image quality assessment based on joint network," in *ACM International Conference on Multimedia*, 2022, pp. 943–951.

[8] Y. Zhou, Y. Sun, L. Li, K. Gu, and Y. Fang, "Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1767–1777, 2022.

[9] X. Sui, K. Ma, Y. Yao, and Y. Fang, "Perceptual quality assessment of omnidirectional images as moving camera videos," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 8, pp. 3022–3034, 2022.

[10] L. Yang, M. Xu, T. Liu, L. Huo, and X. Gao, "TVFormer: Trajectory-guided visual quality assessment on 360° images with transformers," in *ACM International Conference on Multimedia*, 2022, pp. 799–808.

[11] T. Wu, S. Shi, H. Cai, *et al.*, "Assessor360: Multi-sequence network for blind omnidirectional image quality assessment," in *Advances in Neural Information Processing Systems*, 2023.

[12] X. Sui, H. Zhu, X. Liu, Y. Fang, S. Wang, and Z. Wang, "Perceptual quality assessment of 360° images based on generative scanpath representation," *CoRR*, vol. abs/2309.03472, 2023. [Online]. Available: https://arxiv.org/abs/2309.03472.

[13] W. Zhou, J. Xu, Q. Jiang, and Z. Chen, "No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1778–1791, 2022.

[14] H. Jiang, G. Jiang, M. Yu, T. Luo, and H. Xu, "Multi-angle projection based blind omnidirectional image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4211–4223, 2022.

[15] M. Xu, C. Li, S. Zhang, and P. L. Callet, "State-of-the-art in 360° video/image processing: Perception, assessment and compression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 5–26, 2020, ISSN: 1941-0484.

[16] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," in *Optics and Photonics for Information Processing X*, International Society for Optics and Photonics, vol. 9970, SPIE, 2016, pp. 57–65.

[17] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1408–1412, 2017, ISSN: 1558-2361.

[18] Y. Zhou, M. Yu, H. Ma, H. Shao, and G. Jiang, "Weighted-to-spherically-uniform SSIM objective quality evaluation for panoramic video," in *IEEE International Conference on Signal Processing*, 2018, pp. 54–57.

[19] H. G. Kim, H.-T. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 917–928, 2020.

[20] Y. Liu, X. Yin, Y. Wang, Z. Yin, and Z. Zheng, "HVS-based perception-driven no-reference omnidirectional image quality assessment," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[21] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *IEEE International Symposium on Mixed and Augmented Reality*, 2015, pp. 31–36.

[22] Y. Fang, L. Huang, J. Yan, X. Liu, and Y. Liu, "Perceptual quality assessment of omnidirectional images," in *AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 580–588.

[23] A. Sendjasni and M.-C. Larabi, "PW-360IQA: Perceptually-weighted multichannel CNN for blind 360-degree image quality assessment," *Sensors*, vol. 23, no. 9, 2023.

[24] X. Liu, J. Yan, Z. Wan, Y. Fang, and H. Liu, "Blind quality assessment of panoramic images based on multiple viewport sequences," in *IEEE International Symposium on Circuits and Systems*, 2024, pp. 1–5.

[25] H. G. Kim, H. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 917–928, 2019, ISSN: 1558-2205.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[27] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *European Conference on Computer Vision*, 2018.

[28] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[29] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *European Conference on Computer Vision*, 2006, pp. 404–417.

[30] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "ORB: An efficient alternative to SIFT or SURF," *International Conference on Computer Vision*, pp. 2564–2571, 2011.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[32] X. Sui, Y. Fang, H. Zhu, S. Wang, and Z. Wang, "ScanDMM: A deep markov model of scanpath prediction for 360° images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6989–6999.

[33] B. Ni, H. Peng, M. Chen, *et al.*, "Expanding language-image pretrained models for general video recognition," in *European Conference of Computer Vision*, 2022, pp. 1–18.

[34] V. Sitzmann, A. Serrano, A. Pavel, *et al.*, "Saliency in VR: How do people explore virtual environments?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1633–1642, 2018.

Modeling and Analysis of the Interaction between Opinions and Actions among Heterogeneous Agents

Hangjing Zhang and H. Vicky Zhao
Department of Automation, BNRist, Tsinghua University, Beijing, P. R. China 100084
E-mail: hangjing23@mails.tsinghua.edu.cn; vzhao@tsinghua.edu.cn

***Abstract*—In complex networks where agents exchange opinions and influence each other's decisions, their interactions and decisions are shaped by many personal and social factors, such as bounded rationality and agent heterogeneity. In addition, the discrepancy between opinions and actions, known as the value-action gap, significantly impacts agents' decisions. To better model and analyze the dynamics among heterogeneous agents' opinions and behaviors, in this paper, we propose an opinion-action co-evolution model that captures the intertwined interplay between individual opinions and decision-making process. Our model accounts for agents' bounded rationality by considering both limited cognitive ability and non-cognitive subjective interference. We incorporate a cost function based on behaviour reasoning theory to model the value-action gap. We theoretically analyze the steady states of opinion-action co-evolution for all agent types and use simulations to validate our theoretical analysis and gain further insights.**

***Index Terms*—Co-evolution, opinion, action, heterogeneity**

I. Introduction

In complex networks with intricate structure involving diverse agents and interconnected relationships, agents interact dynamically, leading to changes in their attributes and overall characteristics. The dynamics of opinions and actions in these networks have received lots of attentions in various scenarios, from everyday transactions and cooperations to large-scale events such as elections. According to Cambridge Dictionary, an opinion is a thought or belief about something or someone, while an action refers to something done or performed. They do not exist in isolation as analyzed in most existing works. Rather, they are intertwined in a complex and interrelated manner that influences agents' decision-making processes, social norms, and collective behaviors [1]. Studying the complicated dynamics of opinions and actions is critical to understand and predict complex social behaviors and outcomes, addressing critical issues such as polarization, public health compliance, and collective decision-making.

Numerous factors influence the interplay and evolution of opinion and action. Bounded rationality, which suggests that people tend to find adequate solutions rather than optimal solutions due to various limitations [2], is a significant personal factor in this dynamics, providing a more realistic representation of human decision processes. Beyond personal factors, social influence is essential, as agents within a network often affect each other. Moreover, agents exhibit heterogeneity due to differences in statuses, functions and objectives, e.g., publishers and their audience, suppliers and consumers, or candidates and voters. So the mechanisms by which they update their opinions and actions are different, and social influences occur both within homogeneous groups and across heterogeneous types of agents. Additionally, a value-action gap, also known as the attitude-behavior gap, often exists during the co-evolution of opinion and action. This gap refers to the discrepancy between individuals' stated opinions and their actions [3]. For instance, despite preferences for renewable energies, their adoption is slow due to additional costs and other factors [4].

In summary, it is crucial to understand the complicated interaction between opinion dynamics and action change among heterogeneous agents, and model and analyze factors that influence the complex co-evolution of opinion and action.

A. Literature Review

In the following, we review recent works in the evolution of opinion and action, bounded rationality, and value-action gap.

1) Opinion and Action Evolution: Numerous studies have explored separately on opinion dynamics [5] and action dynamics [6]. Research in social psychology has shown that opinion dynamics and behavioral decisions are deeply intertwined [1]. In other words, an agent's opinion plays a key role in his/her decision-making process, and observed behaviors of others can, in turn, shape his/her opinion. However, only a few works have investigated their co-evolution, which can be categorized into two types: those considering one-sided influence and those considering mutual effects. The former type, such as the seminal CODA model [7], generally regards opinions as continuous variables and models their direct impact on discrete actions. The latter type, such as the study by Luo et al. on the co-evolution of opinion and action in advertising contexts, considers mutual effects [8]. Subsequent works have provided viable frameworks and theoretical analyses, such as conditions for convergence, consensus, and stable states [9], [10]. Authors in [11] explored the co-evolution of actions and opinions on a two-layer network using potential game theory. However, all existing studies ignore the heterogeneity of agents and individual features such as bounded rationality and the value-action gap.

2) Bounded Rationality: Compared with previous theories of unbounded rationality, the bounded rationality emphasizes

Supported by Beijing Natural Science Foundation (IS24036).

the limits of agents' ability to make decisions given specific conditions and constraints, thus agents generally can not seek the optimal solution when achieving goals [12]. According to existing works, three kinds of constraints explain this "boundedness": limitations of cognitive ability, interference of non-cognitive subjective factors, and environment conditions [13]. The first type is embodied in the agent's limited ability to get and process information in the face of complexity and uncertainty [14]. The second type reflects the influence of factors such as emotion, motivation and social psychology. For instance, Simon proposed aspiration level and rules capturing subjective emotional responses and motivation adjustments in [15]; aspiration adaptation theory contains similar ideas [16]; and prospect theory models decision-making under risk, emphasizing that people's aversion to loss is much greater than the pleasure brought by gain [17]. However, existing works typically focus on certain individual characteristics in bounded rationality, without simultaneously addressing and analyzing phenomena like the value-action gap.

3) Value-action Gap: The value-action gap is primarily studied in environmental and social psychology [4], [18]. The causes of this gap can be broadly categorized into environmental influences that focus on social values and moral norms [19], and personal factors that are based on self-interest and individual choice [20]. Personal factors are generally considered to be the primary reasons, such as additional costs, sufficient benefits, and alignment with personal values. Several models have been proposed to explain the value-action gap, including the Theory of Planned Behavior (TPB) [21] and Behaviour Reasoning Theory (BRT) [22]. Considering the reasons for and against the products, Claudy et al. applied BRT to provide a new paradigm for explaining the relationship between consumers' attitudes, values and adoption intentions [4]. However, existing research only validates models through causal analysis and lacks quantitative models and corresponding analyses.

B. Our Contribution

In this work, we jointly consider personal factors and social influences in the complex interaction of agents' opinions and actions, and propose a unified framework to study their coevolving process. The main contributions of this paper are:

- We propose a co-evolution model for opinions and actions among heterogeneous agents including dominant agents and normal agents, to model the complicated interaction among different agents as well as their opinion dynamics and behaviors.
- We mathematically model the bounded rationality during agent's decision-making process and their heterogeneity, to address personal factors and social influences in the dynamic change of opinions and actions respectively. Furthermore, we consider the value-action gap to capture the discrepancy between interactive opinions and actions.
- We theoretically analyze the opinion-action co-evolution for two types of agents respectively, and derive its steady state. This investigation provides important insights into the long-term behavior patterns.

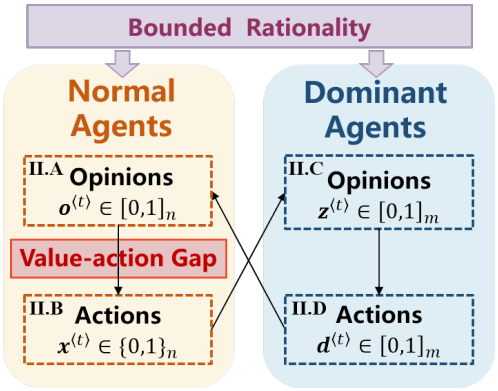


Fig. 1: The model of opinion-action co-evolution for heterogeneous agents.

II. The Opinion-Action Co-evolution Model

In this section, we model the co-evolution of opinions and actions among heterogeneous agents, considering their bounded rationality, social influence and value-action gap. Using continual transactions of products between fixed suppliers and consumers as an example, suppliers initiate product offerings and persuade consumers to make purchases. These agents are heterogeneous due to their differing statuses and objectives. Suppliers' opinions are their emphasis on the product, and their actions reflect the efforts and resources (e.g., advertising intensity and manpower) they invest in encouraging purchases. Consumers' opinions are their thoughts about the product and the supplier, while their actions represent their purchasing choices. As interactions continue, agents adjust their opinions and decisions accordingly. We use a directed graph to model the social connections in complex networks among different agents.

To distinguish heterogeneous agents with various status and objectives, we divide all agents into two categories: dominant agents and normal agents, corresponding to suppliers and consumers in the above example, respectively. Dominant agents are initiators or service providers, aiming to make the audience, i.e., normal agents, approve their objectives and take actions. Normal agents, the targets of dominant agents, primarily focus on their own needs. Empirical evidence and models incorporating the bounded rationality have investigated the heterogeneity of agents [23]. Specifically, some sophisticated agents can process information with better abilities and fewer constraints, aligning with dominant agents; while others have more limited capabilities, corresponding to normal agents. Moreover, research on the value-action gap is typically conducted among target group of specific policy or products, indicating that this gap generally occurs among normal agents [4]. Therefore, we introduce different update mechanisms for the decision-making processes of dominant and normal agents.

The co-evolution model is illustrated in Fig. 1. For two types of agents, they both have two interactive attributes: opinions and actions. Here we consider a simplified scenario that the action of an agent is only influenced by his/her opinion,

and the opinion would be affected by the observed actions of other type of agents. The value-action gap exists when normal agents take actions based on their own opinions. When different types of agents decide on their opinions and actions, bounded rationality plays a role to different degree, thus the four submodels in Fig. 1, i.e., II.A, II.B, II.C and II.D that introduced in the following, operate in different ways.

Assume there are n normal agents and m dominant agents. We divide time into slots, and in each time slot t, every agent updates his/her opinion and action based on the observed opinions and actions of others and his/her own interest. Normal agents' continuous opinion vector and discrete action vector are denoted by $\boldsymbol{o}^{\langle t\rangle} \in [0,1]_n$ and $\boldsymbol{x}^{\langle t\rangle} \in \{0,1\}_n$ respectively. Dominant agents' continuous opinion vector and action vector are denoted by $\boldsymbol{z}^{\langle t\rangle} \in [0,1]_m$ and $\boldsymbol{d}^{\langle t\rangle} \in [0,1]_m$ respectively. Here, a value of 1 in continuous opinion variables signifies strong positivity, while 0 denotes strong negativity. For discrete action variables for normal agents, 1 indicates execution of behaviors such as purchase or vote, and 0 indicates the opposite. For continuous action variables for dominant agents, 1 means maximal effort towards action, while 0 means inaction. We assume that different types of agents can only observe others' actions. Then we introduce submodels of opinions and actions for normal agents and dominant agents respectively.

A. Update Mechanisms for Normal Agents' opinions

We define the update mechanism of normal agents' opinion vector $\boldsymbol{o}^{\langle t\rangle}$ based on Friedkin-Johnsen (FJ) model [24] as:

$$\boldsymbol{o}^{\langle t\rangle} = \boldsymbol{A}_n \boldsymbol{W} \boldsymbol{o}^{\langle t-1\rangle} + (\boldsymbol{I} - \boldsymbol{A}_n)\boldsymbol{s}^{\langle t\rangle}, \tag{1}$$

where $\boldsymbol{W}$ is the graph weight matrix with $w_{i,j} \geq 0$ and $\sum_{j=1}^{n} w_{i,j} = 1$ and represents the influence of connected normal neighbors' opinions. $\boldsymbol{A}_n = \operatorname{diag}(a_1, \cdots, a_n)$ is an $n \times n$ diagonal susceptibility matrix of normal agents indicating their sensitivity to social influences. $\boldsymbol{s}^{\langle t\rangle}$ is normal agents' internal opinion vector, which represents individual current thoughts.

The internal opinion of a normal agent is affected by the observed actions of dominant agents, the agent's initial opinion and his/her previous internal opinion. First, when modeling the impact of observed actions of dominant agents, we consider agents' bounded rationality caused by limited cognitive ability. We assume that normal agents cannot ascertain the true values of dominant agents' actions and have to estimate them based on limited information. Thus, similar to [25], we define $\hat{d}_{ij}^{\langle t-1\rangle} \sim \mathrm{TN}\left(d_j^{\langle t-1\rangle}, \sigma_i^2, 0, 1\right)$ as the estimate of dominant agent j's action generated by normal agent i in time slot $t-1$, sampled from a truncated normal distribution on $[0,1]$, with the actual action value $d_j^{\langle t-1\rangle}$ as the mean and σ_i^2 as the variance, reflecting the diversity in perception. The normal agent i forms an overall impression of all dominant agents' actions via the mean $\overline{\hat{d}_i^{\langle t-1\rangle}} = \frac{1}{m}\sum_{j=1}^{m} \hat{d}_{ij}^{\langle t-1\rangle}$. We then represent the estimate vector of dominant agents' actions generated by all normal agents by $\overline{\hat{\boldsymbol{d}}^{\langle t-1\rangle}} = \left[\overline{\hat{d}_1^{\langle t-1\rangle}}, \overline{\hat{d}_2^{\langle t-1\rangle}}, \cdots, \overline{\hat{d}_n^{\langle t-1\rangle}}\right]^T$. The influence of dominant agents' actions on normal agents is captured by $\boldsymbol{F}_n = \operatorname{diag}(F_{n1}, \cdots, F_{nn})$, which is an $n \times n$ diagonal influence matrix of normal agents representing the susceptibilities to the actions of dominant agents.

The initial opinion vector of normal agents, denoted by $\boldsymbol{s}^{\langle 0\rangle}$, is a crucial factor in opinion dynamics as stated in Friedkin-Johnsen (FJ) model [24]. The influence degree is represented by $\boldsymbol{\Gamma} = \operatorname{diag}(\gamma_1, \cdots, \gamma_n)$, an $n \times n$ diagonal stubbornness matrix of normal agents representing their stubbornness for initial opinions.

Then following prior works in [26], we model the internal opinion vector of normal agents $\boldsymbol{s}^{\langle t\rangle}$ as:

$$\boldsymbol{s}^{\langle t\rangle} = \boldsymbol{F}_n \overline{\hat{\boldsymbol{d}}^{\langle t-1\rangle}} + \boldsymbol{\Gamma}\boldsymbol{s}^{\langle 0\rangle} + (\boldsymbol{I} - \boldsymbol{F}_n - \boldsymbol{\Gamma})\,\boldsymbol{s}^{\langle t-1\rangle}. \tag{2}$$

B. Update Mechanisms for Normal Agents' actions

We assume that normal agents' actions are only affected by their individual opinions, value-action gap and their bounded rationality, without considering the influence of neighbors' actions. Following [27], we assume that normal nodes' actions evolve according to a stochastic process. Specifically, we formulate the probability for agent i to adopt action $x \in \{0,1\}$ at time t based on the mixed logit model [28], considering agents' uncertainty in determining strategy. It is defined as:

$$P\left(x_i^{\langle t\rangle} = x\right) = \frac{e^{\pi_i(x)}}{e^{\pi_i(x)} + e^{\pi_i(1-x)}}, \tag{3}$$

where $\pi_i(x)$ is the payoff when the normal agent i adopting action x given his/her current opinion $o_i^{\langle t\rangle}$. When defining the payoff function, we consider issues including the agent's opinion $o_i^{\langle t\rangle}$, the cost $C_n\left(x, o_i^{\langle t\rangle}\right)$ modeling the value-action gap, and individual bounded rationality represented by the value function $V_n(\cdot)$:

$$\pi_i(x) = V_n\left(\kappa\left[\lambda_i o_i^{\langle t\rangle} + (1-\lambda_i)\, C_n\left(x, o_i^{\langle t\rangle}\right)\right]\right). \tag{4}$$

In (4), $\kappa > 1$ is the expansion coefficient, and $\lambda_i \in [0,1]$ is the trade-off coefficient between individual opinion and the cost.

$V_n(\cdot)$ is the value function in the Prospect Theory [17]:

$$V_n(x) = \begin{cases} x^{\alpha}, & x \geq 0, \\ -\rho(-x)^{\beta}, & x < 0. \end{cases} \tag{5}$$

It means that subjective value deviates from actual value due to cognitive biases and subjective evaluations under risk. Agents exhibit risk aversion and tend to underestimate positive values, while showing risk preference and overestimating negative values. In addition, the sensitivity to negative values is stronger than to positive values. This embodies agents' bounded rationality due to non-cognitive subjective factors. From existing studies [17], we set the parameters as $\alpha = \beta = 0.88, \rho = 2.25$.

$C_n\left(x, o_i^{\langle t\rangle}\right)$, whose value is negative, quantifies the cost of adopting the action x with opinion $o_i^{\langle t\rangle}$, and is designed based on BRT to explain the value-action gap [4]. It represents the price the agent must pay, which discourages the agent from taking the action, and is denoted by:

$$C_n\left(x, o_i^{\langle t\rangle}\right) = \begin{cases} o_i^{\langle t\rangle} - 1 - c, & x = 1, \\ -o_i^{\langle t\rangle}, & x = 0, \end{cases} \tag{6}$$

where $c > 0$ is the additional loss when $x = 1$, as the execution generally costs more. We can see that the absolute value of cost function is the absolute difference between the opinion and action as well as the additional loss, and the cost is lower when the values of opinion and action are closer.

C. Update Mechanisms for Dominant Agents' Opinions

During the interaction, opinions of dominant agents are primarily influenced by the actions of normal agents. Then the opinion vector of dominant agents $\boldsymbol{z}^{\langle t\rangle}$ is updated by:

$$\boldsymbol{z}^{\langle t\rangle} = \boldsymbol{F}_d\,\boldsymbol{\omega}\left(p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1\rangle}\right) + (\boldsymbol{I}-\boldsymbol{F}_d)\boldsymbol{z}^{\langle t-1\rangle}, \tag{7}$$

where $\boldsymbol{F}_d = \mathrm{diag}\,(F_{d1},\cdots,F_{dm})$ is an $m\times m$ diagonal influence matrix of dominant agents representing their susceptibilities to the influence of normal agents' actions.

In (7), $p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1\rangle}$ is the proportion of normal agents adopting the action 1 in the time slot $t-1$. $\boldsymbol{\omega}\left(p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1\rangle}\right) = \left[\omega_1\left(p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1\rangle}\right),\cdots,\omega_m\left(p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1\rangle}\right)\right] : [0,1] \rightarrow [0,1]^{m\times 1}$ is a mapping vector function of $p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1\rangle}$, signifying the effect of normal agents' actions on different dominant agents. When defining the mapping function $\boldsymbol{\omega}\left(p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1\rangle}\right)$, note that based on empirical studies, estimates of the frequency of events by human observers are typically distorted, with the relative low frequency overestimated and the relative high frequency underestimated. Therefore, following the work in [29], the distortion of the proportion $\omega_i\left(p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1\rangle}\right)$ is denoted as the form of linear transformations of the log odds:

$$\mathrm{Lo}\left(\omega_i\left(p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1\rangle}\right)\right) = \epsilon_i\,\mathrm{Lo}\left(p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1\rangle}\right) + (1-\epsilon_i)\mathrm{Lo}(p_0). \tag{8}$$

The log odds $\mathrm{Lo}(p) = \log\frac{p}{1-p}$. $\epsilon_i \in (0,1)$ is the slope of the linear transformation, and p_0 is the fixed point which means that the value of p is mapped to itself. Note that according to [29], the prediction of fixed point p_0 in (8) is the inverse of the number of classes, so here p_0 is set to be 0.5.

D. Update Mechanisms for Dominant Agents' Actions

We adopt the continuous generation model to formulate the actions of dominant agents in [30]. In time slot t, the dominant agent i's action is affected by all his/her previous and current opinions, which can be understood as a process of continuous revision on current action decision from the influence of opinion. He/she will decide on the action by random sampling from the following beta distribution:

$$d_i^{\langle t\rangle} \sim \mathrm{Beta}\left(\sum_{l=0}^{t} z_i^{\langle l\rangle}, t+1-\sum_{l=0}^{t} z_i^{\langle l\rangle}\right). \tag{9}$$

The probability density function of $\mathrm{Beta}\,(a,b)$ is $f(x;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$, and $a = \sum_{l=0}^{t} z_i^{\langle l\rangle}$ and $b = t+1-\sum_{l=0}^{t} z_i^{\langle l\rangle}$ in (9).

III. The Analysis of Opinion and Action Dynamics and Steady State

Based on the proposed co-evolution model, we analyze how both normal and dominant agents' opinions and actions change dynamically, and obtain theoretical results of the evolutionary steady states. Here we give conclusions directly and the detailed derivation can be found in Supplementary [31].

A. Opinion Dynamics

The dynamic results of normal agents' and dominant agents' opinions can be represented by iterations according to their evolution rule in (1), (2) and (7) as:

$$\boldsymbol{o}^{\langle t\rangle} = (\boldsymbol{A}_n\boldsymbol{W})^{t-1}\boldsymbol{s}^{\langle 0\rangle} + \sum_{i=0}^{t-2}(\boldsymbol{A}_n\boldsymbol{W})^{t-i}(\boldsymbol{I}-\boldsymbol{A}_n)\boldsymbol{s}^{\langle i\rangle}, \tag{10}$$

$$\boldsymbol{s}^{\langle t\rangle} = \left[\sum_{i=0}^{t-2}(\boldsymbol{I}-\boldsymbol{F}_n-\boldsymbol{\Gamma})^i\,\boldsymbol{\Gamma} + (\boldsymbol{I}-\boldsymbol{F}_n-\boldsymbol{\Gamma})^{t-1}\right]\boldsymbol{s}^{\langle 0\rangle} + \sum_{i=0}^{t-2}(\boldsymbol{I}-\boldsymbol{F}_n-\boldsymbol{\Gamma})^i\,\boldsymbol{F}_n\overline{\hat{\boldsymbol{d}}^{\langle t-1-i\rangle}}, \tag{11}$$

$$\boldsymbol{z}^{\langle t\rangle} = \sum_{i=0}^{t-1}(\boldsymbol{I}-\boldsymbol{F}_d)^i\,\boldsymbol{F}_d\,\boldsymbol{\omega}\left(p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t-1-i\rangle}\right) + (\boldsymbol{I}-\boldsymbol{F}_d)^t\boldsymbol{z}^{\langle 0\rangle}. \tag{12}$$

where $\boldsymbol{s}^{\langle i\rangle}$ in (10) is substituted by (11). Thus when the initial opinion vector $\overline{\boldsymbol{s}^{\langle 0\rangle}}, \boldsymbol{z}^{\langle 0\rangle}$, the vector of observed dominant agents' action $\overline{\hat{\boldsymbol{d}}^{\langle t\rangle}}$ and the proportion of normal agents' actions $p_{\boldsymbol{x}=\boldsymbol{1}}^{\langle t\rangle}$ in each time slot are known, the opinions of both types of agents in current time slot can be directly obtained.

B. The Steady State

Note that when the actions of normal agents evolve to a steady state, although some agents would change their actions with probabilities, the proportion of normal agents who take actions 1 will remain asymptotically stable. Therefore, we pay attention to the proportion of agents adopting the action 1 as the steady state of normal agents' actions. We can show that:

$$\mathbb{E}[\boldsymbol{o}^*] = (\boldsymbol{I}-\boldsymbol{A}_n\boldsymbol{W})^{-1}(\boldsymbol{I}-\boldsymbol{A}_n)(\boldsymbol{F}_n+\boldsymbol{\Gamma})^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{s}^{\langle 0\rangle}+\mu^*\boldsymbol{f}_n\right), \tag{13}$$

$$\mathbb{E}\left[p_{\boldsymbol{x}=1}^*\right] = \frac{1}{n}\sum_{i=1}^{n} P\left(x_i^{\langle t\rangle}=1|o_i^{\langle t\rangle}=o_i^*\right), \tag{14}$$

$$\mathbb{E}\left[\boldsymbol{z}^*\right] = \boldsymbol{\omega}\left(\mathbb{E}\left[p_{\boldsymbol{x}=\boldsymbol{1}}^*\right]\right), \tag{15}$$

$$\mathbb{E}\left[\boldsymbol{d}^*\right] = \mathbb{E}\left[\boldsymbol{z}^*\right], \tag{16}$$

where all derivations are in the Supplementary [31]. $\mu^* = \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[d_j^*]$ is the mean of steady states for dominant agents' actions, $\boldsymbol{f}_n = (F_{n1},\cdots,F_{nn})^T$ is an $n\times 1$ column vector whose elements are the diagonal elements of $\boldsymbol{F}_n$, $P\left(x_i^{\langle t\rangle}=1|o_i^{\langle t\rangle}=o_i^*\right)$ is the probability of adopting the action 1 when the opinion is asymptotically stable, and its detailed representation is in the Supplementary [31], and $\omega_i\left(\mathbb{E}\left[p_{\boldsymbol{x}=\boldsymbol{1}}^*\right]\right) = \left[1+\left(\frac{1}{\mathbb{E}\left[p_{\boldsymbol{x}=\boldsymbol{1}}^*\right]}-1\right)^{\epsilon_i}\right]^{-1}$. To simplify the analysis in this work, we consider the simple scenario where the trade-off coefficients of different normal agents are equal, i.e.,

$\lambda_1 = \cdots = \lambda_n = \lambda$. In this simple scenario, we can see that the expectation of dominant agents' actions are the same as that of opinions, and the dominant agents' opinions are the result of the distortion for the proportion of normal agents' action. We will investigate the more complicated scenario where $\lambda_i \neq \lambda_j, \exists i \neq j$ in our future work.

Due to the large number of equations and the complexity of (13)-(16), we cannot obtain the theoretical solutions. We propose Algorithm 1 to numerically find the steady states (SS).

Algorithm 1: SS-learning Algorithm

Input: W, A_n, F_n, Γ, learning and decay rate r_l, r_d, error threshold e_t

Output: the expectation of steady states for normal agents' and dominant agents' opinions and actions $\mathbb{E}[\boldsymbol{o}^*], \mathbb{E}\left[p^*_{\boldsymbol{x}=1}\right], \mathbb{E}[\boldsymbol{z}^*], \mathbb{E}[\boldsymbol{d}^*]$

Set initial error $e = +\infty$, initial predicted proportion of normal agents' action $p^0_{\boldsymbol{x}=1} = 0.5$, initial gradient direction $D = 0$ and iteration times $It = 0$;

while $e > e_t$ **do**

  $It = It + 1$;

  Update current predicted proportion of normal agents' action by $p^{It}_{\boldsymbol{x}=1} = p^{It-1}_{\boldsymbol{x}=1} + D * r_l * (r_d)^{It}$;

  while $p^{It}_{\boldsymbol{x}=1} \notin [0,1]$ **do**

    $It = It + 1$, $p^{It}_{\boldsymbol{x}=1} = p^{It-1}_{\boldsymbol{x}=1} + D * r_l * (r_d)^{It}$

  Use the current predicted proportion $p^{It}_{\boldsymbol{x}=1}$ to compute $\mathbb{E}[\boldsymbol{z}^*]$ using (15);

  Use the current predicted $\mathbb{E}[\boldsymbol{d}^*]$, i.e. $\mathbb{E}[\boldsymbol{z}^*]$, to compute $\mathbb{E}[\boldsymbol{o}^*]$ using (13);

  Use the current predicted $\mathbb{E}[\boldsymbol{o}^*]$ to compute $P\left(x_i^{\langle t \rangle} = 1 | o_i^{\langle t \rangle} = o_i^*\right)$ using simplified formulas in Supplementary [31];

  Compute the error $e = \frac{1}{n}\sum_{i=1}^{n} P\left(x_i^{\langle t \rangle} = 1 | o_i^{\langle t \rangle} = o_i^*\right) - p^{It}_{\boldsymbol{x}=1}$;

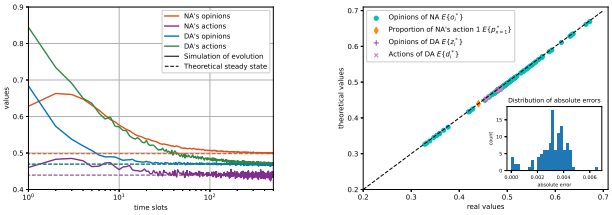
  $D = \mathrm{sgn}(e)$, where $\mathrm{sgn}(\cdot)$ is the sign function;

return $p^{It}_{\boldsymbol{x}=1}$ as $\mathbb{E}\left[p^*_{\boldsymbol{x}=1}\right]$ and compute other steady states by (13), (15) and (16)

IV. Simulations Results

In this section, we run simulations to verify our theoretical analysis. We assume that there are $n = 100$ normal agents and $m = 5$ dominant agents, and normal agents are in an Erdős–Rényi (ER) random weighted network with connecting probability $p = 0.4$. Normal agents' initial opinions $\boldsymbol{s}^{\langle 0 \rangle}$ are set randomly in $[0, 1]$. Diagonal elements of the influence matrices and other coefficient matrices, including $\boldsymbol{F}_n$, $\boldsymbol{F}_d$, $\boldsymbol{A}_n$ and $\boldsymbol{\Gamma}$, take random values in $[0, 1]$.The expansion coefficient is $\kappa = 6$. Dominant agents' distortion of the proportion of the action ϵ_i is independently and uniformly distributed in $[0.2, 1]$, which reflects dominant agents' subjective cognitive ability. In each simulation run, the update process is repeated until opinions and actions reach the stable state.

We first observe the whole co-evolution process and validate our theoretical analysis. Fig. 2 shows the simulation and theoretical results when $\lambda = 0.9$ and $c = 0.5$. In Fig. 2(a), solid lines are the mean results with 200 repeated simulations, and dotted lines are the predicted theoretical steady states. We can see that as time goes, simulation results align well with the theoretical steady states. Note that dominant agents' opinions and actions converge to the same point, consistent with (16). Also, the steady states of normal agents' opinions and actions are inconsistent, validating the value-action gap. We can also see that the rate of convergence of dominant agents' opinions is quicker than others. In Fig. 2(b), we scatter plot pairs of real steady states and theoretical analysis results derived by (13)-(16) for each normal agent's and dominant agent's opinions and actions. They are all on or very close to the dashed black line $y = x$, validating our theoretical analysis again.

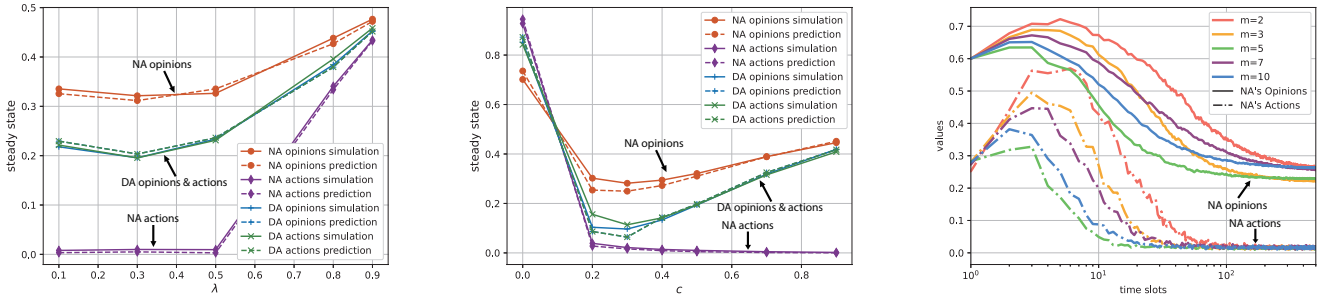


(a) Evolution process and theoretical steady state of normal and dominant agents' opinions and actions.

(b) The comparison of real values and theoretical values and the distribution of all absolute errors.

Fig. 2: Simulation and theoretical results with $\lambda = 0.9$, $c = 0.5$. NA represents normal agents. DA represents dominant agents.

We then show the impact of different parameters in our model on the final steady state and evolution process. Fig. 3(a) shows the influence of λ on the steady states, which is the trade-off coefficients between normal agent's opinion and the cost, in 5 situations: $\lambda = 0.1, 0.3, 0.5, 0.8, 0.9$. As expected, the steady states of all agents' opinions and actions increase with λ since the cost is more unimportant. The simulation results and predicted steady states match well, which validates our analysis again. Fig. 3(b) shows the effect of c to steady states, which is the additional loss for the action 1, in 7 situations: $c = 0, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9$. We can see that when c increases, fewer normal agents adopt the action 1 due to the rise of additional loss, while other variables including normal agents' opinions, dominant agents' opinions and actions decrease first and then increase. This indicates that the additional loss aggravates normal agent's value-action gap. To avoid the deep gap, dominant agents should take actions to diminish the loss of execution. Fig. 3(c) shows the effect of m to the evolution process, or equivalently m/n, which is the proportion of dominant agents to normal agents, in 5 situations: $m = 2, 3, 5, 7, 10$ with $n = 100$. The results of normal agents' opinions and actions are shown as the example. It can be found that the speed of evolution reaching the final steady states does not monotonically increase with the number of agents. In these 5 situations, normal agents achieve the optimal results from the perspective of dominant agents when $m = 5$, as the co-evolution being stable more quickly. This suggests that dominant agents should arrange their resources and scale in a reasonable range.



(a) The effect of different trade-off coefficients between normal agent's opinion and the cost λ = action 1 $c = 0.0, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9$. $0.1, 0.3, 0.5, 0.8, 0.9$.

(b) The effect of different additional loss for the agents $m = 2, 3, 5, 7, 10$, i.e., the effect of the proportion of dominant agents to normal agents.

(c) The effect of different numbers of dominant agents $m = 2, 3, 5, 7, 10$, i.e., the effect of the proportion of dominant agents to normal agents.

Fig. 3: The simulations and theoretical results of steady states and evolution process in different settings.

V. Conclusions

In this paper, we propose a unified co-evolution model of opinions and actions among dominant and normal agents to study how heterogeneous agents' opinions and actions interact. We incorporate bounded rationality and the value-action gap into our framework. We theoretically analyze the dynamics and steady state of both normal and dominant agents' opinions and actions. Simulations are conducted to validate our analysis, and show the effect of model parameters including λ, c and m/n, indicating that dominant agents should take actions to avoid deeper value-action gap and also arrange their resources properly to accelerate the process to the steady state, providing references for the practical decision.

References

[1] S. Gavrilets and P. J. Richerson, "Collective action and the evolution of social norm internalization," *Proceedings of the National Academy of Sciences*, vol. 114, no. 23, pp. 6068–6073, 2017.

[2] A. L. Olsen, "" simon said," we didn't jump," *Pub. Admin. Rev.*, vol. 75, p. 325, 2015.

[3] *Value-action gap*, Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Value-action_gap.

[4] M. C. Claudy, M. Peterson, and A. O'Driscoll, "Understanding the attitude-behavior gap for renewable energy systems using behavioral reasoning theory," *Journal of Macromarketing*, vol. 33, no. 4, pp. 273–287, 2013.

[5] H. Noorazar, "Recent advances in opinion propagation dynamics: A 2020 survey," *The European Physical Journal Plus*, vol. 135, pp. 1–20, 2020.

[6] M. Li, X. Wang, K. Gao, and S. Zhang, "A survey on information diffusion in online social networks: Models and methods," *Information*, vol. 8, no. 4, p. 118, 2017.

[7] A. C. Martins, "Continuous opinions and discrete actions in opinion dynamics problems," *International Journal of Modern Physics C*, vol. 19, no. 04, pp. 617–624, 2008.

[8] G.-X. Luo, Y. Liu, Q.-A. Zeng, S.-M. Diao, and F. Xiong, "A dynamic evolution model of human opinion as affected by advertising," *Physica A: Statistical Mechanics and its Applications*, vol. 414, pp. 254–262, 2014.

[9] L. Zino, M. Ye, and M. Cao, "A coevolutionary model for actions and opinions in social networks," in *2020 59th IEEE Conference on Decision and Control (CDC)*, IEEE, 2020, pp. 1110–1115.

[10] M. Zhan, G. Kou, Y. Dong, F. Chiclana, and E. Herrera-Viedma, "Bounded confidence evolution of opinions and actions in social networks," *IEEE Transactions on Cybernetics*, vol. 52, no. 7, pp. 7017–7028, 2021.

[11] H. D. Aghbolagh, M. Ye, L. Zino, Z. Chen, and M. Cao, "Coevolutionary dynamics of actions and opinions in social networks," *IEEE Transactions on Automatic Control*, 2023.

[12] H. A. Simon, "Theories of bounded rationality, decision and organization," *CBR a. R. Radner. Amsterdam, NorthHolland*, 1972.

[13] L. Yongfang, "The essence of bounded rationality and debate over its value," *Acta Psychologica Sinica*, vol. 54, no. 11, p. 1293, 2022.

[14] H. A. Simon, "Barriers and bounds to rationality," *Structural Change and Economic Dynamics*, vol. 11, no. 1-2, pp. 243–253, 2000.

[15] ——, "A behavioral model of rational choice," *The quarterly journal of economics*, pp. 99–118, 1955.

[16] R. Selten, "Aspiration adaptation theory," *Journal of mathematical psychology*, vol. 42, no. 2-3, pp. 191–214, 1998.

[17] D. Kahneman and A. Tversky, "Choices, values, and frames.," *American psychologist*, vol. 39, no. 4, p. 341, 1984.

[18] R. H. Fazio and D. R. Roskos-Ewoldsen, "Acting as we feel: When and how attitudes guide behavior.," 2005.

[19] M. Press and E. J. Arnould, "Constraints on sustainable energy consumption: Market system and public policy challenges and opportunities," *Journal of Public Policy & Marketing*, vol. 28, no. 1, pp. 102–113, 2009.

[20] A. Prothero, P. McDonagh, and S. Dobscha, "Is green the new black? reflections on a green commodity discourse," *Journal of Macromarketing*, vol. 30, no. 2, pp. 147–159, 2010.

[21] I. Ajzen, "The theory of planned behavior," *Organizational behavior and human decision processes*, vol. 50, no. 2, pp. 179–211, 1991.

[22] J. D. Westaby, "Behavioral reasoning theory: Identifying new linkages underlying intentions and behavior," *Organizational behavior and human decision processes*, vol. 98, no. 2, pp. 97–120, 2005.

[23] J. Haltiwanger and M. Waldman, "Rational expectations and the limits of rationality: An analysis of heterogeneity," *The American Economic Review*, vol. 75, no. 3, pp. 326–340, 1985.

[24] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *Journal of mathematical sociology*, vol. 15, no. 3-4, pp. 193–206, 1990.

[25] T. Choi and P. J. Robertson, "Contributors and free-riders in collaborative governance: A computational exploration of social motivation and its effects," *Journal of Public Administration Research and Theory*, vol. 29, no. 3, pp. 394–413, 2019.

[26] S. Gavrilets, D. Tverskoi, and A. Sánchez, "Modelling social norms: An integration of the norm-utility approach with beliefs dynamics," *Philosophical Transactions of the Royal Society B*, vol. 379, no. 1897, p. 20 230 027, 2024.

[27] L. Zino, M. Ye, and M. Cao, "A two-layer model for coevolving opinion dynamics and collective decision-making in complex social systems," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 8, 2020.

[28] Y. Wang, A. V. Vasilakos, J. Ma, and N. Xiong, "On studying the impact of uncertainty on behavior diffusion in social networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 2, pp. 185–197, 2014.

[29] H. Zhang and L. T. Maloney, "Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition," *Frontiers in neuroscience*, vol. 6, p. 1, 2012.

[30] Y. Gu, Y. Sun, and J. Gao, "The co-evolution model for social network evolving and opinion migration," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 175–184.

[31] H. Zhang and H. V. Zhao, *Supplementary information: Modeling and analysis of the interaction between opinions and actions over heterogeneous agents*. [Online]. Available: https://cloud.tsinghua.edu.cn/f/bead9f27d4d2489fa943/.

ViP-CBM: Reducing Parameters in Concept Bottleneck Models by Visual-Projected Embeddings

Ji Qi, Huisheng Wang, H. Vicky Zhao
Department of Automation, Tsinghua University, Beijing, China
E-mail: {qij21,whs22}@mails.tsinghua.edu.cn, vzhao@tsinghua.edu.cn

***Abstract*—With the widespread application of deep neural networks (DNN) in key areas of personal and property security, the interpretability and trustworthiness of DNN models become increasingly significant. Concept Bottleneck Models (CBM) are popular interpretable models in which hidden layer neurons are trained to predict human-understandable concepts of the final task. However, existing CBMs encounter low efficiency and interpretability in the multi-label classification (MLC) of concepts: correlations of concepts are ignored by many CBMs, while others expressing correlations with complex models have very limited improvements in the model's performance. To address the challenge of massive parameters and limited interpretability in concept MLC problem, we propose a novel *Visual-Projected CBM (ViP-CBM)*, which transforms MLC of concepts into an input-dependent binary classification problem of concept embeddings using visual features for projection. Our ViP-CBM model reduces the training parameter set by more than 50% with the number of training parameters when compared to other *embedding-based* CBMs. Experimental results show that our ViP-CBM achieves comparable or even better performance in concept prediction.**

I. Introduction

As deep neural networks (DNNs) are increasingly used in scenarios concerning personal and property security such as healthcare, automatic driving, and financial services, the trustworthiness and interpretability of DNNs have been increasingly significant, bringing explainable AI (XAI) to the forefront. Researches on XAI mainly follow two routines: the post-hoc explanation that looks into already-trained black-box models, and human-interpretable modeling that aims to build a DNN with an entirely human-understandable inference process. Traditional post-hoc explanation methods such as LIME [1] and SHAP [2] try to identify the most important parts of the input data that support the model's decision. Network Dissection [3] attempts to explain image classification by convolutional neural networks (CNNs) through features in the intermediate layers, but cannot align these features perfectly with human cognition. In summary, end-to-end trained black-box DNNs usually cannot be entirely explained in a human-interpretable way, leading us to achieve interpretability by model construction from scratch. A natural conceit is to train models to provide information in the intermediate layers that support the decisions, which requires step-by-step design in reasoning with additional supervision. Concept-based models are popular human-interpretable models that explain the decisions of models with supporting high-level concepts. Concept Whitening [4] performs affine transformations in the latent space to align axes with concepts of interests, and research in [5] assigns each layer of a CNN to capture certain concepts from low-level (e.g., colors, textures) to high-level (e.g., objects). The above methods provide layer-wise concept explanations with massive annotation and high training costs.

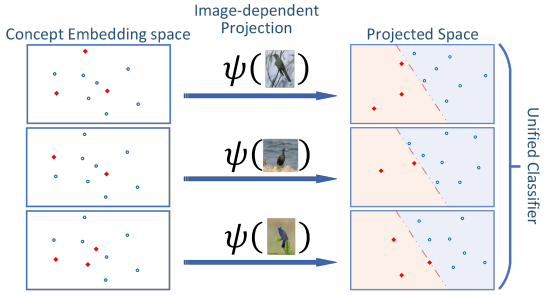


Fig. 1. Our ViP-CBM converts multi-label classification of concepts to unified binary classification of concept embeddings projected by visual features of input images. $\psi(\cdot)$ denotes the visual feature extractor.

In this work, we focus on a certain type of interpretable concept-based model for image classification, the Concept Bottleneck Model (CBM) [6]. CBMs split end-to-end prediction into two steps: first to predict the concepts and then to predict classes using only concepts. CBM is a simple and useful interpretable deep model since it enables humans to understand the decisions of the model with concept predictions and allows test-time intervention to improve accuracy in downstream tasks by correcting false concept predictions.

As the bottleneck of CBM, concept prediction in conventional CBMs usually learn concepts from an N-way binary classifier for all N concepts. Since the classifier treats all labels disconnected, correlations between concepts are neglected, leading to a loss of accuracy, especially in cases where the concept label set is large and a loss of interpretability as well. To address this issue, the work in [7] uses an autoregressive architecture inspired by the classifier chain [8] in multi-label classification (MLC) to capture correlations among concepts, which improves concept accuracy and task accuracy significantly. ECBM [9] employs an energy-based model to learn the relevance of concepts through inferences on a probability graph, where gradient descent is applied to minimize the energy function for prediction. In summary, current studies on solving the MLC problems in CBM involve more parameters and a complex model structure and require extensive samplings and calculations with limited performance improvements.

To improve the efficiency and interpretability of the MLC problem in concept prediction, we propose the **Visual-Projecting Concept Bottleneck Models (ViP-CBM)**. In ViP-CBM, visual features extracted from the input image are used as projecting matrices on the concept embedding space. The projected embedding vectors of the concepts are binary classified as activated or not by a unified linear classifier, as is shown in Figure 1. Through the visual projecting (ViP) module in ViP-CBM, we convert the MLC problem in concept prediction into an input-dependent binary classification problem, which intuitively explains the classification with the projection mechanism and attains better interpretability. The contributions of this work are as follows:

- We propose an interpretable ViP module for image multi-label classification to convert MLC problems into binary classification problems in the embedding space of labels.
- We propose ViP-CBM, which reduces the number of training parameters to that of the minimal scalar CBM, which is less than 50% of that of other embedding-based CBMs. Our ViP-CBM achieves similar performance in both concept prediction and class prediction when compared to other CBMs.

II. Related Works

In this section, we review related works in CBMs and visual-semantic embedding that inspire our work.

A. Concept Bottleneck Models

Concept Bottleneck Models (CBM) [6] consists of two parts, the concept predictor and the class predictor. The concept predictor predicts human-specified concepts from the input image, while the class predictor receives predictions from the concept predictors as input to predict the classes. Earlier studies in CBMs employ scalar variables in the last layer before downstream tasks as concept predicting logits or probabilities. Assuming that the incompleteness of the concept set prevents the CBM from achieving higher task accuracy, side channels [7] are introduced in hard CBMs to represent undiscovered binary concepts to enhance model performance. Coop-CBM [10] adds a side branch before the concept prediction layer for immediate task prediction, leading to higher accuracy in concept prediction. Post hoc CBM [11] predicts concepts by projecting extracted features on Concept Activation Vectors (CAVs) trained from other supporting datasets by SVM or multimodal models, and Label-free CBM [12] further employ GPT-3 [13] for concept annotation to eliminate the need for densely annotated data. Both models convert pre-trained black-box models into CBMs and maintain task accuracy comparable to the original black-box networks.

To improve the expressivity of concepts in the model, the Concept Embedding Model (CEM) [14] learns a pair of positive and negative embeddings for each concept from the input to extend feature representations to higher dimensions. ProbCBM [15] and ECBM [9] use individually trained concept embeddings for concept prediction through their relationship with features extracted from the original input, the former using Euclidean distance in space and the latter using Boltzmann energy models. However, all the above methods introduce more parameters or even other large models and additional data to attain higher concept and task accuracy, which makes model structures and training complicated and increases training costs, and deviates from the original intent of achieving interpretability on basic small models by feature supervision.

B. Visual-semantic Embeddings Models

Visual-semantic embedding models exploit label semantic relationships by leveraging the textural data from the label set to map visual features into a rich-semantic embedding space and are frequently used in image-text matching problems. In image classification with large-number labels, the Deep Visual Semantic Embedding (DeViSE) model [16] outperforms conventional one-way models since the latter treats all labels as disconnected and unrelated. The DeVisE model uses a pre-trained word2vec [17] model to embed words into vectors with semantic information preserved such as synonymy and a pre-trained visual model to extract feature vectors from input images. Classifying an input image is to assign the most relevant label to the image based on the similarities between the image and labels, which is measured by a generalized dot product of the visual feature vectors and concept embedding vectors where the metric matrix is trainable.

Visual-semantic embedding models are extended to sentence-level problems such as image description generation in [18], [19] using RNNs, and are used in zero- and few-shot learning due to the continuity of visual space and the use of unannotated text in [20]. In [21], the DeViSE model is extended for MLC, where the compatibility of an image and a label is represented by the embedding vectors projected by matrices of visual features extracted from the images. However, this model predicts only the k most possible labels from the highest k matching scores and does not provide predictions on the entire set of labels. In summary, all the above visual-semantic embedding models require a large corpus to learn semantic embeddings of the label texts from semantic relations and syntactic components. Furthermore, these semantic embeddings are independent of the final tasks, thus the performance of the model relies entirely on the training of the visual part, while the concept part is not involved in improving performance. On the other hand, BotCL [22] learns a visual-semantic concept bottleneck in SENN [23] with nonsemantic embeddings of implicit concepts where concept embeddings are learnable during training. However, due to self-supervised implicit concepts in SENN that are not understandable to humans, BotCL lacks credibility in comparison to CBM and does not support intervention.

III. The Proposed ViP-CBM Methods

A. Motivation

To address the challenge of massive parameters and low interpretability in MLC of concepts, we propose ViP-CBM to transfer the MLC problem into a binary classification problem

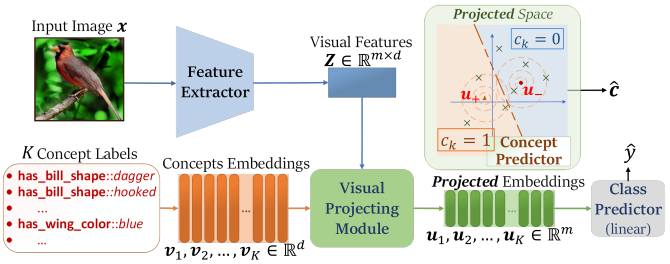


Fig. 2. Model structure of our ViP-CBM.

of all concept embeddings projected by the visual features of each input image individually. We modify the model in [21] and propose the ViP module to predict all concepts where embeddings are trained simultaneously with visual features.

Our main idea is as follows: The MLC problem in concept prediction can be transformed into finding bisections of the concept label set for different inputs, which is equivalent to finding binary classifications of concept embedding vectors for different input images in the concept embedding space. With a well-defined input-dependent projection that projects different classification surfaces for each input image into a unified hyperplane, the original concept bottleneck layer, which individually predicts each concept as positive or negative, can be transferred into a unified binary classifier for projected concept embedding vectors. By constructing a projection function based on the visual features, we can build a concept bottleneck model that predicts concepts from visual-projected concept embeddings. Therefore, our ViP-CBM leverages both visual and semantic information from the data, where both visual and concept representations are learnable.

As a variant of CBM, our ViP-CBM requires a fully supervised dataset denoted as $\mathcal{D} = \left\{\mathbf{x}^{(i)}, \mathbf{c}^{(i)}, y^{(i)}\right\}_{j=1}^{N}$ with N data points, K binary concepts and M classes, where the i-th data point consists of the input $\mathbf{x}^{(i)} \in \mathcal{X}$, the concepts $\mathbf{c}^{(j)} \in \{0,1\}^K$ and the label $y^{(j)} \in \{1, \ldots, M\}$.

B. Model Structure

Figure 2 shows the general prediction flow of our ViP-CBM. Our ViP-CBM includes concept embeddings the visual projecting (ViP) module, the concept predictor, and the task predictor, which we will introduce one by one.

Concept embeddings and the ViP module. For concept embedding, we embed all K concepts $\{c_1, \ldots, c_K\}$ in $\mathbf{R}^d$ space as $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$. Given an input image $\mathbf{x}$, a backbone CNN $\phi(\cdot)$ extracts visual features as a matrix $\mathbf{Z} = \phi(\mathbf{x}) \in \mathbf{R}^{m \times d}$. We use a nonlinear projection $\mathbf{u}_k = \theta(\mathbf{Z}\mathbf{v}_k) \in \mathbf{R}^m, k = 1, \ldots, K$, by adding a nonparametric nonlinear function $\theta(\cdot)$ on a simple linear projection. In this work, we propose the nonlinear projection as

$$u_{k,j} = ReLU\left(\mathbf{z}_j^T \mathbf{v}_k + \frac{\mathbf{z}_j^T}{\|\mathbf{z}_{j,\cdot}\|_2}\mathbf{v}_k\right), j = 1, \ldots, m, \quad (1)$$

where $\mathbf{z}_j \in \mathbf{R}^d, j = 1, \ldots, m$ is the j-th row of the matrix $\mathbf{Z}$ and $u_{k,j}$ denotes the j-th element of $\mathbf{u}_k$. This projection function uses ReLU as the activation function and introduces "normalized linear projections", which is the latter term of the inputs of the ReLU function, to increase nonlinearity.

Concept predictor. ViP-CBM converts MLC in concept prediction into a unified binary classification in projected concept embeddings $\mathbf{u}_1, \ldots, \mathbf{u}_K$. We propose a linear classifier with a pair of anchor points for the concept predictor. Define a pair of trainable anchor points $\mathbf{u}_+, \mathbf{u}_- \in \mathbf{R}^m$ and classify the concept c_k according to the Euclidean distance between its corresponding projected embedding $\mathbf{u}_k$ and the two anchors, which adds nonlinearity to a linear model, with projections of positive concepts closer to $\mathbf{u}_+$ and negative concepts closer to $\mathbf{u}_-$. Inspired by Triplet Loss in [24], the probability of the concept c_k being activated is

$$p(\hat{c}_k = 1 | \mathbf{Z}, \{\mathbf{v}_i\}_{i=1}^K) = \sigma(a(\|\mathbf{u}_k - \mathbf{u}_-\|_2 - \|\mathbf{u}_k - \mathbf{u}_+\|_2 - m_{c_k})), \quad (2)$$

where $\sigma(\cdot)$ represents the sigmoid function, $a > 0$ is a learnable scaling parameter, $m_{c_k} \geq 0$ is an optional decision margin depending on true label c_k. For example, we can set m_{c_k} to a positive constant m to penalize false positives. To encourage the model to project both activated and inactivated concepts closer to the corresponding anchors for inputs from each class, we set the margins as

$$m_{c_k} = \mathbf{1}_{(c_k=1)} m + \mathbf{1}_{(c_k=0)}(-m), m > 0, \quad (3)$$

where $\mathbf{1}_{(\cdot)}$ is the instructional function.

Class predictor. To reduce the model parameter set to the minimum, we use a simple linear classifier as the task predictor to predict the M final classes. Different from conventional CBMs, we employ the K projected concept embeddings $[\mathbf{u}_1, \ldots, \mathbf{u}_K]$ as inputs of the class predictor instead of concept predictions $p(\hat{c}_k = 1 | \mathbf{Z}, \{\mathbf{v}_i\}_{i=1}^K)$, to improve model's performance and smooth the training process by leveraging the richer information in embeddings.

Training strategy and loss function. Since we calculate concept probabilities and task probabilities, we apply Binary Cross-Entropy (BCE) loss to concept prediction and Cross-Entropy (CE) loss to class label prediction. In this work, we employ the *joint* CBM training strategy, which is to train both concepts and labels simultaneously by minimizing a weighted sum of the two losses:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{x},\mathbf{c},y)}\left[\mathcal{L}_{CE}(y, \hat{y}) + \alpha \mathcal{L}_{BCE}(\mathbf{c}, \hat{\mathbf{c}})\right]. \quad (4)$$

C. Parameter Reduction in ViP-CBM

Our ViP-CEM reduces parameters mainly in concept prediction. Consider a minimal CBM with the same backbone CNN as our ViP-CBM model, binary concepts are predicted from extracted visual features $\mathbf{Z} \in \mathbf{R}^{m \times d}$ by a simple linear model:

$$p(\hat{c}_k = 1 | \mathbf{Z}) = \sigma\left(\mathbf{w}^T \bar{\mathbf{z}} + b\right), \quad (5)$$

where $\bar{\mathbf{z}} \in \mathbf{R}^{md}$ represents the flattened matrix $\mathbf{Z}$, and the parameters of the linear model are $(\mathbf{w}, b) \in \mathbf{R}^{Kmd} \times \mathbf{R}$. In comparison, the parameters in the concept predictor of our ViP-CEM are $(\{\mathbf{v}_i\}_{i=1}^K, \mathbf{u}_+, \mathbf{u}_-) \in \mathbf{R}^{K \times d} \times \mathbf{R}^m \times \mathbf{R}^m$. Therefore, the number of training parameters in the concept predictor in our ViP-CBM is approximately m times less than that in the minimal CBM.

In addition, the concept predictor in (2) without margins is equivalent to a linear classifier in the final decision. Thus, we can rewrite the concept predicting step with a linear classifier of the form $p(c_k = 1|\mathbf{u}_k) = \tilde{\mathbf{w}}^T\mathbf{u}_k + \tilde{b}, \tilde{\mathbf{w}} \in \mathbf{R}^m, \tilde{b} \in \mathbf{R}$ as follows, neglecting the nonlinearity in projection:

$$p\left(\hat{c}_k = 1|\mathbf{Z}, \{\mathbf{v}_i\}_{i=1}^K\right) = \sigma(\tilde{\mathbf{w}}^T(\mathbf{Z}\mathbf{v}_k) + \tilde{b}) = \sigma(\text{tr}[(\mathbf{v}_k\tilde{\mathbf{w}}^T)\mathbf{Z}] + \tilde{b}), \quad (6)$$

which is similar to (5) with the weight matrix $\mathbf{v}_k\tilde{\mathbf{w}}^T$ of rank 1. This reveals that our model uses a rank-1 classifier for concept prediction while using the two anchors to encourage the separation of the positive and negative samples, and the nonlinear function $\theta(\cdot)$ to preserve the performance.

When using projected embeddings for class predicting instead of probabilities, the number of parameters of the class predictor in our ViP-CBM is m times that in the minimal CBM that predicts classes from scalars, which makes the total number of training parameters comparable to the minimal *scalar* CBM. Nevertheless, due to the parameter reduction in concept predicting, our ViP-CBM has only less than half as many training parameters as the minimal CEM and ProbCBM.

IV. Experimental Setup

A. Datasets

CUB-200-2011. The Caltech-UCSD Birds-200-2011 (CUB-200-2011) [25] dataset contains 11,788 images of 200 subcategories belonging to birds annotated with 312 binary concepts. We use the preprocessed dataset in [6] where the number of concepts is reduced to $K = 112$ and concepts are denoised to class-level, which means images from the same class share the same concept annotations. Concept labels in CUB-200-2011 are of the form *"{general_concept}::{detail}"*, so that we can naturally group concepts into 28 groups based on the general concepts, with the largest group having 6 concepts.

AwA2. Animals with Attributes 2 (AwA2) [26] dataset contains 37,322 images of 50 categories of animals with 85 binary attributes, e.g., color, stripe, etc. AwA2 provides a category-attribute matrix that contains concept labels for each category so that concepts are also class-level. We artificially summarize these 85 concepts into 30 groups of color, pattern, habit, etc., with the largest group having 14 concepts.

B. Experimental Setup

Data augmentation. For data augmentation, we first perform color jittering and random horizontal flipping on the images, then resize them to 256×256. We randomly crop the images with a scale of (0.8, 1.0) and resize images from CUB to 224×224 and images from AwA2 to 256×256.

ViP-CBM model settings. The visual feature extractor of our ViP-CBM is shown in Figure 3. We use a ResNet34 [27] pre-trained on ImageNet-1k [28] as the backbone and extract outputs of the layer before the global average pooling with a size of (512, l, l) where l denotes the side length of the feature map. We then use a 1×1 convolution layer with d channels and flatten the outputs to get a feature representation of size (d, l^2), where d is the dimension of concept embeddings. Then we

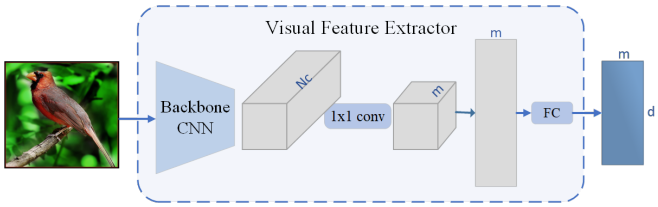


Fig. 3. Detailed structure of the visual feature extractor in ViP-CEM.

use a Fully Connected (FC) layer to reduce the l^2-entry inputs to m entries to get the feature matrix $\mathbf{Z} \in \mathbf{R}^{m \times d}$. The settings of the concept and class predictor follow the descriptions in Section III-B. We set symmetric margins following (3) with $m = 0.1$ to test the effects of margins.

Baselines. We compare our ViP-CBM to conventional joint CBM, CEM, and ProbCBM with equivalent parameters. For all baseline models, we use the same pre-trained ResNet34 and 1×1 d-channel convolution layer to get a feature representation of size (d, 7, 7) and flatten it to a $49d$ dimensional vector. For baseline CBM, we use a 2-layer MLP for the concept predictor with a hidden layer size of 128 and apply ReLU as the activation function. We use a linear model to predict class labels directly from concept predictions as the class predictor.

For baseline CEM, we use a simple linear model to predict the positive and negative d-dimensional embeddings of each concept to align the dimensions with our model. We use a shared scoring function and linear class predictors as in [14].

For baseline ProbCBM, we use a linear model to generate K visual embeddings of d dimensions from the original features of size (d, l, l), and learn K pairs of positive and negative anchors for concept prediction in the embedding space of d dimensions. since our only concern is the model's performance, we omit the sampling step and refine the class predictor to a simple linear model where the inputs are the K visual embeddings for this work.

Hyperparameters settings. We set the weight between the two losses $\alpha = 5$. We use an SGD optimizer with a learning rate of 0.01 for the CUB dataset and 0.002 for the AwA2 dataset, both with momentum of 0.9 and weight decay of 5×10^{-4}. We train 400 epochs on the training split and compare model performance on the test split for the CUB dataset, and 250 epochs for the training split of the AwA2 dataset.

C. Metrics

We use class accuracy as the criterion of the model's task performance. Denote the k-th concept prediction of the i-th sample as $c_k^{(i)} \in \{0, 1\}$. For the MLC of concepts, we define two metrics as follows.

- To evaluate the model's accuracy on each concept individually, we use the **Hamming score (HS)**:

$$HS = \sum_{i=1}^N \sum_{k=1}^K \mathbf{1}_{(\hat{c}_k^{(i)} = c_k^{(i)})}/NC. \quad (7)$$

- To evaluate the model's ability to predict *all* concepts correctly, we use the **exact match ratio (EMR)**:

$$EMR = \sum_{i=1}^N \mathbf{1}_{(\hat{\mathbf{c}}^{(i)} = \mathbf{c}^{(i)})}/N. \quad (8)$$

TABLE I
PERFORMANCE OF OUR VIP-CBM IN COMPARISON WITH OTHER BASELINE CBMS FOR CUB AND AWA2 DATASETS

Model	CUB				AwA2			
	Hamming Score	Overall EMR	Min Group EMR	Class Accuracy	Hamming Score	overall EMR	Min Group EMR	Class Accuracy
scalar-CBM	**0.9529**±0.0004	**0.4270**±0.0069	**0.7819**±0.0027	0.7197±0.0033	**0.9708**±0.0006	0.7816±0.0040	0.8288±0.0020	0.8782±0.0046
CEM	0.9506±0.0003	0.3887±0.0078	0.7651±0.0026	0.7186±0.0048	0.9696±0.0007	0.7646±0.0035	0.8288±0.0020	0.8794±0.0028
ProbCBM	0.9517±0.0010	0.4032±0.0132	0.7743±0.0058	**0.7268**±0.0113	0.9704±0.0008	0.7798±0.0084	0.8366±0.0049	**0.8822**±0.0026
ViP-CBM (ours)	0.9496±0.0009	0.3784±0.0223	0.7646±0.0037	0.7169±0.0048	0.9702±0.0010	0.7857±0.0055	0.8403±0.0047	0.8818±0.0031
+margin	0.9500±0.0013	0.3898±0.0090	0.7670±0.0053	0.7115±0.0075	0.9701±0.0009	**0.7906**±0.0071	**0.8431**±0.0047	0.8807±0.0037
LP	0.9398±0.0014*	0.3647±0.0148*	0.7248±0.0337*	0.6462±0.0463*	0.9692±0.0006	0.7741±0.0037	0.8335±0.0050	0.8801±0.0024

TABLE II
COMPARISON OF THE NUMBER OF TRAINING PARAMTERS

Model	scalar-CBM	CEM	ProbCBM	ViP-CBM (ours)
Training Params #	254296	744681	746185	289625

We can also use **group EMR** to evaluate the model's performance on each concept group. In summary, we use the Hamming score to measure the individual concept accuracy, EMR for all concepts to represent the overall concept accuracy, and the minimum group EMR to evaluate concept prediction in the hardest group.

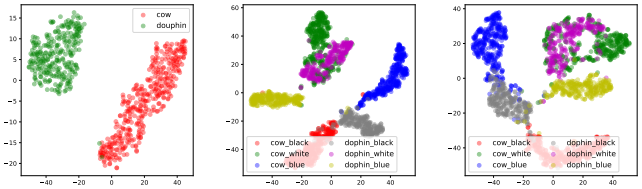
V. EXPERIMENTAL RESULTS

A. Model Performance

We set embedding dimensions $d = 32$ and the dimension of the projected space $m = 12$ for the experiments. We add additional ablation studies of margins and nonlinearity, denoting ViP-CBM using margins as "+margin" and ViP-CBM using linear projection $\mathbf{u}_k = \mathbf{Z}\mathbf{v}_k$ in ViP module as "LP". We conduct experiments with 5 different random seeds on the CUB dataset and 4 on the AwA2 dataset to compute the average scores and standard errors, marked as "mean±std" in our results. Table II shows that our model has only 40% of the training parameters of other *embedding*-based CBM, which is comparable to the scalar-CBM.

Table I shows the performance in concept and class prediction of our ViP-CBM and other baseline models. For each metric, we mark the highest score in **bold**, the second highest in purple, and the third highest in cerulean. Note that we directly use the output of the backbone CNN of size (d, l, l) as the visual features for the baseline models as described in Section IV-B, which is larger than the visual features used in our ViP-CBM, suggesting that we are comparing with larger CBMs than we proposed in Section III-C.

For the experiments on the CUB dataset with a relatively large concept set and a small amount of data, due to the reduction of parameters and rank in concept prediction, our ViP-CBM model underperforms CBM significantly in overall concept accuracy but is comparable to CBM in individual concept accuracy and class accuracy with a loss of less than 0.003. Our ViP-CBM achieves comparable or slightly superior performance to CEM with low embedding dimensions and slightly inferior performance to ProbCBM by less than 0.015 in overall concept accuracy and class accuracy. Compared to CEM which learns $2K$ embeddings in total for each input image and ProbCBM which requires $2K$ concept anchors in total with K individually extracted visual features of d dimensions to embed in each concept space, our ViP-CBM learns only K concept embeddings independent to inputs and the dimension of projected space m is much smaller than K. Thus, our ViP-CBM improves the efficiency of concept learning with concept representations of the same dimensions. For the experiments on the AwA2 dataset with fewer concept and class labels and a larger amount of data, our ViP-CBM model ranks third in individual concept accuracy and second in all other metrics, and outperforms CEM by over 0.02 in concept overall accuracy and over 0.002 in class accuracy.



(a) Visual features $\mathbf{Z}$. (b) $\mathbf{u} = \mathbf{Z}\mathbf{v}$. (c) $\mathbf{u} = \theta(\mathbf{Z}\mathbf{v})$.

Fig. 4. t-SNE plot of the original visual features $\mathbf{Z}$ and concept representations $\mathbf{u}$ by both linear and nonlinear projections.

The ablation study for margins shows that symmetric margins in ViP-CBM enhance the overall accuracy for concepts and the stability for different initializations, consistent with our intent to penalize projected embeddings close to the classifying surface. In the ablation study of nonlinearity in visual projection, we mark a "*" on some of the results of the "LP" model to show that more than half of the experiment failed due to gradient explosion in training, indicating that nonlinearity in visual projection also allows larger learning rates, which increase the convergence speed and stability in training. Besides, the results of the successful experiments also prove that nonlinearity is necessary to bridge the performance gap due to the reduction of parameter numbers.

B. Interpretability

To reveal the interpretability of our ViP module proposed for MLC on a certain label set, we look into the spatial distributions of the projected concept embeddings. With the hypothesis of the continuity of visual feature space, the projected embeddings should cluster by class regardless of concept labels. For AwA2 dataset with class-level concepts, we select 3 concepts *"black", "white" and "blue"* and 2 classes *"cow"* and *"dophin"*, where *"black"* and *"white"* are positive for *"cow"* and *"white"* and *"blue"* are positive for *"dophin"*. We visualize the spatial distributions of the original visual features $\mathbf{Z}$ and the linear and nonlinear projected embeddings of the 3 concepts $\mathbf{u}$ by visual features of images from both 2 classes

in the test dataset using a 2-dimensional t-SNE [29] plot, as is shown Figure 4. Original visual features are clearly clustered in Figure 4a, and so are projected embeddings in Figures 4b and 4c. In Figures 4b and 4c, projected embeddings of each concept for inputs in the same class (represented by points with labels in each column of the legend) are clearly separated from each other, indicating the separation in original concept embeddings $\mathbf{v}$. Clusters of common activated concept *"white"* overlap significantly, indicating that the ViP module gathers the same concepts in the projected space for all input images containing the concept. Furthermore, activated and inactivated concept embeddings are much more separated in Figure 4c than in Figure 4b, which reveals that nonlinear projection enhances performance in concept prediction.

VI. Conclusions

We propose ViP-CBM to reduce the number of parameters and enhance interpretability in the MLC problem in concept learning for CBMs. Experiments show that our ViP-CBM, whose number of training parameters is comparable to a minimal scalar CBM, achieves comparable performance to conventional CBMs, and outperforms CEM in concept learning with low embedding dimensions. The interpretability of ViP-CBM is also shown in experiments by visualizing the projected space. Our ViP-CBM is a low-parameter substitution for embedding-based CBMs with more interpretability.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should i trust you?' explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[3] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[4] Z. Chen, Y. Bei, and C. Rudin, "Concept whitening for interpretable image recognition," *Nature Machine Intelligence*, vol. 2, no. 12, pp. 772–782, 2020.

[5] W. Pan and C. Zhang, "The definitions of interpretability and learning of interpretable models," *arXiv preprint arXiv:2105.14171*, 2021.

[6] P. W. Koh, T. Nguyen, Y. S. Tang, *et al.*, "Concept bottleneck models," in *International conference on machine learning*, PMLR, 2020, pp. 5338–5348.

[7] M. Havasi, S. Parbhoo, and F. Doshi-Velez, "Addressing leakage in concept bottleneck models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 386–23 397, 2022.

[8] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, pp. 333–359, 2011.

[9] X. Xu, Y. Qin, L. Mi, H. Wang, and X. Li, "Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations," in *The Twelfth International Conference on Learning Representations*, 2024.

[10] I. Sheth and S. Ebrahimi Kahou, "Auxiliary losses for learning generalizable concept-based models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[11] M. Yuksekgonul, M. Wang, and J. Zou, "Post-hoc concept bottleneck models," *arXiv preprint arXiv:2205.15480*, 2022.

[12] T. Oikarinen, S. Das, L. M. Nguyen, and T.-W. Weng, "Label-free concept bottleneck models," *arXiv preprint arXiv:2304.06129*, 2023.

[13] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[14] M. E. Zarlenga, P. Barbiero, G. Ciravegna, *et al.*, "Concept embedding models," in *NeurIPS 2022-36th Conference on Neural Information Processing Systems*, 2022.

[15] E. Kim, D. Jung, S. Park, S. Kim, and S. Yoon, "Probabilistic concept bottleneck models," *arXiv preprint arXiv:2306.01574*, 2023.

[16] A. Frome, G. S. Corrado, J. Shlens, *et al.*, "Devise: A deep visual-semantic embedding model," *Advances in neural information processing systems*, vol. 26, 2013.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[18] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[19] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[20] L. Chen, H. Zhang, J. Xiao, W. Liu, and S.-F. Chang, "Zero-shot visual recognition using semantics-preserving adversarial embedding networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1043–1052.

[21] M.-C. Yeh and Y.-N. Li, "Multilabel deep visual-semantic embedding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 6, pp. 1530–1536, 2019.

[22] B. Wang, L. Li, Y. Nakashima, and H. Nagahara, "Learning bottleneck concepts in image classification," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2023, pp. 10 962–10 971.

[23] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[25] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[26] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2019. DOI: 10.1109/TPAMI.2018.2857768.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[29] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

BEES: A New Acoustic Task for Blended Emotion Estimation in Speech

Xingfeng Li[1], Xiaohan Shi[2], Yuke Si[3], Zilong Zhang[4], Feifei Cui[5],
Yongwei Li[6], Yang Liu[7], Masashi Unoki[8], Masato Akagi[9]
[1] City University of Macau, Macau. E-mail: xfli@cityu.edu.mo
[2] Nagoya University, Nagoya. E-mail: xiaohan.shi@g.sp.m.is.nagoya-u.ac.jp
[3] China Mobile Research Institute, Beijing. E-mail: siyuke@163.com
[4, 5] Hainan University, Haikou. E-mail: {zhangzilong, feifeicui} @hainanu.edu.cn
[6] Institute of Psychology, Chinese Academy of Sciences, Beijing. E-mail: liyw@psych.ac.cn
[7] Qingdao University of Science and Technology, Qingdao. E-mail: yangliu@qust.edu.cn
[8, 9] Japan Advanced Institute of Science and Technology, Nomi. E-mail: {unoki, akagi} @jaist.ac.jp

***Abstract*—Acoustic emotion recognition (AER) is a challenging yet crucial area in affective computing. Despite the clear importance of perceiving vocal emotion blends with varying intensities for understanding the emotional complexity of everyday life, most existing studies remain limited to single-label emotion classification tasks. This paper introduces a novel acoustic task for blended emotion estimation in speech (BEES), which aims to capture the intricate emotional distributions of co-existing emotions with varied intensities in voices. The BEES task stands out by defining a precise reference for the emotion intensity of vocal blends, introducing innovative evaluation metrics and providing robust baseline results. Additionally, our study highlights the significant impact of self-supervised learning representations, voice naturalness, and gender on BEES performance. By addressing the limitations of traditional AER methods and offering a comprehensive framework for analyzing blended emotions, this work paves the way for advanced computational AER systems capable of handling the complexity of vocal emotion blends.**

I. Introduction

Emotion is a psychological state brought on by neurophysiological changes, variously associated with thoughts, feelings, and behavioral responses [1]. It can function to communicate positively important information to individuals in significant external events, such as values and ethics [2]. Also, it is sometimes internally regarded as part of a mental illness and thus possibly of negative value, for instance, anxiety or depression [3]. There is widespread evidence supporting that emotion is of vital importance for social competence, decision-making, and well-being [4]. Within its research area of interest, acoustic emotion recognition (AER) aims to give machines emotional intelligence underlying how humans feel emotionally towards their voice has drawn great attention and could be beneficial to wide applications such as customer care service, human-machine interaction, and many others [5], [6].

Most of the existing works toward AER can be regarded as a single-label classification task, assuming that each voice stimulus only evokes a dominant emotion [7]–[9]. Despite substantial progress made in this area, these works however over-simplify the complexity of human emotions and neglect the ambiguity and subjectivity that lie in them. Historically and today, advances in psychology research substantiate the existence of the expression and perception of emotion-blends, indicating that humans can frequently experience more than one emotion at the same time [10]–[13]. For instance, Juslin et al. (2021) conducted a study capturing naturalistic vocal expressions in a field setting [14]. Speakers themselves provided the ground truth of emotional expressions in each recording, revealing that 41% of the recordings depicted situations where blended emotions were experienced. Also, other nonverbal communication research has further investigated various types of blended emotional experiences such as the co-activation of happiness and sadness [13], fear and happiness [15], disgust and amusement [16], as well as hope and fear [17]. It is grounded that treating AER as a single-label classification problem is oversimplified and lacks practical applicability.

To address acoustical emotion-blends estimation, multi-label learning has been extensively investigated, which usually selects a threshold for the output of a classifier, then labels emotions with scores higher than the threshold as existing and the others as not existing [18], [19]. Unfortunately, these methods failed to estimate the intensity of each specific emotion. On the other hand, recent advancements in machine learning suggest the existence of label distribution learning which can be promising to represent the degree to which each label describes the instance [20], [21]. For any instance, the sum of the description degrees of all labels is one, indicating a full description of this instance. Inspired by the above studies, we first introduce an acoustic task for BEES, assuming each voice stimulus contains a blend of multiple emotions with varying intensities. Specifically, we label each stimulus by an emotion vector, where each element corresponds to a specific emotion and the value of each element is the intensity of that emotion. In this context, the emotion vectors can be interpreted as emotion distributions, and BEES aims to map human voices to their corresponding emotion distributions by maximizing the similarities between distributions obtained by human evaluations and system estimations.

To our knowledge, this paper takes one step beyond current AER algorithms and is the first attempt at BEES to investigate machine learning-based emotion-blends estimation in human voices. The main contributions of this paper are multi-fold: (1) we introduce an important human assessment criterion to produce blended emotion references in voice stimulus, (2) we identify efficient evaluation metrics to assess the BEES performance, (3) we discuss the effects of acoustic features, emotion production styles, and genders on the BEES performance. The remainder of this article is structured as follows: Section 2 outlines the utilized data and human assessment methodology for identifying blended emotion references in human voices. Section 3 presents the foundational acoustic features of self-supervised learning representations (SSLs), estimation methodologies, and evaluation metrics applied to the tasks of BEES. Section 4 provides an elaborate breakdown of the BEES results along with corresponding discussions. Finally, Section 5 presents pivotal conclusions drawn from this study.

II. Blended Emotion Estimation in Speech

A. Emotional Speech Database

We use the popular IEMOCAP corpus for this study for the following four reasons [22]. First, the evaluators annotating the IEMOCAP corpus were allowed to tag more than one emotional category per speech stimulus, to account for mixtures of emotions such as frustration and anger, etc, which are frequently observed in real-life scenarios. Secondly, IEMOCAP defines N-category labels by carefully balancing the trade-off between the number of emotion categories, providing a more accurate and detailed emotional description with higher agreement among evaluators. It contains nine categorical labels of emotion, including anger (ang.), sadness (sad.), happiness (hap.), disgust (dis.), fear (fea.), and surprise (sur.) which are known as basic emotions, plus frustration (fru.), excited (exc.), and neutral (neu.) states. Third, this corpus was deliberately chosen to contain two different emotion production styles. The spontaneous speech subset supports application-oriented research on authentic emotions, while the acted speech subset allows for state-of-the-art emotion categorization comparisons. Fourth, in contrast to existing emotional databases mainly contain only isolated sentences or short dialogs, without taking into account the discourse context, which is known to be an important component. The IEMOCAP corpus was designed from a dialog perspective, eliciting sequential emotions with adequate context. Most interestingly, from an application point of view, this corpus is well-suited for studying the dynamic progression of blended emotions, enabling the detection of when and how a user's blended affective state changes, which can contribute to improving human-computer interactions. This corpus has 12 hours of speech data from ten subjects and is pre-segmented into shortcuts, resulting in a total of 10,039 utterances. Three different evaluators assessed each utterance using the aforementioned nine categorical labels of emotion. It is worth noting that only 22.8% of the utterances in the IEMOCAP corpus are pure examples of a single perception category. Significantly, the remaining 77.2% of utterances present blends of different emotions, which in turn confirms its capability to approach BEES.

B. Task Definition

In contrast to the majority of existing studies in AER, BEES takes a further step by estimating the intensity of each emotion that may concurrently exist in an individual voice stimulus. Different measures have been developed to capture the simultaneity and the subjective experience of emotion-blends [11], [23]. We start by giving a more formal definition of emotional intensity following [24], [25], namely inferring the concurrent experience and intensity of blended emotions by quantifying the frequency with which individuals experience each measured emotion.

Formally, let emoMat be a matrix of experienced and tagged emotional labels for the nth voice stimulus given by three evaluators.

$$\text{emoMat} = \begin{bmatrix} \epsilon^n_{1,1} & \epsilon^n_{1,2} & \cdots & \epsilon^n_{1,\lambda} \\ \epsilon^n_{2,1} & \epsilon^n_{2,2} & \cdots & \epsilon^n_{2,\lambda} \\ \epsilon^n_{3,1} & \epsilon^n_{3,2} & \cdots & \epsilon^n_{3,\lambda} \end{bmatrix} \tag{1}$$

More specifically, each row of the emoMat is correspondingly given by the ith evaluator and λ is the number of emotion classes. Each element of the emoMat is defined as:

$$\epsilon^{\text{n}}_{\text{i},\lambda} = \begin{cases} 1, & \text{if a specific emotion is experienced,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The BEES then calculates the ρ^n_λ as a reference to the emotional vector of the intensity of a specific categorical label λ in the nth voice stimulus as follows:

$$\rho^{\text{n}}_\lambda = \frac{\sum \text{emoMat}\,(:,\lambda)}{\sum_i \sum_\lambda \text{emoMat}_{\text{i},\lambda}} \tag{3}$$

Figure 1 shows an example of the ρ_λ relative to two voice stimuli in the IEMOCAP corpus. The X-axis represents the

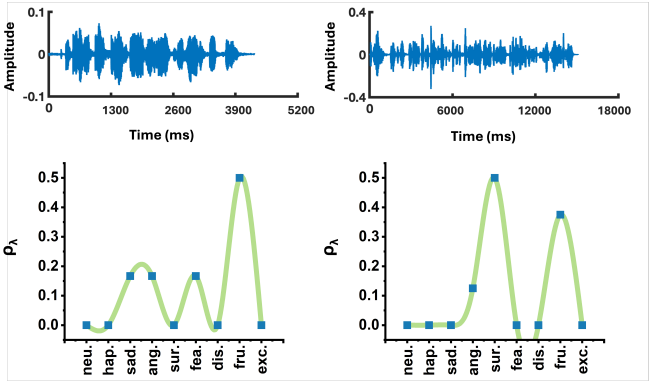


Fig. 1: Two voice stimulus examples with different ρ_λ from the IEMOCAP corpus (left: Ses02M_impro02_F011.wav; right: Ses02F_impro05_F005.wav). Rather than a dominant emotion, human voices often evoke multiple emotions with different perception intensities.

nine kinds of emotions and the Y-axis represents the ρ_λ of each kind of emotion. Note that ρ_λ^n denotes the proportion that λ accounts for in a full emotion distribution of the nth voice stimulus. This differs from the probability of λ being the correct emotion label for the nth voice stimulus. Probability distribution implies that only one emotion label is correct for each sentence, whereas BEES allows for the possibility of multiple emotions co-exist within a single voice.

Specifically, the goal of BEES to identify the component of intensity in terms of each specific emotion λ, ρ_λ^n, can be formulated by computing the optimal parameter θ^* via solving the problem as follows:

$$\theta^* = \text{argmin}_\theta \sum^{N} \text{Dist}(\rho_\lambda^n, \hat{\rho_\lambda^n}) \tag{4}$$

where Dist$(\cdot)$ is the distance function measuring the similarity of the predicted emotion intensity, $\hat{\rho_\lambda^n}$, and the ground truth intensity ρ_λ^n. Additionally, N is the total number of voice stimuli in the IEMOCAP corpus.

III. Baseline Approaches

The overall architecture of BEES is shown in Fig. 2, starting from the human voices and ending with its emotion distribution estimation. The subsequent sections detail the BEES as follows: Section 3.1 introduces the acoustic features that we used to represent voices. Section 3.2 defines a deep neural network-based baseline model to estimate emotion intensity distribution. Section 3.3 provides the metrics to evaluate the BEES performance.

A. Acoustic Features

An important issue in the design of a BEES system is the extraction of suitable features that best reflect different emotions and their intensities and should be robust against other speech diversities of speaking styles, speakers, genders, etc. While deep learning started in this field in the early 2010s, a widespread trend exists to explore the neural representation of acoustics [26], [27]. Within its areas of interest, SSLs have recently gained considerable attention due to their advances in capturing many acoustic characteristics comprehensively and reliably [26], [28]. The importance of the SSLs of speech is increasingly evident in AER, and also many others, like speech

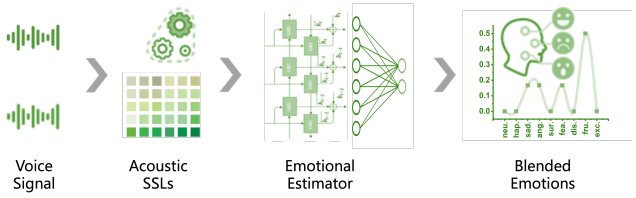


Fig. 2: Illustration of the BEES system. Given a voice evaluated with emotional intensities, our approach represents the voice with a vector matrix using SSLs and then employs a deep neural network-based estimator on it. The mean square error loss is applied to blended emotion estimation.

TABLE I: The BEES results in terms of MCS values to the emoMat references for the overall evaluated emotion intensity obtained by using different SSLs. The top performance is highlighted, and ↑ indicates "the larger the better".

Features	MCS ↑
Data2vec	0.7190
Unispeech	0.6886
HuBERT	0.7087
Wav2vec 2.0	0.6888
WavLM	**0.7303**

recognition [29], speaker diarization [30], gender recognition [31], and more [32], [33]. Five SSLs applied to BEES are thus explored in this work including Wav2vec 2.0[1], HuBERT[2], WavLM[3], Unispeech[4], and Data2vec[5] based on their top performance in well-known benchmarks such as the Speech processing Universal PERformance baseline (SUPERB) [34], and the Holistic Evaluation of Audio Representations [35], as well as recent SER studies [36], [37]. Given that the larger models in general deliver superior performances, achieving state-of-the-art scores on SUPERB tasks [34]. This study therefore approaches MER using large Wav2vec 2.0, HuBERT, WavLM, Unispeech, and Data2vec models, where the last hidden states are transformed into 1024-dimensional vectors for individual speech. For consistent input audio processing, all models are uniformly sampled at 16kHz.

B. Model Configuration

For all SSLs, we employed the same deep neural network-based estimator. We conducted experiments using a hybrid model by combining a bidirectional gated recurrent unit (BiGRU) with a deep neural network. To be more explicit, the BiGRU is utilized to process the acoustic features of SSLs in both forward and backward directions, it could capture contextual information effectively, which is crucial for understanding the dynamic nature of vocal emotion. The bidirectional nature of the BiGRU exhibits robustness to variations in voice stimuli, including differences in speaking rate, accent, and pronunciation [38], [39], enabling it to adapt to diverse speaking styles and linguistic nuances inherent in different speakers. This BiGRU is set to two layers with 256 units and a dropout rate of 0.5. In the training process, the optimizer is set to the Adam optimizer with a learning rate of 0.0001. In particular, this study follows [40], which uses a global average pooling layer instead of a fully connected layer to obtain the SER results. Most notably, this study uses the loss function of mean square error for BEES tasks. Experiments are conducted to estimate blended emotional states via leave-one-speaker-out cross-validation, where we use all utterances of each speaker once as the test set and each time use the utterances of their pair as the validation set.

[1]https://huggingface.co/facebook/wav2vec2-large
[2]https://huggingface.co/facebook/hubert-large-ls960-ft
[3]https://huggingface.co/microsoft/wavlm-large
[4]https://huggingface.co/microsoft/unispeech-sat-large
[5]https://huggingface.co/facebook/data2vec-audio-large-960h

TABLE II: The BEES results in terms of CCC and MAE values to references for the evaluated emotion intensity considering each specific emotional state in the emoMat obtained by different SSLs. ↑ indicates "the larger the better" and indicates ↓ "the smaller the better".

Features	ang.	dis.	exc.	fea.	fru.	hap.	neu.	sad.	sur.
CCC ↑									
Data2vec	0.6581	0.0140	0.5465	0.0487	0.4305	0.4575	0.4250	**0.6839**	0.3815
Unispeech	0.5633	0.0304	0.3900	**0.1952**	0.2106	0.2498	0.2933	0.5519	0.2158
HuBERT	0.6592	0.0084	0.5573	0.0044	0.4520	0.4437	0.4350	0.6220	0.0477
Wav2vec 2.0	0.6181	0.0213	0.4631	0.0145	0.3724	0.3473	0.3758	0.6144	0.3166
WavLM	**0.6788**	**0.0682**	**0.5976**	0.0167	**0.5028**	**0.5188**	**0.4764**	0.6651	**0.4327**
MAE ↓									
Data2vec	0.1130	**0.0060**	0.1163	0.0223	0.2251	**0.0928**	0.1866	**0.0950**	**0.0224**
Unispeech	0.1438	0.0245	0.1413	**0.0157**	0.2382	0.1184	0.2120	0.1111	0.0268
HuBERT	0.1206	0.0133	0.1201	0.0223	0.2108	0.1003	0.1789	0.1027	0.0437
Wav2vec 2.0	0.1307	0.0084	0.1229	0.0216	0.2204	0.1123	0.1976	0.1100	0.0316
WavLM	**0.1039**	0.0070	**0.1142**	0.0178	**0.1954**	0.0980	**0.1770**	0.1059	0.0238

C. Evaluation Metrics

The mean cosine similarity (MCS), concordance correlation coefficient (CCC), and mean absolute error (MAE) between a system's estimations and human evaluations, are calculated as three metrics, in order to evaluate the BEES performance. In particular, the MCS is merely a preferred metric to evaluate the BEES results, in view of the fact that BEES explores the intensities of multiple emotions that co-exist simultaneously as an emotional vector, where smaller MAE and greater CCC measure the agreement between the outputs of the system and the ground truth considering each specific emotional intensity individually might not result in a good BEES performance.

Formally, $\hat{\rho_\lambda^n}$ represents the estimated intensity reference value for a specific emotion λ in the emoMat of the nth voice stimulus from the IEMOCAP corpus, as determined by a BEES system, and the corresponding value of that given by three human estimators is ρ_λ^n. The MCS, CCC, and MAE are accordingly obtained by:

$$\text{MCS} = \frac{1}{N}\sum_{1}^{N} \frac{\sum_\lambda \rho_\lambda^n \hat{\rho_\lambda^n}}{\sqrt{\sum_\lambda \left(\rho_\lambda^n\right)^2}\sqrt{\sum_\lambda \left(\hat{\rho_\lambda^n}\right)^2}} \tag{5}$$

$$\text{CCC} = \frac{2\wp\sigma_{\hat{\rho_\lambda^n}}\sigma_{\rho_\lambda^n}}{\sigma^2_{\hat{\rho_\lambda^n}} + \sigma^2_{\rho_\lambda^n} + \left(\mu_{\hat{\rho_\lambda^n}} - \mu_{\rho_\lambda^n}\right)^2} \tag{6}$$

$$\text{MAE} = \frac{1}{N}\sum_{1}^{N} \mid \rho_\lambda^n - \hat{\rho_\lambda^n} \mid \tag{7}$$

where $\mu_{(\cdot)}$ and $\sigma^2_{(\cdot)}$ are the mean values and variances, respectively, and $\wp$ is the correlation coefficient. Notably, MCS and CCC assign values that trend to 1 for a closer system's estimation to human evaluations; and MAE assigns values that trend to 0 for a better BEES performance of a system's estimations.

IV. Experiment Results and Discussions

In this section, we introduce the BEES baseline using the IEMOCAP corpus and present the main results in Tables I and II. Table I clearly demonstrates that the MCS value consistently turned out to receive a notable gain from the wavLM in comparison with that of other SSLs, providing the highest MCS reaching up to 0.7303. The main reason might be first attributed to the fact that the WavLM utilizes utterance mixing augmentation by adding both interfering speech and noise during the training process. This augmentation strategy significantly contributes to enhancing the model's robustness by exposing it to a more diverse array of inputs. Secondly, given that WavLM extends the HuBERT framework by integrating a gated relative position bias into the Transformer structure, it is poised to introduce advancements in effectively managing relative positions within the sequence that have been shown to play a vital role in conveying emotional information [41]–[43].

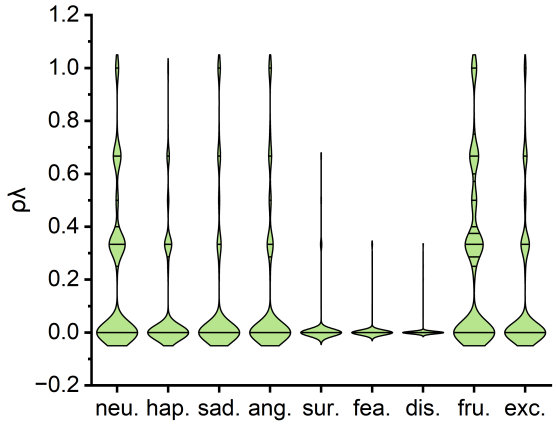


Fig. 3: The ρ_λ distributions for each specific emotion.

Moreover, Table II details the BEES results obtained using different SSLs, with CCC and MAE values provided for each specific emotion, respectively. It highlights that estimating emotional intensities for disgust, fear, and surprise poses notably greater challenges compared to other emotional states. Figure 3 shows a data visualization, illustrating that such difficulty in estimating these three emotions mainly stems from

TABLE III: Comparison results of MCS values on genders (male vs. female) and voice naturalness (acted vs. spontaneous). n.s. indicates that the MCS results do not differ significantly between male and female voices; * indicates that the MCS results differ significantly between acted and spontaneous voices ($p<0.05$).

MCS	Genders$^{n.s.}$		Voice naturalness*	
	male	female	acted	spontaneous
Data2Vec	0.7187	0.7194	0.7380	0.6982
Unispeech	0.6875	0.6898	0.6884	0.6888
HuBert	0.7100	0.7072	0.7293	0.6860
Wav2vec 2.0	0.6932	0.6840	0.6943	0.6828
WavLM	0.7293	0.7314	0.7422	0.7172

their limited distribution of emotional intensity, which results in an unbalanced data problem recognized as one of the main obstacles in the field of machine learning [44].

For further analysis, Table III summarizes two aspects of the BEES results in terms of MCS values, focusing on the impact of genders and voice naturalness. It is important to note that the t-test conducted on all SSLs indicated no statistically significant difference in performance between male and female voices (p=0.56). This observation shows the potential of SSLs in effectively characterizing gender diversity in acoustic emotion analysis, a task known for its complexity in many emotion studies [45], [46]. Additionally, as reasonably expected, the MCS values were significantly higher for the acted voices compared to the spontaneous ones ($p<0.05$), which aligns with findings from prior studies [4], [46]. These results are primarily attributed to the acted voices generally exhibiting fewer clear emotions with higher intensities, while the spontaneous ones tend to be more ambiguous and imprecise, which may be harder to estimate and thus limit their performance.

V. Conclusions

This paper introduces a groundbreaking acoustic task for BEES, addressing the limitations of traditional single-label classification approaches in AER. Our task is well-matched with the general psycho-evolutionary theory of emotion, which posits that human emotions often comprise blends of different emotions with varying intensities. We have offered a straightforward and easily replicable guide for investigating BEES. Future research endeavors could delve into exploring dynamic shifts in BEES within conversational dialogues, providing deeper insights into real-time emotional fluctuations. Moreover, there is promising potential for studying the intricate interplay between emotions and related psychological phenomena, such as stress and depression detection. This proposition not only paves the way for pioneering research in AER but also holds considerable promise for applications in mental health analysis. We hope that BEES will inspire further research and development, leading to powerful emotional agents capable of nuanced emotional understanding and response.

Reference

[1] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions.* Oxford University Press, 1994.

[2] P. A. Thoits, "The sociology of emotions," *Annual review of sociology*, vol. 15, no. 1, pp. 317–342, 1989.

[3] L. F. Barrett and J. A. Russell, *The psychological construction of emotion*. Guilford Publications, 2014.

[4] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Comm. of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[5] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *AI Rev.*, vol. 43, pp. 155–177, 2015.

[6] X. Shi, S. Li, and J. Dang, "Dimensional emotion prediction based on interactive context in conversation.," in *INTERSPEECH*, 2020, pp. 4193–4197.

[7] X. Shi, X. Li, and T. Toda, "Multimodal fusion of music theory-inspired and self-supervised representations for improved emotion recognition," in *Proc. Interspeech*, 2024, pp. 2024–2350.

[8] J. Tian, D. Hu, X. Shi, *et al.*, "Semi-supervised multimodal emotion recognition with consensus decision-making and label correction," in *Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing*, 2023, pp. 67–73.

[9] X. Shi, X. Li, and T. Toda, "Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation," in *Proc. Interspeech*, 2023, pp. 765–769.

[10] K. R. Scherer, "Analyzing emotion blends," in *Proceedings of the 10th Conference of the International Society for Research on Emotions*, ISRE Publications Würzburg, 1998, pp. 142–148.

[11] R. Berrios, P. Totterdell, and S. Kellett, "Eliciting mixed emotions: A meta-analysis comparing models, types, and measures," *Frontiers in psychology*, vol. 6, p. 428, 2015.

[12] V. Y. Oh and E. M. Tong, "Specificity in the study of mixed emotions: A theoretical framework," *Personality and Social Psychology Review*, vol. 26, no. 4, pp. 283–314, 2022.

[13] A. Israelsson, A. Seiger, and P. Laukka, "Blended emotions can be accurately recognized from dynamic facial and vocal expressions," *J. Nonverbal Behav.*, pp. 1–18, 2023.

[14] P. N. Juslin, P. Laukka, L. Harmat, and M. Ovsiannikow, "Spontaneous vocal expressions from everyday life convey discrete emotions to listeners.," *Emotion*, vol. 21, no. 6, p. 1281, 2021.

[15] E. B. Andrade and J. B. Cohen, "On the consumption of negative feelings," *J. Consumer Res.*, vol. 34, no. 3, pp. 283–300, 2007.

[16] S. H. Hemenover and U. Schimmack, "That's disgusting, but very amusing: Mixed feelings of amusement

and disgust," *Cognition and Emotion*, vol. 21, no. 5, pp. 1102–1113, 2007.

[17] C. C. Bee and R. Madrigal, "Consumer uncertainty: The influence of anticipatory emotions on ambivalence, attitudes, and intentions," *J. Consum. Behav.*, vol. 12, no. 5, pp. 370–381, 2013.

[18] X. Li, Z. Zhang, C. Gan, and Y. Xiang, "Multi-label speech emotion recognition via inter-class difference loss under response residual network," *IEEE TMM*, 2022.

[19] A. Slimi, N. Hafar, M. Zrigui, and H. Nicolas, "Multiple models fusion for multi-label classification in speech emotion recognition systems," *Proc. Comput. Sci.*, vol. 207, pp. 2875–2882, 2022.

[20] X. Geng, "Label distribution learning," *IEEE TKDE*, vol. 28, no. 7, pp. 1734–1748, 2016.

[21] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proceedings of the IEEE/CVF conference on CVPR*, 2020, pp. 13 984–13 993.

[22] C. Busso, M. Bulut, C. C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[23] U. Schimmack, "Pleasure, displeasure, and mixed feelings: Are semantic opposites mutually exclusive?" *Cognition & Emotion*, vol. 15, no. 1, pp. 81–97, 2001.

[24] K. Oatley and P. N. Johnson-Laird, "The communicative theory of emotions: Empirical tests, mental models, and implications for social interaction," in *Striving and feeling*, Psychology Press, 2014, pp. 363–393.

[25] A. P. McGraw and C. Warren, "Benign violations: Making immoral behavior funny," *Psychol. Sci.*, vol. 21, no. 8, pp. 1141–1149, 2010.

[26] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Trans. Affect. Comput.*, 2022.

[27] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. on ASLP*, vol. 26, no. 12, pp. 2423–2435, 2018.

[28] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, "Using semi-supervised learning for monaural time-domain speech separation with a self-supervised learning-based si-snr estimator," in *Interspeech 2023*. ISCA, 2023.

[29] H. Shi, M. Mimura, and T. Kawahara, "Waveform-domain speech enhancement using spectrogram encoding for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[30] V. P. Minotto, C. R. Jung, and B. Lee, "Multimodal multi-channel on-line speaker diarization using sensor fusion through svm," *IEEE TMM*, vol. 17, no. 10, pp. 1694–1705, 2015.

[31] I. E. Livieris, E. Pintelas, and P. Pintelas, "Gender recognition by voice using an improved self-labeled algorithm," *MLKE*, vol. 1, no. 1, pp. 492–503, 2019.

[32] K. Fujita, T. Ashihara, H. Kanagawa, T. Moriya, and Y. Ijima, "Zero-shot text-to-speech synthesis conditioned using self-supervised speech representation model," *arXiv preprint arXiv:2304.11976*, 2023.

[33] E. Ekstedt and G. Skantze, "Voice activity projection: Self-supervised learning of turn-taking events," *arXiv preprint arXiv:2205.09812*, 2022.

[34] S. Yang, P. Chi, Y. Chuang, *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[35] J. Turian, J. Shier, H. R. Khan, *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 CD Track*, PMLR, 2022, pp. 125–145.

[36] B. T. Atmaja and A. Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022.

[37] H. Sun, S. Zhao, X. Wang, W. Zeng, Y. Chen, and Y. Qin, "Fine-grained disentangled representation learning for multimodal emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 051–11 055.

[38] A. Mutawa, "An end-to-end tacotron model versus pre trained tacotron model for arabic text-to-speech synthesis," *J. of Engineering Research*, 2023.

[39] N. Saleem, J. Gao, M. I. Khattak, H. T. Rauf, S. Kadry, and M. Shafi, "Deepresgru: Residual gated recurrent neural network-augmented kalman filtering for speech enhancement and recognition," *Knowledge-Based Systems*, vol. 238, p. 107 914, 2022.

[40] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[41] R. Lesser, "Verbal comprehension in aphasia: An english version of three italian tests," *Cortex*, vol. 10, no. 3, pp. 247–263, 1974.

[42] X. Wu, Y. Cao, H. Lu, *et al.*, "Speech emotion recognition using sequential capsule networks," *IEEE/ACM Trans. on ASLP*, vol. 29, pp. 3280–3291, 2021.

[43] X. Li, X. Shi, D. Hu, *et al.*, "Music theory-inspired acoustic representation for speech emotion recognition," *IEEE/ACM Trans. on ASLP*, vol. 31, pp. 2534–2547, 2023.

[44] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *PIAI*, vol. 5, no. 4, pp. 221–232, 2016.

[45] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *PR*, vol. 44, no. 3, pp. 572–587, 2011.

[46] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10-11, pp. 787–800, 2007.

Sampling Pattern Augmentation to Enhance Deep Learning-based Image Reconstruction of MRI

Kazuki Yamato* and Satoshi Ito*
* Department of Fundamental Engineering, School of Engineering, Utsunomiya University
E-mail: {yamato, itohst}@is.utsunomiya-u.ac.jp Tel/Fax: +81-286896277

***Abstract*—The application of compressed sensing (CS) to MRI enables the acceleration of imaging times. CS can approximate an original signal from measurements that do not satisfy the sampling theorem. While CS can reduce the sampling time of k-space signals by reconstructing an MR image from limited measurement signals, there are still issues such as reconstruction time and optimal parameter search due to the CS optimization algorithm. In contrast, unrolling-based network models that can learn the parameters of the CS optimization algorithm to appropriate values for the subject are proposed, and the reconstruction performance is improved by applying the model to CS. However, the unrolling-based network models are trained based on a specific sampling pattern, and the reconstruction performance is adversely affected when the reconstruction is carried out using an untrained pattern. In this study, a sampling pattern augmentation method that employs multiple sampling patterns for learning is proposed as a means of developing a model that is not dependent on specific sampling patterns. Simulation experiments demonstrated that the reconstructed images using the sampling pattern augmentation exhibited superior quality compared to those reconstructed using conventional methods, while requiring a significantly shorter learning time. Moreover, it was confirmed that the proposed method yields high reconstruction performance even when reconstructing with untrained patterns.**

I. Introduction

Magnetic resonance imaging (MRI) is a non-invasive technique that can provide high-contrast cross-sectional images of the body. However, the requisite relaxation time for signal recovery results in imaging times that are typically between 20 and 60 minutes. Consequently, numerous researchers have explored avenues for accelerate MRI, including parallel imaging [1], [2] and compressed sensing [3], [4]. In recent years, the reconstruction of under-sampled k-space data obtained by compressed sensing using deep learning [5] has enabled the generation of high-quality images in a time-efficient manner, when compared to traditional iterative reconstruction methods.

The field of deep learning-based reconstruction methods can be divided into four principal categories. The first category comprises image domain-based methods [6]–[9], which employ an inverse Fourier transform (IFT) to the under-sampled signal in order to obtain an initial image, which serves as the primary input to the deep learning process. The second category comprises k-space based methods [10], in which an under-sampled signal is directly input to the deep learning model, and then IFT is applied to obtain the reconstructed image. However, methods belonging to this category have not been widely studied due to the potential for any artifacts introduced by the deep learning to be spread over the reconstructed image. The third category comprises iterative unrolling methods [11]–[14], which commence with an optimization problem whose solution is the image to be reconstructed. The optimization algorithm is then unrolled into the deep learning. Notably, the ADMM-Net proposed by Yang et al. [11] has relatively good reconstruction performance in the same category. The fourth category comprises methods that learn the image directly from an under-sampled signal [15]. This category typically necessitates the incorporation of fully connected layers, resulting in a network of considerable scale.

In the context of deep learning reconstruction methods, the generation of artifacts resulting from the under-sampling of k-space data, utilizing a specific sampling pattern, is addressed through the implementation of a learning process. This enables the removal of artifacts that may otherwise manifest in the reconstructed image. In other words, the characteristics of the artifacts that appear in the reconstructed image are different for sampling patterns not used in training, which makes effective artifact removal difficult. Liu et al. proposed a novel sampling pattern augmentation technique [8], [9], which employs a combination of multiple sampling patterns during training with the objective of enhancing robustness in the event of disagreement between sampling patterns. Models trained with different sampling patterns are capable of estimating and removing a diverse range of artifacts. In contrast, the network proposed by Liu et al. is based on image domain-based methods, including U-Net [16] and Generative Adversarial Networks (GANs) [17]. It thus appears that there is scope for further enhancement of reconstruction performance through the incorporation of sampling pattern augmentation into iterative unrolling models such as ADMM-Net.

In this paper, we introduce sampling pattern augmentation to ADMM-Net, which has particularly high reconstruction capability among iterative unrolling models, with the aim of developing a model that is robust to changes in sampling patterns. To validate the effectiveness of the proposed approach, we conducted experimental simulations in which image reconstruction with sampling pattern augmentation was compared with that without sampling pattern augmentation, i.e., using a single sampling pattern. The experimental outcomes demonstrate the effectiveness and utility of the sampling pattern augmentation, even in the context of iterative unrolling models.

II. METHODS

A. ADMM-Net

Compressed sensing is a technique that allows the reconstruction of an original signal from fewer under-sampled signals than required by the Nyquist-Shannon sampling theorem, provided the target signal has sparsity. Thus, compressed sensing can accelerate MRI imaging time by under-sampling the Fourier transform imaging signals. The ADMM-Net proposed by Yang et al. was developed by using unrolling ADMM optimization [18], [19] to train the regularization parameters of compressed sensing. The reconstructed image ρ is estimated by solving the optimization problem as follows:

$$\hat{\boldsymbol{\rho}} = \arg\min_{\boldsymbol{\rho}} \left\{ \frac{1}{2}\|\mathbf{PF}\boldsymbol{\rho} - \mathbf{m}\|_2^2 + \sum_{l=1}^{L} \lambda_l g[\mathbf{D_l}\boldsymbol{\rho}] \right\}, \quad (1)$$

where $\mathbf{m}$ is an under-sampled signal, $\mathbf{P}$ is an under-sampling matrix, $\mathbf{F}$ is the Fourier transform, $\mathbf{D_l}$ is an undetermined linear transform, $g[\cdot]$ is a nonlinear regularization function, and λ_l is a regularization parameter.

By introducing the independent auxiliary variables $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L\}$ in the image domain and using the Lagrangian multiplier β, the iterations of the ADMM algorithm are represented as follows:

$$\begin{cases} \arg\min_{\boldsymbol{\rho}} \frac{1}{2}\|\mathbf{PF}\boldsymbol{\rho} - \mathbf{m}\|_2^2 - \frac{r}{2}\|\boldsymbol{\rho} + \boldsymbol{\beta} - \mathbf{z}\|_2^2, \\ \arg\min_{\mathbf{z}} \sum_{l=1}^{L} \lambda_l g(\mathbf{D}_l\mathbf{z}) + \frac{r}{2}\|\boldsymbol{\rho} + \boldsymbol{\beta} - \mathbf{z}\|_2^2, \\ \boldsymbol{\beta} \longleftarrow \boldsymbol{\beta} + \tilde{\eta}(\boldsymbol{\rho} - \mathbf{z}), \end{cases} \quad (2)$$

where the parameter $\tilde{\eta}$ is an update rate, and r is a penalty parameter. If $\mathbf{I}$ is a unit matrix, these iterations have the following solutions:

$$\begin{cases} \mathbf{R}^{(n)} : \boldsymbol{\rho}^{(n)} = \mathbf{F}^T \dfrac{\left[\mathbf{P}^T\mathbf{y} + r\mathbf{F}(\mathbf{z}^{(n-1)} - \boldsymbol{\beta}^{(n-1)})\right]}{(\mathbf{P}^T\mathbf{P} + r\mathbf{I})}, \\ \mathbf{Z}^{(n)} : \mathbf{z}^{(n,k)} = \mu_1\mathbf{z}^{(n,k-1)} + \mu_2(\boldsymbol{\rho}^{(n)} + \boldsymbol{\beta}^{(n-1)}) \\ \qquad\qquad - \sum_{l=1}^{L} \lambda_l \mathbf{D}_l^T \mathcal{H}[\mathbf{D}_l\mathbf{z}^{(n,k-1)}], \\ \mathbf{M}^{(n)} : \boldsymbol{\beta}^{(n)} = \boldsymbol{\beta}^{(n-1)} + \tilde{\eta}(\boldsymbol{\rho}^{(n)} - \mathbf{z}^{(n)}), \end{cases} \quad (3)$$

where $\mathcal{H}[\cdot]$ denotes a non-linear transform corresponding to the gradient of the regularization function $g[\cdot]$, and k denotes the k-th sub-iteration of $\mathbf{z}^{(n)}$. In the solution $\mathbf{Z}$, $\mu_1 = 1 - \delta r$, $\mu_2 = \delta r$, $\lambda_l = \delta\lambda_l$ where δ is the gradient descent step size. As the ADMM algorithm iterates, the solutions represented in (3) are updated alternately in the order of $\mathbf{R}^{(n)}$, $\mathbf{Z}^{(n)}$, and $\mathbf{M}^{(n)}$.

Figure 1 shows the data-flow graph of ADMM-Net. As shown in Fig. 1, the n-th stage in the data-flow graph corresponds to the n-th iteration of the ADMM algorithm. In ADMM-Net, the solutions of (3), i.e. $\mathbf{R}$, $\mathbf{Z}$, and $\mathbf{M}$, act as three kinds of layers. The layers $\mathbf{R}$, $\mathbf{Z}$, and $\mathbf{M}$ are the reconstruction layer, the noise reduction layer, and the multiplier update layer, respectively, and a set of these layers is treated as a stage, which is equivalent to a solving step of the ADMM algorithm in (3). ADMM-Net trains the penalty parameter r in the reconstruction layers, the convolution filters of the two convolution layers, which are sub-layers of the noise reduction layers $\mathbf{Z}$, the multiplier update parameter $\tilde{\eta}$, and the regularization parameter λ_l.

In the training of ADMM-Net, under-sampled signals are input to the network, and the loss between the reconstructed image output from the network and the full-data image is calculated. The loss function of the ADMM-Net uses the normalized mean squared error (NMSE).

B. Sampling Pattern Augmentation

In general, deep learning reconstruction methods employ a single sampling pattern for model training. It is therefore challenging to obtain high-quality images from an under-sampled MR signal using an untrained sampling pattern. By augmenting the sampling patterns utilized in model training from a single pattern to multiple patterns as illustrated in Fig. 2, the constraint imposed by employing a single sampling pattern in deep learning reconstruction methods can be alleviated. In the sampling pattern augmentation process, a library containing N sampling patterns are generated, and a sampling pattern to be used for each image is randomly selected from the library. The selected pattern is reselected with each change in epoch. Accordingly, the number of sampling patterns in the library, designated as N, can be expressed as follows:

$$N = EI_{\text{train}}, \quad (4)$$

where E and I_{train} represent the number of epochs in model training and the number of training images, respectively. In this paper, the generated sampling patterns are 1D Cartesian patterns sampled along the phase-encoding direction. Fig. 3 shows an example of the generated sampling patterns.

C. Experimental Configuration

To verify the effectiveness of the proposed method, we conducted experimental simulations. In this study, the robustness of the sampling pattern augmentation applied to the unrolling model is evaluated and its reconstruction performance is compared with that of the image domain-based model. In the experiments, ADMM-Net is used as the unrolling model method and U-net is used as the image domain-based model, and 485 T1-weighted MR images of the heads of consenting volunteers from IXI Dataset [20] were used. 400 randomly selected images were used for training and the rest were for testing. All images were 256×256 pixels.

In the configurations of ADMM-Net, the number of epochs, the number of stages, the number of convolution filters, filter size, the activation function, optimizer, and learning rate were set to 300, 10, 64, 5×5, Rectified Linear Unit (ReLU) [21], Adam [22], and 10^{-3}, respectively. In the configurations of U-Net, the following parameters were set: the number of epochs, the number of stages, the number of initial convolution filters, filter size, the activation function , optimizer, and learning rate. These were set to 300, 5, 64, 3×3, ReLU, Adam, 10^{-4},

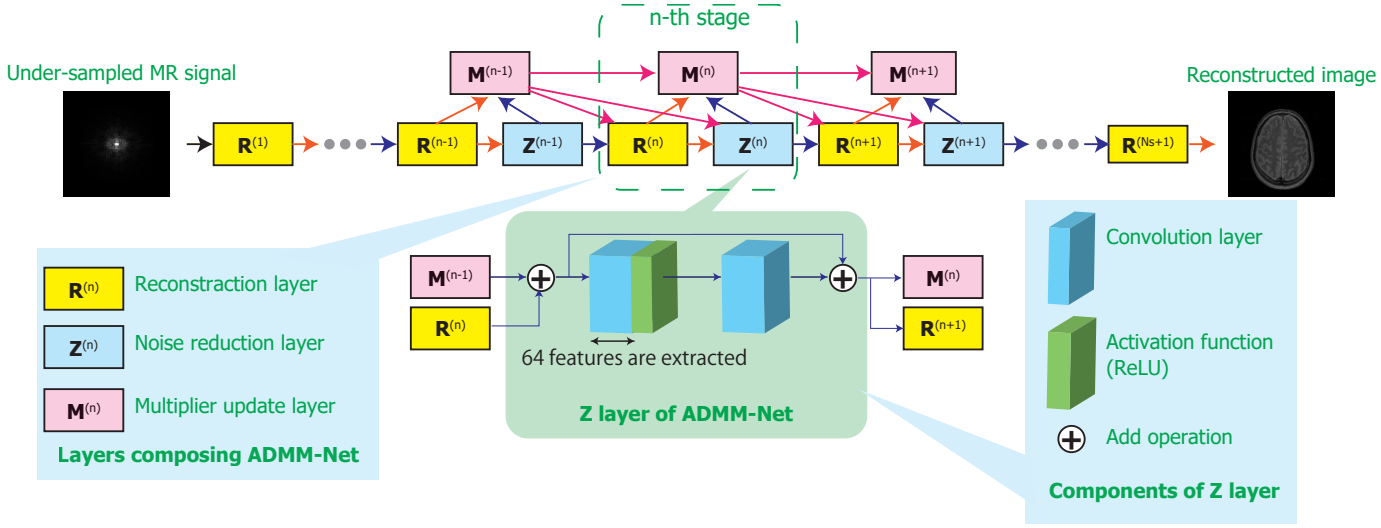


Fig. 1. The data flow graph for ADMM-Net. ADMM-Net comprises three distinct types of layers: the reconstruction layer, the noise reduction layer, and the multiplier update layer. The noise reduction layer comprises two convolution layers and an activation layer situated between them.

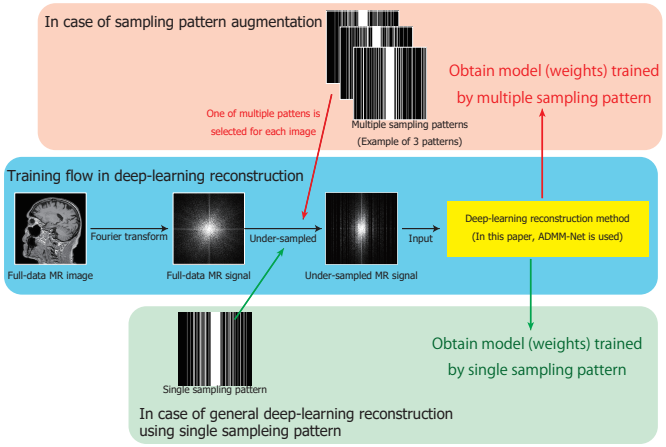


Fig. 2. Training flow of deep-learning reconstruction when under-sampling with single sampling pattern and the sampling pattern augmentation.

respectively. Furthermore, the reconstructed images by U-Net are subjected to the data consistency (DC), thereby improving the quality of the reconstructed images. The reconstructed image is then subjected to a calculation of the data consistency, denoted as $\boldsymbol{\rho}_{DC}$, as follows:

$$\boldsymbol{\rho}_{\mathrm{DC}} = \mathbf{F}^{T}[\mathbf{Pm} + (\mathbf{I} - \mathbf{P})\mathbf{F}\boldsymbol{\rho}_{\mathrm{U-Net}}] \tag{5}$$

where $\boldsymbol{\rho}_{U-Net}$ represents a reconstructed image by U-Net. $\mathbf{F}$, $\mathbf{P}$, and $\mathbf{I}$ represent the Fourier transform, a sampling pattern matrix, and a unit matrix, respectively. $\mathbf{m}$ represent the under-sampled MR signal. Therefore, the quality of the reconstructed images can be enhanced by replacing the under-sampled points of the MR signal of $\boldsymbol{\rho}_{U-Net}$ with $\mathbf{Pm}$.

The number of sampling patterns contained in the sampling pattern library are $400 \times 300 = 120000$. The sampling patterns used in the experiments are based on a Gaussian distribution. The standard deviation of the Gaussian distribution σ is as

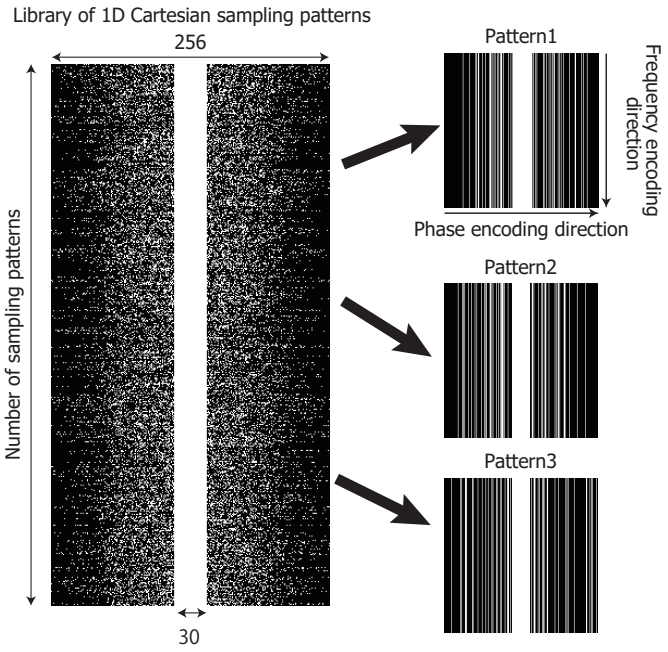


Fig. 3. Example of the 1D variable density Cartesian random sampling patterns. In this example, the matrix size of sampling patterns is set to 256×256. The White regions of the library and Pattern 1 to 3 represent under-sampling points of MR signal, and black regions are zero-filling points.

follows:

$$\sigma = 256\alpha, \tag{6}$$

where α denotes a parameter that controls the distribution of under-sampled MR signals by the sampling pattern. In addition, the low-frequency components are sampled continuously because the MR signals have a high energy concentration in the low-frequency component. The number of continuous sampling points in the

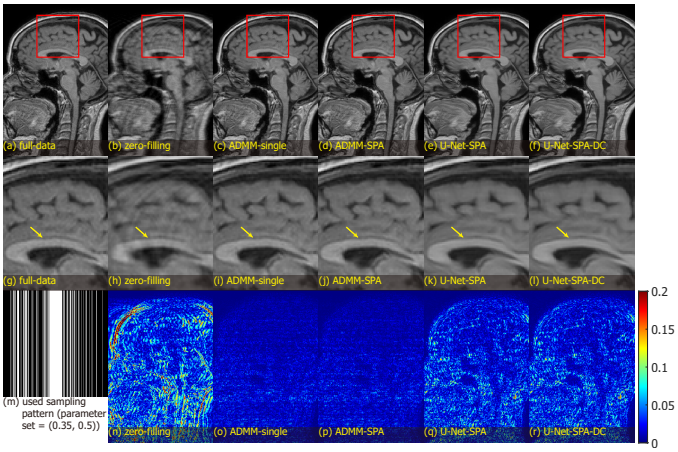


Fig. 4. Reconstructed images by each reconstruction method. The images were reconstructed from the under-sampled MR signal in accordance with a sampling pattern that had not been trained, as shown in (m). (g)-(l) show enlarged views of the red rectangle areas in (a)-(f), respectively. (n)-(r) are the error maps between the full-data and each reconstructed images. (c) was reconstructed by the model that had been trained with a sampling pattern of $(R, \alpha) = (0.3, 0.5)$. In (d)-(f), the sampling pattern augmentation (SPA) is employed.

experiments is set to 30. The sampling patterns used for the sampling pattern augmentation is generated by setting the sampling rate R and α. In the experiments, R and α are considered as a parameter set, and the sampling patterns are generated from 5 parameter sets, i.e., $(R, \alpha) \in \{(0.2, 0.5), (0.3, 0.3), (0.3, 0.5), (0.3, 0.8), (0.4, 0.5)\}$. Note that the signal sampling points are determined randomly.

To evaluate the robustness, the proposed model using the sampling pattern augmentation in ADMM-Net and the conventional model trained with a single pattern were tested by generating 10 different sampling patterns for testing with the parameter set used in single pattern learning. As an example, when the single pattern learning model using the parameter set $(0.2, 0.5)$ is tested, 10 sampling patterns for testing are generated with the parameter set $(0.2, 0.5)$.

The quantitative evaluation of the reconstructed images was conducted using three metrics: the Peak Signal-to-Noise Ratio (PSNR) and the Learned Perceptual Image Patch Similarity (LPIPS) metric [23]. The LPIPS employs pre-trained image feature extraction networks comprising comvolutional neural network (CNN)-based models such as AlexNet [24] and VGG [25]. The LPIPS metric, which employs a pre-trained AlexNet model available on GitHub, was used to assess the results of the experiments. It should be noted that a lower value of the LPIPS indicates a higher quality of reconstructed images.

The experimental simulations were performed on a computer with an Intel(R) Core(TM) i9-11900K (3.5 GHz), with 64.0 GB of RAM, and NVIDIA GeForce RTX 3090 GPU running Windows 10. The software used was Python 3.9.5, PyTorch 2.3.1, and CUDA ToolKit 12.1.

III. Experimental results

Figure 4 shows the reconstructed images when the untrained sampling pattern is used for the reconstruction. Fig. 4(m) shows the used sampling pattern in this experiment, whose sampling rate and a parameter α are set to 0.35 and 0.5, respectively. Table I shows the quantitative evaluation results of the reconstructed images. In the image reconstruction using ADMM-Net, a comparison of the reconstruction results with and without the sampling pattern augmentation indicates that the image reconstructed by the model with the sampling pattern augmentation is more accurate reproducing the indicated texture by the arrows than the image reconstructed by the model trained with a single sampling pattern. As shown in the quantitative evaluation results, the reconstructed images with the sampling pattern augmentation have higher PSNR than those of the single pattern model. On the other hand, the LPIPS of the reconstructed images with the sampling pattern augmentation was equal or higher for $(R, \alpha) = (0.2, 0.5)$ and $(0.35, 0.5)$, but lower for $(R, \alpha) = (0.3, 0.5)$ and $(0.4, 0.5)$ than that of the model trained with the single sampling pattern.

Subsequently, a comparison is conducted between ADMM-Net and U-Net, with the sampling pattern augmentation. As shown in Fig. 4, the reconstructed image by U-Net with the sampling pattern augmentation exhibits a notable degree of smoothing and a discernible loss of detail structure. Notwithstanding the enhancement of the reconstructed image quality through data consistency, no significant improvements were observed. In contrast, the ADMM-Net with the sampling pattern augmentation demonstrated superior performance in preserving detail structure compared to U-Net. As shown in the quantitative evaluation, the evaluation metrics of ADMM-Net with the sampling pattern augmentation exhibited better than those of U-Net.

IV. Discussion

In the sampling pattern augmentation conducted in the experiments, 120,000 different sampling patterns were used for model training. This equates to 120,000 artifacts being learned, with the model parameters optimized to remove these artifacts. It is therefore hypothesised that the application of sampling pattern augmentation improves the image quality of the reconstructed images both quantitatively and qualitatively.

From Fig. 4 and Table I, ADMM-Net with the sampling pattern augmentation has better reconstruction performance than that of using a single sampling pattern in terms of PSNR. In the experiments, model training is performed with 300 epochs, and the model whose reconstructed image quality is the highest among the 300 models is selected as the model for testing. The metric that determines the model for testing is PSNR. In other words, the models are determined such that both the sampling pattern augmentation and the single sampling pattern have the maximum PSNR of the reconstructed images. In addition, the loss functions of ADMM-Net and U-Net are NMSE and MSE, and the PSNR formula includes MSE. Therefore, the changes in MSE directly affect the changes in PSNR. Evaluation by

TABLE I
QUANTITATIVE METRICS COMPARING THE FULL-DATA IMAGES.

(R, α)	(0.2, 0.5)		(0.3, 0.5)		(0.4, 0.5)		(0.35, 0.5)	
	PSNR [dB]	LPIPS	PSNR [dB]	LPIPS	PSNR [dB]	LPIPS	PSNR [dB]	LPIPS
ADMM-Net (SPA)	29.02±0.71	0.016±0.008	34.07±0.93	0.016±0.003	37.37±1.13	0.012±0.002	35.74±0.99	0.007±0.002
ADMM-Net (single)	28.90±0.65	0.038±0.007	33.92±0.97	0.011±0.003	37.22±1.23	0.005±0.002	35.52±1.02	0.007±0.002
U-Net (SPA)	28.06±0.65	0.058±0.011	29.49±0.60	0.045±0.008	30.36±0.59	0.035±0.007	30.03±0.58	0.038±0.007
U-Net (SPA) + DC	28.92±0.67	0.052±0.011	31.32±0.61	0.035±0.008	32.98±0.58	0.024±0.006	32.26±0.59	0.028±0.007

TABLE II
COMPUTATIONAL COMPLEXITY

	Training [min]	Reconstruction [ms]	Parameters
ADMM-Net (SPA)	24.80	3.59	33,823
ADMM-Net (single)	24.80	3.81	33,823
U-Net (SPA)	32.21	725.54	22,023,937

PSNR has a high affinity with the loss functions such as NMSE and MSE. As a result, the effectiveness of the sampling pattern augmentation could be demonstrated in terms of PSNR.

On the other hand, ADMM-Net with the sampling pattern augmentation exhibits suboptimal performance with respect to the LPIPS metric. The LPIPS metric can be evaluated in accordance with human perceptual characteristics. In particular, LPIPS tends to evaluate more highly an image in which random noise-like artifacts remain but the detailed structure has been restored than an image in which the detailed structure has been destroyed by smoothing through reconstruction. In contrast, PSNR is evaluated more highly when the error between the pixel values of the full data and those of the reconstructed image is small. Consequently, the evaluation is higher even if smoothing or blurring has occurred. As previously stated, since the selection criterion for the test model was PSNR, LPIPS of the reconstructed image was not necessarily highly evaluated. Therefore, the evaluation when LPIPS is used as the model selection criterion is a future issue.

Table II shows the processing times for training and reconstruction, as well as the number of parameters associated with each model. As shown in the table, it is clear that the training time, reconstruction time, and number of parameters remain unaltered when the sampling pattern augmentation is introduced or excluded. In the sampling pattern augmentation, a sampling pattern for each image is randomly selected from the sampling pattern library and reselected with each change in epoch. In other words, the number of training data remains constant, in contrast to the data augmentation approach. Furthermore, the sampling pattern augmentation is a form of preprocessing that occurs before the network input for training. Therefore, the number of parameters comprising the model remains unaltered in the presence or absence of the sampling pattern augmentation.

With regard to the issue of computational complexity, it can be observed that the reconstruction of ADMM-Net is capable of begin carried out at a speed that is 190 times faster than that of U-Net. The U-Net model comprises a multitude of convolution layers, with a total of 650 times more parameters than those comprising the ADMM-Net model. Therefore, ADMM-Net demonstrated superior performance in terms of both speed and image quality compared to U-Net.

In this paper, the experiments were performed on real-valued images, but the images acquired from MRI are complex images with phase components. From practical point of view, it is necessary to confirm the effectiveness of sampling pattern expansion also for complex images, and this is a topic for future work.

V. CONCLUSIONS

In this study, we introduced the sampling pattern augmentation to an iterative unrolling model, specifically ADMM-Net in order to enhance the robustness of the trained model in the presence of sampling pattern discrepancies. In the sampling pattern augmentation, a sampling pattern for each image is randomly selected from the sampling pattern library and reselected with each change in epoch. To verify the effectiveness of the sampling pattern augmentation in enhancing the robustness of the trained model, image reconstruction experiments were conducted using untrained sampling patterns. As a consequence, the reconstructed images produced by ADMM-Net with the sampling pattern augmentation exhibited a higher PSNR than those of the model trained with a single sampling pattern. In comparison to the image domain-based model, i.e., U-Net, with the sampling pattern augmentation, it was confirmed that ADMM-Net with the sampling pattern augmentation exhibited superior performance in terms of reconstruction speed and quality relative to U-Net.

ACKNOWLEDGMENT

This work was supported by a Grant for Young Exploratory Research in the School of Engineering, Utsunomiya University. The authors thank the Information eXtraction from Images project for creating the IXI Dataset.

REFERENCES

[1] D. Liang, B. Liu, J. Wang, and L. Ying, "Accelerating sense using compressed sensing," *Magnetic Resonance in Medicine*, vol. 62, no. 6, pp. 1574–1584, 2009.

[2] H. Jung, K. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye, "K-t focuss: A general compressed sensing framework for high resolution dynamic mri," *Magnetic Resonance in Medicine*, vol. 61, no. 1, pp. 103–116, 2009.

[3] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006. DOI: 10.1109/TIT.2006.871582.

[4] M. Lustig *et al.*, "Sparse mri: The application of compressed sensing for rapid mr imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007. DOI: https://doi.org/10.1002/mrm.21391.

[5] S. Wang, Z. Su, L. Ying, *et al.*, "Accelerating magnetic resonance imaging via deep learning," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 514–517. DOI: 10.1109/ISBI.2016.7493320.

[6] G. Yang, S. Yu, H. Dong, *et al.*, "Dagan: Deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1310–1321, 2018. DOI: 10.1109/TMI.2017.2785879.

[7] M. Mardani, E. Gong, J. Y. Cheng, *et al.*, "Deep generative adversarial neural networks for compressive sensing mri," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 167–179, 2019. DOI: 10.1109/TMI.2018.2858752.

[8] F. Liu, L. Feng, and R. Kijowski, "Mantis: Model-augmented neural network with incoherent k-space sampling for efficient mr parameter mapping," *Magnetic Resonance in Medicine*, vol. 82, no. 1, pp. 174–188, 2019. DOI: https://doi.org/10.1002/mrm.27707.

[9] F. Liu, A. Samsonov, L. Chen, R. Kijowski, and L. Feng, "Santis: Sampling-augmented neural network with incoherent structure for mr image reconstruction," *Magnetic Resonance in Medicine*, vol. 82, no. 5, pp. 1890–1904, 2019. DOI: https://doi.org/10.1002/mrm.27827.

[10] Y. Han, L. Sunwoo, and J. C. Ye, "k -space deep learning for accelerated mri," *IEEE Transactions on Medical Imaging*, vol. 39, no. 2, pp. 377–386, 2020. DOI: 10.1109/TMI.2019.2927101.

[11] Y. Yang *et al.*, "Admm-csnet: A deep learning approach for image compressive sensing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 521–538, 2020. DOI: 10.1109/TPAMI.2018.2883941.

[12] K. Hammernik *et al.*, "Learning a variational network for reconstruction of accelerated mri data," *Magnetic Resonance in Medicine*, vol. 79, no. 6, pp. 3055–3071, 2018. DOI: https://doi.org/10.1002/mrm.26977.

[13] J. Zhang and B. Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[14] H. K. Aggarwal *et al.*, "Modl: Model-based deep learning architecture for inverse problems," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 394–405, 2019. DOI: 10.1109/TMI.2018.2865356.

[15] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, 2018.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[18] "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976, ISSN: 0898-1221. DOI: https://doi.org/10.1016/0898-1221(76)90003-1.

[19] J. Eckstein and D. P. Bertsekas, "On the douglas—rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1, pp. 293–318, Apr. 1992. DOI: 10.1007/BF01581204.

[20] *IXI dataset*, https : / / brain - development . org / ixi - dataset/(Accessed: July 16, 2024).

[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

Enhancing YOLOv7 with GLF-Trans for Precision in Small Object Detection

Naohito Yoshikawa* and Masaaki Ikehara†
* Electrical and Information Engineering (of Keio University), Kanagawa, Japan
E-mail: yoshikawa@tkhm.elec.keio.ac.jp
† Electrical and Information Engineering (of Keio University), Kanagawa, Japan
E-mail: ikehara@tkhm.elec.keio.ac.jp

Abstract—Object detection techniques are essential for identifying objects and accurately determining their locations in images and videos. This paper proposes an approach to enhance the performance of YOLOv7, the YOLO (You Only Look Once) series, by improving detection accuracy, particularly for small objects. The accurate detection of small objects is crucial in various real-world applications such as autonomous driving, surveillance, and medical imaging, where the identification of small, often critical objects can have significant implications. The proposed method balances the current one-step and two-step algorithms used in object detection, with the goal of improving the accuracy of small object detection. We apply several improvements to the YOLOv7 framework, including the introduction of an attention module, the Global-Local Fusion Transformer (GLF-Trans), and the addition of a tiny-head. These enhancements are designed to further enhance the accuracy of small object detection. The experimental results show that the proposed method achieves higher accuracy than the previous model, with a specific improvement of 3.8% in $mAP_{0.5}$ on the VisDrone2019 dataset.

I. Introduction

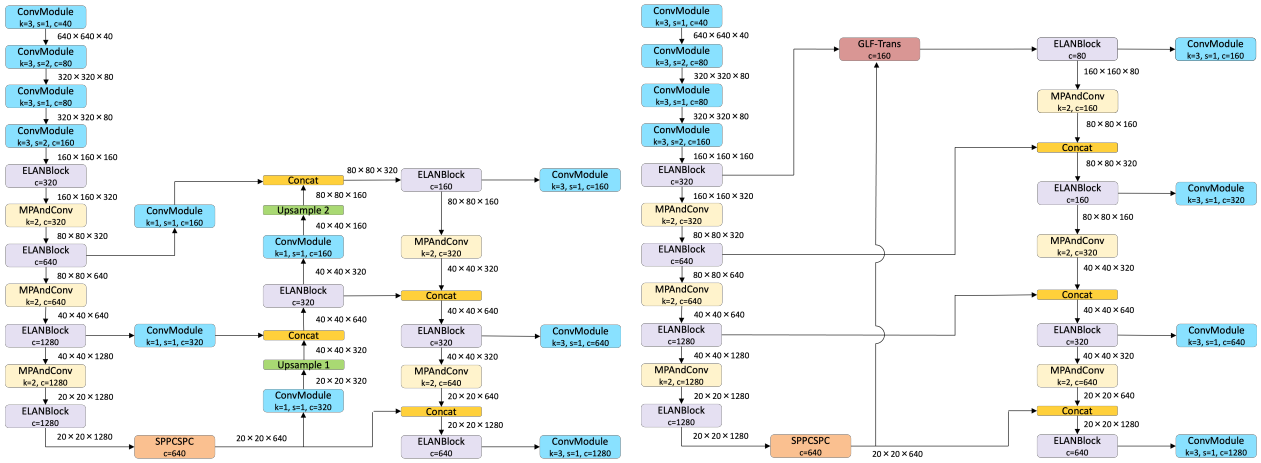
Object detection is a technique used to locate and identify objects in images or videos. This technology is widely used in various applications. In recent years, advances in object detection models based on deep learning have enabled fast and accurate detection of objects. The YOLO (You Only Look Once) series [1][2][3][4][5][6][7] has gained attention due to its balance of speed and accuracy. This study examines YOLOv7 by Wang et al. [7], which has demonstrated superior performance compared to other object detection techniques. However, there is still potential for improvement in detecting small objects.

Object detection algorithms can be broadly divided into two categories: single-stage and two-stage algorithms. The former prioritizes execution speed over accuracy, while the latter achieves higher accuracy at the cost of slower processing times. YOLO is a typical example of a one-stage algorithm, which is suitable for real-time processing due to its high speed. However, it has been noted that its accuracy is inferior to the two-stage approach, particularly in detecting small objects. Small object detection is a challenging task due to the lack of detail and difficulty in distinguishing objects from the background. However, improving the accuracy of small object detection while maintaining real-time performance can have numerous applications, including navigation in automated vehicles, surveillance systems, and robotics. This paper develops new methods that can accurately detect small objects by surpassing the limitations of current one-step algorithms.

In this study, we improved the YOLOv7 [7] object detection framework based on the importance of learning the right balance between detailed information and overall context to enhance the accuracy of small object detection. The existing YOLO model does not learn enough contextual information, lacking the detailed information necessary for small object detection. To address this issue, we introduced two major improvements. First, we added a fourth head, the Tiny Head, to convey sufficient information for detecting small objects. Next, we replaced the neck part of the model with an attention module, the Global-Local Fusion Transformer (GLF-Trans), to reduce the increased parameter count and computational load caused by the Tiny Head while maintaining accuracy and real-time performance. These improvements are expected to enable learning the balance between detailed and contextual information, thereby improving the accuracy of small object detection.

The proposed method reduces the number of parameters from 70.9M to 67.3M without negatively impacting computational cost and efficiency. Experiments on the VisDrone2019 dataset [8] demonstrate a 3.8% accuracy improvement at $mAP_{0.5}$ for small object detection compared to the previous model. The main contributions of this paper are as follows:

- A new Tiny-head specialized for the detection of small objects is added to the existing framework of YOLOv7, enabling more precise identification of small objects. Detection of small objects relies heavily on high-resolution feature maps due to their diminutive size. The implementation of Tiny-head allows for the utilization of finer feature maps to detect small objects, significantly enhancing the accuracy of identification for small objects.
- To maintain accuracy while reducing the number of parameters and computational load, the Global-Local Fusion Transformer (GLF-Trans) is introduced. The detection of small objects necessitates the differentiation of crucial details from background noise



(a) YOLOv7x (b) Proposed model

Fig. 1: Structure of YOLOv7x and the proposed model

and the concurrent understanding of the object and its surrounding context. GLF-Trans enhances this capability by fusing global contextual information with local detail through an attention mechanism, effectively replacing the current neck mechanism.

- By adopting Cluster-NMS in place of the traditional NMS algorithm, we improve the clustering accuracy of detected objects, thereby increasing detection precision, particularly in dense scenes or situations where occlusion occurs.

II. Related Works

A. Advancements in YOLO Architectures for Small Object Detection

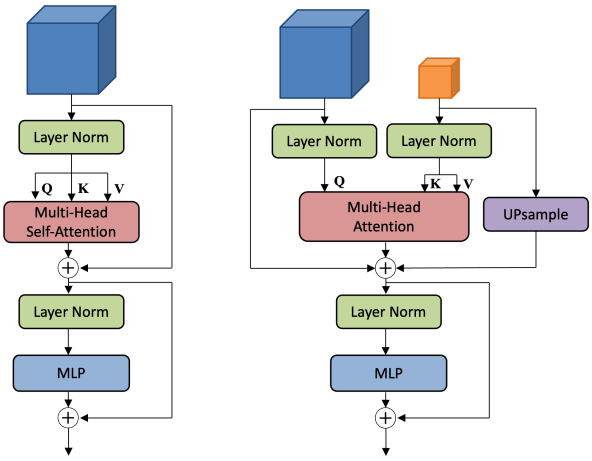
The field of small object detection has significantly evolved alongside the advancements in object detection technologies. Among these developments, the QueryDet approach by Yang et al. [9] represents a specialized methodology for enhancing the detection accuracy of small objects through the utilization of a cascading structure. This approach first identifies the approximate location of objects, then refines the details at a finer scale, thereby improving the detection precision for small objects. The YOLO series is renowned for its simplicity and rapid detection speed, and with each successive version, improvements have been made to enhance small object detection. While early versions struggled with detecting small objects, the introduction of multi-scale detection in YOLOv3 [3], utilizing feature maps of different resolutions, marked a significant improvement in detection accuracy for small objects. YOLOv4 [4] further expanded upon this by adding data augmentation techniques such as CutMix and Mosaic. Additionally, the inclusion of technologies like CSPNet and Spatial Pyramid Pooling further strengthened the detection accuracy and generalization capability of models for small object detection. In YOLOv5 [5], the model architecture was optimized, and the use of anchor boxes was refined, leading to further improvements in detecting performance for small objects. YOLOv7 [7] referenced in this paper, the introduction of the ELAN structure is highlighted, which serves to enhance the network's learning capacity without interrupting the original gradient flow. These improvements have elevated the accuracy of the YOLO series in detecting various objects, including small ones. However, the performance in small object detection has not yet reached a satisfactory level.

B. Attention Modules for Object Detection

The attention mechanism allows models to focus on the important parts of the input, clarifying the distinction between objects and background in object detection. Zhu et al.'s TPH-YOLOv5 [10], developed for drone-captured scenario applications, introduces a Transformer Prediction Head (TPH) leveraging a self-attention mechanism into YOLOv5 to accommodate object scale variations due to different altitudes and motion blur, thereby enhancing the accuracy in dense object detection. This model employs the Convolutional Block Attention Module to concentrate on significant objects while suppressing confusing information in the image. The utilization of attention module to capture detailed information extends beyond object detection. In Han et al.'s BRNet [11], a Global-Detail Fusion Module (GDFM) that combines local and global features was introduced to improve self-supervised monocular depth estimation. This module generates feature representations enriched with more comprehensive information through the fusion of local and global information, enhancing the accuracy of depth estimation.

C. NMS

Non-Maximum Suppression (NMS) is an essential process in object detection, aimed at reducing redundant detection boxes to refine the final detection outcomes. The method proposed by Zheng et al., Cluster-NMS [12],



(a) Transformer Block (b) GLF-Trans Block

Fig. 2: Structure of Transformer Block and GLF-Trans Block

enhances this traditional NMS process by efficiently clustering detection boxes, thereby simultaneously improving the efficiency and accuracy during the inference phase. This technique not only considers the overlapping areas but also incorporates geometric elements such as the distance between the centers of the boxes and their aspect ratios, improving the balance between detection precision (Average Precision, AP) and recall (Average Recall, AR). Notably, it offers performance enhancements in densely populated scenes and situations where occlusion occurs, demonstrating effectiveness not only in detecting large objects but also in small object detection scenarios.

III. Proposed Method

The proposed network is illustrated in Fig. 1(b). While our model builds upon the foundation of YOLOv7x by Wang et al. [7], citing Fig. 1(a) as the basis for improvements, it fundamentally retains the same structure as YOLOv7. YOLOv7 is composed of a Backbone, Neck, and Head, whereas our network consists of a Backbone, GLF-Trans, and Head. The Backbone, similar to traditional YOLOv7, is comprised of ConvModules and ELANBlocks [13], extracting feature maps of various scales from the input image. The GLF-Trans serves as an attention module, replacing the conventional Path Aggregation Network structure in the Neck segment. Feature maps downsampled by factors of 4 and 32 are fed into the attention mechanism. The fusion of feature maps of different scales through attention enables the aggregation of global features required by each pixel and adds these features to the detailed feature map, generating feature representations containing richer information. This allows for the acquisition of detailed information necessary for improving the detection accuracy of small objects, as well as global contextual information, and is also expected to contribute to the suppression of background noise. Traditionally, the Head component utilized three Heads for

TABLE I: Model Performance Comparison

Model	$mAP_{0.5_val}$	$mAP_{0.5_test}$	Parameters	GFLOPS
YOLOv5m	36.7	30.9	20.9M	48.3
YOLOv5l	37.2	31.6	46.1M	107.8
YOLOv5x	40.7	32.7	86.2M	203.9
YOLOv7	48.9	41.1	61.9M	103.3
YOLOv7x	50.1	42.3	70.9M	189.1
Ours	53.9	44.5	67.3M	210.3

detecting large, medium, and small objects, respectively; however, an additional Head is essential for detecting even smaller objects. Therefore, a fourth Head dedicated to detecting tiny objects is added. The introduction of the other head could ordinarily lead to a higher computational cost, but the adoption of GLF-Trans as a replacement for the conventional Neck structure offsets this increase in this model. As a result, the addition of the Tiny-head does not significantly escalate the overall computational burden. Furthermore, in the post-detection NMS phase, the adoption of Cluster-NMS [12] is expected to yield higher accuracy in dense scenes and situations with occlusion. In summary, the main improvements from the traditional YOLOv7 can be identified in three areas.

- The ConvModule and ELANBlock, previously utilized in the Neck section of YOLOv7, are removed and replaced with the GLF-Trans.
- Beyond the conventional three Heads, a fourth Tiny-Head is introduced to detect smaller objects.
- The traditional NMS is replaced with Cluster-NMS [12].

A. Global-Local Fusion Transformer(GLF-Trans)

The Global-Local Fusion Transformer (GLF-Trans) is constructed based on the Global-Detail Fusion Module (GDFM), an attention module utilized in Han et al.'s BRNet [11]. Its structure is depicted in Fig. 2. This architecture adapts the multi-head attention of a basic Transformer block to handle two different-sized inputs. The C3 level feature map (downsampled by a factor of 4) serves as the Query, and the C5 level feature map (downsampled by a factor of 32) acts as both Key and Value for input into the multi-head attention mechanism. Subsequently, the output of the multi-head attention is added to the C3 feature map along with the upsampled C5 feature map. Utilizing this module enables the fusion of global and local features and the suppression of background noise. From the perspective of fusing global and local features, this function can replace the PAN structure embedded in the Neck of the traditional YOLOv7. Therefore, by replacing the entire Neck with GLF-Trans, it is possible to maintain the detection accuracy of small objects while reducing parameters and computational load.

B. Tiny-Head

In small object detection, the size of the feature maps processed by the Head constitutes a critical factor. Net-

TABLE II: Ablation Experiment

Model	Tiny-head	GLF-Trans	Neck	Cluster-NMS	$mAP_{0.5}$	mAP_{small}	Parameters	GFLOPS
Base(YOLOv7x)	-	-	✓	-	42.3	11.9	70.9M	189.1
+Cluster-NMS	-	-	✓	✓	43.0	12.2	70.9M	189.1
+Tiny-Head	✓	-	✓	-	44.6	13.4	72.4M	226.8
+Tiny-Head w/o Neck	✓	-	-	-	41.8	12.4	66.9M	198.4
+Tiny-Head+GLF-Trans	✓	✓	-	-	44.4	13.1	67.3M	210.3
Ours	✓	✓	-	✓	44.5	13.3	67.3M	210.3

works in the RetinaNet lineage, such as Yang et al.'s QueryDet [9] and Chen et al.'s YOLOF [14], have identified the increased computational cost associated with a higher number of heads as a significant issue. However, the design of the YOLO series networks allows for improved detection performance while maintaining high computational speed. RetinaNet [15] employs numerous heads for multi-scale detection, leading to increased computational expenses and a decrease in computational speed. In contrast, YOLO processes the entire image in one go, resulting in an increase in computational cost, but it restrains the decrease in computational speed. The newly added Tiny-head, generated from low-level, high-resolution feature maps, exhibits high sensitivity towards small objects. Consequently, it mitigates the adverse effects of object scale variance, particularly enhancing detection accuracy for smaller objects.

IV. Experiment

A. Set Up

Dataset: In this study, we conduct a comprehensive analysis based on the VisDrone-DET2019 [8] dataset. This dataset is comprised of a diverse collection of images captured by drones, focusing on ten categories of objects that are closely related to daily life in urban environments, including pedestrians and various types of vehicles. With a total of 8,599 images, annotated with over 540,000 bounding boxes, the dataset highlights the diversity of objects and their complex overlaps, captured from various altitudes and locations. The dataset is divided into a training set of 6,471 images, a validation set of 548 images, and a test set of 1,580 images, collected from different locations under similar environmental conditions, facilitating a broad evaluation of the challenges in object detection. The utilization of this dataset represents an attempt to address some of the more difficult challenges in object detection, such as variations in object scale and the identification of overlapping objects. The images contain a rich diversity of overlapping objects, as well as a variety of sizes and shapes, making it an ideal dataset for testing the accuracy and robustness of small object detection technologies.

Metrics: In this study, we place particular emphasis on the evaluation metric $mAP_{0.5}$ for assessing performance. $mAP_{0.5}$ represents the mean Average Precision calculated with an Intersection over Union (IoU) threshold set at 0.5, evaluating detections as accurate if the predicted bounding boxes by the algorithm overlap more than 50% with the actual object bounding boxes. Utilizing this metric allows us to directly reflect the accuracy of predictions regarding the location and size of objects, thereby effectively capturing the actual detection performance of the algorithm. Additionally, we assess the detection accuracy for small objects (mAP_{small}) to validate the performance of the algorithm in small object detection. Moreover, as another crucial aspect of performance, we use GFLOPS and the number of parameters as metrics to evaluate processing speed.

Hyperparameters: In this study, we adopted the following hyperparameters for training the model. Stochastic Gradient Descent (SGD) is utilized for optimization, with an initial learning rate set at 0.01. This learning rate decreases gradually through the cosine annealing method to a final value of 0.1, after a warm-up period of 3 epochs. During the warm-up period, the momentum is set at 0.8 and the learning rate for biases at 0.1. Weight decay is established at 5×10^{-4} to mitigate overfitting. Furthermore, the model is trained on images of size 640x640 pixels, with a batch size of 10.

B. Performance Comparison

To evaluate the performance of the proposed model, we conduct a comparative analysis with other state-of-the-art single-stage detectors, including three variants of YOLOv5 [5] and two variants of YOLOv7 [7]. The results are presented in Tab. I. The proposed model has achieved a 3.8% improvement in $mAP_{0.5}$ on the validation set and a 2.2% improvement on the test set compared to the traditional YOLOv7x. The number of parameters has slightly decreased due to the removal of the Neck component. Regarding computational load, there was a slight increase compared to YOLOv7x, but this increase is not as significant as the increase in computational load from YOLOv7 to YOLOv7x. Overall, the proposed algorithm demonstrates its effectiveness in detecting small and densely packed objects contained within aerial photography datasets.

C. Ablation Study

To investigate the effectiveness of the Tiny-head, GLF-Trans, the existing Neck, and Cluster-NMS [12] in small object detection, we conduct an ablation study using YOLOv7x [7]. The results are shown in Tab. II. The

use of the Tiny-head resulted in a 2.3% improvement in mAP0.5, indicating its significant impact on small object detection. However, both the number of parameters and computational load increased. Therefore, replacing the existing Neck with GLF-Trans can reduce the number of parameters and computational load, although it slightly decreases the accuracy. Additionally, the adoption of Cluster-NMS [12] in YOLOv7x [7] improved mAP0.5 by 0.7% and mAPsmall by 0.3%, demonstrating its contribution to detecting small objects. In the final proposed model, mAP0.5 is improved by 2.2%, and $\text{mAP}_{\text{small}}$ is improved by 1.4%.

V. Conclusions

In this paper, we propose a real-time object detector with high accuracy for small object detection, based on YOLOv7x [7]. To optimize the network for small objects, we introduce an attention module, the Global-Local Fusion Transformer (GLF-Trans), and a fourth head, known as the Tiny-head. By adopting GLF-Trans in place of the traditional Neck, we are able to reduce parameters while learning the contextual information necessary for detecting small objects. The addition of the Tiny-head utilizes high-resolution feature maps, thereby enhancing the network's ability to recognize small objects. The proposed model demonstrates its effectiveness for small objects on the VisDrone-2019 [8] dataset, showing a 3.8% improvement in $\text{mAP}_{0.5}$ compared to conventional models.

References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[2] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263–7271.

[3] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[5] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, and Y. K. et al., ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, version v7.0, 2022. doi: 10.5281/zenodo.7347926. [Online]. Available: https://doi.org/10.5281/zenodo.7347926.

[6] C. Li, L. Li, H. Jiang, et al., "Yolov6: A single-stage object detection framework for industrial applications," arXiv preprint arXiv:2209.02976, 2022.

[7] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7464–7475.

[8] D. Du, P. Zhu, L. Wen, et al., "Visdrone-det2019: The vision meets drone object detection in image challenge results," in Proceedings of the IEEE/CVF international conference on computer vision workshops, 2019, pp. 0–0.

[9] C. Yang, Z. Huang, and N. Wang, "Querydet: Cascaded sparse query for accelerating high-resolution small object detection," in Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2022, pp. 13 668–13 677.

[10] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2778–2788.

[11] W. Han, J. Yin, X. Jin, X. Dai, and J. Shen, "Brnet: Exploring comprehensive features for monocular depth estimation," in European Conference on Computer Vision, Springer, 2022, pp. 586–602.

[12] Z. Zheng, P. Wang, D. Ren, et al., "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," IEEE transactions on cybernetics, vol. 52, no. 8, pp. 8574–8586, 2021.

[13] C.-Y. Wang, H.-Y. M. Liao, and I.-H. Yeh, "Designing network design strategies through gradient path analysis," arXiv preprint arXiv:2211.04800, 2022.

[14] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13 039–13 048.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

Adaptive Subspace Clustering for Matrix Completion

Takuto Wada*, Ryohei Sasaki†, and Katsumi Konishi*
* Hosei University, Tokyo
E-mail: takuto.wada.5w@stu.hosei.ac.jp, konishi@hosei.ac.jp Tel/Fax: +81-332649240
† Tokyo University of Technology, Tokyo
E-mail: sasakirh@stf.teu.ac.jp Tel/Fax: +81-426372111

***Abstract*—This paper deals with subspace clustering for a matrix completion, which is a problem of estimating missing entries in a matrix under the assumption that row or column vectors belongs to multiple low-dimensional linear spaces. Various methods for the problem have been proposed. Some of them assume that row or column vectors in a matrix can be divided into several clusters where vectors span a low-dimensional linear space and provide mathematical optimization techniques which divide a matrix into several smaller low rank matrices. However, the performance of these approaches depends on the initial values and conditions, and the accuracy of a matrix completion becomes worse when each subspace is not strictly low-dimensional due to noise. In order to improve the accuracy of a matrix completion, this paper proposes a method to reduce the dependence of initial values and to enhance robustness against noise. Numerical examples show that the proposed method achieves higher accuracy in subspace clustering and matrix completion compared to conventional methods.**

I. Introduction

This paper deals with subspace clustering for matrix completion, which is a problem of estimating missing entries in a matrix under the assumption that row or column vectors belongs to multiple low-dimensional linear spaces. While low-rank matrix approaches [1], [2], [3] have been proposed for a matrix completion, these approaches are effective for the case that row or column vectors belong to a single low-dimensional linear space and do not work well when they belong to multiple low-dimensional linear spaces. Since a matrix completion with multiple low-dimensional linear spaces has more practical applications, this paper deals with a matrix completion method with subspace clustering. Recently, a matrix completion with subspace clustering has been used in various signal processing applications, including speech restoration [4] and image restoration [5].

Some mathematical optimization-based methods have been proposed for a matrix completion with subspace clustering [5], [6]. In [6] kernel method is applied and estimates missing entries in a matrix by minimizing the rank of the kernel matrix. This method works well when both the dimensions of each subspace and the degree of a monomial are small, and requires a huge number of data (number of row or columns in a matrix) to achieve high completion accuracy. In [5] a matrix completion with subspace clustering is formulated as a problem of finding combinations of row or column vectors to divide a given matrix into several low-rank submatrices with estimating missing entries. An alternating optimization method is proposed to solve this problem and works well when a matrix is correctly divided into submatrices. However, its performance depends on the initial values and noise.

In order to improve the accuracy of a matrix completion, this paper proposes a method to reduce the dependence of initial values and to enhance robustness against noise. Numerical examples show that the proposed method achieves higher accuracy in subspace clustering and matrix completion compared to existing methods [5].

II. Subspace Clustering

We consider subspace clustering which a problem of finding L low-dimensional linear subspaces consisting of column vectors in a matrix $X \in \mathbb{R}^{M\times N}$. Takahashi et al. [5] formulated this problem as follows using the truncated nuclear norm [7],

$$\begin{aligned} \text{Find} \quad & \left\{D^{(i)}\right\}_{i=1}^{L} \\ \text{subject to} \quad & \left\|XD^{(i)}\right\|_{*,r_i} \le \varepsilon_i, i = 1,2,\dots,L, \\ & \sum_{i=1}^{L} D^{(i)} = I, \\ & (D^{(i)})_{j,j} \in \{0,1\}, \end{aligned} \tag{1}$$

where I is the identity matrix, $D^{(i)} \in [0,1]^{N\times N}$ denotes diagonal matrices, and $\|\cdot\|_{*,r}$ represents the truncated nuclear norm, which is the sum of the first (largest) singular value to the rth singular value of a matrix. $D^{(i)}$ has diagonal elements that are 0 or 1, and their sum equals the identity matrix. Using such $D^{(i)}$, $XD^{(i)}$ becomes a matrix where the columns of X corresponding to the indices where the diagonal elements of $D^{(i)}$ are 0 become zero vectors. Consequently, $XD^{(i)}$ represents a matrix that retains only the elements of the original column vectors corresponding to the indices where the diagonal elements of $D^{(i)}$ are 1, generating pseudo-submatrices corresponding to each subspace. The problem (1) describes the subspace clustering problem by imposing a low-rank constraint on these pseudo-submatrices $XD^{(i)}$ using the truncated nuclear norm. If the constant ε_i is 0, it equals to a strict low-rank constraint where the rank of $XD^{(i)}$ equal to r_i. However, in practical applications, even if $D^{(i)}$ is ideally obtained, $XD^{(i)}$ is rarely exact low-rank. Therefore Takahashi et al. [5] formulated the problem to perform low-rank approximation of $XD^{(i)}$ by setting ε_i to accommodate noise in the data. However, it is difficult to find $D^{(i)}$ such that the truncated nuclear norm of $XD^{(i)}$ is less than a certain value. Therefore they relaxed this problem by formulating it

as an objective function:

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{L} \|XD^{(i)}\|_{*,r_i} \\ \text{subject to} \quad & \sum_{i=1}^{L} D^{(i)} = I, \\ & (D^{(i)})_{j,j} \in \{0,1\}. \end{aligned} \tag{2}$$

However, there are two issues in obtaining a solution of this problem. The first issue arises from the constraints $(D^{(i)})_{j,j} \in \{0,1\}$, that is, the problem is a $0-1$ combinatorial optimization. The second issue is related to the minimization of the truncated nuclear norm. Even if $(D^{(i)})_{j,j}$s take continuous values, there is no practical method to minimize $\|XD^{(i)}\|_{r_i}$. Hence, based on the idea in [7], Takahashi et al. [5] formulated the following optimization problem using auxiliary variables $Z^{(i)}$:

$$\begin{aligned} \text{Minimize} \quad & f\left(X, \{Z^{(i)}\}_{i=1}^{L}, \{D^{(i)}\}_{i=1}^{L}\right) \\ \text{subject to} \quad & \sum_{i=1}^{L} D^{(i)} = I, \\ & (D^{(i)})_{j,j} \in [0,1]. \end{aligned} \tag{3}$$

where, the objective function f is given by

$$\begin{aligned} & f\left(X, \{Z^{(i)}\}_{i=1}^{L}, \{D^{(i)}\}_{i=1}^{L}\right) \\ & = \sum_{i=1}^{L} \left(\frac{1}{2}\|Z^{(i)} - XD^{(i)}\|_F^2 + \lambda \|Z^{(i)}\|_{*,r_i} \right), \end{aligned} \tag{4}$$

where a tehnique of the alternating direction method of multiplier (ADMM) is utilized, and its solution can be obtained through alternating optimization of $D^{(i)}$ and $Z^{(i)}$. Takahashi et al. [5] finally proposed the following subspace clustering method based on the iterative method of mathematical optimization as the minimization algorithm:

$$Z^{(i)} \leftarrow \mathcal{S}_{r_i}(XD^{(i)}) \text{ for } i = 1, \cdots, L, \tag{5}$$

$$\begin{aligned} (D^{(i)})_{j,j} \leftarrow & \frac{1}{L}\left(1 - \sum_{l=1}^{L} \frac{\langle \boldsymbol{x}_j, \boldsymbol{z}_j^{(l)} \rangle}{\langle \boldsymbol{x}_j, \boldsymbol{x}_j^{(l)} \rangle}\right) + \frac{\langle \boldsymbol{x}_j, \boldsymbol{z}_j^{(i)} \rangle}{\langle \boldsymbol{x}_j, \boldsymbol{x}_j \rangle} \\ & \text{for } i = 1, \cdots, L, \quad j = 1, \cdots, N, \end{aligned} \tag{6}$$

where $\langle , \rangle$ denotes the inner product of vectors, and $\boldsymbol{x}_j$ and $\boldsymbol{z}_j^{(i)}$ represent the j-th column vectors of X and $Z^{(i)}$, respectively. $\mathcal{S}_{r,\lambda}$ denotes a soft thresholding operator that decreases the singular values which are smaller than rth singular value by λ [7]. This thresholding operator gives $Z^{(i)}$ which minimizes the objective function for the given $D^{(i)}$. Subsequently, the above procedure provides $(D^{(i)})_{j,j}$ using the cosine similarity $\langle \boldsymbol{x}_j, \boldsymbol{z}_j^{(i)} \rangle / \langle \boldsymbol{x}_j, \boldsymbol{x}_j \rangle$ of (6). The closer the cosine similarity between the reduced non-principal components $\boldsymbol{z}_j^{(i)}$ and the original vector $\boldsymbol{x}_j$ is to 1, the closer $(D^{(i)})_{j,j}$ is to 1. In other words, $XD^{(i)}$ obtained through appropriate partitioning tends to have its non-principal components close to 0, making this cosine similarity close to 1. Equation (6) also finds $D^{(i)}$ that minimizes the objective function similarly to (5), and alternating these operations always leads to a quasi-optimal solution.

Algorithm 1 Matrix completion based on subspace clustering using the truncated nuclear norm [5]

Require: $X, \Omega, L, \alpha, \eta_\alpha, \gamma, t_{max}$

1: $t \leftarrow 0$
2: **repeat**
3: $\quad \alpha \leftarrow \alpha / \eta_\alpha$
4: $\quad t \leftarrow t + 1$
5: $\quad$ **for** $i = 1$ to L **do**
6: $\quad\quad [U, S, V] \leftarrow \text{svd}(XD^{(i)})$
7: $\quad\quad r \leftarrow \underset{\hat{r}}{\text{argmax}}\, S_{\hat{r},\hat{r}}$ s.t. $S_{\hat{r},\hat{r}} > \alpha S_{1,1}$
8: $\quad\quad S_{j,j} \leftarrow \min\left(S_{j,j} - \gamma S_{r,r}, 0\right)$ for $j > r$
9: $\quad\quad Z^{(i)} \leftarrow USV^T$
10: $\quad$ **end for**
11: $\quad$ **for** $i = 1$ to L, $j = 1$ to N **do**
12: $\quad\quad (D^{(i)})_{j,j} \leftarrow \frac{1}{L}\left(1 - \sum_{l=1}^{L} \frac{\langle \boldsymbol{x}_j, \boldsymbol{z}_j^{(l)} \rangle}{\langle \boldsymbol{x}_j, \boldsymbol{x}_j^{(l)} \rangle}\right) + \frac{\langle \boldsymbol{x}_j, \boldsymbol{z}_j^{(i)} \rangle}{\langle \boldsymbol{x}_j, \boldsymbol{x}_j \rangle}$
13: $\quad$ **end for**
14: $\quad X \leftarrow \left(\sum_{i=1}^{L} Z^{(i)} D^{(i)}\right)\left(\sum_{l=1}^{L} {D^{(l)}}^2\right)^{-1}$
15: $\quad x_{m,n} \leftarrow x^*_{m,n}$ for $(m,n) \in \Omega$
16: **until** $t_{max} < t$

Ensure: X.

Next we consider a matrix completion with subspace clustering. For matrix completion, we can modify the above method by adding the constraint $x_{m,n} = x^*_{m,n}$ for $(m,n) \in \Omega$ to (3), where Ω denotes the index set of known elements, and $x^*_{m,n}$ represents the known elements. As an optimization method for the problem with the additional constraints, Takahashi et al. [5] proposed an optimization method for X using the $Z^{(i)}$ and $D^{(i)}$ obtained from (5) and (6). For the optimization of X, a simple quadratic optimization can be performed using the following update formula:

$$\begin{aligned} X \quad & \leftarrow \quad \left(\sum_{i=1}^{L} Z^{(i)} D^{(i)}\right)\left(\sum_{l=1}^{L} {D^{(l)}}^2\right)^{-1} \\ x_{m,n} \quad & \leftarrow \quad x^*_{m,n} \quad \text{for } (m,n) \in \Omega \end{aligned} \tag{7}$$

Finally, Algorithm 1 is obtained. Note that α and η_α in Algorithm 1 represent parameters for estimating the matrix rank r in line 7. Details are written in [7]. While this works well, the clustering accuracy deteriorates if the initial values of $D^{(i)}$ are not appropriate, For example, if $D^{(i)}_{j,j}$ is set to 0, $\boldsymbol{x}_j D^{(i)}_{j,j}$ becomes a zero vector, and $\boldsymbol{z}_j^{(i)}$ is also calculated as a zero vector, resulting in a cosine similarity of 0 regardless

of whether $\boldsymbol{x}_j$ belongs to the i-th cluster. Therefore, this algorithm depends on the initial values of $D^{(i)}$. Furthermore, numerical examples show that the algorithm often fails to estimate $D^{(i)}$ if there exists noise in X.

III. Proposed Method

This section proposes a new subspace clustering method. The original idea of subspace clustering is to find the subspaces related to each cluster and the data belonging to those subspaces. In the method by Takahashi et al. [5], auxiliary variables $Z^{(i)}$ are used to represent these subspaces. Ideally, the column space of $Z^{(i)}$ matches each target subspace. However, as mentioned above, the deterioration of clustering accuracy derives from the use of auxiliary variables $Z^{(i)}$. Therefore, this paper formulates the following optimization problem using auxiliary variables $U^{(i)} \in \mathbb{R}^{M \times r}$ that can directly represent the column space to represent the subspaces.

$$\begin{aligned} \text{Minimize} \quad & g\left(X, \{D^{(i)}\}_{i=1}^{L}, \{U^{(i)}\}_{i=1}^{L}\right) \\ \text{subject to} \quad & \sum_{i=1}^{L} D^{(i)} = I \\ & D^{(i)} \in \{0,1\} \\ & {U^{(i)}}^T U^{(i)} = I \end{aligned} \tag{8}$$

$$\begin{aligned} & g\left(X, \{D^{(i)}\}_{i=1}^{L}, \{U^{(i)}\}_{i=1}^{L}\right) \\ =& \sum_{i=1}^{L}\sum_{j=1}^{N} D_{j,j}^{(i)} \left\| \left(I - U^{(i)}{U^{(i)}}^T\right) \boldsymbol{x}_j \right\|_2^2 \\ =& \sum_{i=1}^{L} \left\| \left(I - U^{(i)}{U^{(i)}}^T\right) X D^{(i)} \right\|_F^2 \end{aligned} \tag{9}$$

The idea of minimizing the Frobenius norm to minimize the matrix rank is originally proposed in [3], and minimizing the objective function g is equivalent to minimizing the rank of the submatrix $XD^{(i)}$. A solution of this problem can be obtained by alternating optimization as follows,

$$\begin{aligned} [\tilde{U}, \tilde{S}, \tilde{V}] &\leftarrow \text{svd}(XD^{(i)}) \\ U^{(i)} &\leftarrow \tilde{U}_{:,1:r} \\ & \text{for } i = 1, \cdots, L \end{aligned} \tag{10}$$

$$(D^{(i)})_{j,j} \leftarrow \begin{cases} 1 & \text{if } i = \underset{\hat{i}}{\text{argmin}} \left\| \left(I - U^{(\hat{i})}{U^{(\hat{i})}}^T\right) \boldsymbol{x}_j \right\|_2^2 \\ 0 & \text{otherwise} \end{cases} \\ \text{for } j = 1, \cdots, N \tag{11}$$

where $A_{:,1:r}$ denotes the matrix formed by the first to r-th columns of the matrix A. The singular values $\tilde{S}_{j,j}$ obtained by the singular value decomposition $\text{svd}(XD^{(i)})$ are assumed to be arranged in descending order. The above equations are methods for alternately solving for $U^{(i)}$ and $D^{(i)}$. Let us consider the problem of minimizing (9) with respect to $U^{(i)}$ for given $D^{(i)}$. Based on classical theorems concerning matrices [8], the optimal solution can be obtained through singular value decomposition of the matrix. Let $XD^{(i)}$ be decomposed as $XD^{(i)} = \tilde{U}\tilde{S}\tilde{V}^T$, then the optimal solution $U^{(i)}$ is equal to the matrix formed by the column vectors corresponding to the first to r-th principal components of $\tilde{U}$. Next we consider the problem of minimizing (9) with respect to $D^{(i)}$ for given $U^{(i)}$. This problem can be performed independently for each j, and the minimum solution is obtained by setting $(D^{(i)})_{j,j} = 1$ for the i that minimizes $\|(I - U^{(i)}{U^{(i)}}^T)\boldsymbol{x}_j\|_2^2$ and 0 otherwise. By repeating the above two schemes until $D^{(i)}$ converges, a solution can be obtained. In practice, since r is unknown, we propose a method that starts with $r = 1$ and iteratively repeats the above schemes. Finally this paper proposes Algorithm 2. An advantage of Algorithm 2 is that it requires fewer parameters to be specified compared to existing methods.

Algorithm 2 Adaptive subspace reconstruction algorithm (proposed method)

Require: X, L, t_{max}
  for $r = 1$ to $M - 1$ **do**
    $t \leftarrow 0$
    repeat
      $t \leftarrow t + 1$
      $D_{old}^{(i)} \leftarrow D^{(i)}$ for $i = 1, \cdots, L$
      for $i = 1$ to L **do**
        $[\tilde{U}, \tilde{S}, \tilde{V}] \leftarrow \text{svd}(XD^{(i)})$
        $U^{(i)} \leftarrow \tilde{U}_{:,1:r}$
      end for
      for $i = 1$ to L, $j = 1$ to N **do**
        $(D^{(i)})_{j,j} \leftarrow 1$
        if $i = \underset{\hat{i}}{\text{argmin}} \left\| \left(I - U^{(\hat{i})}{U^{(\hat{i})}}^T\right) \boldsymbol{x}_j \right\|_2^2$
        0 otherwise
      end for
    until $t_{max} < t$ or $D^{(i)} = D_{old}^{(i)}$
  end for
Ensure: $\{D^{(i)}\}_{i=1}^{L}$.

As described in section 2, we consider performing matrix completion using subspace clustering with Algorithm 2. If $D^{(i)}$ is appropriately obtained by Algorithm 2, the matrix completion problem can be solved by imposing the constraint $x_{m,n} = x_{m,n}^*$ for $(m,n) \in \Omega$ on (8), and minimizing (9) with respect to X, which performs matrix rank minimization within each cluster. Based on this idea, we present Algorithm 3 for matrix completion with subspace clustering, where the truncated nuclear norm minimization is applied. In this algorithm, it is not necessary to update $D^{(i)}$ if there is no significant change in X, and therefore the algorithm executes Algorithm 2 only when there is a significant change in X.

Algorithm 3 Matrix completion using adaptive subspace reconstruction algorithm (proposed method)

Require: $X, \Omega, L, \alpha, \eta_\alpha, \gamma, t_{max}$

$t \leftarrow 0$

$r_{max} \leftarrow 1, r_{max\,old} \leftarrow 0$

repeat

$t \leftarrow t+1$

if $r_{max} \neq r_{max\,old}$ **then**

obtain $\{D^{(i)}\}_{i=1}^{L}$ using Algorithm 2

end if

for $i = 1$ to L **do**

$[U, S, V] \leftarrow \mathrm{svd}(XD^{(i)})$

$r_i \leftarrow \underset{\hat{r}}{\mathrm{argmax}}\, S_{\hat{r},\hat{r}}$ s.t. $S_{\hat{r},\hat{r}} > \alpha S_{1,1}$

$S_{j,j} \leftarrow \min\left(S_{j,j} - \gamma S_{r_i,r_i}, 0\right)$ for $j > r_i$

$Z^{(i)} \leftarrow USV^T$

end for

$X \leftarrow \sum_{i=1}^{L} Z^{(i)}$

$x_{m,n} \leftarrow x^*_{m,n}$ for $(m,n) \in \Omega$

$r_{max\,old} \leftarrow r_{max}$

$r_{max} \leftarrow \max r_i$

until $t_{max} < t$

Ensure: X.

IV. Numerical Examples

This section demonstrates the effectiveness of the proposed algorithm by comparing with Algorithm 1. In order to confirm how the estimation accuracy changes with the matrix rank, we apply algorithms with matrix ranks $r = 10, 11, \cdots, 15$. In all experiments, matrix size is set to 150×150, and the total number of clusters is set to 50.

First, we compare the estimation accuracy of subspace clustering. Assuming that the true matrix X is given, we examine performances of subspace clustering for matrices with no missing entries. White noise $SNR = 20$ is added to each column vector of the matrix, and Algorithm 1 and Algorithm 2 are applied. The maximum number of iterations t_{max} for both algorithms is set to $t_{max} = 1,000$, and the estimation accuracy of $D^{(i)}$ is compared. Since the $D^{(i)}$ obtained by Algorithm 1 can take values in $[0, 1]$, we calculate the confusion matrix C with respect to the true value $D^{(i)}_{true}$ as $C_{k,l} = \sum_{j=1}^{N} (D^{(k)})_{j,j} (D^{(l)}_{true})_{j,j}$, and compare the inverse condition number of this matrix. If $D^{(i)}$ has even one different entry from an entry of $D^{(i)}_{true}$, the inverse condition number of C will be 0.972 or less. Table 1 shows the average of the inverse condition numbers of C obtained over 100 trials whose initial entries of $D^{(i)}$ are generated according to uniform random distribution for each method. From this table, it can be seen that the proposed method performs subspace clustering with sufficient accuracy for noise existing case.

TABLE I
Average of the inverse condition numbers of confusion matrix C obtained by subspace clustering in 100 trials

Method \ rank r	10	11	12	13	14	15
Algorithm 1 [5]	0.16	0.19	0.22	0.18	0.19	0.20
Algorithm 2	1.00	1.00	1.00	1.00	0.98	0.97

Next we compare Algorithm 1 with Algorithm 3 for matrix completion with subspace clustering ith a missing rate of 50%. Similar to Table 1, we conducted 100 trials and calculated the average of normalized mean square error (NMSE) between the matrix X obtained by each algorithm and the true value X^*.

$$NMSE = \sum_{(m,n)\notin\Omega} (x_{m,n} - x^*_{m,n})^2 \;/\; \sum_{(m,n)\notin\Omega} (x^*_{m,n})^2$$

Figure 1 shows the average of NMSE for each algorithm. As can be seen from the figure, Algorithm 1 gives a significant deterioration in completion accuracy for $r \geq 12$. In contrast, the average of NMSE for the proposed method is less than 0.1 for all ranks. These results demonstrate the effectiveness of the proposed method.

V. Conclusions

This paper deals with subspace clustering for matrix completion. One of the existing methods is the subspace clustering method using an optimization problem with the truncated nuclear norm. This method is based on mathematical optimization and has the advantage of guaranteeing a quasi-optimal solution. However, due to the nature of the algorithm, it tends to depend on initial values and is strongly affected by noise, which can result in degraded matrix completion accuracy. This paper proposes a new subspace clustering method using auxiliary variables representing each subspace and apply it to matrix completion. Numerical experiments demonstrated that the proposed method showed higher estimation accuracy compared to existing methods.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP24K14845.

References

[1] J. F. Cai, E. J. Candés, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optimization*, vol.20, no.4, pp.1956–1982, 2010.

[2] K. Mohan and M. Fazel, "Iterative reweighted algorithms for matrix rank minimization," *J. Machine Learning Research (JMLR)*, vol.13, no.1, pp.3441–3473, 2012.

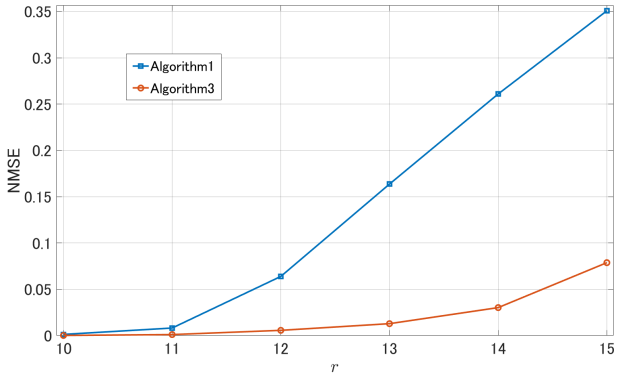


Fig. 1. Average NMSE comparison of each method

[3] K. Konishi, "Parallel GPU implementation of null space based alternating optimization algorithm for large-scale matrix rank minimization," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing.*, pp.3689-3692, 2012.

[4] R. Sasaki, K. Konishi, T. Takahashi, and T. Furukawa, "Multiple matrix rank minimization approach to audio declipping," *IEICE Transactions on Information and Systems*, vol.E101-D, no.3, pp.821–825, 2018.

[5] T. Takahashi, K. Konishi, K. Uruma, and T. Furukawa: Multiple Subspace Model and Image-Inpainting Algorithm Based on Multiple Matrix Rank Minimization, *IEICE Trans. on Info. and Syst.*, vol.E103.D, no.12, pp.2682–2692, 2020

[6] G. Ongie, R. Willett, R. D. Nowak, and L. Balzano, "Algebraic variety models for high-rank matrix completion," *Proc. of the 34th Int. Conf. on Machine Learning (PMLR)*, vol.70, 2691-2700 (2017)

[7] K. Konishi, K. Uruma, T. Takahashi, and T. Furukawa, "Iterative partial matrix shrinkage algorithm for matrix rank minimization," *Signal Processing*, vol.100, pp.124–131, 2014.

[8] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol.1, pp.211–218, 1936.

Ablation Study to Derive a Computationally Efficient Deep Learning-Based Super-Resolution Approach

Asfa Jamil* and Alessandro Artusi†
* CYENS- Center of Excellence E-mail: asfajamil164@outlook.com Tel/Fax: +353-852006837
† CYENS- Center of Excellence E-mail: a.artusi@cyens.org.cy

***Abstract*—Recent research on single-image super-resolution (SR) shows that deep learning-based methods outperform state-of-the-art techniques but at the cost of increased memory consumption and computational complexity. This results in longer training and inference times and higher GPU memory requirements compared to traditional approaches. Network architecture modifications can impact performance, complexity, and memory needs. This paper explores enhancing SR performance using a simple yet efficient deep-learning SR model, focusing on local and global connections in residual networks, channel attention mechanisms, and up-sampling techniques. Our efficient, lightweight, locally dense residual SR architecture achieves performance comparable to state-of-the-art models, reducing spatial complexity by up to 1/6 and inference time by half compared to the baseline.**

I. Introduction and Related Work

Single Image Super-Resolution (SR) enhances the resolution of low-resolution images to generate high-resolution counterparts [2]. This process focuses on restoring high-frequency information lost during image acquisition or compression [3], significantly improving visual quality. SR has diverse applications in computer vision, including enhancing medical images, improving surveillance footage, and analyzing satellite imagery [4],

Deep learning-based SR methods have emerged as superior, leveraging neural networks to learn complex mappings from low-resolution to high-resolution images, producing sharper outputs. Convolutional Neural Network (CNN)-based architectures, such as SRCNN [5] and FSRCNN [6], stack convolutional layers to map low to high-resolution images. ResNet and Recurrent Neural Network (RNN)-based architectures, like DRCN [7] and DRRN [8], use recursion and residual learning for better efficiency. EDSR [9] simplifies SRResNet, and CARN [10] offers a lightweight network for fast SR. Dense-connection-based architectures, such as SRDenseNet [11], improve learning by propagating information from early to later layers. Generative Adversarial Networks (GANs), like Relativistic GAN [12] and ESRGAN [13], produce realistic textures by using a generator and discriminator network. Channel attention-based architectures, like RCAN [14], focus on informative features using attention mechanisms.

Despite their effectiveness, deep learning methods face challenges due to their complexity, leading to high memory usage and computational inefficiency during inference. These multi-layered neural networks demand substantial computational resources during training and execution, posing scalability constraints. Studies show that modifying neural network architecture, loss function, and learning approach can affect performance quality, network complexity, and memory requirements [12]. Previous research [15] demonstrated that minimizing the complexity of state-of-the-art deep-learning SR methods, like DRLN [1], can achieve cost-effective solutions without sacrificing performance. This approach reduced spatial complexity by up to 1/6 and inference time by half compared to the baseline model.

Inspired by this, we propose a novel deep-learning SR model that minimizes spatial complexity and memory usage while enhancing computational efficiency. We tested the model on various datasets and up-sampling scales (e.g., 2×, 3×, 4×, 8×) and compared it with top-performing SR methods. Its efficiency makes it suitable for integration into scenarios such as encoding/decoding standards, editing tools, satellite imaging, and embedded systems.

The main contributions of our work are: noitemsep

1) An ablation study to identify optimal combinations for a DNN-based SR method.
2) Investigating strategies like local and global dense residual connections, pre/post and iterative up-sampling methods, and various pooling strategies in Laplacian attention modules to enhance SR performance.
3) Evaluating how these strategies impact the network's structural complexity, spatial complexity, and computational cost.
4) Developing an SR model that achieves comparable results to state-of-the-art models while reducing spatial complexity by one-sixth and halving training/inference time.

This paper is structured as follows: Section II presents our proposed model. Section III outlines experimental procedures and results. Section IV concludes the paper.

II. Proposed Architecture

The overall deep-learning structure of the proposed model is depicted in Figure 2. The architecture incorporates local dense cascading residual connections, enabling the network to effectively capture complex patterns. An attention mechanism is also employed that selectively emphasizes important regions of the input image, enhancing feature relevance. The network employs an iterative up-sampling technique to incrementally improve image resolution and a sub-pixel up-sampling technique to rearrange pixel values and enhance resolution. Finally, the network is trained using an L1 loss function, which

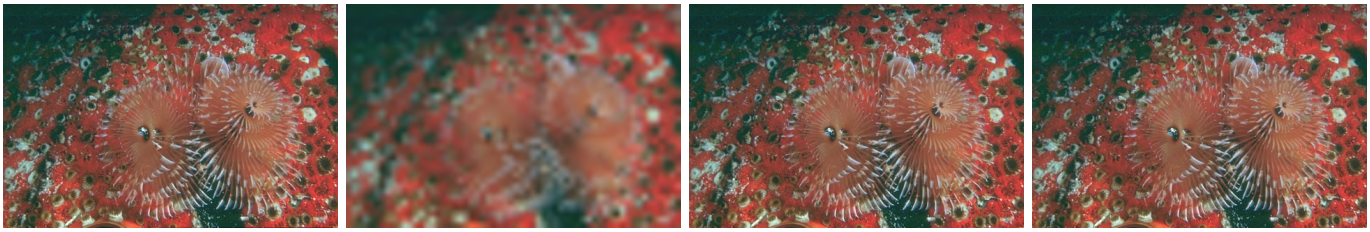


Fig. 1. Super-resolution is a process to increase the spatial resolution of an input image, with the intent of improving its overall spatial information. Here we show results comparing the ground truth high resolution image (left), compared to the up-sampled images using (left-center) bi-cubic interpolation, (right-center) a state-of-the-art deep-learning based approach [1] and (right) the proposed solution.

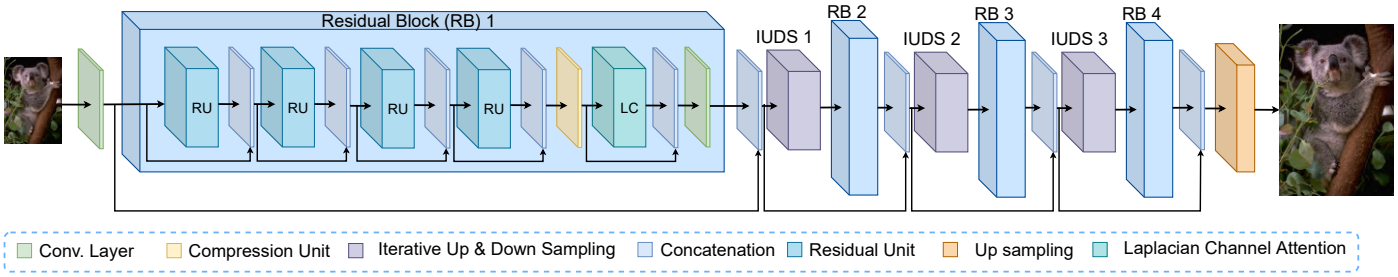


Fig. 2. The proposed SR architecture includes key components: the Residual Block (RB) and Iterative Up Down-Sampling (IUDS). The RB contains the Residual Unit (RU) and Laplacian Channel Attention (LC), which uses a Laplacian pyramid mechanism for multi-scale behavior. The IUDS incorporates up-sampling and down-sampling mechanisms.

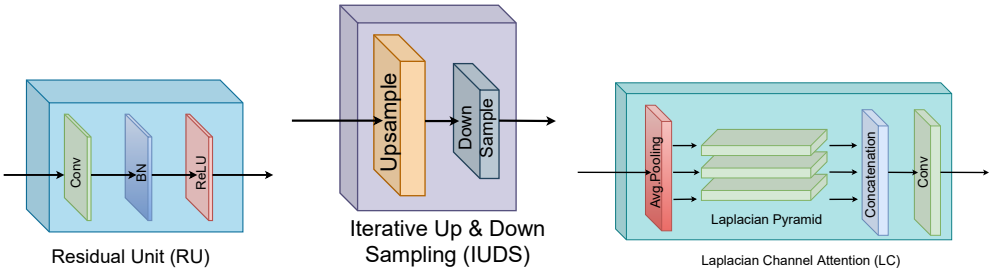


Fig. 3. Major components of the Residual Block: (left) Iterative Up-Sampling; (center) Residual Unit; (right) Laplacian Channel Attention.

measures the absolute differences between the predicted and ground-truth images. Overall, our proposed architecture is designed to produce high-quality output images while minimizing the network's complexity and memory requirements.

Cascading Residual over Residual Block Our model is composed of four cascading residual over residual blocks, which we will continue to refer to as Residual Blocks (RB) for simplicity. These blocks are designed to gradually learn complex features in the input image, allowing the model to generate high-quality super-resolved images. Each RB in our model consists of 4 Residual Units (RU), each followed by a concatenation layer. This design allows the model to learn residual functions more effectively, by incorporating information from earlier into later ones.

Residual Unit Each Residual Block (RB) is built from a lightweight Residual Unit (RU), consisting of one convolutional layer, one ReLU layer, and one batch normalization layer (Figure 3, left). This streamlined design enhances efficiency and ease of training. As shown in Table I for the 2X up-sampling factor, our model achieves comparable PSNR scores while halving memory usage and being 25% more computationally efficient. Figure 4(a) shows the visual qualitative result of the lightweight RU compared to the traditional one, with clearer details and less noise on the zebra skin.

Laplacain Channel (LC) attention module Following the last Residual Unit (RU), we include a compression layer, a Laplacian attention unit, a concatenation layer, and a convolutional layer.

The Laplacian Channel (LC) attention module enhances significant image features using a Laplacian pyramid approach, which weights important sub-band features. It employs a global descriptor for image statistics and a gating mechanism for channel dependencies (Figure 3, right).

The global descriptor processes the output from the compression unit, which has c feature maps of size $h \times w$. The Laplacian pyramid weights are computed using convolutional layers with specified kernel dilation. The pyramid output is convolved, down-sampled, up-sampled, and passed through a sigmoid activation function, adaptively re-scaling the output.

In our findings, tt has demonstrated effectiveness on the

MANGA109 dataset, providing crisper and better-represented details, as shown in Figure 4(b). The upper zoomed area in Figure 4(b) shows the improved detail using the LC module compared to the bottom zoomed area without it.

Local vs Global dense connections Our architecture emphasizes localized dense connections over global ones, featuring more dense local connections between residual units than global connections between residual blocks. Our findings highlight that this approach preserves intricate details, optimizes parameter efficiency, enhances gradient flow, and maintains feature hierarchy with reduced computational overhead. This underscores the importance of focusing on local dense mechanisms for tasks that require high-resolution detail conservation. Figure 4(c) shows the results for the 2x up-sampling factor case. When long residual connections are used (top zoomed area), crisper details are observed compared to the bottom area in the same figure, where long residual connections are not used.

Iterative Up/Down-Sampling Framework

We propose iterative up-sampling and down-sampling modules (Figure 3, center) to enhance network performance. These modules adjust the resolution of feature maps during processing.

Up-Sampling Module: The up-sampling module uses a sub-pixel layer to rearrange pixels and increase feature map resolution:

$$F_{up} = P \times (R \times F_{low})$$

where F_{up} is the up-sampled feature map, P is the pixel shuffle operation, R is the up-scaling factor, and F_{low} is the input low-resolution feature map. This is followed by batch normalization and ReLU activation.

Down-Sampling Module: The down-sampling module uses a MaxPool2d layer to reduce feature map resolution:

$$F_{down} = \max_{i,j \in W} (F_{up}(i,j))$$

where F_{down} is the down-sampled feature map, F_{up} is the input high-resolution feature map, W is the pooling window, and i and j are indices in the pooling window.

Table I and Figure 4(d) show that the IUDS module improves image details without increasing model complexity, memory usage, or computation time.

TABLE I
PERFORMANCE COMPARISONS FOR 2X UP-SAMPLING FACTOR ON THE B100 DATASET [16].

Metrics	TRU	LRU	No IUS	With IUS
GPU memory(MB)	4414	2555	2417	2555
Run-time (s)	0.27	0.187	0.18	0.187
PSNR	32.097	32.069	32.055	32.069

TRU = Traditional RU, LRU = Lightweight RU, IUS = Iterative up-sampling

III. EXPERIMENTAL RESULTS

We have tested the proposed model on a variety of standard benchmark data sets, i.e., URBAN100, B100, SET14, and SET5 . To demonstrate its robustness we have applied the proposed model to different types of up-sampling factors from $2\times$ to $8\times$. We have evaluated the proposed models on the following characteristics: image quality (PSNR and SSIM scores), spatial complexity, and inference time. Spatial complexity is measured as occupied on-board GPU memory, and the methodology to calculate GPU memory occupation is the one used in [25].

Training Settings: The experiments have been performed on a computer with an Intel Core i7-8700k processor running at 3.70 GHz, using 32 GB RAM, and NVIDIA GeForce RTX 2080Ti GPU. The model has been trained using the DIV2k[26] data set. In our training settings, we maintain a batch size of 4 and start with an initial learning rate of 10^{-4}. This rate is halved every 2^4 iterations. We employ the ADAM optimizer, setting the initial parameters to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$.

Overall Comparison for $2\times$, $3\times$, $4\times$, and $8\times$ Up-Sampling: This section compares the proposed model with state-of-the-art SR methods in terms of PSNR, SSIM, inference time, and GPU memory consumption for $2\times$, $3\times$, $4\times$, and $8\times$ up-sampling factors. First, the quantitative analysis of the proposed method is shown in Section III-1. Then, the spatial complexity and efficiency is analyzed in Section III-2, and finally, Section III-3 provides the visual quality comparison at different up-scaling factors, all compared to the state-of-the-art methods.

1) Quantitative Analysis of Proposed Model: Tables III to VI compare the PSNR and SSIM scores of our model against existing high-performance methods.

For 2x upsampling, our model performs competitively across all datasets, achieving PSNR scores of 33.50 dB on SET14 and 31.96 dB on URBAN100, with SSIM scores of 0.92 for both. These results are comparable to EDSR and RCAN. At 3x upsampling, our model records PSNR values of 30.28 dB on SET14 and 29.04 dB on BSD100, with SSIM scores of 0.84 and 0.80, respectively, closely matching RDN and RCAN. For 4x upsampling, our model achieves PSNR scores of 28.72 dB on SET14 and 26.34 dB on URBAN100, with SSIM scores of 0.78 and 0.80, slightly trailing DRLN and RCAN. At 8x upsampling, our model reaches PSNR scores of 24.84 dB on SET14 and 22.97 dB on URBAN100, with SSIM scores of 0.62 and 0.61, slightly below DRLN and RCAN. These comparisons show that our model consistently delivers high-quality images across various up-sampling sizes, demonstrating its effectiveness and adaptability.

2) Spatial Complexity and Efficiency Analysis: The spatial complexity of a deep-learning model, measured by on-board GPU memory consumption [25], is crucial for practical applications. Figures 5 and 7 show these results.

Figure 5 highlights the spatial complexity of our proposed model compared to SOTA methods across resolutions of 320x240, 720x480, and 1920x1080. Our model uses significantly less memory, about $1/13^{th}$ and $1/5^{th}$ of the GPU memory compared to EDSR and RCAN at 320x240, maintaining efficiency even at 1920x1080.

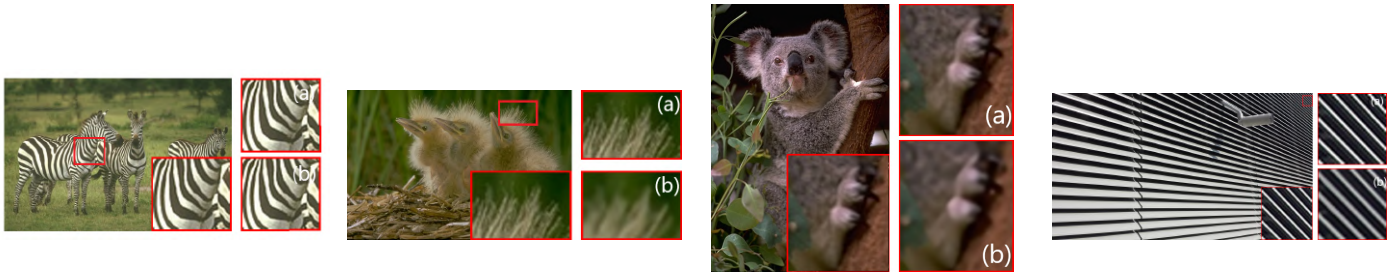


Fig. 4. Qualitative comparisons on 3x up-sampling factor: (a) Traditional vs. lightweight RU, (b) Incorporating Laplacian attention unit, (c) SR technique with and without global residual connections, (d) Proposed model with and without IUDS unit.

TABLE II

QUANTITATIVE EVALUATION OF COMPETING METHODS ON 2X AND 3X UP-SAMPLING FACTORS. WE REPORT THE PERFORMANCE OF STATE-OF-THE-ART ALGORITHMS AND THE PROPOSED MODEL ON WIDELY USED PUBLICLY AVAILABLE DATASETS, IN TERMS OF PSNR (IN DB) AND SSIM SCORES.

TABLE III

2X UP-SAMPLING FACTOR

Method	SET5		SET14		BSD100		URBAN100	
Bicubic	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403
SRCNN [17]	36.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946
FSRCNN [18]	37.05	0.9560	32.66	0.9090	31.53	0.8920	29.88	0.9020
VDSR [19]	37.53	0.9590	33.05	0.9130	31.90	0.8960	30.77	0.9140
LapsRN [20]	37.52	0.9591	33.08	0.9130	31.08	0.8950	30.41	0.9101
EDSR [21]	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351
RDN [22]	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353
RCAN [23]	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384
CARN[24]	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256
DRLN[1]	38.27	0.9616	34.28	0.9231	32.44	0.9028	33.37	0.9390
Our Model	37.95	0.9611	33.50	0.9212	32.11	0.9015	32.96	0.9266

TABLE IV

3X UP-SAMPLING FACTOR

Method	SET5		SET14		BSD100		URBAN100	
Bicubic	30.39	0.8682	27.55	0.7742	27.21	0.7385	24.46	0.7349
SRCNN[17]	32.75	0.9090	29.30	0.8215	28.41	0.7863	26.24	0.7989
FSRCNN[18]	33.18	0.9140	29.37	0.8240	28.53	0.7910	26.43	0.8080
VDSR[19]	33.67	0.9210	29.78	0.8320	28.83	0.7990	27.14	0.8290
LapsRN[20]	33.82	0.9227	29.87	0.8320	28.82	0.7980	27.07	0.8280
EDSR [21]	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653
RDN[22]	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653
RCAN [23]	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702
CARN[24]	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493
DRLN [1]	34.78	0.9303	30.73	0.8488	29.36	0.8117	29.21	0.8722
Our Model	34.30	0.9300	30.28	0.8460	29.04	0.8112	28.10	0.8689

TABLE V

4X UP-SAMPLING FACTOR

Method	SET5		SET14		BSD100		URBAN100	
Bicubic	28.42	0.8104	26.00	0.7027	25.96	0.6675	23.14	0.6577
SRCNN [17]	30.48	0.8628	27.50	0.7513	26.90	0.7101	24.52	0.7221
FSRCNN [18]	30.72	0.8660	27.61	0.7550	26.98	0.7150	24.62	0.7280
VDSR [19]	31.35	0.8830	28.02	0.7680	27.29	0.7276	25.18	0.7540
LapSRN [20]	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560
EDSR [21]	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033
RDN [22]	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028
RCAN [23]	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087
CARN [24]	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837
DRLN [1]	32.63	0.9002	28.94	0.7900	27.83	0.7444	26.98	0.8119
Our Model	32.19	0.8987	28.72	0.7860	27.57	0.7435	26.34	0.8067

TABLE VI

8X UP-SAMPLING FACTOR

Method	SET5		SET14		BSD100		URBAN100	
Bicubic	24.40	0.6580	23.10	0.5660	23.67	0.5480	20.70	0.5160
SRCNN [17]	25.33	0.6900	23.76	0.5910	24.13	0.5660	21.29	0.5440
FSRCNN [18]	20.13	0.5520	19.75	0.4820	24.21	0.5680	21.32	0.5380
VDSR **vdsr**	25.93	0.7240	24.26	0.6140	24.49	0.5830	21.70	0.5710
LapSRN [20]	26.15	0.7380	24.35	0.6200	24.54	0.5860	21.81	0.5810
EDSR [21]	26.96	0.7762	24.91	0.6420	24.81	0.5985	22.51	0.6221
RCAN [23]	27.31	0.7878	25.23	0.6511	24.98	0.6058	23.00	0.6452
DRLN [1]	27.36	0.7882	25.34	0.6531	25.01	0.6057	23.06	0.6471
Our Model	26.94	0.7700	24.84	0.6290	24.71	0.5990	22.97	0.6151

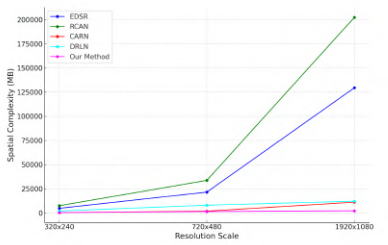


Fig. 5. Comparison of spatial complexity of the proposed model and state-of-the-art at resolutions 320x240, 720x480, and 1920x1080.

Figures 7 provide a detailed comparison of spatial complexity. At 4x upscaling on the B100 dataset (Figure 7), our model operates with approximately $1/5^{th}$ the spatial complexity of EDSR and RCAN. For 8x upscaling on the Urban100 dataset, our model requires just $1/6^{th}$ the spatial complexity of models like RCAN, while delivering impressive PSNR and SSIM scores.

Overall, our model consistently delivers high-quality super-resolution with significantly reduced spatial complexity compared to SOTA models such as RCAN, EDSR, and DRLN.

We also compared our model's computational efficiency in terms of inference time. Figure 8 shows the inference time per image and PSNR score on the B100 dataset [16]. The proposed model reduces inference time by approximately 50% across 2x, 3x, 4x, and 8x resolutions.

3) Comparing Visual Quality at Different Up-sampling Levels: Figure 6 visually compares our model's performance at 2x, 3x, 4x, and 8x upscaling against benchmarks: (a) ground truth, (b) bicubic interpolation, (c) SRCNN, (d) CARN, (e) EDSR, (f) RCAN, and (g) DRLN, with our model shown in (h). These comparisons highlight our model's ability to enhance image clarity, outperforming methods (c) and (d), and competing well with methods (e), (f), and (g). At 8x upscaling, our model matches ground truth and DRLN quality while improving computational efficiency by 50% and reducing spatial complexity by nearly $1/5^{th}$.

IV. CONCLUSIONS

This paper analyzes the architecture of state-of-the-art deep learning SISR models, focusing on performance, memory needs, and computational costs. We propose a lightweight

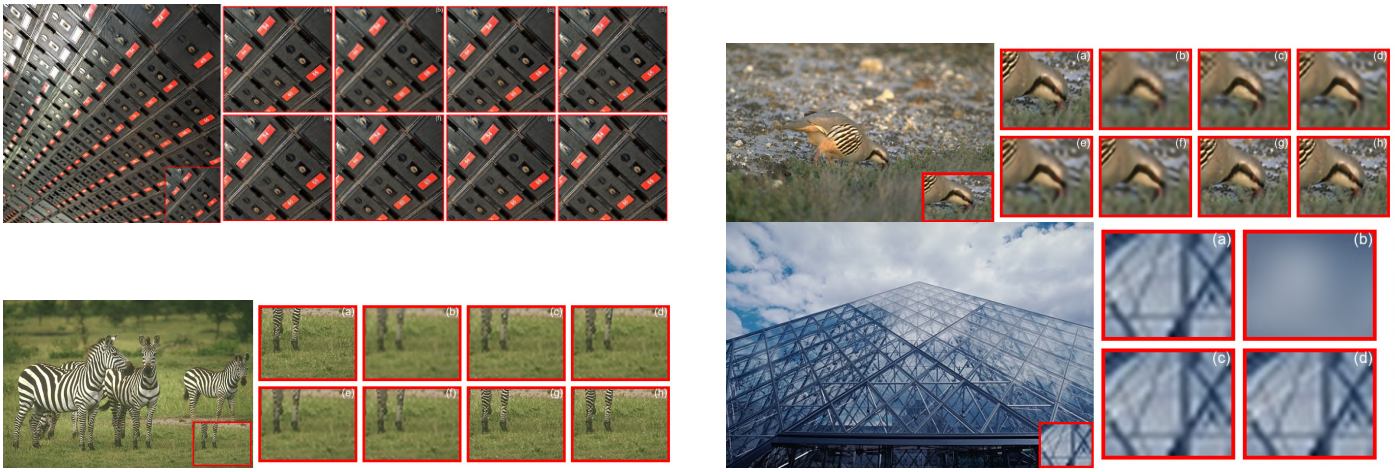


Fig. 6. Qualitative comparison of our models with other works on super-resolution for different up-sampling factors: (a) 2x, (b) 3x, (c) 4x, and (d) 8x. Each subfigure demonstrates the performance of our proposed model relative to existing methods, featuring (a) ground truth, (b) bicubic interpolation, and results from various methods including SRCNN, CARN, EDSR, RCAN, and DRLN, with our proposed model shown last.

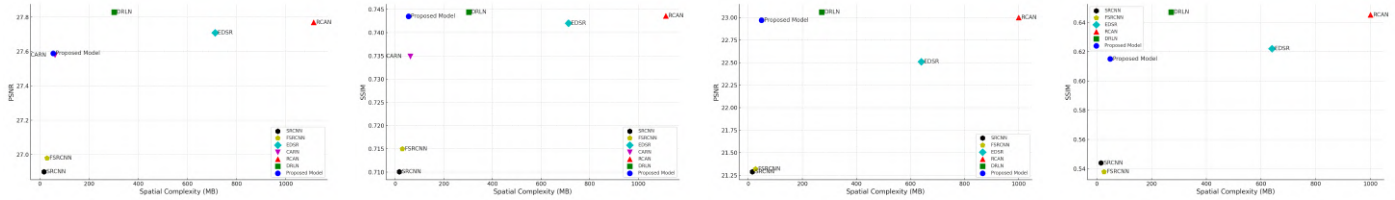


Fig. 7. (a) PSNR vs. Spatial Complexity at 4x resolution for B100 dataset. (b) SSIM vs. Spatial Complexity at 4x resolution for B100 dataset. (c) PSNR vs. Spatial Complexity at 8x resolution for Urban100 dataset. (d) SSIM vs. Spatial Complexity at 8x resolution for Urban100 dataset.

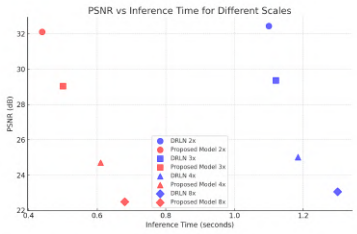


Fig. 8. Comparison of PSNR vs. inference time between state-of-the-art DRLN and proposed model at 2x, 3x, 4x, and 8x resolution for B100 dataset.

SISR model that is spatially efficient and computationally effective. Our model reduces inference time by approximately 50% compared to DRLN and supports upscaling factors from 2x to 8x, as demonstrated on benchmark datasets. It consumes only 1/6th of the GPU memory of state-of-the-art models while maintaining competitive PSNR and SSIM scores. Our model achieves superior visual quality compared to other lightweight alternatives and matches the performance of more computationally intensive models like EDSR, RCAN, and DRLN.

REFERENCES

[1] S. Anwar and N. Barnes, "Densely residual laplacian super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[2] C. Zhong and Y. Zhou, "Progressive low/high-resolution space attention fusion network for single image super-resolution," *Journal of Physics: Conference Series*, vol. 1828, no. 1, pp. 12–21, Feb. 2021. DOI: 10.1088/1742-6596/1828/1/012021. [Online]. Available: http://dx.doi.org/10.1088/1742-6596/1828/1/012021.

[3] H. Liu, Z. Fu, J. Han, L. Shao, S. Hou, and Y. Chu, "Single image super-resolution using multi-scale deep encoder–decoder with phase congruency edge map guidance," *Information Sciences*, vol. 473, pp. 44–58, 2019, ISSN: 0020-0255. [Online]. Available: https://doi.org/10.1016/j.ins.2018.09.018.

[4] Y. Li, B. Sixou, and F. Peyrin, "A review of the deep learning methods for medical images super resolution problems," *Irbm*, vol. 42, no. 2, pp. 120–133, 2021.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[6] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Eu-*

ropean conference on computer vision, Springer, 2016, pp. 391–407.

[7] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[8] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3147–3155.

[9] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[10] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 252–268.

[11] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4799–4807.

[12] S. Zhang, D. Cheng, D. Jiang, and Q. Kou, "Least squares relativistic generative adversarial network for perceptual super-resolution imaging," *IEEE Access*, vol. 8, pp. 185 198–185 208, 2020.

[13] X. Wang, K. Yu, S. Wu, *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[14] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[15] A. Jamil and A. Artusi, "Computational efficient deep learning-based super resolution approach," in *Real-time Processing of Image, Depth and Video Information 2023*, M. F. Carlsohn, Ed., International Society for Optics and Photonics, vol. 12571, SPIE, 2023, 125710E. DOI: 10.1117/12.2665645. [Online]. Available: https://doi.org/10.1117/12.2665645.

[16] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, IEEE, vol. 2, 2001, pp. 416–423.

[17] C. M. Ward, J. Harguess, B. Crabb, and S. Parameswaran, "Image quality assessment for determining efficacy and limitations of super-resolution convolutional neural network (srcnn)," in *Applications of Digital Image Processing XL*, SPIE, vol. 10396, 2017, pp. 19–30.

[18] L. S. Passarella, S. Mahajan, A. Pal, and M. R. Norman, "Reconstructing high resolution esm data through a novel fast super resolution convolutional neural network (fsrcnn)," *Geophysical Research Letters*, vol. 49, no. 4, e2021GL097571, 2022.

[19] S. Hitawala, Y. Li, X. Wang, and D. Yang, "Image super-resolution using vdsr-resnext and srcgan," *arXiv preprint arXiv:1810.05731*, 2018.

[20] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.

[21] B. M. Kuriakose *et al.*, "Edsr: Empowering super-resolution algorithms with high-quality div2k images," *Intelligent Decision Technologies*, no. Preprint, pp. 1–15,

[22] G. Chen, L. Zhang, M. Sun, Y. Gao, P. N. Michelini, and Y. Wu, "Single-image hdr reconstruction with task-specific network based on channel adaptive rdn," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 398–403.

[23] Z. Lin, P. Garg, A. Banerjee, *et al.*, "Revisiting rcan: Improved training for image super-resolution," *arXiv preprint arXiv:2201.11279*, 2022.

[24] Y. Li, E. Agustsson, S. Gu, R. Timofte, and L. Van Gool, "Carn: Convolutional anchored regression network for fast and accurate single image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[25] A. Jamil and A. Artusi, "Computational efficient deep learning-based super resolution approach," in *Real-time Processing of Image, Depth and Video Information 2023*, SPIE, vol. 12571, 2023, pp. 114–121.

[26] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 114–125.

Postoperative Delirium Prediction Based on Preoperative Electrocardiogram and Electroencephalogram

Shogo Mito*, Miho Miyajima†, Hirofumi Tomioka†, Hitomi Sato†, Takashi Takeuchi†, Hitoshi Muto†, Yuji Kabasawa†, Hiroyuki Harada†, Kana Eguchi*, Shota Kato*, Manabu Kano*
* Kyoto University, Japan
E-mail: manabu@human.sys.i.kyoto-u.ac.jp
† Tokyo Medical and Dental University, Japan

***Abstract*—Preventing postoperative delirium is of great importance because its onset causes significant social and health problems. We aimed to develop a model to predict the postoperative delirium using data measured preoperatively with electrocardiogram (ECG) sensors and simplified electroencephalogram (EEG) sensors in a less invasive situation. The data were collected from patients undergoing oral surgery, specifically from 59 patients (including 8 who developed delirium) from 9 AM until 1 PM and 56 patients (including 8 who developed delirium) from 4 PM until 9 PM. We developed binary classification models using support vector machine (SVM) and random forest (RF). We tested three patterns of input variables: EEG indicators, heart rate variability (HRV) indicators, and both. Finally, the SVM model with both indicators was able to predict with a sensitivity of 0.63 and a specificity of 0.69.**

I. Introduction

Delirium is a severe neuropsychiatric syndrome characterized by a sudden decline in cognitive function and attention. Patients often experience altered states of wakefulness, ranging from reduced responsiveness to severe agitation. This condition is frequently accompanied by symptoms such as delusions, hallucinations, and mood changes [1].

The onset of delirium results in numerous adverse effects. In the United States, it is estimated that the extension of hospital stays due to delirium costs an additional 38 billion dollars annually [2]. Furthermore, an increase in mortality rates has been reported, with delirium patients having a mortality rate 2.9 times higher than non-delirium patients [3]. Long-term cognitive decline is also a concern, as the onset of delirium has been reported to be a cause of dementia [4]. Additionally, delirium poses significant psychological distress to patients as well as to their caregivers [5]. Due to the numerous problems caused by delirium, the importance of preventive medicine is emphasized. Predicting the onset of delirium in advance can lead to efficient prevention.

Most predictions of delirium use electronic health records and clinical evaluations by physician's subjectivity [6]. One such model achieved a sensitivity of 72.9% and a specificity of 77.5% using 115 preoperative features [7]. However, clinical evaluations can sometimes differ between trainee doctors and other medical professionals. Hence, highlighting the need for predictions based on objective indicators. Biological markers provide objective measures, and recent studies have successfully predicted delirium with 86% accuracy using preoperative electroencephalogram (EEG) [8]. Additionally, changes in preoperative heart rate variability (HRV) are likely associated with delirium onset [9]. Therefore, this study aims to develop a model for predicting the onset of delirium using these biological indicators.

This work was supported by JSPS KAKENHI Grant Numbers JP20K07915, JP21H03855, and JP23K07006.

II. Methods

This study has been approved by the research ethics committee of Tokyo Medical and Dental University.

In the following text, we defined the period from 9 AM until 1 PM as the AM and the period from 4 PM until 9 PM as the PM.

A. Participants

A total of 80 patients aged 16 and older who underwent oral surgery at the Department of Oral and Maxillofacial Surgery, Tokyo Medical and Dental University Hospital, March 23, 2022, and March 25, 2024, were participated in this study. Electrocardiogram (ECG) and EEG were measured twice a day, once in the AM and once in the PM, each for more than 3 minutes, from the day of admission until the day before surgery. We excluded 11 subjects in the AM and 12 subjects in the PM who did not have a Holter ECG attached during EEG measurement. Then, we excluded 7 subjects in the AM and 9 subjects in the PM whose ECG waveforms at the target time were not recorded for some reason. Additionally, Data from 3 subjects in the AM and 3 subjects in the PM, whose ECG waveforms could not be confirmed due to noise, were excluded. As a result, data from 59 subjects in the AM and 56 subjects in the PM were included in the analysis. Tables I and II show the demographic data of the subjects for the AM and PM, respectively.

TABLE I
DEMOGRAPHIC DATA OF STUDY SUBJECTS
(FROM 9 AM UNTIL 1 PM)

Item	Delirium Positive	Delirium Negative
Number of subjects	8	51
Gender (male/female)	5/3	30/21
Age(years±SD)	64.8±12.7	79.3±10.6

TABLE II
DEMOGRAPHIC DATA OF STUDY SUBJECTS
(FROM 4 PM UNTIL 9 PM)

Item	Delirium Positive	Delirium Negative
Number of subjects	8	48
Gender (male/female)	6/2	27/21
Age(years±SD)	79.8±10.1	63.3±12.9

B. Data Collection

During the measurements, participants were asked to lie down on a bed and close their eyes. ECG was measured at 250 Hz using a Holter monitor Simplo (Technology Inc.). EEG was measured at 128 Hz using a single-channel electroencephalograph ZA-X (Proassist Ltd.). The evaluation of delirium was conducted by physicians over several days postoperatively.

C. EEG Data Preprocessing

We analyzed 3-minute EEG segments with minimal noise that had been preselected by a physician. For each patient, there is one segment in the AM and one segment in the PM. Each segment was divided into 90 non-overlapping epochs, each lasting 2 seconds.

For each epoch, we calculated the kurtosis, skewness, and the difference between the maximum and minimum voltages. For each patient's 3-minutes data segment, we applied the three sigma method to these three metrics to exclude epochs considered to contain noise. Subsequently, for each remaining epoch, we calculated the power ratio for the Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma1 (30–59 Hz) frequency bands, as well as the ratios between these variables, and then computed the average of them over the 3-minutes period.

D. ECG Data Preprocessing

In this study, we analyzed 3-minute segments of ECG taken at the same time as the clipped EEG. We preprocessed the selected data by applying a bandpass filter of 0.5–150 Hz. R-waves were detected by the maximal overlap discrete wavelet transform and RR intervals (RRIs) were calculated from them. During this process, visual inspection was conducted. It is known that accurate HRV indicators cannot be calculated from RRIs containing outliers such as arrhythmias [10]. Hence, we employed the Hampel filter to detect these outliers and replaced them with linearly interpolated values. Subsequently, we calculated the time-domain indicators shown in Table III, the frequency-domain indicators shown in Table IV, and the Poincare plot indicators shown in Table V.

TABLE III
HRV TIME DOMAIN INDICATORS

Feature Name	Definition
meanNN	mean RRI
meanHR	mean heart rate
SDNN	standard deviation of RRI
RMSSD	root mean square of successive RRI differences
SDSD	standard deviation of successive RRI differences
NN50	number of RRI pairs differing by over 50 ms
NN20	number of RRI pairs differing by over 20 ms
pNN50	percentage of RRI pairs differing by over 50 ms
pNN20	percentage of RRI pairs differing by over 20 ms

TABLE IV
HRV FREQUENCY DOMAIN INDICATORS

Feature Name	Definition
HF	Power in the 0.15–0.4 Hz range
LF	Power in the 0.04–0.15 Hz range
LF/HF	ratio of LF to HF

TABLE V
HRV POINCARE PLOT INDICATORS

Feature Name	Definition
SD1	standard deviation of points vertical to identity line
SD2	standard deviation of points horizontal to identity line
SD1*SD2	product of SD1 and SD2
SD2/SD1	ratio of SD2 to SD1

III. MODEL

To distinguish between delirium and non-delirium, we constructed a binary classification model. We designed three types of input variable patterns: EEG indicators, HRV indicators, and both. Considering the circadian rhythm, separate models were built for AM and PM. We used 59 patients (including 8 with delirium) in the AM and 56 patients (including 8 with delirium) in the PM to construct predictive models.

The methods used were support vector machines (SVMs) with a radial basis function (RBF) kernel and random forests (RFs), which were chosen based on their effectiveness for small datasets with approximately 100 samples in ECG and EEG analysis [11] [12]. For hyperparameters, we used the L2 regularization term and RBF kernel parameters for SVMs, and the maximum number of decision branches, the minimum number of samples per leaf node, and the number of input variables randomly selected for each split for RFs. To test classification performance, we employed leave-one-patient-out cross-validation.

After standardizing the training data by setting the mean of each input variable to 0 and the variance to 1, we increased the number of samples for delirium patients to match the number of non-delirium samples using an oversampling method, SMOTE [13]. Oversampling is used to adress significant imbalances between the numbers of samples of minority and majority classes. We performed 10-fold cross-validation on the training data and used Bayesian optimization to determine the hyperparameters that maximize area under the precision-recall curve (PR-AUC). We then made predictions on the

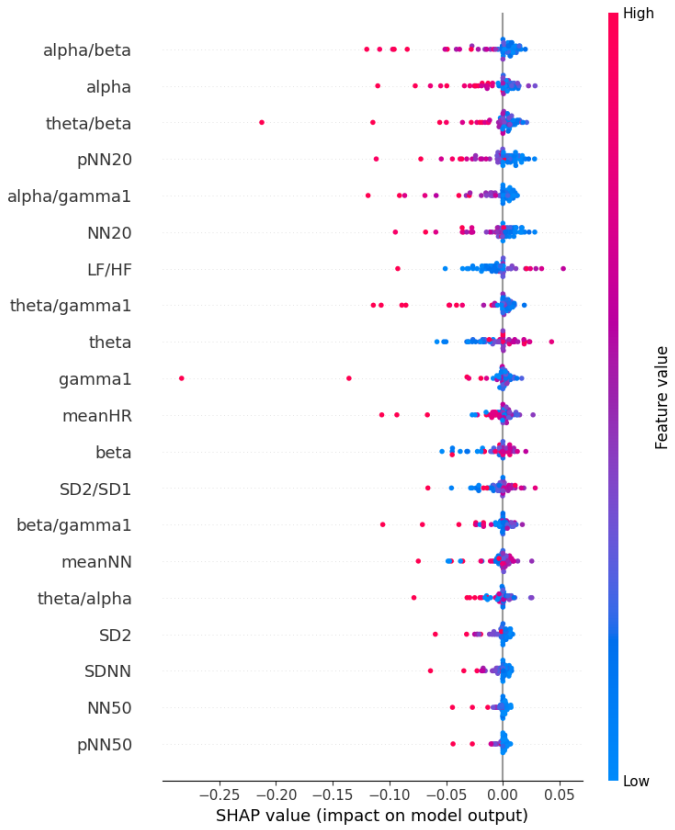


Fig. 1. Shap Value on SVM Model (from 9 AM until 1 PM)

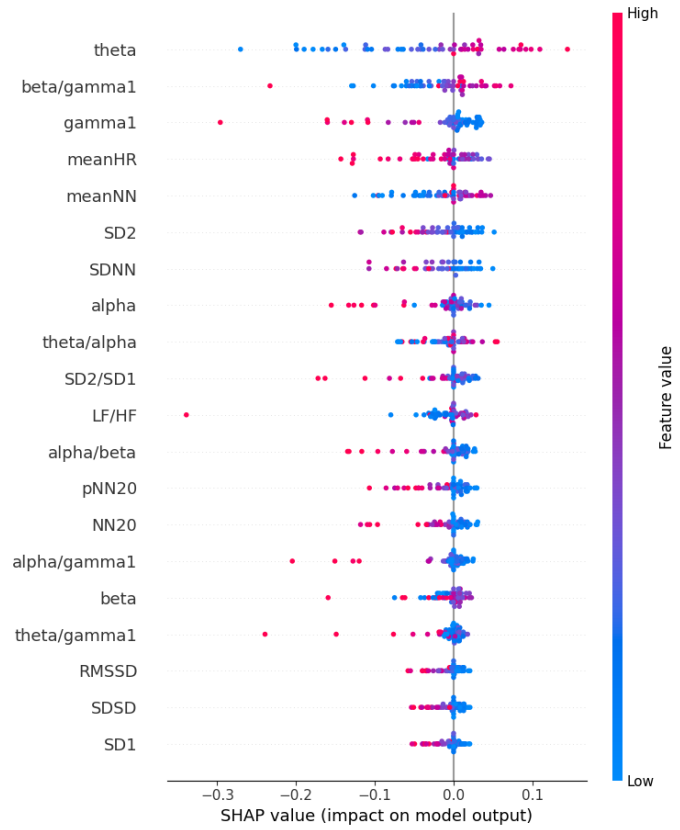


Fig. 2. Shap Value on SVM Model (from 4 PM until 9 PM)

remaining one sample and validated the performance. Finally, we calculated SHAP values [14] for each model to determine the importance of the variables.

IV. Results

Tables VI and VII show the results in the AM and the PM, respectively, indicating the average accuracy, sensitivity, specificity, and PR-AUC on the training data. Tables VIII and IX summarize the prediction results for the test data in the AM and PM, respectively. Changing the input variable pattern did not lead to significant changes in performance.

Figs. 1 and 2 plot the SHAP values for all test data predicted by the SVM model with the highest sensitivity, using both EEG and HRV indicators in the AM and PM, respectively. The SHAP values are sorted in descending order of their mean absolute values. The horizontal axis represents the SHAP values, where larger values indicate a positive contribution to the model and smaller values indicate a negative contribution. The relative magnitude of SHAP values for different variables indicates their comparative importance in influencing the model's output. The redder plot shows the value of the respective feature is higher. The bluer plot shows the value of the respective feature is lower.

V. Discussion

Since sensitivity is important for predicting delirium, we focus on sensitivity in the test data. We conclude that the SVM model using both EEG and HRV indicators or only EEG indicators for AM, and the Random Forest model using only HRV indicators for PM, performed best in our data set. However, compared to the previous study, which reported an accuracy of 86% [8], the accuracy in our study is lower. This discrepancy may be attributed to the fact that 30% of the patients in the previous study were delirium patients, while only 12.5% of the patients in our study were delirium patients, resulting in greater class imbalance. Additionally, while the previous study used three EEG channels, our study only used one EEG channel, potentially failing to capture the EEG changes associated with the onset of delirium.

Next, We discuss the top-ranked variables in importance. First, we will discuss the EEG indicators. The low power of alpha waves contributes to delirium, which is consistent with the study showing that preoperative EEG alpha waves are inversely correlated with the severity of postoperative delirium [15]. The increase in theta wave power contributes to the onset of delirium, consistent with research indicating that delirium patients experience delayed brain waves and increased theta waves [16], suggesting that delayed brain waves may occur preoperatively. The increase in gamma wave activity

TABLE VI
TRAIN AVERAGE RESULTS (FROM 9 AM UNTIL 1 PM)

Method	Input Variable	Accuracy	Sensitivity	Specificity	PR-AUC
SVM	HRV	0.84	0.83	0.86	0.92
	EEG	0.82	0.92	0.72	0.94
	HRV and EEG	0.81	0.90	0.71	0.90
RF	HRV	0.85	0.80	0.90	0.96
	EEG	0.89	0.86	0.89	0.96
	HRV and EEG	0.84	0.84	0.85	0.93

TABLE VII
TRAIN AVERAGE RESULTS (FROM 4 PM UNTIL 9 PM)

Method	Input Variable	Accuracy	Sensitivity	Specificity	PR-AUC
SVM	HRV	0.88	0.91	0.85	0.96
	EEG	0.85	0.86	0.84	0.88
	HRV and EEG	0.86	0.99	0.73	0.96
RF	HRV	0.86	0.80	0.90	0.96
	EEG	0.83	0.92	0.74	0.93
	HRV and EEG	0.89	0.92	0.86	0.98

TABLE VIII
TEST RESULTS (FROM 9 AM UNTIL 1 PM).
FRACTIONS IN PARENTHESES INDICATE HOW MANY PATIENTS ANSWERED CORRECTLY

Method	Input Variable	Accuracy	Sensitivity	Specificity	PR-AUC
SVM	HRV	0.81	0.50 (4/8)	0.86 (44/51)	0.20
	EEG	0.66	0.63(5/8)	0.67(34/51)	0.16
	HRV and EEG	0.68	0.63(5/8)	0.69(35/51)	0.20
RF	HRV	0.73	0.38(3/8)	0.78(40/51)	0.18
	EEG	0.73	0.50(4/8)	0.76(39/51)	0.21
	HRV and EEG	0.68	0.38(3/8)	0.69(36/51)	0.12

TABLE IX
TEST RESULTS (FROM 4 PM UNTIL 9 PM).
FRACTIONS IN PARENTHESES INDICATE HOW MANY PATIENTS ANSWERED CORRECTLY

Method	Input Variable	Accuracy	Sensitivity	Specificity	PR-AUC
SVM	HRV	0.75	0.38(3/8)	0.81(39/48)	0.14
	EEG	0.75	0.38(3/8)	0.81(39/48)	0.23
	HRV and EEG	0.64	0.63(5/8)	0.65(31/48)	0.18
RF	HRV	0.68	0.63(5/8)	0.69(33/48)	0.24
	EEG	0.61	0.38(3/8)	0.65(31/48)	0.25
	HRV and EEG	0.64	0.50(4/8)	0.69(33/48)	0.18

indicates a heightened state of arousal [17], implying that preoperative arousal levels may be lower in delirium patients compared to non-delirium patients.

Finally, we will discuss the indicators related to HRV. The decrease in SD2 suggesting the onset of delirium is consistent with the study indicating that SD2 decreases under stress conditions [18]. It is noted that higher SDNN indicates greater resistance to stress [19], and lower SDNN is associated with a state of stress, aligning with the prediction that a decrease in SDNN contributes to the onset of delirium. Additionally, when SDNN decreases, the variability of NN intervals diminishes, leading to reductions in NN20 and pNN20. Therefore, the decrease in NN20 and pNN20 due to stress load is consistent with their contribution to the onset of delirium. A higher LF/HF ratio contributes to delirium, indicating a state of mental stress [20], consistent with studies suggesting that stress is partly responsible for the onset of delirium.

VI. CONCLUSIONS

In this study, our model was able to predict with a sensitivity of 0.63, a specificity of 0.69, and an accuracy of 0.68. However, due to the limited number of our data set, we need to conduct further studies. Although our models did not achieve the accuracy reported in other studies, the result allowed us to investigate the relationship between physiological markers and the delirium. In the future work, we plan to collect additional subject data and compare the results with those obtained using various machine learning techniques.

REFERENCES

[1] J. E. Wilson, M. F. Mart, C. Cunningham, *et al.*, "Delirium," *Nature Reviews Disease Primers*, vol. 6, p. 90, Nov. 2020. DOI: 10.1038/s41572-020-00223-4.

[2] D. L. Leslie, E. R. Marcantonio, Y. Zhang, L. Leo-Summers, and S. K. Inouye, "One-year health care costs associated with delirium in the elderly population," *Archives of Internal Medicine*, vol. 168, pp. 27–32, Jan. 2008. DOI: 10.1001/archinternmed.2007.4.

[3] D. K. Kiely, E. R. Marcantonio, S. K. Inouye, *et al.*, "Persistent delirium predicts greater mortality," *Journal of the American Geriatrics Society*, vol. 57, pp. 55–61, Jan. 2009. DOI: 10.1111/j.1532-5415.2008.02092.x.

[4] T. E. Goldberg, C. Chen, Y. Wang, *et al.*, "Association of delirium with long-term cognitive decline: A meta-analysis," *JAMA Neurology*, vol. 77, pp. 1373–1381, Nov. 2020. DOI: 10.1001/jamaneurol.2020.2273.

[5] S. T. Williams, J. K. Dhesi, and J. S. L. Partridge, "Distress in delirium: Causes, assessment and management," *European Geriatric Medicine*, vol. 11, pp. 63–70, Feb. 2020. DOI: 10.1007/s41999-019-00276-z. eprint: 2019Dec9.

[6] T. Strating, L. Shafiee Hanjani, I. Tornvall, R. Hubbard, and I. A. Scott, "Navigating the machine learning pipeline: A scoping review of inpatient delirium prediction models," *BMJ Health Care Informatics*, vol. 30, e100767, Jul. 2023. DOI: 10.1136/bmjhci-2023-100767.

[7] A. Bishara, C. Chiu, E. L. Whitlock, *et al.*, "Postoperative delirium prediction using machine learning models and preoperative electronic health record data," *BMC Anesthesiology*, vol. 22, p. 8, Jan. 2022. DOI: 10.1186/s12871-021-01543-y.

[8] M. Hata, Y. Miyazaki, C. Nagata, *et al.*, "Predicting postoperative delirium after cardiovascular surgeries from preoperative portable electroencephalography oscillations," *Frontiers in Psychiatry*, vol. 14, p. 1 287 607, Nov. 2023. DOI: 10.3389/fpsyt.2023.1287607.

[9] G. Ernst, L. O. Watne, M. Rostrup, and B. E. Neerland, "Delirium in patients with hip fracture is associated with increased heart rate variability," *Aging Clinical and Experimental Research*, vol. 32, pp. 2311–2318, Nov. 2020. DOI: 10.1007/s40520-019-01447-5. eprint: 2020Jan8.

[10] "Heart rate variability. standards of measurement, physiological interpretation, and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology," en, *Eur. Heart J.*, vol. 17, pp. 354–381, Mar. 1996.

[11] A. M. Tăuan, E. P. Casula, M. C. Pellicciari, *et al.*, "TMS-EEG perturbation biomarkers for Alzheimer's disease patients classification," *Scientific Reports*, vol. 13, p. 7667, 2023. DOI: 10.1038/s41598-022-22978-4.

[12] M. Martinez-Alanis, E. Bojorges-Valdez, N. Wessel, and C. Lerma, "Prediction of Sudden Cardiac Death Risk with a Support Vector Machine Based on Heart Rate Variability and Heartprint Indices," *Sensors*, vol. 20, p. 5483, 2020. DOI: 10.3390/s20195483.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, ISSN: 1076-9757. DOI: 10.1613/jair.953. [Online]. Available: http://dx.doi.org/10.1613/jair.953.

[14] S. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, 2017. arXiv: 1705.07874 [cs.AI]. [Online]. Available: https://arxiv.org/abs/1705.07874.

[15] L. Acker, M. K. Wong, M. C. Wright, *et al.*, "Preoperative electroencephalographic alpha-power changes with eyes opening are associated with postoperative attention impairment and inattention-related delirium severity," *British Journal of Anaesthesia*, vol. 132, pp. 154–163, Jan. 2024. DOI: 10.1016/j.bja.2023.10.037. eprint: 2023Dec12.

[16] C. S. Guay, M. Kafashan, E. R. Huels, *et al.*, "Postoperative delirium severity and recovery correlate with electroencephalogram spectral features," *Anesthesia Analgesia*, vol. 136, pp. 140–151, Jan. 2023. DOI: 10.1213/ANE.0000000000006075. eprint: 2022May13.

[17] C. Summerfield, A. I. Jack, and A. P. Burgess, "Induced gamma activity is associated with conscious awareness of pattern masked nouns," *International Journal of Psychophysiology*, vol. 44, pp. 93–100, May 2002. DOI: 10.1016/s0167-8760(02)00003-x.

[18] S. Rahman, M. Habel, and R. J. Contrada, "Poincaré plot indices as measures of sympathetic cardiac regulation: Responses to psychological stress and associations with pre-ejection period," *International Journal of Psychophysiology*, vol. 133, pp. 79–90, Nov. 2018. DOI: 10.1016/j.ijpsycho.2018.08.005. eprint: 2018Aug11.

[19] H. G. Kim, E. J. Cheon, D. S. Bai, Y. H. Lee, and B. H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Investigation*, vol. 15, pp. 235–245, Mar. 2018. DOI: 10.30773/pi.2017.08.17. eprint: 2018Feb28.

[20] W. von Rosenberg, T. Chanwimalueang, T. Adjei, U. Jaffer, V. Goverdovsky, and D. P. Mandic, "Resolving ambiguities in the lf/hf ratio: Lf-hf scatter plots for the categorization of mental and physical stress from hrv," *Frontiers in Physiology*, vol. 8, p. 360, Jun. 2017. DOI: 10.3389/fphys.2017.00360.

A High-Isolation Sub-6 GHz In-Band Full-Duplex Communication System

Chengzhe Shi*, Wensheng Pan*, Wanzhi Ma*, Ying Liu*, Qiang Xu*, Zhiya Zhang† and Shihai Shao*
*National Key Laboratory of Wireless Communications,
University of Electronic Science and Technology of China, Chengdu, P. R. China
Email: scz@std.uestc.edu.cn, panwens@uestc.edu.cn, mawanzhi@uestc.edu.cn,
yliu85@uestc.edu.cn, xuqiang06@uestc.edu.cn, ssh@uestc.edu.cn
† Xidian university, Xi'an 710071, P. R. China
E-mail: zyzhang@xidian.edu.cn

***Abstract*—In 5G networks, the sub-6 GHz band offers broader user coverage and more reliable connections compared to its millimeter-wave counterpart. The increasing spectrum congestion in sub-6 GHz has driven research into the in-band full-duplex (IBFD) communication. This work explores the design principles of a high-isolation IBFD transceiver prototype for future 5G-Advanced networks. The designed IBFD transceiver employs a high-isolation spiral antenna design, which offers high spatial isolation while avoiding the complicated radio frequency self-interference cancellation. Prototype test results demonstrate that the provided spatial isolation is over 70 dB, and the overall self-interference cancellation performance exceeds 120 dB combined with nonlinear digital self-interference cancellation.**

I. Introduction

In 5G networks, the sub-6 GHz frequency band (ranging from 600 MHz to 6 GHz) is highly regarded due to its balanced attributes of coverage and capacity. Compared to the millimeter-wave (mmWave) band, the sub-6 GHz band offers a broader coverage range and superior penetration through buildings and vegetation, thus ensuring reliable connectivity in urban, suburban, and rural environments [1], [2].

Currently, the sub-6 GHz band is not only favored by mobile network operators but also widely used by other wireless services such as Wi-Fi, satellite communications, and military applications. This intense competition for spectrum resources may lead to congestion and limit the potential of 5G networks to fully exploit this band [3]. To facilitate flexible use of radio resources, 3GPP Release 18, the initial release of 5G-Advanced, included a study item on the feasibility of simultaneous uplink and downlink in the same frequency band, also known as in-band full-duplex (IBFD) [4]. Potential IBFD features are anticipated to be incorporated into the specifications from Release 19 onward. Recently, the 3rd Generation Partnership Project (3GPP) completed the study of subband full-duplex, which may pave the way for commercial deployments of IBFD in 5G-Advanced and 6G.

IBFD technology allows simultaneous transmission and reception on the same frequency channel, theoretically doubling spectral efficiency compared to traditional half-duplex mode, making it a crucial technology to meet the high data rate demands of 5G and 5G-Advanced [5], [6]. However, achieving full duplex communication presents significant technical challenges, primarily related to self-interference (SI), where the transmitter interferes with its own receiver. Overcoming these challenges requires advances in self-interference cancellation (SIC) techniques, typically involving three-stage cancellation in the spatial, RF, and digital domains [6]. The successful integration of these cancellation techniques is crucial for the development of high-performance IBFD transceiver systems.

Recently, a mmWave full-duplex prototype testing platform, developed in accordance with 5G Integrated Access and Backhaul (IAB) specifications, demonstrated the feasibility of a mmWave SIC transceiver [7]. Experimental results confirmed that full-duplex throughput in IAB scenarios could potentially double compared to half-duplex systems. By leveraging the high path loss of the mmWave frequency band and a passive antenna design with multiple isolators, spatial isolation exceeding 80 dB was achieved.

As a complementary frequency band to mmWave in 5G networks, the sub-6 GHz band lacks the advantage of high path loss and available spatial degrees of freedom. Consequently, designing high-isolation IBFD transceiver prototypes in the sub-6 GHz band is essential for 5G deployment. In this work, we investigate the design principles of SIC for IBFD systems for the sub-6 GHz band, and a full-duplex prototype with high-isolation spiral antennas is designed, which provides high spatial isolation without the complicated radio frequency (RF) SIC designs. Prototype testing results demonstrate the spatial SI isolation of over 70 dB, and the overall SIC capability exceeding 120 dB combined with nonlinear digital SIC (DSIC).

II. High-Isolation IBFD Transceiver Design Principles

The coupling between transmit and receive antennas in IBFD radios results in significant SI at the receiver, overshadowing the desired signal from remote uplink users. Therefore, a primary challenge in implementing IBFD radios is to cancel the SI to the level of background noise. For instance, 5G base stations typically operate with an effective isotropic radiated power (EIRP) of 40 to 50 dBm. At such a transmit power level, the overall required SIC capability to cancel SI to noise

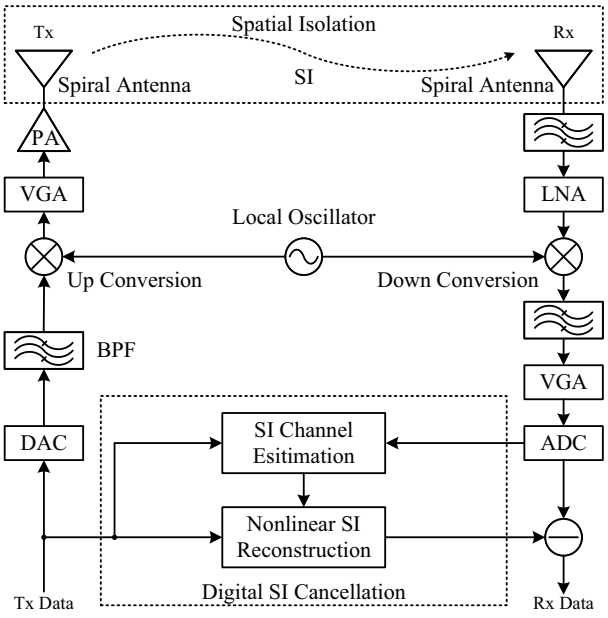
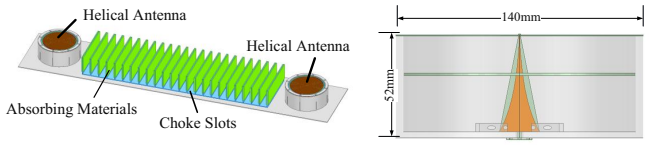
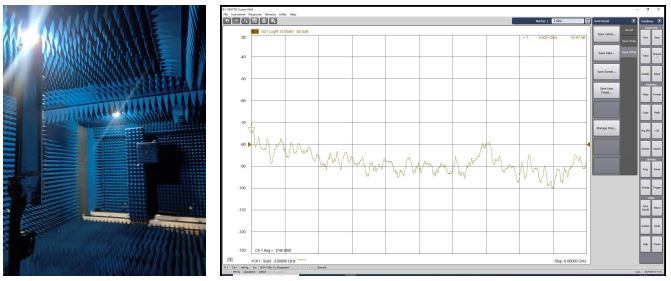


Fig. 1. Schematic diagram of the high-isolation IBFD SIC transceiver.



(a) Antenna structure. (b) Cross-section of the spiral antenna.

Fig. 2. High-isolation wideband spiral antennas. (a) Antenna structure. (b) Cross-section of the spiral antenna.



(a) Testing setup. (b) Testing results.

Fig. 3. The anechoic chamber testing setup and results for the high-isolation spiral antenna. (a) Testing setup. (b) Testing results.

floor could be as high as 129 dB given -79 dBm noise floor at 300 MHz bandwidth with 10 dB noise figure.

Given such strong SI, it is necessary to implement cancellation in both the spatial domain and the receive RF front-end to reduce SI; otherwise, it would severely saturate the receiver. Typically, the saturation power of the receive RF front-end, determined by the low noise amplifier (LNA), is in the range of -20 to -10 dBm, which implies that the spatial domain and RF front-end must be able to mitigate more than 70 dB of SI.

Moreover, transmit error vector magnitude (EVM) distortion resulting from impairments in the transmit chain also degrades the performance of IBFD radios. For instance, in 5G networks, typical transmit powers are 20 dBm for user equipment and 46 dBm for base stations. Hence, a 17.5% transmit EVM distortion in the base station exceeds the transmit power of user equipment by an order of magnitude[1] [8]. Considering the path loss, the signal received at the base station from user equipment is much weaker than the EVM distortion introduced by the base station transmitter, necessitating the DSIC.

Taking into account the SIC requirements outlined above, we design a SIC transceiver for IBFD, as shown in Fig 1. Our IBFD prototype design employs joint spatial and digital SIC schemes. The complex RF SIC has been omitted due to the high isolation provided by the circular polarized spiral antenna design, which ensures that the receive RF front-end does not saturate due to strong SI, thus preserving sufficient dynamic range for the analog-to-digital converter (ADC) to capture the desired signal. Furthermore, through nonlinear DSIC, we anticipate reducing the residual SI after ADC sampling, dominated by distortions in the transmit chain, to the noise floor level.

[1]In 5G, the maximum allowed EVM levels for base station transmitters are 17.5% for QPSK, corresponding to -15 dB power levels for the distortion of the transmitted signal.

A. *High Isolation spiral Antenna*

In IBFD systems, the SI arising from the transmit antenna coupling to the receive antenna poses a significant risk of saturating the receive RF front-end, particularly in the sub-6 GHz band where spatial attenuation is relatively low. Specifically, these mutual coupling effects can be categorized into surface and space wave coupling. The latter can be subdivided into direct radiation coupling (line-of-sight component) and space reflection coupling (reflection component) [6]. In high-density array systems, the high spatial degrees of freedom can be leveraged to employ active beamforming SIC techniques to mitigate near-field coupling between transmit and receive arrays [9]–[11]. However, in sub-6 GHz, the limited spatial degrees of freedom make it crucial to enhance the spatial isolation through specialized antenna design.

In our IBFD prototype design, we adopt a solution involving high-isolation wideband spiral antennas, as depicted in Fig 2. The high-isolation wideband spiral antennas comprise two co-polarized spiral antenna elements, choke slots, and absorbing materials.

A spiral antenna is constructed from a helix-shaped metallic conductor, typically made from wire or tubing. It is commonly fed with a coaxial cable, where the inner conductor connects to one end of the helix and the outer conductor is grounded to a metal plate. This ground plate diminishes induced currents on the coaxial cable's surface, enhancing the antenna's radiation characteristics and minimizing backward radiation.

As shown in Fig 2 (a), a choke slot structure is integrated between the two antenna elements, providing several advantages. Firstly, unlike a smooth metal ground plate, the metal plate with choke slots directly impedes surface wave propagation, effectively reducing coupling between the transmit and receive antennas. Secondly, since the height of the choke slots is higher

than that of the spiral antennas, they can significantly suppress the direct radiation coupling of electromagnetic waves. Moreover, absorbing materials are applied onto the dielectric floor and the surface of the choke slots, which absorb space-reflected waves, further reducing electromagnetic interference between antenna elements.

Fig 3 (a) illustrates the anechoic chamber testing setup for the high-isolation spiral antenna. The S21 parameter, which characterizes the isolation, is also depicted in Fig 3 (b). It can be seen that within the 3∼6 GHz testing frequency range, the antenna isolation is generally above 80 dB, with a slight degradation between 3∼3.5 GHz, indicating overall good isolation performance.

B. Withouting RF SIC

Typically, the RF SIC is an effective method to prevent saturation or blocking of the receive RF front-end caused by SI, although it necessitates additional complex circuitry for implementation. Benefiting from the design of the high-isolation spiral antenna, spatially coupled SI is effectively suppressed, greatly avoiding the risk of RF front-end saturation. This also ensures that ADCs have sufficient dynamic range to sample both SI and the desired signal simultaneously. As a result, our IBFD prototype design obviates the complex RF SIC.

C. Nonlinear Digital SIC

In practical engineering, constrained by hardware complexity and cost, merely reducing SI through spatial isolation and RF cancellation is insufficient. The residual SI after ADC sampling still contaminates the desired signal, resulting in decreased receive sensitivity. Therefore, it is necessary to further cancell the residual SI in the digital domain to reach the demodulation threshold of the desired signal.

After upconverting the baseband signal to the RF domain, inevitable nonlinear interference components arise, induced by factors such as phase noise, I/Q imbalance, and power amplifier (PA) nonlinearities. As transmit power increases, the impact of PA-dominated nonlinear interference becomes more pronounced. Let $x\left(n\right)$ denote the digital baseband signal. Considering the PA's nonlinear distortion characteristics and memory effects, the discrete form of the transmitted signal can be expressed according to the widely used memory polynomial (MP) model as

$$\begin{aligned} x_{\mathrm{T}}\left(n\right) &= f_{\mathrm{MP}}\left[x\left(n\right)\right] + z_{\mathrm{T}}\left(n\right) \\ &= \sum_{p=0}^{P-1}\sum_{m=0}^{M-1}\alpha_{mp}\phi_{m,p}\left[x\left(n\right)\right] + z_{\mathrm{T}}\left(n\right), \end{aligned} \tag{1}$$

where $f_{\mathrm{MP}}\left[\cdot\right]$ represents the nonlinear MP function, P denotes the maximum order of nonlinearity, M is the maximum memory depth, and $\phi_{m,p}\left[\cdot\right]$ is the MP basis function with respect to the nonlinear order p and memory depth m, which is given by

$$\phi_{m,p}\left[x\left(n\right)\right] = x\left(n-m\right)\left|x\left(n-m\right)\right|^{2p}, \tag{2}$$

and α_{mp} is the corresponding coefficient.

After propagating through a multipath channel, the digital model of the SI received by the local antenna is

$$\begin{aligned} y\left(n\right) &= h_{\mathrm{SI}}\left(n\right) \otimes f_{\mathrm{MP}}\left[x\left(n\right)\right] + z_{\mathrm{R}}\left(n\right) \\ &= \sum_{i=1}^{L} h_i f_{\mathrm{MP}}\left[x\left(n-D_i\right)\right] + z_{\mathrm{R}}\left(n\right) \\ &= \sum_{i=1}^{L} h_i \sum_{p=0}^{P-1}\sum_{m=D_{I,i}}^{M+D_{I,i}-1}\alpha_{mp}\phi_{m+D_{F,i},p}\left[x\left(n\right)\right] + z_{\mathrm{R}}\left(n\right), \end{aligned} \tag{3}$$

where $z_{\mathrm{R}}\left(n\right)$ represents the receive noise, h_i is the complex gain of the i-th path, and $D_i = D_{I,i} + D_{F,i}$ is the normalized delay (with respects to the sampling time) of the i-th path, including an integer part $D_{I,i}$ and a fractional part $D_{F,i}$.

Since $\phi_{m+D_{F,i},p}\left[x\left(n\right)\right]$ is the signal $\phi_{m,p}\left[x\left(n\right)\right]$ delayed by a fractional delay $D_{F,i}$, it can be modeled with an ideal fractional delay filter as

$$\begin{aligned} \phi_{m+D_{F,i},p}\left[x\left(n\right)\right] &= \phi_{m,p}\left[x\left(n\right)\right] \otimes h_{D_{F,i}}\left(n\right) \\ &= \sum_{j=-K}^{+K} h_{D_{F,i}}\left(j\right)\phi_{m+j,p}\left[x\left(n\right)\right], \end{aligned} \tag{4}$$

where $h_{D_{F,i}}\left(n\right) = \mathrm{sinc}\left(n - D_{F,i}\right)$ is the impulse response of the ideal fractional delay filter. Moreover, the filter coefficients are truncated from $-K$ to K by a rectangular window to meet practical engineering requirements.

After nonlinear distortion and multipath channel propagation, the received SI comprises linear and nonlinear components, alongside noise, which is omitted in the subsequent derivation for convenience. The received SI can be expressed in polynomial form

$$\begin{aligned} y\left(n\right) &= \sum_{i=1}^{L}\sum_{j=-K}^{K}\sum_{p=0}^{P-1}\sum_{m=D_{I,i}}^{M+D_{I,i}-1}\beta_{ijmp}\phi_{m+j,p}\left[x\left(n\right)\right] \\ &= \sum_{i=1}^{L}\sum_{k=m-K}^{m+K}\sum_{p=0}^{P-1}\sum_{m=D_{I,i}}^{M+D_{I,i}-1}\beta_{ikmp}\phi_{k,p}\left[x\left(n\right)\right], \end{aligned} \tag{5}$$

where $\beta_{ijmp} = h_i\alpha_{mp}h_{D_{F,i}}\left(j\right)$ and $\beta_{ikmp} = h_i\alpha_{mp}h_{D_{F,i}}\left(k-m\right)$. To further simplify (5), the following transformation is performed

$$\begin{aligned} y\left(n\right) &= \sum_{i=1}^{L}\sum_{k=-K+D_{I,i}}^{K+D_{I,i}+M-1}\sum_{p=0}^{P-1}\sum_{m=D_{I,i}}^{M+D_{I,i}-1}\beta_{ikmp}^{'}\phi_{k,p}\left[x\left(n\right)\right] \\ &= \sum_{i=1}^{L}\sum_{k=-K+D_{I,\min}}^{K+D_{I,\max}+M-1}\sum_{p=0}^{P-1}\sum_{m=D_{I,i}}^{M+D_{I,i}-1}\beta_{ikmp}^{''}\phi_{k,p}\left[x\left(n\right)\right] \\ &= \sum_{k=-K+D_{I,\min}}^{K+D_{I,\max}+M-1}\sum_{p=0}^{P-1} w_{k,p}\phi_{k,p}\left[x\left(n\right)\right], \end{aligned} \tag{6}$$

where

$$\beta_{ipmk}^{'} = h_i\alpha_{mp}h_{D_{F,i}}^{'}\left(k-m\right), \tag{7}$$

$$\beta_{ipmk}^{''} = h_i\alpha_{mp}h_{D_{F,i}}^{''}\left(k-m\right), \tag{8}$$

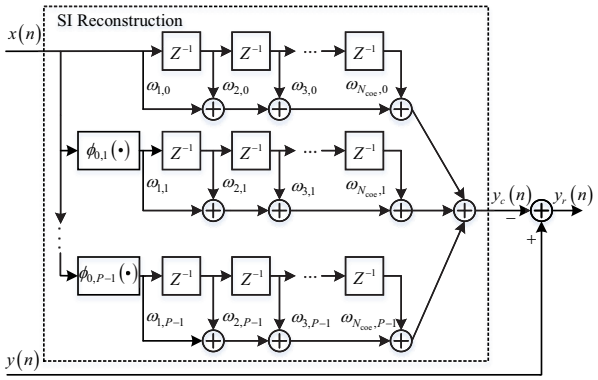


Fig. 4. Nonlinear DSIC adaptive filter structure.

$$h'_{D_{F,i}}(k-m)=\begin{cases}h_{D_{F,i}}(k-m), -K\le k-m\le K\\ 0, otherwise\end{cases}, \quad (9)$$

$$h''_{D_{F,i}}(k-m)=\begin{cases}h'_{D_{F,i}}(k-m),\\ -K+D_{I,i}\le k\le K+M+D_{I,i}-1,\\ 0, otherwise\end{cases} \quad (10)$$

$D_{I,\max}=\max_i D_{I,i}$ and $D_{I,\min}=\min_i D_{I,i}$ represent the maximum and minimum normalized arrival times of the SI after multipath channel propagation, respectively. $w_{k,p}$ is the simplified basis function coefficients, which is given by

$$w_{k,p}=\sum_{i=1}^{L}\sum_{m=D_{I,i}}^{M+D_{I,i}-1} h_i\alpha_{mp}h''_{D_{F,i}}(k-m) \quad (11)$$

Assuming there are N sampling points, (6) can be expressed in matrix form as

$$\mathbf{Y}=\mathbf{X}\mathbf{w}, \quad (12)$$

where

$$\mathbf{Y}=[y(1),y(2),\cdots,y(N)]^T, \quad (13)$$

$$\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2,\cdots,\mathbf{X}_N]^T, \quad (14)$$

$$\mathbf{w}=\{w_{-K+D_{I,\min},0},\cdots,w_{K+M+D_{I,\min}-1,0},\cdots,w_{K+M+D_{I,\min}-1,P-1}\}^T, \quad (15)$$

$$\mathbf{X}_n=\{\phi_{-K+D_{I,\min},0}[x(n)],\cdots,\phi_{K+M+D_{I,\min}-1,0}[x(n)],\cdots,\phi_{K+M+D_{I,\min}-1,P-1}[x(n)]\}^T. \quad (16)$$

The optimal coefficient estimation with the least squares criterion can be expressed as

$$\hat{\mathbf{w}}=\left(\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{Y}. \quad (17)$$

As shown in Fig. 4, the nonlinear DSIC structure can be realized by configuring P one-dimensional adaptive filters in parallel, where each dimension has N_{coe} coefficients. Specifically, when the transmit power is relatively low, and the nonlinearity can be ignored, i.e., $P=0$, this structure can be simplified to a one-dimensional horizontal filter.

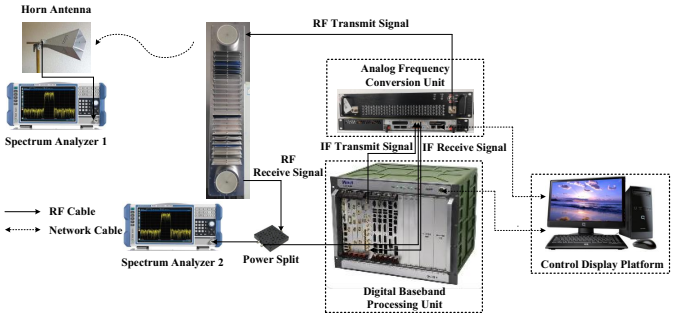


Fig. 5. Connection Diagram of the High-Isolation IBFD System.

Particularly, based on (6), the minimum number of taps for the adaptive filter structure can be derived as

$$N_{\min}=P(2K+M+\Delta D_I), \quad (18)$$

where $\Delta D_I=D_{I,\max}-D_{I,\min}$ denotes the maximum delay spread of the SI channel. It can be seen that the minimum number of taps depends on the nonlinearity order P, the memory depth M, the window length K of the fractional delay filter, and the maximum delay spread ΔD_I. However, it is independent of the number of multipaths L.

III. High-Isolation IBFD Protoyping

The high-isolation IBFD system comprises a digital baseband processing unit (responsible for generating and receiving baseband signals, and implementing nonlinear DSIC), an analog frequency conversion unit (used for up-conversion and down-conversion of analog signals), the high-isolation spiral antennas and control display platform, as shown in Fig. 5.

A. Analog Frequency Conversion Unit

The transmit chain employs a combination of up-conversion and down-conversion schemes. An intermediate frequency (IF) signal with a center frequency of 300 MHz and a bandwidth of 300 MHz is first up-converted to 13 GHz. Subsequently, this signal is then down-converted to 1∼6 GHz, resulting in a transmitted signal with center frequencies ranging from 1.15 GHz to 5.85 GHz.

Both receive chains also utilize a combination of up-conversion and down-conversion schemes. The received IF signal within the 1∼6 GHz is first up-converted to 13 GHz, and then down-converted to a received signal with a center frequency of 300 MHz.

Importantly, both the transmit and receive chains use the same local oscillator to reduce the frequency offset and phase noise of SI.

B. Digital Baseband Processing Unit

The digital baseband processing unit generates the transmitted IF signal, receives the IF signal, executes the nonlinear DSIC algorithm, and outputs the signal after DSIC. Moreover, it also controls various modules via the serial peripheral interface (SPI). This unit includes one FPGA, six ADC channels, three DAC channels, one input 100 MHz reference clock, two

Ethernet ports, and five SPI control interfaces. The ADC channels are used to sample wideband signals with a bandwidth of 300 MHz and an IF of 300 MHz. The FPGA performs the nonlinear DSIC algorithm. The DAC channels are utilized to output the IF analog signal with a bandwidth of 300 MHz and an IF of 300 MHz.

The bandwidth signal generated by the baseband unit is formed by a Chu sequence, which has the advantages of constant envelope, zero cyclic autocorrelation, and low peak-to-average power ratio. When synchronizing with the received signal, a distinct correlation peak can easily be found by performing a sliding correlation between the locally stored known transmitted signal and the received signal. The Chu sequence is expressed as

$$x\left(n\right)=\begin{cases}\exp\left\{\dfrac{j\pi\mu\left[n\left(n+1\right)+2qn\right]}{N}\right\}, N\ is\ odd\\ \exp\left\{\dfrac{j\pi\mu\left(n^2+2qn\right)}{N}\right\}, N\ is\ even\end{cases}, \quad (19)$$

where $\mu\in\{1,2,\cdots,N-1\}$, and q is any integer.

Specifically, the baseband unit uses Chu sequences with $\mu=1$ and $q=0$, assembled into a 2048-point OFDM signal for transmission. The bandwidth of the OFDM signal is 300 MHz with a sampling rate of 245 MSPS. The number of effective subcarriers is $\frac{200\times2048}{245}\approx1672$, which is the length of the Chu sequence. The length of the OFDM pilot sequence is 72, making the total length of one OFDM signal 2120. After 4x upsampling, the OFDM signal with a sampling rate of 980 MSPS is converted from serial to parallel, resulting in four 245 MSPS signals, which are then upconverted and output by the DAC.

IV. Prototype Test

A. Test Enviorment

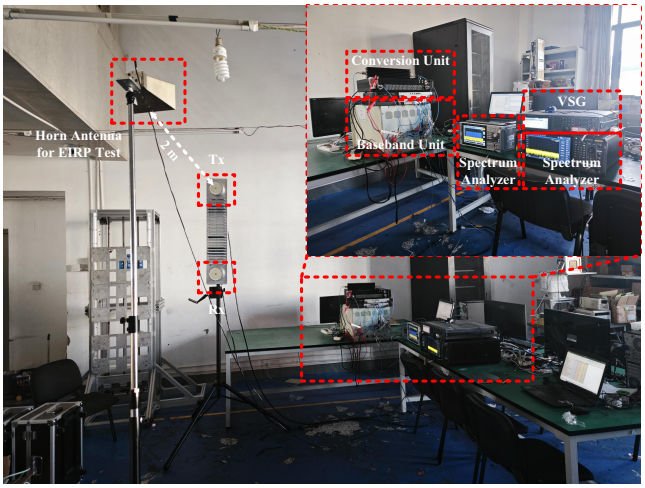


Fig. 6. High-Isolation IBFD Prototype Testing Scenario.

The testing process is conducted in a typical laboratory environment, as shown in Fig. 6. The baseband signal is transformed into a 300 MHz analog signal through the digital baseband processing unit and sent to the frequency conversion unit. This IF analog signal then undergoes a two-stage up-conversion process to achieve the desired RF frequency range of 1∼6 GHz. Subsequently, the signal is transmitted through the spiral antenna. The receive antenna captures the SI and outputs it after spatial isolation. The output signal is split into two paths by a power splitter. One path is sent to the spectrum analyzer to observe the interference power after spatial isolation, and the other path is fed into the DSIC unit, which can be enabled or disabled. After DSIC, the processed output is displayed on a control display interface.

Auxiliary equipment required for the experimental testing includes a vector signal generator (VSG) to calibrate the receive chain gain. A receive horn antenna is placed 2 meters away from the experimental system to receive the transmitted signal for EIRP testing. All auxiliary equipment is connected to a switch via Ethernet cables to facilitate centralized control.

B. Metric Definition and Test Method

The SIC test procedure for the high-isolation IBFD prototype is as follows:

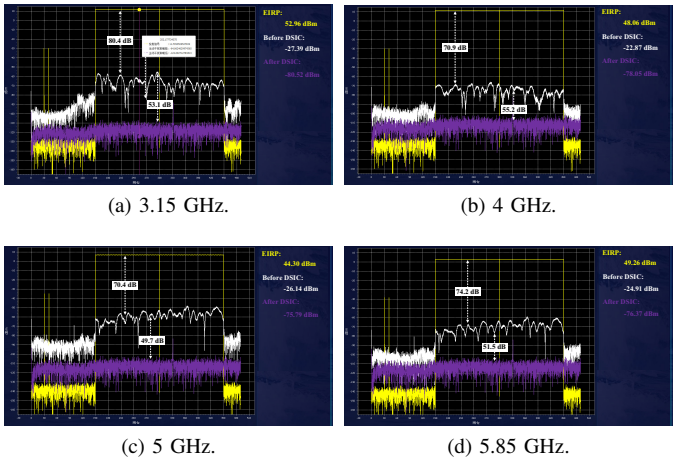
1) Connect the test equipment as depicted in Fig. 5, load the transmission waveform, and set the operating center frequency of the prototype f_{c} to 3.15 GHz.
2) Set the integration bandwidth of all spectrum analyzers to 300 MHz.
3) Measure the power of spectrum analyzer 1, denoted as P_{meas}, and calculate the EIRP:
$$\mathrm{EIRP}\left(\mathrm{dB}\right)=P_{\mathrm{meas}}+L_{\mathrm{path}}+L_{\mathrm{cable}}-G_{\mathrm{horn}}+3, \quad (20)$$
where L_{cable} is the cable loss from the horn antenna to the spectrum analyzer, G_{horn} is the gain of the horn antenna, and $L_{\mathrm{path}}=20\lg D\left(\mathrm{km}\right)+20\lg f_{\mathrm{c}}\left(\mathrm{MHz}\right)+32.5$ is the free-space loss between the transmit spiral antenna and the horn antenna with a distance of 2 meters. Moreover, 3 dB is attributed to the polarization loss, where the transmit antenna exhibits circular polarization and the receive antenna is linear polarization.
4) Measure the power of spectrum analyzer 2, denoted as P_{r1}, and the air interface equivalent residual SI power is $I_{\mathrm{air}}\left(\mathrm{dBm}\right)=P_{\mathrm{r1}}-G_{\mathrm{r1}}$, where G_{r1} represents the receiver front-end gain (mainly including the antenna, cables, and power dividers).
5) Calculate the spatial isolation $\mathrm{ISO}\left(\mathrm{dB}\right)=\mathrm{EIRP}-I_{\mathrm{air}}$.
6) Enable the DSIC unit and record the received power after DSIC as P_{r2}. Then, the air interface equivalent residual SI power after DSIC is $I_{\mathrm{res}}\left(\mathrm{dBm}\right)=P_{\mathrm{r2}}-G_{\mathrm{r2}}$, where G_{r2} represents the total receive gain (mainly including the frequency conversion unit and the digital baseband unit).
7) Calculate the DSIC capability $\mathrm{SIC}_{\mathrm{digital}}\left(\mathrm{dB}\right)=I_{\mathrm{air}}-I_{\mathrm{res}}$.
8) Calculate the overall SIC capability of the prototype $\mathrm{SIC}_{\mathrm{total}}\left(\mathrm{dB}\right)=\mathrm{EIRP}-I_{\mathrm{res}}=\mathrm{ISO}+\mathrm{SIC}_{\mathrm{digital}}$.

TABLE I
THE TEST RESULTS OF THE HIGH-ISOLATION FULL-DUPLEX PROTOTYPE.

Parameter	Vaule			
Center Frequency (GHz)	3.15	4	5	5.85
Bandwidth (MHz)	300	300	300	300
Receive power of spectrum analyzer 1 (dBm)	2.9	-1.0	-6.6	-2.0
Horn antenna gain (dB)	5.2	9.5	10.5	12.3
Cable loss (dB)	3.9	4.7	5.9	6.7
Path loss (dB)	48.4	50.5	52.4	53.8
EIRP (dBm)	53.2	47.7	44.2	49.1
SI Power before DSIC (dBm)	-24.0	-23.2	-26.1	-24.9
Spatial isolation (dB)	77.2	70.9	70.3	74.0
SI Power after DSIC (dBm)	-77.4	-77.8	-76.5	-78.3
DSIC capability (dB)	53.3	54.6	50.4	53.4
Prototype SIC capability(dB)	130.5	125.5	120.7	127.4

9) Change the operating frequencies to 4 GHz, 5 GHz, and 5.85 GHz, and repeat steps (3) through (8) to complete performance testing at different center frequencies.

C. Test Results



(a) 3.15 GHz. (b) 4 GHz.

(c) 5 GHz. (d) 5.85 GHz.

Fig. 7. The power spectral density of SIC at each stage for different center frequencies shown on the control display interface. (a) 3.15 GHz. (b) 4 GHz. (c) 5 GHz. (d) 5.85 GHz.

Fig.7 shows the power spectral density of SIC at each stage on the control display interface, where the yellow solid line represents the transmitted EIRP, and the white and purple solid lines represent the SI before and after DSIC, respectively. It can be observed that the SI after spatial isolation exhibits significant in-band frequency selectivity. This implies that RF SIC would require numerous analog taps and delay lines with wide tunability, thereby increasing the complexity of practical implementation.

However, due to the high-isolation design of the spiral antenna, our IBFD prototype obviates the requirement for complex RF SIC circuits. The measured spatial isolation of the designed spiral antenna structure is approximately 70∼80 dB, which is slightly lower than the anechoic chamber test results, possibly due to the rich indoor reflection components. Furthermore, by combining the nonlinear DSIC algorithm, a digital cancellation performance of over 50 dB has been achieved, reducing the residual SI power to near the noise floor level (approximately -79 dBm). The detailed test results of the prototype across different frequency bands from 3 to 6 GHz are provided in Table I. According to the test results, the designed prototype demonstrates an overall SIC capability exceeding 120 dB across all tested frequency bands.

V. CONCLUSION

In this work, we explore the design principles of SIC for IBFD systems in the sub-6 GHz band and develop a high-isolation IBFD prototype for future 5G-Advanced networks. The designed IBFD prototype employs a high-isolation spiral antenna design, which provides substantial spatial isolation without the need for complex RF SIC design. Prototype test results indicate that the provided spatial isolation is over 70 dB, and the overall SIC performance exceeds 120 dB combined with the nonlinear DSIC.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62071094, and Grant 61901396, and the National Key Laboratory of Wireless Communications Foundation under Grant 6142102222504.

REFERENCES

[1] Y. Huang and Y. Huang, "Challenges and opportunities of sub-6 ghz integrated sensing and communications for 5g-advanced and beyond," *Chinese Journal of Electronics*, vol. 33, no. 2, pp. 323–325, 2024.
[2] S.-M. Kim, J. Kim, H. Cha, M. S. Sim, J. Choi, S.-W. Ko, C.-B. Chae, and S.-L. Kim, "Opportunism in spectrum sharing for beyond 5g with sub-6 ghz: A concept and its application to duplexing," *IEEE Access*, vol. 8, pp. 148 877–148 891, 2020.
[3] B. Smida, R. Wichman, K. E. Kolodziej, H. A. Suraweera, T. Riihonen, and A. Sabharwal, "In-band full-duplex: The physical layer," *Proceedings of the IEEE*, 2024.
[4] 3GPP, "Study on evolution of nr duplex operation," *Technical Specification 38.858, 3rd Generation Partnership Project (3GPP)*, 2023.
[5] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-driven characterization of full-duplex wireless systems," *IEEE Transactions on Wireless Communications*, vol. 11, no. 12, pp. 4296–4307, 2012.
[6] K. E. Kolodziej, B. T. Perry, and J. S. Herd, "In-band full-duplex technology: Techniques and systems survey," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 7, pp. 3025–3041, 2019.
[7] B. Yu, C. Qian, J. Lee, S. Shao, Y. Shen, W. Pan, P. Lin, Z. Zhang, J. Jung, S. Choi *et al.*, "Realizing high power millimeter wave full duplex: Practical design, prototype and results," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 3803–3808.
[8] 3GPP, "Evolved universal terrestrial radio access (e-utra); base station (bs) radio transmission and reception," *Technical Specification 36.104, 3rd Generation Partnership Project (3GPP)*, vol. 10, 2023.
[9] I. P. Roberts and S. Vishwanath, "Beamforming cancellation design for millimeter-wave full-duplex," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.
[10] C. Shi, W. Pan, Y. Shen, and S. Shao, "Robust transmit beamforming for self-interference cancellation in star phased array systems," *IEEE Signal Processing Letters*, vol. 29, pp. 2622–2626, 2022.
[11] X. Huang, A. T. Le, and Y. J. Guo, "Transmit beamforming for communication and self-interference cancellation in full duplex mimo systems: A trade-off analysis," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3760–3769, 2021.

A Parameter-free model for long-term concrete creep prediction

Conghui Li, Chern Hong Lim and Xin Wang
Monash University Malaysia, Sunway, Malaysia
E-mail: {Conghui.Li, lim.chernhong, wang.xin}@monash.edu

***Abstract*—This paper presents a novel approach to predicting long-term concrete creep performance by leveraging early creep observations and eliminating the need for many sensing devices to collect material parameters. Our hybrid encoder-decoder model integrates a Convolutional Neural Network (CNN) for feature compression and noise reduction and a transformer for learning long-distance dependencies and global feature understanding. Additionally, the adaptive data completion method is designed to handle varying data intervals, ensuring robust predictions despite irregular sampling. Our results indicate that the proposed model significantly improves the consistency and reliability of creep predictions and comprehensively understands concrete material creep behavior over time.**

I. Introduction

Creep of concrete refers to the gradual long-term deformation of concrete under continuous load. This phenomenon is an inherent property of concrete as a composite material, and when subjected to constant stress, concrete will continue to deform over time [1]. Many factors can affect concrete creep, including the type and amount of cement, the water-cement ratio, the type of aggregate used, environmental conditions, and the age of the concrete when the load is applied [2], [3]. Concrete creep can cause many types of damage to structures. Prolonged creep can cause severe deformation of beams, plates, and other structural members, leading to serviceability problems and potential structural failures. Differential creep between the various parts of the structure can cause stress, which can lead to cracking, compromising the durability and appearance of the structure [4], [5]. In the prestressed concrete structure, creep will lead to the gradual loss of prestress and affect the ability of the structure to carry the load according to the design. More seriously, it may also lead to the failure of concrete structural members [6].

The current methods are parameter-based and divided into empirical, machine, and deep learning methods. With the iteration of the methods, they have achieved good results. However, the biggest challenge of these methods is that many different parameters are required for prediction, which means many sensor resources are consumed. Secondly, because of the complexity of building materials, limited parameters cannot represent the properties of their materials. Therefore, a new parameter-free method based on the early creep of concrete is proposed to predict the long-term creep of concrete. The contributions are: (1). Analyze the time series data of early creep to learn and represent the material properties. (2). A hybrid cnn-transformer method without parameters is proposed. (3). An encoder based on multi-task learning and CNN is designed to reduce the influence of noise generated by creep data in sample collection.

II. Related work

The current mainstream creep prediction methods are divided into empirical analysis, statistical machine learning analysis, and deep learning.

Empirical methods utilize physical and empirical models to predict concrete creep, drawing on established theories of material behavior. Parameters such as stress, temperature, humidity, and concrete age are integral to these models. The CEB-FIP Model Code 1990, one of the pioneering models, laid the groundwork by offering detailed equations for creep and shrinkage deformations, though it may not accurately account for various factors in atypical environments [7]. Bazant et al.'s B3 model, inspired by the CEB-FIP Model Code 1990, provides a scientific approach that captures both essential and drying creep while considering a more comprehensive range of factors [8]. Based on the success of B3, the B4 model uses a more extensive database of experimental results to refine predictions and better account for environmental and material property influences [9]. However, despite their accuracy, both B3 and B4 require significant mathematical and material knowledge, making them less accessible for actual design work without specialized software. In addition, while these models are extensively used in the industry, their accuracy is constrained by the complex behavior of concrete.

Machine learning methods employ statistical techniques to develop predictive models based on historical data, capturing complex relationships without explicit physical models. Yunze et al. developed a model with 13 internal parameters for comprehensive SHAP analysis, considering the impact of environmental temperature and humidity on concrete [10]. Minfei et al.'s ensemble machine learning model, based on Bayesian optimization, achieved high accuracy but suffered from overfitting, resulting in non-smooth creep compliance curves [11]. Other studies further explored the performance variations of creep under different conditions and developed models for specific scenarios, providing various solutions in the field [12]–[14]. However, these methods are limited by the quality and completeness of input parameters, requiring improvements in robustness and generalization.

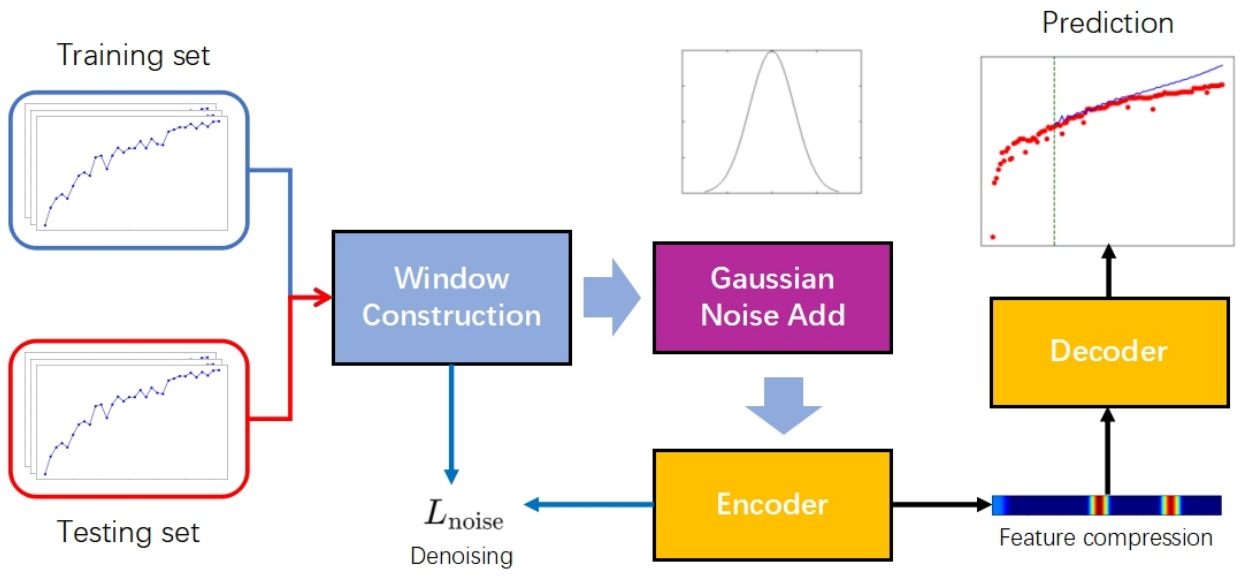


Fig. 1. The overall pipeline of the proposed parameter-free model.

Deep learning methods use neural networks with multiple layers to model complex relationships in data. These methods can automatically learn features from raw data, reducing the need for manual feature engineering. Previous research has developed various neural network models for predicting concrete creep, but traditional neural networks had limited feature extraction capabilities and did not significantly outperform ensemble learning [15]–[19]. In 2021, Jingsong et al. applied convolutional neural networks (CNNs) to unify the tasks of predicting creep and shrinkage, significantly improving prediction performance through convolutional downsampling [20]. However, this approach has shortcomings, including inadequate noise filtering from data collection errors and the training of separate models for creep and shrinkage, preventing weight sharing and deeper feature understanding. To address these issues, Conghui et al. proposed the Denoising Residual Neural Network (DRNN), which uses an encoder-decoder structure for soft sorting of tabular features, denoising, and global feature sampling, significantly improving generalization and accuracy [21].

The above three prediction methods require various internal and external factors of the concrete material, such as the type and amount of cement, aggregate, water-cement ratio, and temperature and humidity, which require many sensors to detect the material. However, obtaining these sensors for detection in the construction environment may take much work. In addition, building materials are highly complex and cannot be expressed through simple two dozen factors. A common example in the real world is two concrete materials with the same parameters, whose actual properties may differ slightly. Based on these two difficulties, this paper proposes a new method to extract and express the complexity of concrete materials while reducing the dependence on sensors.

III. Methodology

The proposed method requires only early concrete creep observation to predict long-term performance, eliminating the need for multiple sensing devices to collect material parameters. Our model captures complex nonlinear material properties by learning from detail in the early creep process using a sliding window. The detailed process is illustrated in Figure 1. The calculation process of this method is more intuitively understood through (1).

$$\hat{x}_{t+k} = f(\hat{x}_{t+k-1}, \hat{x}_{t+k-2}, \ldots, \hat{x}_{t+1}, x_t, x_{t-1}, \ldots, x_{t+k-n}) \tag{1}$$

Where $x^t + k\hat{x}_{t+k}x^t + k$ is the predicted value of the time series at a future time step $t + k$. f denotes the proposed method to make the prediction. The model uses the concept of sliding window to apply both past actual values $(\hat{x}_{t+k-1}, \hat{x}_{t+k-2}, \ldots, \hat{x}_{t+1})$ and past predicted values $(x_t, x_{t-1}, \ldots, x_{t+k-n})$ to predict long-term creep by early creep data.

Our proposed model, can be refereed at Fig 2, is a hybrid encoder-decoder architecture. The encoder, based on a Convolutional Neural Network (CNN), is responsible for feature compression and data noise reduction because of its capability of local feature sensing capability. On the other hand, the transformer decoder leverages the multi-head attention mechanism to handle long-distance dependencies, learns the extracted features, and effectively predicts outcomes.

A. Encoder

Feature acquisition in real-world scenarios often leads to significant data loss and noise. This can be attributed to various factors such as the skill level and experimental procedures of the tester collecting the data, the accuracy of facilities,

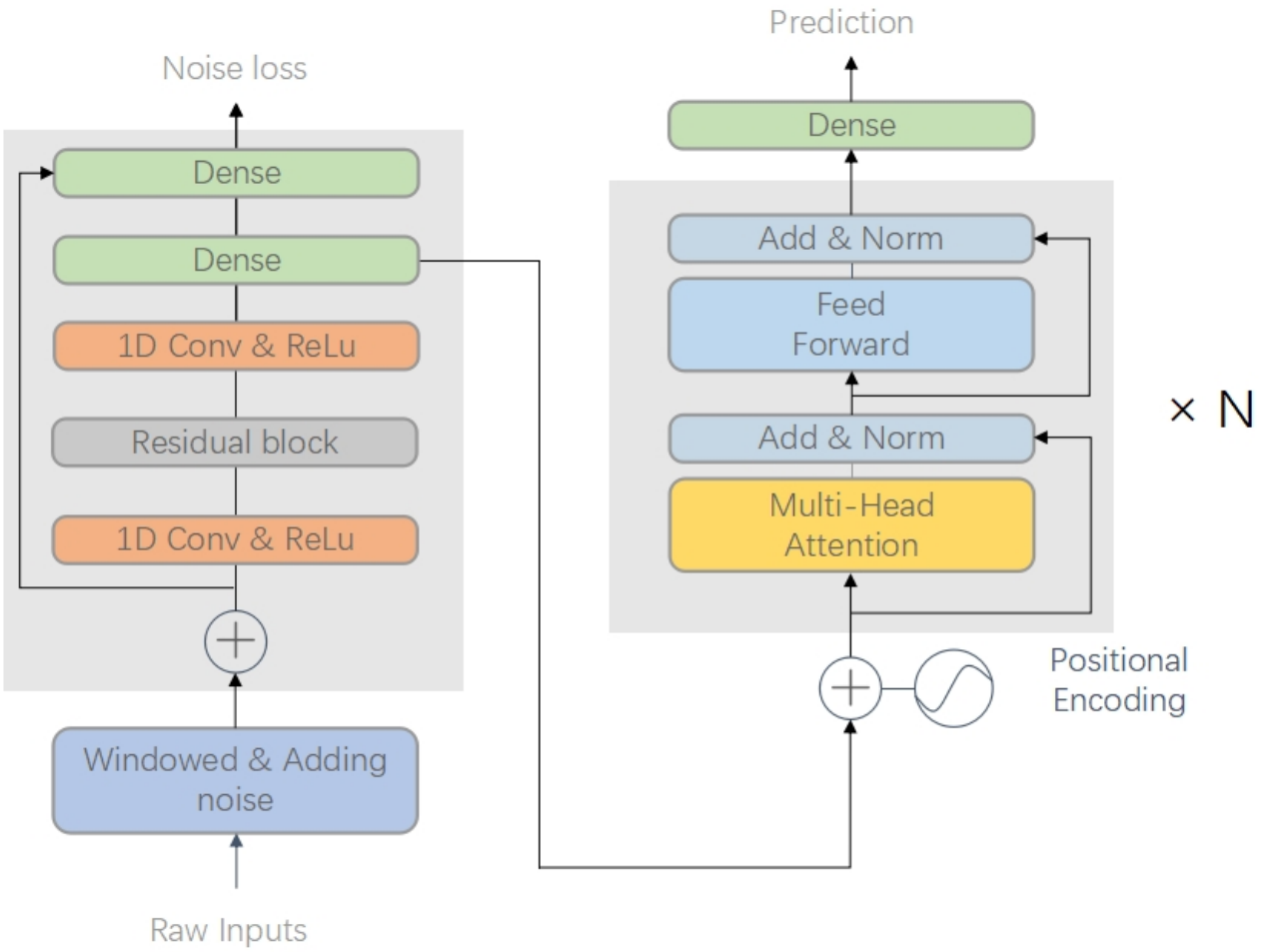


Fig. 2. The proposed hybrid encoder-decoder architecture

and changes in the engineering environment. Additionally, transformers are highly sensitive to slight variations in time series data, where even minor noise can have significant impacts. The accumulation of errors generated by time series predictions can further enhance the effects of these noises in long-term forecasts. Therefore, a critical function of the encoder module is to reduce noise.

Noise reduction is primarily achieved through residual connections and noise learning. Residual neural networks utilize simple skip connections to reduce the vanishing gradient problem, enhance learning efficiency, and preserve more noise-free information. The effectiveness of skip connections can be demonstrated through (2).

$$H(h) = F(h, W) + h \tag{2}$$

where h is the input feature to the residual block. W represents the weights of these layers. Finally, $F(h, W)$ denotes the residual function. The original information is preserved by adding the h directly to the residual function output. The network can focus on the relevant features by adjusting the weights W in $F(h, W)$, while the residual connection ensures that the original input h is still part of the final output. This ensures the network focuses on learning the difference between the input and the desired output, making reducing small errors and noise easier.

Another significant benefit of utilizing CNN encoders is their capability for feature compression. In this context, the dimension ratio between the compressed feature and the original input feature stands at 1:2. This compression reduces the computational complexity and enhances subsequent processing stages' efficiency. By diminishing the sequence length, understanding time series data becomes more manageable for the transformer decoder. The transformer architecture is known for its prowess in capturing global features and modeling long-distance dependencies. A shorter sequence length reduces the workload on the transformer decoder, enabling it to focus more effectively on learning the complicated patterns and relationships in the time series data.

Furthermore, the local feature extraction capability of CNNs is an excellent complement to the Transformer. CNNs excel at identifying and processing fine-grained, local patterns within the data through their convolutional layers. This local focus allows CNNs to capture detailed features efficiently. On the other hand, transformers are good at synthesizing these local features

into a coherent global understanding because of self-attention mechanisms. Thus, the synergy of CNNS and transformers in processing time series data is obvious. This complementary relationship improves the model's overall performance in tasks requiring detailed local analysis and a comprehensive global background, making it particularly suitable for complex creep analysis.

B. Objective function

In addition, we introduced Gaussian noise to the input data and multi-objective learning for the model to enhance noise resistance. The multi-objective function applies Mean Squared Error (MSE). The specific formula is as follows (3):

$$\mathcal{L} = \sum_{t=T+1}^{n} \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|_2^2 + \sum_{i=1}^{N} \ell(\widetilde{\mathbf{x}}_i, \widehat{\mathbf{x}}_i) + \lambda\Omega(\Theta) \quad (3)$$

The objective function $\mathcal{L}$ is composed of three main terms by addition: prediction error term, noise reduction term, and regularization term. The prediction error term measures the prediction error for the time series. $\mathbf{x}$ represents the actual value at time t, while $\widehat{\mathbf{x}}$ represents the predicted value at time t. The $|\cdot|_2^2$ denotes the squared Euclidean (L2) norm, which calculates the squared difference between the actual and predicted values. The noise reduction term is designed to reduce noise in the data. $\widetilde{\mathbf{x}}$ represents the noised input data, while $\widehat{\mathbf{x}}$ represents the predicted value. The function $\ell(\cdot)$ is a loss function that measures the difference between the input and predicted data. The regularization term represents a regularization component, where $\Omega(\Theta)$ is a regularization function applied to the model parameters Θ, and λ is a hyperparameter that controls the weight of this regularization term.

C. Decoder

We use a transformer decoder in the proposed model for global feature perception and creep prediction. Transformers are a widely used deep learning architecture. Its self-attention mechanism effectively handles long-distance dependencies and interactions within sequences. By capturing both local and global patterns through their self-attention mechanisms and multi-head attention, transformers provide a powerful ability for analyzing and predicting creep trend.

D. Data preprocessing

The Northwestern University (NU) database is a database focused on creep and shrinkage testing of concrete [22]. The creep data in this database are measured and collected using standardized experimental methods to ensure the consistency and reliability of the data. The database contains approximately 1,400 creep experiments and 1,800 shrinkage experiments. It contains a wide range of concrete components, with various cement types, aggregate sizes, and admixtures, and considers different environmental factors that affect creep, such as temperature and humidity. Some of these experiments also include the influence of admixtures in contemporary concrete formulations.

Due to the inconsistent time intervals present in the NU dataset during data collection, we employed the random forest method to supplement the missing data in this study. This process can ensure that the time interval after supplementation was uniformly set to 1 day. The performance of data completion is 2.55 in MAE and 4.07 in RMSE.

IV. RESULT ANALYSIS

TABLE I
PERFORMANCE METRICS FOR DIFFERENT MODELS ON NU DATABASE

Model	Metric	60 days	90 days	120 days	Overall
B4[9]	$\mathbf{R^2}$	0.62	0.64	0.64	0.70
	MAE	14.26	14.92	16.33	19.14
	RMSE	15.29	18.66	21.41	38.17
SVR[12]	$\mathbf{R^2}$	0.89	0.89	0.89	0.91
	MAE	5.77	5.94	6.22	6.94
	RMSE	10.83	10.95	11.64	13.25
LGBM[11]	$\mathbf{R^2}$	0.92	0.93	0.93	0.94
	MAE	5.13	5.20	5.21	5.41
	RMSE	10.24	12.15	12.81	14.81
XGBoost[11]	$\mathbf{R^2}$	0.93	0.93	0.94	0.95
	MAE	4.61	4.60	**4.65**	**4.83**
	RMSE	8.63	8.85	11.98	13.90
C&S CNN[20]	$\mathbf{R^2}$	0.91	0.92	0.92	0.93
	MAE	5.41	5.44	5.87	6.48
	RMSE	10.50	11.58	12.09	13.64
DRNN[21]	$\mathbf{R^2}$	0.94	0.94	0.94	0.95
	MAE	4.96	5.16	5.37	5.83
	RMSE	6.23	7.58	8.70	8.88
Transformer[23]	$\mathbf{R^2}$	0.94	0.94	0.95	0.95
	MAE	3.02	5.18	7.10	7.10
	RMSE	3.68	5.80	9.66	9.66
Proposed method	$\mathbf{R^2}$	**0.95**	**0.95**	**0.95**	**0.95**
	MAE	**2.63**	**4.12**	6.85	6.85
	RMSE	**3.04**	**4.77**	**8.28**	**8.28**

We evaluated the performance of both the baseline and proposed models over various periods, explicitly dividing the experiment into 30 to 60-day, 90-day, and 120-day creep predictions. The overall performance indicates the result of using all the data in experiments to predict. For the parameter-based method, we use the data of the whole life cycle for analysis. However, for the parameter-free method, because the model is based on time series, we set the maximum period of each experiment to 120 days. As shown in Table I, the proposed method and Transformer initially show very high performance, but over time, their performance begins to significantly decrease due to the accumulation of errors in time series predictions. In contrast, the parameter-based model starts with a high error. However, as the test data increases and incorporates more variety of creep states, its performance has a very small decrease. This evidence suggests that the parameter-based model cannot effectively leverage the time series characteristics in the creep process. Despite this, the proposed model achieves state-of-the-art performance compared to other baseline methods.

V. CONCLUSION

This research proposes a parameter-free method to predict long-term concrete creep performance. While the proposed method achieves state-of-the-art results, some limitations may require further improvement.

1) Long-term predictions accumulate errors with initial inaccuracies magnifying over time to reduce the reliability.
2) The random forest method introduces errors in filling in missing data, leading to imprecise predictions due to incomplete or unevenly spaced data points.
3) Ignoring the influence of critical static parameters like water-cement and aggregate-to-cement ratios results in prediction variances, limiting the model's accuracy in capturing concrete's material properties.

In the future, we will implement the following improvements:

1) Develop models that integrate inner static material parameters with early creep data to improve prediction accuracy and reliability, providing a more comprehensive understanding of material behavior.
2) Create an adaptive model capable of handling varying data intervals, ensuring robust and accurate predictions despite irregular data sampling.

REFERENCES

[1] Z. P. Bazant and R. L'Hermite, *Mathematical modeling of creep and shrinkage of concrete*. Wiley, 1988.

[2] Z. P. Bažant and M. Jirásek, *Creep and Hygrothermal Effects in Concrete Structures* (Solid Mechanics and Its Applications), 1st ed. Dordrecht: Springer Dordrecht, Feb. 2018, pp. L, 918, 325 b/w illustrations, 10 illustrations in colour, ISBN: 978-94-024-1136-2. DOI: 10.1007/978-94-024-1138-6. (visited on 01/17/2018).

[3] D. J. Hannant, "The mechanism of creep in concrete," *Matériaux et Construction*, vol. 1, no. 5, pp. 403–410, 1968, ISSN: 1871-6873. DOI: 10.1007/BF02473738. [Online]. Available: https://doi.org/10.1007/BF02473738.

[4] Z. P. Bažant, "Prediction of concrete creep and shrinkage: Past, present and future," *Nuclear Engineering and Design*, vol. 203, no. 1, pp. 27–38, 2001, ISSN: 0029-5493. DOI: https://doi.org/10.1016/S0029-5493(00)00299-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029549300002995.

[5] Z. P. Bazant and J.-C. Chern, "Bayesian statistical prediction of concrete creep and shrinkage," *ACJ Journal, Proceedings*, vol. 81, no. 4, pp. 319–330, 1984.

[6] Y. Peng, R. Li, D. Bie, X. Yao, X. Liu, and Y. Duan, "A coupled creep and damage model of concrete considering rate effect," *Journal of Building Engineering*, vol. 45, p. 103 621, 2022, ISSN: 2352-7102. DOI: https://doi.org/10.1016/j.jobe.2021.103621.

[7] C. E.-I. du Béton, *CEB-FIP Model Code 1990: Design Code*. London: Thomas Telford Publishing, 1993.

[8] Z. P. Bazant and S. Baweja, "Creep and shrinkage prediction model for analysis and design of concrete structures: Model b3," *ACI Special Publications*, vol. 194, pp. 1–84, 2000.

[9] M. H. Hubler, R. Wendner, and Z. P. Bažant, "Statistical justification of model b4 for drying and autogenous shrinkage of concrete and comparisons to other models," *Materials and Structures*, vol. 48, no. 4, pp. 797–814, 2015, ISSN: 1871-6873. DOI: 10.1617/s11527-014-0516-z. [Online]. Available: https://doi.org/10.1617/s11527-014-0516-z.

[10] Y. Liu, Y. Li, J. Mu, H. Li, and J. Shen, "Modeling and analysis of creep in concrete containing supplementary cementitious materials based on machine learning," *Construction and Building Materials*, vol. 392, p. 131 911, 2023, ISSN: 0950-0618. DOI: https://doi.org/10.1016/j.conbuildmat.2023.131911.

[11] M. Liang, Z. Chang, Z. Wan, Y. Gan, E. Schlangen, and B. Šavija, "Interpretable ensemble-machine-learning models for predicting creep behavior of concrete," *Cement and Concrete Composites*, vol. 125, p. 104 295, 2022, ISSN: 0958-9465. DOI: https://doi.org/10.1016/j.cemconcomp.2021.104295.

[12] K. Li, Y. Long, H. Wang, and Y.-F. Wang, "Modeling and sensitivity analysis of concrete creep with machine learning methods," *Journal of Materials in Civil Engineering*, vol. 33, no. 8, p. 04 021 206, 2021. DOI: 10.1061/(ASCE)MT.1943-5533.0003843.

[13] Y. Bouras and L. Li, "Prediction of high-temperature creep in concrete using supervised machine learning algorithms," *Construction and Building Materials*, vol. 400, p. 132 828, 2023, ISSN: 0950-0618. DOI: https://doi.org/10.1016/j.conbuildmat.2023.132828.

[14] N. Tošić, A. de la Fuente, and S. Marinković, "Creep of recycled aggregate concrete: Experimental database and creep prediction model according to the fib model code 2010," *Construction and Building Materials*, vol. 195, pp. 590–599, 2019, ISSN: 0950-0618. DOI: https://doi.org/10.1016/j.conbuildmat.2018.11.048.

[15] M. R. Taha, A. Noureldin, N. El-Sheimy, and N. G. Shrive, "Artificial neural networks for predicting creep with an example application to structural masonry," *Canadian Journal of Civil Engineering*, vol. 30, no. 3, pp. 523–532, 2003. DOI: 10.1139/l03-003.

[16] A. El-Shafie and S. Aminah, "Dynamic versus static artificial neural network model for masonry creep deformation," *Proceedings of the Institution of Civil Engineers - Structures and Buildings*, vol. 166, no. 7, pp. 355–366, 2013. DOI: 10.1680/stbu.11.00024.

[17] M. M. Abed, A. El-Shafie, and S. A. Bt. Osman, "Creep predicting model in masonry structure utilizing dynamic neural network," *Journal of Computer Science*, vol. 6, no. 5, pp. 597–605, May 2010. DOI: 10.3844/jcssp.2010.597.605.

[18] A. El-Shafie, T. Abdelazim, and A. Noureldin, "Neural network modeling of time-dependent creep deformations in masonry structures," *Neural Computing and Applications*, vol. 19, no. 4, pp. 583–594, 2010, ISSN: 1433-3058. DOI: 10.1007/s00521-009-0318-3. [Online]. Available: https://doi.org/10.1007/s00521-009-0318-3.

[19] L. Bal and F. Buyle-Bodin, "Artificial neural network for predicting creep of concrete," *Neural Computing and Applications*, vol. 25, no. 6, pp. 1359–1367, 2014, ISSN: 1433-3058. DOI: 10.1007/s00521-014-1623-z. [Online]. Available: https://doi.org/10.1007/s00521-014-1623-z.

[20] J. Zhu and Y. Wang, "Convolutional neural networks for predicting creep and shrinkage of concrete," *Construction and Building Materials*, vol. 306, p. 124 868, 2021, ISSN: 0950-0618. DOI: https://doi.org/10.1016/j.conbuildmat.2021.124868.

[21] C. Li, M. Zhang, and X. Zhang, "Enhancing concrete creep prediction with deep learning: A soft-sorted one-dimensional cnn approach," *IEEE Access*, vol. 11, pp. 139 314–139 325, 2023. DOI: 10.1109/ACCESS.2023.3340425.

[22] M. Hubler, R. Wan-Wendner, and Z. Bazant, "Comprehensive database for concrete creep and shrinkage: Analysis and recommendations for testing and recording," *ACI Materials Journal*, vol. 112, Aug. 2015. DOI: 10.14359/51687453.

[23] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Learnable Cross-Correlation based Filter-and-Sum Networks for Multi-channel Speech Separation

Xianrui Wang*†, Shiqi Zhang†‡, Bo He†, Shoji Makino†, and Jingdong Chen*
* CIAIC and Shaanxi Provincial Key Laboratory of Artificial Intelligence,
Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
† Graduate School of Information, Production and Systems, Waseda University, Kitakyushu 808-0135, Japan
‡ Audio Research Group, Tampere University, Tampere, Finland

***Abstract*—Multichannel source separation plays an important role in audio and speech signal processing. With recent advancements in deep neural networks (DNN), numerous DNN-based beamforming algorithms have been developed. To leverage spatial information, a time domain filter-and-sum network (FaSNet) was introduced, and the transform average concatenate (TAC) technique was subsequently adopted to further enhance separation performance. FaSNet captures spatial information by assessing cosine similarity between different channels; but this approach may have limited spatial resolution and could exhibit bias in noisy, reverberant environments, thereby potentially compromising performance. Motivated by the efficacy of the generalized cross-correlation (GCC) method in achieving reliable source localization in adverse environments, this paper introduces a learnable cross-correlation (LCC) module for FaSNet and FaSNet-TAC. By offering improved flexibility and robustness across diverse environments, LCC enhances source separation performance, which is validated by several simulations.**

***Index Terms*—Multichannel source separation, neural network based beamfroming, spatial information, learnable cross-correlation.**

I. Introduction

Separating different source signals in noisy and reverberant environments [1]–[3] is of vital importance. Beamforming is one of the most widely used methods to address this problem [4]–[7]. With recent advancements in deep neural networks (DNN), numerous DNN-based beamforming algorithms have been proposed [8]–[10]. As shown in [3], [7], [11], incorporating spatial information can help significantly improve the separation performance. To this end, a filter-and-sum network (FaSNet) [12] was proposed. The original FaSNet comprises several Temporal Convolutional Network (TCN) modules [13], which are employed to separate source signals. Additionally, cosine similarity modules are utilized to capture spatial information. To further enhance separation performance, TCN modules are replaced with dual-path recurrent neural network (DPRNN) modules [14], [15]. Moreover, the transform average concatenate (TAC) technique was adopted [15]. TAC first transforms each channel's feature into a latent space and then the average of transformed features is concatenated into following blocks.

While FaSNet [12] and FaSNet-TAC [15] have demonstrated promising separation performance, they rely on cosine similarity to extract spatial information, which may suffer from limited spatial resolution and bias in noisy, reverberant environments. Generalized cross-correlation (GCC) is one of the most widely used methods in source localization [16]–[20]. GCC modifies the cross-correlation through a frequency domain weight function. There is a number of member algorithms in the GCC family, which consider different weight functions and their proprieties depend on the weight functions. For example, phased transform (PHAT) [21] is robust to reverberation while Roth processor [22] and smoothed coherence transform (SCOT) [23] are robust to noise. Inspired by the principles in GCC, we propose to design in this paper a learnable cross-correlation (LCC) module, in which a time domain convolution function is learned as the weight function from input data adaptively. By replacing the cosine similarity with the proposed LCC module, we derive two new frameworks, i.e., LCC-FaSNet and LCC-FaSNet-TAC. These frameworks offer enhanced separation performance, particularly in scenarios where two target speakers are closely distributed, due to LCC's ability to achieve higher spatial

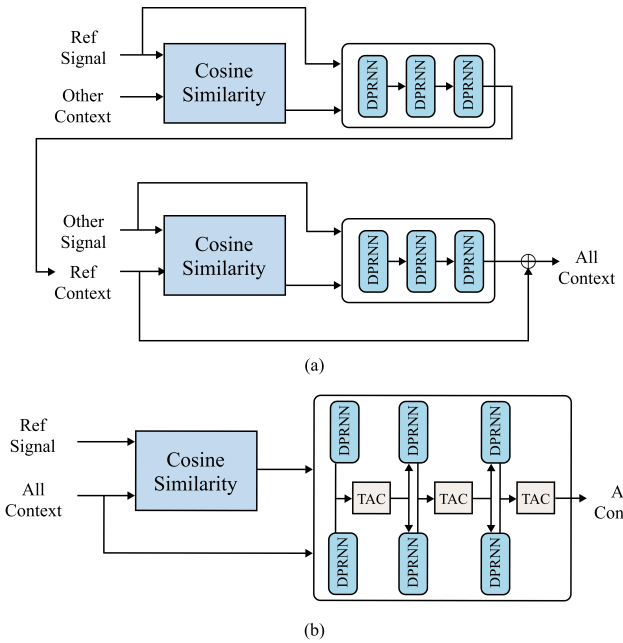


Fig. 1. Structures of: (a). two-stage FaSNet, (b). single-stage FaSNet-TAC.

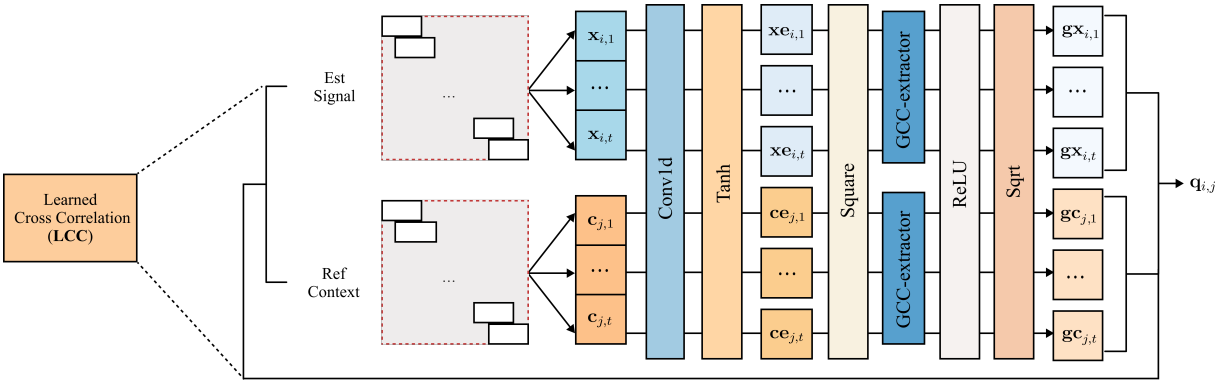


Fig. 2. The structure of the proposed LCC module.

resolution. Moreover, as the LCC module dynamically learns the weight function from input data, the proposed frameworks are anticipated to outperform conventional approaches across diverse environments.

The remainder of the paper is organized as follows. In Section II, we present a brief overview of the cosine similarity feature used in the original FaSNet and FaSNet-TAC methods. We then introduce the proposed LCC module in Section III. In Section IV, we present the training configuration and compare the proposed methods with their conventional counterparts. Finally, we draw the conclusion in Section V.

II. Cosine Similarity Feature

Incorporating spatial information properly is shown to be an effective approach to improve source separation performance [3], [7]. In this paper, we consider the original two-stage FaSNet structure [12] and single-stage FaSNet-TAC structure [15], both of which incorporate DPRNN modules. As depicted in Fig. 1 (a) and Fig. 1 (b), FaSNet and FaSNet-TAC extract the spatial information using the cosine similarity feature [12], [15], which is named normalized cross-correlation (NCC). Mathematically, NCC is defined as

$$\mathrm{NCC}_{\mathbf{s}_1,\mathbf{s}_2} = \frac{\mathbf{s}_1^T \mathbf{s}_2}{||\mathbf{s}_1||_2 ||\mathbf{s}_1||_2}, \tag{1}$$

where $\mathbf{s}_1$, $\mathbf{s}_2$ are two signal vectors and $||\cdot||_2$ denotes ℓ_2 norm.

For the two-stage FaSNet, in the initial stage, a single microphone is selected as the reference. The NCCs between the signals picked up by the other microphones and the reference microphone are computed to estimate the target signal. In the subsequent stage, the NCCs between the microphone signals and the estimated target signal are employed to further enhance the separation performance. While for the single-stage FaSNet-TAC, NCCs are only calculated at the beginning and the spatial information are maintained with TAC blocks.

III. Proposed Learnable Cross-Correlation Module

While FaSNet and FaSNet-TAC have demonstrated promising separation performance, they rely on cosine similarity to extract spatial information, which may suffer from limited spatial resolution and bias in noisy, reverberant environments. To circumvent this limitation, we introduce a new, lightweight LCC module in this work. The LCC module, as illustrated in Fig. 2, is a dual-layer convolutional neural network operating at the time-domain frame level. It first partitions the input signals $\mathbf{x}_i$, $i = 1, ..., N$, into short frames (named "chunk") with a frame length of L and a hop size of $H \in [0, L-1]$, i.e.,

$$\mathbf{x}_{i,t} = \mathbf{x}_i[tH : tH + L - 1], \tag{2}$$

where t is the frame index. We additionally form a signal vector at the corresponding time index, referred to as the "**context**", consisting of extra W immediate future and past samples relative to $\mathbf{x}_{i,t}$, i.e.,

$$\mathbf{c}_{j,t} = \mathbf{x}_j[tH - W : tH + L + W - 1]. \tag{3}$$

For simplicity, we omit the frame index t in subsequent discussions, as it is unambiguous. Following input signal segmentation, we utilize a pair of convolutional layers with shared parameters, labeled as $Conv_1(\cdot)$, to encode both the context signal $\mathbf{c}_j$ and the chunk signal $\mathbf{x}_i$. Subsequently, in order to preserve the dynamic range of subsequent neural network inputs, we scale their outputs to the range $[-1, 1]$ using the $\tanh(\cdot)$ function. Mathematically, this process is expressed as

$$\begin{aligned} \mathbf{ce}_j &= \tanh\left[Conv_1(\mathbf{c}_j)\right], \\ \mathbf{xe}_i &= \tanh\left[Conv_1(\mathbf{x}_i)\right], \end{aligned} \tag{4}$$

where $\mathbf{ce}_j \in \mathbb{R}^{O_{c1} \times (2W+L-I_{c1}+1)}$ denotes the encoded context signal, I_{c1} and O_{c1} denote, respectively, the number of input and output channels of the first convolutional layer (In this paper, we set $I_{c1} = O_{c1} = L$ for simplicity in computation), $\mathbf{xe}_i \in \mathbb{R}^{O_{c1} \times 1}$ is the encoded chunk signal.

Subsequently, the encoded vector is squared and passed through the GCC extractor (a single convolutional layer), denoted as $Conv_2(\cdot)$, to dynamically adjust its information in the O_{c1} dimension. The ReLU activation function is then applied to remove negative values, and the square root operation is

TABLE I
EXPERIMENTAL RESULTS BASED ON A 6-ELEMENT CIRCULAR MICROPHONE ARRAY WITH DIFFERENT SPEAKER ANGLES.

Model	# Parameters	SDRi (dB)					SI-SNRi (dB)				
		Speaker angle				Overall	Speaker angle				Overall
		<15°	15-45°	45-90°	>90°		<15°	15-45°	45-90°	>90°	
FaSNet	3.70M	4.77	5.50	6.18	6.59	5.81	2.93	3.58	4.15	4.48	3.83
+LCC (proposal)	+4.22K	**5.02**	**5.96**	**6.75**	**7.05**	**6.26**	**3.99**	**4.84**	**5.57**	**5.79**	**5.11**
FaSNet-TAC	2.76M	8.48	10.55	12.19	12.88	11.06	7.75	9.85	11.49	12.23	10.48
+LCC (proposal)	+4.22K	**9.69**	**11.58**	**12.98**	**13.60**	**12.09**	**8.94**	**10.87**	**12.28**	**12.94**	**11.39**

TABLE II
EXPERIMENT RESULTS BASED ON A 6-ELEMENT CIRCULAR MICROPHONE ARRAY WITH DIFFERENT OVERLAP RATES

Model	# Parameters	SDRi (dB)				SI-SNRi (dB)			
		Overlap rate				Overlap rate			
		<25%	25-50%	50-75%	>75%	<25%	25-50%	50-75%	>75%
FaSNet	3.70M	7.98	5.87	4.57	3.54	5.60	3.88	2.85	1.94
+LCC (proposal)	+4.22K	**8.74**	**6.21**	**4.85**	**3.80**	**7.31**	**5.11**	**3.86**	**2.84**
FaSNet-TAC	2.76M	15.82	10.70	8.74	7.17	14.97	9.98	8.02	6.48
+LCC (proposal)	+4.22K	**16.63**	**11.51**	**9.59**	**8.18**	**15.98**	**10.79**	**8.86**	**7.45**

performed to yield the corresponding Pearson coefficient-like term [24], [25], i.e.,

$$\begin{aligned} \mathbf{gc}_j &= \sqrt{\mathrm{ReLU}\big[Conv_2(\mathbf{ce}_j^2)\big]}, \\ \mathbf{gx}_i &= \sqrt{\mathrm{ReLU}\big[Conv_2(\mathbf{xe}_i^2)\big]}. \end{aligned} \tag{5}$$

Finally, the learnable cross-correlation feature is calculated with following equation:

$$\begin{cases} \mathbf{c}_{j,p} = \mathbf{c}_i[p : p + L - 1] \\ \mathbf{f}_{i,j,p} = \dfrac{\mathbf{x}_i \mathbf{c}_{j,p}^T}{\|\mathbf{gc}_{j,p}\|_2 \|\mathbf{gx}_i\|_2} \end{cases}, \quad j = 1, ..., 2W + 1, \tag{6}$$

where $p \in [0, 2W - I_{c1} + 1]$ represents the position of the focused same-length segment in $\mathbf{c}_j$.

IV. EXPERIMENT

In this section, we provide a detailed description of the training configuration and conduct several simulations in both noisy and reverberant environments. To validate the effectiveness of the proposed LCC module, we conduct a comparative analysis of the proposed networks exclusively against their conventional counterparts, i.e., FaSNet and FaSNet-TAC. The proposed module could also be integrated into other networks.

A. Dataset

The target signals are randomly selected from LibriSpeech [26]. Noise signals from CHiME3 [27] are chosen to control the signal-to-noise ratio (SNR). The sampling rate for both target and noise signals is $16\,\mathrm{kHz}$. We consider scenarios where there are two target speakers with the relative signal-to-interference ratio (SIR) ranging from $0\,\mathrm{dB}$ to $5\,\mathrm{dB}$. One point noise is added and the corresponding SNR is set to range from $-5\,\mathrm{dB}$ and $30\,\mathrm{dB}$.

We consider a room with randomly sampled dimensions, where the length and width fall between 3 and 10 m, and the height ranges from 2.5 to 4 m. The reflection coefficients of all the room surfaces are assumed identical and they are randomly generated to achieve a reverberation time T_{60} ranging from $100\,\mathrm{ms}$ to $500\,\mathrm{ms}$. A uniform circular array with a radius of $5\,\mathrm{cm}$ and consisting of 6 omnidirectional microphones, is utilized. The array is positioned randomly within the room, ensuring that the minimum distance between the array center and each wall is $0.5\,\mathrm{m}$. Speaker positions are then sampled to ensure the average speaker angle relative to the microphone center is uniformly distributed between $0°$ and $180°$. Noise source position is sampled without additional constraints. Subsequently, the microphone observation signals are generated by convolving the speech and noise source signals with their corresponding room impulse responses (RIRs) generated using the toolbox gpuRIR [28], which is based on the image model [29] and adding the results together with the specified SIR and SNR values. In total, 100,000, 10,000, and 5,000 4-second-long utterances are generated as the training, validation, and evaluation sets, respectively.

B. Experimental configuration

Using the generated datasets, we trained and tested the proposed LCC-FaSNet and LCC-FaSNet-TAC, as well as the original baseline models without LCC. The training objective

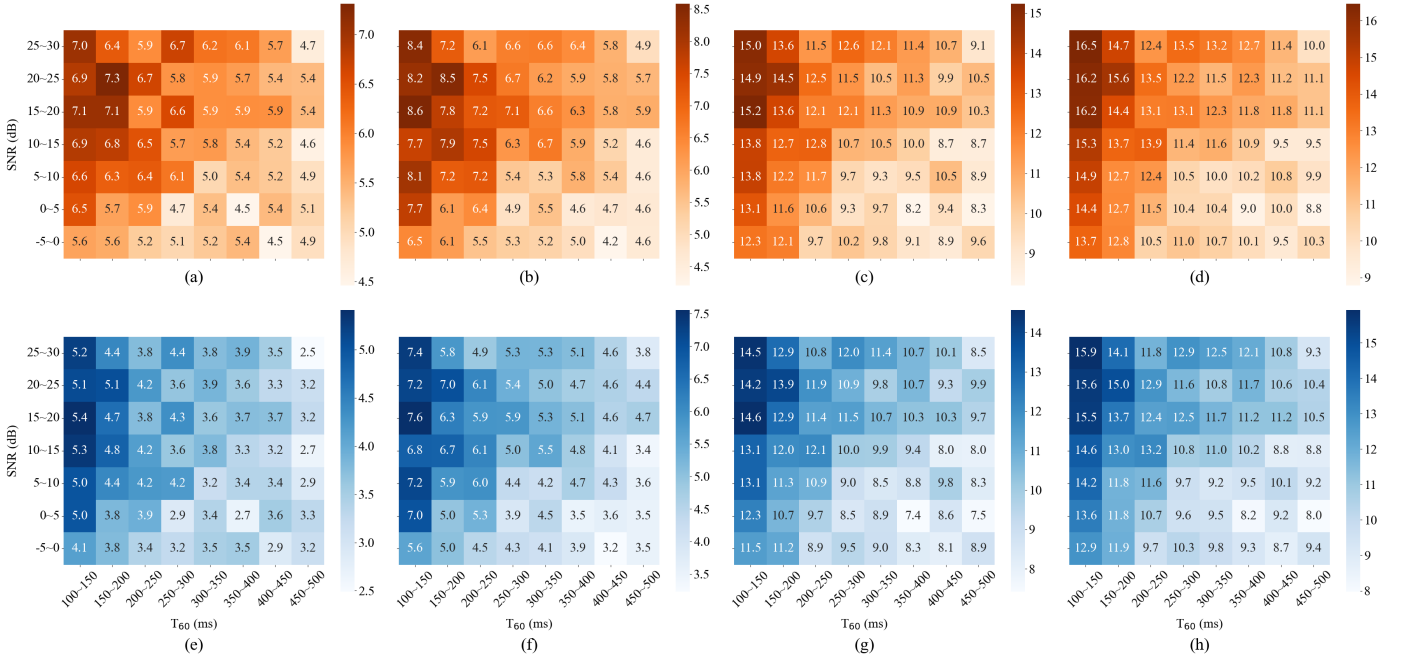


Fig. 3. Performance of the compared methods in noisy and reverberant environments: (a) SDRi of FaSNet, (b) SDRi of LCC-FaSNet, (c) SDRi of FaSNet-TAC, (d) SDRi of LCC-FaSNet-TAC, (e) SI-SNRi of FaSNet, (f) SI-SNRi of LCC-FaSNet, (g) SI-SNRi of FaSNet-TAC, and (h) SI-SNRi of LCC-FaSNet-TAC. The horizontal axis represents T_{60} and the vertical axis represents SNR (dB).

is scale-invariant SNR (SI-SNR) [30]. Models were trained using the Adam [31] optimizer with a momentum of 0, batch size of 8, and 120 epochs. The initial learning rate was set to 2×10^{-3}. If the loss did not decrease after 3 consecutive epochs, the learning rate was halved. Training was halted either when the learning rate did not decrease for two successive epochs or when the maximum number of iterations was reached. The model with the lowest loss on the validation set was chosen for testing and results recording. All experiments were conducted on an NVIDIA A40 with 48 GB of GPU memory.

C. Results and Discussions

Before examining performance, we analyze the additional memory cost of the proposed LCC module. As shown in Tab. I and Tab. II, the proposed LCC module contributes only $4.22\,\mathrm{k}$ to the parameter count, which is less than 0.2% of the parameter volume of the baseline models. Specifically, for FaSNet, LCC utilizes an additional $4.22\,\mathrm{k}$ (0.11%) parameters, and for FaSNet-TAC, LCC uses only 0.15% more parameters.

Now, we investigate the separation performance with variations in the angle between two target speakers. Besides SI-SNR improvement (SI-SNRi), we also show signal-to-distortion ratio improvement (SDRi) [32]. The results are shown in Tab. I. As seen, the proposed methods consistently outperform their conventional counterparts across all cases. Furthermore, the TAC technique significantly improves separation performance. Notably, as the angle between the two speakers decreases, especially below 15°, the proposed frameworks achieve notably superior separation performance. This improvement is attributed to the limited spatial resolution of cosine similarity, which leads to performance degradation in FaSNet and FaSNet-TAC. Conversely, the proposed LCC module offers higher spatial resolution, thereby enhancing separation performance. Additionally, as seen in Tab. II, the proposed methods achieve superior separation performance across various overlap ratios between the two speakers, indicating the suitability of the LCC module for diverse applications.

Moreover, segmented tests were conducted across different SNR levels and T_{60} values. The results are presented in Fig. 3 where darker colors indicate larger gain under the current conditions, while lighter colors signify less gain. It is evident that the proposed methods consistently outperform their counterparts across all configurations, which validates the robustness and flexibility of the proposed LCC module.

V. Conclusion

Separating source signals in noisy and reverberant environments is an important task in audio signal processing. In the so-called FaSNet and FaSNet-TAC, cosine similarity is chosen to represent the inter-channel feature. However, this choice possesses limited spatial resolution and may not be optimal across varied environments. To take full advantage of spatial information, in this paper, we designed a learnable cross-correlation (LCC) module, which dynamically generates a weighted correlation from input observations. We integrated the proposed module into FaSNet and FaSNet-TAC. Simulations confirm that the proposed LCC module enhances separation performance, particularly in scenarios with closely

distributed target speakers, owing to its high resolution. Furthermore, improvements are observed across different overlap ratios, SNR levels, and reverberation levels, underscoring the robustness and flexibility of the proposed LCC module.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0201502, and in part by the Major Program of National Science Foundation of China (NSFC) under Grant 62192713, and in part by the China Scholarship Council (CSC).

REFERENCES

[1] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 715–726, 2007.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: signal processing techniques and applications*, Berlin Heidelberg, Germany: Springer, 2001, pp. 157–180.

[3] X. Wang, A. Brendel, G. Huang, Y. Yang, W. Kellermann, and J. Chen, "Spatially informed independent vector analysis for source extraction based on the convolutive transfer function model," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[5] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

[6] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer, 2008.

[7] Y. Yang, X. Wang, W. Zhang, and J. Chen, "Independent vector analysis assisted adaptive beamfomring for speech source separation with an acoustic vector sensor," in *Proc. IWAENC*, 2022, pp. 1–5.

[8] X. Xiao, S. Watanabe, H. Erdogan, *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE ICASSP*, 2016, pp. 5745–5749.

[9] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *Proc. IEEE ICASSP*, 2017, pp. 271–275.

[10] T. N. Sainath, R. J. Weiss, K. W. Wilson, *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 965–979, 2017.

[11] X. Wang, K. Mo, Y. Yang, L. Zhang, S. Makino, and J. Chen, "A cascaded semi-blind source separation method for joint acoustic echo cancellation, interference suppression, and noise reduction," in *Proc. IWAENC*, 2024, pp. 369–373.

[12] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. ASRU*, 2019, pp. 260–267.

[13] Y. Luo and N. Mesgarani, "Conv-Tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[14] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE ICASSP*, 2020, pp. 46–50.

[15] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE ICASSP*, 2020, pp. 6394–6398.

[16] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors," *J. Signal Process. Sys.*, vol. 63, pp. 265–275, 2011.

[17] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, 2019.

[18] X. Wang, G. Huang, J. Benesty, J. Chen, and I. Cohen, "Time difference of arrival estimation based on a kronecker product decomposition," *IEEE Signal Process. Lett.*, vol. 28, pp. 51–55, 2020.

[19] Y. Zou and H. Liu, "TDOA localization with unknown signal propagation speed and sensor position errors," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1024–1027, 2020.

[20] M. Cobos, F. Antonacci, L. Comanducci, and A. Sarti, "Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1270–1281, 2020.

[21] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Process.*, vol. 92, no. 8, pp. 1950–1960, 2012.

[22] P. R. Roth, "Effective measurements using digital signal analysis," *IEEE spectrum*, vol. 8, no. 4, pp. 62–70, 1971.

[23] E. A. Cohen and A. T. Walden, "A statistical study of temporally smoothed wavelet coherence," *IEEE Trans. on Signal Process.*, vol. 58, no. 6, pp. 2964–2973, 2010.

[24] J. Benesty, J. Chen, and Y. Huang, "On the importance of the pearson correlation coefficient in noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 4, no. 16, pp. 757–765, 2008.

[25] I. Cohen, Y. Huang, J. Chen, *et al.*, "Pearson correlation coefficient," in *Noise reduction in speech processing*, Springer, 2009, pp. 1–4.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain

audio books," in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.

[27] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, 2015, pp. 504–511.

[28] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "GpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimed. Tools and Appl.*, vol. 80, no. 4, pp. 5653–5671, Oct. 2020.

[29] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, Apr. 1979.

[30] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" In *Proc. IEEE ICASSP*, 2019, pp. 626–630.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

Is Corpus Suitable for Human Perception?: Quality Assessment of Voice Response Timing in Conversational Corpus through Timing Replacement

Sadahiro Yoshikawa* Ryo Ishii† Shogo Okada*
* Japan Advanced Institute of Science and Technology, Japan
E-mail: {s2230033, okada-s}@jaist.ac.jp
† Human Informatics Laboratories, NTT Corporation, Japan
E-mail: {ryo.ishii.ct}@hco.ntt.co.jp

***Abstract*—The timing estimation models in spoken dialogue systems (SDSs) have typically been trained by human responses in order to achieve the appropriate response timing. However, human response timings are not always appropriate: in a previous experiment in which annotators listened to responses with the timings replaced by fixed values, some responses with the mode value sounded more realistic than actual human responses. Since this previous experiment was a small-scale preliminary one that only showed that some speakers tended to be significantly preferred, in the current study, we conducted an experiment on about 1,700 human responses, and scored whether they could be replaced with the mode value. The results showed that the annotators tended to feel that mode (or perhaps from 0 ms to 400 ms) responses are more appropriate than actual overlappings. We determined the responses that could and could not be replaced with the mode value by a chi-square test and then formulated a detection task to predict them from the scores. The evaluation results showed that our proposed simple model outperformed random selection with the AUC of 0.650. On the basis of these results, we present examples of SDSs, using the score to predict which responses or response timings are appropriate for the SDS users. Our findings may suggest a more efficient way to determine the appropriate response timing for SDSs compared to training models by corpus data.**

I. Introduction

To achieve natural responses in spoken dialogue systems (SDSs), it is crucial to predict the response time that humans perceive as appropriate and then have the system respond at the appropriate timing. However, in daily conversations, the appropriate response time for humans is not always clear, depending as it does on human perception. In addition, the speaker who responds to the interlocutor's utterances does not always pay attention to the response time, so some responses may sound inappropriate (early or late) to the observer. A previous study reported that some responses with the timing replaced by a fixed value sound more like real conversations than actual human responses [1]. This implies that even if a response generation model is trained using response times recorded in corpus data (i.e., actual conversations), it is not always possible to respond with the appropriate timing. Therefore, to achieve the appropriate response timing, it is crucial to not only train models by corpus data, but also to analyze what response times humans perceive as appropriate for each response.

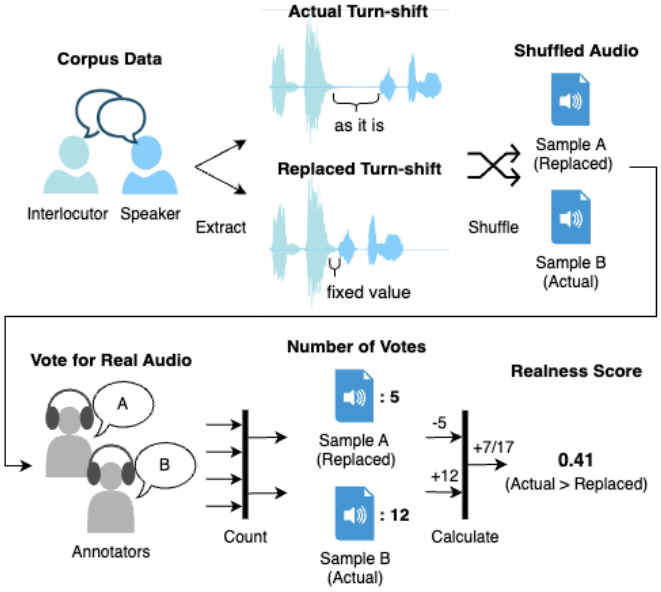


Fig. 1: Procedure of listening test and calculation of realness score (RS). Extraction of Corpus Data is explained in Section III-A, and Vote for Real Audio and calculation of RS is explained in Section III-B.

In the current study, we expanded on previous research [1] with two methodologies. First, since the previous study only showed that some speakers tended to be significantly preferred between an actual response and a response in which the timing was replaced by a fixed value, we utilized more data to analyze the differences for each response to explore the common characteristics of response timing across speakers in Section III. Second, we built a model based on the acquired data and explored how well the model could predict whether response times could be replaced by a fixed value in Section IV. Next, on the basis of these results, we suggested applications to SDSs, using the model to predict which responses or response timings are appropriate for the users in Section VI.

First, we conducted a similar listening test to the previous study to analyze which response times are considered appro-

priate for humans. The procedure of the listening test is shown in Fig. 1. For the corpus data of the listening test, as we aim to explore the diversity of responses for each speaker, we chose a corpus of non-task conversations in Japanese between a speaker and multiple interlocutors. Through several filter conditions, we made a dataset of about 1,700 turn-changes. For voting for the dataset, we gathered 17 annotators for each speaker, and asked them to vote which response voice in the audio samples sounded more realistic: the actual human response (called *actual*) or the response with the timing replaced by a fixed value (called *replaced*). We then defined a realness score (RS) as the average of the votes of the annotators for each response and utilized it to assess the extent to which humans perceived each response time as a real conversation. Since a medium RS contains some randomness, we also performed a statistical test to determine which response timings have statistically significant preferences (SSPs) between *actual* and *replaced*: in other words, which responses can be replaced by a fixed value and which cannot.

As a prediction task for RS, we evaluated how well a prediction model based on neural networks detected which turn-changes have SSPs. The recall rate of SSPs for *actual* indicates the ability of the model to detect the responses for which timing estimation is needed, and that for *replaced* reflects its ability to detect when it is not needed. We evaluated the model using speakers/interlocutor's voice to explore which feature is useful for the prediction, and based on the results, we present examples of SDSs using the model.

II. Related Work

One of the most representative human assessment scores in voice is the mean opinion score (MOS). The annotators of MOS evaluate the quality of voice subjectively and absolutely on a five-point scale ranging from 1 (bad) to 5 (excellent). In the VoiceMOS Challenge [2], a contest to automatically evaluate synthesized speech by building MOS prediction models, there is an application study that utilizes these models to learn only the audio, which is useful for speech synthesis from dark data such as YouTube [3]. Our listening test and RS also aim to achieve high-quality results through automatic evaluation.

There has been some prior work on response time estimation, including an experiment with the users of an automated call system for bus information [4], the development of a estimation model using LSTM [1], and efforts to reduce the effect of speech recognition delays [5], all of which featured models trained by human response timings. Interestingly, it has been reported that some responses replaced by the mode value of the corpus (called *mode*) sound more realistic than actual human responses. In one study [1], the authors defined several filter conditions to extract early responses (called *early*) and late responses (called *late*) from responses across corpus data and then replaced the response times with the mean of the entire *late* if responses were *early* and vice versa (called *opposite*) or *mode*. Then, annotators listened to both the *actual* (called *true* in the paper) and the replaced responses and answered the question "Which response timing sounds like it was produced in a real conversation?". Finally, they performed a statistical test based on the answers to confirm which speakers have SSPs between *actual* and the replaced responses (*opposite* or *mode*) at the $p<0.05$ level.

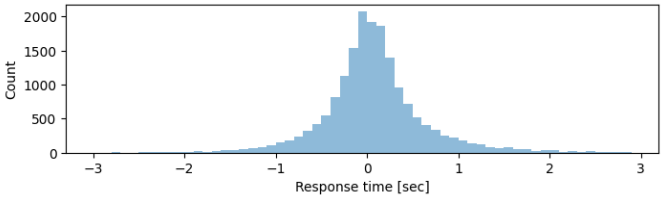


Fig. 2: Number of responses by response time in the target speakers in the corpus of this study.

The test results showed that ten out of 16 speakers of *true* were significantly preferred to *opposite*, while three out of 16 speakers of *actual* and three out of 16 speakers of *mode* were significantly preferred to the comparison. Therefore, in the current study, we explored *true*(*actual*) vs *mode*(*replaced*), which was the controversial result of the previous study.

III. Listening Test

To analyze which response times are considered appropriate for humans, we collected voice data and conducted the listening test. The procedure of the listening test is shown in Fig. 1. We used the mode value of the dataset from the corpus data (0 ms) as a fixed value. Annotators listened to both the actual response (*actual*) and the response whose timing was replaced with the mode value (*replaced*) and then answered the same question as the previous study [1] to vote which one was closer to a real conversation. For analysis and model prediction, we defined the realness score (RS) as the average of the answers of annotators for each response. We also performed a statistical test to confirm which responses humans perceived as a real response timing.

A. Dataset for Listening Test

We used the corpus of Hayashi et al. [6], which includes dyadic interactions in face-to-face conversations in Japanese. A key features of this corpus is the variety of voice samples for each speaker. The speakers to be recorded consisted of groups of four friends, and each speaker in a group recorded one-on-one conversations with three friends and with three other persons the speaker was meeting for the first time (called strangers). The difference between friends and strangers is outside the scope of this study. Each conversation between a speaker and an interlocutor was 1 hour.

To make a dataset of responses, we extracted the audio of eight friends in two groups as speakers, and all (six) interlocutors for each speaker. The distribution of the responses is shown in Fig. 2. We first segmented each conversation audio into voices. These were segmented by silences longer than 200 ms as inter-pausal units (IPUs), following the previous study [1]. We utilized a voice activity detection (VAD) tool https://github.com/wiseman/py-webrtcvad and a low-amplitude

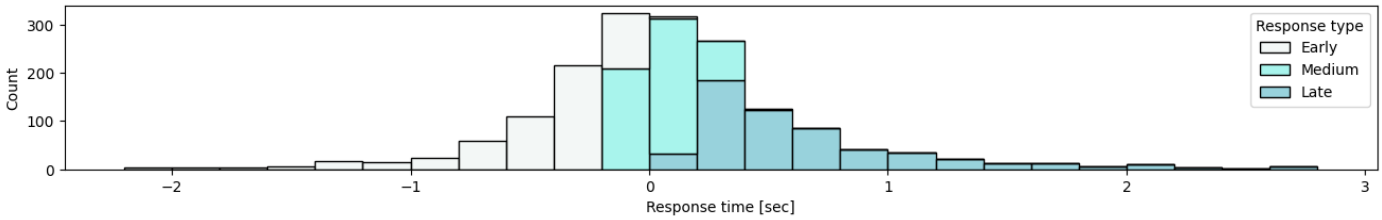


Fig. 3: Number of responses by response time for listening test.

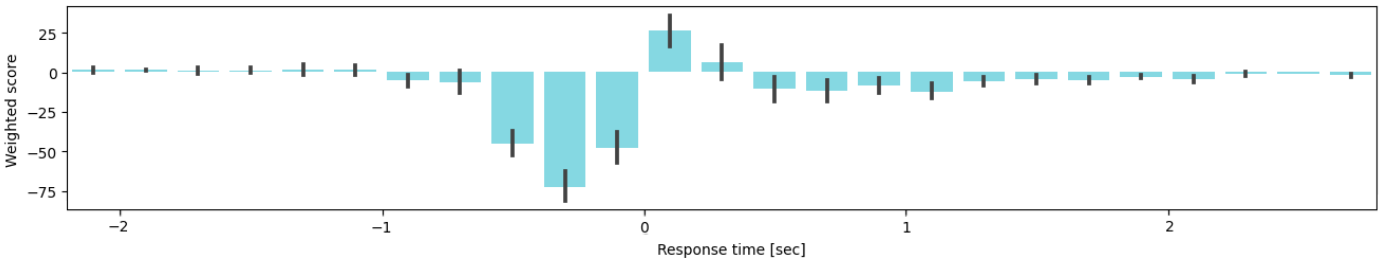


Fig. 4: Distribution of the weighted mean of realness score (RS) by response time multiplied by the number of responses. Vertical line on each bar represents the standard deviation from the mean.

detection script to divide audio into voice segments with 20-ms splits. We then extracted turn-changes from each speaker, defined as when an interlocutor speaks and a speaker responds. We also defined the response as when the speaker's voice is shorter than that of the interlocutor in a turn-change, and excluded those that did not meet this criterion. The response includes not only answers to interlocutors but also backchannels, laughter, reactions, and parts of a speaker's turn.

To make a dataset for the listening test, we extracted 36 turn-changes from the eight speakers and six interlocutors for each speaker. Due to errors during our experimental operations, some turn-changes were excluded. In total, 1,720 turn-changes were targeted in this study, which corresponds to about 10 % of the dataset of responses. The distribution of responses for the listening test is shown in Fig. 3. To extract the 1,720 turn-changes, we set the following filter conditions for this analysis.

Variety of Response Time: To collect a wide variety of response times for each speaker, we divided the responses into three types: *early*, *late*, and *medium*. We calculated response time statistics for each speaker and extracted those below the 30th percentile (as *early*), above the 70th percentile (as *late*), and between *early* and *late* (as *medium*). We collected four responses evenly on the basis of these types. There was a lot of overlap, but most of it occurred without interruption [7].

Extreme Value: We excluded responses later than –20 ms and earlier than 20 ms because we felt there would be almost no difference in such cases even if they were replaced with 0 ms. To avoid extreme values, we also excluded responses below the 0.1 percentile or above the 99.9 percentile.

Unnatural IPU: We excluded turn-changes in which the replacement of response time could be easily perceived by the factors other than voice quality, such as unnatural interruptions in IPUs.

As a result, the distribution of the response times ranged from –2140 ms to 3180 ms. The thresholds between *early* and *medium* for each speaker ranged from –280 ms to 60 ms, and those between *medium* and *late* ranged from 100 ms to 400 ms. The mean of the lengths of the IPUs was 2265 ms for the interlocutors and 908 ms for the speakers.

TABLE I: Statistics of realness score (RS) by response type.

Response type	Mean	SD	Minimum	Maximum
Early	–0.25	0.38	–1.00	0.88
Medium	0.02	0.32	–0.88	0.88
Late	–0.14	0.38	–1.00	0.77
Overall	–0.13	0.37	–1.00	0.88

B. Listening Test Settings

There were 17 annotators for each speaker in the listening test. All annotators were Japanese native speakers. Each used headphones to listen to all turn-changes of two speakers. We shuffled the order of the turn-changes to evaluate each turn-change without time series bias. For each response, annotators listened to both *actual* and *replaced*, and we asked them to answer "Which response timing sounds like it was produced in a real conversation?". We then calculated the RS from the answers, with +1 for answering *actual* and –1 for answering *replaced* and dividing by the number of annotators.

C. Listening Test Results

The results of the listening test are shown in Fig. 4 and Table I. As seen in Fig. 4, the mean of RS was positive from 0 ms to 400 ms and was negative from –1000 ms to 0 ms. From Table I, we found that the maximum RS was 0.88, indicating that there were no responses where all annotators chose *actual*. We also found that the mean RS of *early* and *late* responses was negative, indicating that the annotators preferred *replaced* in the two types.

D. Statistical Test

Since a medium RS contains some randomness, we performed a statistical test on the listening test results to determine which turn-changes had statistically significant preferences (SSPs) between *actual* and *replaced*. Specifically, a chi-square test was conducted to confirm whether there was any bias in the frequency of answers to *actual* or *replaced* for each turn-change at the $p<0.05$ level. The results revealed that there were SSPs in 52 turn-changes for *actual* and in 193 turn-changes for *replaced*. In the turn-changes with SSPs for *actual*, the response time closest to 0 ms was 60 ms, and the furthest was –2140 ms. For *replaced*, the closest was –20 ms, and the furthest was 2800 ms.

The turn-changes that had SSPs are shown in Table II. *Early* responses resulted in the least SSPs for *actual* and the most for *replaced*. The distribution by response time is shown in Fig. 5. In *medium*, all turn-changes that had SSPs for *actual* were positive, i.e. non-overlapping. For all the turn-changes examined in the listening test, in the responses earlier than 0 ms, i.e. overlap, there were fewer SSPs for *actual* than for non-overlap, and there were more SSPs for *replaced*. These results suggest that, in spoken dialogue, humans tend to feel that *mode* (or perhaps from 0 ms to 400 ms) responses are more appropriate than overlap.

TABLE II: Number of turn-changes that have statistically significant preferences (SSPs) between *actual* and *replaced*.

p <0.05	**Response type**			**Total (%)**
	Early	**Medium**	**Late**	
Actual	14	19	19	52 (3.0)
Replaced	107	12	74	193 (11.2)

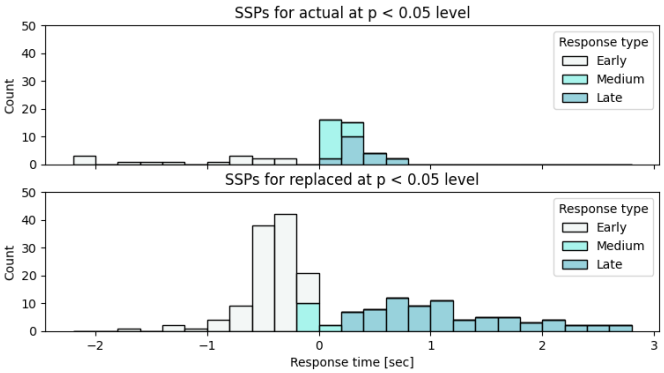


Fig. 5: Number of turn-changes that have SSPs between *actual* and *replaced* by response time.

IV. Realness Score Prediction

We built a model to predict realness score (RS) and explored how well the predicted scores could detect whether response timing estimates are needed or not. As a reference when applying RS to SDSs, we focused on evaluation using only the speaker's/interlocutor's voice. In this study, we define the turn-changes that have SSPs for *actual* as needing response time estimation, and those for *replaced* as not needing it.

A. Model

The VoiceMOS Challenge [8] has shown that self-supervised voice models are useful for predicting the quality of voice. The self-supervised learning framework learns a large amount of audio features, and we used it here as a audio feature extractor. We then implemented a simple model featuring a projection layer added to the self-supervised model as a baseline. For the projection layer, we utilized a 1536-dimensional linear layer that adopted ReLU [9] as the activation function and a linear layer for score output. As the self-supervised model, we used a public HuBERT [10] pretrained with Japanese speech https://huggingface.co/rinna/japanese-hubert-base.

B. Features

Following prior research [8], we input only audio waveforms to the model. We used the audio of the speakers and the interlocutors for a comparison. The number of interlocutors is different from the number of speakers because the speakers consist of friends whereas the interlocutors include both friends and strangers. Therefore, while not a strict comparison, it is sufficient for determining how effective a speaker's features are by using the features of another. We utilized RS normalized from 0 to 1 as the training label.

C. Experimental Settings

We used k-fold cross-validation and measured the out-of-fold prediction scores for each fold. We adopted k = 6 and equalized the number of speakers and interlocutors in each fold. Since we confirmed that strong regularization leads to excessive average results on training data, we set the learning rate to 1e–06 and the weight decay to 1e–05 for model fitting. The model was trained for 70 epochs using Adam [11] without scheduler. The loss function was L1 loss. To enhance reliability, we performed the cross validation with five seeds, and used the average of metrics calculated for each seed as the performance measures.

For evaluation, we defined three evaluation metrics: the recall of SSPs for *actual* (R_a), the recall of SSPs for *replaced* (R_r), and the balanced accuracy (BA). These were calculated as follows:

$$R_a = \frac{T_a}{T_a + F_r} \quad R_r = \frac{T_r}{T_r + F_a} \quad BA = \frac{1}{2}(R_a + R_r) \quad (1)$$

where:

- $T_{a/r}$ is the number of predicted responses that are necessary/unnecessary to estimate timings, which have SSPs for *actual*/*replaced*.
- $F_{a/r}$ is the number of predicted responses that are necessary/unnecessary to estimate timings, which have SSPs for *replaced*/*actual*.

For these evaluation metrics, we set the prediction score threshold as the mean of the score labels.

In addition, we defined a true actual rate (TAR) and a false actual rate (FAR) to evaluate the performance for each prediction score threshold using ROC curve and AUC:

$$TAR = \frac{T_a}{T_a + F_r} \qquad FAR = \frac{F_a}{F_a + T_r} \tag{2}$$

D. Experimental Results

The evaluation results are provided in Table III. R_a shows 0.669 and BA shows 0.618 using the speaker's voice as input. R_r is higher when learning using the interlocutor's voice. This result may be caused by using the features of the interlocutors or by the number of the interlocutors. The ROC curves are shown in Fig. 6. AUC was 0.650 using the speaker's voice, which indicates that the prediction model was more accurate than random selection.

As an ablation study, we also show the prediction results of each response type in Table IV. We found that the prediction of *early* responses had the highest AUC and the biggest difference of predicted RS between *actual* and *mode* of all response type.

V. Discussion and Future Work

In this study, there were some biases in the accuracy of the RS prediction to *early*. Towards the automatic evaluation of response timing, exploring how to predict the other patterns is required. Several aspects remain to be analyzed:

1) Difference between friends and strangers
2) Difference between each speaker
3) Dialogue act analysis (e.g., backchannel, laughter, etc.)
4) 5-point-scale evaluations gathered from annotators

It may be possible to uncover new patterns that can be detected as proper response timings for humans from the results of these analyses. To explore the characteristics of RS, experiments using the fixed values other than 0 ms are also required.

A. Modalities and Contexts

It is well known that the overlap frequency is different between video conferences and face-to-face meetings [7]. Since we conducted the listening test with a limited number of modalities (audio only) and contexts, conducting a listening test with multimodal perception or several contexts might reveal different results or model performances. In addition, there are differences in response times depending on the language [12], research in other languages is needed.

VI. Towards Developing Spoken Dialogue Systems with Appropriate Response Timing

We explored the prediction accuracy of realness score (RS) and found that it may be possible to achieve the automatic assessment of response timing quality using an RS prediction model. This section discusses the challenges in applying the RS model to SDSs. Since the realness of SDS response timing is determined by the SDS user, we need to explore how a user actually feels when interacting with an SDS. We therefore explore some of the topics relating to the development of SDSs using RS prediction models.

TABLE III: Comparison of evaluation metrics when using the mean of score labels as the prediction score threshold.

Audio feature	$\mathbf{R_a}$	$\mathbf{R_r}$	BA	AUC
Speaker's voice	**0.669**	0.567	**0.618**	**0.634**
Interlocutor's voice	0.446	**0.624**	0.535	0.559

Note: The number of interlocutors is more than that of speakers.

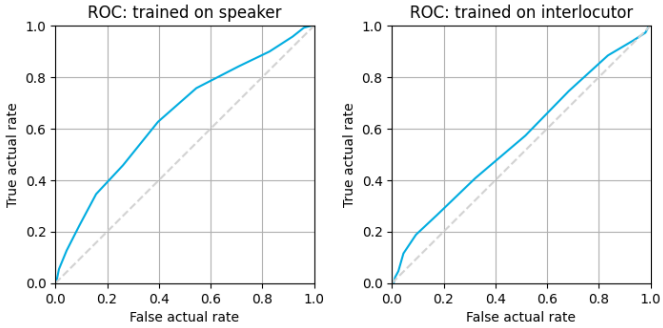


Fig. 6: ROC curves for realness score (RS) prediction model using the voice of speakers and interlocutors.

TABLE IV: Prediction results of each response type inputting the voice of speakers.

Response type	AUC	Mean of predicted RS	
		Actual	Replaced
Early	0.795	0.115	–0.181
Medium	0.540	–0.115	–0.153
Late	0.629	–0.023	–0.126
Overall	0.650	–0.019	–0.159

A. Fixed Value for Timing Replacement

We should first mention the fixed value for the timing replacement. SDSs and VAD tools incur latency when reacting to user utterances. As the fixed value is 0 ms in this study, we cannot utilize the current model directly unless we have prediction models of future conversational events [13]. However, the mode value was just an example of RS measurement and there were high RSs at the response timing from 0 ms to 400 ms; therefore, it would be interesting to evaluate RSs within ranges other than 0 ms.

B. Example of SDSs using RS: Natural Fixed Timing Filler

Under our experimental settings, we can predict RS using only system voices. If we extract system voices in advance, we would be able to predict the voice response timing quality before starting an interaction. In addition, as stated, a high RS response needs a response time estimation, while a low one can be replaced by a fixed value. Therefore, we can extract the responses that can be replaced by a fixed value before starting an interaction as low RS voices using an RS prediction model.

Fig. 7 shows an example of the turn-change using the model. The green zone indicates a low RS voice predicted by the model at a fixed timing, and SDSs can use it as a filler with natural response timing. This will make the quality of response timing stable and increase the time for generating responses.

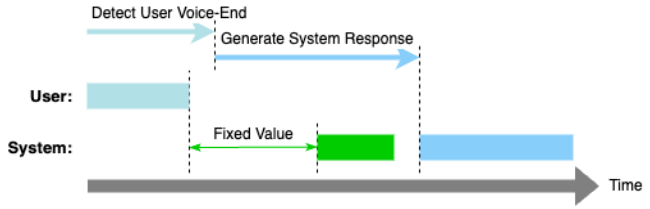


Fig. 7: Example of turn-change of SDSs using realness score (RS) prediction model. The model predicts natural voices at a fixed timing before starting the interaction. In the interaction, SDSs use them as fillers at the fixed timing (green zone).

C. Example of SDSs using RS: Multi-Timing Learning

We can use the model to perform inference during conversations. By predicting the RSs of multiple response timings by the same model, we can predict the most proper fixed timings, including human perceptions. We leave this for future work.

D. VAD Accuracy

Since the end timing detection of IPUs will be different depending on the annotators (including VAD tools), we recommend using the same VAD model not only in the training data but also in the applications. In addition, VAD models must deal with different cases such as phonemes, noises, and languages, so if the RS prediction model uses the interlocutor's features as input, we also have to deal with different cases in the model as well as VAD models. Even if using the system's ones, we need to pay attention to the accuracy of VAD models in tens to hundreds of milliseconds depending on the requirements of the SDS.

VII. Conclusion

Through an experiment on about 1,700 human responses, we clarified the distribution of RS and the possibility of detecting which responses can be replaced with fixed values for appropriate response timing by a simple model. Although there were some patterns that could not be observed due to the filtering process, the voice patterns of the speakers in this study were generally covered. The RS we proposed can also be utilized with previous timing estimation models or with other voice models. We believe the usage of this score will help in the construction of SDSs that are comfortable for humans.

References

[1] M. Roddy and N. Harte, "Neural Generation of Dialogue Response Timings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2020, pp. 2442–2452. DOI: 10.18653/v1/2020.acl-main.221.

[2] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The voicemos challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2023.

[3] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, "Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection," *ArXiv*, vol. abs/2210.14850, 2022.

[4] A. Raux and M. Eskenazi, "Optimizing the turn-taking behavior of task-oriented spoken dialog systems," *ACM Transactions on Speech and Language Processing*, vol. 9, no. 1, 1:1–1:23, 2012, ISSN: 1550-4875. DOI: 10.1145/2168748.2168749.

[5] J. Sakuma, S. Fujie, and T. Kobayashi, "Response Timing Estimation for Spoken Dialog Systems Based on Syntactic Completeness Prediction," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 369–374. DOI: 10.1109/SLT54892.2023.10023458.

[6] T. Hayashi, C. O. Mawalim, R. Ishii, *et al.*, "A Ranking Model for Evaluation of Conversation Partners Based on Rapport Levels," en, *IEEE Access*, vol. 11, pp. 73 024–73 035, 2023, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3287984.

[7] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review," en, *Computer Speech & Language*, vol. 67, p. 101 178, May 2021, ISSN: 08852308. DOI: 10.1016/j.csl.2020.101178.

[8] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, 2021.

[9] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10, Madison, WI, USA: Omnipress, 2010, pp. 807–814, ISBN: 978-1-60558-907-7.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A.-r. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[12] T. Stivers, N. J. Enfield, P. Brown, *et al.*, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, Jun. 2009. DOI: 10.1073/pnas.0903616106.

[13] E. Ekstedt and G. Skantze, "Voice Activity Projection: Self-supervised Learning of Turn-taking Events," in *Proc. Interspeech 2022*, 2022, pp. 5190–5194. DOI: 10.21437/Interspeech.2022-10955.

Enhancing Branchformer with Dynamic Branch Merging Module for Code-Switching Speech Recognition

Hong-Jie Hu*, Yu-Chiao Lai† and Chia-Ping Chen‡
* National Sun Yat-sen University, Taiwan
E-mail: jie12504469@gmail.com
‡ National Sun Yat-sen University, Taiwan
E-mail: cpchen@cse.nsysu.edu.tw

***Abstract*—Branchformer combines the self-attention and convolutional gating multi-layer perceptron blocks as two branches to capture global and local features. To merge both features, Branchformer applies concatenation and weighted average methods. In this work, extending from Branchformer, we propose DBM-Branchformer in a novel way for the combination of both features. DBM-Branchformer employs Dynamic Branch Merging modules (DBMs) to merge both branches. These modules fuse two branches with dynamic weights for a more flexible combination than Branchformer. On TALCS, the code-switching dataset, DBM-Branchformer obtains a word error rate of 6.9%, outperforming the Branchformer with two methods by 0.1% and 0.2%.**

I. Introduction

Nowadays, there are many applications of ASR around us, such as Google Maps, SIRI, and auto-generated subtitles on YouTube The main goal of ASR is to convert audio into corresponding text. Learning from the datasets, the model can perform stochastic probability calculation or neural model decoding to decide the possible sentence with the acoustic features.

In commercial or academic contexts, speakers sometimes resort to code-switching when it is hard to translate some specialized words into their mother tongue. Although monolingual ASR systems are mature for various applications, code-switching ASR systems remain challenging. It is difficult for the model to identify the phonetic boundaries when different languages have similar phonemes. The model may output the wrong words using incorrect language but have similar pronunciation.

End-to-end models have become popular in ASR in recent decades due to their impressive performance, such as Convolutional Neural Networks (CNNs) and Transformers [1]. The main idea of CNNs is to apply convolution on the input sequences with multiple kernels, extracting the local features with several sliding windows. In contrast, the main component of the Transformer is the self-attention layer, which can capture long-range dependencies within the sequence based on global relationships.

Recently, many studies [2]–[7] indicate that fusing self-attention and CNN will gain more improvement. Conformer [2] and Branchformer [3], both Transformer-based systems, use the self-attention layer for global features and convolution layers for local features. Conformer, known as the state-of-the-art structure for ASR tasks, stacks the self-attention and convolution layers in series. It considers the local and global features equally. The static single-branch architecture of Conformer makes it hard to explain and modify the usage of local and global relationships across the encoder layer. To solve these problems, Yifan Peng et al. [3] propose Branchformer that has comparable performance to Conformer. It connects the self-attention and convolution layers in parallel, making the model more flexible in deciding whether to pay more attention to the local or global context.

Shim et al. [8] demonstrated the relationship between phonetic and linguistic localization and explained that phonetic localization can be subdivided into two behavioral modes: one where similar phonemes are assigned high correlations with each other, and another where highly correlated feature frames are transformed into similar ones. They also noted that the diagonality of the attention matrix increases in the upper layers. This suggests that local feature extraction predominantly occurs in the upper layers during speech recognition tasks. Additionally, when merging branches with the weighted average method, it is evident that the upper layers in the Branchformer primarily focus on the local branch [3, Sec. 4.6]. Like Conformer, interleaving self-attention and convolution layers may not be the most efficient approach.

Moreover, according to the findings of Yifan Peng et al. on Aishell-1 [9] and Switchboard [10], we hypothesize that merging weight patterns vary with different languages. The low standard deviation of merging weights indicates that different speeches in the same language exhibit similar merging weight patterns. In other words, the merging behavior remains consistent even when the weight is computed based on the input speeches. Furthermore, when the merging weight is low, computational costs are still incurred on the branches that are ultimately ignored.

In code-switching speeches, since different languages exhibit unique distributions of local and global context, the merging weight pattern varies across encoder layers and changes

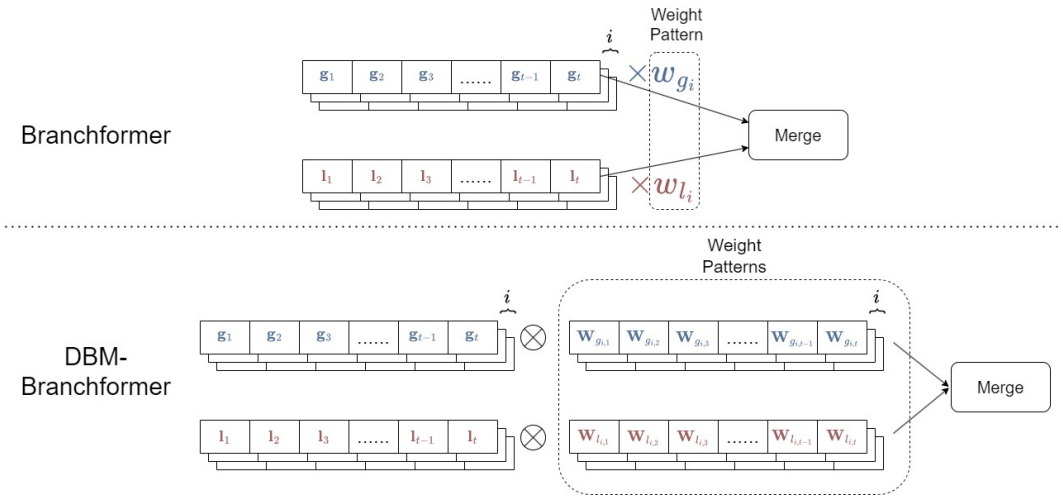


Fig. 1. The concept of dynamic branch merging. The upper part of the figure shows the branch merging method in Branchformer, called "Weighted Average," while the lower part depicts the method we proposed in this study. $g_i, l_i \in \mathbb{R}^{1\times d}$ are the outputs from the global and local branches, t denotes the frame number of the hidden vectors, and d is the feature dimension of the hidden vectors. i indicates the block number of the encoder. The weight pattern represents the merging weight distribution of the encoder. In Branchformer, each weight is calculated from the entire vector sequence from both branches, resulting in a consistent weight pattern over the time domain. In DBM-Branchformer, since the merging method weights are generated separately, each input frame has a unique weight pattern.

with language shifts. Therefore, if a model could adaptively switch weight patterns, it might improve parameter efficiency and performance for code-switching speeches, rather than using a consistent merging method. We believe that a flexible structure capable of adjusting the importance of local and global features is essential for optimal code-switching ASR systems.

In this work, we present a new method that is more flexible and dynamic in tuning the importance of local and global features, specifically aimed at code-switching ASR tasks, as shown in Fig. 1. Extended from Branchformer, we propose Dynamic Branch Merging-enhanced Branchformer (DBM-Branchformer), which employs the Dynamic Branch Merging module (DBM) to decide the weights of the local and global branches. The DBM evaluates the weights of the local and global features by considering the hidden vectors from the previous layer and the features extracted from the current layer. Similar to the learned weight method in Branchformer, which calculates the merging weight with the input vectors, DBM-Branchformer dynamically adjusts the weights for each frame of the hidden feature instead of applying uniform weights across the whole hidden sequence.

In Section II, we will introduce the methods used in this work. The details of the experiment will be mentioned in Section III. In Section IV, we will analyze and discuss the result. Finally, the conclusion will be made in Section V.

II. Method

This section takes a deep dive into the core structure of Branchformer and our merge method. We extend the Branchformer with DBM to calculate the weight of the output from local and global branches for the merging process.

A. Local Branch

Branchformer employs convolutional gating multi-layer perceptron (cgMLP) [11] to extract local features. cgMLP deploys a Convolutional Spatial Gating Unit (CSGU) [11] to get the local information. After projecting to the higher dimension, the input is split into two parts, one is the residual part $\mathbf{X}_r$, and the other one is the gate part $\mathbf{X}_g$.

After layer normalization, $\mathbf{X}_g$ will be depth-wise convoluted. Then, use the linear gating by directly element-wise multiplication with $\mathbf{X}_r$ and $\mathbf{X}_g$ without activation function. Finally, use a linear layer to project the latent vectors back to the d dimension. The processes of the local branch are written as follows,

$$[\mathbf{X}_r, \mathbf{X}_g] = \text{GELU}(\mathbf{X}\mathbf{U}) \tag{1}$$

$$\begin{aligned}\mathbf{Z} &= \text{Dropout}(\text{CSGU}(\mathbf{X}_g)) \\ &= \text{Dropout}(\mathbf{X}_r \otimes \text{DWConv}(\text{LN}(\mathbf{X}_g))\end{aligned} \tag{2}$$

$$\mathbf{L} = \mathbf{Z}\mathbf{V} \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{T\times d}$, $\mathbf{X}_r, \mathbf{X}_g \in \mathbb{R}^{T\times d_{\text{latent}}}$ are latent vectors, and $\mathbf{U} \in \mathbb{R}^{d\times 2d_{\text{latent}}}$, $\mathbf{V} \in \mathbb{R}^{d_{\text{latent}}\times d}$ are the weights of two projections before and after the CSGU. d_{latent} is the latent dimension in both projections.

B. Global Branch

To extract the global feature, Branchformer uses multi-head self-attention layers with relative position encoding [12]. The main idea of self-attention is to project the input to matrices $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{T\times d}$, calculating the relation score between $\mathbf{Q}$ and $\mathbf{K}$ for each frame, forming a square matrix $\mathbf{A}_{\text{h}}$. $\mathbf{A}_{\text{h}}$

contains the global information which shows the relationship between each frame. Then, multiply $\mathbf{V}$ with the attention matrix $\mathbf{A_h}$. As for multi-head self-attention, it splits the self-attention process into h groups. h is the number of heads in each self-attention layer, and each head calculates the relation score independently. The multi-head self-attention formula is shown as below,

$$\text{Score}_h = \frac{(\mathbf{Q}_h + \mathbf{u}_h)\mathbf{K}_h^\top}{\sqrt[2]{d/h}} + \frac{\text{RelShift}((\mathbf{Q}_h + \mathbf{v}_h)\mathbf{P}_h^\top)}{\sqrt[2]{d/h}} \quad (4)$$

$$\mathbf{A}_h = \text{Softmax}(\text{Score}_h) \quad (5)$$

$$\mathbf{G} = \text{Dropout}(\mathbf{A}_h)\mathbf{V}_h \quad (6)$$

where $\mathbf{Q}_h$, $\mathbf{K}_h$, and $\mathbf{V}_h \in \mathbb{R}^{h\times T\times d/h}$ are the projection that split into h groups, $\mathbf{u}_h$ and $\mathbf{v}_h \in \mathbb{R}^{h\times d/h}$ are the trainable bias, $\mathbf{P}_h^\top \in \mathbb{R}^{h\times(2T-1)\times d/h}$ is relative positional embedding. After the multiplication with $\mathbf{P}_h^\top$ and $\mathbf{Q}_h$, the relative shifting is applied. d_k/h represents the level of multi-head, used to normalize the feature to avoid the condition that a huge value occupies the output of Softmax, making others become zero.

C. Merge method

Branchformer applies two types of merge methods. One is the concatenation, which combines $\mathbf{L}$ and $\mathbf{G}$, then directly projects the combination to the original size. The process is shown below,

$$\mathbf{Y}_{\text{merge}} = \text{Concat}(\mathbf{L}, \mathbf{G})\mathbf{W} \quad (7)$$

where $\mathbf{W} \in \mathbb{R}^{2d\times d}$ is the trainable matrix, $\mathbf{L}$ and $\mathbf{G}$ are the output of local and global branches.

The other method is the weighted average. The weighted average method calculates the importance score on local and global branches with attention-based pooling. Attention-based pooling squeezes the hidden vectors into a single dimension, then applies softmax on the time domain to get the score. After multiplying the score and the output from both branches, we can get the pooling feature with the d dimension. After projecting the pooling feature to a single dimension, we can get the weight on local or global branches. The formula is shown as follows,

$$\text{Score}_G = \text{Softmax}(\text{PoolProj}_G(\mathbf{G})^\top) \quad (8)$$

$$w_G = \text{Score}_G\mathbf{G}\mathbf{U}_G \quad (9)$$

$$\text{Score}_L = \text{Softmax}(\text{PoolProj}_L(\mathbf{L})^\top) \quad (10)$$

$$w_L = \text{Score}_L\mathbf{L}\mathbf{U}_L \quad (11)$$

$$w'_G, w'_L = \text{Softmax}(w_G, w_L) \quad (12)$$

$$\mathbf{Y}_{\text{merge}} = w'_G\mathbf{G} + w'_L\mathbf{L} \quad (13)$$

where the learnable weight in $\text{PoolProj}_G, \text{PoolProj}_G \in \mathbb{R}^{d\times 1}$, $\mathbf{U}_G, \mathbf{U}_L \in \mathbb{R}^{d\times 1}$ are matrix to project the pooling features to one dimension. L and G are the output from local and global branches. The weights for the local and global branches are denoted by w'_G and w'_L, respectively, where $w'_G, w'_L \in \mathbb{R}$.

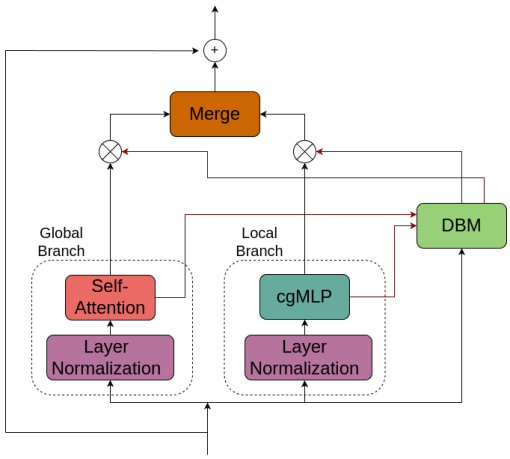


Fig. 2. DBM-Branchformer encoder block architecture. Instead of evaluating $w_G, w_L \in \mathbb{R}$ with the whole hidden sequence which is the weighted average method used in Branchformer, using the DBM module to calculate the $W \in \mathbb{R}^{T\times 2d}$, each element in the hidden vectors not only has the own weight individually but also multiplies with the dynamic weights depending on the input.

Although merging with the concatenation method has proven that it can perform better than the weighted average on monolingual datasets, it can not show the importance of local or global features in each layer.

D. Dynamic Branch Merging

One of the challenges in code-switching speech recognition is that the model tends to output words with similar pronunciations but in the wrong language. According to the research by Shim et al., we speculate that one of the reasons for this issue is the lack of sufficient data. Different languages often have unique distributions of local and global features, meaning that these feature distributions in code-switching speech are not fixed but change over time. However, when a model processes all acoustic features with fixed weights equally, it can easily interpret similar-sounding acoustic features from different languages as the same parts.

In the lower layers, due to insufficient data, these similar-sounding feature frames tend to have high correlations among themselves without generating sufficient correlations with adjacent blocks. Consequently, in the upper layers, these features gradually assimilate, leading to errors in judgment.

Therefore, when using the concatenation method, since the projection merely merges two branches with static weights W, the same projection is applied despite language changes. Similarly, using the weighted average method to merge branches might not be the best choice either. Although w_G and w_L in Section II-C are calculated based on the outputs from the two branches, this calculation is done based on the entire sequence, not on each frame individually. Hence, even if the language switches, the proportion of the two branches remains unchanged.

To enable the model to adjust its recognition strategy according to input variations, and to make the branch compu-

tations more efficient and flexible, preventing some branches from remaining unused, we propose dynamic branch merging (DBM).

DBM-Branchformer, shown as Fig. 2, applies the DBM module deciding the weights to merge both branches. The DBM module determines the attention each frame in the current layer should pay to global and local branches based on the output of each frame from the previous layer.

The process of the DBM module is shown as follows,

$$\mathbf{Y}_{\text{cat}} = \text{Concat}(\mathbf{X}, \mathbf{L} \otimes \mathbf{G}) \tag{14}$$

$$\mathbf{W} = \text{GELU}(\mathbf{Y}_{cat}\mathbf{D}) \tag{15}$$

$$\mathbf{Y}_{\text{merge}} = \mathbf{W} \otimes \text{Concat}(\mathbf{L}, \mathbf{G}) \tag{16}$$

where $\mathbf{W} \in \mathbb{R}^{T \times 2d}$ is the weight for merging the local and global branches, $\mathbf{D} \in \mathbb{R}^{2d \times 2d}$ is the learnable weight. Since the weight $\mathbf{W}$ varies with $\mathbf{X}$, $\mathbf{L}$, and $\mathbf{G}$, the DBM module merges both branches, depending on each hidden vector instead of the static value or the score of each branch.

To achieve flexible branch merging and allow the model to skip certain branches, we use GELU for calculating the merging weights instead of Softmax. This prevents the DBM module from being forced to choose between two branches when it considers both branch outputs unsuitable, thereby increasing training stability.

To make the DBM module consider not only the output of the previous layer but also the position information of local and global features extracted in the current layer, $\mathbf{Y}_{\text{cat}}$ is composed by $\mathbf{X}$ and the tensor product of the outputs from both branches. We found that this approach makes the generated weights $\mathbf{W}$ more dynamic and improves recognition performance on code-switching speeches.

III. Experiment

This section introduces the experiment method, such as parameter setting, dataset, and data preprocessing.

A. Dataset

In this work, TALCS [13] is used to evaluate the performance which has about 587 hours of Mandarin-English code-switching speech audio, provided by TAL Education Group. All of the audio in the dataset are sampled at 16 kHz.

B. Data Preprocessing and Augmentation

For data augmentation, we apply speed perturbation with the ratios 0.9, 1.0, and 1.1 [14]. We also utilize SpecAug [15] with 5 time-warp-windows, 2 frequency-masks which range in [0,27], and 10 time-masks which range in [0,0.05].

To lower the dictionary size, we apply the byte encode module provided by Icefall [5] to encode the Chinese characters. For example, "你好 ITS'S OKAY 的" is converted to "ţN̩ ţ ITS'S OKAY L'ĥ". With the byte encode module, the dictionary size decreases to 1,000.

C. Experiment setting

We use NVIDIA RTX A4500 with 20GB memory for the training process. In this study, we utilize the ESPnet toolkit [16] for the training process.

The ASR model we use consists of three main parts: an encoder, a decoder, and a CTC (Connectionist Temporal Classification) module [17].

We present the total loss function of the entire model below,

$$Loss_{\text{asr}} = (1-\lambda)Loss_{\text{dec}} + \lambda Loss_{\text{ctc}} \tag{17}$$

$Loss_{\text{asr}}$ is the total loss of the entire ASR model. $Loss_{\text{dec}}$ is the loss from the decoder. $Loss_{\text{ctc}}$ is the loss which CTC accounting for. λ is the hyperparameter to adjust the influence from the decoder and the CTC model. In this work, we don't use a language model to analyze the model performance. In the training process, we set λ to 0.3.

During training, we set the label smoothing weight to 0.1 and use Adam [18] as the optimizer with weight decay 1e-6.

1) Encoder Setup: In Section II-B, it is mentioned that multi-head self-attention is used in the calculation of the global branch. In this study, we set the number of heads h to 8 and the feature dimension d to 512.

We set the number of encoder layers for the DBM-Branchformer to 18, and to ensure a fair comparison with a similar number of parameters, we set the number of encoder layers for the Branchformer to 22.

The cgMLP in Branchformer and DBM-Branchformer has kernel size 31 and latent dimension d_{latent} 3072.

To speed up the process and decrease memory usage in self-attention, we apply the subsampling module in the front of the encoder. The subsampling module consists of two 2d-convolution layers, each with kernel size (3x3) and stride (2x2) to downsample the input sequence by a factor of 4.

2) Decoder Setup: In this study, the Transformer decoder is employed in all cases, maintaining the same dimension d as the encoder. The decoder would calculate the relationship between the acoustic features by the multi-head cross-attention and between the previous output texts by the multi-head causal self-attention. Then, similar to the CTC module, outputs the most likely element as a vector from the dictionary.

The causal self-attention is similar to the original self-attention method but has a multiplication of an upper triangular matrix. This modification aims to eliminate the relationship between the hidden vectors and their successors.

In this study, all the models used for comparison employed the same decoder architecture. We set the number of heads to 8, the feature dimension to 512, matching the encoder, and the number of layers to 6.

In the inference stage, we use "Rescoring" as the decoding method [19]. First, it uses beam search decoding to take the top n_{bs} list of hypotheses from the decoder's output. Second, it rescores the hypotheses with the log probabilities of the output from CTC and the decoder. We set n_{bs} to 10, and CTC weight to 0.4.

IV. Results

In this chapter, we compare the performance differences between DBM-Branchformer and Branchformer in code-switching tasks to demonstrate the impact of the dynamic branch merging strategy on the recognition ability of code-switched speeches. Additionally, in Section II-C, we mentioned that the concatenation method of Branchformer performs better than the weighted average method on monolingual datasets. We are also curious whether their performance on code-switching datasets will be consistent with the results on monolingual datasets, so we also include them in the comparison.

A. Main Result

Table I compares the result of both Branchformer and DBM-Branchformer oh TALCS. Both methods in Branchformer are trained for 60 epochs, while the DBM-Branchformer goes through 49 epochs of training. The DBM-Branchformer outperforms the Branchformer with the concatenation merging method in WER by 0.1%. It also outperforms the weighted average method by 0.2%.

B. Analysis of Merging Weight

In the TALCS test set, the average weight distribution for each layer when Branchformer uses weighted averaging to merge branches is shown in the Fig. 3. According to our measurements, the highest variance of branch merging weights across layers on the TALCS test set is in the first layer, at approximately 0.0012. This indicates that even in the case of code-switching, the branch merging weights calculated by the weighted average method remain nearly static, similar to the monolingual result, despite these values being based on the branches' outputs at each layer. We believe this is because the pooling computation in the weighted average method overlooks the dynamic of the features, leading the merging modules to disregard the uniqueness of different inputs.

In contrast, DBM-Branchformer shows a minimum variance of 0.0057 in the merging weights across layers on the TALCS test set. This suggests that DBM-Branchformer can handle global and local dependencies more flexibly than Branchformer.

The variance distribution of the merge weights of each layer is shown in Fig. 4. It can be seen that both DBM-Branchformer and Branchformer have higher variance at the first layer. We believe that this is because the encoder in the bottom layer focuses on phonetic localization, which is more variable than linguistic localization.

Level	Local	Global
22	0.9432	0.0568
21	0.9282	0.0718
20	0.9397	0.0603
19	0.9553	0.0447
18	0.9587	0.0413
17	0.1904	0.8096
16	0.1062	0.8938
15	0.1086	0.8914
14	0.1135	0.8865
13	0.0091	0.9909
12	0.1119	0.8881
11	0.1113	0.8887
10	0.0816	0.9184
9	0.0903	0.9097
8	0.0753	0.9247
7	0.0977	0.9023
6	0.9377	0.0623
5	0.1000	0.9000
4	0.9458	0.0542
3	0.0897	0.9103
2	0.9582	0.0418
1	0.0012	0.9988

Fig. 3. The weight pattern of Branchformer with the weighted average method on TALCS Test set. When training on code-switching datasets, we can observe that the branch merging strategy in the lower layers of Branchformer can be roughly divided into three parts. The lower part alternates between using global and local branches. The middle part tends to use the self-attention layer to capture global features. The upper part focuses on using cgMLP to process local information.

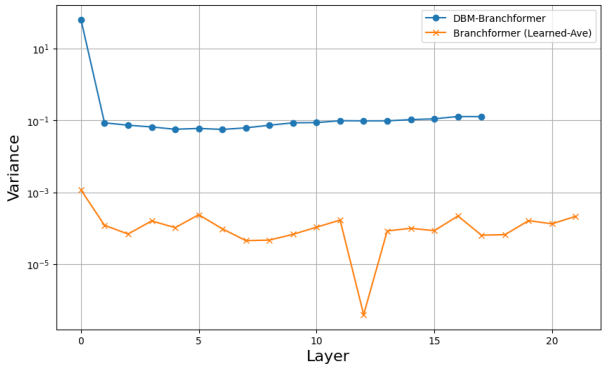


Fig. 4. The distribution of the variances of the merge weights of each layer in DBM-Branchformer and Branchformer.The variances of the merge weights of each layer in DBM-Branchformer are higher than those of Branchformer.

V. Conclusions

In this study, we propose a different method for merging the local and global branches in Branchformer. Aiming at code-switching auto speech recognition, extending from the merging methods used in Branchformer, DBM-Branchformer uses a Dynamic Branch Merging module to calculate the importance of each element in the hidden vectors. Instead of considering the whole input to weigh the local and the global branches, the DBM module calculates the weights independently to adapt the code-switching speech. By incorporating the DBM module, the DBM-Branchformer outperforms the Branchformer in TALCS without using a language model.

TABLE I
COMPARISON OF BRANCHFORMER AND DBM-BRANCHFORMER IN THE TALCS CODE-SWITCHING DATASET. WE ANALYZE THE PERFORMANCE WITH WORD ERROR RATE (WER) AND APPLY THE BYTE ENCODE MODULE.

Model Type	Params(M)	Test set
Branchformer		
With Concat	127.76	7.0
With Learned Ave	122.04	7.1
DBM-Branchformer	129.64	**6.9**

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[3] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding," in *International Conference on Machine Learning*, PMLR, 2022, pp. 17 627–17 643.

[4] K. Kim, F. Wu, Y. Peng, *et al.*, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 84–91.

[5] Z. Yao, L. Guo, X. Yang, *et al.*, "Zipformer: A faster and better encoder for automatic speech recognition," *arXiv preprint arXiv:2310.11230*, 2023.

[6] J. Li, Z. Duan, S. Li, X. Yu, and G. Yang, "Gncformer enhanced self-attention for automatic speech recognition," *arXiv preprint arXiv:2305.12755*, 2023.

[7] A. Andrusenko, R. Nasretdinov, and A. Romanenko, "Uconv-conformer: High reduction of input sequence length for end-to-end speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[8] K. Shim, J. Choi, and W. Sung, "Understanding the role of self attention for efficient speech recognition," in *International Conference on Learning Representations*, 2021.

[9] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, IEEE, 2017, pp. 1–5.

[10] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, speech, and signal processing, ieee international conference on*, IEEE Computer Society, vol. 1, 1992, pp. 517–520.

[11] J. Sakuma, T. Komatsu, and R. Scheibler, "Mlp-based architecture with variable length input for automatic speech recognition," 2021.

[12] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[13] C. Li, S. Deng, Y. Wang, *et al.*, "Talcs: An open-source mandarin-english code-switching corpus and a speech recognition baseline," *arXiv preprint arXiv:2206.13135*, 2022.

[14] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[15] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[16] S. Watanabe, T. Hori, S. Karita, *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[17] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 4835–4839.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] T. Hori, S. Watanabe, and J. R. Hershey, "Joint ctc/attention decoding for end-to-end speech recognition," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 518–529.

Enhancing Neural Speech Embeddings for Generative Speech Models

Doyeon Kim * and Yanjue Song† and Nilesh Madhu† and Hong-Goo Kang*
* Dept. of Electrical and Electronic Eng., Yonsei University, Seoul, Korea
E-mail: ehyeon24@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr
† IDLab, Ghent University - imec, Ghent, Belgium
E-mail: Yanjue.Song, Nilesh.Madhu@ugent.be

***Abstract*—We explore a speech enhancement framework where neural speech embeddings, derived from pre-trained self-supervised learning (SSL) models applied to noisy signals, are used as inputs to a neural vocoder to generate the corresponding clean speech. The primary innovation lies in enhancing these latent neural embeddings to mitigate distortions caused by noise and reverberation, resulting in a superior quality of the synthesized signal. By dividing the process into Separate phases for embedding enhancement and speech generation, the approach allows for greater flexibility in network design. We also examine the advantage of integrating hidden states from the SSL model in a learnable manner to create a more robust embedding for the vocoder input. Additionally, we investigate various loss functions for training the neural vocoder. Experimental results confirm the effectiveness of our proposed approach, particularly in environments with simultaneous background noise and reverberation.**

***Index Terms*—embedding enhancement, speech embeddings, generative speech enhancement, automatic speech recognition, self-supervised learning**

I. Introduction

The limited availability of large labeled and paired datasets in deep learning for speech has led to a surge in the development of self-supervised learning (SSL) models [1]–[3], which are designed to extract meaningful representations from unlabeled data. Pre-trained SSL models have shown their potential as feature extractors across various tasks, allowing a small back-end network to generalize effectively with minimal training for different downstream applications, as reported in the SUPERB benchmark [4].

Several prior studies have explored the use of SSL models for speech enhancement. For instance, Zhao et al. [5] conducted a denoising task by initializing encoder and bottleneck parameters with a pre-trained WavLM model. In [6], [7], SSL features were fine-tuned specifically for speech enhancement enabling the estimation of masks from noisy SSL embeddings to recover clean speech. In general, SSL models typically transform speech input into low-dimensional embedding features. Neural vocoders, commonly used in text-to-speech (TTS) systems, generate synthesized speech waveforms from low-dimensional speech features such as log-mel spectrograms [8]–[10]. To leverage the benefits of pre-trained SSL models for generating compact representations and the neural vocoder's capabilities in speech generation, the Denoising Vocoder [11] was proposed. This approach directly synthesizes clean speech from the distorted embeddings extracted by the pre-trained HuBERT model [3] from noisy input signals. The vocoder utilizes the HiFi-GAN architecture [12] to produce a time-domain signal through a series of upsampling blocks from the embedding features.

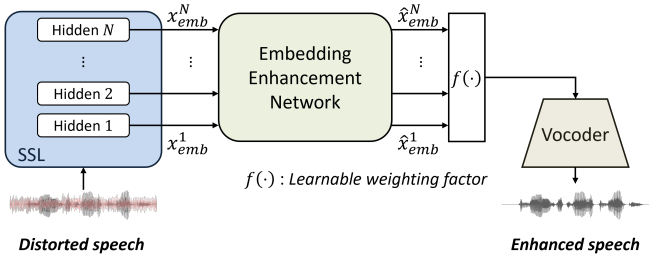


Fig. 1: Proposed framework overview

A recent study [13] explored the semantic information encoded by SSL models by tokenizing and de-tokenizing wav2vec2.0 embeddings. The HiFi-GAN vocoder was then used to synthesize the clean speech estimate from these processed embeddings. Miipher [14] also utilized SSL feature information for speech synthesis, but this model generates speech by conditioning enhanced features on a combination of the ground truth transcript and the SSL embeddings. In this paper, we investigate the model's performance when only the noisy speech signal is available.

Embeddings extracted from noisy inputs are likely to be distorted, resulting in lower-quality synthesized speech. To address this, we propose dividing the process into two stages: a) enhancing the distorted embeddings and b) synthesising clean speech from these improved embeddings. We aim to optimize the respective networks for each stage individually. The *embedding enhancement* network transforms SSL features extracted from distorted signals into features that resemble those from clean speech. Since the network preserves the dimension of the pre-trained SSL features, these enhanced features can seamlessly replace the SSL embeddings from noisy inputs in downstream tasks, thereby straightforwardly improving the system's robustness to noise.

To achieve high-quality speech, we investigate methods to further optimize the HiFi-GAN vocoder within the context of speech enhancement (SE). We begin by comparing the baseline mel-spectrogram loss with alternative loss functions commonly

used in speech enhancement, such as signal-to-distortion ratio (SDR) [15], compressed linear spectrogram [16], and Huber loss [17]. Regarding the *input* to the vocoder, it can theoretically be the latent representation from each hidden state of the SSL model. Thus, we analyze the importance of these latent representations at each hidden state and investigate the impact of a learnable weighted combination of all states. This approach offers deeper insights into the information from each layer that is pertinent to the task of speech enhancement.

The remainder of the paper is organized as follows. Section II describes the selection of pre-trained SSL models based on environmental distortions. In Section III, We outline the proposed framework and the optimization of the vocoder for speech enhancement. Section IV demonstrates the effectiveness of the approach and related metrics, while Section V provides the conclusions drawn from our study.

II. Selection of the pre-trained SSL model

Given the growing application of SSL models for generative speech enhancement, our recent study [18] analyzed these neural embeddings to gain insights into the information they contain. Specifically, we quantified the preservation of speaker and phonetic information, as well as the robustness of embedding features to noise and reverberation, extracted from various SSL speech models, including TERA [19], wav2vec2.0 [1], wavLM [2], and HuBERT [3]. The analysis revealed that HuBERT and wavLM preserved more phonetic information compared to TERA and wav2vec2.0, while TERA demonstrated the highest preservation of speaker information among the models compared. For assessing robustness, we considered two types of input distortions: additive background noise and reverberation, achieved by convolving with room impulse responses. The benchmark employed instrumental metrics, including cosine similarity and normalized mean square error (MSE), to compare embeddings extracted from clean speech with those from the corresponding distorted versions. The main finding was that additive noise caused more significant distortion to the embedding features than reverberation, with the distortion worsening at lower signal-to-noise ratios (SNRs). Among the four pre-trained models, TERA generated the most robust embedding features across all conditions. In the context of speech enhancement, where the input speech is subject to similar distortions, this robustness offers a significant advantage, making TERA a strong candidate as the SSL feature extractor. This choice is further supported by our results for the speech enhancement task, as described in Section IV-D.

III. Methodology

Figure 1 provides a schematic overview of the proposed method. First, we segment the signal and extract pre-trained SSL feature embeddings x_{emb}^{N}, where N denotes the number of layers, from the distorted speech data x for each frame. For simplicity, N will be omitted in the subsequent sections. The embedding enhancement network is then employed to transform the distorted features into their clean counterparts. Subsequently, a learnable weighted sum function within the $f(\cdot)$ layer combines the enhanced representations from the various hidden layers of the SSL model into a single refined feature per frame. These refined features, denoted as $\hat{x}_{emb}$, are then fed into the vocoder network to produce the enhanced speech signal.

A. Embedding Enhancement Network

The embedding enhancement network comprises four convolutional gated recurrent unit (ConvGRU) blocks, with each block featuring two bidirectional GRU layers and one convolution layer. The bidirectional GRU layers capture long-term temporal dependencies, while the convolutional layer adjusts the input dimension to match the feature dimension required for the vocoder input. We train the network by minimizing the mean squared error (MSE) between the enhanced embeddings and the clean embeddings, as follows:

$$L_{emb} = \frac{1}{B}\sum_{i=0}^{B}(\hat{x}_{emb} - y_{emb})^2, \tag{1}$$

where $\hat{x}_{emb}$ and y_{emb} represent the feature embeddings from the enhanced and clean speech signals, and B denotes the batch size.

In the case of using the last hidden feature, the loss is computed using only this final hidden feature. For fixed summation and learnable weighted summation, the loss values are averaged across all layers.

Note that the embedding enhancement network is trained in a *task-agnostic* manner to maximize flexibility, enabling its application to various downstream tasks without the need for re-training. Thus, the network directly estimates the clean SSL features. To enable fine-tuning for specific tasks, a learnable weighting factor is introduced and trained according to the requirements of the downstream task.

B. Vocoder

We utilize the HiFi-GAN [12] neural vocoder to synthesize speech from the SSL feature embeddings. Unless stated otherwise, the vocoder architecture adheres to the original HiFi-GAN design. We adjust the vocoder parameters to match the window size and frame shift of the pre-trained SSL model, as described further in Section IV. Additionally, we align the dimension of SSL feature embeddings with the input dimension of the vocoder through a one-dimensional matching layer and adapt the upsampling ratio for each block in the neural vocoder.

We use the original adversarial loss as proposed in [12] and optimize the generator loss specifically for the speech enhancement task. In addition to the original **L1 Mel-spectrogram loss**, we also investigate other loss functions commonly used in speech enhancement, including the **SDR loss**:

$$L_{\text{SDR}} = 10\log_{10}\frac{\|y\|^2}{\|\hat{y}-y\|^2}, \tag{2}$$

where y is the target signal and $\hat{y}$ is the synthesized estimate; the **compressed spectrogram loss**:

$$L_{\text{spec}} = \||\hat{Y}|^c - |Y|^c\|_H, \tag{3}$$

where $\hat{Y}$ is the synthesized speech magnitude spectrogram, Y is the clean speech magnitude spectrogram, and c is a compression factor; and the Huber loss [17], $\|\cdot\|_H$, on the compressed spectrogram, which calculates L1 loss when the error values exceed a predefined threshold and switches to mean squared error (MSE) loss below this threshold, thereby combining the advantages of both loss functions. Lastly, we investigate the **Time-domain Huber loss** applied to the synthesized and clean speech in the time domain to determine the optimal loss function.

Each hidden layer of the SSL model provides latent speech representations, with each layer containing valuable information about the underlying speech that can be utilized by the vocoder. Consequently, the method of combining embeddings is crucial, as the speech generation process involves upsampling this input. We investigate three approaches: (1) utilizing a single, optimal, hidden feature; (2) employing a fixed, equally weighted summation of all embedding layers; and (3) incorporating a learnable weighted summation of all embedding layers. These approaches are tested with different SSL models (see Section IV), with a distinct set of weights independently learned for each model.

IV. Experiments

A. Experimental setup

The vocoder and the proposed embedding network are trained using the Deep Noise Suppression (DNS) Challenge 2021 training set [20]. This dataset includes clean speech samples from the Librivox dataset [21], which features recordings from 10,000 audiobooks in various languages. Additionally, noise samples are sourced from Audioset [22], with supplementary augmentation from noise datasets such as Freesound [23] and DEMAND [24]. For our experiments, we utilized the English language subset of the aforementioned datasets for model training. To simulate reverberation, we employed the openSLR26 and openSLR28 datasets [25]. We adhered to the SNR, reverberation time (RT60), and other hyperparameters specified in the original DNS challenge paper [26]. For the comparison of model performance and analysis, the test set included both background noise and reverberation. Background noise was sourced from the Noisex-92 [27] dataset, with SNR levels set at (-7, 0, 5, 10, 15) dB. We use the MIT Impulse Response Survey dataset [28], which contains recorded room impulse responses, to generate reverberation. For analysis, we utilized three distinct datasets: a noisy test set (*noise*), a reverberant test set (*reverb*), and a combined test set (*noise+reverb*).

The proposed system is evaluated against two baselines. The first baseline, *Clean Vocoder*, is trained to synthesize clean speech signals from embeddings extracted from *clean* data. This naïve neural vocoder serves as an *upper* bound for the proposed system's performance, representing the best available speech quality when the embedding enhancement network perfectly maps distorted embeddings to clean speech embeddings. The second baseline is the Denoising Vocoder [11], which synthesizes clean speech from distorted embeddings derived from noisy input signals.

B. System and training details

We employ the AdamW optimizer [29] with an initial learning rate of 1e-4 and momentum parameters (β_1, β_2) set to (0.8, 0.99). Exponential learning rate scheduling is applied with a decay rate of 0.999. The parameters of the HiFi-GAN vocoder were adjusted to match the input specifications of the SSL model. For example, to align with the pre-trained TERA model and accommodate 16 kHz signals, the DFT size, window size, and hop size in HiFi-GAN are set to 400, 400, and 160 samples, respectively. After the feature matching layer, the channel dimension of the feature embeddings is reduced from 768 to 512, consistent with the original HiFi-GAN vocoder. This reduction allows us to align the SSL model with the baseline models for comparison. For loss and input analysis, we first reduce the input embeddings to 80 dimensions to match the dimension of the original vocoder input, which uses 80-dimensional mel spectrograms. Subsequently, the channel is upsampled using ratios of (8, 5, 2, 2) and kernel sizes of (16, 15, 4, 4). The spectrogram compression factor c in Equation (3) is set to 0.6, as recommended in [16], and the threshold value for the Huber loss is set to 1.0.

C. Metrics

Performance is evaluated using non-intrusive neural network-based metrics, specifically DNSMOS [20] and NISQAv2 [30], as intrusive metrics are less suitable for assessing generative models [31]. DNSMOS is a robust instrumental metric utilized to evaluate model performance in the DNS Challenge. It involves training a combination of convolutional and fully connected layers to estimate perceptual mean opinion score (MOS) results. DNSMOS provides instrumental MOS scores for overall quality (**OVRL**), signal quality (**SIG**), and background noise (**BAK**) of the input signal, with higher scores indicating better performance. Another non-intrusive speech quality assessment metric, NISQAv2, provides a more detailed evaluation of speech quality. It evaluates the given signal based on **MOS**, noisiness (**NOIS.**), discontinuity (**DIS.**), coloration (**COL.**), and loudness (**LOUD.**), with higher scores indicating better speech quality.

In addition, we use word error rate (**WER**) and character error rate (**CER**) [32] to measure the speech intelligibility of the enhanced signals. For this purpose, we employ two different pre-trained ASR backends: wav2vec2.0[1] and Hubert[2]. Both WER and CER follow a "lower is better" principle, meaning that lower values indicate better recognition of the speech signal.

D. Results

Choice of SSL model. To determine the optimal SSL features for vocoder input, we implemented the vocoder using three different pre-trained SSL models: HuBERT, WavLM and TERA.

[1]https://huggingface.co/facebook/wav2vec2-base-960h
[2]https://huggingface.co/facebook/hubert-large-ls960-ft

TABLE I: Vocoder performance with different SSL models, trained with the *generator loss only*. Use learnable weight summation of each model as input.

SSL Model	DNSMOS			NISQAv2				
	OVRL	**SIG**	**BAK**	**MOS**	**NOIS.**	**DIS.**	**COL.**	**LOUD.**
Target	3.668	3.951	4.209	4.550	4.251	4.596	4.301	4.476
HuBERT	3.435	3.725	4.131	3.256	3.637	3.284	3.331	3.777
WavLM	3.477	3.755	4.158	3.354	3.760	3.355	3.373	3.817
TERA	**3.607**	**3.922**	**4.160**	**4.021**	**3.893**	**4.184**	**3.823**	**4.224**

TABLE II: Analysis with various types of generator loss functions using *noise+reverb*.

Loss function	DNSMOS			NISQAv2
	OVRL	**SIG**	**BAK**	**MOS**
Distorted	1.819	2.545	1.812	1.659
Mel-spectrogram	2.893	3.241	3.859	2.035
SDR	1.841	2.047	4.014	1.117
Time-domain Huber	2.878	3.218	3.842	1.589
Compressed spec.	**3.170**	**3.483**	**4.044**	**2.145**

We utilized a clean input waveform and employed hidden features obtained through learnable weighted summation as input to the vocoder network. The results are summarized in Table I. It is important to note that the vocoders consistently utilize the learnable weighted summation of all available hidden states, including all four layers of TERA or all twelve layers of HuBERT. Speech generated from pre-trained TERA embeddings demonstrates superior scores in terms of speech intelligibility, quality, and denoising, even though HuBERT extracts three times more hidden features than TERA. This supports the findings from the task-agnostic comparison of SSL models in [18], which again highlighted TERA as a highly suitable option. Based on these results, TERA is used in the subsequent studies.

Generator loss function. Table II presents the instrumental metric scores for different loss functions used in the vocoder generator loss. The evaluation is conducted on both reverberant and noisy data. It can be observed that the compressed spectrogram loss yields the best results across all metrics. Thus, this loss function was selected as the generator loss function, replacing the mel-spectrogram loss. Due to space constraints, only the MOS score from NISQAv2 is presented. However, we confirmed that the other NISQAv2 metrics showed similar trends to those observed with DNSMOS.

Consequently, both the Clean Vocoder and Denoising Vocoder baselines are trained using this generator loss with the original HiFi-GAN adversarial loss. It is important to reiterate their main difference: the Clean Vocoder synthesizes speech using embeddings extracted from clean speech, while the Denoising Vocoder aims to synthesize the underlying clean speech from embeddings extracted directly from the noisy input signal.

Embedding enhancement network. Table III benchmarks the performance of the proposed two-stage framework, which includes the embedding enhancement network. This network comprises four ConvGRU blocks, a configuration that has been shown to provide the most significant improvement in speech signal quality and noise reduction compared to other setups. To ensure that the enhanced embeddings are accurately mapped to clean speech, we use the pre-trained Clean Vocoder model to generate the underlying clean speech.

Overall, the **Proposed.** framework consistently outperforms the Denoising Vocoder, demonstrating that the task-agnostic embedding enhancement is indeed beneficial. The ASR performance results in Table III focus only on background noise, as prior research [18] has shown that background noise has a more pronounced impact on embeddings than reverberation. We verified this finding: using wav2vec2.0 as backend, the noise test set yielded WER and CER scores of 47.60% and 35.95%, respectively, whereas the reverberant test set yielded scores of 17.10% and 7.53%. This trend is also observed with the HuBERT-based backend. The proposed system outperforms the Denoising Vocoder and generally provides better than directly using the noisy input.

Layer-wise analysis. To assess the impact of different layers, we measure the distortion of the representation at *each* layer using normalized MSE against the representation from clean speech, following the methodology outlined in [18]. As shown in Fig. 2, noise affects the embeddings more significantly than reverberation, as indicated by larger deviations in the distances. The representation from the last layer shows the lowest normalized MSE.

Next, we investigate the layer-wise weighting factors relative to the input distortion, as detailed in Table IV. The first row shows the learnable weighting factors for distorted speech input, while the second row represents the weighting factors for clean speech input. The robustness analysis of the last hidden feature, as depicted in Figure 2, aligns with the weighting factor results, which show that the last hidden feature is given the highest weight. Therefore, for evaluating model performance using a *single* embedding, we select the embedding from the last layer.

Table V presents the performance results for different combinations of TERA feature embeddings. The performance with the last hidden feature is comparable to that with fixed summation, indicating that fixed weighted summation might be less effective than using a single robust feature. The learnable sum approach achieved the highest performance, suggesting

TABLE III: Comparison with proposed embedding enhancement network. Evaluate with the combination of *noise+reverb* for † mark and tested with each metric of DNSMOS and NISQAv2, and use *noise* for $*$ mark and tested with pre-trained model of wav2vec2.0 and Hubert

Model	DNSMOS†			NISQAv2†					wav2vec2.0-based ASR*		Hubert-based ASR*	
	OVRL	SIG	BAK	MOS	NOIS.	DIS.	COL.	LOUD.	WER (%)	CER (%)	WER (%)	CER (%)
Clean	3.668	3.951	4.209	4.550	4.251	4.596	4.301	4.476	9.27	3.25	5.78	1.73
Distorted	1.819	2.545	1.812	1.659	1.698	3.073	2.328	2.505	47.60	35.95	31.73	21.69
Clean Vocoder	3.569	3.886	4.102	3.751	3.546	3.996	3.574	4.038	14.85	6.12	7.69	2.80
Denoising Vocoder	3.038	3.397	3.863	2.698	2.966	**3.579**	2.699	3.384	51.08	30.94	41.63	24.85
Proposed.	**3.285**	**3.606**	**4.027**	**2.779**	**3.204**	3.009	**2.737**	**3.597**	**45.73**	**27.50**	**37.34**	**22.68**

TABLE IV: Weighting of each latent feature from the two baselines, Denoising- and Clean Vocoder, for the corresponding inputs.

Input	1st layer	2nd layer	3rd layer	4th layer
Distorted	-0.018	0.113	-0.078	**1.360**
Clean	-0.025	0.171	0.244	**1.738**

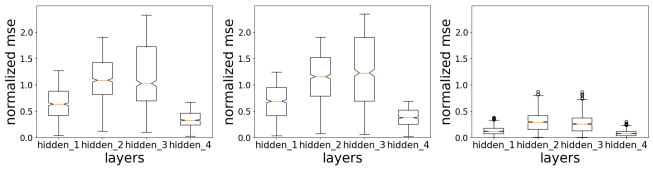
TABLE V: Analysis with various types of input embedding combinations, trained with all loss terms.

Input combination	DNSMOS			NISQAv2
	OVRL	SIG	BAK	MOS
Last hidden	2.718	3.059	3.673	2.008
Fixed sum	2.789	3.130	3.763	1.990
Learnable sum	**3.170**	**3.483**	**4.044**	**2.145**

that even though a single robust hidden feature is valuable, the model still benefits from incorporating information from multiple embeddings to synthesize speech more effectively.

V. Conclusions

We proposed an embedding enhancement network that maps distorted speech embeddings, obtained from pre-trained SSL models applied to noisy and reverberant inputs, to the embeddings of the underlying clean speech. These enhanced embeddings are then used by a vocoder to generate clean speech. After selecting TERA as the SSL model, experiments were conducted to identify the optimal generator loss function for the vocoder, resulting in the adoption of the compressed spectrogram loss. Additionally, a layer-wise analysis was performed to evaluate the robustness of embeddings from different hidden layers under various input distortions. This analysis revealed two key points: first, that additive noise has a greater impact on embeddings than reverberation, and second, that the final layer embedding is the most robust. These findings were supported by the analysis of weighting factor according to distortion type, which indicates contribution of each hidden layer to speech enhancement. The best performance was achieved using a learnable weighted combination of all layers, demonstrating that incorporating information from multiple embeddings remains beneficial. The proposed system, which includes the embedding enhancement, consistently outperforms the Denoising Vocoder baseline in both ASR tasks and speech quality metrics, even under challenging conditions.



(a) Noise+reverb (b) Noise (c) Reverb

Fig. 2: Standardized MSE between the clean and distorted embeddings from pre-trained TERA models. Three types of distortions (*noise+reverb*, *noise*, *reverb*) are simulated.

VI. Acknowledgment

This work has been supported by Global Education and Research Center for Convergence Technology, a Brain Korea 21 four program, Yonsei University.

References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. Adv. in Neural Inf. Process. Syst.*, vol. 33, pp. 12 449–12 460, 2020.

[2] S. Chen, C. Wang, Z. Chen, J. Li, and et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. of Sel. Topics in Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[3] W. Hsu, B. Bolte, Y. Tsai, K. Lakhotia, and et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 29, pp. 3451–3460, 2021.

[4] S. Yang, P. Chi, Y. Chuang, C. Lai, and et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. INTERSPEECH*, 2021, pp. 1194–1198.

[5] X. Zhao, Q. Zhu, and J. Zhang, "Speech enhancement using self-supervised pre-trained model and vector quantization," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 330–334.

[6] K. Hung, S. Fu, H. Tseng, H. Chiang, and et al., "Boosting Self-Supervised Embeddings for Speech Enhancement," in *Proc. INTERSPEECH*, 2022, pp. 186–190.

[7] Z. Huang, S. Watanabe, S. Yang, P. García, and et al., "Investigating self-supervised learning for speech enhancement and separation," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2022, pp. 6837–6841.

[8] S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and et al., "BigVGAN: A universal neural vocoder with large-scale training," in *Proc. Intl. Conf. on Telecommun. and Signal Process.*, 2023.

[9] Y. Li, C. Han, V. Raghavan, G. Mischler, and et al., "StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Proc. Adv. in Neural Inf. Process. Syst.*, vol. 36, 2024.

[10] W. Jang, D. Lim, J. Yoon, B. Kim, and et al., "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation," in *Proc. INTERSPEECH*, 2021, pp. 2207–2211.

[11] B. Irvin, M. Stamenovic, M. Kegler, L. Yang, and et al., "Self-supervised learning for speech enhancement through synthesis," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2023, pp. 1–5.

[12] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. Adv. in Neural Inf. Process. Syst.*, vol. 33, pp. 17 022–17 033, 2020.

[13] Z. Wang, X. Zhu, Z. Zhang, Y. Lv, and et al., "SELM: Speech enhancement using discrete tokens and language models," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2024, pp. 11 561–11 565.

[14] Y. Koizumi, H. Zen, S. Karita, Y. Ding, and et al., "Miipher: A robust speech restoration model integrating self-supervised speech and text representations," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023, pp. 1–5.

[15] R. Scheibler, "SDR — medium rare with fast computations," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2022, pp. 701–705.

[16] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. Intl. Conf. on Telecommun. and Signal Process.*, IEEE, 2021, pp. 72–76.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[18] Y. Song, D. Kim, N. Madhu, and H. Kang, "On the disentanglement and robustness of self-supervised speech representations," in *Intl. Conf. on Electron., Inf. and Commun.*, 2024, pp. 662–665.

[19] A. Liu, S. Li, and H. Lee, "TERA: Self-supervised learning of transformer encoder representation for speech," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 29, pp. 2351–2366, 2021.

[20] C. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2021, pp. 6493–6497.

[21] J. Kearns, "Librivox: Free public domain audiobooks," *Reference Rev.*, vol. 28, no. 1, pp. 7–8, 2014.

[22] J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, and et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2017, pp. 776–780.

[23] E. Fonseca, J. P. Puig, X. Favory, F. F. Corbera, and et al., "Freesound datasets: A platform for the creation of open audio datasets," in *Proc. Intl. Soc. for Music Inf. Retrieval*, ISMIR, 2017.

[24] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. of Meetings on Acoust.*, vol. 19, 2013.

[25] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and et al., "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2017, pp. 5220–5224.

[26] C. Reddy, H. Dubey, K. Koishida, A. Nair, and et al., "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *Proc. INTERSPEECH*, 2021, pp. 2796–2800.

[27] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[28] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proc. of the National Academy of Sci.*, vol. 113, no. 48, E7856–E7865, 2016.

[29] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Intl. Conf. on Learning Representations*, 2019.

[30] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. INTERSPEECH*, 2021, pp. 2127–2131.

[31] J. Pirklbauer, M. Sach, K. Fluyt, W. Tirry, and et al., "Evaluation metrics for generative speech enhancement methods: Issues and perspectives," in *SC; 15th ITG Conference*, VDE, 2023, pp. 265–269.

[32] A. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition," in *Proc. Intl. Conf. on Spoken Language Process.*, 2004.

Generalized Graph Signal Sampling under Subspace Priors by Difference-of-Convex Minimization

Keitaro Yamashita*, Kazuki Naganuma*, and Shunsuke Ono*
*Institute of Science Tokyo, Japan
E-mail: yamashita.k.45d1@m.isct.ac.jp; naganuma.k.491a@m.isct.ac.jp; ono.s.5af2@m.isct.ac.jp

***Abstract*—This paper proposes an effective approach for sampling graph signals under the subspace prior. Unlike conventional methods that assume bandlimited signals, our method, based on generalized sampling theory, designs a sampling operator suitable for general graph signals beyond bandlimitedness. We formulate a feasibility problem for designing the sampling operator for graph signals under the subspace prior, and then transform it into a difference-of-convex (DC) minimization problem. The DC problem involves a tight relaxation of the intractable rank constraint in the original problem via the nuclear norm. To solve this DC problem, we present an algorithm based on the general double-proximal gradient DC algorithm (GDPGDC), which ensures convergence to a critical point of our DC problem. The effectiveness of our approach is validated by sampling and recovery experiments on various graph signal models.**

I. Introduction

Graph Signal Processing (GSP) is a novel framework for analyzing data represented as signals located at the vertices of a graph. Currently, GSP is a thriving area of research within the signal processing field, covering a wide range of studies from theoretical approaches to practical applications. For a comprehensive overview of GSP, see, for example, [1]–[3].

In the realm of GSP, one of the fundamental technologies is *graph signal sampling* [4], [5]. A significant difference from traditional sampling methods is that graph signal sampling does not have a regular sampling pattern. Consequently, extensive research has been devoted to extending the principles of Shannon-Nyquist sampling theory to accommodate graph signals. Most of the existing sampling methods are based on ensuring bandlimitedness with respect to graph frequencies, as shown in previous works such as [4], [6]–[14]. However, there is a wide variety of graph signals *not* assuming bandlimitedness, which can also be perfectly recovered by extending the *generalized sampling* theory [15], [16].

In generalized sampling, there are two main sampling approaches, one in the *graph frequency domain* and the other in the *graph vertex domain*. For the graph frequency domain, the authors of [17] pioneered the extension of generalized sampling theory to sample and reconstruct graph signals beyond bandlimitedness, which led to branches in the graph frequency domain, one focusing on smoothness and subspace priors [18], and the other on the stochastic prior [19]. While approahes in the graph frequency domain are elegant, they have limitations, such as requiring graph Fourier transforms, so approaches in the vertex domain have also been explored.

For sampling approaches in the graph vertex domain, the study in [20] proposes a vertex-wise sampling method for signals under arbitrary priors, including subspace, smoothness, and stochastic priors. However, since the method is based on a greedy algorithm, it may select biased sampling positions and is sensitive to noise. On the other hand, the sampling method proposed in [21], while targeting only the subspace prior, takes an approach of directly designing a sampling operator through convex optimization. This method allows for more efficient and robust sampling because it can generate samples by mixing multiple signal values. However, in order to transform the sampling operator design problem into a convex optimization problem, this method significantly relaxes a certain invertibility constraint involving the sampling operator to be designed. This relaxation is undesirable from the perspective of generalized sampling theory.

A natural question arises: *Can the flexible sampling operator design problem with the constraint be reduced to a tractable optimization problem without significant relaxation?* In this paper, we propose a novel sampling operator design method for graph signals under the subspace prior by utilizing difference-of-convex (DC) optimization techniques, which we aim to design *flexible* sampling operator that mixes the signal values of multiple vertices to create a sampled signal.[1] The key contributions of our study are summarized as follows:

- Formulate a feasibility problem for designing the flexible sampling operator for graph signals under the subspace prior, incorporating a rank constraint derived from the generalized sampling theory.
- Reformulate the original problem as a tractable DC problem via a tight relaxation of the rank constraint by using the nuclear norm.
- Develop an efficient algorithm based on the *general double-proximal gradient DC algorithm* (GDPGDC) [24] to solve the DC problem with guaranteed convergence to a critical point.

Finally, we perform sampling and recovery experiments on different types of graph signals to validate the effectiveness of our method.

[1]In [22], DC-based graph signal sampling for the smoothness prior is also discussed. DC optimization is also attractive to other signal processing applications [23].

II. Preliminaries

A. Notation and Definitions

Bold lowercase letters reperesent vectors, and bold uppercase letters represent matrices. We use x_i and X_{ij} to represent the i-th element of a vector $\mathbf{x}$ and the element in the i-th row and j-th column of a matrix $\mathbf{X}$, respectively. The ℓ_2 norm of a vector $\mathbf{x}$ is denoted by $\|\mathbf{x}\|_2 := \sqrt{\sum_i x_i^2}$. The transpose of a matrix $\mathbf{X}$ is denoted as $\mathbf{X}^\top$. The inverse and the pseudo-inverse of a matrix $\mathbf{X}$ are denoted by $\mathbf{X}^{-1}$ and $\mathbf{X}^\dagger$, respectively. The inner product of two matrices $\mathbf{X}$ and $\mathbf{Y}$ is denoted by $\langle \mathbf{X}, \mathbf{Y} \rangle := \sum_i \sum_j X_{ij} Y_{ij}$. The i-th singular value of $\mathbf{X}$ is denoted by $\sigma_i(\mathbf{X})$. The ℓ_1 norm, the Frobenius norm, and the nuclear norm of $\mathbf{X}$ are denoted by $\|\mathbf{X}\|_1 := \sum_i \sum_j |X_{ij}|$, $\|\mathbf{X}\|_F := \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$, and $\|\mathbf{X}\|_* = \sum_i \sigma_i(\mathbf{X})$, respectively. A diagonal matrix with $\cdot$ as its principal diagonal is denoted by $\mathrm{diag}(\cdot)$. The indicator function of a set $\mathcal{C}$ is denoted by $\iota_{\mathcal{C}}(\mathbf{X})$ that is defined such that $\iota_{\mathcal{C}}(\mathbf{X}) = 0$ when $\mathbf{X} \in \mathcal{C}$, and $\iota_{\mathcal{C}}(\mathbf{X}) = \infty$ otherwise. The extended real line $\overline{\mathbb{R}}$ is defined as $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\}$. An identity matrix is denoted by $\mathbf{I}$, and $\mathbf{I}_N$ represents $N \times N$ identity matrix.

B. Generalized Sampling of Graph Signals

In this section, we provide a brief overview of the generalized sampling theory [15], [16] for graph setting [5], [18] which forms the foundation of our method. We consider a weighted undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ and $\mathcal{E}$ denote a set of vertices and a set of edges between the vertices, respectively. The number of vertices is denoted as $N = |\mathcal{V}|$. We define an adjacency matrix $\mathbf{E} \in \mathbb{R}^{N \times N}$, where E_{ij} is the weight of the edge between the i-th and j-th vertices. The degree matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ is a diagonal matrix, where the i-th diagonal element $D_{ii} := \sum_j E_{ij}$ represents the sum of weights connected to vertex i.

We define a graph Laplacian as $\mathbf{L} := \mathbf{D} - \mathbf{E}$ as a graph variation operator for clarity and specificity. Since $\mathbf{L}$ is a real symmetric matrix, it always admits an eigendecomposition $\mathbf{L} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top$, where $\mathbf{U} = [\mathbf{u}_1 \dots \mathbf{u}_N]$ forms a unitary matrix containing the eigenvectors $\mathbf{u}_i$, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$ comprises the eigenvalues λ_i. We denote $\mathbf{U}$ and λ_i as the graph Fourier basis and the graph frequency, respectively, and the graph frequency is smaller as λ_i is smaller.

There are two approaches for sampling and recovering: sampling and recovering in the vertex domain [4], [7] and those in the frequency domain [10]. We describe the sampling approach in the vertex domain as our proposal in this paper focuses in it.

Fig. 1 shows the outline of sampling in the vertex domain. Let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^N$, $\mathbf{c} \in \mathbb{R}^M (M \leq N)$, and $\tilde{\mathbf{x}} \in \mathcal{X}$ be an original graph signal, the sampled signal, and the recovered graph signal, respectively. The graph signal $\mathbf{x}$ undergoes sampling by a sampling operator $\mathbf{S}^\top \in \mathbb{R}^{M \times N}$, i.e., $\mathbf{c} := \mathbf{S}^\top \mathbf{x}$. Subsequently, the sampled signal $\mathbf{c}$ is filtered with a correction operator $\mathbf{H}$ to reduce any errors or distortions introduced during the sampling and recovering process. Following this, it is further filtered by a reconstruction operator $\mathbf{W}$ to map

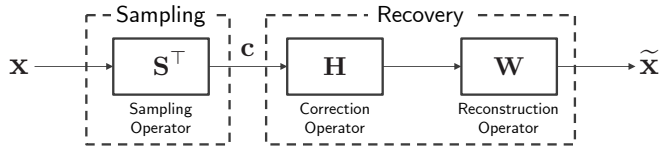


Fig. 1: Framework for generalized sampling of graph signals. Here, $\mathbf{x}$, $\mathbf{c}$, and $\tilde{\mathbf{x}}$ represent the original, sampled, and recovered graph signals, respectively.

the sampled and corrected signal back onto the original graph. Hence, the recovered signal $\tilde{\mathbf{x}}$ is represented as follows:

$$\tilde{\mathbf{x}} = \mathbf{W}\mathbf{H}\mathbf{c} = \mathbf{W}\mathbf{H}\mathbf{S}^\top \mathbf{x}. \tag{1}$$

The reconstruction operator $\mathbf{W}$ may be constrained, i.e., it may be predefined for some reason, such as computational or hardware constraints. We call the cases where $\mathbf{W}$ is constrained as *predefined case* and the cases where $\mathbf{W}$ is not constrained as *unconstrained case*. The recovery problem entails finding the optimal $\mathbf{H}$ (and $\mathbf{W}$ if not predefined) based on the assumed priors of graph signals. This framework encompasses various situations involving sampling and recovery, including bandlimited graph signals (refer to [5] for more details).

C. Graph Signal Sampling under Subspace Prior

In this paper, we consider sampling and recovering graph signals under the subspace prior [5] for the unconstrained case. We provide the description of the subspace prior and the desigining method for the best $\mathbf{H}$ and $\mathbf{W}$ under the prior from the established results in this subsection.

Under the the subspace prior, we suppose that a graph signal $\mathbf{x} \in \mathbb{R}^N$ is characterized by a linear model as follows:

$$\mathbf{x} := \mathbf{A}\mathbf{d}, \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{N \times K}$ $(K \leq N)$ is a known generator matrix and $\mathbf{d} \in \mathbb{R}^K$ are expansion coefficients. For simplisity, we consider cases that $M \leq K$ in this paper.

The operators $\mathbf{H}$ and $\mathbf{W}$ are designed based on well-established strategies; the least-squares (LS) strategy and the minimax (MX) strategy. The LS strategy aims to find the recovered signal $\tilde{\mathbf{x}}$ that minimizes the ℓ_2 norm of the difference between the recovered signal $\tilde{\mathbf{x}}$ and the sampled signal $\mathbf{c}$:

$$\tilde{\mathbf{x}}_{\mathrm{LS}} = \underset{\tilde{\mathbf{x}} \in \mathcal{X},\ \mathbf{S}^\top \mathbf{x} = \mathbf{c}}{\mathrm{argmin}} \|\mathbf{S}^\top \tilde{\mathbf{x}} - \mathbf{c}\|_2^2, \tag{3}$$

The MX strategy attempts to directly control the recovery error $\|\tilde{\mathbf{x}} - \mathbf{x}\|_2$ for minimizing the error for the worst feasible signal:

$$\tilde{\mathbf{x}}_{\mathrm{MX}} = \underset{\tilde{\mathbf{x}} \in \mathcal{X}}{\mathrm{argmin}} \max_{\mathbf{x} \in \mathcal{X}} \|\tilde{\mathbf{x}} - \mathbf{x}\|_2^2. \tag{4}$$

The solutions of the LS and MX strategy are both given by

$$\tilde{\mathbf{x}} = \mathbf{A}(\mathbf{S}^\top \mathbf{A})^\dagger \mathbf{S}^\top \mathbf{x}, \tag{5}$$

followed by the correction operator $\mathbf{H}$ and the reconstruction operator $\mathbf{W}$ as

$$\mathbf{H} = (\mathbf{S}^\top \mathbf{A})^\dagger, \quad \mathbf{W} = \mathbf{A}. \tag{6}$$

Here are three examples of graph signal models assuming the subspace prior:

- Bandlimited (BL) signal [5] is one of the most studied signal models, which are characterized as:

$$\mathbf{x} = \sum_{i=0}^{K-1} d_i \mathbf{u}^{(i)} = \mathbf{U}_{\mathcal{BL}}\mathbf{d}, \tag{7}$$

where $\mathbf{U}_{\mathcal{BL}} \in \mathbb{R}^{N\times K}$ is the submatrix of $\mathbf{U}$ whose rows are extracted within $\mathcal{BL} = \{1, \ldots, K\}$. In this case,

$$\mathbf{W}_{\mathrm{BL}} = \mathbf{A}_{\mathrm{BL}} = \mathbf{U}_{\mathcal{BL}}. \tag{8}$$

- Periodic graph spectrum (PGS) signal [18] assumes the periodicity of the graph spectrum as follows:

$$\mathbf{x} = \mathbf{U}A(\mathbf{\Lambda})\mathbf{D}_{\mathrm{samp}}^{\top}\mathbf{d}, \tag{9}$$

where $A(\mathbf{\Lambda}) := \mathrm{diag}(A(\lambda_0), \ldots, A(\lambda_{N-1}))$ represents a graph spectral response of the generator, and $\mathbf{D}_{\mathrm{samp}} = [\mathbf{I}_K \ \mathbf{I}_K \ \ldots] \in \mathbb{R}^{K\times N}$ is the matrix for the graph Fourier transform domain upsamling. In this case,

$$\mathbf{W}_{\mathrm{PGS}} = \mathbf{A}_{\mathrm{PGS}} = \mathbf{U}A(\mathbf{\Lambda})\mathbf{D}_{\mathrm{samp}}^{\top}. \tag{10}$$

- Piecewise constant (PWC) signal [25] is characterized by constant values in separated vertex regions and are defined as follows with the number of pieces K:

$$\mathbf{x} = \sum_{i=1}^{K} d_i \boldsymbol{\tau}^{(i)} = [\boldsymbol{\tau}^{(1)} \ \ldots \ \boldsymbol{\tau}^{(K)}]\mathbf{d}, \tag{11}$$

where $\tau^{(i)}$ for any $i = 1, \ldots, K$ is defined as $\tau_n^{(j)} = 1$ when the node j is in the i-th piece and $\tau_j^{(i)} = 0$ otherwise for any $j = 1, \ldots, N$. In this case,

$$\mathbf{W}_{\mathrm{PWC}} = \mathbf{A}_{\mathrm{PWC}} = [\boldsymbol{\tau}^{(1)} \ \ldots \ \boldsymbol{\tau}^{(K)}]. \tag{12}$$

D. General Double-Proximal Gradient Difference-of-Convex Algorithm

After discussing the recovery of graph signals using the subspace prior based on generalized sampling theory, we now turn our attention to an algorithm designed to solve *difference-of-convex* (DC) minimization problems, which form the basis of the optimization problems we will develop. In this section, we introduce the *General Double-Proximal Gradient Difference-of-Convex* (GDPGDC) algorithm, a flexible tool for solving DC minimization problems.

The GDPGDC [24] can solve DC minimization problems in the form of

$$\min_{\mathbf{X}} f_1(\mathbf{X}) + f_2(\mathbf{X}) - h(\mathbf{Z}) \quad \text{s.t. } \mathbf{Z} = \mathbf{B}\mathbf{X}, \tag{13}$$

where $f_1 : \mathbb{R}^{n\times m} \to \mathbb{R}$ is a differentiable convex function with $1/\beta$-Lipschitz continuous gradient for some $\beta > 0$, $f_2 : \mathbb{R}^{n\times m} \to \overline{\mathbb{R}}$ and $h : \mathbb{R}^{k\times m} \to \overline{\mathbb{R}}$ are proper lower-semi-continuous convex functions, and $\mathbf{B} : \mathbb{R}^{k\times n}$ is a matrix.

We introduce the *proximity operator* of a proper lower-semi-continuous convex function f with a parameter $\gamma > 0$ as follows:

$$\begin{aligned} &\mathrm{prox}_{\gamma f} : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m} : \\ &\mathbf{Y} \mapsto \operatorname*{argmin}_{\mathbf{X}} f(\mathbf{X}) + \frac{1}{2\gamma}\|\mathbf{Y} - \mathbf{X}\|_F^2. \end{aligned} \tag{14}$$

Then, Prob. (13) can be solved by GDPGDC by the following procedures: for $\gamma_1 > 0$ and $\gamma_2 > 0$, iterate

$$\left\lfloor \begin{array}{l} \mathbf{X}^{(t+1)} \leftarrow \mathrm{prox}_{\gamma_1 f_2}(\mathbf{X}^{(t)} - \gamma_1(\nabla f_1(\mathbf{X}^{(t)}) - \mathbf{B}^*\mathbf{Z}^{(t)})); \\ \mathbf{Z}^{(t+1)} \leftarrow \mathrm{prox}_{\gamma_2 h^*}(\mathbf{Z}^{(t)} + \gamma_2\mathbf{B}\mathbf{X}^{(t+1)}); \\ t \leftarrow t + 1; \end{array} \right. . \tag{15}$$

Here, the *Fenchel–Rockafellar conjugate function* of h is defined as

$$h^*(\mathbf{X}) := \max_{\mathbf{Y}} \langle \mathbf{X}, \mathbf{Y}\rangle - h(\mathbf{Y}). \tag{16}$$

Thanks to Moreau's Identity[26, Theorem 14.3(ii)], the proximity operator of h^* is calculated with a parameter $\gamma_2 > 0$ as follows:

$$\mathrm{prox}_{\gamma_2 h^*}(\mathbf{X}) = \mathbf{X} - \gamma_2 \mathrm{prox}_{\frac{1}{\gamma_2}h}\left(\frac{1}{\gamma_2}\mathbf{X}\right). \tag{17}$$

We summarize the theoretical results for the convergence of GDPGDC as follows:

Theorem 1 ([24, Proposition 4] Convergence of the sequence generated by GDPGDC)**.** *Let* $\inf\{f_1(\mathbf{X}) + f_2(\mathbf{X}) - h(\mathbf{B}\mathbf{X}) \mid \mathbf{X} \in \mathcal{H}\} > -\infty$*, where* $\mathcal{H}$ *is a Hilbelt space, and let* $0 < \gamma_1 < 2\beta$ *and* $0 < \gamma_2 < +\infty$ *be satisfied. If* $\{\mathbf{X}^{(t)}\}_{t\in\mathbb{N}}$ *and* $\{\mathbf{Z}^{(t)}\}_{t\in\mathbb{N}}$*, generated by Algorithm* (15)*, are bounded, then every cluster point of* $\{\mathbf{X}^{(t)}\}_{t\in\mathbb{N}}$ *is a critical point of Prob.* (13)*.*

III. Proposed Method

The generalized sampling theory works under the assumption of a predefined sampling operator $\mathbf{S}^{\top}$. This assumption, which may be determined by hardware for example, does not provide any direction for the specific design of $\mathbf{S}$. In contrast, when dealing with graph signal sampling, the main challenge revolves around the strategic construction of $\mathbf{S}$. To address this, we formulate a feasibility problem to design the flexible sampling operator for graph signals under the subspace prior, which is then reformulated and relaxed into a DC problem. Finally, we develop an effective GDPGDC-based algorithm to solve this problem.

A. Problem Formulation

Suppose $\mathbf{W}$ and $\mathbf{H}$ are defined as shown in Eq. (6), our current task is to devise an appropr iate $\mathbf{S}$ that is consistent with the sampling and recovery process described in Eq. (1). To achieve the best possible recovery, a promising strategy is to *find* $\mathbf{S}$ *for which the correction operator* $\mathbf{H} = (\mathbf{S}^{\top}\mathbf{A})^{\dagger}$ *has the full column rank, i.e.,* $\mathbf{S}^{\top}\mathbf{A}$ *has the full row rank.*

The question then arises how to design such $\mathbf{S}$. To address this, we first formulate the sampling operator design problem

as the following feasibility problem:

$$\text{find } \mathbf{S} \text{ s.t. } \begin{cases} \|\mathbf{S}\|_F \leq \varepsilon, \\ \mathbf{S}^\top \mathbf{A} \text{ has full column rank.} \end{cases} \tag{18}$$

The first constraint serves to control the absolute values of the elements of the sampling operator, ensuring stable sampling within a radius of Frobenius norm $\varepsilon > 0$. As $\mathbf{S}^\top \mathbf{A}$ is $M \times K$ matrix, the second constraint should be

$$\operatorname{rank}(\mathbf{S}^\top \mathbf{A}) = \operatorname{rank}(\mathbf{A}^\top \mathbf{S}) = M. \tag{19}$$

Then, Prob. (18) can be rewritten as follows:

$$\text{find } \mathbf{S} \text{ s.t. } \begin{cases} \mathbf{S} \in \mathcal{B}_{F,\varepsilon} := \{\mathbf{X} \mid \|\mathbf{X}\|_F \leq \varepsilon\}, \\ \operatorname{rank}(\mathbf{A}^\top \mathbf{S}) = M. \end{cases} \tag{20}$$

Prob. (20) is still a challenge because of the rank constraint. To overcome this, as it is known that the nuclear norm of a matrix is the tightest convex envelope of the rank function [27], we relax the problem into a constrained convex *maximization* problem as follows:

$$\max_{\mathbf{S}} \|\mathbf{A}^\top \mathbf{S}\|_* \text{ s.t. } \mathbf{S} \in \mathcal{B}_{F,\varepsilon}. \tag{21}$$

By introducing the indicator function of $\mathcal{B}_{F,\varepsilon}$, i.e. $\iota_{\mathcal{B}_{F,\varepsilon}}$, and reversing maximization to minimization, we can reformulate Prob. (21) as depicted as follows:

$$\min_{\mathbf{S}} \iota_{\mathcal{B}_{F,\varepsilon}}(\mathbf{S}) - \|\mathbf{A}^\top \mathbf{S}\|_*. \tag{22}$$

Here, $\iota_{\mathcal{B}_{F,\varepsilon}}$ is a proper lower-semicontinuous convex function as $\mathcal{B}_{F,\varepsilon}$ is a nonempty closed convex set, and $\|\mathbf{A}^\top \mathbf{S}\|_*$ is also a proper lower-semicontinuous convex function. Thus, this problem is regarded as the minimization of the difference between two convex functions and is reduced to Prob. (13).

B. Optimization

The algorithmic procedure for solving Prob. (22) is summarized in Algorithm 1, where $f_1(\mathbf{X}) = 0$, $f_2(\mathbf{X}) = \iota_{\mathcal{B}_{F,\varepsilon}}(\mathbf{X})$, $h(\mathbf{Z}) = \|\mathbf{Z}\|_*$, $\mathbf{X} = \mathbf{S}$, and $\mathbf{B} = \mathbf{A}^\top$ in (13).

In what follows, we derive specific computations of each step of the algorithm. Since $\mathcal{B}_{F,\varepsilon}$ in the step 2 of the Algorithm 1 is a nonempty closed convex set, the proximity operator of its indicator function is equal to the metric projection[2] onto $\mathcal{B}_{F,\varepsilon}$, i.e.,

$$\operatorname{prox}_{\iota_{\mathcal{B}_{F,\varepsilon}}}(\mathbf{X}) = \mathcal{P}_{\mathcal{B}_{F,\varepsilon}}(\mathbf{X}) = \begin{cases} \mathbf{X}, & \text{if } \mathbf{X} \in \mathcal{B}_{F,\varepsilon}; \\ \frac{\varepsilon \mathbf{X}}{\|\mathbf{X}\|_F}, & \text{otherwise.} \end{cases} \tag{23}$$

The proximity operator of h^* in the step 3 in Algorithm 1 is calculated by (17), and the proximity operator for the nuclear norm $\|\cdot\|_*$ is calculated by

$$\operatorname{prox}_{\gamma\|\cdot\|_*}(\mathbf{X}) = \mathbf{U}_\mathbf{X} \mathcal{S}_\gamma(\mathbf{\Sigma}_\mathbf{X}) \mathbf{V}_\mathbf{X}^\top, \tag{24}$$

where $\mathbf{X} = \mathbf{U}_\mathbf{X} \mathbf{\Sigma}_\mathbf{X} \mathbf{V}_\mathbf{X}^\top$ is the SVD of $\mathbf{X}$, and $\mathcal{S}_\gamma(\cdot)$ is the soft-thresholding operator applied to the singluar values, defined as:

$$\mathcal{S}_\gamma(\sigma_i) = \max(\sigma_i - \gamma, 0), \tag{25}$$

and $\mathcal{S}_\gamma(\mathbf{\Sigma}_\mathbf{X})$ is a diagonal matrix with entries $\mathcal{S}_\gamma(\sigma_i(\mathbf{X}))$, i.e., $\mathcal{S}_\gamma(\mathbf{\Sigma}_\mathbf{X}) = \operatorname{diag}(\mathcal{S}_\gamma(\sigma_i(\mathbf{X})))$. With (17) and (24), the step 3 in Algorithm 1 can be computed.

[2]The proximity operator of $\iota_\mathcal{C}$, where $\mathcal{C}$ is a nonempty closed convex set, is equivalent to the metric projection onto $\mathcal{C}$, denoted by $\mathcal{P}_\mathcal{C}$ [28].

Algorithm 1 Algorithm for designing sampling operator

Input: $\mathbf{S}^{(0)}, \mathbf{Z}^{(0)}, \varepsilon > 0, \gamma_1 > 0, \gamma_2 > 0$

1: **while** until a stopping criterion is satisfied **do**
2: $\quad \mathbf{S}^{(t+1)} \leftarrow \mathcal{P}_{\mathcal{B}_{F,\varepsilon}}\left(\mathbf{S}^{(t)} + \gamma_1 \mathbf{A}\mathbf{Z}^{(t)}\right)$ by (23);
3: $\quad \mathbf{Z}^{(t+1)} \leftarrow \operatorname{prox}_{\gamma_2 h^*}\left(\mathbf{Z}^{(t)} + \gamma_2 \mathbf{A}^\top \mathbf{S}^{(t+1)}\right)$ by (17) and (24);
4: $\quad t \leftarrow t + 1$;
5: **end while**

Output: $\mathbf{S}^{(t)}$

Remark 1 (Convergence of Prob. (22))**.** Notice that a variable whose Frobenius norm is less than ε is always returned by $\mathcal{P}_{\mathcal{B}_{F,\varepsilon}}$. This indicates that the sequence $\{\mathbf{S}^{(t)}\}_{t\in\mathbb{N}}$ generated by Algorithm 1 is bounded. Consequently, the sequence is ensured to converge to a critical point of Prob. (22) by Theorem 1.

IV. Experiments

We validate the performance of our method through sampling and recovery experiments on various types of graph signals. All experiments were carried out using MATLAB (R2024a) on a Windows 11 system with an Intel Core i9-12900 3.19 GHz processor, 32 GB RAM, and an NVIDIA GeForce RTX 3090 GPU. We compare our method with the following graph signal sampling methods: NLPD [4], SP [7], AVM [14], SASB [21], and GSSS [20]. NLPD, SP, and AVM are methods for bandlimited graph signals, while SASB and GSSS are applicable to graph signals under the subspace prior.

A. Setup

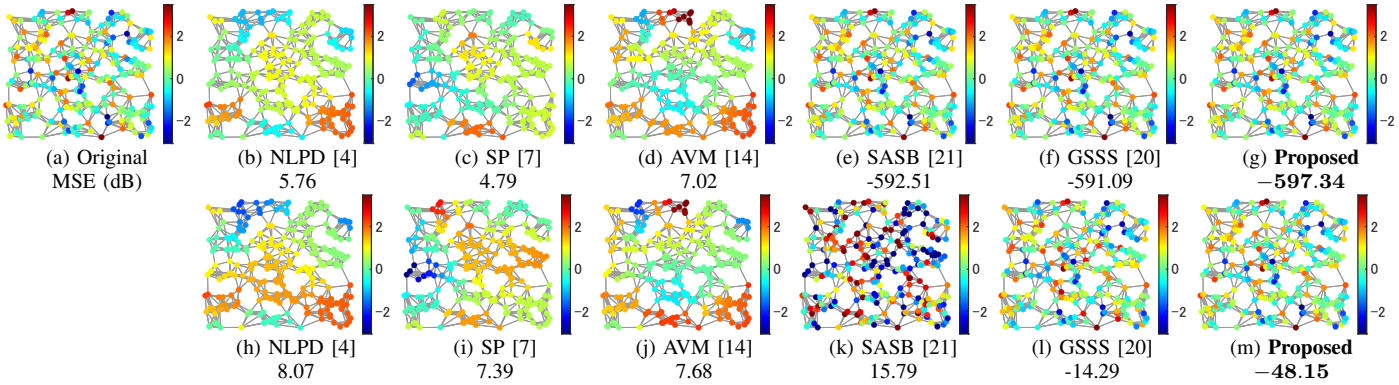
We experienced with random sensor graphs consisting of $N = 256$ vertices by using GSPBox [29]. The size of the sampled signal was set to $M = 16$. We have generated the following types of graph signals:

- Bandlimited (BL) signals (7) with $K = 16$;
- Periodic graph spectrum (PGS) signals (9) with following [18] that the graph spectral response of the generator $\mathbf{A}$ was set to $A(\lambda_i) := \exp(-1.5\lambda_i/\lambda_{\max})$ for any $i = 1, \ldots, K$, where λ_i and $\lambda_{\max}$ are the i-th graph frequency and the largest graph frequency, respectively, and $K = 16$;
- Piecewise constant (PWC) signals (11) with $K = 16$.

For all types above, we set the expansion coefficients $\mathbf{d}$ as their element $d_i \sim \mathcal{N}(1, 1)$ for all $i = 1, \ldots, K$, and we used the generator matrix $\mathbf{A}$ and the reconstruction matrix $\mathbf{W}$ as presented in (8), (10), and (12) for each type of signals. We also experimented with noisy sampled signal $\mathbf{y} := \mathbf{c} + \boldsymbol{\eta}$, where the noise $\boldsymbol{\eta} \in \mathbb{R}^M$ is generated as a white Gaussian noise with its variance $\sigma^2 = 0.1$, for all types of signals. In this case, we recovered from $\mathbf{y}$, i.e., $\tilde{\mathbf{x}} = \mathbf{W}\mathbf{H}\mathbf{y} = \mathbf{W}\mathbf{H}(\mathbf{S}^\top \mathbf{x} + \boldsymbol{\eta})$.

TABLE I: Average MSEs in Decibel of the Recoveries for 20 Independent Runs.

Signal Model	Method					
	NLPD [4]	SP [7]	AVM [14]	SASB [21]	GSSS [20]	**Proposed**
BL	-607.38	-51.80	-44.44	-316.55	-605.38	$\mathbf{-607.58}$
BL + noise	-13.14	-15.03	-2.61	13.44	-12.73	$\mathbf{-46.74}$
PGS	3.53	2.30	24.69	-578.67	-586.66	$\mathbf{-591.99}$
PGS + noise	5.29	3.29	23.91	18.17	-12.39	$\mathbf{-46.48}$
PWC	-5.67	0.06	6.82	-590.25	-591.17	$\mathbf{-591.73}$
PWC + noise	-2.91	1.76	14.30	-11.08	-12.85	$\mathbf{-46.22}$



(a) Original MSE (dB) | (b) NLPD [4] 5.76 | (c) SP [7] 4.79 | (d) AVM [14] 7.02 | (e) SASB [21] -592.51 | (f) GSSS [20] -591.09 | (g) **Proposed** $\mathbf{-597.34}$

(h) NLPD [4] 8.07 | (i) SP [7] 7.39 | (j) AVM [14] 7.68 | (k) SASB [21] 15.79 | (l) GSSS [20] -14.29 | (m) **Proposed** $\mathbf{-48.15}$

Fig. 2: An example of PGS signal and recovered graph signals on a random sensor graph ($N = 256, M = 16$). Fig. (a) shows the original PGS signal. Fig. (b)-(g) show the recovered signals from the sampled signal without noise under each method. Fig. (h)-(m) show the recovered signals from the noisy sampled signal under each method. The color of each vertex indicates the magnitude of the signal value.

The parameter ε in Prob. (22) was set to $\sqrt{NM}/8$, and γ_1 and γ_2 in our algorithm were set to $\gamma_1 = \gamma_2 = 0.001$. We defined $\mathbf{S}^{(0)}$ as a matrix with random Gaussian entries and the stopping criterion as $\|\mathbf{S}^{(t+1)} - \mathbf{S}^{(t)}\|_F / \|\mathbf{S}^{(t)}\|_F \leq 10^{-5}$. The existing methods used parameters as described in their papers.

For the quantitative evaluations, we used the mean squared error (MSE): $\mathrm{MSE} = \|\tilde{\mathbf{x}} - \mathbf{x}\|_2^2 / N$.

B. Results and Discussion

Table I presents the averaged MSEs in decibel obtained from 20 independent runs. The results in the table are expressed in decibels, with lower numbers indicating better results.

Notably, the average MSEs for NLPD, SP, and AVM assuming bandlimitedness, were lower compared to that of our method, especially for the non-bandlimited signals, PGS and PWC signals. Since SASB and GSSS have the ability to sample graph signals assuming the subspace prior, their results were better than those of the three methods, especially for the PGS and PWC signals without the noise. However, our method outperformed SASB and GSSS in all cases. In particular, it can be seen that the recovery results from noisy sampled signals were particularly superior. This difference can be attributed to the limitation of GSSS, which restricts its strategy to local sampling, as opposed to our method, which allows the integration of non-local signal values into the sampling. In other words, the samples generated by our method tend to preserve more information of the original signal. The comparison with SASB also shows that our method was able to relax the formulation more appropriately.

Fig. 2 visualizes an example of PGS signal and recovered graph signals on a random sensor graph. The distribution of colors on the graphs shown in the figure also visually confirms that our method can recover the graph signal closer to the original signal than the other methods.

V. Conclusion

In this paper, we addressed the challenge of designing a sampling operator for graph signals under the subspace prior beyond bandlimitedness. Our approach involved formulating a feasibility problem based on generalized sampling theory to determine the optimal sampling operator. To deal with the rank constraint, we reformulated the problem in a DC form and developed an algorithm based on GDPGDC that guarantees convergence to a critical point.

In contrast to prevailing methods that select individual vertices for sampling, our method introduced the novel concept of combining non-local signal values at multiple vertices to generate a sampled signal. This innovation gave our approach a degree of flexibility and efficiency not found in traditional techniques. To validate our approach, we performed sampling and recovery experiments on various graph signal models. The results confirmed that our method consistently outperformed existing methods in terms of recovery accuracy.

Acknowledgment

This work was supported in part by JST PRESTO under Grant JPMJPR21C4, JST AdCORP under Grant JPMJKB-2307, JST ACT-X Grant JPMJAX23CJ, JSPS KAKENHI under Grant 22H03610, 22H00512, 23H01415, 23K17461, 24K03119, and 24K22291, and Grant-in-Aid for JSPS Fellows under Grant 23KJ0912.

REFERENCES

[1] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, 2013.

[2] A. Sandryhaila and J. M. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, 2013.

[3] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[4] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6510–6523, 2015.

[5] Y. Tanaka, Y. C. Eldar, A. Ortega, and G. Cheung, "Sampling signals on graphs: From theory to applications," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 14–30, 2020.

[6] A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Sampling of graph signals with successive local aggregations," *IEEE Trans. Signal Process.*, vol. 64, no. 7, pp. 1832–1843, 2015.

[7] A. Anis, A. Gadde, and A. Ortega, "Efficient sampling set selection for bandlimited graph signals using graph spectral proxies," *IEEE Trans. Signal Process.*, vol. 64, no. 14, pp. 3775–3789, 2016.

[8] M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo, "Signals on graphs: Uncertainty principle and sampling," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4845–4860, 2016.

[9] D. Valsesia, G. Fracastoro, and E. Magli, "Sampling of graph signals via randomized local aggregations," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, no. 2, pp. 348–359, 2018.

[10] Y. Tanaka, "Spectral domain sampling of graph signals," *IEEE Trans. Signal Process.*, vol. 66, no. 14, pp. 3752–3767, 2018.

[11] G. Puy, N. Tremblay, R. Gribonval, and P. Vandergheynst, "Random sampling of bandlimited signals on graphs," *Appl. Comput. Harmon. Anal.*, vol. 44, no. 2, pp. 446–475, 2018.

[12] A. Sakiyama, Y. Tanaka, T. Tanaka, and A. Ortega, "Eigendecomposition-free sampling set selection for graph signals," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2679–2692, 2019.

[13] Y. Bai, F. Wang, G. Cheung, Y. Nakatsukasa, and W. Gao, "Fast graph sampling set selection using gershgorin disc alignment," *IEEE Trans. Signal Process.*, vol. 68, pp. 2419–2434, 2020.

[14] A. Jayawant and A. Ortega, "Practical graph signal sampling with log-linear size scaling," *Signal Process.*, vol. 194, p. 108 436, 2022.

[15] Y. C. Eldar and T. Michaeli, "Beyond bandlimited sampling," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 48–68, 2009.

[16] Y. C. Eldar, *Sampling theory: Beyond bandlimited systems*. Cambridge University Press, 2015.

[17] S. P. Chepuri, Y. C. Eldar, and G. Leus, "Graph sampling with and without input priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 4564–4568.

[18] Y. Tanaka and Y. C. Eldar, "Generalized sampling on graphs with subspace and smoothness priors," *IEEE Trans. Signal Process.*, vol. 68, pp. 2272–2286, 2020.

[19] J. Hara, Y. Tanaka, and Y. C. Eldar, "Graph signal sampling under stochastic priors," *IEEE Trans. Signal Process.*, vol. 71, pp. 1421–1434, 2023.

[20] J. Hara and Y. Tanaka, "Sampling set selection for graph signals under arbitrary signal priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 5732–5736.

[21] J. Hara, K. Yamada, S. Ono, and Y. Tanaka, "Design of graph signal sampling matrices for arbitrary signal subspaces," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 5275–5279.

[22] S. Ono, K. Naganuma, and K. Yamashita, "Graph signal sampling under smoothness priors: A difference-of-convex approach," *arXiv preprint arXiv:2306.14634*, 2023.

[23] K. Sato, K. Naganuma, and S. Ono, "Enhancing hyperspectral anomaly detection by difference-of-convex sparse anomaly modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, IEEE, 2024, pp. 9921–9925.

[24] S. Banert and R. I. Boţ, "A general double-proximal gradient algorithm for dc programming," *Mathematical programming*, vol. 178, no. 1-2, pp. 301–326, 2019.

[25] S. Chen, R. Varma, A. Singh, and J. Kovačević, "Representations of piecewise smooth signals on graphs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, IEEE, 2016, pp. 6370–6374.

[26] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 2nd ed. Cham: Springer International Publishing AG, 2017.

[27] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, PhD thesis, Stanford University, 2002.

[28] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

[29] N. Perraudin, J. Paratte, D. Shuman, *et al.*, "GSPBOX: A toolbox for signal processing on graphs," *arXiv:1408.5781*, 2014, [Online]. Available: https://arxiv.org/abs/1408.5781.

Gamma-VAE: Speech representation based on VAE assuming gamma distribution for both latent variables and observation

Nanako Imaichi* and Toru Nakashika†
* University of Electro-communication, Chofu, Tokyo
E-mail: n.imaichi@uec.ac.jp Tel/Fax: +81 3-6876-7836
† University of Electro-communication, Chofu, Tokyo
E-mail: nakashika@uec.ac.jp Tel/Fax: +81 42-443-5602

***Abstract*—In recent years, deep generative models that can represent complex relationships in data generation have attracted significant attention. One example is the variational autoencoder (VAE), which usually assumes a Gaussian distribution for both the encoder and decoder and often utilizes amplitude spectra because they are suitable as input features. However, the amplitude spectrum is always positive because it is a Fourier transform of the aperiodic waveform and is taken in the absolute value. Thus, a Gaussian distribution is not appropriate to assume a Gaussian distribution when using VAE with amplitude spectra as the input features. As an alternative, we propose the Gamma-VAE, which is a VAE that assumes a gamma distribution; a probability distribution that considers only positive values; and a method of back-propagation by the reparameterization trick, which also assumes a gamma distribution for latent variables. Experiments using various evaluation methods showed the effectiveness of Gamma-VAE.**

I. Introduction

In recent years, generative models such as Chat GPT [1] and Stable Diffusion [2] have been attracting significant attention becouse of their ability to generate images and text as if they were intelligent and experienced humans. Deep generative models, which are a cross between generative models and deep neural networks(DNNs), can generate realistic, high-dimensional image and audio data. Representative deep generative models include variational autoencoder (VAE) [3] (a type of autoencoder (AE) [4]), generative adversarial networks (GAN) [5], Boltzmann machines [6]. VAE has a disadvantage in that the generated data are more prone to over-smoothing than in GAN, but it has the advantage that knowledge can be easily incorporated because the distribution can be explicitly assumed. VAE also has derivatives such as the conditional VAE (CVAE) [7], vector quantized VAE (VQ-VAE) [8], VQ-VAE-2 [9], and stochastically quantized VAE (SQ-VAE) [10].

In this paper, we focus on the probability distribution, which is a characteristic of VAE. In conventional VAEs, the distributions of both the encoder and decoder are often assumed to be Gaussian. Furthermore, there are various input features utilized in VAE, such as mel-frequency cepstral coefficients (MFCC), mel-cepstra, and amplitude spectra. Amplitude spectra are often used in VAE because it does not compress the data, has less missing information, and is more compatible with DNNs with many parameters than other features. However, since the amplitude spectrum is the absolute value of the Fourier transform of the speech signal, it is always positive. Thus, the assumption of a Gaussian distribution with a range of random variables from $-\infty$ to $+\infty$ is not appropriate. In terms of modeling the amplitude spectra, it is better to assume a probability distribution that considers only non-negative values with a range of 0 to $+\infty$ for the random variable, rather than a Gaussian distribution. There are various probability distributions that represent non-negative random variables, such as the gamma, Lévy, Weibull, and Gompertz distributions. In this work, we model the amplitude spectra of speech using VAE assuming the gamma distribution, considering the ease of formula deformation and the fact that the gamma distribution is suitable for modeling the amplitude spectrum of speech [11]. In addition, although some studies have used the gamma distribution in VAE, such as those applying the gamma distribution to VAE utilized in drum score estimation [12] and those using the gamma distribution in the conjugate prior distribution estimated by the decoder [13], to the best of our knowledge, no studies have assumed the gamma distribution for latent variables by the encoder as well as observations by the decoder, and there have been no reports related to basic speech representation utilizing the VAE that assumes a gamma-distributed observation. In addition, although various reparameterization trick methods are applicable to a wide range of distributions including the gamma distribution based on approximation or implicit approaches [14], [15], no reparameterization trick has yet been proposed for the gamma-assumed VAE that enables back-propagation without any approximation. In this work, we propose Gamma-VAE, a novel VAE that assumes a gamma distribution for both latent variables output by the encoder and observed data output by the decoder, and evaluate its effectiveness through speech reconstruction experiments. We also propose a new reparametrization trick to enable back-propagation of gradients from the decoder to the encoder that outputs the distribution of latent variables following the gamma distribution.

II. Conventional Gaussian VAE

The conventional VAE that assumes a Gaussian distribution is trained to reconstruct the observed data $\boldsymbol{x} \in \mathbb{R}^D$($D$ is the dimension of the observation) via the latent variables $\boldsymbol{z} \in \mathbb{R}^Z$ (Z is the dimension of the latent variable). The system consists of two components: 1) the encoder of the parameter ϕ that outputs the mean $\boldsymbol{\mu}_\phi \in \mathbb{R}^Z$ and variance $\boldsymbol{\sigma}_\phi \in \mathbb{R}^Z$, which are the parameters of the Gaussian distribution assumed for the latent variable, and 2) the decoder of the parameter θ that outputs the mean $\boldsymbol{\mu}_\theta \in \mathbb{R}^D$, which is a parameter of the Gaussian distribution assumed for the observed data. A unit variance is often assumed for the variance of the observation.

A. ELBO

VAE simultaneously learns a set of encoder parameters ϕ and a set of decoder parameters θ on the basis of maximizing the logarithmic marginal likelihood [16]. However, since the log marginal likelihood cannot generally be computed, we consider maximizing the variational lower bound, which can also be called the evidence lower bound (ELBO) [17], through latent variables. The ELBO can be computed using variational Bayes:

$$\mathcal{L}(\boldsymbol{x};\phi,\theta) = \mathbb{E}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{\mathrm{KL}}[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})]. \quad (1)$$

In Eq. (1), the Kullback–Leibler (KL) divergence[18], the negative distance between the prior and posterior distributions of $\boldsymbol{z}$, takes on non-negative values; hence, maximizing the first term, which is the expected value of the conditional log-likelihood, maximizes the variational lower bound. The distribution of the first term can be calculated as in Eq. (3), assuming a Gaussian distribution with a variance of 1 for each dimension in Eq. (2). The expression for the expectation is omitted here.

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\sum_{d=1}^{D} \frac{(x_d - \mu_{\theta d}(\boldsymbol{z}))^2}{2}\right) \quad (2)$$

$$\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) \propto -\frac{1}{2}\sum_{d=1}^{D}(\mu_{\theta d}(\boldsymbol{z}) - x_d)^2 \quad (3)$$

The KL divergence in the second term of the right-hand side of Eq. (1), assuming the prior distribution $p(\boldsymbol{z})$ to be the standard normal distribution, can be derived as

$$D_{\mathrm{KL}}[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})] = -\frac{1}{2}\sum_{m=1}^{Z}(\log \sigma_{\phi m}^2 - \mu_{\phi m}^2 - \sigma_{\phi m}^2). \quad (4)$$

The combination of Eqs. (3) and (4) is the objective function, and the parameters are optimized to maximize it.

B. Reparameterization trick

In training based on back-propagation, the gradients computed on each layer are back-propagated from the output layer down to the input layer. However, in VAE based on the Gaussian distribution, the encoder outputs the parameter mean $\boldsymbol{\mu}_\phi \in \mathbb{R}^Z$ and variance $\boldsymbol{\sigma}_\phi \in \mathbb{R}^Z$ of the Gaussian distribution followed by randomly sampled latent variables $\boldsymbol{z}$, which prevents back-propagation. To solve this, the reparameterization trick is usually utilized. Instead of directly sampling $\boldsymbol{z} \sim N(\boldsymbol{\mu}_\phi(\boldsymbol{x}), \boldsymbol{\sigma}_\phi(\boldsymbol{x}))$, the latent variables are computed by transforming it as $\boldsymbol{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\epsilon} \circ \sqrt{\boldsymbol{\sigma}_\phi}$, where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, I)$ is randomly sampled, to allow back-propagation.

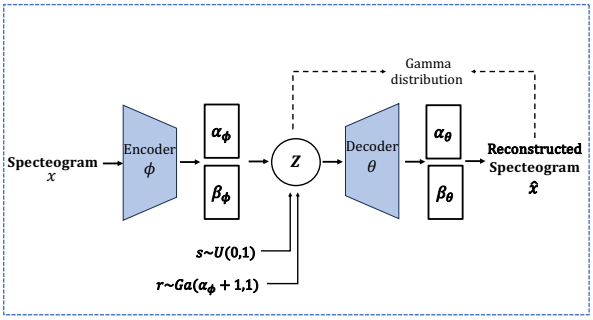


Fig. 1. Architecture of proposed Gamma-VAE.

III. Proposed Gamma-VAE

Conventional VAE assumes a Gaussian distribution for the observed data, making it unsuitable for modeling amplitude spectra that take positive values. On the other hand, using a logarithmic amplitude spectrum as a feature fits a Gaussian distribution as a range of random variables, but the generated data are excessively smoothed. One possible approach is to make the decoder predict the variance of the data distribution to prevent over-smoothing, but the variance parameter is reportedly divergent and difficult to predict [13]. Therefore, we propose Gamma-VAE based on a gamma distribution that directly represents the amplitude spectra taking positive values. In Gamma-VAE, the decoder not only outputs the parameters of the gamma distribution but also applies the gamma distribution to the distribution assumption of the latent variable output by the encoder. This is based on the idea that a variable that follows a gamma distribution and the data that follows a gamma distribution generated from that variable are compatible due to the reproducibility of the gamma distribution.

The architecture of the proposed VAE assuming a gamma distribution is shown in Figure 1. The model is trained to reconstruct the observed data $\boldsymbol{x} \in \mathbb{R}^D$ (D is the dimension of the feature) via the latent variable $\boldsymbol{z} \in \mathbb{R}^Z$ (Z is the dimension of the latent variable). The system consists of two components: 1) the encoder of the parameter ϕ that outputs $\boldsymbol{\alpha}_\phi \in \mathbb{R}^Z$ and $\boldsymbol{\beta}_\phi \in \mathbb{R}^Z$, which are the parameters of the gamma distribution assumed for the latent variables, and 2) the decoder of the parameter θ that outputs $\boldsymbol{\alpha}_\theta \in \mathbb{R}^D$ and $\boldsymbol{\beta}_\theta \in \mathbb{R}^D$, which are the parameters of the gamma distribution assumed for the observed data.

A. ELBO

As stated in II-A, the parameters of Gamma-VAE are optimized by maximizing ELBO in Eq. (1). Assuming independent

gamma distributions for all dimensions, the objective function of the decoder takes the logarithm of the probability density function of the gamma distribution when the output of the decoder is $\boldsymbol{\alpha}_\theta \in \mathbb{R}^D$ and $\boldsymbol{\beta}_\theta \in \mathbb{R}^D$. Thus, we define the conditional likelihood given latent variables $\boldsymbol{z}$ as

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = Ga(\boldsymbol{x}; \boldsymbol{\alpha}_\theta(\boldsymbol{z}), \boldsymbol{\beta}_\theta(\boldsymbol{z})) \tag{5}$$

$$= \prod_{d=1}^{D} \frac{\beta_{\theta d}}{\Gamma(\alpha_{\theta d})} x_d^{\alpha_{\theta d}-1} e^{-\beta_{\theta d} x_d}, \tag{6}$$

$$\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \sum_{d=1}^{D} -\beta_{\theta d} x_d - \log\Gamma(\alpha_{\theta d}) + \alpha_{\theta d}\log\beta_{\theta d} + (\alpha_{\theta d}-1)\log x_d, \tag{7}$$

where $\alpha_{\theta d} \in \boldsymbol{\alpha}_\theta$ and $\beta_{\theta d} \in \boldsymbol{\beta}_\theta$, and Ga and Γ respectively denote the gamma distribution and the gamma function.

The objective function of the encoder is represented by the KL divergence between the posterior and prior distributions of the latent variable output by the encoder. A simple but effective gamma distribution is preferred for the prior distribution of $\boldsymbol{z}$. $Ga(\mathbf{1}, \mathbf{1})$ and $Ga(\mathbf{2}, \mathbf{1})$ could be reasonable options. In this work, we define the prior and posterior distributions as

$$p(\boldsymbol{z}) = Ga(\boldsymbol{z}; \mathbf{1}, \mathbf{1}), \tag{8}$$

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = Ga(\boldsymbol{z}; \boldsymbol{\alpha}_\phi(\boldsymbol{x}), \boldsymbol{\beta}_\phi(\boldsymbol{x})). \tag{9}$$

With Eqs. (8) and (9), we can derive the KL divergence, which is the encoder-related term of the objective in Eq. (1), as

$$D_{\mathrm{KL}}[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})] = \sum_{m=1}^{Z} -\frac{\beta_{\phi m}-1}{\beta_{\phi m}}\alpha_{\phi m} + \log\frac{\beta_{\phi m}}{\Gamma(\alpha_{\phi m})} + (\alpha_{\phi m}-1)\psi(\alpha_{\phi m}), \tag{10}$$

where $\alpha_{\phi d} \in \boldsymbol{\alpha}_\phi$ and $\beta_{\phi d} \in \boldsymbol{\beta}_\phi$, and ψ indicates a digamma function. Based on the above, the combination of Eqs. (7) and (10) is the objective function, and the parameters are optimized to maximize it.

B. Reparameterization trick

As stated in II-B, if latent variables sampled on the basis of the probability distribution output by the decoder are utilized to calculate the gradient during training, the gradient cannot be back-propagated to the encoder side. As stated in I, sampling-based methods have been proposed to allow reparameterization for larger classes of variational distributions [14], or to solve the problem of the inapplicability of conventional methods stated in II-B to continuous distributions, an alternative approach has also been proposed to compute reparameterized gradients on the basis of implicit derivatives, which can be applied to gamma, Beta, Dirichlet, and von Mises distributions [15]. However, since there is no direct reparameterization trick methods that can be applied to the gamma distribution, in this work, we propose a reparameterization trick for Gamma-VAE that enables the calculation of gradients even for latent variables following a gamma distribution.

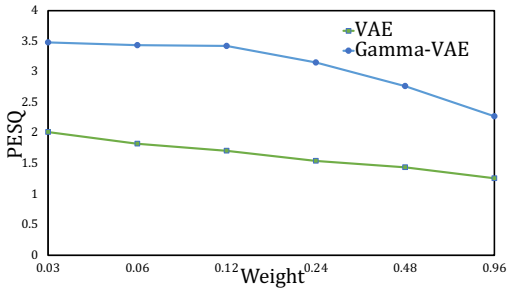


Fig. 2. Performance with various trade-off weights.

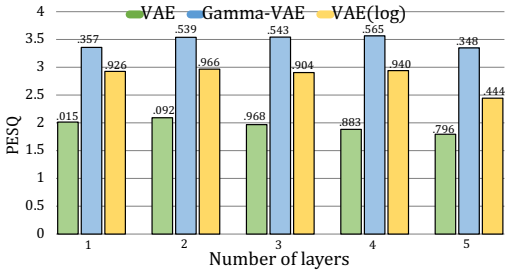


Fig. 3. Performance when changing the number of layers L.

First, we sample two mutually independent random variables $\boldsymbol{r} \in \mathbb{R}^Z$ and $\boldsymbol{s} \in \mathbb{R}^Z$ as $\boldsymbol{r} \sim Ga(\boldsymbol{r}; \boldsymbol{\alpha}_\phi(\boldsymbol{x}) + \mathbf{1}, \mathbf{1})$ and $\boldsymbol{s} \sim U([0,1])$, where $U([a,b])$ is the dimension-independent uniform distribution in the interval $[a,b]$. Second, we compute $\boldsymbol{z}$ by the change of variables, as

$$\boldsymbol{z} = \frac{1}{\boldsymbol{\beta}_\phi(\boldsymbol{x})} \circ \boldsymbol{r} \circ \boldsymbol{s}^{\frac{1}{\boldsymbol{\alpha}_\phi(\boldsymbol{x})}}, \tag{11}$$

which results in $\boldsymbol{z} \sim Ga(\boldsymbol{z}; \boldsymbol{\alpha}_\phi(\boldsymbol{x}), \boldsymbol{\beta}_\phi(\boldsymbol{x}))$. Accordingly, the gradients $\frac{\partial z_m}{\partial \alpha_{\phi m}}$ and $\frac{\partial z_m}{\partial \beta_{\phi m}}$ can be calculated and back-propagated to the encoder side after sampling $\boldsymbol{z}$ that follows the desired probability distribution.

C. Introducing a trade-off weight of the objective function

Training the model to minimize the KL divergence (Eq. (10)) implies that each dimension of the latent variable $\boldsymbol{z}$ approaches the standard exponential distribution and their respective scales become aligned and uncorrelated. However, as in the case of non-negative matrix factorization (NMF) [19], correlated underlying latent variables are sometimes desirable under non-negative constraints. Therefore, as in β-VAE [20], the degree of correlation of the underlying latent variables can be adjusted by assigning a weight $w > 0$ to each cost function in the variational lower bound, as

$$\mathcal{L}(\boldsymbol{x}; \phi, \theta) = \mathbb{E}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - w \cdot D_{\mathrm{KL}}[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})]. \tag{12}$$

IV. Experiments

A. Set-up

We conducted experiments on speech encoding and reconstruction to investigate how well Gamma-VAE can model

TABLE I

EXPERIMENTAL CONDITIONS AND RESULTS OF SUBJECTIVE EVALUATION (MOS) AND OBJECTIVE EVALUATION (PESQ, STOI, AND UT-MOS) FOR EACH METHOD. THE ABBREVIATION "AMP" REFERS TO AMPLITUDE SPECTRUM AND "LOG" TO LOGARITHMIC AMPLITUDE SPECTRUM. "UPPER BOUND" INDICATES AN EXPERIMENT TO RESTORE SPEECH USING THE AMPLITUDE SPECTRUM OF THE CORRECT ONE.

Method	**Phase recovery**	**VAE**		**enc-Gamma-VAE**		**dec-Gamma-VAE**	**Gamma-VAE**	Upper bound
Input feature		amp	log	amp	log	amp	amp	amp
PESQ	Griffin-Lim	2.09	2.96	1.93	2.67	3.42	**3.58**	4.29
	Correct phase	2.47	3.71	2.27	3.36	4.07	**4.15**	4.55
STOI	Griffin-Lim	0.93	0.91	0.90	0.90	0.96	**0.96**	0.97
	Correct phase	0.95	0.94	0.92	0.93	0.98	**0.98**	1.00
UTMOS	Griffin-Lim	1.57	1.67	1.40	1.61	**2.31**	2.26	4.11
	Correct phase	2.87	3.30	2.64	3.13	3.85	**3.86**	
MOS	Correct phase	2.35	3.59	2.34	3.29	3.79	**3.89**	4.25

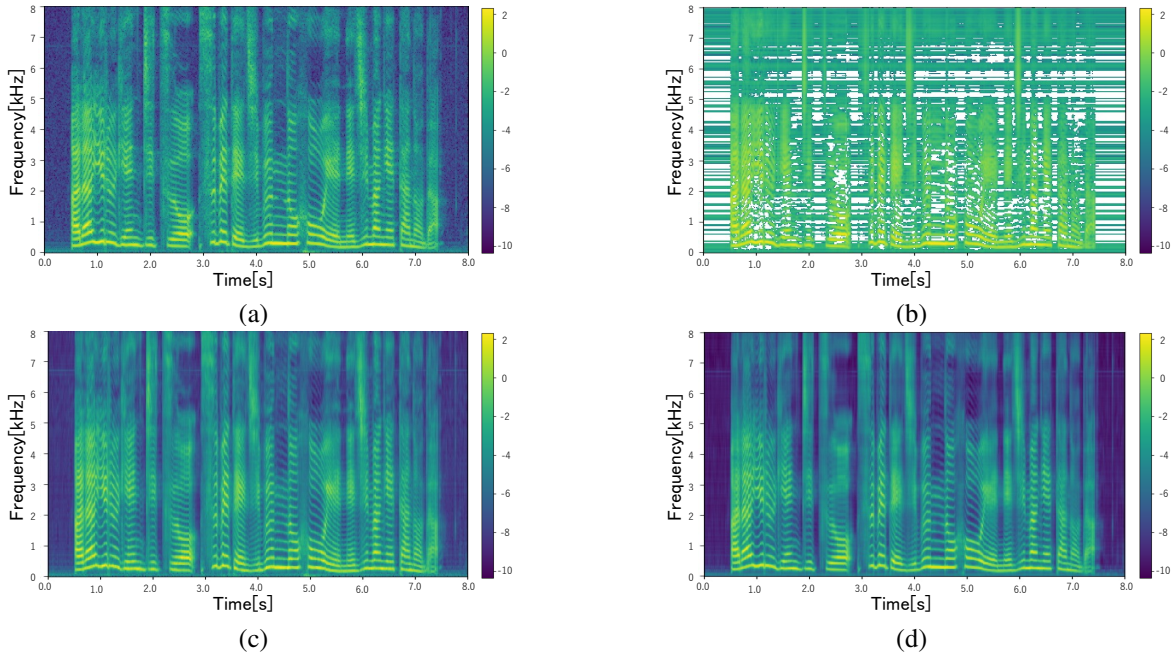


Fig. 4. Logarithm spectrograms of (a) original speech, (b) reconstructed by VAE, (c) by VAE(log), and (d) by Gamma-VAE. White in (b) indicates N/A caused by calculating the logarithm on minus values under a Gaussian assumption.

speech spectra. First, the amplitude spectra or logarithmic amplitude spectra were obtained from the speech by short-time Fourier transform, and the spectra reconstructed by each method were restored to the speech signals. Various methods for recovering the phase have been proposed, such as the Griffin-Lim algorithm [21] and von Mises distribution DNN-based methods such as [22], and in recent years, studies have also explored ways of predicting speech waveforms directly by using methods called neural vocoders, such as WaveNet [23] and HiFiGAN [24]. In this study, we focus on the accuracy of amplitude spectrum representation, so we used the simplest Griffin-Lim algorithm [21] and the correct phase to restore speech. We conducted our experiments on the ATR corpus [25]. Spectrograms were extracted with a fast Fourier transform (FFT) size of 512 and a hop size of 128. The sampling frequency was resampled to 16 kHz. We set the learning rate to 0.001 and the number of epochs to 1,000. Assuming $Z = 100$, we set the numbers of units of each layer of the encoder and decoder to 129, 64xL, 200 (for $\boldsymbol{\alpha}_\phi$ and $\boldsymbol{\beta}_\phi$), and to 100, 64xL, 258 (for $\boldsymbol{\alpha}_\theta$ and $\boldsymbol{\beta}_\theta$) from input to output, respectively, where L is the number of intermediate layers with 64 units (discussed later). Although the accuracy could be improved by using convolutional neural network (CNN), long short-term memory (LSTM) [26], the residual structure [27], etc., we here adopted a simple, fully connected structure for each layer, as we wanted to focus on the differences in the distributions output by the encoder and decoder.

For these evaluations, we not only compared Gamma-VAE to a conventional VAE assuming a Gaussian distribution but also implemented two additional ablation methods: `enc-Gamma-VAE`, which assumes a Gamma distribution for the latent variables and a Gaussian distribution for the observed data only, and `dec-Gamma-VAE`, which assumes a Gaussian distribution for the latent variables and a gamma distribution

for the observed data only. In addition, since the logarithmic amplitude spectrum is a better fit than the amplitude spectrum when a Gaussian distribution is assumed for the latent variable, experiments were conducted for VAE and enc-Gamma-VAE with two kinds of input features, one for the amplitude spectra and the other for the logarithmic amplitude spectra.

1) Evaluation metrics: PESQ, STOI, UTMOS and MOS: In this work, perceptual evaluation of speech quality (PESQ) [28], short-time objective intelligibility (STOI) [29], and UTokyo-SaruLab mean opinion score (UTMOS) [30] are utilized as the objective metrics. PESQ evaluates the subjective quality of speech signals with values ranging from –1 to 4.5 by quantifying the spectral difference between the original and degraded speech signals. STOI utilizes clean speech and speech in noisy environments to estimate the subjective sound quality of speech in noisy environments. The lowest rating is 0, and the highest is 1. UTMOS is a prediction of the MOS value of a given speech signal.

In addition, mean opinion score (MOS), a subjective evaluation metric, was also utilized in this work. MOS evaluates the synthesis results in terms of the naturalness of the converted speech and the similarity of the speakers. A total of 30 listeners needed to score each sample in a test case in accordance with the criteria: 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, and 5 = Excellent.

2) Determination of hyper-parameters: For the trade-off weight w in Eq. (12), we searched for the optimal parameter exponentially in the range of around 0 to 1. Figure 2 shows PESQ [28] when changing the value of the weight. As seen here, both models obtained higher PESQ from smaller weight. This is presumably due to the fact that the smaller the KL divergence weight, the larger the reconstructed weight, and thus the better the reconstructed performance. Further reducing the weights would be expected to further improve PESQ; however, since the latent variables would become less generalizable, we chose $w = 0.03$ for the following experiments. Next, the number of intermediate layers L was increased from 1 to 5 to find the best models. As shown in Figure 3, VAE, VAE with logarithmic amplitude spectra (`VAE(log)`), and Gamma-VAE had the best accuracy with 2, 2, and 5 layers, respectively, which will be used for the final comparison.

B. Results and discussion

As shown in Table I, the PESQ, STOI, and MOS results show similar trends. The amplitude spectrum is compatible with the gamma distribution while the logarithmic amplitude spectrum is compatible with the Gaussian distribution. The higher values for conventional VAE than for enc-Gamma-VAE indicate that the accuracy of reconstruction is better when the encoder and decoder distributions are aligned. In addition, Gamma-VAE having a higher value than conventional VAE indicates that modeling the amplitude spectrum with the gamma distribution achives higher accuracy than modeling the logarithmic amplitude spectrum with the normal distribution. A pair-wise t-test was conducted among all pairs of methods, and the proposed Gamma-VAE performed best by a significant margin (p-value ≤ 0.05) except for dec-Gamma-VAE. Lastly, Figure 4 depicts the logarithm spectrogram reconstructed by each method, where it is clear that the one reconstructed by Gamma-VAE is considerably clear and has fine spectral structures. This contributes to the effectiveness of the proposed method in PESQ, STOI, and MOS.

V. Conclusion

In this paper, we proposed gamma variational autoencoder (Gamma-VAE), a method that rewrites the distribution assumption to a gamma distribution to improve the accuracy of VAE with a Gaussian distribution, which is a conventional method in terms of modeling the amplitude spectra of speech. The results demonstrate that Gamma-VAE has higher reconstruction accuracy conventional VAE in both subjective and objective evaluations. Thus, Gamma-VAE was effective at modeling amplitude spectra. In future work, we plan to apply Gamma-VAE to other speech processing tasks such as speech synthesis, voice conversion, and speech recognition.

VI. Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 24H00715.

References

[1] "ChatGPT:A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121–154, 2023.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10 684–10 695, 2022.

[3] M. W. Diederik P Kingma, "Auto-Encoding Variational Bayes," *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

[4] G.E.Hinton,R.R.Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *SCIEENCE*, vol. 313, no. 1-3, pp. 504–507, 2006.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial net," *Proc. NIPS 2014*, pp. 2672–2680, 2014.

[6] Ackley,D.H.,Hinton,G.E.andSejnowski,T.J, "Efficient learning of deep Boltzmann machine," *Cognitive Science*, vol. 9, pp. 147–169, 1985.

[7] shon,K., Lee,H., and Yan, X., "Learnig structured output representation using deep conditional generative models," pp. 3483–3491, 2015.

[8] K. K. Aaron van den Oord Oriol Vinyals, "Neural Discrete Representation Learning," *Advances in Neural Information Process- ing Systems (NIPS) CoRR, abs/1711.00937*, 2017.

[9] V. O. Razavi A van den Oord A, "Generating diverse high-fidelity images with VQ-VAE-2.," *Advances in Neural Information Process-ing Systems (NIPS)*, vol. 32, pp. 1–11, 2019.

[10] Yuhta Takida, Takashi Shibuya, WeiHsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, Yuki Mitsufuji, "SQ-VAE: Variational Bayes on Discrete Representation with Self-annealed Stochastic Quantization," *arXiv preprint arXiv:2205.07547*, 2022.

[11] T. Nakashika and K. Yatabe, "Gamma Boltzmann Machine for Audio Modeling," *EEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2591–2605, 2021.

[12] Shunn UedaYoshiaki BandoKatsutoshi ItoyamaKazuyoshi Yoshii, "Drum score estimation for musical acoustic signals using NMF with VAE as a prior distribution," *Collection of Lectures at the 80th National Convention*, vol. 1, pp. 161–162, 2018.

[13] Takahashi HiroshiIwata TomoharuYamanaka YukiYamada Masanori, Yagi SatoshiKashima Hisashi, "Robust probability density estimation by Student-t VAE," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 36, pp. 1–9, 2021.Japanese.

[14] Naesseth, C., Ruiz, F., Linderman, S., Blei, D., "Reparameterization gradients through acceptance-rejection sampling algorithms," *Proc. Artificial Intelligence and Statistics,*, vol. 54, pp. 489–498, 2017.

[15] Figurnov, M., Mohamed, S., Mnih, A., "Implicit reparameterization gradients. advances in neural information processing systems," *Proc. NeurIPS*, 2018.

[16] R. Linske, "An application of the principle of maximum information preservation to linear systems.," *Advances in Neural Information Processing Systems*, vol. 1, pp. 186–194, 1989.

[17] M.D.Hoffman and M.J.Johnson, "Elbo surgery: Yet another way to carve up the variational evidence lower bound," *in Workshop in Advances in Approximate Bayesian Inferenc*, vol. 1, p. 2, 2016.

[18] I. Csisza, "Information type measures of difference of probability dis- tributions and indirect observations.," *Stud. Sci. Math. Hung.*, vol. 2, pp. 299–318, 1967.

[19] D.Lee, and H.S.Seung, "Algorithms for non-negative matrix factorization," *IAdvances in neural information processing systems*, pp. 556–562, 2001.

[20] I. Higgins, *et al*, "β-vae: Learning basic visual concepts with a constrained variational framework," *International conference on learning representations*, 2016.

[21] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *EEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 236–243, 1984.

[22] Shinnosuke Tkaamichi, Yuki Saito, Norihiro TakamuneDaichi KitamuraHiroshi Saruwatari, "Phase Recovery from Amplitude Spectrograms Based on von Mises Distribution DNN," *IPSJ Research Report*, vol. 54, pp. 1–6, 2012.

[23] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *arXiv preprint arXiv:1609.03499*, 2016.

[24] J. K. et al., "HiFi-GAN: generative adver- sarial networks for efficient and high fidelity speech synthesis," *in Proc. NeurIPS*, pp. 17 022–17 033, 2020.

[25] S. Y. Genichiro Kikui Toshiyuki Takezawa, "Multilingual Corpora for Speech-to-Speech Translation Research," *ICSLP*, vol. 2, pp. 357–360, 2004.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Technical Report, FKI-207-95, Fakultat fur Informatik, Technische Universitat Munche*, pp. 1735–1780, 1997.

[27] Hochreiter,S. and Schmidhuber,J., "Deep Residual Learning for Image Recognition," *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 770–778, 2016.

[28] A.W.Rix, J.G.Beerends, M.P.Hollier, A.P.Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *Proc.ICASSP*, vol. 2, no. 2, pp. 749–752, 2001.

[29] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech.," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[30] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," *in Proc. INTERSPEECH*, pp. 4521–4525, 2022.

A method for classification NEO–FFI answers fabricated and advantageous due to psychological bias using brainwave specific brain activity networks

Yuto Ashikawa*, Takashi Ito, Shohei Ishizu and Yosuke Kurihara
* University of Aoyama Gakuin, Kanagawa, Japan
E-mail: c5624226@aoyama.jp Tel: +81-80-5899-9702
University of Aoyama Gakuin, Kanagawa, Japan
E-mail: ito@ise.aoyama.ac.jp Tel: +81-42-759-6423
University of Aoyama Gakuin, Kanagawa, Japan
E-mail: ishizu@ise.aoyama.ac.jp Tel: +81-44-953-1525
University of Aoyama Gakuin, Kanagawa, Japan
E-mail: kurihara@ise.aoyama.ac.jp Tel: +81-42-759-6371

Abstract— **The Neuroticism Extraversion Openness–Five Factor Inventory(NEO–FFI) is a psychological scale that assesses personality through five factors. Although widely used, it is susceptible to biases, which reduces reliability. This study proposes a method to classify biased answers using brainwave measured while answering NEO–FFI questions. Brainwave networks are constructed from alpha, beta, delta, theta, and gamma waves measured at eight brain regions (frontal, temporal, parietal, and occipital lobes). Fast Fourier transform and coherence are applied to obtain feature values for nodes and edges of the network. Support vector machines are then trained to classify answers as biased or unbiased based on these features for each personality factor. An experiment with 23 participants aged 19–23 reveals an accuracy of 0.67, precision of 0.72, negative predictive value of 0.62, recall of 0.71, specificity of 0.64, and F1 score of 0.70. This indicates that the proposed method effectively classifies biased answers and improves the reliability of NEO–FFI personality assessments.**

I. INTRODUCTION

The Neuroticism Extraversion Openness‐Five Factor Inventory(NEO–FFI) is a psychological scale developed by P.T. Costa and colleagues based on L.R. Goldberg's Big Five theory[1][2]. The theory measures personality in terms of five factors (N: neuroticism, E: extraversion, O: openness, A: agreeableness, and C: conscientiousness). Each factor is evaluated by 12 questions. The answers are rated from zero to four. The total score for each factor is used to assess personality. The NEO–FFI is widely used in psychological research to quantitatively evaluate personalities [3–7].

However, it has been demonstrated that respondents can fabricate their answers owing to psychological biases to present themselves favorably[8][9]. The inclusion of such fabricated answers may compromise the reliability of personality assessments. Therefore, classify fabricated answers in NEO–FFI items can enhance the interpretative reliability of personality evaluations.

Psychological biases are induced by emotional stimuli[10][11], which activate the brain activity[12–14]. Studies have revealed that the brain blood flow can predict the scores for personality questionnaires such as the state‐trait anxiety inventory[15][16]. Additionally, studies have analyzed brain network characteristics in answer to emotional stimuli[17].

Based on these observations, this study proposes a method to classify fabricated answers in the NEO–FFI by constructing brain activity networks using brainwave data from eight brain regions for alpha, beta, theta, delta, and gamma waves and applying an SVM trained on personality factors.

II. PROPOSED METHOD

Fig. 1 shows a schematic of the signal processing conducted in the proposed method.

If the set of personality factors targeted in the NEO–FFI is $F \in \{F_{\mathrm{N}}, F_{\mathrm{E}}, F_{\mathrm{O}}, F_{\mathrm{A}}, F_{\mathrm{C}}\}$, each of the 60 questions in the questionnaire belongs to one of the factors $i \in F$. If an arbitrary question is $q = (1, 2, \dots, 60)$, the brainwaves of the respondent while answering q are measured at brain region p. Let brain region p be a set of brain regions $p \in \{\mathrm{Fpz}, \mathrm{Fz}, \mathrm{Cz}, \mathrm{Pz}, \mathrm{T3}, \mathrm{T4}, \mathrm{O1}, \mathrm{O2}\}$ based on the international 10–20 method. Let $x_n(k)$ be the signal without hum noise obtained by applying a notch filter to brainwaves measured at arbitrary $n \in p$. $k(= 1, 2, 3, \dots, N_d)$ is discretized with the sampling interval Δt. Here, N_d is the number of data points. However, because the answer times differ for each question, N_d also differs.

Based on the measured $x_n(k)$ at each region, a brain activation network is constructed with region p as a node for each of the five types of brainwaves $B \in \{\alpha\text{ wave}, \beta\text{ wave}, \theta\text{ wave}, \delta\text{ wave}, \gamma\text{ wave}\}$. If the frequency

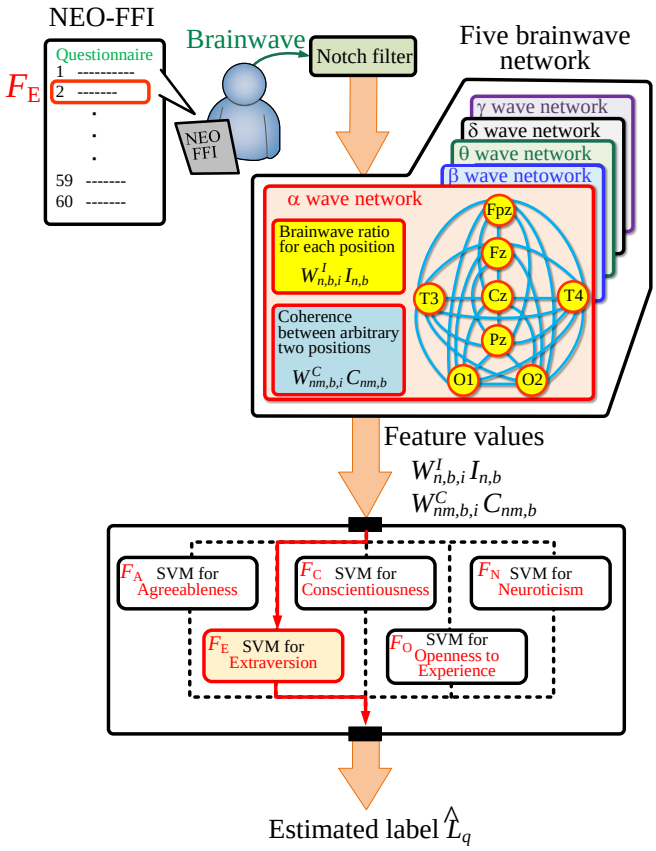


Fig.1 The signal processing flow to classify the answer is affected by psychological bias.

band of each brainwave is f_{low}– f_{high}, the frequency band of the brainwaves are as follows: α wave: 8–13 Hz, β wave: 14–29 Hz, θ wave: 4–7 Hz, δ wave: 0.5–3 Hz, and γ wave: 30 –70 Hz. A brain activation network for any brainwave $b \in B$ is defined by the value $I_{n,b}$ at a node and the value $C_{nm,b}$ of an edge between any two nodes n, $m \in p(m \neq n)$. The $I_{n,b}$ of the node at region n is calculated using Equation (1) for the amplitude spectrum $S_n(f)$ obtained by applying a fast Fourier transform to $x_n(k)$. Here, $f(=0, \Delta f, 2\Delta f, \ldots, f_N)$ is a discrete frequency, Δf is the frequency resolution, and f_N is the Nyquist frequency.

$$I_{n,b} = \sum_{f=\tilde{f}_{\text{low}}}^{\tilde{f}_{\text{high}}} S_n(f) \,/\, (\sum_{f=0}^{F_N} S_n(f)) \qquad (1)$$

Because f is a discrete frequency with frequency resolution Δf, the amplitude spectra corresponding to f_{low} and f_{high} in each brainwave is not necessarily obtained. Therefore, we let $\tilde{f}_{\text{low}}$ and $\tilde{f}_{\text{high}}$ be the discrete frequencies f closest to f_{low} and f_{high}, respectively, and calculate the amplitude spectrum. In addition, the edge value $C_{nm,b}$ is the average of the coherence values $C(f_{\text{low}}\text{:}\, f_{\text{high}})$ in the frequency– band f_{low}– f_{high} at the arbitrary region n and m. These are determined by Equation (2).

$$C(f_{\text{low}}\text{:}\, f_{\text{high}}) = \left|S_n(f_{\text{low}}\text{:}\, f_{\text{high}}) S_m(f_{\text{low}}\text{:}\, f_{\text{high}})\right|^2 \,/\, \left(\left|S_n(f_{\text{low}}\text{:}\, f_{\text{high}})\right|^2 \left|S_m(f_{\text{low}}\text{:}\, f_{\text{high}})\right|^2\right) \qquad (2)$$

In this proposed method, the degree of influence of factor i for $I_{n,b}$, $C_{nm,b}$ are $w^I_{n,b,i}$, $w^C_{nm,b,i} (= 0\text{–}1)$. And then, $w^I_{n,b,i} I_{n,b}$ and $w^C_{nm,b,i} C_{nm,b}$ are the feature value for the brain activation network in brainwave b.

Using an SVM, we classify whether the answer to question q was artificial from the features obtained from the network for all the brainwaves. In the learning phase of the SVM, N_s names are asked to answer the NEO–FFI. The questions answered honestly are labeled "0", and those answered intentionally are labeled "1". Here, the label assigned to the question is $L_q \in \{0, 1\}$. In the proposed method, for each personality factor i, the features are $w^I_{n,b,i} I_{n,b}$, $w^C_{nm,b,i} C_{nm,b}$. Moreover, the label constructs a learning model to determine L_q. Let M_i denote the SVM learning model for each factor i. In the classification phase of the SVM, the feature values $w^I_{n,b,i} I_{n,b}$, $w^C_{nm,b,i} C_{nm,b}$ and inputs it into the learning model M_i trained for personality factor i to determine the estimated classify label $\hat{L}_q$ for question q.

III. Verification Experiment

A. *Experimental Procedures to Obtain Dataset*

To validate the proposed method, a verification experiment was conducted with 23 participants aged 19–23 years.

The participants were asked to answer an NEO–FFI questionnaire containing 60 randomly arranged questions. The questions were displayed on a PC screen. The participants answered using a wireless keyboard and scored each question from zero to four. During this process, the participants were instructed to intentionally provide deceptive answers to questions wherein they wished to present themselves more favorably. This set of answers was labeled (A1).

To verify the proposed method, it was necessary to determine whether each question q in (A1) was answered genuinely or deceptively. Therefore, in this experiment, each participant was asked to answer another set of NEO–FFI questions in a manner similar to that for (A1), but honestly. These honest answers were labeled (A0).

For each question q, the difference in scores between (A0) and (A1) was used to determine the label L_q. For the factors F_{E}, F_{O}, F_{A}, and F_{C}, higher scores are generally considered more favorable. Meanwhile, for F_{N}, lower scores were considered more favorable [9]. Therefore, the absolute difference in scores between (A0) and (A1) was calculated. If this difference was equal to or larger than a threshold s, the answers to that question in (A1) was labeled deceptive ($L_q = 1$). If the difference was less than s, it was labeled as genuine ($L_q = 0$). The order of the (A0) and (A1) responses was randomized for

each participant. Additionally, during the (A1) answers, brainwave data from head position p was recorded using the polymate(AP1532) brainwave system from Miyuki Giken to extract feature values. This study was approved by the Ethics Review Committee of the Aoyama Gakuin University (Approval Number H23–025).

B. Hyper Parameters for SVM

In this validation experiment, the hyperparameters used for the SVM were of six types: the kernel function; degree d in the polynomial kernel; and parameters BC (adjusts the penalty for classification errors), G (adjusts the complexity of the decision boundary), C_0 (determines the importance of the negative class), and C_1 (determines the importance of the positive class).

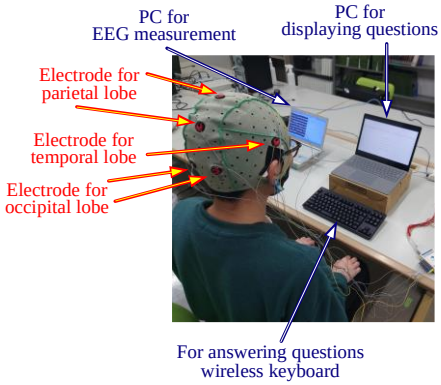


Fig.2 Experimental system.

C. Evaluation Method and Parameter Tuning by Genetic Algorithm (GA)

To evaluate the accuracy of the SVM–based classification, we performed leave–one–subject–out cross–validation using data from 22 of the 23 participants ($N_s = 22$) for training and the remaining participant's data for testing. We randomly selected data to balance the number of L_q labels (zero and one) five times in the training set and generated a confusion matrix for each test participant. This yielded 115 confusion matrices (23 participants × 5 times). We calculated the accuracy, Positive Predictive Value(PPV), Negative Predictive Value(NPV), sensitivity, specificity, and F1–score using Equations (3)–(8) and averaged these over 115 matrices for validation.

For the overall evaluation, a confusion matrix was constructed from the results of each question for each participant averaged over 115 repetitions. We optimized the weights $w^I_{n,b,i}$, $w^C_{nm,b,i}$; threshold s; kernel type (linear, RBF, polynomial); and parameters BC, G, d, C_0, C_1 using a GA. The range for $w^I_{n,b,i}$ and $w^C_{nm,b,i}$ was 0–1; that for BC, G, C_0, and C_1 was 0–3000; and that for d was 2–4. The GA fitness function is the product of the six performance metrics from Equations (3)–(8) with 10,000 generations and a number of individuals size of 40.

Table 1 Confusion matrix.

		Actual	
		Negative	Positive
Classification	Negative	TN	FN
	Positive	FP	TP

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN) \quad (3)$$

$$\text{PPV} = TP / (TP + FP) \quad (4)$$

$$\text{NPV} = TN / (TN + FN) \quad (5)$$

$$\text{recall} = TP / (TP + FN) \quad (6)$$

$$\text{specificity} = TN / (TN + FP) \quad (7)$$

$$\text{F1–score} = 2 \times \text{PPV} \times \text{recall} / (\text{PPV} + \text{recall}) \quad (8)$$

IV. Results

Table 1 presents the results of the verification experiments. It is evident from the table that for the factor N, the highest and lowest accuracy six performance indices were the PPV and NPV, respectively. This indicates that for this factor, the method cannot classify questions unaffected by psychological bias. Similarly, for the factors E, A, and C, the highest and lowest accuracy six performance indices were positive and negative recall, respectively. This indicates that the method cannot classify questions unaffected by psychological bias (similar to the factor N). Meanwhile, for the factor O, the highest and lowest accuracy indices were specificity and recall, respectively. This indicates that the method could not classify questions affected by psychological bias.

Overall, the highest and lowest accuracy indices were the PPV and NPV, respectively. This indicates that the method could not classify questions unaffected by psychological bias. The highest accuracy for each performance index was as follows: O, accuracy; N, PPV; O, NPV; A, recall; O, specificity; and N, F1–score. This indicates that O showed the highest accuracy for questions unaffected by psychological bias. Conversely, N and C showed high accuracies for questions affected by psychological bias, and A and E showed a balanced classification (see Table 2).

Table 2 Classification results by SVM.

	NEO–FFI measurement factors					
	N	E	O	A	C	whole
Accuracy	0.67	0.67	0.71	0.68	0.64	0.67
PPV	0.83	0.68	0.66	0.74	0.72	0.72
NPV	0.45	0.62	0.71	0.65	0.52	0.62
Recall	0.73	0.75	0.59	0.76	0.75	0.71
Specificity	0.55	0.61	0.78	0.61	0.46	0.64
F1–score	0.73	0.66	0.56	0.69	0.68	0.70

V. DISCUSSION

In the proposed method, we considered the causes of misclassification. For O, 15.1% of the questions in conditions (A0) and (A1) had an identical upper limit of four scores. Because the answers are on a five–score scale, those affected by psychological bias cannot select more than four scores. Thus, these fail to reflect the bias. This may cause misclassification.

Moreover, the balanced classification of A and E can be attributed to respondents attempting to appear favorable by considering their relationships with others (such as harmony and communication), which yields balanced results for these factors. In contrast, other factors focused more on self–perception, which may have resulted in a biased classification.

Additionally, the influence of psychological bias does not manifest immediately after starting the questions but appears after a certain time. Therefore, it is necessary to construct a brain activity network that considers the time–dependent effects. The current network does not account for these temporal variations.

VI. CONCLUSION

In this study, we proposed a method for classify fabricated answers caused by psychological bias in NEO–FFI answers. This method uses brain activity networks characterized by five types of brainwave signals for each personality factor. The validation experiments with 23 participants revealed the feasibility of classify the fabricated answers with an accuracy of 0.67, PPV of 0.72, NPV of 0.62, recall of 0.71, specificity of 0.64, and F1–score of 0.70.

Brain activity networks were constructed using the brainwave spectral content and average of coherence values from different brain regions to classify fabricated answers owing to psychological bias. Future work should consider the time required for participants to be influenced by psychological biases and incorporate response time into the features for a more accurate classification method.

REFERENCES

[1] D. Peabody and L. R. Goldberg, “Some determinants of factor structures from personality–trait descriptors,” *Journal of Personality and Social Psychology*, Vol. 57, No. 3, p.552, 1989.

[2] P. T. Costa and R. R. McCrae, “Neo personality inventory–revised (NEO PI–R),” *Psychological Assessment Resources*, 1992.

[3] R. Riemann, A. Angleitner, and J. Strelau, “Genetic and environmental influences on personality: A study of twins reared together using the self‐and peer report NEO‐FFI scales,” *Journal of Personality*, Vol. 65, No. 3, pp.449-475, 1997.

[4] N. Archer, R. G. Brown, H. Boothby, C. Foy, H. Nicholas, and S. Lovestone, “The NEO–FFI is a reliable measure of premorbid personality in patients with probable Alzheimer's disease,” *International Journal of Geriatric Psychiatry: A Journal of the Psychiatry of Late Life and Allied Sciences*, Vol. 21, No. 5, pp.477-484, 2006.

[5] M. Z. Podolska, M. Bidzan, M. Majkowicz, J. Podolski, O. S. Szmigiel, and E. R. Walknowska, “Personality traits assessed by the NEO Five–Factor Inventory (NEO–FFI) as part of the perinatal depression screening program,” *Medical Science Monitor*, Vol. 16, No. 9, pp.77-81, 2010.

[6] K. Lee and M. C. Ashton, “Prediction of self–and observer report scores on HEXACO–60 and NEO–FFI scales,” *Journal of Research in Personality*, Vol. 47, No. 5, pp.668-675, 2013.

[7] S. Sharma, “The impact of neuroticism and conscientiousness on sleep quality: exploring personality traits and sleep disturbances in young adults,” *International Journal of Interdisciplinary Approaches in Psychology*, Vol. 2, No. 5, pp.1515-1536, 2024.

[8] R. R. Holden and C. E. Lambert, “Response latencies are alive and well for identifying fakers on a self–report personality inventory: A reconsideration of van Hooft and Born 2012,” *Behavior Research Methods*, Vol. 47, pp.1436-1442, 2015.

[9] S. Horio and K. Takahashi, “Roles of social desirability scales of the Big–Five Personality Inventory in faking settings,” *Japanese Association of Industrial/Organization Psychology Journal*, Vol. 17, pp.65–77, 2004.

[10] G. Ceschi, S. Meylan, C. Rowe, and A. H. Boudoukha, “Psychological profile, emotion regulation, and aggression in police applicants: A Swiss cross-sectional study,” *Journal of Police and Criminal Psychology*, Vol. 37, pp.962–971, 2022.

[11] I. Krumpal, “Social desirability bias and context in sensitive surveys,” *Encyclopedia of quality of life and well–being research*, Cham: Springer International Publishing, pp.6527–6532, 2024.

[12] L. Xia, et al. "The brain activation of anxiety disorders during emotional stimulations: A coordinate–based activation likelihood estimation meta–analysis," 2021.

[13] K. Juliane, et al. "The effect of emotional content on brain activation and the late positive potential in a word n–back task." *PloS one*, Vol. 8, No. 9, 2013: e75598.

[14] N. A. Puccetti, et al. “Linking amygdala persistence to real–world emotional experience and psychological well–being,” *Journal of Neuroscience 21*, Vol. 41, No. 16, pp.3721-3730, April 2021.

[15] K. Hasegawa, Y. Hamada, and Y. Kurihara, “Using granger causal test associated with trait anxiety to estimate STAI scores by constructing a brain activity independence network,” The 11th IIAE International Conference on Industrial Application Engineering 2023, 2023.

[16] K. Hasegawa, Y. Hamada, and Y. Kurihara, “Verifying the effects of ceberal blood flow pulse component and dependance probability of brain activity in the prefrontal cortex on empathy estimation accuracy,” The SICE Annual Conference 2022, 2022.

[17] A. Muramatsu, S. Kobayashi, and Y. Mizuno–Matsumoto, “Complex network analysis of electroencephalography elicited by emotional stimuli presented on the smartphone,” *Journal of Affective Engineering*, Vol. 18, No. 4, pp.263-271, 2019.

Optimizing Multi-Speaker Speech Recognition with Online Decoding and Data Augmentation

Yizhou Peng*† and Eng Siong Chng*
* Nanyang Technological University, Singapore
E-mail: peng.yizhou@ntu.edu.sg
† National University of Singapore, Singapore

***Abstract*—This paper addresses missing transcription of short speeches from changed speakers in streaming ASR. This problem may be attributed to two factors: existing training data consists of only a single speaker in each training utterance, and the transformer memory of a single speaker during streaming decoding may impact the decoding of segments from the new speaker. To improve, we propose to leverage limited left-context streaming decoding and to include data augmentation of multi-speaker in training utterances. Our experimental results on Librispeech, Aishell-1, and SEAME corpus demonstrate 32%, 28%, and 9% relative improvements in word error rates (WER), character error rates (CER), and mixed error rates (MER) across English, Chinese, and code-switching datasets. These findings suggest robust solutions for real-time ASR applications in complex audio environments.**

I. Introduction

Since the advent of large-scale models like GPT [1] in natural language processing (NLP) and Whisper [2] in speech recognition as well as speech translation, there has been a significant surge in research aimed at leveraging Large Models to address a variety of tasks [3], [4]. In the meantime, to utilize abundant information in speech signals, SpeechGen [5] explores the application of prompt tuning to enhance the generative capabilities of Speech Language Models for various tasks. Also, the Qwen-Audio [6] model creates a universal audio-language model capable of processing diverse audio types, from human speech to natural sounds and music. These models have shown remarkable capabilities in understanding and generating human language. The final goal is to develop a single model that can seamlessly handle various tasks while achieving state-of-the-art performance for each mission.

However, despite the impressive advancements of these large-scale models, they often fall short of achieving the best performance on specific tasks when compared to specialized models with a similar number of parameters. This highlights that specialized systems are still critical for most industry applications.

In the area of ASR, the capability to handle as many scenarios as possible, as well as computational efficiency, are two crucial factors when it comes to real applications. Specifically, traditional reading-style single-speaker speech recognition systems are no longer sufficient to meet the demands of modern applications because of the rapid development of the Internet, and the increasing prevalence of virtual meetings that cause continuous conversation with multiple speakers has become increasingly common. These scenarios require ASR systems to accurately transcribe conversations involving multiple participants, sometimes speaking simultaneously, which makes it more complex than that in single-speaker contexts [7]. Studies like those on streaming speaker-attributed ASR [7] and Qwen-Audio-Chat, which integrates audio and text inputs for multi-turn dialogues, emphasize the importance of these developments [6].

In addition to handling conversational and multi-speaker scenarios, there is another type of ASR system that is designed to process and transcribe speech in real-time, making it indispensable for applications that require immediate feedback, such as live captioning, virtual assistants, and real-time communication platforms, namely streaming ASR. The need for low-latency, high-accuracy ASR solutions has driven significant research and development efforts in this area, such as RNN-Transducer [8], [9], U2++ Conformer [10], Zipformer-Transducer [11], even zero-shot streaming implementation for Whisper model [12].

In this paper, we focus on dealing with scenarios containing multiple speakers in a single decoding segment in the streaming ASR task. Specifically, we found that for single-speaker ASR systems, the transcription for the changed speakers could be partially or totally missing when performing offline decoding on sentences containing rapid speaker changes. We propose that online streaming decoding can alleviate these issues to different extents with diverse configurations without requiring any model fine-tuning. Also, experimental results show that the multi-speaker data augmentation method [13] with our refinement strategy further improves the model performance on both single-speaker and multi-speaker testsets.

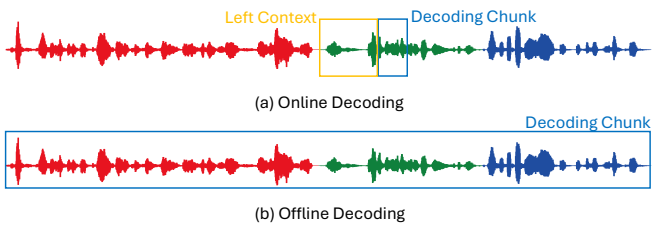


Fig. 1. Comparison between Offline and Online decoding for a sentence

This paper is organized as follows. Section I shows the background of this topic. Section II illustrates related works

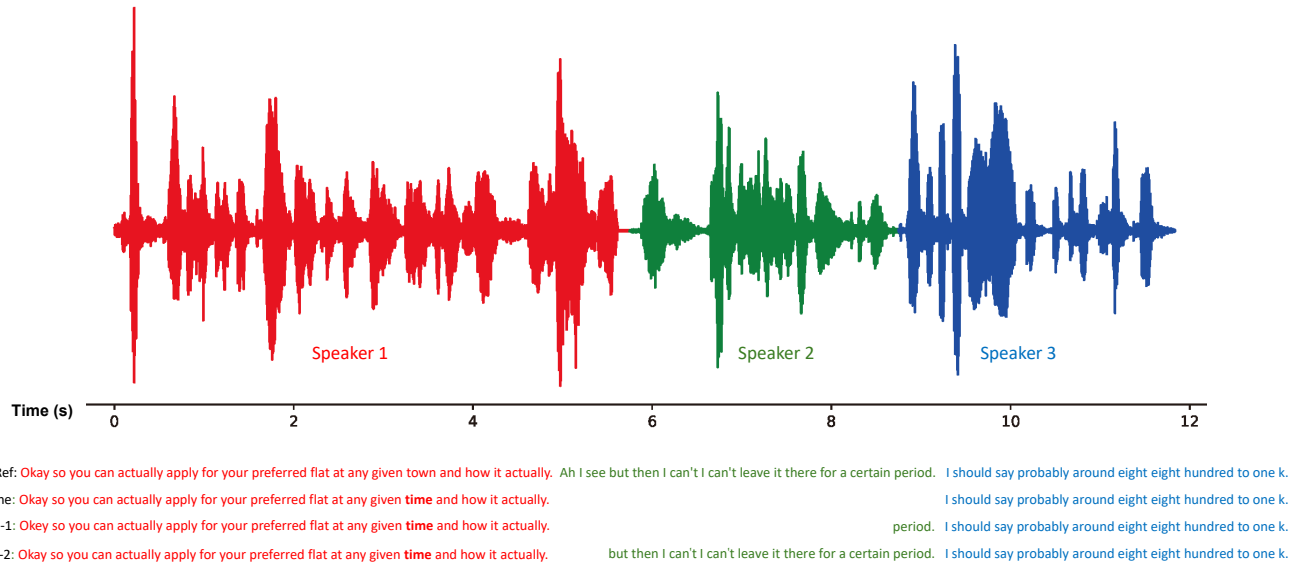


Fig. 2. An inference example of a multi-speaker speech segment using single-speaker ASR model, where the silence gap between each speaker is less than 0.5 seconds. **Ref** stands for reference text that is manually transcribed, **Offline** means the model sees the entire sentence when performing decoding. **Online-1** and **Online-2** perform online stream decoding whose decoding chunksize is 640ms, where **Online-1** uses 2.56s left context while **Online-2** sees only 1.28s.

to multi-speaker ASR and streaming ASR. Section III then includes methods, corpus information, model architectures, and training configurations. Succeedingly is section IV for experimental results and analysis. Finally, section V concludes the work.

II. Related Works

Multi-speaker speech recognition has become an essential focus area with the increasing prevalence of conversational and meeting scenarios where multiple participants may speak simultaneously. A model using token-level serialized output training (t-SOT) [14] to accurately identify and transcribe speech from multiple speakers in real-time is proposed [15], which significantly improves performance in overlapping speech scenarios. Additionally, a mixture encoder for joint speech separation and recognition [16] was proposed, leveraging explicit speech separation while incorporating cross-speaker context information to mitigate error propagation, achieving significant improvements.

Streaming ASR is critical for real-time transcription applications like live captioning and virtual assistants. Introducing cumulative attention mechanisms in streaming Transformers reduces latency while maintaining high accuracy by synchronizing attention heads within the model [17]. MiniStreamer framework [18] enhances the Conformer model with chunked-context masking, optimizing it for edge devices with limited computational resources. These advancements illustrate the ongoing efforts to improve the efficiency and effectiveness of streaming ASR systems in various real-world applications.

III. Experiments

A. Methodologies

1) Limited Left-Context Streaming Decoding: In streaming ASR, chunk-size and left-context are critical parameters for balancing latency and accuracy during decoding. Fig. 1 (a) illustrates an example of online stream decoding where the Decoding Chunk is audio frames to form a specific chunk-size of the chunk, and the Left Context involves the amount of previous audio data considered when processing the current Decoding Chunk. Chunk-size refers to the length of audio data processed at one time, with smaller chunks reducing latency but potentially decreasing accuracy due to less context. In comparison, larger chunks increase accuracy at the cost of higher latency. To simplify, Fig. 1 (b) indicates that Offline Decoding includes the whole speech segment into one Decoding Chunk.

However, in multi-speaker scenarios, accuracy is not necessarily improved when given more left context. On the contrary, when the model sees less left context, it produces more accurate recognition results, even compared with the Offline decoding method, which is generally believed to generate the best annotations. As shown in Fig. 2, for a single-speaker ASR model using Zipformer architecture [11] with a causal input layer, the Offline and Online-1 decoding methods produce results that almost lose all content from **Speaker 2**. In contrast, the Online-2 decoding method only misses a few words. This suggests that for a trained single-speaker causal model, decoding with limited left context in streaming style can resolve the multi-speaker issues in some respect.

2) Multi-Speaker Data Augmentation: In [13], simply and randomly combining several speakers' speech segments significantly improves the model performance on multi-speaker speech recognition while maintaining unharmed single-speaker

recognition accuracy. Technically, this augmentation method works on-the-fly after speech features are extracted for batches during training, combining speech features of segments in one batch from two or three different speakers to form an augmented new batch. For transcriptions in the augmented training batch, it would simply add <SC> label, which stands for Speaker-Change, in between the transcription of combined segments while remaining unchanged for those not chosen for augmentation. Fig. 3 shows how the augmented annotation formed by given segments from three different speakers. However, to perform an on-the-fly style of augmentation, the combination can only occur in each batch during training, which might significantly reduce the diversity of combinations between different speakers.

To solve the above issue, we simplify the augment process by generating the augmented dataset before training and randomly combining speakers from the whole training set for each epoch, which maximizes the diversity of speaker combinations.

RAW-1: Okay so you can actually apply for your preferred flat at any given town and how it actually
RAW-2: Ah I see but then I can't I can't leave it there for a certain period
RAW-3: I should say probably around eight eight hundred to one k
AUG: Okay so you can actually apply for your preferred flat at any given town and how it actually **<SC>** Ah I see but then I can't I can't leave it there for a certain period **<SC>** I should say probably around eight eight hundred to one k.

Fig. 3. An augmentation example of a multi-speaker speech segment. **RAW-{1,2,3}** stand for three speech segments from three different speakers. **AUG** means the resulting augmented segment that combines the three separate utterances with <**SC**> label in between where <**SC**> stands for Speaker-Change

B. Data

We perform our experiments on diverse datasets in languages including English, Chinese, and Code-switching scenarios to show the effectiveness of our methods. Correspondingly, we choose Librispeech [19], Aishell-1 [20], and SEAME [21] datasets for each language setting since these datasets are the most widely used for each language for research purposes in the last few years. Librispeech is a large-scale corpus of approximately 1,000 hours of English speech derived from audiobook recordings, containing read speech recorded by native speakers from the United States. Aishell-1 is an open-source Mandarin Chinese speech corpus consisting of about 178 hours of speech data recorded from 400 speakers with various accents across China. The data covers several topics, such as news reports and conversational content, and it is recorded in reading style. SEAME (South East Asia Mandarin-English) is a conversational speech corpus featuring code-switching between Mandarin and English, collected from naturally occurring conversations and interviews in Singapore and Malaysia, in both formal and informal settings. Table I shows the statistics information for each dataset.

For Librispeech, we only report results on **Test-C**lean and **Test-O**ther. Dev_{Man} and Dev_{Sge} are two official subsets for evaluation purposes of the SEAME corpus. All sets with **-M** label stand for simulated Speaker-Mixed sets, consisting of two or three speakers in one segment, keeping the exact same duration and number of words as the original sets, following the multi-speaker data augmentation method mentioned in section III-A2.

TABLE I
OVERALL SPEECH DATA DISTRIBUTION FOR BOTH ASR MODEL TRAINING AND TESTING. **EN** STANDS FOR ENGLISH WHILE **CN** MEANS CHINESE. **EN-CN CS** CORRESPONDS TO ENGLISH CHINESE CODE-SWITCHING. **TRAIN-S** MEANS SINGLE-SPEAKER TRAINING SET WHILE **TRAIN-M** STANDS FOR MULTI-SPEAKER TRAINING SET.

Corpus	Language	Subset	Duration(Hrs)
Librispeech	EN	Train-S / Train-M	961.1
		Test-C / Test-C-M	5.4
		Test-O / Test-O-M	5.3
Aishell-1	CN	Train-S / Train-M	150.9
		Dev / Dev-M	18.1
		Test / Test-M	10.0
SEAME	EN-CN CS	Train-S / Train-M	93.6
		Dev_{Man} / Dev_{Man}-M	7.5
		Dev_{Sge} / Dev_{Sge}-M	3.9

C. Model

We perform all of our experiments using zipformer-transducer architecture, with icefall[1] toolkit that is supported by K2[2] project. The model architecture involves a zipformer [11] encoder, a stateless transducer [22] decoder, and a simple joiner network. The encoder is configured with layers of 2, 2, 3, 4, 3, and 2, each having subsampling factors of 1, 2, 4, 8, 4, and 2, respectively, following the official configuration of zipformer recipe. The input feature is 80-dim Mel frequency bins, computed on 25-ms windows with a stride of 10 ms. SpecAugment [23] is enabled for all experiments, while speed-perturb is disabled, as well as mix precision training is also disabled. We use 4000 secs, 2000 secs, and 800 secs as model batch sizes for Librispeech, Aishell-1, and SEAME experiments, respectively. We use BPE [24] as modeling units for English while using characters for Chinese, which results in 500 BPEs for Librispeech, 4338 units for Aishell-1, and 3136 units for SEAME corpus. The context size is set to one for Aishell-1 and SEAME, while it remains 2 for Librispeech. The learning rate scheduler and optimizer use the same configuration proposed in zipformer [11], named Eden LRScheduler and ScaledAdam optimizer, which achieve faster convergence and better performance than Adam and significantly reduce warmup steps during training. Our models are trained with a maximum of 8x A100-40GB GPUs according to different batch size configurations for each corpus.

The single-speaker models were trained for 40 epochs, 60 epochs, and 60 epochs for the Librispeech, Aishell-1, and SEAME corpus, respectively. For multi-speaker, the models were trained with only half of the epochs for single-speaker

[1]https://github.com/k2-fsa/icefall
[2]https://github.com/k2-fsa/k2

TABLE II
WER/CER/MER RESULTS FOR MODELS TRAINED WITH **TRAIN-S** SETS. **OFFLINE-S** AND **ONLINE-S** STAND FOR OFFLINE AND ONLINE DECODING FOR SINGLE-SPEAKER MODELS. **INF** SPECIFIED IN CHUNK-SIZE AND LEFT-CONTEXT MEANS UNLIMITED DECODING CHUNK. **BOLD** NUMBERS ARE THE BEST RESULTS FOR EACH TESTSET.

Decoding Configs			Librispeech				Aishell-1				SEAME			
Methods	Chunk-size	Left-Context	Test-C	Test-O	Test-C-M	Test-O-M	Dev	Test	Dev-M	Test-M	Dev_{Man}	Dev_{Sge}	Dev_{Man}-M	Dev_{Sge}-M
Offline-S	inf	inf	**2.74**	**6.44**	**3.27**	9.97	**5.08**	**5.39**	6.64	7.24	**16.55**	**23.68**	18.78	**25.74**
Online-S	320 ms	1280 ms	3.66	9.57	3.74	10.23	5.90	6.46	6.03	6.67	19.09	26.97	20.04	28.17
	320 ms	2560 ms	3.55	9.17	3.84	11.14	5.85	6.31	6.30	7.31	18.89	26.83	20.32	28.58
	320 ms	5120 ms	3.52	9.07	4.11	12.24	5.85	6.30	6.70	8.33	18.79	26.83	20.74	28.93
	640 ms	1280 ms	3.40	8.59	3.44	**9.20**	5.57	6.11	**5.72**	**6.34**	17.94	25.37	**18.69**	26.54
	640 ms	2560 ms	3.29	8.29	3.47	9.91	5.5	6.04	5.92	6.82	17.84	25.26	18.95	26.81
	640 ms	5120 ms	3.28	8.18	3.69	11.08	5.5	6.01	6.27	7.63	17.82	25.19	19.47	27.13

because we combined the Train-S and Train-M to form the final training dataset, keeping similar updating steps for the single and multi-speaker models.

IV. EXPERIMENTAL RESULTS

The experimental results mainly consist of two parts: single-speaker and multi-speaker. Each model is tested under both single-speaker and multi-speaker scenarios with offline and online streaming decoding configurations. Word Error Rate (WER) is reported for Librispeech, Character Error Rate (CER) is reported for Aishell-1, and Mix Error Rate (MER) where CER is calculated for Chinese parts and WER is calculated for English parts, is reported for SEAME.

A. Single-Speaker Models

In this section, all of our models are trained with **Train-S** set of each corpus. Table II shows the results of these models on the corresponding testsets. In Table II, we can conclude that for a single-speaker ASR model, testing on single-speaker testsets would always obtain the best recognition results, either decoding with offline or online settings, compared with multi-speaker testsets. It also shows significant performance gaps in the offline decoding config between corresponding multi-speaker and single-speaker testsets, mainly observed from Test-Other of Librispeech, Test of Aishell-1, and Dev_{Man} of SEAME, where the gaps reach the range of (13%, 55%) in relative, respectively.

From online decoding results, it is observed that a larger decoding chunk-size gives better WER results regardless of left-context size given for all situations in our experiments. However, left-context size shows a reversed influence on single-speaker versus multi-speaker testsets. Specifically, when changing left-context configurations for single-speaker testsets, it shows reasonably better performance each time scaling up. In contrast, in multi-speaker scenarios, the models produce much worse recognition results, which is around 10% relative performance degradation for each time left-context size scaling up. We even find that most of the best results for multi-speaker testsets come from chunk-size of 640 ms and left-context size of 1280 ms, such as Test-O-M of Librispeech, Dev-M and Test-M from Aishell-1, and Dev_{Man}-M from SEAME. In this configuration, the performance gap between single-speaker and multi-speaker testsets is significantly reduced to a range of (4%, 8%). The results show consistency with the situation we present in Fig. 2 that indicates single-speaker ASR systems produce better recognition results when given less left-context for multi-speaker speech data.

TABLE III
WER RESULTS FOR MODELS TRAINED WITH THE COMBINATION OF **TRAIN-S** AND **TRAIN-M** SETS FOR LIBRISPEECH. **SINGLE** AUGMENT CONFIG STANDS FOR THE MODELS TRAINED WITH THE SAME COPY OF **TRAIN-M** FOR EACH EPOCH, WHILE **MULTIPLE** MEANS THAT IN EACH EPOCH OF TRAINING, THE MODELS USE DIFFERENT **TRAIN-M** GENERATED SEPARATELY. **OFFLINE-M** AND **ONLINE-M** MEAN OFFLINE AND ONLINE DECODING FOR MULTI-SPEAKER MODELS. **CHUNK-SIZE** FOR ONLINE DECODING IS FIXED TO 640 MS HERE.

Configs			Librispeech			
Augment	Methods	Left-Context	Test-C	Test-O	Test-C-M	Test-O-M
Single	Offline-M	inf	2.72	6.60	2.80	6.76
	Online-M	1280 ms	3.37	8.53	3.39	8.68
		2560 ms	3.27	8.32	3.27	8.47
		5120 ms	3.28	8.20	3.28	8.44
Multiple	Offline-M	inf	2.73	6.59	2.79	6.74
	Online-M	1280 ms	3.32	8.47	3.34	8.49
		2560 ms	3.23	8.30	3.24	8.40
		5120 ms	3.17	8.24	3.20	8.28

B. Multi-Speaker Models

In this section, all of our models are trained with the combination of **Train-S** and **Train-M** sets. To begin with, we propose two augmentation strategies for **Train-M** generation for Librispeech corpus, **Single** augment means that we only generate one set of **Train-M** for all epochs, while **Multiple** means that we re-generate **Train-M** for each epoch, aiming to obtain more diverse speaker combination cases. Table III shows the performance of the resulting models. No noticeable difference is observed in comparing the offline decoding performance between the **Single** and **Multiple** augmentation configurations. However, when it comes to online streaming decoding, **Multiple** augmentation shows some improvement among most of the testsets and left-context configurations. This indeed suggests that a diversity of combinations of speakers

TABLE IV
WER/CER/MER RESULTS FOR MODELS TRAINED WITH THE COMBINATION OF **TRAIN-S** AND **TRAIN-M** SETS, USING **MULTIPLE TRAIN-M** AUGMENT CONFIG. **OFFLINE-S** STANDS FOR OFFLINE DECODING FOR SINGLE-SPEAKER MODELS, WHILE **OFFLINE-M** AND **ONLINE-M** CORRESPOND TO OFFLINE AND ONLINE DECODING FOR MULTI-SPEAKER MODELS. **INF** SPECIFIED IN CHUNK-SIZE AND LEFT-CONTEXT MEANS UNLIMITED DECODING CHUNK.

Decoding Configs			Librispeech				Aishell-1				SEAME			
Methods	Chunk-size	Left-Context	Test-C	Test-O	Test-C-M	Test-O-M	Dev	Test	Dev-M	Test-M	Dev_{Man}	Dev_{Sge}	Dev_{Man}-M	Dev_{Sge}-M
Offline-S	inf	inf	2.74	6.44	3.27	9.97	5.08	5.39	6.64	7.24	16.55	23.68	18.78	25.74
Offline-M	inf	inf	2.73	6.59	2.79	6.74	4.80	5.17	4.81	5.24	16.63	23.53	16.78	23.52
Online-M	320 ms	1280 ms	3.58	9.23	3.65	9.45	5.54	5.96	5.48	5.96	18.55	25.96	18.60	26.15
	320 ms	2560 ms	3.52	9.05	3.53	9.20	5.40	5.89	5.40	5.87	18.14	25.84	18.56	25.98
	320 ms	5120 ms	3.49	8.91	3.51	9.08	5.38	5.87	5.38	5.91	18.34	25.80	18.45	25.95
	640 ms	1280 ms	3.32	8.47	3.34	8.49	5.18	5.68	5.19	5.69	17.70	24.79	17.73	24.74
	640 ms	2560 ms	3.23	8.30	3.24	8.40	5.13	5.60	5.15	5.62	17.57	24.68	17.57	24.67
	640 ms	5120 ms	3.17	8.24	3.20	8.28	5.13	5.58	5.16	5.65	17.55	24.68	17.57	24.68

generalizes the models more effectively. Consequently, **Multiple** augmentation is applied in our following experiments.

Table IV shows the performance for our multi-speaker ASR models for each corpus using **Multiple** data augmentation configuration. First of all, when comparing offline decoding results between single and multi-speaker models, significantly improved results for multi-speaker testsets are observed. In the meantime, improvements were also made regarding some of the single-speaker testsets. Specifically, Test-Clean of Librispeech, Dev, and Test of Aishell-1, and Dev_{Sge} of SEAME show improved results. In contrast, the results of Test-Other of Librispeech and Dev_{Man} of SEAME become a bit worse but still acceptable. Another interesting finding is that when it comes to online streaming decoding, consistent performance is found for all test sets and decoding configurations; that is, when decoding chunk size and left-context size are expanded wider, better recognition accuracy is obtained for not only single-speaker testsets but also multi-speaker counterparts. Finally, the results of multi-speaker testsets compared with single-speaker counterparts are now similar among all decoding configurations using the multi-speaker ASR models compared against single-speaker models shown in table II.

The results suggest that when models are exposed to varying conditions and combinations of different speakers, the effects of streaming decoding configurations on multi-speaker testsets are consistent with those observed for single-speaker testsets. This consistency indicates that the improvements in recognition accuracy with larger decoding chunk sizes and expanded left-context sizes apply similarly to both single and multi-speaker scenarios, highlighting the robustness of the model in handling diverse and complex audio.

V. CONCLUSION

This study demonstrates effective strategies to improve single-speaker ASR systems in multi-speaker scenarios through online streaming decoding and multi-speaker data augmentation. By optimizing left-context configurations and employing diverse speaker combinations, we achieved significant improvements in speech recognition accuracy. Our methods offer practical solutions for real-time ASR applications, enhancing their ability to handle complex and dynamic audio environments. The results underscore the importance of specialized configurations and data augmentation in advancing ASR technology to meet the demands of modern communication systems. Future work will focus on further refining these techniques for tasks such as online speaker diarization and voice activity detection, together with speech recognition function.

VI. ACKNOWLEDGEMENTS

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore, and the National Research Foundation Singapore under its AI Singapore Programme grant number AISG2-100E-2023-105. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore.

The computational work for this article was partially performed on resources of the National SuperComputing Centre (NSCC), Singapore.

REFERENCES

[1] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI*, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *OpenAI*, 2022. [Online]. Available: https://cdn.openai.com/papers/whisper.pdf.

[3] J. Wu, X. Hu, Y. Wang, B. Pang, and R. Soricut, "Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts," *arXiv preprint arXiv: 2312.00968*, 2023.

[4] P. K. Rubenstein, C. Asawaroengchai, D. Nguyen, *et al.*, "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv: 2306.12925*, 2023.

[5] H. Wu, K.-W. Chang, Y.-K. Wu, and H.-y. Lee, "Speechgen: Unlocking the generative power of speech language models with prompts," *arXiv preprint arXiv: 2306.02207*, 2023.

[6] Y. Chu, J. Xu, X. Zhou, *et al.*, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv: 2311.07919*, 2023.

[7] N. Kanda, J. Wu, Y. Wu, *et al.*, "Streaming Speaker-Attributed ASR with Token-Level Speaker Embeddings," in *Proc. INTERSPEECH 2022*.

[8] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," in *Proc. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

[9] S. Wang, P. Zhou, W. Chen, J. Jia, and L. Xie, "Exploring rnn-transducer for chinese speech recognition," in *Proc. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

[10] D. Wu, B. Zhang, C. Yang, *et al.*, *U2++: Unified two-pass bidirectional end-to-end model for speech recognition*, 2021. arXiv: 2106.05642.

[11] Z. Yao, L. Guo, X. Yang, *et al.*, "Zipformer: A faster and better encoder for automatic speech recognition," in *Proc. The Twelfth International Conference on Learning Representations*, 2023.

[12] D. Macháček, R. Dabre, and O. Bojar, "Turning whisper into real-time transcription system," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, Nov. 2023.

[13] S. Thomas, H. Kuo, G. Saon, and B. Kingsbury, "Multi-speaker data augmentation for improved end-to-end automatic speech recognition," *Proc. ICASSP*, 2023.

[14] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized Output Training for End-to-End Overlapped Speech Recognition," in *Proc. Interspeech 2020*.

[15] N. Kanda, J. Wu, Y. Wu, *et al.*, "Streaming Multi-Talker ASR with Token-Level Serialized Output Training," in *Proc. Interspeech 2022*.

[16] S. Berger, P. Vieting, C. Boeddeker, R. Schlüter, and R. Haeb-Umbach, "Mixture Encoder for Joint Speech Separation and Recognition," in *Proc. INTERSPEECH 2023*.

[17] M. Li, S. Zhang, C. Zorila, and R. Doddipatla, "Transformer-based streaming asr with cumulative attention," *Proc. ICASSP*, 2022.

[18] H. Gulzar, M. R. Busto, T. Eda, K. Itoyama, and K. Nakadai, "Ministreamer: Enhancing small conformer with chunked-context masking for streaming asr applications on the edge," *INTERSPEECH 2023*, 2023. DOI: 10.21437/interspeech.2023-1162.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *Proc. ICASSP*, 2015.

[20] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017.

[21] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "Mandarin–english code-switching speech corpus in south-east asia: Seame," *Language Resources and Evaluation*, 2015.

[22] M. Ghodsi, X. Liu, J. Apfel, R. Cabrera, and E. Weinstein, "Rnn-transducer with stateless prediction network," in *Proc. ICASSP*, 2020.

[23] D. S. Park, W. Chan, Y. Zhang, *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. INTERSPEECH 2019*, 2019.

[24] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

Adapting OpenAI's Whisper for Speech Recognition on Code-Switch Mandarin-English SEAME and ASRU2019 Datasets

Yuhang Yang*, Yizhou Peng†§, Hao Huang‡, Eng Siong Chng§ and Xionghu Zhong*¶
* Hunan University, China
¶ Corresponding Author. Email: xzhong@hnu.edu.cn
† National University of Singapore, Singapore
‡ School of Computer Science and Engineering, Xinjiang University, China
§ Nanyang Technological University, Singapore

***Abstract*—This paper reports on SOTA results achieved using openAI's Whisper model with adaptation on different adaptation corpus sizes for two established code-switch Mandarin/English corpus - namely SEAME and ASRU2019 corpora.**

Two key experiments were conducted: a) using adaptation data from 1 to 100/200 hours to demonstrate the effectiveness of adaptation, b) examining different language ID setups on Whisper prompt. The Mixed Error Rate results show that the amount of adaptation data may be as low as $1 \sim 10$ hours to achieve saturation in performance gain (SEAME), while the ASRU task continued to show performance with more adaptation data (>100 hours). For the language prompt without adaptation, the results show that various prompting strategies produce different outcomes. However, after adaptation, the Whisper model uniformly improves its performance, and the language prompt becomes less critical. We believe that these results can help researchers study the adaptation of Whisper to other code-switch Languages.

I. Introduction

Code-switching is a pervasive linguistic phenomenon in multilingual communities, where speakers alternate between two or more languages within a single speech or conversation. The prevalence of code-switching poses a unique challenge for Automatic Speech Recognition (ASR) systems [1], [2], as it requires the model to have a nuanced understanding of multiple languages simultaneously. Unfortunately, current ASR models often underperform in recognizing code-switching speech due to the scarcity of labeled training data.

In the past few years, the research community has introduced several approaches to tackle the challenges posed by code-switching ASR systems. These approaches can be broadly categorized into three technical aspects: speech, text, and modeling methods. From the speech perspective, strategies have been developed to implement monolingual speech into code-switching ASR systems [3]. Additionally, some researchers propose augmenting pronunciation models to accommodate accents, mispronunciations, and pronunciation variations, thereby addressing the issue of data sparsity [4]. On the text front, multiple techniques have been explored, ranging from augmenting code-switching text from monolingual corpora to build language models (LM) [5]–[8], to employing methods like speech T5 [9], [10], multilingual word-embedding [11]–[14], Internal LM estimation [15]–[17], and LM rescoring [18]. Lastly, from the modeling standpoint, various frameworks have been suggested, such as the Mixture of Experts (MoE) which uses separate encoders and decoders for different languages [19], [20], frame-level Language Identification or Diarization as an auxiliary task [21], [22], and the incorporation of self-supervised models as frontend models for ASR [23].

When superlarge parameter models show their emergent ability to understand and generate language when given a suitable prompt [24], researchers are turning to large-scale foundational models trained on extensive multilingual datasets, e.g., Whisper [25], USM [26] and MMS [27]. These models aim to encapsulate various linguistic rules and contexts, offering more robust performance across different languages and dialects. Using large-scale training data and advanced modeling techniques, these foundational models have the potential to revolutionize the field of ASR, making it more inclusive and accurate for multilingual and code-switching populations.

In this paper, we concentrate on applying varied language labels as prompts during the Whisper model's training and decoding phases. Our investigation centers on the efficacy of these language label prompts in the model's fine-tuning process. Additionally, we propose a prompting approach that considers code-switching as a distinct language. This method derives language embeddings through a weighted combination of the respective language embeddings, such as Mandarin and English, and attempts to enhance model performance under a code-switching scenario.

The paper is organized as follows. Section II is to review recently proposed adaptation methods based on the Whisper model to Code-Switching or specific languages. Section III presents the methods used in our experiments. Section IV briefly summarizes the datasets used and the overall experimental setup. Section V shows our experimental results and section VI shows our ablation study on the training data size. After that, we draw conclusions in Section VII.

TABLE I
PROMPT USED IN PROMPTINGWHISPER FOR CS-ASR AND OUR PROPOSED LANGUAGE-FUSION PROMPT FOR BOTH FINETUNING AND DECODING. ⟨|sot|⟩ STANDS FOR ⟨|startoftranscript|⟩ TOKEN AND ⟨|asr|⟩ MEANS THE ⟨|transcribe|⟩ TOKEN IN WHISPER TOKENIZER.

Languages	Default	PromptingWhisper	Language-Fusion
Zh+En	⟨|sot|⟩⟨**|zh|**⟩ *or* ⟨**|en|**⟩⟨|asr|⟩	⟨|sot|⟩⟨**|zh|**⟩⟨**|en|**⟩ *or* ⟨**|en|**⟩⟨**|zh|**⟩⟨|asr|⟩	⟨|sot|⟩⟨**|en-zh|**⟩⟨|asr|⟩

II. WHISPER APPLICATIONS

The Whisper model [25] is an advanced speech recognition system created by OpenAI that can accurately transcribe audio into text. It is trained on a wide-ranging dataset, enabling it to handle multiple languages and dialects effectively. It is also capable of speech translation and language identification, except for speech recognition.

Whisper stands out for its robust performance in various acoustic settings and its contextual understanding of improved transcription, which makes it distinct for various speech and text research tasks that benefit from the Whisper encoder and decoder separately or simultaneously. Specifically, following the same training pipeline and finetuning the entire Whisper model, performance improvement is obtained for several low-resource language speech recognition [28], [29]. TCPGen also shows its effectiveness of contextual biasing for the Whisper model [30]. With the utilization of the Whisper encoder, deepfake detection can benefit from input features composed of the embeddings extracted from the last layer of the Whisper encoder and traditional MFCC features [31]. With a similar strategy applied, these embeddings can also improve the performance of infant cry classification models compared to the MFCC features [32]. Also, researchers indicate that the embeddings from the Whisper encoder are not only noise robust to ASR task but also contain information that can help classify the noise types, e.g., audio event tagging [33].

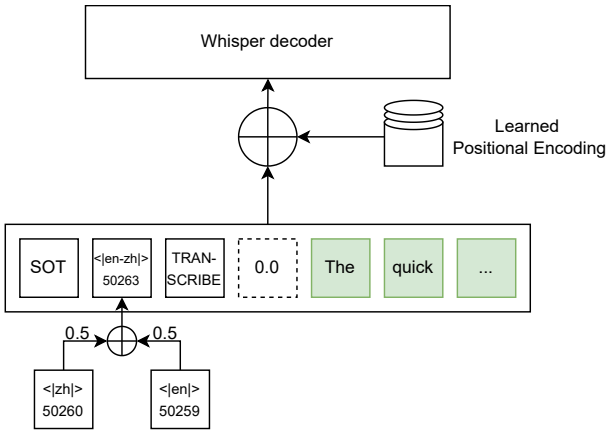


Fig. 1. The proposed Language Fusion method

Researchers usually try to prompt the decoder for various applications when it comes to the Whisper decoder. By replacing the speech recognition transcription label with several Spoken Language Understanding (SLU) labels while keeping the same decoding prompt as the ASR task, Whisper can perform SLU through transfer learning and multitask learning [34]. Also, when customized prompts are fed into the Whisper decoder, improved performance is observed for several zero-shot tasks such as Code-Switching ASR and Speech Translation, which had been previously shown to underperform with the default Whisper prompts. [35].

III. METHODS

A. PromptingWhisper for CS-ASR

PromptingWhisper [35] is an innovative approach to speech recognition, focusing on adapting the Whisper model to new and untrained tasks through prompt engineering. This method involves strategically using prompts - specific instructions or inputs - to guide the Whisper model in processing and responding to tasks beyond its original training. PromptingWhisper primarily explores three tasks: audio-visual speech recognition (AVSR), where the model transcribes speech from videos with related visual content; code-switched speech recognition (CS-ASR), which involves recognizing speech that alternates between different languages; and speech translation (ST) for language pairs that the model has not previously encountered.

For CS-ASR, PromptingWhisper suggests combining two Language Prompts instead of using only one of the two Languages, e.g., Table I shows all the prompts for Mandarin-English CS-ASR. Precisely, following the default prompting rule of Whisper [25], the resulting prompt for speech recognition would be either ⟨|sot|⟩⟨**|zh|**⟩⟨|asr|⟩ or ⟨|sot|⟩⟨**|en|**⟩⟨|asr|⟩ where only one language should be specified, however, PromptingWhisper suggest that given a prompt like ⟨|sot|⟩⟨**|zh|**⟩⟨**|en|**⟩⟨|asr|⟩ where sequentially inserting two languages would introduce 19% relative performance improvement in average for CS-ASR task among several CS corpora such as SEAME [36] and ASCEND [37]. Likewise, we think ⟨|sot|⟩⟨**|en|**⟩⟨**|zh|**⟩⟨|asr|⟩which only reverses the order of two languages should also have similar outcome. Therefore, we apply both ⟨**|en|**⟩⟨**|zh|**⟩ and ⟨**|zh|**⟩⟨**|en|**⟩ language prompts in the following experiments, treating them as PromptingWhisper's suggested prompts.

B. Language-Fusion Prompt

Inspired by the **PromptingWhisper**, we introduce a new Language Prompt that fuses the pre-trained English and Chinese Language Embeddings, which is named **Language-Fusion Prompt**, to investigate if this could be beneficial to English-Mandarin CS-ASR task. Typically, as shown in Figure 1, the new embedding is obtained by weighting the pre-trained embeddings corresponding to English and Chinese Language Prompt tokens that are ⟨**|en|**⟩ (*Token id in Whisper*

Tokenizer is 50259) and ⟨|**zh**|⟩ (*Token id in Whisper Tokenizer is 50260*) with the same weighting factor which is set to 0.5. To keep the same output dimension as the original Whisper Decoder, the ⟨|**ru**|⟩ (*Stands for Russian Language. Token id in Whisper Tokenizer is 50263*) is then replaced with the resulting new Language Prompt which is named as ⟨|**en-zh**|⟩. We selected the Russian language prompt as the replacement because of the substantial acoustic and linguistic differences between Russian and English/Chinese. The decoding prompt is finalized as ⟨|sot|⟩⟨|**en-zh**|⟩⟨|asr|⟩shown in Table I.

TABLE II
OVERALL SPEECH DATA DISTRIBUTION FOR ASR MODEL TRAINING AND TESTING.

Corpus	Subset	Duration(Hrs)
SEAME	Train	93.6
	Dev_{Man}	7.5
	Dev_{Sge}	3.9
ASRU	Train	193.0
	Valid	6.8
	Dev1	20.4
	Dev2	21.3
	Test	20.6

IV. EXPERIMENTS

A. Data

We select two Mandarin-English code-switching ASR data sets to verify the effectiveness of all the methods mentioned in Section III. One is SEAME [36], a conversational Mandarin-English corpus from SouthEast Asia, i.e., Malaysia and Singapore. Another is a Mandarin-English CS data set from China Mainland, released by `Datatang` for a Mandarin-English CS ASR challenge in ASRU2019 [38]. For brevity, we name it ASRU in what follows. Though both data sets are Mandarin-English CS, they are hugely different. Firstly, they are from different areas, which means CS is influenced by different cultural backgrounds. More importantly, SEAME data is conversational speech, while ASRU is read speech, and hence, it is much simpler. Table II reports overall speech data distributions in detail, where Dev_{Man} and Dev_{Sge} are two officially defined test sets for SEAME corpus. Dev_{Man} is dominated by Mandarin and vice versa; the other is dominated by English. Mandarin dominates all ASRU datasets.

B. Model

Due to the limitation of computing resources, all of our experiments are performed with the Whisper-small multilingual model. The encoder is configured with 12 layers, and the decoder consists of 12 layers with 8-head attention. The input feature is 80-dim Mel frequency bins, which are computed on 25-ms windows with a stride of 10 ms. Our models are trained on one A40 GPU with 48GB VRAM. The original batch size is set to 6, and the gradient accumulation is 12, for a total batch size of 72. We use AdamW optimizer with the peak learning rate of $1e^{-5}$, and the warmup lasts 200 steps. The max updating step is set to 30k. Also, mixed-precision training strategy is applied in our experiments. When decoding, we use an average model from the last five epochs, and the decoding beam size is set to 1.

V. RESULTS

The experimental results include mainly two parts. First, we show the results before finetuning the Whisper model called the Zero-shot Prompt. Then, by finetuning the Whisper model following different Prompt styles, we show the huge improvements for all Prompt styles among the two Code-Switching datasets.

TABLE III
MEAN ERROR RATES (MERS)(%) WITH DIFFERENT LANGUAGE PROMPT FOR ZERO-SHOT CODE-SWITCHING ASR.

Type	L-Prompt	SEAME		ASRU		
		Dev_{Man}	Dev_{Sge}	Dev1	Dev2	Test
Conformer	N/A	**16.6**	**23.3**	**8.6**	**14.0**	**13.2**
Official	⟨|en|⟩	101.9	83.6	96.0	98.9	105.3
	⟨|zh|⟩	80.8	157.5	27.0	25.3	25.0
	Auto	67.8	84.9	31.1	29.9	29.4
Custom	⟨|en|⟩⟨|zh|⟩	84.0	81.3	98.4	101.1	99.4
	⟨|zh|⟩⟨|en|⟩	98.8	81.2	33.2	32.2	32.3
	⟨|en-zh|⟩	74.0	101.7	33.8	31.9	31.6

A. Zero-Shot Prompts on Whisper-Small

We follow the instructions of PromptingWhisper and apply the suggested Language Prompts and our proposed method to Whisper-small, and the results are shown in Table III. Specifically, **Conformer** follows the configuration from recipe [39] in the ESPnet2 toolkit. **Official** type represents the original Whisper Language Prompt style, where ⟨|en|⟩ and ⟨|zh|⟩ stand for specifying English and Mandarin for the entire test set respectively. *Auto* denotes that we don't manually state the Language Prompt when performing decoding, which means that Whisper would automatically recognize the Language Label for each sentence. **Custom** type includes two combined Language Prompts following PromptingWhisper where ⟨|en|⟩⟨|zh|⟩ stands for English first and Mandarin second in the combined Language Prompt and vice versa. The proposed weighted-sum method is shown as ⟨|en-zh|⟩.

The results show that for all testsets, the Whisper-small model with various Language-Prompts underperforms Conformer models that are trained with corresponding training data. However, different Language-Prompts do significantly affect the performance of Code-Switching speech recognition. Specifically, for the SEAME dataset, ⟨|**en**|⟩ prompt gives better results for Dev_{Sge} while ⟨|**zh**|⟩ shows better performance on Dev_{Man}. The *Auto* prompt shows better performance compared with those with specific language prompts for Dev_{Man} and similar results for Dev_{Sge}. When given customized prompt, ⟨|**en**|⟩⟨|**zh**|⟩ and ⟨|**zh**|⟩⟨|**en**|⟩ prompts show best results on Dev_{Sge}, while Dev_{Man} underperforms the official prompts. Our proposed Language-Fusion prompt ⟨|**en-zh**|⟩ also performs poorly in the zero-shot scenario.

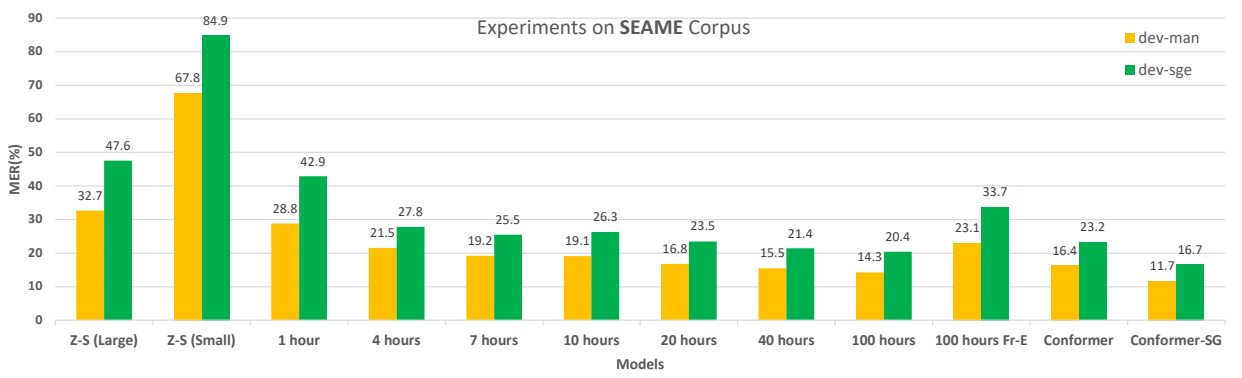


Fig. 2. MER(%) results of SEAME corpus with diverse training data size finetune on Whisper-Small model. **Z-S (Large)** stands for the PromptingWhisper-Large model, and **Z-S (Small)** corresponds to directly decode Whisper-Small model with **Auto** Language Prompt mentioned in Table III. {1,4,7,10,20,40,100} hours indicate the Whisper-Small models that are finetuned with corresponding training data size and ⟨|en|⟩ is fixed during both finetuning and decoding stages. **Fr-E** denotes Freeze-Encoder when fintuning the Whisper-Small model. **Conformer-SG** is our in-house English-Mandarin-Malay trilingual model trained with over 40k hours of Singaporean-accented data.

TABLE IV
MERS(%) WITH DIFFERENT LANGUAGE PROMPT FOR WHISPER-SMALL FINETUNED MODEL. ⟨|ru|⟩ STANDS FOR RUSSIAN LANGUAGE PROMPT.

Type	L-Prompt	SEAME		ASRU		
		Dev_{Man}	Dev_{Sge}	Dev1	Dev2	Test
Conformer	N/A	**16.6**	23.3	8.6	14.0	13.2
Official	⟨|en|⟩	**14.3**	**20.4**	**6.3**	10.9	10.3
	⟨|zh|⟩	14.8	20.6	**6.3**	10.8	**10.1**
	⟨|**ru**|⟩	15.5	21.5	6.3	10.8	10.3
Custom	⟨|en|⟩⟨|zh|⟩	15.1	21.1	**6.3**	**10.6**	**10.1**
	⟨|zh|⟩⟨|en|⟩	15.0	21.0	6.5	11.0	10.5
	⟨|en-zh|⟩	15.1	20.9	**6.3**	10.8	**10.1**

When it comes with ASRU dataset, ⟨|**zh**|⟩ and *Auto* prompts show significant performance improvement compared with ⟨|**en**|⟩ **and** ⟨|**en**|⟩⟨|**zh**|⟩ prompts which shows consistency to data composition of ASRU.

B. Finetuning Whisper Model

We finetune the Whisper-Small model given different language prompts and then specify the same language prompt when performing decoding. Table IV shows the MER results of all Whisper models we finetuned on SEAME and ASRU datasets.

The results show that regardless of which Language Prompt we use for finetuning and decoding, all the testsets obtain significant performance improvements and outperform the Conformer models that are trained with corresponding training set.

Specifically, ⟨|**en**|⟩ prompt achieves best performance for both test sets of SEAME corpus and also introduces 2.3 $\sim$ 2.9% absolute MER reduction compared with Conformer model, while ⟨|**en**|⟩⟨|**zh**|⟩ prompt yields optimal results on all Dev and Test sets of ASRU corpus which demonstrates 2.3 $\sim$ 3.4% absolute MER reduction compared with Conformer model. However, We observe that regardless of whether the language prompt is ⟨|**en**|⟩, ⟨|**zh**|⟩, both, or even our proposed fusion prompt, the results are quite similar, with less than a 1% MER gap across all models for each test set. To further explore the effect of using different language prompts, we introduced the Russian Language Prompt ⟨|**ru**|⟩ as a reference in this experiment. The rationale behind using ⟨|**ru**|⟩, a language not present in the training or test data, is to examine whether the choice of an unrelated language prompt would influence the model's performance after finetuning. Our findings are confirmed by the results of ⟨|**ru**|⟩ which demonstrate that regardless of what language prompt is given to the Whisper decoder, once the finetuning process is complete, the performance achieved will be similar.

VI. ABLATION STUDY

In this section, we primarily investigate how the size of the training dataset affects the performance outcomes of finetuning the Whisper model. Additionally, we explore the effects on Whisper's performance when the encoder parameters are frozen during the finetuning process and solely update the decoder parameters.

First, for all the finetuned Whisper models, we fix the language prompts ⟨|**en**|⟩ for SEAME and ⟨|**zh**|⟩ for ASRU2019 experiments. We then randomly subset $(1, 4, 7, 10, 20, 40)$ hours from both SEAME and ASRU training sets and update the Whisper-Small model for 8000 steps to examine the smallest size of data that could obtain a practical CS-ASR system. Also, we try to freeze the entire encoder of the Whisper model while performing finetuning on the whole dataset to determine if finetuning the decoder alone can effectively bridge the gap between multilingual and Code-Switching scenarios. Additionally, this can help to verify whether the Whisper Encoder can overcome the acoustic mismatch that often arises in such diverse linguistic environments.

Figure 2 shows the results on the SEAME corpus. The results show that only 1 hour of training data can produce a model that outperforms PromptingWhisper-Large by around 10% and obtains around 50% MER reduction. When the amount of training data reaches 20 and 40 hours, the model will yield results comparable to or surpass the conformer

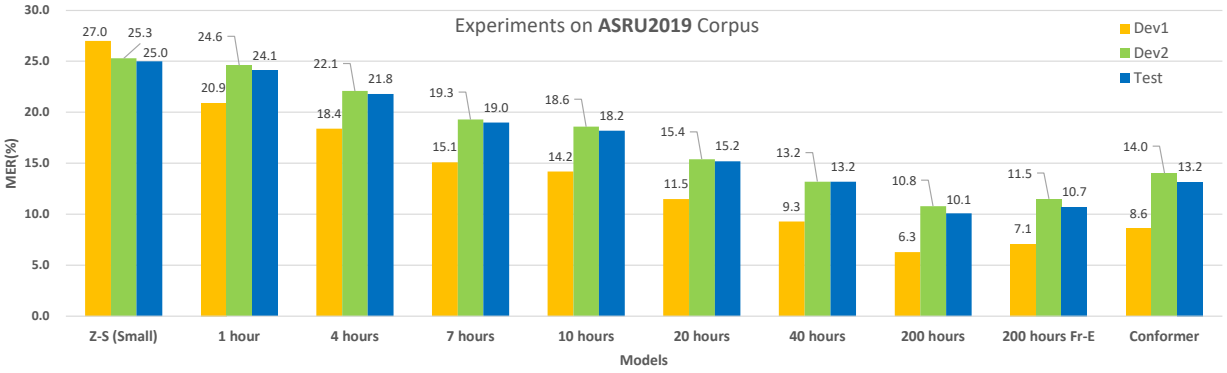


Fig. 3. MER(%) results of ASRU corpus with diverse training data size finetune on Whisper-Small model. **Z-S (Small)** corresponds to directly decode Whisper-Small model with ⟨|zh|⟩ Language Prompt. {1,4,7,10,20,40,200} hours indicate the Whisper-Small models that are finetuned with corresponding training data size and ⟨|zh|⟩ is fixed during both finetuning and decoding stages. **Fr-E** denotes Freeze-Encoder when fintuning the Whisper-Small model.

model trained with the entire training set. However, when the encoder parameters are frozen, the performance drop could be at most 65%, which suggests that the encoder of the Whisper-Small model may exhibit a substantial mismatch with SEAME corpus (Singaporean Accents), potentially leading to degradation in its performance.

In Figure 2, we also present the Conformer-SG, which is a U2++ Conformer streaming ASR model as referenced in [40]. This model, trained on over 40,000 hours of Mandarin-English-Malay speech data, achieves state-of-the-art performance on the SEAME corpus and surpasses the finetuned Whisper-Small model by approximately 18%.

Figure 3 shows the results on the ASRU corpus. The results show a similar conclusion to the one we obtained from experiments with the SEAME corpus. However, the frozen encoder experiment for the ASRU corpus shows much lower performance degradation compared with the SEAME corpus, which suggests the performance gap tends to diminish when dealing with scenarios, such as speech without obvious accents, that closely align with the conditions and characteristics of the Whisper model's training set.

VII. Conclusion

This paper reveals that adapting the Whisper model with code-switching datasets significantly enhances its capability for code-switching speech recognition (CS-ASR), even in contexts with limited resources. Our experiments, which span a variety of linguistic backgrounds, demonstrate that while different prompting strategies yield varied performances prior to adaptation, after adapting the Whisper model with code-switching speech data, these strategies result in similarly enhanced performance, effectively mitigating the complexities inherent in code-switching environments. These adaptations not only bolster the model's overall performance but also align with the broader goal of developing ASR systems that are more inclusive and precise for multilingual users. This research thus marks a crucial step forward in understanding the potential of large foundational models for navigating the intricate dynamics of code-switching in various linguistic scenarios.

VIII. Acknowledgements

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore.

References

[1] G. Ma, W. Wang, Y. Li, Y. Yang, B. Du, and H. Fu, "Lae-st-moe: Boosted language-aware encoder using speech translation auxiliary task for e2e code-switching asr," in *ASRU*, 2023.

[2] Y. Peng, J. Zhang, H. Xu, H. Huang, and E. S. Chng, "Minimum word error training for non-autoregressive transformer-based code-switching asr," in *ICASSP*, 2022.

[3] B. Yan, C. Zhang, M. Yu, *et al.*, "Joint modeling of code-switched and monolingual asr via conditional factorization," in *ICASSP*, 2022.

[4] Y. Long, S. Wei, J. Lian, and Y. Li, "Pronunciation augmentation for mandarin-english code-switching speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, 2021.

[5] C.-T. Chang, S.-P. Chuang, and H.-Y. Lee, "Code-switching sentence generation by generative adversarial networks and its application to data augmentation," *arXiv preprint arXiv:1811.02356*, 2018.

[6] Y. Li and P. Fung, "Code-switch language model with inversion constraints for mixed language speech recognition," in *Proceedings of COLING 2012*, 2012.

[7] X. Hu, Q. Zhang, L. Yang, B. Gu, and X. Xu, "Data augmentation for code-switch language modeling by fusing multiple text generation methods.," in *INTERSPEECH*, 2020.

[8] Y. Li and P. Fung, "Code switch language modeling with functional head constraint," in *ICASSP*, 2014.

[9] J. Ao, R. Wang, L. Zhou, *et al.*, "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," *arXiv preprint arXiv:2110.07205*, 2021.

[10] B. Yusuf, A. Gandhe, and A. Sokolov, "Usted: Improving asr with a unified speech and text encoder-decoder," in *ICASSP*, 2022.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[13] S. Santy, A. Srinivasan, and M. Choudhury, "Bertologicomix: How does code-mixing interact with multilingual bert?" In *Proceedings of Adapt-NLP*, 2021.

[14] G. Lee and H. Li, "Modeling code-switch languages using bilingual parallel corpus," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[15] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *ICASSP*, 2018.

[16] Y. Liu, R. Ma, H. Xu, Y. He, Z. Ma, and W. Zhang, "Internal language model estimation through explicit context vector learning for attention-based encoder-decoder asr," *arXiv preprint arXiv:2201.11627*, 2022.

[17] Y. Peng, Y. Liu, J. Zhang, *et al.*, "Internal language model estimation based language model fusion for cross-domain code-switching speech recognition," *arXiv preprint arXiv:2207.04176*, 2022.

[18] G. Liu and L. Cao, "Code-switch speech rescoring with monolingual data," in *ICASSP*, 2021.

[19] Y. Lu, M. Huang, H. Li, J. Guo, and Y. Qian, "Bi-encoder transformer network for mandarin-english code-switching speech recognition using mixture of experts.," in *INTERSPEECH*, 2020.

[20] T. Song, Q. Xu, M. Ge, *et al.*, "Language-specific characteristic assistance for code-switching speech recognition," *arXiv preprint arXiv:2206.14580*, 2022.

[21] C. Shan, C. Weng, G. Wang, *et al.*, "Investigating end-to-end speech recognition for mandarin-english code-switching," in *ICASSP*, 2019.

[22] H. Liu, H. Xu, L. P. Garcia, A. W. Khong, Y. He, and S. Khudanpur, "Reducing language confusion for code-switching speech recognition with token-level language diarization," in *ICASSP*, 2023.

[23] L.-H. Tseng, Y.-K. Fu, H.-J. Chang, and H.-y. Lee, "Mandarin-english code-switching speech recognition with self-supervised speech representation models," *arXiv preprint arXiv:2110.03504*, 2021.

[24] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, 2023.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[26] Y. Zhang, W. Han, J. Qin, *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[27] V. Pratap, A. Tjandra, B. Shi, *et al.*, "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[28] K. Bhogale, S. Sundaresan, A. Raman, T. Javed, M. M. Khapra, and P. Kumar, "Vistaar: Diverse Benchmarks and Training Sets for Indian Language ASR," in *INTERSPEECH*, 2023.

[29] B. Talafha, A. Waheed, and M. Abdul-Mageed, "N-Shot Benchmarking of Whisper on Diverse Arabic Speech Recognition," in *INTERSPEECH*, 2023.

[30] G. Sun, X. Zheng, C. Zhang, and P. C. Woodland, "Can Contextual Biasing Remain Effective with Whisper and GPT-2?" In *INTERSPEECH*, 2023.

[31] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved DeepFake Detection Using Whisper Features," in *INTERSPEECH*, 2023.

[32] M. Charola, A. Kachhi, and H. A. Patil, "Whisper Encoder features for Infant Cry Classification," in *INTERSPEECH*, 2023.

[33] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers," in *INTERSPEECH*, 2023.

[34] M. Wang, Y. Li, J. Guo, *et al.*, "WhiSLU: End-to-End Spoken Language Understanding with Whisper," in *INTERSPEECH*, 2023.

[35] P. Peng, B. Yan, S. Watanabe, and D. Harwath, "Prompting the hidden talent of web-scale speech models for zero-shot task generalization," *arXiv preprint arXiv:2305.11095*, 2023.

[36] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "Mandarin–english code-switching speech corpus in south-east asia: Seame," *Language Resources and Evaluation*, 2015.

[37] H. Lovenia, S. Cahyawijaya, G. I. Winata, *et al.*, "Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation," *arXiv preprint arXiv:2112.06223*, 2021.

[38] X. Shi, Q. Feng, and L. Xie, "The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results," *arXiv preprint arXiv:2007.05916*, 2020.

[39] *ESPnet2-SEAME-recipe*, https://github.com/espnet/espnet/tree/master/egs2/seame/asr1, Accessed: 2023-11-27.

[40] D. Wu, B. Zhang, C. Yang, *et al.*, *U2++: Unified two-pass bidirectional end-to-end model for speech recognition*, 2021. arXiv: 2106.05642.

Transformer Attention Matrix Multiplication Design using 4×4 Systolic Arrays

Muhammad Sayyid Afif[‡], Infall Syafalni[‡§], Nana Sutisna[‡§], and Trio Adiono[‡§]
[‡]School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia
[§]University Center of Excellence on Microelectronics, Bandung Institute of Technology, Indonesia

***Abstract*—The transformer has brought a groundbreaking transformation to natural language processing (NLP) with its multihead attention mechanism, enabling efficient parallel processing of text data. This paper enhances transformer performance using a systolic array architecture for matrix multiplication. We optimize the computation process by leveraging parallel partitioning to accelerate self and multihead-attention mechanisms within transformer models. Utilizing the BERT model, we partition its 768-dimensional vectors and process to make multiple smaller 4×4 structures. The experiments used software simulations and FPGA Zybo 7020 hardware design synthesis. The primary objective is to improve computational throughput and resource efficiency, leading to the development of lightweight and high-performance transformer models. This study highlights the potential of parallel partitioning with systolic arrays in matrices multiplication, advancing the transformer's computation processing tasks. Our experiment simulates the parallel partition and synthesizes the hardware architecture design. By comparing the conventional nested loop and parallel partition method, we find a 66.9% speed increase, with a maximum frequency of 91.7 Mhz. Our findings contribute to optimizing transformer models for various applications, emphasizing the role of hardware acceleration.**

***Index Terms*—Transformer, Multihead Attention, Accelerator, Systolic Array, Natural Language Processing.**

I. Introduction

TRANSFORMER has brought a groundbreaking transformation to natural language processing. The transformer model introduced a novel mechanism called multihead attention, which has proven highly effective in capturing contextual relationships and dependencies in text data [1]. By eliminating the need for sequential operations like those in recurrent neural networks or convolutional neural networks, multihead attention allows the transformer model to process text in parallel, [2], significantly improving tasks such as machine translation. [3].

The use of transformer models is not limited to natural language processing (NLP) but has also expanded to image processing. One prominent implementation is using transformer models in image processing, known as Vision Transformer (ViT) [4]. A specialized accelerator known as ViT Accelerator (ViA) [5] aims to enhance the performance of ViT.

The transformer model itself is still in the stage of efficiency and optimization, ranging from accuracy and speed to computational flexibility. Researchers have researched transformer models using Field Programmable Gate Array (FPGA) [6], opening up opportunities for further study. We conducted experiments as follows:

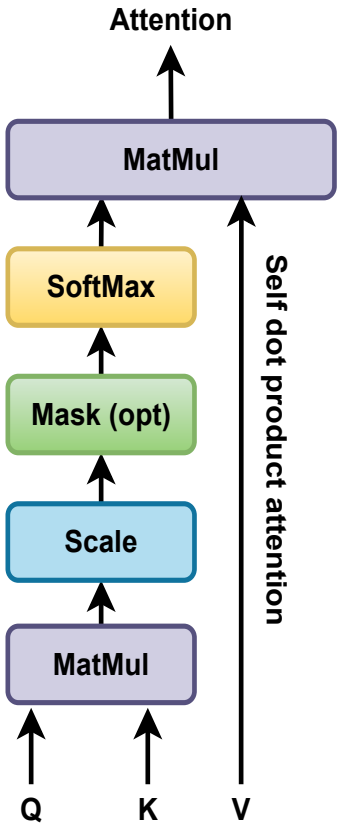


Fig. 1: Self head attention [1]

- We utilize the systolic array property in matrix multiplication by evaluating and developing the matrix computation process using parallel partitions.
- We will use the proposed architecture to experiment with the FPGA Zybo 7020 hardware.

Therefore, the objectives of conducting these experiments are as follows:

- The main objective is accelerating the transformer model, specifically regarding self and multi-head attention.
- It aims to enhance performance with limited resources by evaluating and developing the matrix computation process.
- By minimizing the available resources, we can create a lightweight model.

The continuation of this paper consists of several chapters, arranged as follows. Chapter II will discuss the theoretical foundations that support the methods to be used, along with

several related works relevant to the proposed method. Chapter III presents the proposed method, including the method and architecture. Chapter IV explains how we experiment. Chapter V provides the experimental results. Chapter VI discusses the conclusion and future development.

II. Related works

A. Self-Head Attention

The transformer model consists of three parts: word embedding, encoder blocks, and decoder blocks. Within the decoder block, there is a component called attention used to compute the relationships between words simultaneously. In this process, we collect vectors to form matrices of query (Q), key (K), and value (V). These matrices are then computed into Self-Head Attention using the following formula [1]:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Equation (1) represents the scaled dot product attention mechanism. It compares the query matrix (Q) with the key matrix (K) and calculates the dot product (QK^T) to measure the similarity between each query and each key. Using the dimension key (d_K) ensures the value of the dot product is within the manageable range, avoiding a large number that could push the softmax results to become too small to be helpful during training. Once the dot product is scaled, its result passes the softmax function. The softmax function transforms the scaled value into a probability distribution where the total of all scaled results is 1. The softmax output then multiplies the value matrix (V), which includes the specific data the model aims to prioritize. The model gives attention to the values corresponding to higher softmax scores, which reflect the relevance of the keys to the query. The higher the softmax score for a key, the more weight its corresponding value will have in the final output.

B. Multihead Attention

Multihead attention is the application of several self-head attentions running in parallel with separate parameters. After completing the calculation of each head, we combine the results. Combining the results can help the model capture more complex relationships between words and enable the model to represent the input data better, resulting in more affluent and contextual outcomes, thus addressing more complex tasks. We compute multihead attention using the following formula [1]:

$$Multihead(Q, K, V) = Concat(head1, ..., head_n) \quad (2)$$

Equation (2) computes the multihead mechanism by concatenating $head1$ to $headn$, with each result from equation (1) being each head.

C. Systolic Array

A systolic system is an architecture designed to perform processes in parallel, using multiple simple processing elements. Simplicity with orderly communication and control

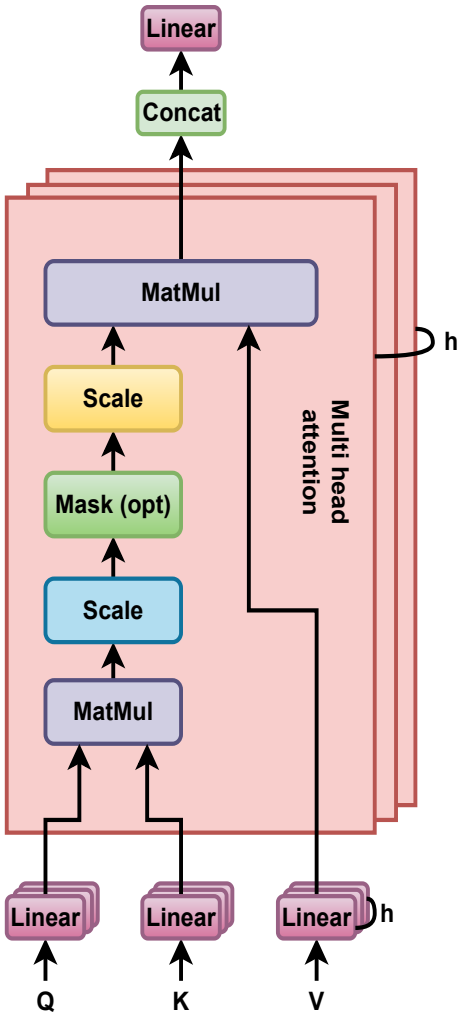


Fig. 2: Multi head attention [1]

structures has excellent advantages over complicated structures in design and implementation. The cells in the systolic system are usually interconnected to form a systolic array or systolic tree.

Since early 1984, CMU has been working on the design of Warp engines using systolic processors. The designers intended these processors to be fast and powerful computing engines for many low-level signal-processing and image-processing tasks [7]. The Warp processor family uses a systolic array to efficiently implement various systolic algorithms, including Fast Fourier Transform, matrix multiplication, and two-dimensional convolution [8].

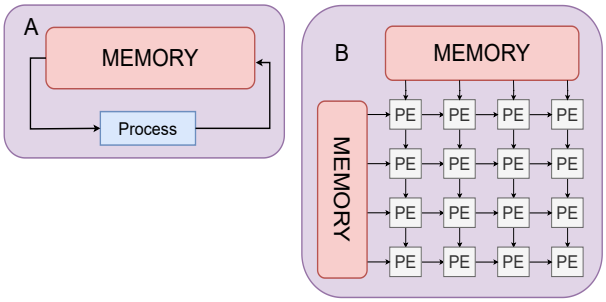


Fig. 3: A) Normal computation. B) Systolic Array

Replacing processing elements (PE) with PE arrays or cells, as shown in Fig. 3B, can achieve higher computing throughput without increasing memory bandwidth. Each data point from the main memory is processed several times and returned to memory [9].

III. Proposed Method

A. Parallel Partition

Parallelization here involves partitioning and paralleling the computation of large-dimensional matrices. Matrix partitioning consists of breaking down large-dimensional matrices into several smaller-dimensional matrix parts. In this case, we use a 4×4 smaller matrix. We use the BERT model as the foundation of our method. The BERT model utilizes a 768-dimensional vector [10]. With the dimensions 4×768, we will partition it into 4×4, which will iterate 192 times.

In Fig. 4, we will multiply a matrix Q with dimensions of 4×768 by a matrix K^T with dimensions of 4×768. We will partition the Q and K^T matrices into 4×4 small matrices. We will then multiply Q_{1-192} and K^T_{1-192} pairwise to form temporary matrices G_{1-192} with dimensions of 4×4. Afterward, we will sum G_{1-192} entirely to form the matrix G with dimensions of 4×4.

In Fig. 5, we will partition the matrix V with dimensions of 4×768 into small matrices V_{1-192} with dimensions of 4×4. Each V_{1-192} will be multiplied by the matrix G. The multiplication result is Att_{1-192}, each with dimensions of 4×4. To obtain the final Attention matrix, matrices Att_{1-192} will be sequentially concatenated until the Attention matrix has dimensions of 4×768.

B. Architecture

Fig. 6 provides an overview of our hardware setup, utilizing 2 Block RAM (BRAM) modules as input and 1 BRAM module as output. The IP Core contains signal control and a systolic array of multiple Processing Elements (PEs).

Fig. 7 is our proposed processing element. We keep the input from matrices A and B in the register and forward it in the next cycle. We multiply A by B to get output C, and the result is maintained and accumulated in the next cycle to achieve the desired output.

IV. Experiment

A. Systolic Array & Bram

In this experiment, we use Vivado 2019.1. First, we created the processing element and integrated it with the systolic array, using Algorithm 1 as the processing element. If the $reset$ signal is true, all outputs will equal 0. Otherwise, $Output_C$ is calculated with $Output_C + (Input_A \times Input_B)$, while $Forward_A$ and $Forward_B$ are filled with $Input_A$ and $Input_B$, respectively.

We implement a finite state machine (FSM), as shown in Fig. 8, into the main code. The system will start in an idle state, waiting for the process to begin with a $start$ signal. If the $start$ signal is active and the num_of_input parameters are zero, it moves to the Done state, resetting the $start$ signal and moving again to the Idle state. Otherwise, it moves to the Read state, where we read the data from $Bram_A$ and $Bram_B$, and the $write-enabled$ signal is activated. We write the processed data from $Bram_A$ and $Bram_B$ into $Bram_C$ as the output in the Write state. If the data that came in does not equal the num_of_input parameter, it will return to the Read state. If the data equals, the next phase is in the Done state, resetting the $start$ signals and moving to Idle states awaiting the following process or cycle.

Algorithm 1 Processing Element

Input : $reset$, $clock$, $InputA$, $InputB$
Output : $Forward_A$, $Forward_B$, $Output_C$

procedure ALWAYS @(POSEDGE CLOCK)
  if $reset$ IS TRUE **then**
    SET $Forward_A$, $Forward_B$, $Output_C$ TO 0
  else
    SET $Output_C$ TO $Output_C + (Input_A \times Input_B)$
    SET $Forward_A$ TO $Input_A$
    SET $Forward_B$ TO $Input_B$
  end if
end procedure

B. Parallel Partition

We simulate the parallel partition in Python. For the first part, we multiply the matrix Q and K^T. Same as Fig. 4 with Algorithm 2. The second part multiplies the matrix result from the first part (matrix G) to matrix V, the same as in Fig. 5 with Algorithm 3.

Algorithm 2 $Q \times K^T$ partition

Input : $QMatrix$, $K^TMatrix$
Output : $GMatrix$

SET $block_size$ TO 4
for i **do** FROM 0 TO 768 STEP 4 DO
  SET $block_Query$ TO $QMatrix$ [ALL ROW, i TO $i+block_size$]
  SET $block_Key^T$ TO $K^TMatrix$ [i TO $i+block_size$, ALL COLUMN]
  SET $block_result$ TO DOT PRODUCT OF $block_Query$, $block_Key^T$)
  ADD $block_result$ TO $GMatrix$
end for
PRINT $GMatrix$

V. Experimental Results

Fig. 9 results from the simulation of Algorithms 2 and 3 from Python using Google Colab. Meanwhile, Fig. 10 uses C in CodeBlocks. For comparison, we include matrices multiplied only using a conventional nested loop in Python. From conducting 20 times runs, we get the slowed method is a conventional nested loop with 31.60 ms as the slowest and 13.40 ms as the fastest, averaging 18.79 ms. Then, it was

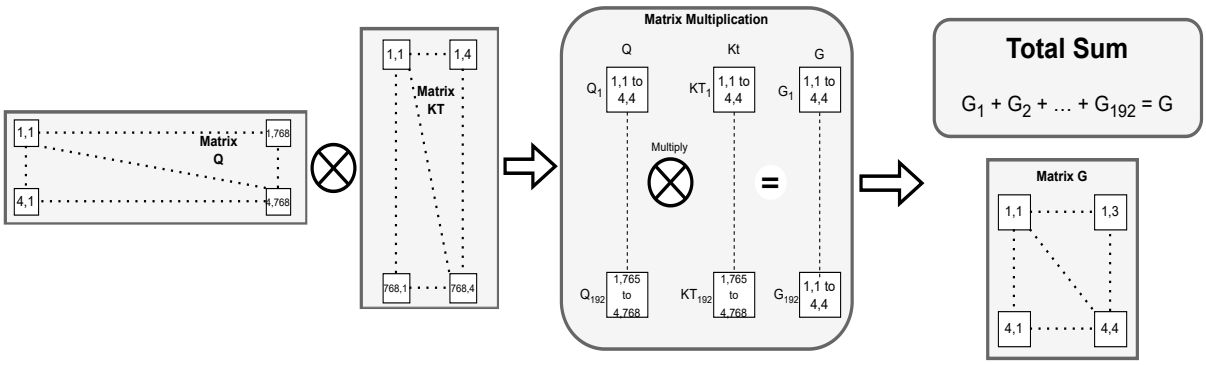


Fig. 4: Q and Kt partition to form matrix G

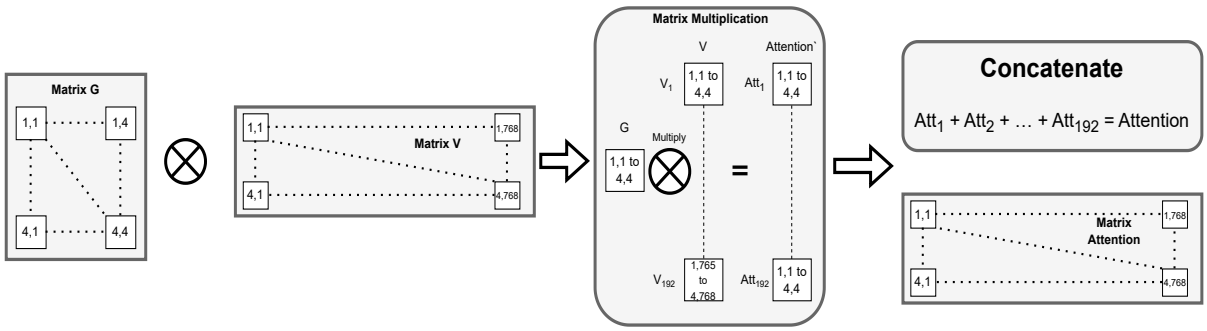


Fig. 5: V partition to form Attention

Algorithm 3 $G \times V$ partition

Input : $GMatrix$**,** $VMatrix$
Output : $Attention$

SET $block_size$ TO 4
for s **do** FROM 0 TO 768 STEP 4 DO
SET $block_Value$ TO $VMatrix$ [ALL ROW, s TO $s+block_size$]
SET $block_attention$ TO DOT PRODUCT OF ($GMatrix$, $block_Value$)
CONCATENATE $block_attention$ TO $Attention$
end for
PRINT $Attention$

followed by parallel partition in Python, with 16.10 ms as the slowest and 7.27 ms as the fastest, averaging 11.26 ms. The parallel partition in C is faster, with 0.28 ms as the slowest and 0.19 ms as the fastest, averaging 0.22 ms.

Using the design in Fig. 6, we run synthesis and implementation in Vivado 2019.1 without any problem. The worst negative slack (WNS) result is 9.1 ns with a clock period of 20 ns. We use equation 3 to find the maximum frequency. We get 0.0917 Ghz or 91.7 Mhz. With the junction temperature at $43.5°C$, Table II breaks down the utilization and report timing, and Table I breaks down the power usage.

$$F_{max} = 1/(T - WNS) \tag{3}$$

The Equation (3) is to calculate the max frequency. F_{max} is the maximum frequency, T is the clock period, and WNS is the worst negative slack.

Site Type	Used	Fixed	Available	Util%
Slice LUTs	8354	0	53200	15.70
LUT as Logic	6622	0	53200	12.45
LUT as Memory	1732	0	17400	9.95
LUT as Distributed RAM	1308	0		
LUT as Shift Register	424	0		
Slice Registers	9309	0	106400	8.75
Register as Flip Flop	9309	0	106400	8.75
Register as Latch	0	0	106400	0.00
F7 Muxes	6	0	266000	0.02
F8 Muxes	0	0	13300	0.00
Block RAM Tile	8	0	140	5.71
RAMB36/FIFO	8	0	140	5.71
RAMB36E1 only	8			
RAMB18	0	0	280	0.00
DSPs	0	0	220	0.00

TABLE I: Timing and utilization report

Parameter	Power usage
Clocks	0.025 W
Signal	0.023 W
Logic	0.016 W
Bram	0.023 W
PS7	1.370 W
Total	1.606 W

TABLE II: On-Chip Power

VI. Conclusions and Future Works

A. Conclusion

This experiment simulation and hardware test demonstrate that partitions are faster than the conventional nested loops in Python. The process speed increase averaged around 66.9%. Calculating parallel partitions with the C language proved

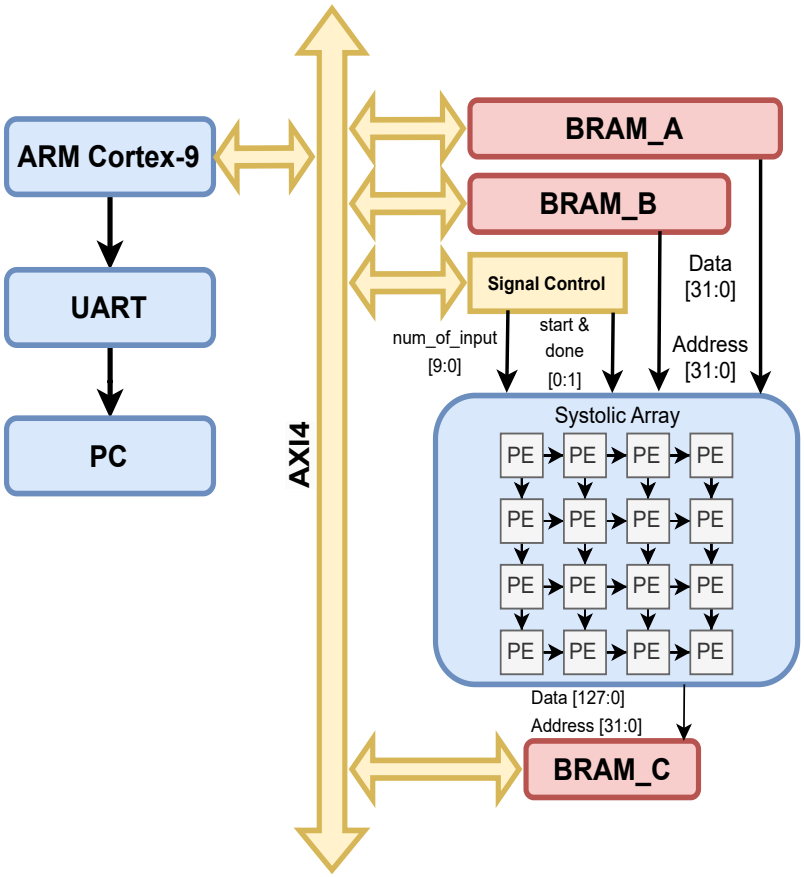


Fig. 6: Proposed Architecture Overview

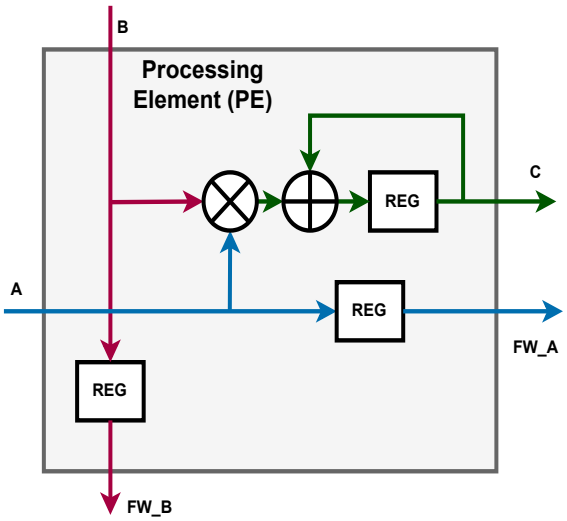


Fig. 7: Architecture of processing element

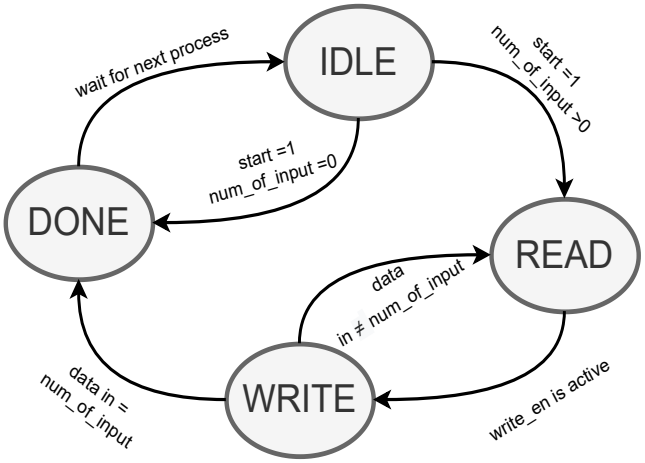


Fig. 8: Finite state machine

even faster, with an average speed of 0.22 ms. Additionally, the proposed architecture synthesis and implementation design using Zybo 7020 with Vivado 2019.1 showed a maximum frequency of 91.7 Mhz without issues.

B. Future Works

This experiment identifies several upgrades for future work that researchers can pursue to enhance the performance and capability of the transformer model, such as:

- Hardware implementation: Our future work is implementing the design and architectures into the Zybo 7020 hardware.
- Scaling to larger, more complex transformer models: Researchers may expand the experiment to deal with larger data sets. With more layers and parameters, the model

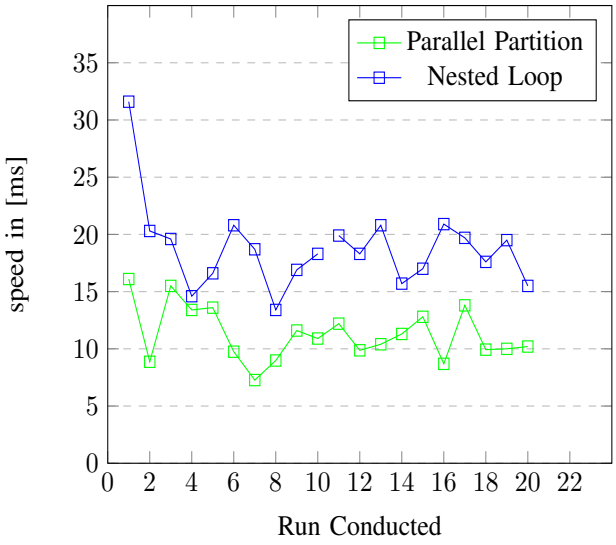


Fig. 9: Python simulation results

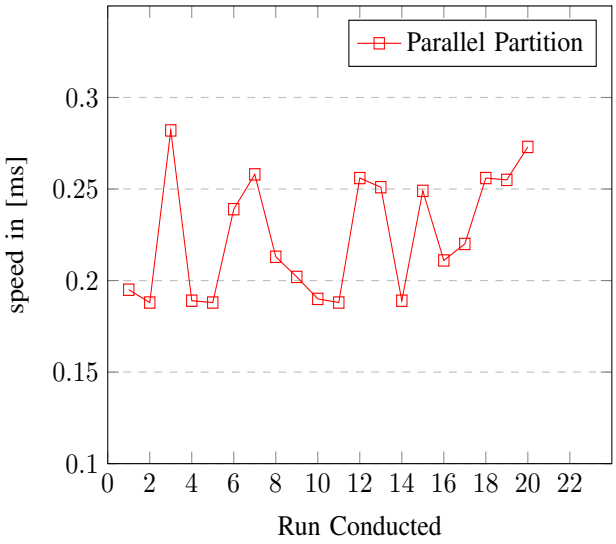


Fig. 10: Parallel Partition in C

would learn complicated patterns and dependencies on the data more reliably and accurately.

- Other machine-learning models using matrix multiplication can work on this parallel partition. Optimizing these models with parallel partitions will give us performance gains.
- Optimized other than Natural Language Processing: It can be optimized for natural language processing, extended, and applied in broader domains such as robotics and computer vision, making it more useful in general applications.

ACKNOWLEDGEMENT

This work is supported by the 2024 ITB Young Researcher Program provided by LPPM, Bandung Institute of Technology under grant no. [STEI1.PN-6-04-2024].

REFERENCES

[1] A. Vaswani et al., *Attention Is All You Need* arXiv, Aug. 01, 2023. Accessed: Dec. 03, 2023. [Online]. Available: http://arxiv.org/abs/1706.03762

[2] H. Peng, R. Schwartz, D. Li, and N. A. Smith, *A Mixture of $h-1$ Heads is Better than h Heads* arXiv, May 13, 2020. Accessed: Dec. 03, 2023. [Online]. Available: http://arxiv.org/abs/2005.06537

[3] S. M. Lakew, M. Cettolo, and M. Federico, *A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. In Proceedings of the 27th International Conference on Computational Linguistics*, New Mexico, USA: Aug.2018 pp. 641–652.

[4] Dosovitskiy, Alexey et al. *An* Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv abs/2010.11929 (2020)

[5] T. Wang et al., *ViA*: A Novel Vision-Transformer Accelerator Based on FPGA IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 41, no. 11, pp. 4088–4099, Nov. 2022, doi: 10.1109/TCAD.2022.3197489.

[6] B. Li et al., *FTRANS: Energy-Efficient Acceleration of Transformers using FPGA* vol. 7, 2020, doi: https://doi.org/10.1145/3370748.3406567.

[7] Kung, H.T. *Systolic Algorithms for the CMU Warp Processor*, *In Proceedings of the Seventh International Conference on Pattern Recognition*, pages 570-577.1984.

[8] E. Arnould, H. Kung, O. Menzilcioglu and K. Sarocky, *A systolic array computer*, ICASSP '85. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tampa, FL, USA, 1985, pp. 232-235, doi: 10.1109/ICASSP.1985.1168488.

[9] Kung, *Why systolic architectures?*, in *Computer, vol. 15*, no. 1, pp. 37-46, Jan. 1982, doi: 10.1109/MC.1982.1653825.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding arXiv*, May 24, 2019. Available: http://arxiv.org/abs/1810.04805

Affine Combination of General Adaptive Filters

Danqi Jin*, Yitong Chen†, Jie Chen†, Gongping Huang*
* School of Electronic Information, Wuhan University, Wuhan, China
E-mail: danqijin@mail.nwpu.edu.cn, gongpinghuang@whu.edu.cn
† CIAIC, Northwestern Polytechnical University, Xi'an, China
E-mail: chenyitong@mail.nwpu.edu.cn, dr.jie.chen@ieee.org

***Abstract*—Combination schemes are very effective for adaptive filters; yet, they are still constrained to use a linear model and the mean-square error (MSE) cost. All types of models, both linear and nonlinear, are accommodated in this work by expanding combination schemes to generic cost functions. Specifically, in this kind of combination framework, two candidate filters are firstly designed by optimizing separate general cost functions; then, to optimize overall performance, the combination coefficients that are given to these filters are fine-tuned by minimizing a third general cost function. Separately, we set and optimize each of the three generic cost functions. After that, we design an affine combination scheme by using the stochastic sign-gradient descent to adjust combination coefficients. Also, to enhance its performance, we include a weight-copying trick. Finally, simulation results are provided to validate their effectiveness for both the MSE and the logistic risk cost functions.**

***Index Terms*—Adaptive filter, combination scheme, nonlinear model, general cost function, weight-copying trick.**

I. INTRODUCTION

Adaptive filters are powerful tools for online parameter estimation from streaming data, and typical algorithms include the least-mean-square (LMS), the recursive least-square (RLS), the Kalman filter, and the affine projection algorithm (APA) [1], [2], [3], [4], [5]. Speech signal processing [6], [7], radar signal processing [8], earthquake detection [9], and many other applications have made extensive use of these adaptive filters.

Combination strategies have been suggested and implemented for adaptive filters [10], [11], [12] and distributed adaptive networks [13], [14], [15] with great success in the last few decades. In these studies, it is demonstrated that filters may be created by combining the benefits of all candidate filters via convex or affine combinations. Most often, these combination schemes are employed to make it easier to choose filter settings, to make it more resilient to an unknown environment, or to potentially improve performance beyond what any one filter could do [16], [17]. Acoustic echo cancellation [18], [19], beamforming [20], [21], and speech dereverberation [22], [23] are only a few examples of the actual applications where combination techniques have demonstrated promising results in providing superior solutions.

Use cases like online logistic learning from nonlinear data [24] have prompted the need for combination of general adaptive filters that follow nonlinear models. The majority of the research on adaptive filter combination strategies is devoted to linear models using the MSE cost function [25]. Although some studies have taken into account the combination of nonlinear filters in more general nonlinear models, these models are suboptimal since their combination coefficients are set by minimizing the MSE cost [18], [19]. For this reason, developing a method to integrate nonlinear models with generic adaptive filters is crucial.

In this work, we propose a strategy for combining two general adaptive filters to handle this problem, and we do it by using streaming data that could adhere to nonlinear models. Two candidate filters are built at the adaptive filter layer by minimizing various general cost functions. In order to optimize the overall performance, the combination layer uses the third general cost function to adjust combination coefficients. Every one of the three generic cost functions may be fine-tuned on its own. After that, we use the stochastic sign-gradient descent to design an affine combination scheme. Its performance is further improved by including a weight-copying trick. Finally, their effectiveness has been validated by simulation results. Before concluding this section and proceeding to the next, we provide the notation used throughout this work.

Notation. All scalars are represented by the regular typeface x with the exception of $F^{(i)}$, which indicates the i-th candidate filter. Column vectors are denoted by boldface little letters $\boldsymbol{x}$. The superscript $(\cdot)^\top$ denotes the transpose operator. The mathematical expectation is denoted by $\mathbb{E}\{\cdot\}$. The $\mathrm{mod}(\cdot,\cdot)$ function returns the remainder after division.

II. GENERAL ADAPTIVE FILTERS

Consider the following real-valued, strongly convex cost functions $J^{(i)}(\boldsymbol{w}^{(i)})$ for $i = 1, 2$, representing the expectation of loss functions $Q^{(i)}_{n+1}(\boldsymbol{w}^{(i)}; \boldsymbol{\chi}_{n+1})$:

$$J^{(i)}(\boldsymbol{w}^{(i)}) \triangleq \mathbb{E}\big\{Q^{(i)}_{n+1}(\boldsymbol{w}^{(i)}; \boldsymbol{\chi}_{n+1})\big\}, \tag{1}$$

where the expectation is assessed using the distribution of random data $\boldsymbol{\chi}_{n+1}$, which could adhere to nonlinear models. The subscript $n+1$ indicates the current time instant, and $\boldsymbol{w}^{(i)} \in \mathbb{R}^L$ is a real-valued parameter vector.

By iteratively minimizing $J^{(i)}(\boldsymbol{w}^{(i)})$ with $i = 1, 2$ over iterations, we may obtain two generic adaptive filters $F^{(1)}$ and $F^{(2)}$, respectively:

$$\boldsymbol{w}^{(i)}_{n+1} = f^{(i)}\Big\{\boldsymbol{w}^{(i)}_n, Q^{(i)}_{n+1}\big(\boldsymbol{w}^{(i)}_n; \boldsymbol{\chi}_{n+1}\big), \mu^{(i)}_{n+1}, \boldsymbol{\beta}^{(i)}_{n+1}\Big\} \tag{2}$$

where the generic update formula for the adaptive filter $F^{(i)}$ is denoted by $f^{(i)}\{\cdot\}$, the non-negative step-size parameter is denoted by $\mu^{(i)}_{n+1}$, and the state vector $\boldsymbol{\beta}^{(i)}_{n+1}$ integrates any extra

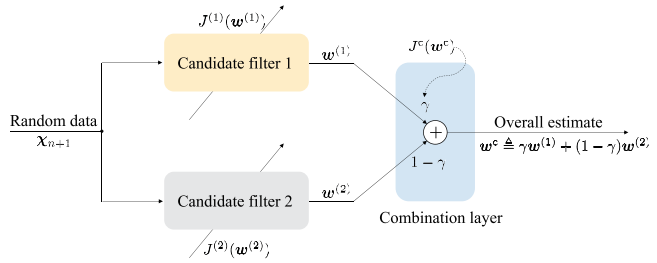


Fig. 1. Illustration of the combination framework for two general adaptive filters.

information necessary for filter adaptation. Also, take notice that in (2) we have used an adaptive technique to approximatively find the unavailable expectation term $\mathbb{E}\{Q_{n+1}^{(i)}(\cdot;\boldsymbol{\chi}_{n+1})\}$ by introducing the stochastic quantity $Q_{n+1}^{(i)}(\cdot;\boldsymbol{\chi}_{n+1})$.

III. AN AFFINE COMBINATION SCHEME

A. A combination framework for general adaptive filters

Fig. 1 shows the architecture of the combination framework for two general adaptive filters, which includes a combination layer and an adaptive filter layer that operate in tandem with one another. Two separate adaptive filters, $F^{(1)}$ and $F^{(2)}$, are considered for the adaptive filter layer; each is defined by an update equation (2). Each of these filters estimates the associated optimum weight vectors under cost functions $J^{(i)}(\boldsymbol{w}^{(i)})$ using the same data set $\boldsymbol{\chi}_{n+1}$. Then, in the combination layer, we link candidate filters $F^{(1)}$ and $F^{(2)}$ with combination coefficients γ and $1-\gamma$, respectively.

The goal of the combination layer is to learn which adaptive filter works better at each time instant, by adjusting γ in order to optimize the criterion $J^{c}(\boldsymbol{w}^{c})$ defined as:

$$J^{c}(\boldsymbol{w}^{c}) \triangleq \mathbb{E}\big\{Q_{n+1}^{c}(\boldsymbol{w}^{c};\boldsymbol{\chi}_{n+1})\big\}, \tag{3}$$

where $Q_{n+1}^{c}(\cdot;\boldsymbol{\chi}_{n+1})$ is a strongly convex loss function, and $\boldsymbol{w}^{c} \triangleq \gamma\boldsymbol{w}^{(1)} + (1-\gamma)\boldsymbol{w}^{(2)}$, with the superscript c highlighting that these are quantities at the combination layer. Please take note that in $J^{c}(\boldsymbol{w}^{c})$, the optimization variable is γ alone, whereas $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$ are estimates of candidate filters. It is also possible to establish and optimize each of the three generic cost functions, namely $J^{(i)}(\boldsymbol{w}^{(i)})$ for $i = 1, 2$ and $J^{c}(\boldsymbol{w}^{c})$, independently. Determining an approach to assess γ by minimizing cost function (3), given $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$, becomes the subsequent task.

B. An affine sign-gradient scheme

We propose that γ be assessed by attempting to minimize (3) using the values $\boldsymbol{w}_n^{(1)}$ and $\boldsymbol{w}_n^{(2)}$, that is:

$$\gamma_{n+1}^{\star} = \operatorname{argmin}_{\gamma}\ J^{c}\big(\gamma\boldsymbol{w}_n^{(1)} + (1-\gamma)\boldsymbol{w}_n^{(2)}\big). \tag{4}$$

We may determine that the best solution $\gamma_{n+1}^{\star}$ meets the above requirement by setting the derivative of $J^{c}\big(\gamma\boldsymbol{w}_n^{(1)} + (1-\gamma)\boldsymbol{w}_n^{(2)}\big)$ with respect to γ to zero:

$$\mathbb{E}\Big\{\nabla_{\gamma}Q^{c}\big(\gamma_n^{\star}\boldsymbol{w}_n^{(1)} + (1-\gamma_n^{\star})\boldsymbol{w}_n^{(2)};\boldsymbol{\chi}_{n+1}\big)\Big\} = 0, \tag{5}$$

where the derivative of the function $Q^{c}(\cdot)$ with respect to γ is represented by $\nabla_{\gamma}Q^{c}(\cdot)$. Because it necessitates knowledge of statistics, evaluating $\gamma_{n+1}^{\star}$ using (5) is not feasible.

We provide an adaptive strategy to solve this challenge. One must first determine an estimate at the combination layer and time instant $n+1$ before proceeding with this task:

$$\boldsymbol{w}_{n+1}^{c} \triangleq \big[\gamma\boldsymbol{w}_{n+1}^{(1)} + (1-\gamma)\boldsymbol{w}_{n+1}^{(2)}\big]\Big|_{\gamma=\gamma_{n+1}}, \tag{6}$$

where γ_{n+1} is an estimate of γ at time instant $n+1$.

By minimizing $J^{c}\big(\gamma\boldsymbol{w}_n^{(1)} + (1-\gamma)\boldsymbol{w}_n^{(2)}\big)$ in the direction of negative gradient with respect to γ and estimating the expectation terms with instantaneous values, the following affine sign-gradient scheme is obtained:

$$\begin{aligned}\gamma_{n+1} &= \gamma_n - \mu_{\gamma}\operatorname{sgn}\Big\{\nabla_{\gamma}J^{c}\big(\gamma\boldsymbol{w}_n^{(1)} + (1-\gamma)\boldsymbol{w}_n^{(2)}\big)\Big\}\Big|_{\gamma=\gamma_n} \\ &\approx \gamma_n - \mu_{\gamma}\operatorname{sgn}\big\{\nabla_{\boldsymbol{w}^{c}}^{\top}Q_{n+1}^{c}(\boldsymbol{w}_n^{c};\boldsymbol{\chi}_{n+1})\cdot(\boldsymbol{w}_n^{(1)} - \boldsymbol{w}_n^{(2)})\big\}\end{aligned} \tag{7}$$

where $\operatorname{sgn}\{\cdot\}$ is the sign function, and μ_{γ} is a positive step-size. Different from the affine combination of two LMS filters for which the power normalization scheme [16] and the sign regressor scheme have been introduced [11], to make it more resistant to changes in data $\boldsymbol{\chi}_{n+1}$ for the combination of two general adaptive filters, the sign gradient descent method is proposed in this paper, and equation (7) is derived by using the chain rule of derivatives.

C. A weight-copying trick

The weight-copying trick, first suggested in [26], is used to the combination of two generic adaptive filters to further improve its performance. To be more precise, at every time instant $n+1$, the parameter γ_{n+1}, which can be found using equation (7), shows which candidate filter performs better locally. As a result, the poorest candidate filter can take the weight vector from the best one. The weight-copying trick writes to:

$$\boldsymbol{w}_{n+1}^{(1)} = \begin{cases}\boldsymbol{w}_{n+1}^{(2)}, & \text{if } \gamma_{n+1} < 1-\beta \text{ and } \operatorname{mod}(n+1,N) == 0\\ \boldsymbol{w}_{n+1}^{(1)}, & \text{otherwise}\end{cases} \tag{8}$$

and

$$\boldsymbol{w}_{n+1}^{(2)} = \begin{cases}\boldsymbol{w}_{n+1}^{(1)}, & \text{if } \gamma_{n+1} > \beta \text{ and } \operatorname{mod}(n+1,N) == 0\\ \boldsymbol{w}_{n+1}^{(2)}, & \text{otherwise,}\end{cases} \tag{9}$$

where $0 \ll \beta < 1$ is a pre-defined threshold value with typical value 0.9, notation "==" denotes the logical equivalence operator, and a free integer $N \geq 2$. In equations (8) – (9), conditions $\gamma_{n+1} > \beta$ and $\gamma_{n+1} < 1-\beta$ imply that the weight-copying can occur when one candidate filter greatly outperforms another, and condition $\operatorname{mod}(n+1,N) = 0$ means that this trick is used periodically with period $N \geq 2$. The last step is to execute the weight-copying trick after evaluating (6). The findings will be employed in the next iteration.

For the ease of reference, the proposed algorithm is presented in Algorithm 1.

Algorithm 1 The proposed affine sign-gradient scheme with a weight-copying trick

Initialization: Set step-sizes $\mu_{n+1}^{(i)}$ and μ_γ, set integer $N \geq 2$ and $0 \ll \beta < 1$, initialize $\boldsymbol{w}_0^{(1)}$, $\boldsymbol{w}_0^{(2)}$ and γ_0; evaluate $\boldsymbol{w}_0^{\mathrm{c}} = \gamma_0 \boldsymbol{w}_0^{(1)} + (1-\gamma_0)\boldsymbol{w}_0^{(2)}$.

Iteration: Repeat the following steps for $n = 0, 1, 2 \cdots$:

1) Evaluate two generic adaptive filters $F^{(1)}$ and $F^{(2)}$ by using (2) for $i = 1, 2$.
2) Evaluate the combination coefficient γ_{n+1} by using (7).
3) Evaluate the estimate $\boldsymbol{w}_{n+1}^{\mathrm{c}}$ by using (6).
4) If conditions $\mod(n+1, N) == 0$ and $\gamma_{n+1} < 1 - \beta$ are satisfied simultaneously, then evaluate $\boldsymbol{w}_{n+1}^{(1)} = \boldsymbol{w}_{n+1}^{(2)}$.
5) If conditions $\mod(n+1, N) == 0$ and $\gamma_{n+1} > \beta$ are satisfied simultaneously, then evaluate $\boldsymbol{w}_{n+1}^{(2)} = \boldsymbol{w}_{n+1}^{(1)}$.

Output: Vectors $\boldsymbol{w}_{n+1}^{\mathrm{c}}$ for all n.

IV. SIMULATION RESULTS

In this section, we present simulation results to illustrate the proposed affine sign-gradient scheme and the weight-copying trick. All simulated curves were averaged over 100 Monte Carlo runs.

A. Combination of two sparsity-promoting adaptive filters

In the first experiment, we considered the following linear model as a special case:

$$y_{n+1} = \boldsymbol{x}_{n+1}^\top \boldsymbol{w}^\star + s_{n+1}, \tag{10}$$

in order to ensure that our proposed scheme works for linear model. A regression vector $\boldsymbol{x}_{n+1} \in \mathbb{R}^{35}$ was generated from a zero-mean Gaussian distribution with a unit covariance matrix, and an unknown sparse parameter vector $\boldsymbol{w}^\star \in \mathbb{R}^{35}$ was specified in equation (10). The additive noise s_{n+1} was also an i.i.d. zero-mean white Gaussian noise with variance $\sigma_s^2 = 0.01$. For candidate filters $F^{(1)}$ and $F^{(2)}$, two distinct cost functions were taken into account, specifically:

$$J^{(1)}(\boldsymbol{w}^{(1)}) = J_{\mathrm{MSE}}\big(\boldsymbol{w}^{(1)}\big) + \lambda \sum_{i=1}^{35} \log\Big(1 + \big|w_i^{(1)}\big|/\varepsilon\Big) \tag{11}$$

$$J^{(2)}(\boldsymbol{w}^{(2)}) = J_{\mathrm{MSE}}\big(\boldsymbol{w}^{(2)}\big) + \lambda \sum_{p=1}^{7} \log\Big[1 + \frac{\|\boldsymbol{w}_{\mathcal{G}_p}^{(2)}\|_2}{\varepsilon}\Big] \tag{12}$$

where $\lambda \geq 0$ was the regularization parameter, $\varepsilon > 0$ was a parameter for reweighted penalty, $w_i^{(1)}$ was the i-th entry of vector $\boldsymbol{w}^{(1)}$, $\boldsymbol{w}_{\mathcal{G}_p}^{(2)} \in \mathbb{R}^5$ denoted a subvector of $\boldsymbol{w}^{(2)}$ with entries indexed by $\mathcal{G}_p$, $\{\mathcal{G}_p\}_{p=1}^7$ was a partition of the index set $\mathcal{G} \triangleq \{1, \ldots, 35\}$, and $J_{\mathrm{MSE}}(\cdot)$ was the MSE cost defined as:

$$\begin{aligned} J_{\mathrm{MSE}}(\boldsymbol{w}) &\triangleq \mathbb{E}\big\{Q_{n+1}(\boldsymbol{w}; \boldsymbol{\chi}_{n+1})\big\} \\ &= \frac{1}{2}\,\mathbb{E}\big\{\|y_{n+1} - \boldsymbol{w}^\top \boldsymbol{x}_{n+1}\|^2\big\}, \end{aligned} \tag{13}$$

with quantities $\{y_{n+1}, \boldsymbol{x}_{n+1}\}$ standing for the random data $\boldsymbol{\chi}_{n+1}$. By using the stochastic subgradient descent to minimize (11) and (12), RZA-LMS algorithm and GRZA-LMS algorithm for sparse system identification had been derived in [27], [28], [29], [30] as:

$$\begin{aligned} \boldsymbol{w}_{n+1}^{(1)} &= \boldsymbol{w}_n^{(1)} + \mu^{(1)}\,(y_{n+1} - \boldsymbol{w}_n^{(1)\top}\boldsymbol{x}_{n+1})\boldsymbol{x}_{n+1} \\ &\quad - \lambda\mu^{(1)}\,\frac{\mathrm{sgn}\{\boldsymbol{w}_n^{(1)}\}}{\boldsymbol{\varepsilon} + |\boldsymbol{w}_n^{(1)}|} \end{aligned} \tag{14}$$

$$\begin{aligned} \boldsymbol{w}_{n+1,\mathcal{G}_p}^{(2)} &= \boldsymbol{w}_{n,\mathcal{G}_p}^{(2)} + \mu^{(2)}\,(y_{n+1} - \boldsymbol{w}_n^{(2)\top}\boldsymbol{x}_{n+1})\boldsymbol{x}_{n+1,\mathcal{G}_p} \\ &\quad - \lambda\mu^{(2)}\,\beta_{n,p}\boldsymbol{v}_{n,\mathcal{G}_p}, \quad \forall p = 1, 2, \cdots, 7 \end{aligned} \tag{15}$$

respectively, where the weighting coefficient $\beta_{n,p}$ was defined as:

$$\beta_{n,p} \triangleq \frac{1}{\|\boldsymbol{w}_{n,\mathcal{G}_p}^{(2)}\|_2 + \varepsilon} \tag{16}$$

subvector $\boldsymbol{v}_{n,\mathcal{G}_p}$ was defined as:

$$\boldsymbol{v}_{n,\mathcal{G}_p} = \begin{cases} \frac{\boldsymbol{w}_{n,\mathcal{G}_p}}{\|\boldsymbol{w}_{n,\mathcal{G}_p}\|_2} & \text{when } \|\boldsymbol{w}_{n,\mathcal{G}_p}\|_2 \neq 0 \\ 0 & \text{when } \|\boldsymbol{w}_{n,\mathcal{G}_p}\|_2 = 0, \end{cases} \tag{17}$$

and $\boldsymbol{x}_{n+1,\mathcal{G}_p} \in \mathbb{R}^5$ denoted a subvector of $\boldsymbol{x}_{n+1}$ with entries indexed by $\mathcal{G}_p$. Note that the division operator and the absolute value operator $|\cdot|$ in the third term on right-hand-side of (14) were applied in an element-wise manner, and $\boldsymbol{\varepsilon} \in \mathbb{R}^{35}$ was a vector with each entry being the scalar ε. Filters (14) and (15) were used as candidate filters $F^{(1)}$ and $F^{(2)}$, respectively, with step-sizes being 0.003 and 0.01. Regarding the combination layer, we determined that $J^{\mathrm{c}}(\boldsymbol{w}^{\mathrm{c}}) = J_{\mathrm{MSE}}(\boldsymbol{w}^{\mathrm{c}})$ and $\mu_\gamma = 0.015$. These values were then used to estimate γ_{n+1} by substituting them into equation (7), leading to:

$$\begin{aligned} &\gamma_{n+1} \approx \gamma_n + \\ &\mu_\gamma \mathrm{sgn}\big\{(y_{n+1} - \boldsymbol{x}_{n+1}^\top \boldsymbol{w}_n^{\mathrm{c}}) \cdot \boldsymbol{x}_{n+1}^\top (\boldsymbol{w}_n^{(1)} - \boldsymbol{w}_n^{(2)})\big\}. \end{aligned} \tag{18}$$

Fig. 2(a) displays the learning curves for two candidate filters, the affine sign-gradient scheme, and the weight-copying trick with respect to mean-square deviation (MSD). A better steady-state performance is associated with the RZA-LMS filter, but the GRZA-LMS filter exhibits a quicker initial convergence rate. The affine sign-gradient scheme proves its efficacy in the special case of linear model by integrating the RZA-LMS and GRZA-LMS algorithms to produce a learning curve with a lower steady-state MSD and a quicker initial convergence rate. In addition, the affine sign-gradient scheme can benefit from the weight-copying trick, which can enhance its convergence rate. Once again, the efficiency of the suggested affine sign-gradient scheme is validated by the evolution of parameter γ_n in Fig. 2(b).

B. Combination of two logistic regression learners

We took a nonlinear model into account in the second experiment. Assume that $\boldsymbol{x}_{n+1} \in \mathbb{R}^{15}$ was a real-valued random vector and that z_{n+1} was a streaming series of binary random variables with values of $+1$ or -1. In a machine learning

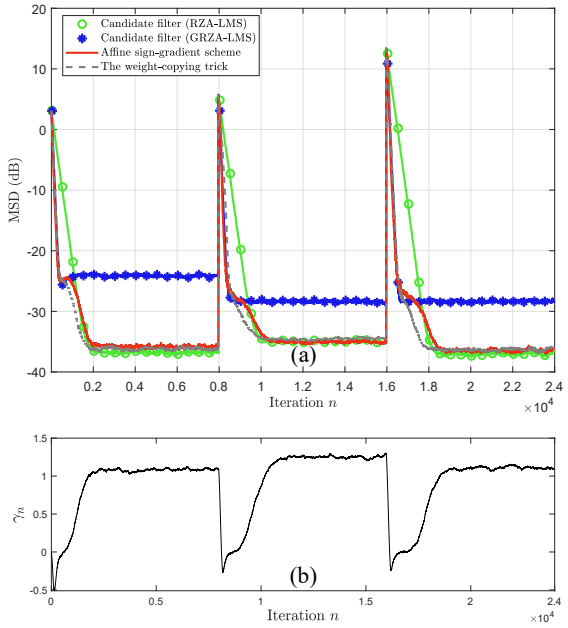


Fig. 2. Simulation results of the affine sign-gradient scheme and the weight-copying trick. (a) MSD learning curves and (b) evolution of combination coefficient γ_n for the affine sign-gradient scheme.

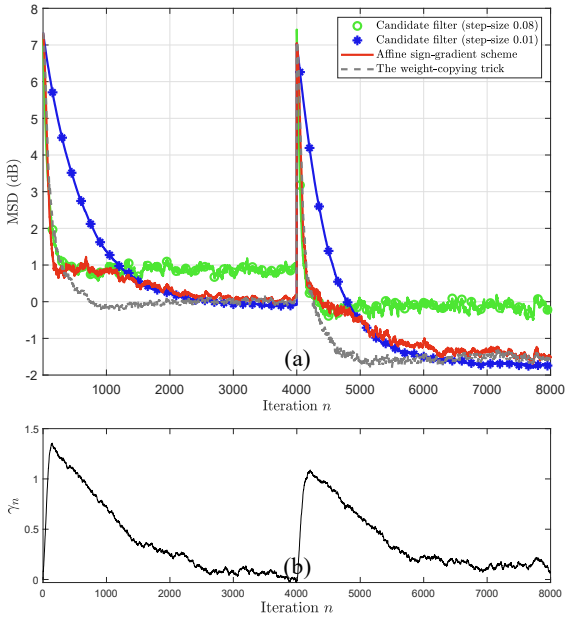


Fig. 3. Simulation results of the affine sign-gradient scheme and the weight-copying trick. (a) MSD learning curves and (b) evolution of combination coefficient γ_n for the affine sign-gradient scheme.

setting, the class to which feature vector $\boldsymbol{x}_{n+1}$ belonged was denoted by variable z_{n+1}. In these problems, we sought a vector $\boldsymbol{w}^\star$ that minimized the following regularized, strongly convex logistic risk function:

$$J_{\text{logis}}(\boldsymbol{w}) \triangleq \mathbb{E}\big\{Q_{\text{logis}}(\boldsymbol{w};\boldsymbol{\chi}_{n+1})\big\}, \tag{19}$$

with:

$$Q_{\text{logis}}(\boldsymbol{w};\boldsymbol{\chi}_{n+1}) \triangleq \frac{\rho}{2}\cdot\|\boldsymbol{w}\|_2^2 + \ln\big(1+e^{-z_{n+1}\boldsymbol{x}_{n+1}^\top\boldsymbol{w}}\big), \tag{20}$$

where $\rho > 0$ was the regularization parameter, notations $\ln(\cdot)$ and $e^{(\cdot)}$ were the logarithmic and exponential functions, respectively, and the combined quantities $\{z_{n+1},\boldsymbol{x}_{n+1}\}$ represented the random variable $\boldsymbol{\chi}_{n+1}$ appeared in (19). Usually, data $\{z_{n+1},\boldsymbol{x}_{n+1}\}$ were generated under an underlying optimal vector $\boldsymbol{w}^\star \in \mathbb{R}^{15}$ in a nonlinear fashion [24]. Additionally, we assumed a non-stationary environment in this experiment, which meant that $\boldsymbol{w}^\star$ may take on two different values at time instants $n \in [0, 3999]$ and $n \in [4000, 7999]$. We examined the merging of two logistic regression learners, $F^{(1)}$ and $F^{(2)}$, to verify the affine sign-gradient scheme. In particular, the stochastic gradient descent approach was used to minimize (19) in both $F^{(1)}$ and $F^{(2)}$. We could easily get to:

$$\boldsymbol{w}_{n+1}^{(i)} = \big(1-\mu_{n+1}^{(i)}\cdot\rho\big)\boldsymbol{w}_n^{(i)} + \frac{\mu_{n+1}^{(i)} z_{n+1}\boldsymbol{x}_{n+1}}{1+e^{z_{n+1}\boldsymbol{x}_{n+1}^\top\boldsymbol{w}_n^{(i)}}} \tag{21}$$

with $i = 1, 2$ for $F^{(1)}$ and $F^{(2)}$, respectively, since the derivative of $Q_{\text{logis}}(\boldsymbol{w};\boldsymbol{\chi}_{n+1})$ with regard to $\boldsymbol{w}$ was provided by:

$$\nabla_{\boldsymbol{w}} Q_{\text{logis}}(\boldsymbol{w};\boldsymbol{\chi}_{n+1}) = \rho\boldsymbol{w} - \frac{z_{n+1}\boldsymbol{x}_{n+1}}{1+e^{z_{n+1}\boldsymbol{x}_{n+1}^\top\boldsymbol{w}}}. \tag{22}$$

At all time instants, filters $F^{(1)}$ and $F^{(2)}$ had step-sizes of $\mu_{n+1}^{(1)} = 0.08$ and $\mu_{n+1}^{(2)} = 0.01$, respectively, which was the sole difference between them. Additionally, the cost function of combination layer was also set to (19), namely:

$$J_{\text{logis}}^{\text{c}}\big(\gamma\boldsymbol{w}_n^{(1)} + (1-\gamma)\boldsymbol{w}_n^{(2)}\big) \triangleq \mathbb{E}\big\{Q_{\text{logis}}\big(\gamma\boldsymbol{w}_n^{(1)} + (1-\gamma)\boldsymbol{w}_n^{(2)};\boldsymbol{\chi}_{n+1}\big)\big\}, \tag{23}$$

which was substituted into (7) to evaluate γ, resulting in:

$$\gamma_{n+1} \approx \gamma_n - \mu_\gamma \text{sgn}\Big\{\big(\rho\boldsymbol{w}_n^{\text{c}} - \frac{z_{n+1}\boldsymbol{x}_{n+1}}{1+e^{z_{n+1}\boldsymbol{x}_{n+1}^\top\boldsymbol{w}_n^{\text{c}}}}\big)^\top\big(\boldsymbol{w}_n^{(1)} - \boldsymbol{w}_n^{(2)}\big)\Big\} \tag{24}$$

Fig. 3 displays the outcomes of the simulation. Fig. 3(a) shows that two candidate filters with different steady-state MSDs and convergence rates are the result of using different step-sizes. The affine sign-gradient scheme reduces the steady-state MSD and speeds up the convergence rate by integrating the best features of two candidate filters. The affine sign-gradient scheme achieves a better convergence rate with the aid of the weight-copying trick. Furthermore, for the sake of demonstration, the change in parameter γ_n is shown in Fig. 3(b). All of the experimental findings show that the suggested affine sign-gradient scheme and the weight-copying trick work well with both nonlinear models and generic cost functions.

V. CONCLUSIONS

The affine combination scheme was successfully extended to generic cost functions in this paper, enabling a nonlinear model. All three generic cost functions might be specified and optimized separately inside this combination framework. We

proposed an affine combination scheme by using the stochastic sign-gradient descent, and a weight-copying trick had been introduced to enhance its performance. Simulation results were also provided to validate their effectiveness.

ACKNOWLEDGEMENT

The work of Danqi Jin was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 62401405, and in part by the 'Postdoctoral Excellence Programme' of Wuhan University. The work of Jie Chen was supported in part by the NSFC under Grants Nos. 62171380 and 62192713. The work of Gongping Huang was supported in part by the NSFC under Grant No. 62471340.

REFERENCES

[1] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1985.

[2] S. Haykin, *Adaptive Filter Theory*, Pearson Education India, 4th edition, 2005.

[3] A. H. Sayed, *Adaptive Filters*, John Wiley & Sons, Inc., 2008.

[4] D. Jin, J. Chen, C. Richard, and J. Chen, "Online proximal learning over jointly sparse multitask networks with $\ell_{\infty,1}$ regularization," *IEEE Trans. Signal Process.*, vol. 68, pp. 6319–6335, 2020.

[5] Y. Chen, D. Jin, J. Chen, C. Richard, W. Zhang, G. Huang, and J. Chen, "Online parameter estimation over distributed multitask networks with a rank-one model," in *32nd European Signal Processing Conference*, 2024, pp. 1042–1046.

[6] J. Benesty, T. Gnsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer Publishing Company, Incorporated, 1st edition, 2010.

[7] C. Paleologu, J. Benesty, and S. Ciochina, *Sparse adaptive filters for echo cancellation*, Morgan & Claypool, San Rafael, Calif, 2010.

[8] S. Haykin, *Adaptive Radar Signal Processing*, John Wiley & Sons, 2007.

[9] S. Bose, A. De, and I. Chakrabarti, "Framework for automated earthquake event detection based on denoising by adaptive filter," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 67, no. 9, pp. 3070–3083, 2020.

[10] M. Martinez-Ramon, J. Arenas-Garcia, A. Navia-Vazquez, and A. R. Figueiras-Vidal, "An adaptive combination of adaptive filters for plant identification," in *Proc. Int. Conf. Digital Signal Process.*, 2002, vol. 2, pp. 1195–1198.

[11] L. A. Azpicueta-Ruiz, A. R. Figueiras-Vidal, and J. Arenas-Garcia, "A normalized adaptation scheme for the convex combination of two adaptive filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Mar. 2008, pp. 3301–3304.

[12] S. Zhang, H. Zhao, and H. So, "Delayed combination of adaptive filters in colored noise," *IEEE Trans. Signal Process.*, vol. 70, pp. 1918–1931, 2022.

[13] D. Jin, J. Chen, C. Richard, J. Chen, and A. H. Sayed, "Affine combination of diffusion strategies over networks," *IEEE Trans. Signal Process.*, vol. 68, no. 1, pp. 2087–2104, Dec. 2020.

[14] D. Jin, J. Chen, C. Richard, J. Chen, and A. H. Sayed, "Convex combination of diffusion strategies over networks," *IEEE Trans. Signal Inf. Process. Networks.*, vol. 6, pp. 714–731, 2020.

[15] D. Jin, J. Chen, and J. Chen, "Convex combination of diffusion strategies over distributed networks," in *Proc. Asia-Pacific Signal Inf. Process. Association*, Hawaii, USA, Nov. 2018, pp. 224–228.

[16] J. Arenas-Garcia, L. A. Azpicueta-Ruiz, M. T. M. Silva, V. H. Nascimento, and A. H. Sayed, "Combinations of adaptive filters: Performance and convergence properties," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 120–140, Jan. 2016.

[17] S. S. Kozat, A. T. Erdogan, A. C. Singer, and A. H. Sayed, "Steady-state MSE performance analysis of mixture approaches to adaptive filtering," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4050–4063, Aug. 2010.

[18] L. Azpicueta-Ruiz, M. Zeller, A. Figueiras-Vidal, J. Arenas-Garcia, and W. Kellermann, "Adaptive combination of volterra kernels and its application to nonlinear acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 97–110, 2011.

[19] W. Lu, L. Zhang, J. Chen, and J. Chen, "Generalized combined nonlinear adaptive filters: From the perspective of diffusion adaptation over networks," *Signal Process.*, vol. 172, pp. 107507, 2020.

[20] C. Pan, J. Chen, J. Benesty, and G. Shi, "On the design of target beampatterns for differential microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1295–1307, 2019.

[21] J. Jin, X. Luo, G. Huang, J. Chen, and J. Benesty, "Beamforming through online convex combination of differential beamformers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 8561–8565.

[22] G. Huang, J. Benesty, I. Cohen, E. Winebrand, J. Chen, and W. Kellermann, "Switching kronecker product linear filtering for multispeaker adaptive speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[23] G. Huang, J. Benesty, and J. Chen, "Kronecker product multichannel linear filtering for adaptive weighted prediction error-based speech dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1277–1289, 2022.

[24] A. H. Sayed, *Adaptation, Learning, and Optimization over Networks*, vol. 7, Now Publishers Inc., Hanover, MA, USA, Jul. 2014.

[25] N. J. Bershad, J. C. M. Bermudez, and J. Y. Tourneret, "An affine combination of two LMS adaptive filters—transient mean-square analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1853–1864, May 2008.

[26] V. H. Nascimento and R. C. de Lamare, "A low-complexity strategy for speeding up the convergence of convex combinations of adaptive filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 3553–3556.

[27] Y. Chen, Y. Gu, and A. O. Hero, "Sparse LMS for system identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 3125–3128.

[28] J. Chen, C. Richard, Y. Song, and D. Brie, "Transient performance analysis of zero-attracting LMS," *IEEE Signal Process. Letters*, vol. 23, no. 12, pp. 1786–1790, Dec. 2016.

[29] D. Jin, J. Chen, C. Richard, and J. Chen, "Model-driven online parameter adjustment for zero-attracting LMS," *Signal Process.*, vol. 152, pp. 373–383, Nov. 2018.

[30] D. Jin, J. Chen, C. Richard, and J. Chen, "Adaptive parameters adjustment for group reweighted zero-attracting LMS," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, April 2018, pp. 4294–4298.

Psychological Driving Style Estimation from GPS Sensor Data Alone

Hiroto Horimoto*, Ryusei Kimura*, Takahiro Tanaka† and Shogo Okada*
* School of Computer Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: s2310148@jaist.ac.jp, ryusei_kimura@jaist.ac.jp, okada-s@jaist.ac.jp
† Institutes of Innovation for Future Society, Nagoya University, Aichi, Japan
E-mail: takahiro.tanaka@mirai.nagoya-u.ac.jp

***Abstract*—Automobiles are essential to society, but accidents involving older drivers are increasing. Driving Assistance Systems (DASs) have been developed to prevent such accidents. An essential goal is to build on this by implementing a system that provides adaptive assistance without burdening drivers with different characteristics. Accurately estimating driver characteristics is essential for this purpose. In contrast to studies that use in-vehicle sensor data through a Controller Area Network (CAN)-bus data, which requires additional equipment for data collection and is accessible to only a limited number of people, a more efficient and accessible approach is needed. Previous studies have shown that driving styles can be estimated using easily accessible Global Positioning System (GPS) data alone. We posit that the psychological driving style, which is the attitude and orientation for driving and affects the driver's behavior, can also be estimated from GPS data. In this study, we focus on developing a Driving Style Questionnaire (DSQ) as an indicator of the psychological driving style recognition model. The experimental results reveal differences in some features of all the GPS data between the high and low groups, which are divided by the median for each DSQ item on the two road types. Additionally, the findings indicate that the model with GPS data achieves the F1-macro greater than the random-assignment baseline on five out of eight of the DSQ items in the majority, and the average macro F1-score is 0.596, a difference of 0.036 from 0.632 for the model with CAN-bus data. The proposed GPS-based recognition model contributes to the development of personalized assistance systems without the burden of excessive sensor installation costs.**

I. Introduction

Automobiles are crucial to society, but they cause traffic accidents. This issue is more severe for older adults because as people age, their vulnerability to accidents or loss of life in a traffic accident increases. According to a Centers for Disease Control and Prevention report, about 7,500 older adults were killed in traffic crashes, and almost 200,000 were treated in emergency departments for crash injuries in 2020 [1].

Driving Assistance Systems (DASs), which provide vital information and warnings to drivers about circumstances that can improve driving safety and reduce traffic accidents, constitute a solution for preventing these issues.

However, these systems are generally designed for drivers in the current market [2]. This system leads to a potential mismatch between the driver and the system. One solution is to design support systems on the basis of understanding driver behavior [3], which is influenced by various factors, such as traffic, conditions, the environment, context, and personal characteristics [4]. Marafie et al. [5] reported that a personality-based driving agent improves the driver's experience. Kimura et al. [6] also reported that personalizing assistance and the basis of driver characteristics possibly improve drivers' behavior more effectively. These results indicate that an essential goal of DASs is to provide adaptive driving assistance suitable for drivers with different driving characteristics. To achieve this goal, using a Machine Learning (ML) model to accurately estimate drivers' characteristics, which are crucial in determining their behavior on the road, is the first step in implementing adaptive DASs without burden on a driver.

Some previous studies have focused on estimating driver characteristics using Controller Area Network (CAN)-bus data, a commonly used source for analyzing driving behavior [7], to provide a new direction for adaptive DASs. Wang et al. [8] proposed a prediction model for Big Five personality traits that with CAN-bus data. This study shows that driving data can automatically identify individual personality traits. Kimura et al. [9] estimated a driver's psychological characteristics, such as cognitive function, psychological driving style, and workload sensitivity from on-road driving data through a CAN-bus and a Global Positioning System (GPS) sensor. This study shows that the proposed model can accurately estimate a driver's cognitive function and characteristics.

While CAN-bus data are relatively accurate and frequent, obtaining driving sensor data through CAN-bus requires laborious processes like using data loggers for data collection [7]. Thus, it is difficult for all but a few companies or universities to collect scale data through a CAN-bus. A smartphone has a GPS sensor, and an accelerometer is a favorite platform for sensing a driver's behavior as another data source since it does not require additional equipment and is more straightforward than CAN-bus data. Furthermore, it is often used as hardware for a system to provide feedback according to analysis [5]. GPS sensor data are more reliable than accelerometer data since the orientation of a smartphone is not essential for obtaining GPS data [10]. Furthermore, they are easy to access and applicable to many vehicles/drivers [7].

In this context, some previous studies have focused on learning driving style representations, which generally aim to obtain latent representations of a driver's fine-grained driving habits [11] from GPS data and make use of such representations to identify drivers because the learned driving style

TABLE I
IN-VEHICLE SENSOR DATA THROUGH A CAN-BUS AND GPS SENSOR.

	Sensor Data	Unit
	CAN-bus	
1	Steering angle	deg
2	EPS torque	Nm
3	Forward acceleration	m/s^2
4	Lateral acceleration	m/s^2
5	Yaw rate	deg/s
6	Speed	km/h
7	Accelerator position	$\%$
8	Brake pressure	MPa
9	Fuel consumption	ml
	GPS sensor	
10	Longitude	deg
11	Latitude	deg

representations act as the "driver DNA" [12]. Chowdhury et al. [13] built an ML model for solving unique driver identification problems using smartphone GPS data as an indirect measurement of vehicle speed and acceleration. This study implies that a driver's driving style can be analyzed using GPS data. Additionally, accelerator pedal and steering wheel signals, which can be measured directly by a CAN-bus, reflect the interaction between the driver and the vehicle. In contrast, speed and acceleration, which can be determined from both CAN-bus data and GPS data, are measures of driving styles that reflect specific driving preferences and habits [8].

On the basis of these results, we posit that psychological driving style, which is the attitude and orientation for driving and affects a driver's behavior, can also be estimated from GPS data alone. In this study, we use the Driving Style Questionnaire (DSQ) [3], which is based on a self-report questionnaire, as an indicator of psychological driving style. In [6], the effectiveness of the driving support is suggested to depend on the drivers' scores on the DSQ scale. We classified drivers with high and low scores of DSQ items.

Hence, as a novel challenge, we aim to develop a recognition model of psychological driving style from on-road driving data through a GPS sensor via ML models in this study.

II. DATASET

In this study, we use a dataset provided by the Institutes of Innovation for Future Society of Nagoya University [14], the same as that used in [9]. The dataset was collected from 24 older adults (12 males and 12 females) aged 50 to 79 years (the average age was 66 years). They drove a car equipped with several in-vehicle sensors through the CAN-bus and real-time kinematics GPS sensor (JAVAD Delta-3) twice each on the same public road in Japan, and 48 sets of driving session data were collected. We use 9 CAN-bus data and GPS sensor data for comparison experiments. Table I details 11 CAN-bus and GPS sensor data.

However, after removing incomplete or obviously inferior data, we retained 32 sets of driving session data from 23 drivers. Hence, for some drivers, one set of driving sessions is utilized. The Ethical Committee of Nagoya University approved all procedures in this study. Informed consent was obtained from all drivers before the experiments were conducted.

TABLE II
DRIVING STYLE QUESTIONNAIRE (DSQ).

	Item
1	Confidence in driving skill
2	Hesitation for driving
3	Impatience in driving
4	Methodical driving
5	Preparatory maneuvers at traffic signals
6	Importance of automobile for self-expression
7	Moodiness in driving
8	Anxiety about traffic accidents

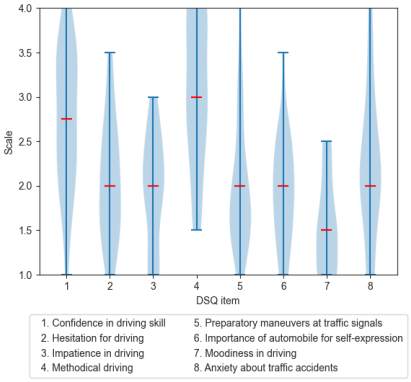


Fig. 1. Violin plots of the DSQ scores. The DSQ unit is on a scale from 1 to 4, and each violin plot represents each DSQ item. Each red line indicates the median of 23 drivers. The values are 2.5, 2.0, 2.0, 3.0, 2.0, 2.0, 1.5 and 2.0, respectively.

The driving tests were conducted on public roads. All participants first drove on an arterial road and then circulated in a residential area. The driving duration ranged from 2,324 s to 4,762 s, with an average of 2,907 s, and the mileage ranged from 10,079 m to 14,810 m, with an average of 12,109 m.

In addition to the driving tests, the drivers answered the DSQ, which is based on a self-report questionnaire and was introduced by [3] for characterizing driving style from a psychological aspect. Table II details the items of the DSQ. The DSQ includes eight items measured on a scale from 1 to 4. Each item is associated with two questions, and their mean value represents the score for the item. Fig. 1 shows violin plots of the scores for each DSQ item.

III. METHOD

This section presents a psychological driving style estimation method that uses GPS data alone. An overview of our model is shown in Fig. 2.

A. Data Segmentation

Estimating psychological characteristics is challenging because we do not know when or where differences in drivers' characteristics are exhibited. Kimura et al. [9] hypothesized that it is possible to estimate drivers' psychological characteristics by focusing on the segmentation of road types

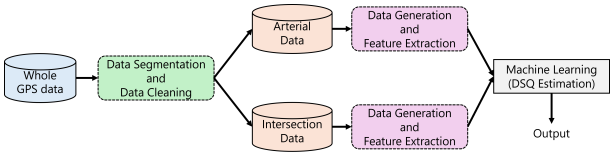


Fig. 2. Overview of our model. Arrows indicate GPS data flow. First, whole GPS data is segmented according to road type and cleaned. Then, primary and secondary GPS data are generated, and features are extracted. Finally, the DSQ is estimated via ML models with processed GPS data.

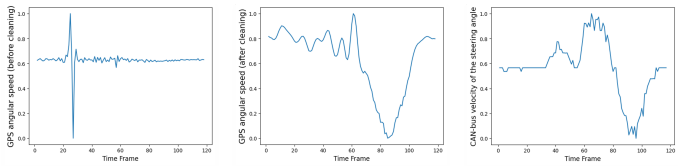


Fig. 3. Example of time-series GPS angular speed before (left) and after (middle) our data cleaning method and CAN-bus velocity of the steering angle (right) when a vehicle curves at an intersection

and partial driving data because the ability to drive safely varies across different road types. This study indicates that the estimation accuracy of the DSQ items by a method based on the hypothesis is higher than that of the method with whole driving data.

On the basis of this result, we largely followed the model introduced by [9]. First, we classified the whole GPS data into two road types, arterial roads and intersections, using the vehicle position obtained through the GPS sensor. The intersections are located in a residential area. Furthermore, we eliminated incomplete intersection data and different driving directions at the same intersection. We treated all these intersections as the same road types, but features were extracted separately because the visibility, ease of driving, and directions were different. Furthermore, when the data are arterial roads, we arranged the segment points so that the average number of seconds for each segment was in the set [ALL, 60, 30, 15, 10, 5, 3]. "All" means that no division is used. In [9], the intersection data were also segmented by brake sensor data. However, we did not do this because we only used GPS data in this study.

B. Data Cleaning

In the following section, we discuss the main challenges that arise from working with GPS data. The following speed, heading, longitudinal acceleration, and angular speed in this section are described in detail in section III-C

The first challenge is to process point cloud data. Point clouds mostly occur at very low speeds or if the vehicle stands still in one place for some time [15]. Those points do not represent any car movement and lead to high angular speed. To address this problem, when the speed was very low ($< 2km/h$), the points were removed, and then the points were filled via a linear interpolation method.

The second challenge is outliers. The set of driving session data we use contains an enormous number of GPS data points with inferior accuracy. GPS signals are affected by events such as urban tunneling and sensor error. Thus, we designed a filtering method to address this problem. First, outliers were removed, and then their points were filled via a linear interpolation method, the same as that used for the point cloud. We used an outlier detection method based on the Interquartile Range (IQR) considering longitudinal acceleration and angular speed data and set k = 2.25 to identify outliers. All the GPS data points were subsequently filtered with a Butterworth filter [16] because the smallest deviations in latitude and longitude can lead to high lateral and longitudinal acceleration [15]. The filter was a low pass filter with a degree of 2 and a cutoff frequency of 1 Hz. The filter was applied bidirectionally in the forward and reverse directions to cancel the phase delay in the data. Finally, for the arterial road data, we used the one with the smallest number of outliers detected from the filtered data for some drivers with two sets of driving sessions because large missing parts or GPS data with inferior accuracy were still found. On the other hand, we removed only the intersection data if outliers were still detected from the filtered data for the same reason as in the arterial road data.

We used one data at the arterial roads and four straight and two curve data at the intersections where all drivers' data were recorded after this cleaning in the experiments.

Fig. 3 shows one example of the time-series GPS angular speed before and after our data cleaning method and the CAN-bus velocity of the steering angle, which is the first derivative with respect to time for the steering angle when a vehicle curves at an intersection. These time-series data had point clouds ranging from 23 to 30 in each frame. Since the CAN-bus data and GPS data are not the same but similar, we scaled these data on a scale of 0 to 1 to focus only on the shape of these time series data. The point clouds considerably affect GPS angular speed before data cleaning. However, after data cleaning, the data behave like the CAN-bus velocity of the steering angle, which can be confirmed.

C. Data Generation and Feature Extraction

In this dataset, GPS data are collected at approximately 10 Hz. They consist of attributes such as the timestamp, longitude, and latitude.

We employed six primary and secondary basic data types to capture driving styles: speed, longitudinal acceleration, angular speed, lateral acceleration, longitudinal jerk, and lateral jerk. When two points' longitudes and latitudes are given, the distance (m) and the heading (deg) between the two points can be computed. The heading represents the current direction compared with North, and its range is $(0, 360]$. We computed the first derivative of the distance with respect to time to obtain the speed (m/s). Furthermore, we implemented a moving mean filter to make them smooth for speed and heading.

Few important secondary data are computed from these primary GPS measurements, namely, speed and heading. The longitudinal acceleration (m/s^2) can be computed from consecutive speed measurement samples via the equation ex-

pressed as:

$$acc_{long} = \frac{\Delta v}{\Delta t} \tag{1}$$

where Δv is the speed variation ($v_n - v_{n-1}$), and Δt is the temporal variation. The GPS data in our dataset have a sampling frequency of approximately 10 Hz, so Δt is approximately 100 msec. The angular speed (rad/s), which represents the rate of change in heading, can be computed from consecutive heading measurement samples via the expression specified by [13]:

$$ang_{speed} = \frac{\Delta \theta}{\Delta t} \cdot \frac{\pi}{180} \tag{2}$$

where $\Delta\theta$ is the heading variation. Given the vehicle's heading, before computing the angular speed we modified the heading variation via the following equation:

$$f(\Delta\theta) = \begin{cases} \Delta\theta - 360 & \text{if } \Delta\theta > 180, \\ 360 + \Delta\theta & \text{if } \Delta\theta < -180, \\ \Delta\theta & \text{otherwise.} \end{cases} \tag{3}$$

The lateral acceleration (m/s^2) can be evaluated via the expression specified in [13] and [17]:

$$acc_{lat} = \frac{v^2}{R} = v \cdot ang_{speed} \tag{4}$$

where v is the current time value and R is the turn radius, as introduced by [17]. The jerk, which represents the rate of change in acceleration, can be computed from consecutive accelerations. Therefore, we compute the first derivative with respect to time for the longitudinal acceleration and the lateral acceleration and use them as the longitudinal jerk (m/s^3) and the lateral jerk (m/s^3), respectively. Thus, the above equations can compute secondary data from primary GPS data.

The statistical features of the generated primary and secondary GPS data were extracted to estimate the DSQ via ML models. These statistical features are the mean, median, skewness, kurtosis, variance, and max. Finally, 36 (six data × six statistics) features were extracted. When a driver has multiple data at the same intersection, we computed the mean value of each feature.

D. Machine Learning Model

We used logistic regression with L2 regularization, a linear support vector machine, and random forest models to estimate the DSQ results. We split scores of the DSQ items on the basis of the median value to create binary classification labels as shown in Fig. 1 and then conduct binary classification. The model outputs the DSQ items' class categories (high or low).

IV. Experimental Settings

As an evaluation criterion for the classification model, we report the macro F1-score. We conduct binary classification for DSQ estimation. The parameter C values of the logistic regression and linear support vector machine are selected from the set [0.001, 0.01, 0.1, 1, 10, 100]. The maximum depth of the tree of the random forest is selected from the set [3, 5, 7, 9, 11]. We perform an exhaustive search of specified parameter values to select the best value in the training set. We compute the feature significance of a real-valued feature to a binary targets (high or low) as a p-value on the Kolmogorov-Smirnov (KS) test, which is a nonparametric hypothesis test that statistically examines differences in distribution. A $p < 0.05$ was considered statistically significant in this study. Therefore, we only include features that are statistically significant for each fold to avoid overfitting and reduce computational cost because there are 9036 features on arterial road data and 216 features on intersection data from the GPS sensor. We use leave-one-person-out cross-validation to evaluate the classification model.

In addition, we evaluate our model's effectiveness in estimating the DSQ from GPS features by comparing its feature set with those from the CAN-bus data detailed in Table I. The input signals from these CAN-bus data are resampled at 10 Hz to match the GPS data. In addition to the above CAN-bus data, we compute and use the first derivative with respect to time for the steering angle, forward acceleration, lateral acceleration, and accelerator position and use them as the velocity of steering angle, forward jerk, lateral jerk, and change of the acceleration position are used as secondary data. We also applied data segmentation and feature extraction to the CAN-bus data. After GPS data cleaning, all feature sets use the same driving data at arterial roads and intersections.

V. Results and Discussion

We applied the KS test to all drivers' statistical features to confirm the importance of GPS data for DSQ estimation (section V-A). Next, we investigate how the model with GPS data can estimate the DSQ results to validate our method and compare feature sets with GPS data and CAN-bus data (section V-B). This study evaluates the estimation macro F1-score by comparing the following two feature sets.

FS (GPS): a GPS feature set that is from all six primary and secondary GPS data, which is our method.

FS (CAN-bus): a CAN-bus feature set that is from all 13 primary and secondary CAN-bus data.

Fig. 4 shows bar plots of the ratio of GPS data with features highly related to the DSQ scores with $p < 0.05$ on the KS test. Table III shows the classification result of each DSQ item of a model with a GPS feature set and a CAN-bus feature set. The bold values represent the highest macro F1-score among each DSQ item in each road type with values greater than 0.5. We compute an average macro F1-score from the highest macro F1-score among each DSQ item from two road types within each feature set.

A. Importance of GPS Data

From the results of the KS test shown in Fig. 4, some features of all the GPS data were different between the high and low groups, with statistically significant differences in the arterial road data and intersection data when all the DSQ items were considered. In particular, the ratio of the speed with features was highest (0.275) on "importance of automobile for self-expression" on arterial road data, while

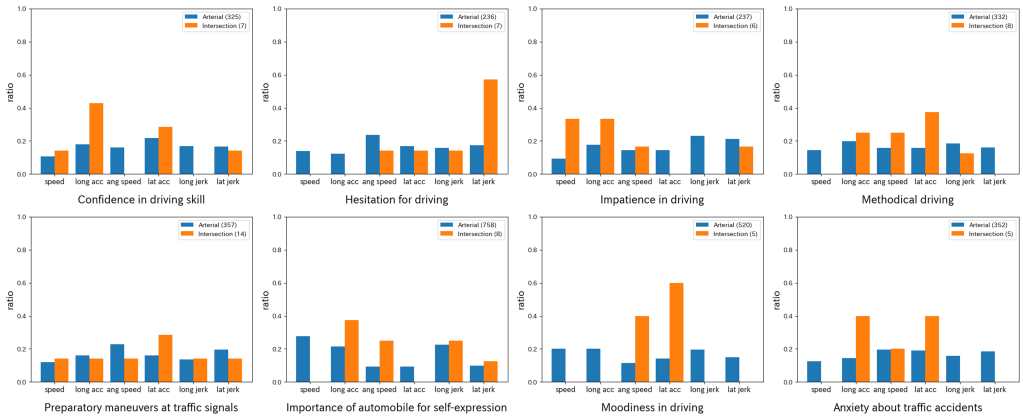


Fig. 4. Bar plots of the ratio of GPS data with features highly related to the DSQ scores with $p < 0.05$ on the Kolmogorov-Smirnov test. Blue (left) and orange (right) bars represent arterial roads and intersections. Legend numbers in parentheses show total statistically significant features. long acc = longitudinal acceleration, ang speed = angular speed, lat acc = lateral acceleration, long jerk = longitudinal jerk, lat jerk = lateral jerk.

the lateral acceleration with features was highest (0.600) on "moodiness in driving" on intersection data. This correlation between each DSQ item, which characterizes driving styles from a psychological aspect, and the primary and secondary GPS data on driving styles has significant implications for understanding and predicting driver behavior. This result also suggests that these important GPS features can be used to classify each DSQ item into groups via ML models while avoiding overfitting. Furthermore, 236 (2.61%) to 758 (8.39%) features from 9036 features on the arterial road data and 5 (2.31%) to 14 (6.48%) features from 216 features on the intersection data remained important GPS features for each DSQ item. This test contributes to reducing computational costs and improving cost-effectiveness.

B. Analysis of Classification Accuracy

From the classification result of each DSQ item of a model with FS (GPS) shown in Table III, three on each road type and five on two road types out of eight DSQ items were estimated, with the highest macro F1-score greater than 0.5, which is the random-assignment baseline. In particular, the macro F1-score of "importance of automobile for self-expression" and "moodiness in driving" were highest (0.625) on arterial road data, whereas "confidence in driving skill" was highest (0.775) on intersection data. Moreover, the model's average macro F1-score was greater than 0.5. These results indicate that the majority of the DSQ items, which represent the psychological driving style, can be adequately estimated from primary and secondary GPS data on driving styles.

In contrast, a model with FS (CAN-bus) achieved the highest macro F1-score greater than 0.5: six on arterial road data, three on intersection data, and seven on two road types out of eight DSQ items. The model's average macro F1-score was higher than the model with FS (GPS). These results indicate that the difference in the average macro F1-score between FS (GPS) and FS (CAN-bus) methods was close to 0.036. However, it also indicates that the model with FS (GPS) was considerably inferior to that with FS (CAN-bus) compared with the number of items greater than 0.5.

In addition, the results reveal a large difference between the macro F1-scores of the model with FS (GPS) and FS (CAN-bus) on "methodical driving" on the arterial road data. Therefore, we applied the KS test to all drivers' FS (CAN-bus) and found that forward acceleration (0.151), lateral acceleration (0.123), and steering angle (0.114) presented high ratios. One possible reason why the macro F1-scores for FS (GPS) and FS (CAN-bus) differ significantly while the top three features in the CAN-bus are of a similar data type to the generated GPS data is that the generated GPS data, which represent the driving style, are not exactly identical in behavior to the CAN-bus data. Fig. III also shows that the model with FS (GPS) had higher macro F1-scores than did the model with FS (CAN) for some DSQ items, suggesting that the generated GPS data are important for these specific DSQ items.

VI. Conclusions

In this paper, we present the results of our study on the psychological driving style estimation approach using only GPS sensor data and compare the models' results with CAN-bus data. The experimental results show differences in some features of all the GPS data between the high and low groups for each DSQ item on two road types via the KS test. Additionally, the findings indicate that the model with a GPS feature set can achieve a macro F1-score greater than 0.5 for five out of eight DSQ items in the majority and the average macro F1-score is 0.596 a difference of 0.036 from 0.632 for the model with a CAN-bus feature set. Our proposed GPS-based recognition model has the potential to significantly reduce the cost of developing personalized assistance systems, as it does not require expensive sensor installations. This study used GPS sensor data collected at approximately 10 Hz, whereas smartphone GPS data are typically collected at 1 Hz. To develop personalized DASs using smartphones alone in the future, it is recommended that we investigate whether a model with the data can estimate psychological driving style.

TABLE III
CLASSIFICATION MACRO F1-SCORE FOR DSQ ITEMS.

DSQ item	FS (GPS)						FS (CAN-bus)					
	Arterial			Intersection			Arterial			Intersection		
	LR	SVM	RF	LR	SVM	RF	LR	SVM	RF	LR	SVM	RF
Confidence in driving skill	0.300	0.258	0.395	0.517	0.436	**0.775**	0.361	0.361	0.410	0.303	0.436	0.343
Hesitation for driving	0.329	0.425	0.361	0.487	0.384	0.343	0.617	**0.654**	0.395	0.281	0.436	0.487
Impatience in driving	0.417	0.327	0.425	0.343	0.361	0.462	0.517	**0.617**	0.425	0.436	0.436	0.395
Methodical driving	0.324	0.258	0.395	0.549	**0.662**	0.343	0.617	**0.744**	0.410	0.329	0.357	0.410
Preparatory maneuvers at traffic signals	0.394	**0.544**	0.324	0.654	**0.692**	0.521	**0.673**	0.617	0.462	**0.712**	0.487	0.673
Importance of automobile for self-expression	0.487	0.555	**0.625**	0.487	0.303	0.357	**0.635**	0.456	0.549	0.300	0.452	0.281
Moodiness in driving	**0.625**	0.487	0.462	0.258	0.324	0.462	**0.589**	**0.589**	0.378	0.673	**0.713**	0.625
Anxiety about traffic accidents	0.238	0.361	0.439	0.378	0.361	0.378	0.410	0.425	0.425	**0.546**	0.378	0.439
Average	0.596						0.632					

LR = Logistic Regression; SVM = Linear Support Vector Machine; RF = Random Forest.
The bold values represent the highest macro F1-score among each DSQ item on each road type with values greater than 0.5. The values in the last row represent the average macro F1-score from the highest macro F1-score among each DSQ item from two road types within each feature set.

REFERENCES

[1] Centers for Disease Control and Prevention, *Web-based Injury Statistics Query and Reporting System (WISQARS)*, https://wisqars.cdc.gov/. Accessed 21 Jul 2024, 2022.

[2] C. M. Martinez, M. Heucke, F.-Y. Wang, B. Gao, and D. Cao, "Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 666–676, 2017.

[3] M. Ishibashi, M. Okuwa, S. Doi, and M. Akamatsu, "Indices for characterizing driving style and their relevance to car following behavior," in *SICE Annual Conference 2007*, IEEE, 2007, pp. 1132–1137.

[4] M. Akamatsu, Y. Sakaguchi, M. Okuwa, T. Kurahashi, and K. Takiguchi, "Database of driving behavior measured with equipped vehicles in the real road environment," in *Proceedings of International Ergonomic Association*, 2003.

[5] Z. Marafie, K.-J. Lin, D. Wang, *et al.*, "Autocoach: An intelligent driver behavior feedback agent with personality-based driver models," *Electronics*, vol. 10, no. 11, 2021.

[6] R. Kimura, H. Horimoto, T. Tanaka, and S. Okada, "Investigating the effect of driving support robot depending on driver characteristics," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 608–612.

[7] M. N. Azadani and A. Boukerche, "Driving behavior analysis guidelines for intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6027–6045, 2022.

[8] Y. Wang, N. Zhao, X. Liu, S. Karaburun, M. Chen, and T. Zhu, "Identifying big five personality traits through controller area network bus data," *Journal of Advanced Transportation*, vol. 2020, p. 8 866 876, 2020.

[9] R. Kimura, T. Tanaka, Y. Yoshihara, K. Fujikake, H. Kanamori, and S. Okada, "Estimating driver personality traits from on-road driving data," *IEEE Access*, vol. 11, pp. 93 679–93 690, 2023.

[10] A. Chowdhury, T. Chakravarty, and P. Balamuralidhar, "Estimating true speed of moving vehicle using smartphone-based gps measurement," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2014, pp. 3348–3353.

[11] N. Lin, C. Zong, M. Tomizuka, P. Song, Z. Zhang, and G. Li, "An overview on study of identification of driver behavior characteristics for automotive control," *Mathematical Problems in Engineering*, vol. 2014, pp. 1–15, 2014.

[12] U. Fugiglando, P. Santi, S. Milardo, K. Abida, and C. Ratti, "Characterizing the "driver dna" through can bus data analysis," in *Proceedings of the 2nd ACM International Workshop on Smart, Autonomous, and Connected Vehicular Systems and Services*, 2017, pp. 37–41.

[13] A. Chowdhury, T. Chakravarty, A. Ghose, T. Banerjee, and P. Balamuralidhar, "Investigations on driver unique identification from smartphone's gps data alone," *Journal of Advanced Transportation*, vol. 2018, p. 9 702 730, 2018.

[14] Y. Yoshihara, E. Takeuchi, and Y. Ninomiya, "Accurate analysis of expert and elderly driving at blind corners for proactive advanced driving assistance systems," in *Transportation Research Board 95th Annual Meeting, no. 16-1992*, 2016.

[15] S. Würtz and U. Göhner, "Driving style analysis using recurrent neural networks with lstm cells," *Journal of Advances in Information Technology*, vol. 11, pp. 1–9, 2020.

[16] S. Butterworth, "On the Theory of Filter Amplifiers," *Experimental Wireless & the Wireless Engineer*, vol. 7, pp. 536–541, 1930.

[17] R. Vaiana, T. Iuele, V. Astarita, *et al.*, "Driving behavior and traffic safety: An acceleration-based safety evaluation procedure for smartphones," *Modern Applied Science*, vol. 8, pp. 88–96, 2014.

Physical Domain Adversarial Attacks Against Source Printer Image Attribution

Nischay Purnekar, Benedetta Tondi, Mauro Barni
University of Siena, Italy
E-mail: nischay.purnekar@student.unisi.it, benedetta.tondi@unisi.it, mauro.barni@unisi.it

***Abstract*—Deep learning networks are vulnerable to adversarial examples, specifically subtle, human-imperceptible modifications that can deceive them. While most research has focused on digital adversarial attacks, in several applications it is necessary that the adversarial examples operate in the physical domain. Physical domain adversarial examples are usually crafted to ensure that the artifacts introduced by the attack survive the digital-to-analog and analog-to-digital transformations involved in such attacks. In this paper, we introduce an approach to generate adversarial examples against a source printer attribution system, aiming to determine which printer was used to print a given image document. With respect to conventional physical domain attacks, attacking a source printer attribution system poses additional challenges since the subtle features the attribution network relies on are introduced again during the Print and Scan (P&S) process that follows the attack, thus possibly nullifying the attack. We address this challenge by applying Expectation Over Transformation, including within the pool of transformations a simulation of the P&S process relying on two Generative Adversarial Network models trained for this purpose. Experimental results demonstrate that our approach yields a significant increase in attack success rates, surpassing those of baseline models.**

I. Introduction

Despite their effectiveness, Artificial Intelligence (AI) systems based on Deep Learning (DL) are vulnerable to various malicious attacks, including adversarial examples [1], backdoor attacks [2], and inversion attacks [3]. Adversarial examples involve subtle, human-imperceptible perturbations that lead to misclassifications or other incorrect behaviors. Most research has focused on pixel-level digital adversarial examples [4], assuming the attacker has full control over the image's digital representation. In contrast, physical adversarial examples [5], [6] exploit variations in texture, shape, and lighting, processed through the system's sensor inputs. Examples include specific patterns applied to physical objects like stop signs, which cause misidentification by autonomous vehicle perception systems. Despite the potential risks, there is significantly less research on generating and defending against physical adversarial attacks compared to digital ones.

This paper focuses on attacks against image-printed document authentication, which is crucial for legal, governmental, and financial sectors that handle sensitive and confidential information. Ensuring document integrity is vital to prevent forgery and fraud, as these can have significant consequences. The Federal Trade Commission reported 2.6 million fraud cases, resulting in $10.3 billion in losses in 2023 [7] due to piracy. Maintaining printed document authenticity and confidentiality is essential for protecting sensitive information and upholding trust in official processes. Within this framework, the goal of this paper is to study the vulnerability of an image printer source attribution classifier based on DL against physical adversarial examples. The classifier is trained to identify a document's originating printer using a diverse set of documents from various printers. We aim to generate adversarial examples that remain effective after reprinting by applying different attack algorithms. Traditionally, adversarial examples in the physical domain are created by adding perturbations directly to digital images, which are then transformed into a physical document or 3D object and fed to the AI model, successfully misleading the system. In our case, the attacked digital images are printed again by the same printer and scanned back before being fed to the classifier. The Print and Scan (P&S) process applied to the attacked images poses several challenges to the creation of effective attacks. Firstly, the P&S process degrades the perturbation introduced by the attack, thus requiring a stronger perturbation. Secondly, and most importantly, the features the attribution network relies on are re-introduced when the attacked digital image is printed for the second time, possibly nullifying the effectiveness of the attack.

Following previous work on the generation of physical adversarial examples, we use Expectation Over Transformation (EOT) [8] to craft perturbations that survive the distortion introduced when transitioning from the physical to the digital domain. As shown in Sect. V, classical EOT alone is not sufficient to maintain the effectiveness of adversarial examples after the P&S process due to the reintroduction of printing artifacts on top of the attacked image. For this reason, we propose incorporating a P&S simulator within the EOT framework to generate an adversarial attack that preemptively accounts for the subsequent reprinting process. In particular, we used a Pix2Pix Generative Adversarial Network (GAN) [9] and a CycleGAN [10] to simulate the P&S transformation. We integrated EOT with P&S into the Iterative Fast Gradient Sign Method (IFGSM) and the Carlini & Wagner (C&W) attacks, achieving a high Attack Success Rate (ASR) even after reprinting.

Given the above, the main contributions of this work are:

1) We developed two P&S simulators utilizing Pix2Pix GAN and CycleGAN image translation models. The simulators are not only applicable for crafting adver-

sarial source printer image attribution but can also be potentially used in digital image forensics to enhance the robustness of synthetic image detectors [11].
2) We integrated the P&S simulators as an additional transformation step in the EOT attack.
3) We were able to generate robust adversarial examples that withstand the reprinting process, successfully deceiving the target source printer attribution classifier.

The paper is organized as follows: Sect. II reviews adversarial attacks in digital and physical domains. Sect. III details the development and performance of the P&S simulators. Sect. IV focuses on the generation of robust adversarial examples. Sect. V analyzes the experimental results. Sect. VI summarizes our findings and suggests directions for future work.

II. Related Work

Adversarial examples, first identified by Szegedy et al. [1], demonstrate that minor perturbations can significantly alter a network's output while remaining nearly imperceptible to humans. These examples often generalize across different models, even those trained with varying hyperparameters or architectures, sparking significant interest in DNNs. Current research explores adversarial examples in both digital and physical domains.

Digital Domain adversarial perturbations are directly applied to the network's input, constraining the l_p-norm (e.g., l_∞-norm [4], l_2-norm, and l_0-norm [12]) of the perturbation to be lower than a certain threshold to maintain the imperceptibility of the attack. Depending on the adversary's knowledge, the adversarial attacks can be categorized as either white-box or black-box. In white-box attacks, the attacker has complete knowledge of the model, allowing full use of the gradient to craft the perturbations [4], [5], [12]. In a black-box attack, the attacker can only query the target model and receive corresponding outputs without access to its internal structure. In this scenario, the attacker can either leverage the generalizability of adversarial examples across different models or deduce the model's internal information through multiple queries.

Physical Domain attacks were pioneered by Kurakin et al. [5], who introduced the first physical-domain attack by printing digitally perturbed images, which were then photographed with a smartphone and fed into a pre-trained Inception v3 classifier. Their results indicated a decline in the effectiveness of the attack after the images underwent printing and photography. Sharif et al. [6] created adversarial eyeglass frames to deceive facial recognition systems by incorporating a non-printability score (NPS) and total variation (TV) loss in their optimization, ensuring printer accuracy and smooth color transitions. [13] used a similar TV loss to generate adversarial stickers for hats to fool the ArcFace system. Lu et al. [14] noted that attack effectiveness diminishes when images are viewed from different angles and distances. To generate physical adversarial examples that withstand the transformations involved when going back and forth from the digital to the physical domain, EOT was introduced in [8]. Eykholt et al. [15] refined EOT with Robust Physical Perturbation (RP2), sampling synthetic and physical transformations to create adversarial stop signs using posters or stickers, although this requires printing and photographing the original image multiple times. Jan et al. [16] proposed D2P, a transformation using a conditional GAN [9], [10] before EOT to simulate printing and photographing effects, but it faces feasibility issues due to the need for extensive printing and photographing to build a training dataset. A work that is somewhat similar to the present work is [17]. Even there, the detector relies on the features that are reintroduced after rebroadcast hence requiring the design of a particular EOT strategy. However, the rebroadcasting artifacts are different from those introduced by P&S, hence the method proposed in [17] cannot be applied in our case.

As a matter of fact, all the attacks based on EOT include natural geometric and color transformations to generate robust adversarial examples. As we will show later, however, this is not enough when the target system is a printer source attribution model. For this reason, we integrated the P&S simulators into the EOT framework. In this way, we were able to significantly improve the ASR, ensuring that the attack remains effective even after reprinting.

III. Print and Scan Simulation

Printing and scanning an image involves converting the digital images to physical copies and back to the digital domain, introducing various distortions and artifacts. Printing can cause color shifts, ink diffusion, and minor geometric distortions due to the printer's mechanical characteristics and type of paper used. Scanning adds further distortions and noise depending on the scanner's resolution, color response, and mechanical misalignments. These steps affect pixel values and introduce artifacts specific to the printer and scanner, along with minor geometric alterations due to imperfect paper positioning within the scanner.

Given the time-consuming and costly nature of manually creating large volumes of printed and scanned images, we developed two P&S simulators to be directly included within the EOT process, enabling the vast generation of training images without the expense and effort of physical P&S. Research on simulating the P&S process by means of deep learning is sparse. A significant contribution in this domain comes from Ferrara et al. [18], who demonstrated that integrating a simulated P&S transformation during training improves the accuracy of face morphing attacks on printed and scanned face images. Their model estimates the pixel distortions incurred during printing and scanning, considering various critical parameters such as the responsivity of the acquisition device, the sampling function characterizing the digitization process of printed images, the point spread function of the printer and scanner, noise levels, and color transformations. However, the presence of device-dependent unknown parameters complicates real-world adaptations, as calculating the point spread functions of printers and scanners is challenging, and fine-tuning each parameter can be time-consuming, especially across multiple devices. Mitkovski et al. [19] also utilized

a Pix2Pix GAN to emulate the P&S process for biometric applications.

To start with, and similarly to [19], we trained a Pix2Pix GAN [9] simulator. Training the Pix2Pix GAN, however, requires pixel-wise alignment of digital and P&S images for effective computation of the mean square error loss. To address this problem, we employed image alignment techniques during training. We also trained a CycleGAN P&S simulator, which supports unpaired image-to-image translation. In fact, CycleGAN does not necessitate paired images, thus greatly simplifying the preparation of the dataset.

A. Architecture of the simulators

Pix2Pix and CycleGAN have been extensively used to address various generative tasks. In our case, the objective of the Pix2Pix generator is to translate the input images from the digital to the P&S domain, while the discriminator is asked to distinguish between real P&S and digital pairs and their synthetic counterparts. The CycleGAN generators aim to translate images from the digital to the P&S domain and vice versa, ensuring cyclic consistency. With respect to classical CycleGAN training, we did not use the identity loss. In fact, printing a printed and scanned image again should not result in the identity operator, as the second P&S process would further degrade the image quality.

B. Dataset

To train the simulators, we used a dataset derived from the second version of the VIPPrint dataset [20]. This dataset consists of human face images printed with various modern color laser printers, each operating at different resolutions. Acquisition was performed using a TaskAlfa 3551 multifunctional scanner at 600×600 dpi resolution, and the images were saved using lossless compression. The size of the digital images is 1024×1024×3, while the P&S images are approximately 2036×2038×3, with slight variations of 5 to 10 pixels in both dimensions introduced during scanning. To align the resolutions of digital and P&S images, the digital images were upsampled to match the P&S image resolution. Our experiments focused on a subset of P&S images printed by one of the 12 printers in the VIPPrint dataset, specifically a Kyocera P5021 CDN Color Laser printer. We used a subset of 200 printed and scanned images from this printer for our experiments. To match the input size of the Pix2Pix and CycleGAN networks, we trained the networks on image patches extracted from 100 printed and scanned images along with their corresponding digital images. The patches were 256×256×3 in size and were extracted without pixel overlap. For Pix2Pix, we aligned the digital and printed and scanned patches using [21]. If alignment was challenging or significant pixel differences were detected, the corresponding patch was skipped. This approach yielded 4,678 aligned digital and P&S patches. In contrast, CycleGAN training utilized unaligned digital and P&S patches, leveraging the ability of CycleGAN's to handle unpaired image data. In total, 4,914 digital and P&S patches were used to train the CycleGAN simulator.



Fig. 1. Examples of digital and simulated P&S images with corresponding ground truth. The first column shows the digital image input, the second and third columns display P&S patches simulated by Pix2Pix GAN and CycleGAN respectively, and the last column shows the ground-truth P&S patches.

C. Training

The Pix2Pix GAN simulator was trained for 800 epochs using the Adam optimizer with parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a learning rate of 1×10^{-4}. The network utilized 64 filters and a Leaky ReLU activation function with a slope of 0.2, while the batch size was restricted to 1. For training CycleGAN, we used the same hyperparameters as the Pix2Pix GAN simulator over 600 epochs. Both the GAN adversarial loss and cyclic consistency loss weights were set to 10. After training, we assessed the performance of both simulators by inputting original digital patches. To introduce variability, we added Gaussian noise with zero mean and variance of 0.0625 to the digital images before feeding them to the simulator. This ensured that multiple inputs of the same digital image yielded slightly different simulated outputs, mimicking real-world variations when an image undergoes printing and scanning multiple times.

We evaluated the quality of the simulated images both visually (Fig. 1) and quantitatively using metrics such as the Structural Similarity Index (SSIM) and Fréchet Inception Distance (FID) (Table I). The SSIM scores are 0.84 for Pix2Pix GAN and 0.87 for CycleGAN, while the FID scores are 47 for Pix2Pix GAN and 45 for CycleGAN. As shown in Fig. 1, the images generated by the P&S simulators closely resemble the corresponding ground-truth images, demonstrating their effective learning of the distortions inherent in the P&S process.

TABLE I
SSIM AND FID BETWEEN SIMULATED AND REAL P&S IMAGES.

P&S Simulator	SSIM Score ↑	FID Score ↓
Pix2Pix GAN	0.84	47
CycleGAN	0.87	45

IV. GENERATION OF PHYSICAL ADVERSARIAL EXAMPLES

A. Threat Model

We consider an attack aiming to alter an image printed by a specific printer, *P*, in such a way that the printer source attribution model can no longer identify *P* as the source printer (untargeted attack) after the image is reprinted by *P*. The challenge is to ensure the attack's effectiveness even after the attacked image undergoes reprinting and scanning. The

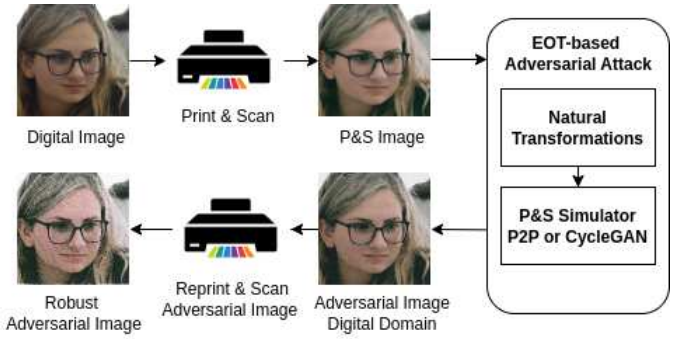


Fig. 2. Attack pipeline for the generation of robust adversarial examples.

attacker has white-box access to the source attribution model, including its weights and architecture. This allows the attacker to optimize and evaluate the adversarial examples in the digital domain before executing the physical-world attack by printing, scanning, and strategically placing the attacked images.

B. Targeted Model

The printer source attribution system targeted by our attack is the one described in [22]. This system, trained on the VIPPrint dataset [20], analyzes the 10 highest-energy 224×224×3 patches of the image and uses a majority voting decision for classification. Preliminary experiments in [22] showed that a basic reprinting black-box attack can deceive the original classifier. To enhance resilience against such attacks, the authors fine-tuned the model using a dataset of reprinted images, resulting in a more robust (hardened) source attribution model, which is the focus of our attack. Since the classifier operates on patches, the adversarial attacks are applied to 224×224×3 image patches. However, because the attack may slightly alter the energy of the patches, the classifier could potentially analyze different patches after the attack. Therefore, we decided to attack all patches in the image. This approach also prevents the introduction of visible discontinuities at patch borders. The success of the attack hinges on inducing sufficient patch misclassifications to misclassify the true printer as the most voted option. Our target is specifically the Kyocera-ecosys P5021cdn laser printer, identified as class #12 in the attribution system's multiclass classification.

C. Attack Pipeline

The attack pipeline (Fig. 2) begins with printing a digital image and applying an adversarial attack in the digital domain. To maintain the effectiveness of the adversarial perturbation after printing and scanning, we used EOT with P&S simulation. The adversarial digital image is then physically printed with the same printer. Finally, the source attribution model scans and analyzes the printed image to identify its origin.

1) Digital Domain Attack: Initially, we assessed the effectiveness of digital domain attacks (without EOT) in inducing misclassifications when the attacked image is subsequently printed and scanned. Adversarial examples were generated using a non-targeted version of I-FGSM [5] and C&W attack [12]. For I-FGSM, we set $\varepsilon = 0.03$, with a step size of 0.01 over 100 iterations. Similarly, for the C&W attack, we let $\varepsilon = 0.1$, with a binary search step size of 9 and a learning rate of 0.01 across 1000 iterations. These hyperparameters were selected to ensure effective attack coverage across most of the patches in the P&S image.

2) Physical Domain Attack: To create robust adversarial examples in the physical world, we integrated the I-FGSM and C&W attacks into an EOT framework, effectively addressing the domain shifts between digital and physical domains. EOT involves defining a pool of transformations T to simulate these shifts. The transformations used in our EOT attack are detailed in Table II, including their parameters, essential for replicating practical domain shifts. Additionally, we incorporated the Pix2Pix and CycleGAN P&S simulators within the transformation set. Results were averaged over 10 transformed samples to assess attack effectiveness. Through extensive experiments, we identified an optimal combination of transformations T (Table II) that consistently produce successful adversarial examples. Our setup includes rotation (2.0 to 10.0 degrees), zoom blur (factors between 1.05 and 1.10), and pixel shifts (5 pixels in all directions) with an inclusion probability of 100%. For color transformations, brightness deltas (10 to 40) and a fixed contrast factor of 0.3 are applied with 50% probability. Additionally, either CycleGAN or Pix2Pix GAN simulators are chosen with a probability of 50% to simulate P&S effects.

The attack algorithms within the EOT framework were configured with the following hyperparameters: for I-FGSM, $\varepsilon = 0.15$, a step size of 0.03, and 500 iterations were used; for C&W, we employed $\varepsilon = 0.15$, 9 binary search steps, a learning rate of 0.01, and 1000 iterations. When we incorporated the P&S simulators into EOT, I-FGSM utilized $\varepsilon = 0.4$, a step size of 0.07, and 500 iterations, while for C&W we used $\varepsilon = 0.4$, 9 binary search steps, a learning rate of 0.01, and 1000 iterations.

TABLE II
SET OF TRANSFORMATIONS USED IN THE EOT ATTACK.

Transformations	Parameter	Values	Probability
Brightness	brightness delta	[10, 40]	50%
Contrast	contrast factor	0.3	50%
Rotation	rotation angle	[2°, 10°]	100%
Zoom	zoom range	[1.05, 1.10]	100%
Shift of Pixels	# pixels (all directions)	5	100%
Pix2Pix P&S Simulator	-	-	50%
CycleGAN P&S Simulator	-	-	50%

V. EXPERIMENTAL RESULTS

To demonstrate the robustness of the hardened source attribution classifier against adversarial examples, we conducted experiments in both the digital and physical domains. Our study involved benchmarking various white-box attacks, including I-FGSM and C&W, both with and without EOT, and incorporating P&S simulators within the EOT transformations. These experiments were performed on a test set of 20 documents, with each document segmented into 81 patches,



(a) First P&S image (b) I-FGSM (c) I-FGSM(EOT) (d) I-FGSM(EOT+PS) (d) C&W (d) C&W(EOT) (d) C&W(EOT+PS)

Fig. 3. Examples of attacked images. From left to right, we show the original P&S image and the generated adversarial examples (after reprinting and scanning) using standard I-FGSM and C&W adversarial attacks, EOT with natural transformations, and EOT including P&S simulation.

totaling 1,620 attacked patches. To measure the strength of the perturbation introduced by the attacks, we computed the Peak Signal-to-Noise Ratio (PSNR) between the original and attacked patches, both before and after the reprinting and scanning process. PSNR calculations were focused only on successfully attacked patches. Additionally, we evaluated the ASR across all patches in the images and on the top 10 highest energy patches, which are generally more challenging to attack. After reprinting, the final classification is determined through majority voting on the results obtained from the top 10 highest energy patches of each document. To assess the overall robustness of the system, we also computed the ASR after the majority voting, where the printer with the largest number of votes among the top energy patches is selected.

The results of our experiments are reported in Table III. Analyzing the second column of the Table III, we observe that all attacks are highly effective when they are applied in the digital domain, achieving nearly 100% ASR. As expected, the attacks incorporating EOT, particularly those utilizing P&S simulation, exhibit lower PSNR values. The fourth column of the table, reports the effectiveness of the attacks in the physical domain, considering the ASR obtained on *all* patches after reprinting. For standard I-FGSM and C&W, the ASR decreases dramatically, while the application of EOT with natural transformations limits the ASR drop. Including the P&S simulators in the EOT transformations further improves the ASR to 80.49% for C&W and 87.83% for I-FGSM, which is a significant advantage with respect to EOT with natural transformations. The main advantage of including P&S simulation within EOT becomes apparent when we limit the analysis to the 10-highest energy blocks of each image. In this scenario, the ASR with natural EOT is only 33.5% for I-FGSM and 28% for C&W, while EOT with P&S simulation allows to attack 69% and 56.5% of the patches (for the attack to be successful it is necessary - though not sufficient - that at least 50% of the blocks are attacked). In the last column of the table, we report the ASR after majority voting on the 10-highest energy blocks, which is crucial for assessing the overall attack effectiveness on the entire image rather than on individual blocks. We observe that the ASR after majority voting drops to negligible values for standard I-FGSM and C&W attacks, showing only slight improvement with EOT using natural transformations [1]. However, when the P&S simulator is incorporated to EOT, the ASR significantly increases for both I-FGSM and C&W attacks. Specifically, the ASR for I-FGSM rises from 25% to 70%, and from 20% to 65% for C&W. These results highlight the effectiveness of incorporating the P&S simulator, given the complexity of creating adversarial examples that survive the reprinting process. Our experiments also suggest that patches with dark backgrounds tend to reintroduce stronger artifacts upon reprinting, thus requiring an excessive distortion.

TABLE III
EFFECTIVENESS OF VARIOUS ATTACKS IN BOTH THE DIGITAL AND PHYSICAL DOMAIN. ASR'S ARE AVERAGED ACROSS ALL PATCHES OF THE IMAGES, ON THE TOP 10 HIGHEST ENERGY PATCHES OF EACH IMAGE AND AFTER MAJORITY VOTING ON THE 10 HIGHEST ENERGY PATCHES.

Attack Method	ASR Digital	PSNR (dB)	ASR Printed **All Patches**	ASR Printed **Top10 Patches**	PSNR (dB)	ASR Printed **Majority Voting**
I-FGSM	100%	36.28	26.72%	15.5%	28.89	10%
I-FGSM (EOT)	96.41%	20.55	77.16%	33.5%	17.25	25%
I-FGSM (EOT+P&S)	100%	13.02	87.83%	69%	11.89	70%
CW	100%	34.25	21.48%	14%	25.53	10%
CW(EOT)	97.16%	19.86	63.70%	28%	16.96	20%
CW (EOT+P&S)	100%	12.28	80.49%	56.5%	11.18	65%

In Fig. 3, we present adversarial examples after reprinting, generated using various attacks. The images include the initial P&S image (the attack target), adversarial examples produced by standard attacks, EOT attacks with natural transformations, and EOT attacks incorporating P&S simulations. Comparing the initial P&S images to the reprinted adversarial examples generated by standard I-FGSM or C&W attacks we see that reprinting weakens the perturbation. The examples produced by I-FGSM(EOT+PS) and CW(EOT+PS) demonstrate the importance of the P&S simulation in creating robust adversarial examples that withstand reprinting.

[1]These results indirectly support the choice made in [22] to base the classification only on the highest energy patches.

VI. Concluding remarks

In our research, we addressed the challenges associated with generating robust adversarial examples against a source printer attribution system. We introduced a novel attack that integrates P&S simulations within the EOT framework. By employing Pix2Pix GAN and CycleGAN models, we developed two simulators that accurately replicate the P&S transformations. The integration of these simulators into the EOT framework significantly increased the ASR, demonstrating the method's effectiveness in producing adversarial examples that survive reprinting. Our work underscores the importance of physical domain adversarial attacks in AI security research and provides a foundation for future efforts to counteract such threats.

Future work will focus on developing defenses, such as adversarial training techniques that incorporate examples of images subjected to the proposed physical domain attack. We also plan to expand our simulators to address various image processing tasks under diverse environmental conditions. Additionally, we aim to develop advanced P&S simulators using diffusion models for enhanced realism and accuracy. Finally, we will explore vision transformers to challenge existing printer attribution systems.

Acknowledgment

This work was partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan, funded by the European Union - NextGenerationEU, and the FOSTERER project, funded by the Italian Ministry of University and Research (PRIN 2022 program, contract 202289RHHP).

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations*, 2014.

[2] W. Guo, B. Tondi, and M. Barni, "An overview of backdoor attacks against deep neural networks and possible defences," *CoRR*, vol. abs/2111.08429, 2021.

[3] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations*, 2015.

[5] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *5th International Conference on Learning Representations*, 2017.

[6] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[7] Federal Trade Commission, "Consumer sentinel network databook," Tech. Rep., 2023.

[8] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning,*, 2018.

[9] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017.

[10] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision, ICCV*, 2017.

[11] N. Purnekar, L. Abady, B. Tondi, and M. Barni, "Improving the robustness of synthetic images detection by means of print and scan augmentation," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec*, 2024.

[12] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy, SP 2017*, 2017.

[13] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face ID system," in *25th International Conference on Pattern Recognition*, 2020.

[14] J. Lu, H. Sibai, E. Fabry, and D. A. Forsyth, "NO need to worry about adversarial examples in object detection in autonomous vehicles," vol. abs/1707.03501, 2017.

[15] I. Evtimov, K. Eykholt, E. Fernandes, *et al.*, "Robust physical-world attacks on machine learning models," *CoRR*, vol. abs/1707.08945, 2017.

[16] S. T. K. Jan, J. Messou, Y. Lin, J. Huang, and G. Wang, "Connecting the digital and physical world: Improving the robustness of adversarial attacks," in *The Thirty-Third AAAI Conference on Artificial Intelligence,*, 2019.

[17] B. Zhang, B. Tondi, and M. Barni, "Adversarial examples for replay attacks against cnn-based face recognition with anti-spoofing capability," *Comput. Vis. Image Underst.*, vol. 197-198, p. 102 988, 2020.

[18] M. Ferrara, A. Franco, and D. Maltoni, "Face morphing detection in the presence of printing/scanning and heterogeneous image sources," *IET Biometrics*, 2021.

[19] A. Mitkovski, J. Merkle, C. Rathgeb, *et al.*, "Simulation of print-scan transformations for face images based on conditional adversarial networks," in *BIOSIG*, 2020.

[20] A. Ferreira, E. Nowroozi, and M. Barni, "Vipprint: A large scale dataset of printed and scanned images for synthetic face images detection and source linking," *CoRR*, vol. abs/2102.06792, 2021.

[21] G. D. Evangelidis and E. Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1858–1865, 2008.

[22] A. Ferreira and M. Barni, "Attacking and defending printer source attribution classifiers in the physical domain," in *Pattern Recognition, Computer Vision, and Image Processing. ICPR*, 2022.

An Annealing-Inspired Gradient-Descent Based Suboptimal Solver for Combinatorial Problems

Shu-Ping Chang[1], Cheng-Che Lee[2], Hsin-Jung Lee[3], Chieh-Hsiung Kuan[4]
Jason Gemsun Young[5], Chia-Yu Yao[6], and Jian-Jiun Ding[7]*
[1, 4] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University
[2, 3, 4] Graduate Institute of Electronics Engineering, National Taiwan University
[5] Electronic and Optoelectronic System Research Laboratories, Industrial Technology Research Institute
[6] Department of Electrical Engineering, National Taiwan University of Science and Technology
[7]* Graduate Institute of Communication Engineering, National Taiwan University
E-mail: r10945044@ntu.edu.tw [1], d06943010@ntu.edu.tw [2], d04943010@ntu.edu.tw [3], chkuan@ntu.edu.tw [4], jasonyoung@itri.org.tw [5], chyao@mail.ntust.edu.tw [6], jjding@ntu.edu.tw [7*], Tel: +886-2-33669652

***Abstract*—Combinatorial optimization problems, such as IC layout and industrial scheduling, have significant industrial applications but are challenging due to exponential time complexity. In this work, we propose a novel annealing-inspired heuristic algorithm that treats combinatorial problems as function optimization problems using nonlinear programming. The proposed gradient-descent-based solver significantly improves the convergence rate and includes a new regularization constraint to escape local minima by increasing convexity. Applied to the Traveling Salesman Problem (TSP) with various city counts, the proposed algorithm demonstrates polynomial time complexity. It much reduces the complexity from $(n-1)!/2$ to n^4 and has a marked improvement in computation efficiency. Notably, for a 50-city TSP, the relative error is just around 5%, indicating the accuracy and efficiency of the proposed algorithm in solving high-dimensional instances.**

I. Introduction

In nature, many processes seek stable states, such as the evolution and changes in ecological environments or macroscopic phenomena resulting from the balance of microphysical laws. These processes can be considered natural optimization mechanisms. Over the past 50 years, various global optimization algorithms have been developed to simulate these processes, including Simulated Annealing, Neural Networks, and Genetic Algorithms [1]. Combinatorial optimization problems are crucial due to their wide range of applications, such as IC layout, industrial scheduling, and isomer analysis. The Traveling Salesman Problem (TSP) is a prominent example frequently studied in mathematics and computer science. Traditional methods like dynamic programming and branch-and-bound aim to solve these problems but often require significant computational resources.

However, many combinatorial optimization problems are NP-hard or NP-complete, making exhaustive search impractical within a reasonable time frame. The exponential growth in computational complexity remains a significant challenge, especially for high-dimensional instances.

In this study, we develop an annealing-inspired algorithm to address the exponential growth in computational complexity in combinatorial optimization problems. Our method treats these problems as function optimization tasks using nonlinear programming, with the TSP as a key example. Inspired by Finnila's Quantum Annealing in 1994 [2], which utilizes quantum tunneling, we propose leveraging pathological system characteristics to achieve states close to multiple solutions. This approach, followed by continuous unconstrained optimization, transforms the problem into a universal combinatorial solver, applicable to TSP, IC layout, industrial scheduling, and heterogeneous structure analysis.

Our approach employs gradient descent to solve ill-conditioned systems, freely choosing methods like the conjugate gradient to improve convergence speed. Inspired by the framework in [3], we introduce a regularization term similar to thermal annealing, enabling the solver to escape local minima by gradually increasing cost function convexity. We validate the feasibility of our algorithm through TSP of different dimensions, a standard test model in mathematics and computer science. This paper emphasizes the algorithm's uniqueness and generality, presenting it as an effective strategy for improving combinatorial optimization solutions.

II. Basic Principles Of Our Approach

The Traveling Salesman Problem (TSP) can be transformed into a quadratic polynomial problem through nonlinear programming, representing it as a binary variable optimization problem [4]. Let TSP be defined as n cities in the set $N \equiv \{1,2,\dots n\}$. Without loss of generality, let the nth city be both the starting point and the endpoint, referred to as the "Hometown," since the degree of freedom for an n-city TSP is effectively only $n-1$.

A. *Using Route Length for Unconstrained Optimization*

After excluding the hometown city, the remaining number of cities is denoted as $m = n - 1$, forming a city set $C \equiv \{1,2,\dots m\}$, and $T \equiv \{1,2,\dots m\}$ as the ordinal set representing travel time. These correspond to the order of visiting cities in the set C. Through these two sets, we generate a set of binary variables x_{tj} $(t \in T, i \in C)$, defined as follows:

$$x_{ti} \equiv \begin{cases} 1, & \text{if city } i \text{ is visited at time order } t \\ 0, & \text{otherwise} \end{cases}, 1 \le t, i \le m \tag{1}$$

x_{ti} has k variables, where $k = m^2$, capable of expressing one possible path among all permutations, representing the city i at time t. Finally, all variables are placed into a $k \times 1$ matrix, $\mathbf{x} \in \{0,1\}^k$. The optimized cost function $J_1(\mathbf{x})$ can then be expressed as:

$$\mathbf{x} \equiv [x_{11}\ x_{12}\ \dots\ x_{1m}\ \dots\ x_{m1}\ x_{m2}\ \dots\ x_{mm}]^T \tag{2}$$

$$J_1(\mathbf{x}) \equiv \sum_{t\in T\setminus\{m\}} \sum_{i\in C} \sum_{j\in C\setminus\{i\}} d_{ij} x_{ti} x_{(t+1)j} + \sum_{t\in\{1,m\}} \sum_{i\in C} c_i x_{ti} \tag{3}$$

The set of paths between two cities is denoted as $E = C^2$. Among them, d_{ij} $((i,j) \in E)$ is the distance between city i and city j, and c_i $(i \in C)$ is the distance from city i to the hometown. This way, the optimized cost function can fully express the distance of one possible path.

Since the number of distances d_{ij} between cities is $m(m-1)/2$, and each route contains n distances between cities, the average route length L_{avg} is approximately:

$$L_{\text{avg}} \equiv \frac{2n}{m(m-1)} \left(\sum_{i\in C} \sum_{j\in C\{i\}} d_{ij} + \sum_{i\in C} c_i \right) \tag{4}$$

The new cost function $J_2(\mathbf{x})$ is the normalized $J_1(\mathbf{x})$:

$$J_2(\mathbf{x}) \equiv J_1(\mathbf{x}) / L_{\text{avg}} \tag{5}$$

B. *Using Regularization to Incorporating Constraints*

We utilize regularization to maintain our problem as unconstrained optimization by incorporating regularization terms to incorporate constraints. In Fig. 1, each row is associated with the condition that only one city is visited at a time, meaning that only one binary variable in each row equals 1, and the others are 0. Similarly, each column is associated with the condition that each city is visited only once, meaning that only one binary variable in each column equals 1, and the others are 0. Therefore, we first transform the TSP into a constrained optimization problem:

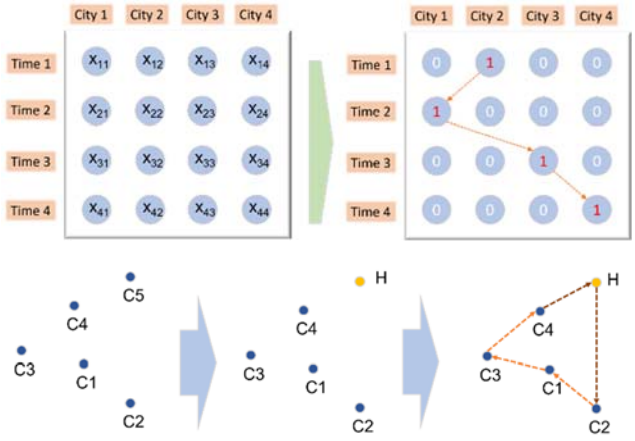


Fig. 1. Traveling Salesman Problem (n = 5) variable presentation form. The arrangement of the remaining cities is expressed using an $n \times n$ matrix, where columns represent the cities and rows represent different time sequences.

Problem TSP:

Minimize

$$J_{\text{TSP}}(\mathbf{x}) \equiv \frac{1}{L_{\text{avg}}} \left(\sum_{t\in T\setminus\{m\}} \sum_{i\in C} \sum_{j\in C\setminus\{i\}} d_{ij} x_{ti} x_{(t+1)j} + \sum_{t\in\{1,m\}} \sum_{i\in C} c_i x_{ti} \right) \tag{6}$$

subject to

$$\sum_{j\in C} x_{tj} = 1, t \in T, \quad \sum_{t\in T} x_{tj} = 1, j \in C,$$
$$x_{tj} \in \{0,1\}, t \in T, j \in C. \tag{7}$$

The above two types of constraints are transformed into the following minimization regularization terms, which require the sum of each row and each column to be close to 1. Therefore, the number of added regularization terms will be the sum of the number of rows and columns, 2m. However, currently, it is not required for all solution elements to be binary values; solutions may be floating-point numbers between 0 and 1:

$$\left(\sum_{j\in C} x_{tj} - 1\right)^2, \ \left(\sum_{t\in T} x_{tj} - 1\right)^2. \tag{8}$$

According to the regularization method, by adding the above regularization terms, a new cost function $J_3(\mathbf{x})$ is generated:

$$J_3(\mathbf{x}) \equiv J_2(\mathbf{x}) + \lambda_1 \left[\sum_{t\in T} \left(\sum_{j\in C} x_{tj} - 1 \right)^2 + \sum_{j\in C} \left(\sum_{t\in T} x_{tj} - 1 \right)^2 \right] \tag{9}$$

Since $J_3(\mathbf{x})$ is a quadratic polynomial function with respect to $\mathbf{x}$, we can transform $J_3(\mathbf{x})$ into matrix form:

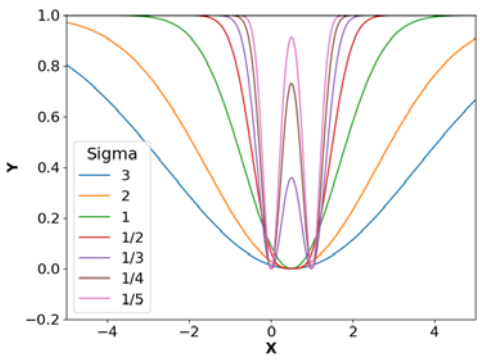


Fig. 2. The normalized regularization term $g_\sigma(x)$ for several σ.

$$J_3(\mathbf{x}) = \mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + r \tag{10}$$

where $\mathbf{P}$ is a symmetric $n \times n$ matrix with each element expressed in d_{ij} and λ_1, $\mathbf{q}$ is an $n \times 1$ column matrix with each element expressed in λ_1, and r is a scalar / constant expressed in λ_1.

C. *Formally Treating Binary-valued Variables as Continuous Variables*

Heading toward possible adoption of the gradient-descent (GD) based methods, such as the Newton-Conjugate-Gradient method, for searching the minimum, we formally treat the binary-valued variables $\mathbf{x}$ as continuous variables such that $\mathbf{x} \in \mathbb{R}^n$. Since the cost function $J_3(\mathbf{x})$ is quadratic in form, $J_3(\mathbf{x})$ is infinitely differentiable, which guarantees the existence of the gradient matrix and Hessian matrix, that are usually required by GD based methods.

In the continuous domain, an $\mathbf{x} \in \mathbb{R}^n$ does not necessarily represent a feasible route. Only a binary-valued $\mathbf{x} \in \{0,1\}^n$ can represent a feasible route. As $\{0,1\}^n$ is a very sparse subset of $\mathbb{R}^n$, we are interested in heuristics that tend to allow many candidates $\mathbf{x}$ in the continuous domain $\mathbb{R}^n$, so that at later stages we can narrow down to those candidates close enough to the binary-valued domain $\{0,1\}^n$. One such "many-candidate" heuristic for picking λ_1 is described in the following subsection.

D. *Picking λ_1 by Heuristics Favoring Near-Singular Systems*

The gradient of the cost function $J_3(\mathbf{x})$ is

$$\nabla J_3(\mathbf{x}) = 2\mathbf{P}\mathbf{x} + \mathbf{q}. \tag{11}$$

A necessary condition for the minimum of $J_3(\mathbf{x})$ to happen at $\mathbf{x} = \hat{\mathbf{x}}$ is "zero gradient"

$$\mathbf{P}\hat{\mathbf{x}} = -\mathbf{q}/2. \tag{12}$$

Since the above is a linear equation, if $\mathbf{P}$ is near-singular / ill-conditioned, many $\hat{\mathbf{x}}$ exist such that $J_3(\hat{\mathbf{x}})$ falls within an ϵ-neighborhood of a local minimum of $J_3(\mathbf{x})$. Because $\mathbf{P}$ is expressed in terms of λ_1, the following "many-candidate" heuristic (as mentioned in the last subsection) is devised.

Heuristic: Picking λ_1 such that $\mathbf{P}$ is near-singular / ill-conditioned, or equivalently, the condition number of $\mathbf{P}$ is large.

E. *Cost Functions for Square of Variables*

The cost function $J_3(\mathbf{x})$ includes the distance form of $d_{ij}x_{ti}x_{(t+1)j}$, which can turn negative when the variables are in the continuous domain $\mathbf{x} \in \mathbb{R}^n$. We simply substitute each variable by its square to arrive at $d_{ij}x_{ti}^2x_{(t+1)j}^2$ to guarantee positive values. Let

$$\mathbf{x^2} \equiv [x_{11}^2 \;\; x_{12}^2 \;\; \ldots \;\; x_{1c}^2 \;\; \ldots \;\; x_{c1}^2 \;\; x_{c2}^2 \;\; \ldots \;\; x_{cc}^2]^T. \tag{13}$$

The new cost function $J_4(\mathbf{x})$ is chosen as

$$J_4(\mathbf{x}) \equiv J_3(\mathbf{x^2}) = \mathbf{x^2}^T\mathbf{P}\mathbf{x^2} + \mathbf{q}^T\mathbf{x^2} + r \tag{14}$$

F. *Using Regularization to Incorporate Constraints Favoring Binary Values*

The general framework in reference [1] has inspired the following regularization term $f_\sigma(x)$ for the constraints favoring binary values $\{0, 1\}$.

$$f_\sigma(x) \equiv 1 - \exp\left(\frac{-x^2}{2\sigma^2}\right) - \exp\left[\frac{-(x-1)^2}{2\sigma^2}\right], \tag{15}$$

$$\lim_{\sigma\to 0} f_\sigma(x) = \begin{cases} 0, & \text{if } x = 0 \text{ or } 1 \\ 1, & \text{otherwise} \end{cases}. \tag{16}$$

Note that $f_\sigma(x)$ is symmetric about $x = 0.5$, and thus $f(0) = f(1)$. For larger σ, $f_\sigma(x)$ has one minimum at $x = 0.5$. For smaller σ, $f_\sigma(x)$ has two minima at $x = 0$ and $x = 1$.

However, the range of $f_\sigma(x)$ varies with different values of σ. When σ increases to the point where there is only one minimum, the minimum value of the function can be far smaller than 0, as shown in Fig. 2. To rescale the range of $f_\sigma(x)$ for all possible σ values to lie between 0 and 1, we normalize the regularization term function $f_\sigma(x)$. The following is the formula for normalizing the function:

$$f_{norm} = (f - f_{min})/(f_{max} - f_{min}) \tag{17}$$

where f_{max} remains unchanged with respect to σ, fixed at 1, and f_{min} varies. Therefore, to define the scaling scalar s_σ for the normalized function $f_\sigma^*(x)$, we generate the normalized function. The function plot is depicted in Fig. 2, where the curve represents the results for several different σ values.

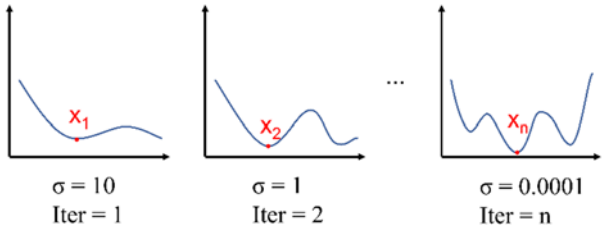


Fig. 3. Visualization of Annealing-Inspired method.

$$s_\sigma \equiv 1 - \min_{x\in\mathbb{R}} f_\sigma(x)$$

$$= \begin{cases} 1 - f_\sigma(0), & f_\sigma(0.5) > f_\sigma(0) \\ 1 - f_\sigma(0.5), & \text{otherwise} \end{cases}, \tag{18}$$

$$f_\sigma^*(x) \equiv (f_\sigma(x) + s_\sigma - 1)/s_\sigma. \tag{19}$$

G. *Annealing-Inspired: Slowly Decreasing σ in Attempts to Escape Local Minima*

The cost function $J_\sigma(\mathbf{x})$ may have many local minima. In Fig. 3, the regularization term $g_\sigma(x)$ is more like a line for larger σ. It is flatter, has lower convexity, less favors binary $\{0, 1\}$, and renders fewer local minima for $J_\sigma(\mathbf{x})$. For smaller σ, $g_\sigma(x)$ has larger convexity, much favors binary $\{0,1\}$, and renders more local minima for $J_\sigma(\mathbf{x})$. In attempts to escape local minima, σ is slowly decreased in each optimization iteration, in which the cost function $J_\sigma(\mathbf{x})$ is optimized using a GD based method.

III. Experimental and Results

This section is divided into four sections. The first subsection illustrates the differences between the algorithm proposed in this paper and existing algorithms. The second section tests the algorithm model with parameter adjustments for the Traveling Salesman Problem (TSP) across various dimensions. The third section presents the results and comparisons of finding the shortest path in a randomly distributed city model. The final section uses the renowned TSPLIB dataset as the benchmark to showcase the algorithm's results and advantages over other algorithms [5]. The algorithm is implemented in Python 3.11 without GPU acceleration. The hardware are an Intel Core i5-7200U with 12GB RAM.

A. *Evaluation of Time Complexity*

The Traveling Salesman Problem is an NP-hard problem, meaning it cannot be solved in polynomial time. Its time complexity increases at least exponentially. For instance, for the Traveling Salesman Problem with n cities, there are $(n-1)!/2$ possible routes. Finding the shortest route among them requires a time complexity of $O(n!)$, known as factorial time. Algorithms of this nature are often referred to as exhaustive or brute-force methods. As the input n increases, the execution time increases dramatically. Just for 11 cities, there are 1,814,400 routes. While cases up to 12 cities can be handled conventionally, 17 cities require the most powerful computational resources available today. However, problems involving 21 cities are nearly impossible to solve.

To mitigate the excessively large computational complexity, dynamic programming can be employed. It breaks down the entire Traveling Salesman Problem into several subproblems, decomposing the entire route into multiple subpaths. Trading space complexity for reduced time complexity is highly effective, reducing the time complexity to $O(n^2 2^n)$. We can observe that beyond 10 cities, dynamic programming's complexity becomes lower than brute-force methods. By the time 20 cities are reached, there is an eight-order-of-magnitude difference between the two approaches. As shown in Fig. 4, this difference accelerates as the dimensionality increases. However, dynamic programming still exhibits exponential complexity and cannot achieve the polynomial time complexity required by modern integrated circuit computers.

The algorithm we developed is based on the matrix P for optimizing computations, with a dimensionality of n^4. The theoretical time complexity is $O(n^4)$. However, during the search for the minimum value, we utilize the Newton conjugate gradient method. Rather than delving into the details of algorithm analysis, we chose a more efficient approach by the program execution time calculations to assess our algorithm's time complexity. This way systematically analyzed the potential factors affecting time complexity and observed the impact as the number of cities increases.

In Fig. 4, we observed that our Annealing-inspired Gradient-Descent Based Suboptimal Solver (ACUO) algorithm's computation time grows with the dimensionality, roughly matching $O(n^4)$. We also analyzed the computation times of two exact solving algorithms, exhaustive search (the brute-force method), and dynamic programming, using results from the package of Python-tsp. They exhibit a trend like the estimated time complexity. Hence, these results can serve as the evaluation standard for our algorithm's time complexity. Despite not being outstanding in terms of computational speed, the time complexity remains within polynomial time complexity. Thus, the problem with high city number can find a solution in finite time for our algorithm.

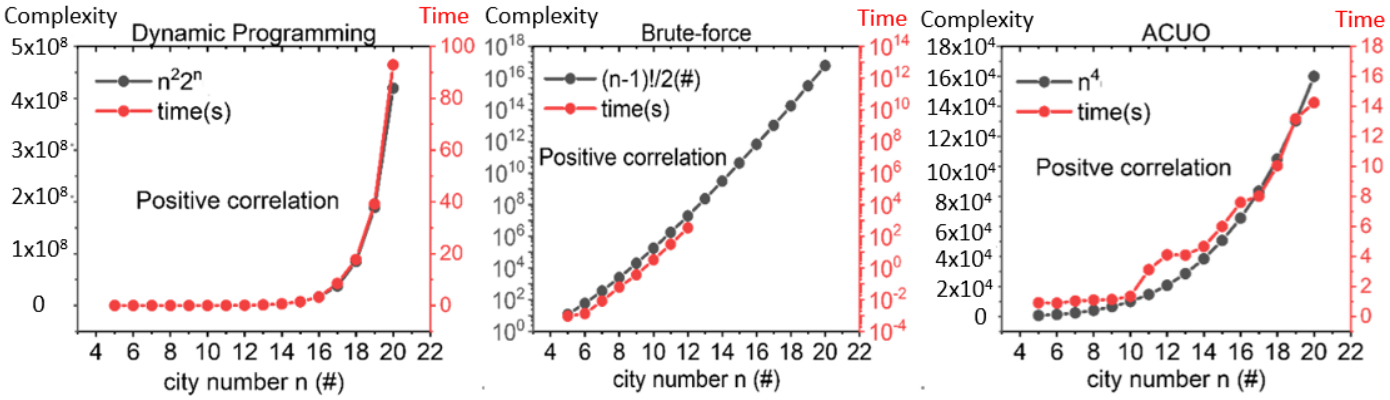


Fig. 4. Comparison of the computation time (red lines) and the complexity (black line) of the (Left) in Dynamic Programming, (Middle) Brute-force, and (Right) proposed ACUO methods. One can see that the proposed algorithm can much reduce the computation time and complexity.

TABLE I. Comparison of ACUO Results with Standard Solutions for Randomly Distributed Cities.

N	ACCU(%)	Brute Force	ACUO	DIFF	DIFF%
5	94	0.8169	0.8186	0.00171	0.2093
6	89	0.7451	0.7465	0.00144	0.1935
7	87	0.6816	0.6833	0.00168	0.2458
8	83	0.6359	0.6381	0.00226	0.3549
9	78	0.5879	0.5910	0.00307	0.5225
10	77	0.5531	0.5562	0.00309	0.5579
11	70	0.5251	0.5285	0.00344	0.6558
12	55	0.5027	0.5100	0.00732	1.4553
13	42	0.4806	0.4902	0.00957	1.9918
14	36	0.4561	0.4676	0.01154	2.5306
15	37	0.4369	0.4511	0.01414	3.2366

B. Results of ACUO for Random Cities Distribution

We extensively tested the optimization algorithm using randomly distributed city data, comparing the optimized paths obtained by our algorithm with the exact solutions obtained by dynamic programming. We examined the accuracy and error of the ACUO in low dimensions. We tested city numbers ranging from n = 5 to n = 15 in various scenarios, generating 100 sets of test data for each scenario using random distributions. The parameters were selected based on the optimal combinations obtained in the previous section. The results are presented in Table I and Fig. 5.

TABLE II. Optimization path length comparison using the TSPLIB data set.

Dataset	*n*	ACUO	SA	MIN	ERR (%)
gr24	24	1340	1350	1272	5.35
bays29	29	2070	2036	2020	2.48
berlin58	58	8006	8135	7542	6.15
eil51	51	450	452	426	5.63

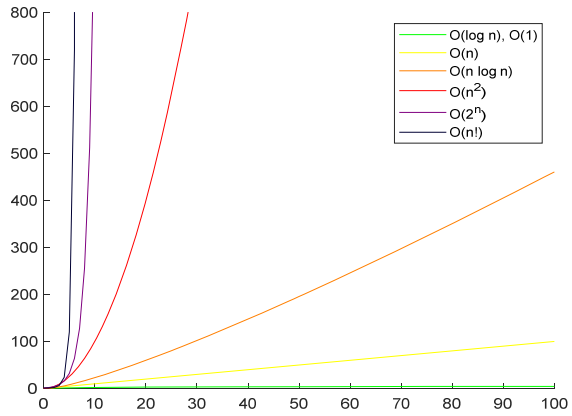


Fig. 5. The time complexity function curve graph.

Considering the error in path length, the difference between the optimized path length obtained by our algorithm and the shortest path is minimal. Regardless of the dimension, the error remains within 6.15%. From the results in Table II and Fig. 6, one can see that there are no obvious undesirable paths with unnecessary overlap. Therefore, the paths selected by our algorithm are considered to be satisfactory choices.

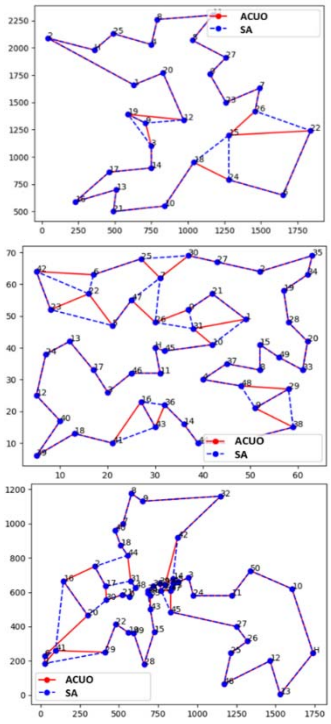


Fig. 6. TSPLIB data set bay29(top), berlin58(middle) and eil51(bottom) optimization path comparison chart

C. *Results and Comparison of TSPLIB Test Sets*

In order to verify whether good results can be achieved even with a higher number of cities and not limited to randomly distributed cities, we obtained five test datasets from TSPLIB for optimization. We compared the results with the simulated annealing (SA) algorithm in Python, as well as with the known shortest path lengths provided by the database, calculating the errors. The results, as shown in TABLE II, indicate that the path lengths obtained by ACUO are close to those of the simulated annealing method, with some parts performing better and others slightly worse. For datasets with city coordinates provided, the relative positions of cities can be plotted to visualize the differences between the paths generated by ACUO and SA in Fig. 6. Compared to the ground truth, the errors in path length compared to the provided shortest path lengths are also within 6.15%.

IV. Conclusions

We introduce the Annealing-inspired Continuous Unconstrained Optimization (ACUO) algorithm, designed to tackle combinatorial optimization problems by transforming them into polynomial function optimization in the continuous domain. This structural innovation sets ACUO apart from traditional discrete iterative methods, enabling the integration of both discrete and continuous constraints, thereby addressing system stability and facilitating the escape from local minima. Using the Traveling Salesman Problem (TSP) as a case study, we demonstrate that ACUO achieves polynomial growth in time complexity, effectively mitigating the curse of dimensionality. While the algorithm maintains high accuracy for low-dimensional problems, the accuracy diminishes as the number of cities increases. Nonetheless, the maximum error is within 6.15%, and path visualization confirms the validity of the solutions. ACUO exhibits broad applicability, a unique model architecture, and the capability to solve combinatorial problems within a specific margin of error.

References

[1] Larranaga, P., et al., Genetic algorithms for the travelling salesman problem: A review of representations and operators. Artificial intelligence review, 1999. 13: p. 129-170.

[2] Finnila, A.B., et al., Quantum annealing: A new method for minimizing multidimensional functions. Chemical Physics Letters, 1994. 219(5): p. 343-348.

[3] Mohimani, H., M. Babaie-Zadeh, and C. Jutten, A fast approach for overcomplete sparse decomposition based on smoothed L0 norm. IEEE Transactions on Signal Processing, 2008. 57(1): p. 289-301.

[4] Diaby, M., The traveling salesman problem: a linear programming formulation. arXiv preprint cs/0609005, 2006.

[5] Reinelt, G., TSPLIB—A traveling salesman problem library. ORSA journal on computing, 1991. 3(4): p. 376-384.

[6] Margalit, D., J. Rabinoff, and L. Rolen, Interactive linear algebra. Georgia Institute of Technology, 2017.

[7] Hoerl, A.E. and R.W. Kennard, Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, 1970. 12(1): p. 69-82.

[8] Louizos, C., M. Welling, and D.P. Kingma, Learning Sparse Neural Networks through L0 Regularization. ArXiv, 2017. abs/1712.01312.

[9] Papadimitriou, C.H. and K. Steiglitz, Combinatorial optimization: algorithms and complexity. 1998: Courier Corporation.

[10] Jünger, M., G. Reinelt, and G. Rinaldi, *The traveling salesman problem.* Handbooks in operations research and management science, 1995. **7**: p. 225-330.

[11] Applegate, D.L., et al., The traveling salesman problem. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2006: p. 1-5.

Quefrency Approach to Audio Deepfake Detection

Kanishq Singhal, Aditya Goyal and Priyanka Gupta
The LNM Institute of Information Technology, Jaipur, India
E-mail: 23ucs605@lnmiit.ac.in, 23ucs512@lnmiit.ac.in, priyanka.gupta@lnmiit.ac.in

***Abstract*—This work aims at audio deepfake detection (ADD) in three different scenarios-only deepfake, re-recorded-deepfake, and replayed-deepfake. To that effect, the existing Fake or Real (FoR) dataset is enhanced by a reverberation topology leading to a novel version of the dataset (called as rev-FoR-rerec) to simulate replayed-deepfakes in varying acoustic environments. Furthermore, this work investigates the significance of quefrency-based representation for ADD. Our findings indicate that as compared to mel-spectrogram representations, the proposed quefrency-based system achieves improved EERs on *both* testing and validation sets of all three attack scenarios, except for the case of replayed-deepfake where the proposed system performs nearly equal to the mel-spectrogram with an absolute difference in testing EER as 0 and an absolute difference in validation EER as 0.09. Furthermore, our sub-band analysis shows that the bands 0-1000 Hz and 7000-8000 Hz are the most discriminating subbands for ADD. Experiments are also performed to investigate the effect of room size and reverberation damping factor on the detection of replayed-deepfakes.**

I. Introduction

Deepfakes pose a significant threat due to their increasing realism and potential for manipulation in various media formats like text, audio, images, and videos [1], [2]. This has detrimental implications for privacy, security, and societal integrity [3]. Audio deepfakes, created through AI-generated or manipulated audio, aim to mimic a target speaker's voice convincingly. Advancements in text-to-speech (TTS) and voice conversion (VC) techniques have facilitated the production of high-quality audio deepfakes [4], [5].

Detecting audio deepfake is a challenging task that involves various techniques from signal processing, machine learning, and deep learning [6], [7]. Previous research has shown promising results [8]–[10]. Many existing works focus on summarizing past spoofing attacks, or protecting automatic speaker verification (ASV) systems [11], [12]. However, datasets released in the ASVSpoof challenges lack usability for research because of undisclosed synthesis algorithms due to the nature of the ASVSpoof challenges. To that effect, we have utilized the Fake-Or-Real dataset for ADD in this work.

So far, most of the neural network architectures for ADD utilize spectrogram analysis [13]–[15], however, quefrency-based representations such as cepstrograms are less explored. To that effect, this work investigates the significance of quefrency approach to ADD. In addition, since TTS and VC technologies have eased the generation of audio deepfakes, deepfake attacks, like the replay attack have become easier to mount [16]. Hence, detecting these two attacks is crucial to ensure security and integrity. To that effect, a scenario of replayed-deepfake detection is also introduced in this work. In particular this paper has the following contributions:

- Quefrency-based features are proposed for ADD, and hence cepstral vs. spectral analysis is done to investigate which domain is more suited for ADD.
- Apart from the scenario where a fake/spoofed utterance is synthesized using deepfake only, we have considered two more scenarios- *rerec-deepfake* and *replayed-deepfake*. Rerec-deepfake considers the situation when the deepfaked utterance is transmitted through a communication channel (like a phone call or voice message). Replayed-deepfake considers the scenario where the deefaked signal is recorded and played back like a replay attack. Given that replay attacks are the easiest to execute from an attacker's perspective [17], it becomes important to evaluate countermeasure systems for replayed-deepfake scenario as well, thereby attempting towards generalization of countermeasure systems.
- To simulate the *replayed-deepfake* scenario, a new dataset (namely rev-FoR-rerec) is created where the deepfaked audio is recorded and replayed in rooms of varying size, with reverberations of varying intensities.
- To that effect, a reverberation topology for simulating the replayed-deepfake scenario is also proposed in this work.
- A lesser-known dataset *Fake-or-Real (FoR)* dataset is used in this work. Unlike the ASVSpoof 2021 DF dataset, this dataset is balanced in terms of gender and class, and normalized in terms of volume and number of channels. It should also be noted that the testing set of the FoR dataset consists of utterances generated using Deepvoice3, Baidu TTS, Amazon AWS Polly and Cloud, WaveNet, which are Google text-to-speech algorithm renowned for producing speech that closely resembles human voices hence testing our model on real-world conditions.
- Analysis on the effect of reverberation parameters affecting room-size and reverberation intensity has also been done the on *replayed-deepfake* scenario.
- Experiments are also performed on the ASVSpoof 2019 Logical Access (LA) dataset, to investigate the performance of the proposed ADD system on LA scenarios.

II. Proposed Work

This work presents three different attacking scenarios. In Fig. 1 (A), deepfake speech is generated using spoofing algorithms such as TTS (Text-to-Speech) or VC (Voice Conversion). In realistic scenarios, as shown in Fig. 1 (B), an attacker can re-record a deepfake utterance using a device like a smartphone and send it to a victim's smartphone through

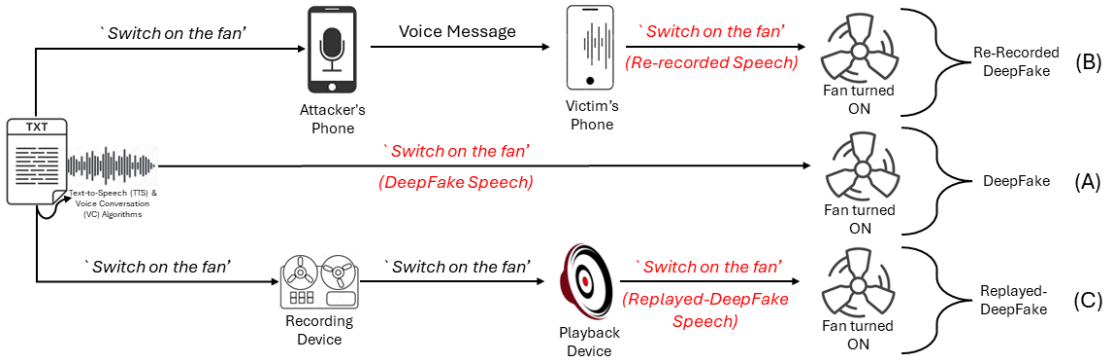


Fig. 1. Different types of attack scenarios: (A) Only Deepfake, (B) Re-Recorded Deepfake, and (C) Replayed-Deepfake

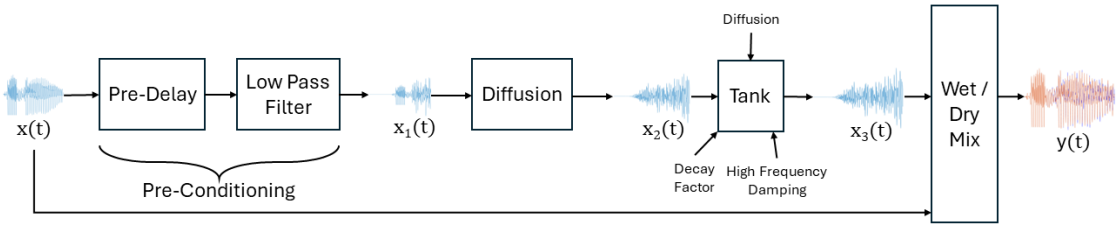


Fig. 2. Reverberated Topology for Replayed-Deepfake Scenario

a voice message or communication channel. Additionally, as illustrated in Fig. 1 (C), we introduced another potential attacking strategy that involves recording and playing back deepfake audio, which we refer to as *"replayed-deepfake"* in this work.

A. Proposed replayed-deepfake with reverberation topology

To further expand the scope of the dataset for anti-spoofing research, a novel "Reverberated FoR Re-Recorded (rev-for-rerec)" dataset is designed in this work. To bridge the gap in the existing ADD dataset we incorporated synthetic reverberations of varying specifications (such as room size, high frequency damping and pre-delay), thereby offering a more comprehensive evaluation platform for anti-spoofing models. This additional layer of intricacy to the dataset, which simulates real-world scenarios, reflects the diverse environments and conditions in which spoofing attacks might occur better. To that effect, we have introduced reverberation in the existing Fake or Real dataset using the reverberation.

The reverberation topology [18] as shown in Fig. 2 begins with *pre-conditioning*. Preconditioning comprises of two stages: 1) Pre-delay, and 2) Low-pass filtering. *Pre-delay* is the time between hearing direct sound and the first early reflection and is expressed as:

$$x_p(t) = x(t - k) \tag{1}$$

where k is the pre-delay constant. A smaller value of k is preferred for short-duration utterances because it ensures that the reverberation effect occurs sooner. Now, the pre-delayed signal is passed through the *low-pass filter* expressed as:

$$LP(z) = \frac{1 - \alpha}{1 - \alpha z^{-1}} \tag{2}$$

where α is estimated as:

$$\alpha_1 = \exp\left(-2\pi \times \frac{f_c}{f_s}\right) \tag{3}$$

where f_c is the *high cut frequency* and f_s is the sampling frequency. After the preconditioning, *diffusion* is performed which controls the density of the reverb tail and is specified as a real positive scalar in the range [0, 1]. To set the diffusion, the preconditioned signal, as shown in step 3 of Algorithm 1, is passed through a cascade of 4 *all-pass filters*, where each filter is expressed as:

$$AP(z) = \frac{\beta + z^{-k}}{1 + \beta z^{-1}} \tag{4}$$

where β is the diffusion constant and k is the delay. Increasing the diffusion pushes the reflections closer together, thickening the sound, whereas reducing it creates more discrete echoes. The filtered output is then fed into a *tank* to adjust the decay factor and high-frequency damping of the reverberation tail. The *high-frequency damping* adjusts the fading of high-frequency reflections, whereas the *decay factor* of the reverb tail determines how quickly the reflections lose energy, providing a longer reverb tail without overwhelming the original signal [18]. For this effect, the resultant signal is passed through a series of filters as demonstrated in step 4 of Algorithm 1. The modulated all-pass filter is:

$$ModulatedAP(z) = \frac{-\beta + z^{-k}}{1 - \beta z^{-1}} \tag{5}$$

where β is the diffusion constant from Eq. 4. Here, in the tank, the low-pass filter constant α (from Eq. 2) is used as the high-frequency damping constant. Finally, "*wet-dry mix*" is performed on the signal. The wet-dry mix ratio W (where, $W \in [0, 1]$) controls the ratio of the reverberated signal to the original, as illustrated in step 5 of Algorithm 1.

B. Proposed Quefrency-based representation for ADD

Most of the work on ADD has been done by analyzing the signal using spectrograms in linear or mel-scale [13]–[15]. However, quefrency-based representations such as cepstro-

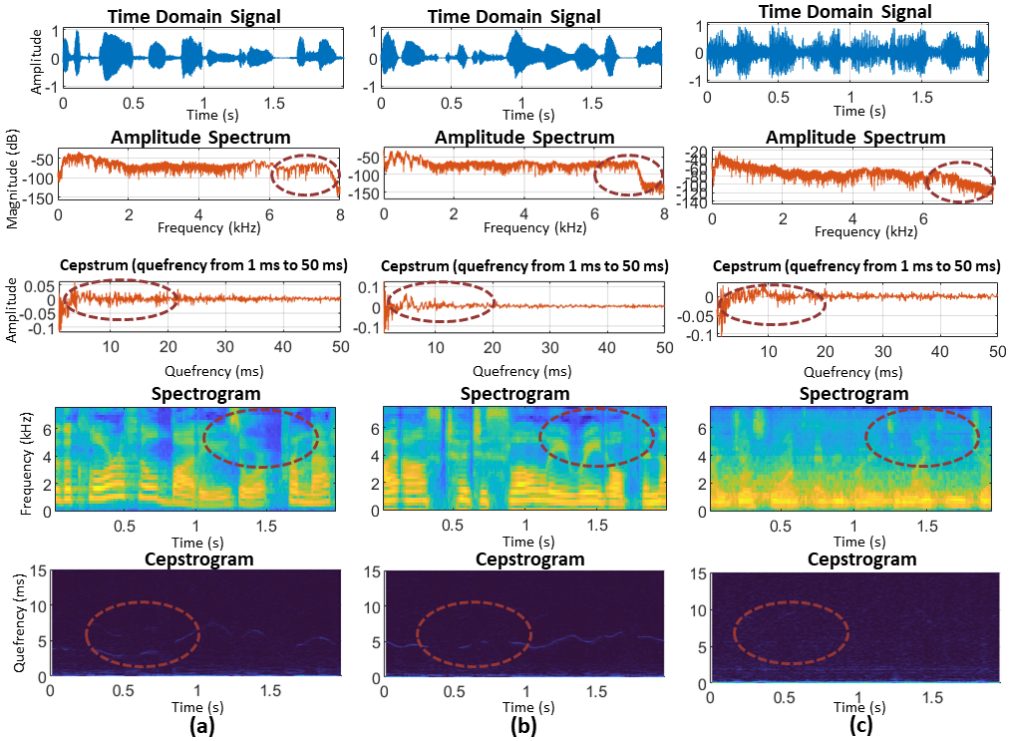


Fig. 3. Analysis of speech signals using various representations. (a) Genuine utterance, (b) Only Deepfake, (c) Replayed-Deepfake

Algorithm 1: Proposed Reverberation Algorithm for replayed-deepfake attack

Input: Input: Speech signal $x(t)$

Output: Output: Reverberated Speech $y(t)$

$x_p(t) \leftarrow Pre-Delay\ [x(t)]$

$x_1(t) \leftarrow LowPass\ Filter\ [x_p(t)]$ } *Pre-Conditioning*

$x_2(t) \leftarrow AP_i\ [x_1(t)]\ where\ AP_i$ is the ith All-Pass Filter and $i = 1, 2, 3, 4$

$x_3(t) \leftarrow AllPassFilter \leftarrow LowPassFilter \leftarrow Modulated\ AllPassFilter\ [x_2(t)]$

$y(t) = (1-k)x(t) + Wx_3(t)$

grams are less explored. These representations have been used in pitch detection [19], echo removal [20], and noise reduction [21], [22]. Since deepfake algorithms are known to introduce artifacts, and the replayed deepfakes have reverberation, this work investigates the importance of quefrency-based representations on ADD.

Quefrency-domain representations measure cepstrum of a signal, which comes from the representation of a homomorphic system for convolution [23]. The word *cepstrum* is called "spectrum of the spectrum", and is estimated as [23] :

$$C = \left| \mathcal{F}^{-1} \left\{ \log \left(|\mathcal{F}\{x(t)\}|^2 \right) \right\} \right|^2 \tag{6}$$

where $x(t)$ is a signal, and $\mathcal{F}$ is its Fourier transformation.

A visual representation of the cepstrum over time is known as *cepstrogram*. It allows one to observe changes in the cepstral features (*rahmonics*) of a speech over time, similar to how a spectrogram shows changes in the spectral content of a speech.

C. *Frequency* vs *Quefrency Analysis*

To analyze the cases of genuine, only deepfake and replayed-deepfake utterances, Fig. 3 shows the differences between frequency and quefrency representations. In particular, Fig. 3 (a) shows the genuine utterance, Fig. 3 (b) shows the only deepfake utterance followed by Fig. 3 (c) which shows replayed-deepfake utterance. On analyzing the amplitude spectrums of the utterances, we observe that in Fig. 3 (a), the trend of the amplitude spectrum is *nearly* constant throughout the frequency range of 6-8 kHz as shown using a dotted circle. However, in Fig. 3 (b), we see that the amplitude spectrum is nearly constant *only* till 7 kHz, and then a sudden drop is observed, as marked by the circle. Meanwhile in Fig. 3 (c), the amplitude spectrum decreases gradually from 6 kHz onwards which is marked using a circle. Furthermore, we also analyze cepstrums of the utterances with quefrency from the range of 1 ms to 50 ms. The quefrency in the range of 1 ms to 50 ms is chosen because most of the variations have been observed in this range. The cepstrum has the highest distortion in Fig. 3 (c) followed by Fig. 3 (a) and Fig. 3 (b) respectively as marked by the circle. This indicates that the cepstrum for *only deepfake* (Fig. 3 (b)) fails to capture the environmental artifacts unlike the cepstrum in Fig. 3 (a) and Fig. 3 (c). This further indicates that the deepfake algorithms are not able to generate *naturalness* like genuine speech.

On analyzing the utterances using spectrogram, we observed that in genuine utterance (Fig. 3 (a)), the energy fading appears natural without any abrupt boundaries as opposed to the case of only deepfake utterance in Fig. 3 (b), where no continuous band of energy is observed as indicated by the circle. Furthermore, we observe that in replayed-deepfake

(Fig. 3 (c)) utterance, there is continuous energy in the region marked by the circle in the figure. This indicates that the genuine speech does not have as much energy as compared to the speech generated due to reverberation. However, when we analyze the utterances using quefrency-based representation or cepstrogram we observe that it is easier to differentiate between the genuine and replayed-deepfake speech than the genuine and only deepfake utterance because more dominant *rahmonic* peaks are observed in genuine speech as compared to replayed-deepfake speech which is marked with the help of the dotted circle.

III. Experimental Setup

A. Datasets

1) Fake or Real (FoR): The Fake or Real (FoR) has 4 versions namely, original, normalized, 2-sec version, and a re-recordered version [24]. In this work, experiments have been performed on the FoR 2-sec and For re-rec datasets. Both of these versions are balanced in terms of the number of utterances in each class, and gender variability of speakers. The FoR 2-sec and FoR re-rec versions correspond to deepfake only, and recorded-deefake attack scenarios, respectively, as discussed in Section II. The partitioning details of both of these datasets are identical [24].

2) Proposed Rev-rerec dataset: The replayed-deepfake attack scenario is emulated by incorporating reverberation into the for-rerec dataset, as discussed in Section II. In particular, for our experiments, we have fixed pre-delay factor at 0.25, diffusion as 0.5, and wet/dry mix at 0.3. The reverberation parameters, like decay factor and high-frequency damping, have been varied to investigate their effect on the performance.

B. Classifier and Performance Metrics

The CNN model used in this work consists of three convolutional blocks: Convolution 1, Convolution 2, and Convolution 3 where each convolutional block includes a 2-D convolutional layer with a kernel size of 3x3, stride of 1 and padding of 1, Rectified Linear Activation (ReLU) layer, and a max-pooling layer with a kernel size of 2x2. The output feature maps from the convolutional blocks are processed by two fully connected layers with 128 hidden units. This layer flattens the feature maps and then passes them through a ReLU activation function. The model is trained using the Adam optimizer with a learning rate of 0.001, and the loss is calculated using cross-entropy loss. The training process runs for 20 epochs, with a batch size of 32. The performance of the model is evaluated using Accuracy and Equal Error Rate (EER).

C. Baseline

In this work, we have considered the baseline from [10]. We investigate the significance of quefrency-based representation and mel-spectrogram. Both the representations are extracted using 30 ms of window length, with a window shift of 15 ms.

IV. Experimental Results & Analysis

A. Comparing Proposed System with the Existing Works

Experiments were performed using the proposed quefrency-based features and compared with the baseline [10] as shown in Table I. It can be observed that the proposed quefrency-based feature outperforms the existing baseline system on all the attack scenarios and datasets. In particular, we observed an absolute increase of 16.64%, 60.05%, 39.70%, and 8.63% in the testing and validation sets of the for-2sec and for-rerec datasets, respectively.

TABLE I
Existing System vs Proposed System

Dataset Used	Existing System [10]		Proposed System	
	Testing	Validation	Testing	Validation
FoR-2sec	71.14	38.29	**87.78**	**98.34**
FoR-rerec	45.96	86.47	**85.66**	**95.10**
rev-FoR-rerec	-	-	**100**	**99.87**

B. Feature Wise Comparison

To compare the performance of quefrency-based representation with a mel-spectrogram, experiments were performed on all the three attack scenarios, as shown in Table II. It can be seen that the performance of quefrency-based representation is better than the mel-spectrogram in most cases except in the rev-FoR-rerec dataset. However, it should be noted that only the best performances are shown in this table, and the detailed comparison is shown in Section IV-D. To further compare the features, experiments were performed on the widely known ASVspoof 2019 LA dataset. The quefrency-based feature achieved an accuracy of 89.63%, while the mel-spectrogram attained an accuracy of 89.60%, indicating no significant difference in performance.

TABLE II
Feature Wise Comparison: (A) Cepstrogram, (B) Mel Spectrogram

Attack Scenario		Only Deepfake		Rerecorded-Deepfake		Replayed-Deepfake	
Dataset Used		FoR-2sec		FoR-rerec		rev-FoR-rerec	
Feature Used		A	B	A	B	A	B
Accuracy (%)	Testing	**87.78**	87.13	**85.66**	81.5	**100**	100
	Validation	98.34	**98.44**	95.10	**96.17**	99.82	**99.87**
EER (%)	Testing	12.5	**8.64**	**13.24**	16.19	**0**	0
	Validation	1.70	**1.56**	**4.11**	4.64	0.26	**0.17**

C. Analysis on Only-Deepfake Scenario: Effect of Sub-Band Frequency Ranges

To observe the effect of the frequency range for deepfake detection, the full spectrum of the utterances in the FoR 2-sec dataset was divided into 8 sub-bands of bandwidth of 1000 Hz each (given the maximum frequency is 8000 Hz).

TABLE III
Effect of Sub-Band Frequency Ranges on ADD

Subband (in Hz)	Accuracy (%)		EER (%)	
	Testing	Validation	Testing	Validation
0 - 1000	86.40	**98.20**	13.05	**1.84**
1000 - 2000	64.98	83.69	33.09	14.79
2000 - 3000	64.15	87.44	32.90	14.79
3000 - 4000	50.92	90.34	46.88	9.77
4000 - 5000	61.76	88.50	34.74	10.62
5000 - 6000	49.45	88.50	51.47	11.46
6000 - 7000	66.27	89.67	34.01	9.98
7000 - 8000	**95.31**	97.38	**2.21**	2.76

TABLE IV
EFFECT OF REVERBERATION PARAMETERS ON REPLAYED-DEEPFAKE DETECTION

Decay Factor	High Frequency Damping	Mel-Spectrogram				Proposed Quefrency-based Representation			
		Accuracy (%)		EER (%)		Accuracy (%)		EER (%)	
		Testing	Validation	Testing	Validation	Testing	Validation	Testing	Validation
0.9	0	92.28	96.75	6.37	3.32	**97.67**	**98.08**	**1.72**	**1.92**
	0.25	90.56	96.43	9.56	3.50	**98.04**	**97.33**	**1.47**	**2.62**
	0.5	91.79	96.79	7.84	2.80	**97.79**	**98.08**	**1.47**	**2.01**
	0.75	98.28	**98.80**	1.72	**1.31**	**99.26**	97.64	**0.74**	2.62
	1	92.03	96.26	7.84	3.67	**97.43**	**97.46**	**2.21**	**2.54**
0.5	0	90.44	96.66	9.31	3.32	**99.63**	**98.26**	**0**	**1.57**
	0.25	97.67	98.31	2.70	1.92	**99.63**	**98.80**	**0**	**1.22**
	0.5	95.83	**97.95**	2.21	**1.92**	**98.90**	96.83	**0.25**	2.62
	0.75	96.81	97.77	3.43	2.01	**99.75**	**98.35**	**0.25**	**1.75**
	1	86.89	**96.52**	11.03	3.41	**98.77**	96.30	**1.47**	**3.32**
0.1	0	**99.88**	**99.91**	0	**0.26**	99.39	99.38	**0**	0.61
	0.25	99.75	**99.82**	0	**0**	**100**	99.64	**0**	0.26
	0.5	100	**99.87**	0	**0.17**	**100**	**99.82**	**0**	**0.26**
	0.75	99.63	**99.82**	0	**0.09**	**100**	99.69	**0**	0.26
	1	89.46	95.94	10.78	4.02	**93.14**	**96.88**	**0.74**	**2.50**

To that effect, experiments were performed by extracting sub-band-wise mel-spectrograms which were fed to the CNN for classification. The obtained performances are shown in Table III. It can be observed that the best performances for both testing and validation sets are observed in the first and last frequency bands (i.e. 0-1000 Hz and 7000-8000 Hz). This indicates that the sub-bands in the extremes of the spectrum play a significant role in ADD. It should also be noted that in the case of Voice Liveness Detection (VLD), the discriminating features w.r.t. pop noise are present in 0-40 Hz, i.e., on the lower extreme of the frequency range [25], [26]. Therefore, our subband analysis for ADD coincides with subband analysis for the VLD problem for low-frequency subbands.

Furthermore, we also observe in Table III that the validation accuracy is greater than the testing accuracy in all the subbands. This discrepancy arises because the testing dataset consists of unseen utterances. Moreover, the testing set was generated using WaveNet- a Google TTS algorithm, which is well-known to generate human-like speech.

D. Analysis on Replayed-Deepfake Scenario: Effect of Reverberation Parameters

To investigate the impact of room size and high-frequency damping on replayed-deepfake detection, experiments were performed on the proposed rev-FoR-rerec dataset. To that effect, Table IV shows the performances using mel-spectrogram and proposed quefrency-based feature (cepstrogram). For each value of decay factor $\left(DecayFactor \propto \frac{1}{RoomSize}\right)$, the high-frequency damping is varied to adjust the high-frequency reflections. The general trend observed in Table IV shows that on the testing set, the quefrency-based feature *consistently* performs better than the mel-spectrogram in *both* accuracy and EER across varying decay factors and damping values. In particular, as room size increases, both the features show improved performances, with the quefrency-based feature frequently achieving near-perfect to perfect accuracy and very low EER of 99% to 100% and 0 to 1% respectively. This indicates that large rooms contribute to better replayed-deepfake detection yielding robust results. The possible reason for this is that in large rooms, the delay between each reflection is more, as the second wave has to travel a larger distance before getting reflected, which makes it easier to distinguish between the replayed-deepfake utterance and the genuine utterance.

In the case of the proposed quefrency-based feature, for all the values of decay factor, and at moderate to high levels of the high-frequency damping factor (0.25 to 0.75), we can observe a high accuracy range $\sim 98 - 100\%$ and low EER values of $\sim 0 - 2\%$ are maintained, across both testing and validation sets. In particular, at a decay factor of 0.1, i.e., a large room, a *near-perfect* performance is observed. Furthermore, if compared to the mel-spectrogram, the proposed quefrency-based representation consistently achieves high performance across all damping levels exhibiting slightly better consistency.

Interestingly, in most of the cases of Table IV, the opposite of the trend mentioned in Section IV-C can be noticed. This is because the training and testing sets share the same intensities of reverberation, i.e., the replay component, because of which this replay component becomes 'seen' even in the testing set, and hence we observe a better performance in the testing set.

V. SUMMARY AND CONCLUSIONS

This study addresses the detection of audio deepfakes in three different scenarios, including replayed-deepfakes. To simulate real-world replayed deepfake attacks, we introduce a novel reverberated version of the FoR dataset (rev-FoR-rerec) with varying acoustic environments through a proposed reverberation topology. The effectiveness of quefrency-based features is investigated for all three scenarios. Furthermore, subband-wise analysis is done for ADD, and reverberation analysis is done for replayed-deepfake detection. This work

demonstrates the superior effectiveness of quefrency-based representations (cepstrograms) over mel-spectrograms, especially in replayed-deepfake attack. The obtained results highlight the importance of room size and high-frequency damping, with larger rooms enhancing the discriminability between genuine and replayed-deepfake utterances. Subband analysis reveals that the subbands in the extreme ranges, i.e., 0-1000 Hz and 7000-8000 Hz are the most informative for deepfake detection, coinciding with the findings of VLD in literature, where subbands in the extremely low-frequency ranges (0 to 40 Hz) are discriminating. In the future, the efficiency of VLD systems can be tested on audio deepfakes.

REFERENCES

[1] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[3] B. Nguyen, "A couple in canada were reportedly scammed out of $21,000 after getting a call from an ai-generated voice pretending to be their son.," *The New York Times*, March 2023 {Last Accessed date : 24th June, 2024}.

[4] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved deepfake detection using whisper features," *arXiv preprint arXiv:2306.01428*, 2023.

[5] J.-w. Jung, H.-S. Heo, H. Tak, *et al.*, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2022, pp. 6367–6371.

[6] A. M. Almars, "Deepfakes detection techniques using deep learning: A survey," *Journal of Computer and Communications*, vol. 9, no. 05, pp. 20–35, 2021.

[7] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE transactions on information forensics and security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[8] G. Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, and M. Prasad, "A comprehensive review of deepfake detection using advanced machine learning and fusion methods," *Electronics*, vol. 13, no. 1, p. 95, 2023.

[9] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," *arXiv preprint arXiv:2308.14970*, 2023.

[10] R. A. M. Reimao, "Synthetic speech detection using deep neural networks," 2019.

[11] A. Mittal and M. Dua, "Automatic speaker verification systems and spoof detection techniques: Review and analysis," *International Journal of Speech Technology*, vol. 25, no. 1, pp. 105–134, 2022.

[12] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "Audio anti-spoofing detection: A survey," *arXiv preprint arXiv:2404.13914*, 2024.

[13] L. Pham, P. Lam, T. Nguyen, H. Nguyen, and A. Schindler, "Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models," *arXiv preprint arXiv:2407.01777*, 2024.

[14] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian Journal for Science and Engineering*, pp. 1–12, 2021.

[15] R. Anagha, A. Arya, V. H. Narayan, S. Abhishek, and T. Anjali, "Audio deepfake detection using deep learning," in *2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART)*, IEEE, 2023, pp. 176–181.

[16] P. Gupta and H. A. Patil, "Linear frequency residual cepstral features for replay spoof detection on asvspoof 2019," in *30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 349–353.

[17] P. Gupta, H. A. Patil, and R. C. Guido, "Vulnerability issues in automatic speaker verification (asv) systems," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 10, 2024.

[18] J. Dattorro, "Effect design, part 1: Reverberator and other filters," *Journal of the Audio Engineering Society*, vol. 45, no. 9, pp. 660–684, 1997.

[19] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE signal processing Magazine*, vol. 21, no. 5, pp. 95–106, 2004.

[20] L. R. Rabiner, R. W. Schafer, *et al.*, "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[21] O. B. Osman and M. H. Arbab, "Mitigating the effects of granular scattering using cepstrum analysis in terahertz time-domain spectral imaging," *PLoS One*, vol. 14, no. 5, e0216952, 2019.

[22] X. Miao, M. Sun, X. Zhang, and Y. Wang, "Noise-robust voice conversion using high-quefrency boosting via sub-band cepstrum conversion and fusion," *Applied Sciences*, vol. 10, no. 1, p. 151, 2019.

[23] B. P. Bogert, "The quefrency alanysis of time series for echoes: Cepstrum, pseudoautocovariance, cross-cepstrum and saphe cracking," in *Proc. Symposium Time Series Analysis, 1963*, 1963, pp. 209–243.

[24] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, IEEE, 2019, pp. 1–10.

[25] P. Gupta and H. A. Patil, "Morse wavelet transform-based features for voice liveness detection," *Computer Speech & Language*, vol. 84, p. 101 571, 2024.

[26] P. Gupta, P. K. Chodingala, and H. A. Patil, "Morlet wavelet-based voice liveness detection using convolutional neural network," in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 100–104.

Wavetable Synthesis Using CVAE for Timbre Control Based on Semantic Label

Tsugumasa Yutani*, Yuya Yamamoto*, Shuyo Nakatani†, Hiroko Terasawa*
* University of Tsukuba, Tsukuba, Japan
E-mail: tsugumasa320@gmail.com, s2130507@u.tsukuba.ac.jp, terasawa@slis.tsukuba.ac.jp
† Cybozu Labs, Tokyo, Japan
E-mail: nakatani@labs.cybozu.co.jp

***Abstract*—Synthesizers are essential in modern music production. However, their complex timbre parameters, often filled with technical terms, require expertise. This research introduces a method of timbre control in wavetable synthesis that is intuitive and sensible and utilizes semantic labels. Using a conditional variational autoencoder (CVAE), users can select a wavetable and define the timbre with labels such as $bright$, $warm$, and $rich$. The CVAE model, featuring convolutional and upsampling layers, effectively captures the wavetable nuances, ensuring real-time performance owing to their processing in the time domain. Experiments demonstrate that this approach allows for real-time, effective control of the timbre of the wavetable using semantic inputs and aims for intuitive timbre control through data-based semantic control. Audio examples are available on the online demo page[1].**

I. Introduction

Synthesizers, electronic musical instruments that create sounds using analog or digital signals, play a significant role in modern music production and performance. Although various methods of synthesizing sounds exist, effectively controlling the synthesis algorithm and obtaining the desired timbre requires specialized knowledge and experience. This complexity arises because numerous synthesis parameters are intricate, low-level, and unrelated to the perceptual and semantic properties of the timbre. This challenge is often called the synthesizer programming problem[1], which can even hinder the creative process for experienced musicians and professionals[2].

To address this problem, the concept of extracting synthesizer features has been proposed. [3]–[6]. This approach involves extracting features from a user-supplied sound (the target sound) and providing macro-controls that enable users to adjust the timbre based on these features, thereby simplifying the process of achieving the desired timbre.

Our research employs wavetable synthesis (WTS), one of the earliest software synthesis methods developed by Max Mathews in the late 1950s[7]. WTS, the core technology behind many synthesis methods, remains popular in numerous software synthesizers today. WTS works by cyclically referencing and replicating a waveform (WT: Wavetable) to produce a sound. The timbre is typically modified using filters, amplifiers, or morphing between different wavetables(WTs).

Furthermore, recent years have seen significant advancements in deep generative models. Among these, autoencoder (AE) [8] and variational autoencoder (VAE) [9] have gained prominence in various fields. These models compress the input data into a compact, low-dimensional latent space and reconstruct the original data from this compressed representation, facilitating the extraction of valuable feature representations from the data. Additionally, conditional variational autoencoder(CVAE)[10], which enables conditional generation based on specified conditions, has also been introduced.

Our research uses a CVAE to generate a WT for a single cycle, which is then applied to the oscillator within WTS. We have engineered a model that allows users to define the desired timbre through semantic labels. Figure 1 depicts a schematic overview of this process. In previous studies, methods for generating WT using deep generative models have been proposed[3], [11], [12]. However, issues regarding the diversity, controllability, and accuracy of the generated WT remain to be solved. Diverging from these methods, our research proposes a technique that enables users to emphasize or attenuate specific harmonic overtones within a given WT using semantic labels. This approach represents the primary contribution of our research.

II. Related Works

A. Synthesizer Programming Problem

The synthesizer programming problem, as outlined in Section I, is a complex issue that experienced musicians and professionals face. This challenge stems from the intricate control mechanisms and the necessity for specialized expertise, particularly in the realm of synthesizer sound creation [2]. Substantial research has focused on automating the estimation of synthesis parameters to address this issue. As a result, two approaches have emerged: the first aims to approximate synthesis parameters to replicate a target sound [13]–[16] and the second utilizes descriptive language concerning timbre for parameter estimation [17]–[19].

The feature synthesizer methodology warrants particular attention. This method is swift and user-friendly, specifically designed for music production and live performances. It involves extracting features from target timbres and providing macro-controls, thereby facilitating the adjustment of timbre based on these extracted features. This approach significantly simplifies the process of achieving the desired timbre. The efficacy of the feature synthesis method in enabling intuitive

[1] https://noto.li/XZKK4q

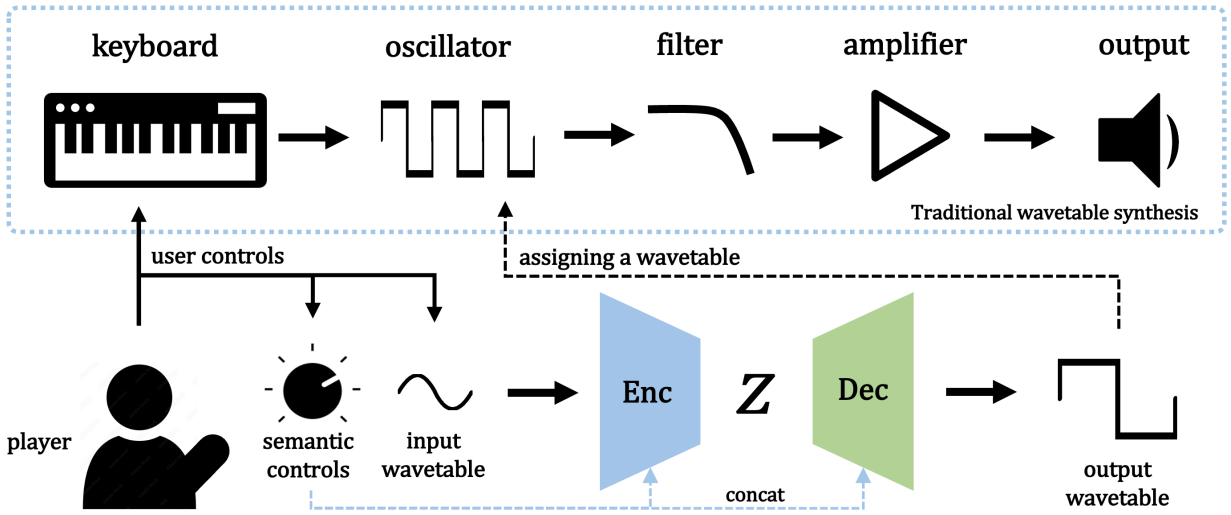


Fig. 1. Schematic showing an overview of the wavetable synthesis (WTS) method demonstrated in this research. A participant creates a wavetable (WT) exhibiting a specific timbre, achieved by conditioning the initial WT with various semantic labels. This generated WT is consistently referenced and emitted as a continuous sound. In this WTS setup, the pitch and duration can be adjusted using a keyboard or similar devices, akin to the conventional WTS operations. Additionally, the timbre of the WT can be modified by using filters and amplifiers.

timbre modification has been substantiated in various research studies[3]–[5], [11], [20]. Our research focuses on this feature synthesis approach, emphasizing its ability to rapidly adapt to changes in timbre during music production and performance.

B. Neural Audio Synthesis

Neural audio synthesis is a technology used for generating or transforming audio data using deep learning. As outlined in Section I, deep generative models are rapidly developing and have become an essential technology in music. Research to generate audio data is also active, and research for applications to synthesizers has been proposed[21]–[23].

Several approaches have been proposed to complement the existing synthesis methods using deep learning. To solve the synthesizer programming problem, Esling et al. [24] showed that VAE can simultaneously process parameter estimation, timbre control, and reconstruction with semantic labels in a single model for subtractive synthesis. Kreković [3] illustrated the feasibility of processing and controlling parameter estimation, timbre control, and reconstruction with semantic labels in a unified WTS model, highlighting the generation of WT from semantic labels.

This research explores the application of deep learning to musical synthesis, akin to the earlier research. It focuses on the latest technical trends in model construction, loss functions, and methods for calculating semantic labels.

C. Advances in Wavetable Synthesis

Wavetable synthesis generates a sound by repeatedly referencing a waveform for one cycle. This method, widely used in various genres, faces the common challenge of timbre generation, as with other synthesis methods. Creating a WT generally involves editing the waveform and frequency spectrum [25] or choosing a preset WT.

However, while these methods significantly impact timbre, they often diverge from human intuition. For instance, subtle changes in the shape or frequency spectrum of a WT can alter its timbre. Nonetheless, discerning the characteristics of the timbre through visual inspection is challenging. Even when selecting a WT from a preset, users must remember the relationship between the name of the preset and its corresponding timbre, which requires a certain level of expertise. Consequently, achieving the desired timbre in WTS can be difficult.

To address this problem, as outlined in Section I, previous research generated WT using VAE[12] and AE[11], and the timbre was changed by randomly manipulating the latent space. Furthermore, Kreković [3] proposed a model that generates WT from three input labels: $bright$, $warm$, and $rich$, using a decoder [26]. However, embedding WT features into a limited number of latent spaces (labels) poses challenges. The diversity and accuracy of the generated WT are still two issues that remain to be solved. To enhance the diversity and accuracy of the WT that can be generated, we have proposed a model that uses CVAE[10]. This model is designed to create WT that exhibits a one-period waveform characteristic.

III. Methods

A. CVAE

The CVAE[10], an extension of the VAE, is a generative model. This model is notably effective in generating data conditioned on a specific variable, denoted as c. In our research, we incorporated the condition, c, into the encoder and the decoder to apply these conditions effectively.

The input waveform, x, and condition, c, (in this research, WT and semantic labels, respectively) were mapped to the latent variable, z, by the encoder, $q_\phi(z|x,c)$. Subsequently, the decoder, $p_\theta(x|z,c)$, generated the output data, $\hat{x}$.

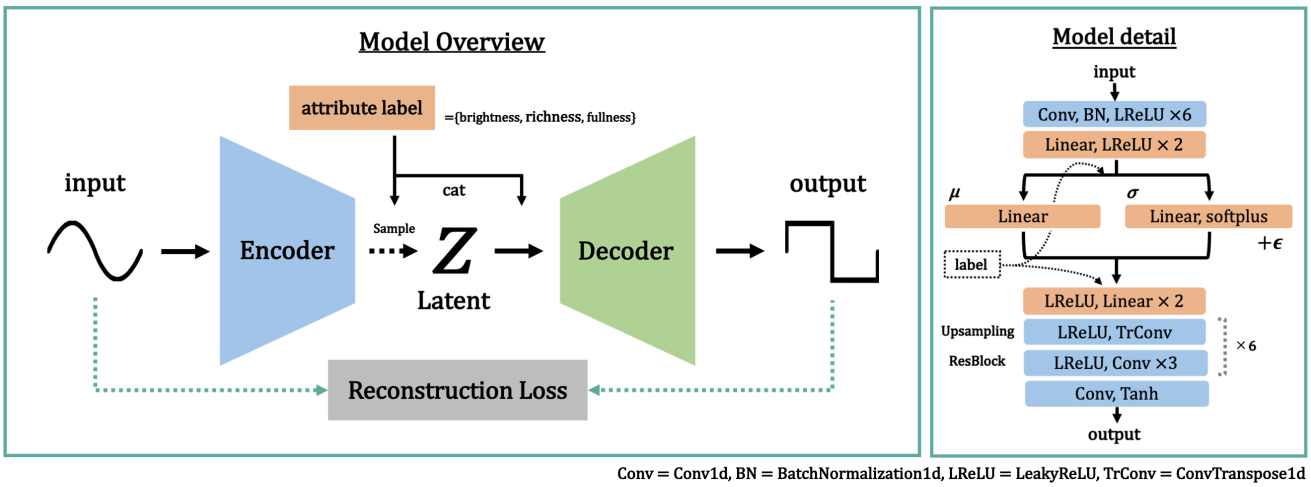


Fig. 2. Schematic of the model configuration. Left: Model overview, right: detailed configuration.

B. Loss Function

The loss function comprised the reconstruction loss and the Kullback–Leibler (KL) divergence term. These components were optimized by minimizing them through the stochastic gradient descent (SGD) method. The loss function in this research was defined as follows:

$$L_{\mathrm{CVAE}}(x) = \mathbb{E}_{\hat{x}\sim p_\theta(x|z)}[d(x,\hat{x})] + \beta \times D_{\mathrm{KL}}[q_\phi(z|x,c)||p(z)] \tag{1}$$

The first term in Equation (1), the reconstruction loss, was used for evaluating the accuracy of the generated data in reconstructing the original data and aimed to minimize the information loss between the encoder and decoder. Previous research has suggested using multiscale spectral distance to estimate the waveform distances[21], [23]. Due to the short data length in our WT approach, we calculated the spectral distance using only the maximum table size (3600 samples in our research) after iterative concatenation.

The function, d, for the reconstruction loss was defined as follows:

$$d(x,\hat{x}) = \frac{||S_x - S_{\hat{x}}||_F}{||S_x||_F} + \log||S_x - S_{\hat{x}}||_1 \tag{2}$$

where S_x and $S_{\hat{x}}$ are the amplitude spectra of x and $\hat{x}$, respectively. Here, $||\cdot||_F$ denotes the Frobenius norm and $||\cdot||_1$ the L1 norm.

The amplitude spectrum of the input waveform, x, and the output waveform, $\hat{x}$, were compared and their loss was calculated. This approach is based on the assumption that the perceptual characteristics of the synthesized sound remain almost invariant between the signal waveforms that are out of phase. Following the convention adopted by Kreković [3], we improved the frequency resolution of the spectrum by concatenating six WTs iteratively, which achieved the best performance in the investigation stage in terms of reconstruction error and controllability of conditioning.

The second term of Equation (1), the KL divergence term, measured the closeness of the latent variable, z, sampled from the encoder, $q_\phi(z|x,c)$, to the predefined multivariate normal distribution, $p(z)$. This term regularized the latent space, ensuring that the model captured the general data characteristics instead of merely memorizing the data.

The hyperparameter, β, adjusted the balance between the reconstruction loss and the KL divergence term. The setting of this parameter significantly influenced the performance of the model, as proposed by Higgins et al. [27].

C. Model Configuration

The model's architecture was adapted from a structure outlined in previous research[23] and was modified to enable the conditional generation of WT. The encoder combined the convolutional layers that convert the WT into a 32-dimensional latent representation. The decoder used an upsampling layer and a residual network[28] to generate the WT. Conditional generation was executed by concatenating the relevant labels to the inputs and outputs within the latent space[12]. Figure 2 illustrates an overview of the model structure.

D. Computation of Semantic Labels

Previous research has explored the computation of semantic labels in sounds, including WTS. In our research, we employed the labels $bright$, $warm$, and $rich$, following the methodology outlined by Kreković [3], [29]. These labels were derived from the spectral centroid, the energy ratio of the odd harmonics, and the spectral density, followed by normalization.

The spectral centroid, representing the center of mass of the spectrum, indicates the balance between the high and low frequencies and correlates with the timbre brightness[30]. The energy ratio of the odd harmonics, i.e., the proportion of energy in the odd harmonics compared to other harmonics, was associated with the warmth of the timbre[31]. The spectral density, calculated as the spectrum's standard deviation, corresponded to the richness of the timbre[32].

Six WTs were concatenated in succession to improve spectral resolution, following the method outlined in Section III-B

for extracting features from a single WT. Audio features were computed using a fast Fourier transform and a Hanning window. Since the derived values do not naturally correspond with human perception, a transformation was applied to ensure linear alignment with human perception, particularly within the range of [0, 1]. However, the energy ratio of the odd harmonics for the $warm$ label, inherently falling within [0, 1], required no additional normalization.

IV. Experiments

A. Dataset

We utilized 4158 WT data from Adventure Kid Research & Technology[2] for the dataset. For the size of the sample for each WT, we select 600, which is larger than that utilized in previous works (i.e., 320 in [11], 327 in [3], and 512 in [12].), to enhance the representativeness of high-frequency components in WTS. The dataset was partitioned into three segments: 80 % for training, 10 % for validation, and 10 % for testing. This distribution ensured a comprehensive evaluation of the model across different data subsets.

B. Training

Training was conducted by feeding the model with training data and the corresponding labels. The Adam optimizer[33] was used with a learning rate 1e-4. The batch size was 32, and the number of epochs was 30000 to minimize the reconstruction and KL divergence.

Over-optimizing the KL divergence between the prior and posterior distributions in the early stages of training can lead to an issue known as KL vanishing[34], in which the latent variables are ignored in the reconstruction. To mitigate this issue, the variable, β, for the KL divergence was gradually increased according to the following Equation:

$$\beta(e, w, \beta_{\min}, \beta_{\max}) = \begin{cases} \beta_{\max} & \text{if } e > w \\ \exp\left(t \cdot (\log \beta_{\max} - \log \beta_{\min}) + \log \beta_{\min}\right) & \text{otherwise} \end{cases} \tag{3}$$

$$t = \frac{e}{w} \tag{4}$$

where e is the current number of epochs, w is the upper bound on the number of epochs, $\beta_{\min}$ is the minimum value set for β, $\beta_{\max}$ is the maximum value set for β, starting from $\beta_{\min} > 0$ and increasing exponentially to $\beta_{\max}$ exponentially as the learning progresses.

We empirically set $w = 10000$, $\beta_{\min} = 1e-4$, $\beta_{\max} = 1e-1$, and 1 epoch for the period of the increasing.

V. Evaluation

The reconstruction quality of the generated WT and the controllability and generation time of the conditioning generation were evaluated. This research aimed to enhance the intuitiveness of timbre control and enable real-time timbre searches. To achieve this, the following three steps were essential: 1) high-quality reconstruction of the WT input to the model 2) conditional generation according to the values of the semantic labels, and 3) short processing time for generation and real-time use. The results obtained by performing the abovementioned steps are shown below. The source code of the model and the weights of the trained model used to obtain the results are also available[3]

A. Reconstruction Quality

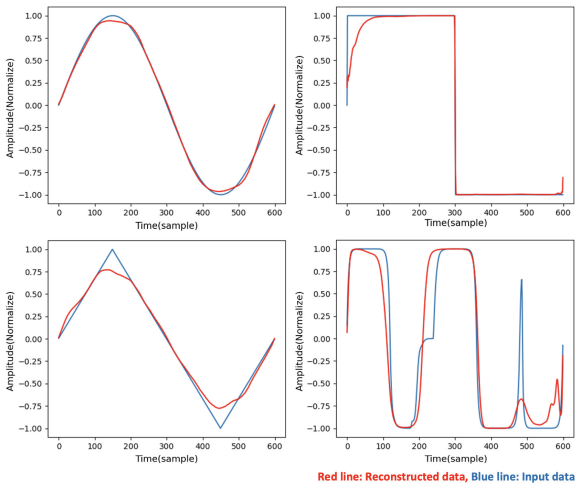


Fig. 3. Example of the qualitative evaluation of the reconstruction quality. The plots show that the reconstruction is done according to the input data.

The mean absolute error (MAE) was calculated to evaluate the reconstruction quality between the reconstructed waveform and the original input waveform. The calculated MAE amounted to 0.082 for the waveform state and 0.013 for the logarithmic amplitude spectrum. The logarithmic amplitude spectrum was calculated by repeatedly concatenating the six WTs as described in Section III-B to facilitate a comparison between the frequency components. The reconstructed waveforms are shown in Figure 3, which reproduce the input waveforms.

These results indicate that the reconstructed waveform successfully emulates the input waveform, capturing the essential characteristics of the WT. Future work will entail a comparison of our method with other models and conducting listening experiments to validate the reconstruction quality quantitatively.

B. Controllability of conditioning generation

To evaluate the controllability of the conditioning generation, the MAE and Pearson correlation coefficients of the output labels concerning the input labels (condition c) were checked. The output labels were computed from the output waveforms using the same procedure as that described in Section III-D. The higher number of minor differences between the input and output labels and higher correlation coefficients indicate better control of conditioning generation. Figure 4

[2]https://www.adventurekid.se/akrt/

[3]https://github.com/tsugumasa320/WavetableCVAE

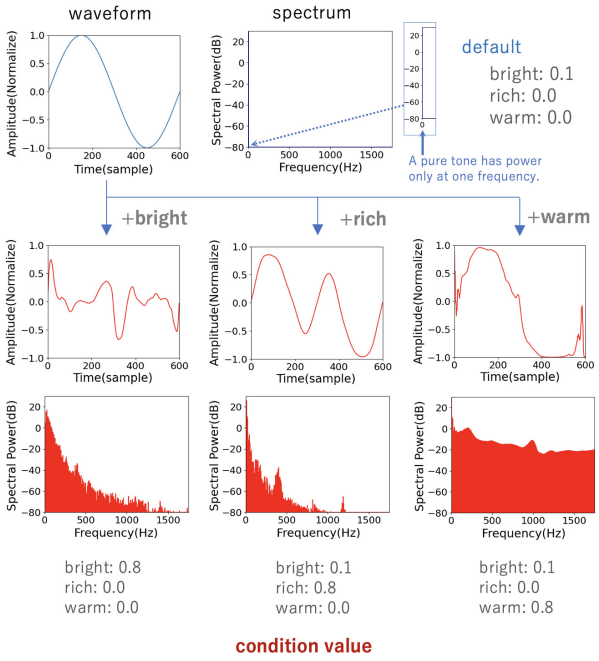


Fig. 4. Outcomes of conditional generation using sine waves as inputs: an analysis based on the Label characteristics. *bright* refers to the spectral centroid, *rich* denotes the energy ratio of the odd harmonics, and *warm* corresponds to the spectral density. The value associated with each label increases with the value of the conditioning.

illustrates a section of the results from the conditioning generation process.

The calculation procedure was as follows: A single input waveform, x, and an input label, c, from the dataset were selected. The input label, c, was a variable that comprised three precomputed values: $bright$, $warm$, and $rich$. From c, one evaluation label, c_e, was selected to evaluate the controllability of conditioning generation. The label c_e was set in order of 20 equally spaced values in the range of [0, 1]. Labels that were not used for evaluation were also needed for conditioning generation. The labels obtained from the input waveforms were calculated using the same procedure as in Section III-D.

The output waveform, $\hat{x}$, and the evaluation label, $\hat{c_e}$, were obtained from the input waveform, x, and the input label, c. This procedure was repeated for the number of divisions of c_e used as input (20 times in this research).

The transition between c_e and $\hat{c_e}$ was stored as an array of 20 points. The MAE and Pearson correlation coefficients were calculated and evaluated from the stored arrays. These processes were repeated for WT, and the averages were obtained with the three parameters, $bright$, $warm$, and $rich$, for the entire dataset.

The value of MAE averaged over $bright$, $warm$, and $rich$ were 0.19, 0.05, and 0.14, respectively. The $warm$ value means the change is consistent with the label, whereas $bright$ and $rich$ remain challenging. The distribution of the training labels may have been biased, limiting the exposure of data to a sufficient range of labels for practical training.

The averaged values of the Pearson correlation coefficients were 0.97 for $bright$, 0.98 for $warm$, and 0.96 for $rich$. Strong positive correlations were confirmed for all parameters. This indicates that conditioning generation responds appropriately to the increase or decrease of each parameter.

C. Generation Time

The time required for generating WT was evaluated to ascertain the real-time performance capabilities of the proposed method. WT was generated 100 times, and the average generation time was calculated to be approximately 2.7 ms. The evaluations were conducted on a MacBook Pro (13-inch, M1, 2020) with an M1 Chip (CPU), utilizing PyTorch Lightning[4] 1.7.7. This notably brief generation time suggests the feasibility of real-time operation of our method using only the CPU without requiring a specialized GPU for deep learning.

VI. Summary and Future Issues

In this research, we proposed a method for semantic timbre control of WT using CVAE, and demonstrated the possibility of improving the synthesizer programming problem. The following is the rationale for the proposed method:

This method demonstrates that the reconstructed waveform can accurately follow the input waveform by effectively capturing the timbre features of the WT. Additionally, we have confirmed that conditioning generation can be controlled according to the intended outcome[5]. However, the inherent complexity of timbre variation indicates that more refined control is essential. Future research should explore increasing the number of conditioning labels to facilitate more nuanced sound manipulation, thereby contributing to the resolution of the synthesizer programming problem.

Research examining the application of deep learning to WTS is difficult to evaluate in a unified manner due to the variety of model architectures and research objectives[3], [11], [12]. As outlined in Section II-C, future research will need to develop quantitative metrics to assess the quality of waveform reconstruction and the controllability of conditioning, enabling better comparisons across different models and providing a more coherent basis for advancing WTS research..

Furthermore, as detailed in Section V-B, the controllability assessment of the conditioning generation indicates significant MAE, particularly with $bright$ and $rich$ labels. The bias in training label distribution is likely the contributing factor, necessitating countermeasures. Results may be enhanced by augmenting the dataset size, altering label analysis, and normalization methods. Another strategy could involve adopting latent space and label feature separation mechanisms akin to those in Fader Networks [35].

With regards to the generation time, the results reveal that real-time use is possible using only the CPU without needing a particular GPU for deep learning.

As for real-time performance, model weight reduction methods such as model parameter quantization and branch pruning could further reduce the processing time of the proposed method.

[4]https://lightning.ai/docs/pytorch

[5]The reconstruction accuracy is shown in Figure 3, and the controllability of conditioning generation is shown in Figure 4

References

[1] J. Shier, *The synthesizer programming problem: Improving the usability of sound synthesizers*, Master's thesis, University of Victoria, 2021.

[2] G. Krekovic, "Insights in habits and attitudes regard-ing programming sound synthesizers: A quantitative study," in *Proc. SMC*, 2019.

[3] G. Kreković, "Deep Convolutional Oscillator: Synthesizing Waveforms from Timbral Descriptors," *Proc. SMC*, 2022.

[4] D. L. Wessel, "Timbre space as a musical control structure," *Computer music journal*, pp. 45–52, 1979.

[5] C. Hourdin, G. Charbonneau, and T. Moussa, "A sound-synthesis technique based on multidimensional scaling of spectra," *Computer Music Journal*, vol. 21, no. 2, pp. 56–68, 1997.

[6] A. Horner, J. Beauchamp, and L. Haken, "Machine tongues xvi: Genetic algorithms and their application to fm matching synthesis," *Computer Music Journal*, vol. 17, no. 4, pp. 17–29, 1993.

[7] M. Puckette, "Max at seventeen," *Computer Music Journal*, vol. 26, no. 4, pp. 31–43, 2002.

[8] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[9] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. ICLR*, 2014.

[10] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. NIPS*, vol. 27, 2014.

[11] G. Kreković, "A concept of a wavetable oscillator based on a neural autoencoder," in *Proc. Audiomostly*, 2021, pp. 240–243.

[12] J. Hyrkas, "WaVAEtable Synthesis," *Proc. CMMR*, p. 263, 2021.

[13] N. Masuda and D. Saito, "Synthesizer sound matching with differentiable dsp.," in *Proc. ISMIR*, 2021, pp. 428–434.

[14] N. Masuda and D. Saito, "Quality-diversity for synthesizer sound matching," *Journal of Information Processing*, vol. 31, pp. 220–228, 2023.

[15] M. Chinen and N. Osaka, "Genesynth: Noise band-based genetic algorithm analysis/synthesis framework," in *Proc. ICMC*, 2007.

[16] M. Macret, "Automatic tuning of the op-1 synthesizer using a multi-objective genetic algorithm," in *Proc. SMC*, 2013.

[17] A. Pošćić and G. Kreković, "Controlling a sound synthesizer using timbral attributes," in *Proc. SMC*, 2013, pp. 467–472.

[18] G. Kreković and D. Petrinović, "Intelligent exploration of sound spaces using decision trees and evolutionary approach," in *Proc. ICMC/SMC*, 2014, pp. 1263–1270.

[19] S. Fasciani, "TSAM: A tool for analyzing, modeling, and mapping the timbre of sound synthesizers," in *Proc. SMC*, Aug. 2016, pp. 129–136.

[20] F. Roche, T. Hueber, M. Garnier, S. Limier, and L. Girin, "Make that sound more metallic: Towards a perceptually relevant control of the timbre of synthesizer sounds using a variational autoencoder," *Transactions of the International Society for Music Information Retrieval*, vol. 4, pp. 52–66, 2021.

[21] J. Engel *et al.*, "Neural audio synthesis of musical notes with wavenet autoencoders," in *Proc. ICML*, PMLR, 2017, pp. 1068–1077.

[22] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "GANSynth: Adversarial neural audio synthesis," in *Proc. ICLR*, 2019.

[23] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

[24] P. Esling, N. Masuda, A. Bardet, R. Despres, and A. Chemla-Romeu-Santos, "Flow Synthesizer: Universal Audio Synthesizer Control with Normalizing Flows," *Applied Sciences*, vol. 10, no. 1, 2020.

[25] R. Bristow-Johnson, "Wavetable synthesis 101, a fundamental perspective," in *Audio engineering society convention 101*, Audio Engineering Society, 1996.

[26] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[27] I. Higgins *et al.*, "Beta-vae: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, 2017.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[29] G. Kreković, A. Pošćić, and D. Petrinović, "An algorithm for controlling arbitrary sound synthesizers using adjectives," *Journal of New Music Research*, vol. 45, no. 4, pp. 375–390, 2016.

[30] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1493–1500, 1978.

[31] K. K. Jensen, "Timbre models of musical sounds - from the model of one sound to the model of one instrument," in *Technical report / University of Copenhagen / Datalogisk institut*, 1999.

[32] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," *CUIDADO Ist Project Report*, vol. 54, no. 0, pp. 1–25, 2004.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2014.

[34] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, "Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing," in *Proc. NAACL*, 2019, pp. 240–250.

[35] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader networks: Manipulating images by sliding attributes," *Proc. NIPS*, vol. 30, 2017.

Frequency & Channel Attention Network for Small Footprint Noisy Spoken Keyword Spotting

Yuanxi Lin and Yuriy Evgenyevich Gapanyuk
Bauman Moscow State Technical University, Russia
linyu@student.bmstu.ru, gapyu@bmstu.ru

***Abstract*—In this paper, we aim to improve the robustness of Keyword Spotting (KWS) systems in noisy environments while keeping a small memory footprint. We propose a new convolutional neural network (CNN) called FCA-Net, which combines mixer unit-based feature interaction with a two-dimensional convolution-based attention module. First, we introduce and compare lightweight attention methods to enhance noise robustness in CNN. Then, we propose an attention module that creates fine-grained attention weights to capture channel and frequency-specific information, boosting the model's ability to handle noisy conditions. By combining the mixer unit-based feature interaction with the attention module, we enhance performance. Additionally, we use a curriculum-based multi-condition training strategy. Our experiments show that our system outperforms current state-of-the-art solutions for small-footprint KWS in noisy environments, making it reliable for real-world use.**

I. Introduction

Keyword Spotting (KWS) [1] is a technology that detects specific keywords in audio streams. It aims to recognize trigger words or commands, allowing smart devices to work without physical contact. This technology is widely used in consumer electronics such as smart speakers, wearables, voice assistants, and Internet of Things devices [2]. We focus on small-footprint KWS in this paper, which requires a small memory footprint and low computational power. This kind of technology is practical because the KWS models are always powered on low-resource devices to provide a constant service.

Recent small-footprint KWS models based on the deep neural network (DNN) have achieved success with less noisy, close-talking audio datasets. Chen et.al [3] first utilize the DNN to regard keyword spotting as a classification problem. Tang and Lin [4] proposed a residual network (ResNet) based KWS system where they used dilated convolution to enlarge the size of the receptive field exponentially with the depth of the network. Choi et.al [5] propose a temporal convolutional neural network (TC-ResNet) for real-time KWS on mobile devices, with temporal convolution reducing computation and increasing performance in terms of accuracy compared to 2D convolutions. However, these systems face significant challenges in noisy environments because background noise can interfere with the detection of keywords, leading to false activations or missed detections.

Improving noise robustness in KWS systems is a crucial task to ensure reliable performance [6]. The motivation to handle this kind of environment is driven by the need for KWS to function accurately in real-world settings where noise is often present, such as in busy households, public spaces, and industrial areas. Enhancing noise robustness will make KWS systems more effective and user-friendly, ensuring they can be relied upon in diverse and challenging auditory conditions.

To augment the noise robustness of acoustic models, curriculum learning has been proposed [7]. As a learning method, curriculum learning increases the difficulty of training samples progressively. It starts with initial training on easy, clean samples and then continues with difficult samples containing loud noise, similar to how a person studies according to a curriculum. Curriculum learning is widely used for noise-robust KWS [8].

Recent work on Ng et.al [9], [10] for small-footprint KWS addresses the challenge of noisy environments. They propose a new encoder architecture based on a mixer module. This mixer computes weighted feature interactions across global channels, improving information flow. The model ConvMixer shows good performance in noisy, far-field conditions. However, there is still room to improve with efficient attention modules to extract local features, such as SE [11] and ECA [12]. This improvement is important because local features can enhance the model's ability to handle varying noise levels and complex audio inputs, leading to more accurate keyword spotting.

In this paper, we aim to enhance the robustness of KWS in noisy environments while ensuring a small memory footprint. We propose a novel network called FCA-Net combining feature interaction in ConvMixer with an efficient attention module. First, we introduce and compare various lightweight attention methods to improve the noise robustness of KWS systems. Then, we propose an efficient two-dimensional convolution-based attention module for noisy KWS. This module produces fine-grained attention weights to capture channel and frequency-specific information, enhancing the model's representation ability for noise-robust speech. Additionally, we implement a curriculum-based multi-condition training strategy. Our

experiments show that our system outperforms existing state-of-the-art solutions for small-footprint KWS in noisy conditions.

II. Related work

Small Footprint Keyword Spotting – With the widespread adoption of voice interfaces in smart consumer electronics, the application of small convolutional neural networks (CNNs) in compact keyword spotting has become increasingly significant. Recent works investigated innovative convolution techniques to improve KWS performance. TC-ResNet [5], [13] applies 1D temporal convolution to enhance efficiency and accuracy. BC-ResNet [14] introduces broadcasted residual learning combining 1D and 2D convolutions. However, most studies do not consider noise robustness, which is a crucial factor for successful voice interfaces.

Noise robust Spoken Keyword Spotting - Data augmentation, especially noise mixture training, has been widely used to enhance the noise robustness of acoustic models [15]. Noise mixture training involves adding environmental noise to the training data, enabling the model to better handle noisy environments. However, small neural networks struggle to learn in high-noise conditions. To address this issue, curriculum learning has been introduced, which gradually increases the difficulty of training samples to improve the model's robustness [7]. Although these methods have improved performance to some extent, they still fall short when dealing with high-noise environments.

Attention mechanisms in CNNs - Attention mechanism address the limitations of traditional CNNs in handling long-range dependencies and feature selection. It aims to enable the model to automatically focus on more important information regions. SE-Net [11], [16] introduces an effective mechanism for learning channel attention, achieving significant and promising performance outcomes for the first time. Subsequently, Woo et al. [17] proposed the Convolutional Block Attention Module (CBAM), which further enhanced model performance by combining channel attention and spatial attention. Attention mechanisms, while powerful, often come with significant computational and memory overheads. These overheads can hinder the deployment of models on edge devices. Consequently, recent research has focused on developing light attention mechanisms. ECA-Net [12] simplifies the attention mechanism by using an efficient channel attention module that avoids dimensionality reduction and performs local cross-channel interaction without significantly increasing the computational complexity. However, integrating these attention mechanisms with the noise-robust spoken keyword spotting systems is still under exploration.

III. Methodology

To overcome the difficulty of learning the noise-robustness property in small-footprint keyword spotting, we propose a joint framework that combines ConvMixer block and attention mechanism.

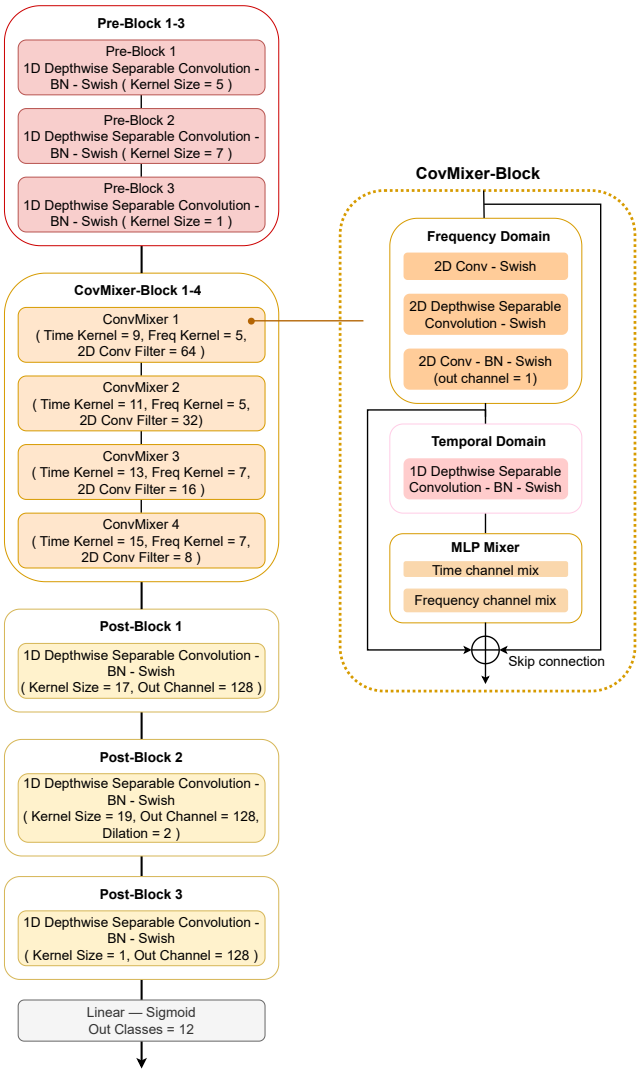


Fig. 1. Overview of the ConvMixer model architecture, which integrates 2D and 1D depthwise separable (DWS) convolutions to extract features from both frequency and temporal domains. The architecture includes pre-convolutional, ConvMixer, and post-convolutional blocks, each performing feature extraction and interaction through different kernel sizes and filters.

A. ConvMixer architecture

The ConvMixer model architecture is divided into three main sections: the pre-convolutional block, the ConvMixer block, and the post-convolutional block. Fig. 1 provides an overview of this architecture.

The pre-convolutional and post-convolutional block includes 1D depthwise separable (DWS) convolution, batch normalization, and swish activation.

The ConvMixer block integrates 2D and 1D DWS convolutions to extract features from both the frequency and

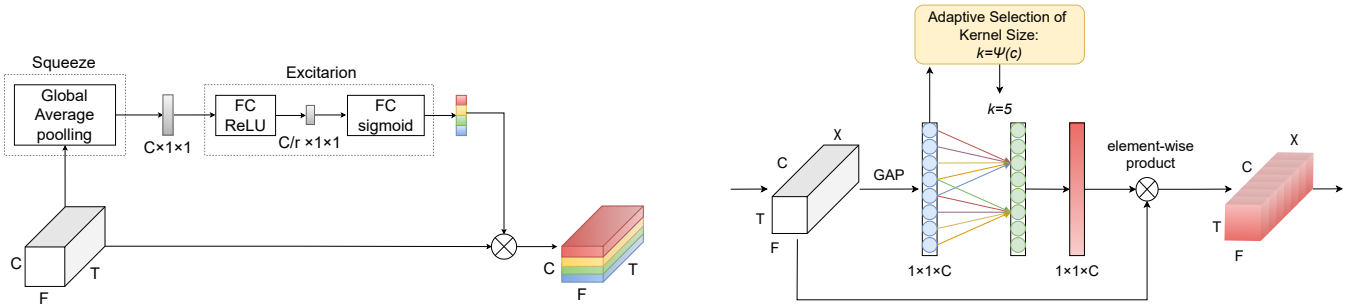


Fig. 2. Illustration of Squeeze-and-Excitation (SE) and Efficient Channel Attention (ECA) mechanisms. SE utilizes global average pooling (GAP) followed by fully connected (FC) layers to generate channel-wise attention weights, enhancing important features. ECA avoids dimensionality reduction by using 1D convolution, adaptively selecting kernel sizes based on channel dimensions to improve efficiency while maintaining strong feature interactions.

temporal domains. It also incorporates a mixer layer to enhance information flow. The following equations describe the ConvMixer block:

$$z = \sigma \circ f_1(\sigma \circ f(x)) \tag{1}$$

In the ConvMixer block, f is a 2D-DWS convolution, and σ is the Swish activation function. First, the input x is convolved by the 2D-DWS convolution function f, and then the result is processed by the activation function σ to obtain the intermediate features. These intermediate features are convolved again by the 2D-DWS convolution f, and processed by the activation function σ, resulting in the frequency domain features z.

$$y_1 = \sigma \circ \text{BatchNorm}(f(z)) \tag{2}$$

Then the frequency domain features z is convolved by a 1D-DWS function f_2. The result is processed by a batch normalization (BatchNorm) layer and finally activated by the function σ, resulting in the temporal domain features y_1.

$$y_2 = \sigma \circ \text{BatchNorm}(f_2(y_1)) \tag{3}$$

The temporal domain features y_1 are processed by another 1D-DWS convolution function f_2, then passed through a BatchNorm layer and activation, resulting in the processed features y_2.

$$\tilde{y} = x + y_1 + f_3(y_2) \tag{4}$$

Finally, the input x, intermediate features y_1, and processed features y_2 are added together to obtain the output $\tilde{y}$ of the ConvMixer module. Equation 4 computes the processed features y_2 of the block with f_3 as the mixer layer. We propose using two types of multi-layer perceptrons (MLP): temporal channel mixing and frequency channel mixing, to induce interaction between the feature space. Each MLP mixing involves two linear layers and a GELU activation unit, independently processing each temporal and frequency channel.

B. Efficient attention module

1) Squeeze-and-excitation block (SE): The Squeeze-and-Excitation (SE) block [11] is an efficient attention mechanism used in CNNs to improve performance. As shown in Fig. 2 (a), the SE block has two main operations: Squeeze and Excitation. The Squeeze operation performs global average pooling on the feature maps across the channel dimension to create a compact intermediate representation. The Excitation operation uses two fully connected (FC) layers to assign importance weights to each channel. The SE module re-weights the output of each channel, highlighting important features and suppressing less significant ones. This helps the network focus on the most informative features, enhancing performance with minimal additional computational cost.

2) Efficient channel attention (ECA): The Efficient Channel Attention (ECA) [12] block is a highly efficient channel attention mechanism for deep CNNs. As illustrated in Fig. 3, the ECA block first uses global average pooling to aggregate features. It then applies a 1D convolution with an adaptively determined kernel size, generating channel weights through a Sigmoid function. Unlike other methods, the ECA block avoids dimensionality reduction via fully connected layers, maintaining a direct correspondence between channels and their weights.

The ECA block captures local cross-channel interactions through 1D convolution, improving performance while maintaining efficiency. The kernel size k of this 1D convolution is adaptively determined based on the channel dimension C , ensuring that higher-dimensional channels interact more broadly, while lower-dimensional channels interact more narrowly. The ECA block achieves effective channel attention with low model complexity, significantly enhancing performance and demonstrating strong generalization in practical tasks.

3) Convolution-based two-dimensional attention (C2D): The C2D block is integrated into CNNs to calculate fine-grained attention weights for various channel and frequency positions. This module enhances the efficiency

and effectiveness of the attention mechanism by using convolution operations instead of the fully connected layers typically used in traditional attention mechanisms.

The input feature map X of C2D block contains C channels, F frequency bins, and T time frames. This feature map is the output of a previous convolutional layer in the network. Then the Global Average Pooling (GAP) is applied along the time dimension of the input feature map X.

$$z(c,f) = \frac{1}{T}\sum_{t=1}^{T} X(c,f,t) \tag{5}$$

The result is a channel-frequency plane z with dimensions $C \times F$, where each element represents the average value of the corresponding channel and frequency bin across the time dimension. Next, the pooled channel-frequency plane z is passed through two consecutive 2D convolutional layers. These convolution operations are used to calculate the attention weights ω for different channel and frequency positions. The convolutional layers include BatchNorm and ReLU activation functions to enhance learning ability and stability. The attention weights ω represent the importance of each channel-frequency pair.

$$\omega = \sigma(\text{Conv2}(\text{ReLU}(\text{BN}(\text{Conv1}(z))))) \tag{6}$$

The calculated attention weights ω are then applied to the original input feature map X through element-wise multiplication. This step re-scales the input features, enhancing important features and suppressing less important ones as the following equation:

$$X' = \omega \odot X \tag{7}$$

This operation re-scales the input features, emphasizing important channel-frequency pairs and suppressing less important ones, thereby enhancing the representation capability of the network.

The C2D block calculates attention weights using convolution operations, which require fewer parameters and less computational cost compared to fully connected layers. By simultaneously considering both channel and frequency dimensions, the C2D block generates more detailed and specific attention weights, enhancing feature representation capability.

C. Proposed FCA-Net

Our proposed FCA-Net combines the ConvMixer with the C2D attention mechanism for keyword spotting. We add the C2D attention after each ConvMixer block, referring to this configuration as FCA-Net-all. Additionally, we explore other variants: FCA-Net-pre, which adds C2D attention only to the pre-convolutional blocks; FCA-Net-post, which adds C2D attention only to the post-convolutional blocks; and FCA-Net-final, which adds C2D attention before the final linear layer to study its impact on the final classification. We setup various experiments for different variants in the ablation study.

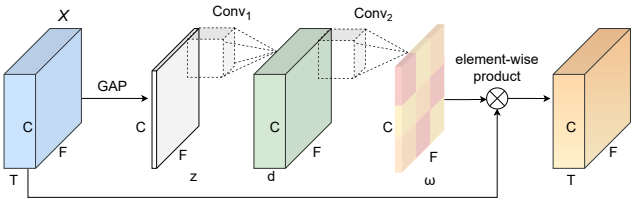


Fig. 3. Network structure of the Convolution-based two-dimensional (C2D) attention block. The C2D block calculates attention weights for each channel and frequency position by applying 2D convolution to the global average pooled feature maps. This method generates fine-grained attention weights, enabling more precise feature selection across both temporal and frequency domains.

IV. Experiment setting

A. Dataset for Noisy Keyword Spotting

We evaluate our proposed system on the Google Speech Commands V2 dataset [22], which contains 105,000 utterances of 35 unique words, each 1 second long and sampled at 16 kHz. We use the official train, validation, and test split for the 12-label classification task. These labels include: 'up', 'down', 'left', 'right', 'yes', 'no', 'on', 'off', 'go', 'stop', 'silence', and 'unknown'. The 'unknown' class includes the remaining words in the dataset.

To simulate a noisy far-field environment, we use noise samples from the MUSAN dataset [23], which has 930 files of assorted noises, about 6 hours total, all sampled at 16 kHz. These noises include technical and non-technical sounds like DTMF tones, thunder, and car horns. We add these noises to our command samples to mimic various noisy conditions. This process is divided into four progressively harder steps. Initially, the model is trained on clean samples without noise. In the following three steps, noise is incrementally introduced to the fixed N samples in decrements of -5dB, with conditions in N samples uniformly distributed, i.e., [clean, 0], [clean, 0, -5], [clean, 0, -5, -10].

B. Implement details

We use Mel-Frequency Cepstral Coefficients (MFCC) with the following parameters: sample rate of 16 kHz, 40 MFCCs, and mel-spectrogram settings of n_fft=400, hop_length=160, n_mels=64, f_min=20, and f_max=8000. Each command is 1 second long; shorter commands are zero-padded to the right. During training, we perform data augmentation with time shifts of -100 to 100 ms and spectrogram masking with maximum lengths of 25 for both time and frequency. We generate noisy data with SNR of 0, -5, and -10 dB.

For stronger regularization, we use input mixup with a ratio of 0.2 on training samples. The model is trained with a batch size of 128 and an initial learning rate of 0.005, reduced by a factor of 0.85 every four epochs after

TABLE I
COMPARISON WITH THE SOTA MODELS.

Model	Num. of Params (K)	MACs (M)	Acc. of V2-12 Official (%)	Accuracy of Noisy Test Command(%)				
				Clean	20 dB	0 dB	-5 dB	-10 dB
MHAtt-RNN[18]	784	141.6	98.04	78.83	74.25	61.98	55.87	50.98
KWT-1[19]	607	53.8	97.72	88.73	85.46	73.52	67.59	59.07
ResNet-15[4]	238	961.2	96.48	89.45	87.34	79.00	73.58	66.73
MatchboxNet-6x2x64[20]	140	36.8	97.60	87.34	85.19	75.58	70.06	62.35
BC-ResNet-6[14]	202	55.4	98.20	93.67	93.30	87.48	83.06	79.40
ConvMixer[9]	119	22.2	98.20	96.32	95.52	90.62	86.37	77.47
FCA-Net (Ours)	119	22.3	98.25	**96.96**	**95.99**	**92.65**	**87.97**	**80.40**
AST-Tiny[21]	5,805	782.2	97.65	91.02	87.71	83.31	78.95	72.32
KWT-3[19]	5,361	526.3	98.54	93.47	91.08	83.97	78.45	71.08

TABLE II
EXPLORATION OF DIFFERENT ATTENTION MODULES AND POSITIONS OF THE PERFORMANCE

Attention	Position	Accuracy of Noisy Test Command (%)				
		Clean	20 dB	0 dB	-5 dB	-10 dB
-	-	96.32	95.52	90.62	86.37	77.47
ECA	pre	95.23	94.49	89.37	85.24	76.97
	post	96.77	95.83	90.75	86.20	78.12
	all	96.73	95.68	91.86	86.60	80.00
	final	96.67	95.85	91.14	85.90	78.27
SE	pre	95.23	94.59	89.70	84.50	75.16
	post	94.70	94.22	89.64	84.89	76.85
	all	95.50	94.86	89.82	85.20	76.42
	final	95.20	94.10	89.97	85.51	76.75
C2D	pre	95.70	94.47	89.08	84.07	75.88
	post	95.27	94.65	89.02	84.00	75.56
	all	96.96	95.99	92.65	87.97	80.40
	final	96.92	96.09	91.18	86.82	78.42

the fifth epoch. We use the Adam optimizer and a learning rate scheduler with Cosine Annealing Warm Restarts. The model is trained for 200 epochs with early stopping if there is no improvement for 20 epochs. This study applied three augmentation methods: mixup [24], SpecAugment [25], and curriculum learning [7], [9].

C. Baselines

We also set up some SOTA keyword spotting models as the baseline to analyze our proposed model.

V. RESULTS

A. Compare study among the SOTA models

We compare the performance of FCA-Net with previously proposed SOTA models. Models are retrained from the official source code provided with our designed data environment. The results are shown in Table I. From the table, we observe that our proposed model achieved the highest accuracy among small models when tested on the official V2-12, with an accuracy improvement of approximately 2% to 3% over other small models like ConvMixer. Furthermore, FCA-Net shows a noticeable reduction in the number of model parameters and MACs, indicating lower memory and computation resource requirements. Most importantly, when evaluated in noisy conditions, FCA-Net demonstrates superior performance, achieving up to 7.4% better accuracy compared to other models with similar memory footprints. The proposed model is also competitive with larger transformer-based models (KWT-3, AST-Tiny) under challenging noisy conditions, showcasing its efficiency and reliability for real-world applications.

B. Ablation study

The results indicate that incorporating efficient attention modules significantly enhances the robustness of KWS systems in noisy environments. As shown in Table II, among the different positions tested, the C2D attention module demonstrates superior performance across various noise levels. Notably, the C2D module achieves the highest accuracy when placed at all positions within the ConvMixer block, consistently outperforming the other attention methods, including ECA and SE. This setup proves to be the most effective strategy for enhancing noise robustness, making the C2D module particularly valuable for improving KWS system performance in challenging auditory conditions.

VI. CONCLUSIONS

In this paper, we proposed FCA-Net, a novel convolutional neural network designed to enhance the robustness of Keyword Spotting (KWS) systems in noisy environments while maintaining a small memory footprint. Our approach combines mixer unit-based feature interaction with an efficient two-dimensional convolution-based attention module, further strengthened by a curriculum-based multi-condition training strategy. Experimental results demonstrate that FCA-Net achieves state-of-the-art accuracy among small models and outperforms existing solutions in noisy conditions. Additionally, FCA-Net significantly reduces the number of parameters and computational resources required, making it both efficient and effective for real-world applications. Our findings highlight the potential of FCA-Net as a reliable and resource-efficient solution for robust KWS in diverse and challenging auditory environments.

References

[1] I. López-Espejo, Z.-H. Tan, J. H. L. Hansen, and J. Jensen, "Deep spoken keyword spotting: An overview," *IEEE Access*, vol. 10, pp. 4169–4199, 2022.

[2] L. Turchet, G. Fazekas, M. Lagrange, H. S. Ghadikolaei, and C. Fischione, "The internet of audio things: State of the art, vision, and challenges," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10 233–10 249, 2020.

[3] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4087–4091.

[4] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5484–5488.

[5] S. Choi, S. Seo, B. Shin, H. Byun, M. Kersner, B. Kim, D. Kim, and S. Ha, "Temporal convolution for real-time keyword spotting on mobile devices," in *Proc. Interspeech 2019*, 2019, pp. 3372–3376.

[6] Y. Lin, T. Zhou, and Y. Xiao, "Advancing airport tower command recognition: Integrating squeeze-and-excitation and broadcasted residual learning," *arXiv preprint arXiv:2406.18313*, 2024.

[7] S. Braun, D. Neil, and S.-C. Liu, "A curriculum learning method for improved noise robustness in automatic speech recognition," in *Proc. European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 548–552.

[8] T. Higuchi, S. Saxena, M. Souden, T. D. Tran, M. Delfarah, and C. Dhir, "Dynamic curriculum learning via data parameters for noise robust keyword spotting," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6848–6852.

[9] D. Ng, Y. Chen, B. Tian, Q. Fu, and E. S. Chng, "Convmixer: Feature interactive convolution with curriculum learning for small footprint and noisy far-field keyword spotting," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3603–3607.

[10] D. Ng, Y. Xiao, J. Q. Yip, Z. Yang, B. Tian, Q. Fu, E. S. Chng, and B. Ma, "Small footprint multi-channel network for keyword spotting with centroid based awareness," in *Proc. Interspeech*, 2023, pp. 296–300.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[12] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 534–11 542.

[13] Y. Xiao, N. Hou, and E. S. Chng, "Rainbow Keywords: Efficient Incremental Learning for Online Spoken Keyword Spotting," in *Proc. Interspeech*, 2022, pp. 3764–3768.

[14] B. Kim, S. Chang, J. Lee, and D. Sung, "Broadcasted Residual Learning for Efficient Keyword Spotting," in *Proc. Interspeech 2021*, 2021, pp. 4538–4542.

[15] V. A. Trinh, H. S. Kavaki, and M. I. Mandel, "Importantaug: A data augmentation agent for speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8592–8596.

[16] Y. Xiao and R. K. Das, "Dual knowledge distillation for efficient sound event detection," *arXiv preprint:2402.02781*, 2024.

[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[18] O. Rybakov, N. Kononenko, N. Subrahmanya, M. Visontai, and S. Laurenzo, "Streaming keyword spotting on mobile devices," *Proc.Interspeech*, 2020.

[19] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword transformer: A self-attention model for keyword spotting," in *Proc.Interspeech*, 2021, pp. 4249–4253.

[20] S. Majumdar and B. Ginsburg, "Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition," in *Proc.Interspeech*, 2020.

[21] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," in *Proc.Interspeech*, 2021, pp. 571–575.

[22] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[23] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[24] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc.International Conference on Learning Representations*, 2018.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc.Interspeech*, p. 2613, 2019.

A Diffusion-Based Approach for Restoring Face-swapped Images

Yuanchen Niu, Yuanman Li*, Guijia Zhang, and Xia Li
University of Shenzhen, Shenzhen
E-mail: yuanmanli@szu.edu.cn

***Abstract*—The misuse of deepfake technology underscores the critical importance of deepfake forensics. Current research in this domain primarily employs binary classification models that are limited to detecting the authenticity of facial images, and are incapable of tracing the original face replaced in the face-swapping process. Consequently, this limitation results in the inability to provide substantial judicial evidence.To address this issue, this paper introduces a novel forensic methodology, which aims to reconstruct the original face from the forged face. Specifically, we propose a novel framework that reformulates the restoration process from the forged face to the original face as a special image editing task guided by a series of conditions. Our research demonstrates that forged face images after face-swapping still retain some features of the original face, and these features can be used to restore the original face.Furthermore, we also construct the first dataset and set up evaluation metrics for this important research problem. Extensive experiments demonstrate the qualitative and quantitative effectiveness of our proposed framework.**

I. Introduction

Recent research primarily focuses on deepfake detection technologies to defense Face-swapping Deepfake [1]–[4]. These techniques identify manipulation artifacts across spatial, frequency, and temporal domains. Deep Neural Networks (DNNs) are then used to predict binary labels (Real or Fake) to discern authenticity [5]–[7]. However, such binary classifications do not yield sufficient persuasive judicial evidence. Moreover, the original identity of the target face, altered in creating fake news, remains obscured and the truth may never be revealed. This issue is particularly prevalent in conflicts such as Russia-Ukraine and Israel-Palestine. To overcome these challenges, we introduce a novel research problem, which seeks to restore the *target face* from the *result face* in the process of face-swapping, offering a more robust forensic solution for addressing the misuse of Face-swapping Deepfake technologies. It's noteworthy that the *result face* in this context refers to the face post-forgery, while the *target face* contributes its attributes and the *source face* provides its identity. We will explore the feasibility of Face-Swapped Image Restoration and introduce a powerful framework to tackle this issue.

We posit that face-swapping deepfake is traceable for several reasons. Firstly, although deepfake algorithms are capable of creating high-quality fake faces, they inevitably left traces on the result face. We believe that target-face-relevant information can be captured from these traces and used for the restoration. Secondly, it is difficult for these algorithms to effectively decouple identity and attribute information in facial images, leading to implicit tampering traces on the synthesized face. Such traces likely retain features associated with the target face, which are beneficial for restoration. Consequently, in restoration of face-swapped images, the critical step involves independently decoupling the explicit and implicit target-relevant features within the result facial image, and then extract useful information to facilitate the restoration of the target face. Motivated by these challenges, we reformulate the restoration process of the result facial image to the target facial image as a special image-to-image editing task.

In this paper, we propose a novel conditional diffusion-based framework named DRFSI (Diffusion Restoration of Face-Swapped Image). DRFSI is able to progressively decouple traceable information related to the target face throughout the denoising process for restoration. Specifically, due to the attribute information on the result face is highly relevant with the target face and should be preserved. We use the multi-scaled attribute information of the result face, compressed by information bottleneck modules, in facilitating the restoration of the target face via the AttributeNet we designed. The AttributeNet employs control mechanisms [8] and serves as a positive guidance to assist the denoising process. These enable our model to adaptively disentangle target-face-relevant information from the result face and accomplish restoration of the target face.

Our main contributions are summarized as follows:

- We introduce a novel research problem, which aims to restore the *target face* from the *result face* in face-swapping deepfake procedure. This approach offers a new forensic method for protecting facial information.
- We propose a novel diffusion-based framework DRFSI to address the challenge of face-swapped image restoration. By designing AttributeNet, we effectively instruct our diffusion model in disentangling target-relevant information, facilitating the restoration of the concealed target face.
- Extensive qualitative and quantitative experiments demonstrate the effectiveness of our method. Moreover, our method can be generalized to unknown face-swapping deepfake algorithms effectively.

II. Related Works

Face-swapping deepfakes employ deep learning techniques to execute the face-swapping task. Current methodologies in

*Corresponding author

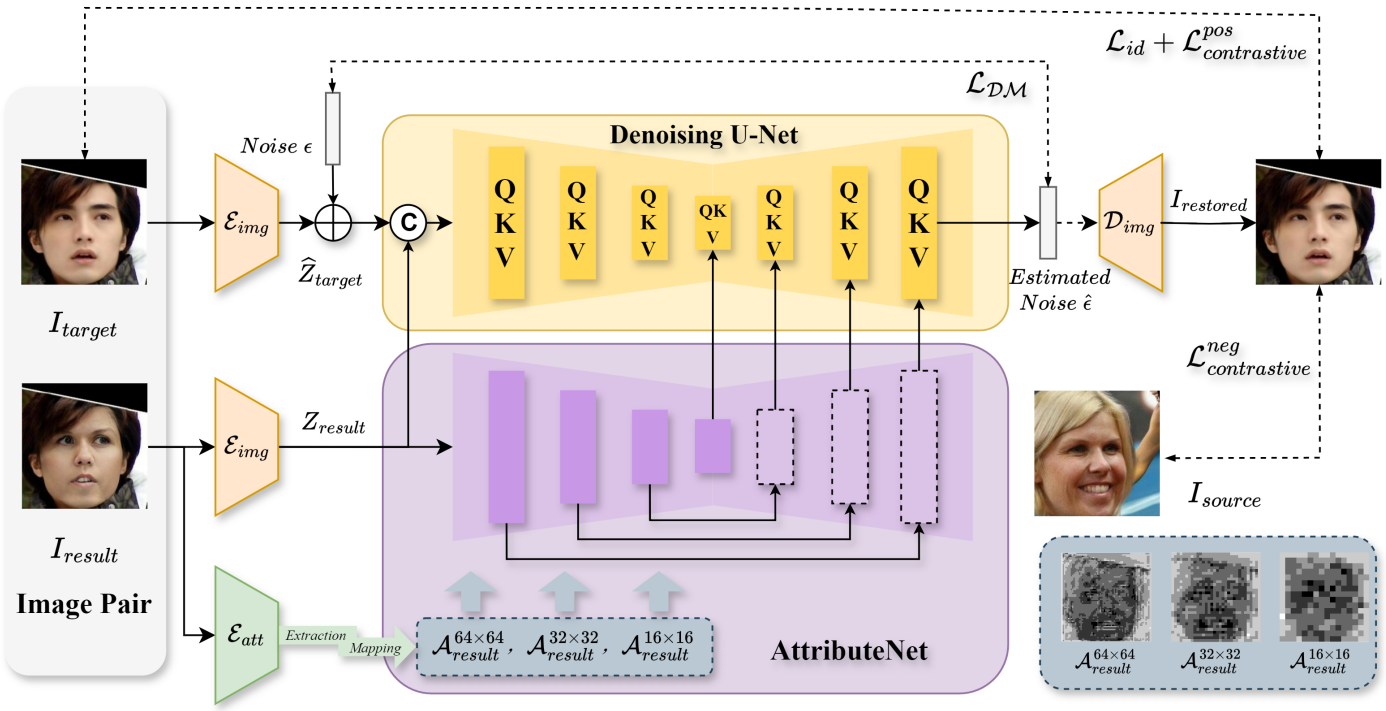


Fig. 1. **The overall pipeline of our proposed DRFSI**. For a given facial image pair comprising the target and result, we first encode target face into the latent space and generate a noisy version of latent target facial image. Subsequently, we employ a conditional diffusion model to denoise the noisy latent target facial image, using resulting facial image as conditional inputs of the AttributeNet.

face-swapping deepfakes can be categorized into two types: source identity-guided and target attribute-guided methods. Source identity-guided face swapping involves extracting identity features from the source face and integrating them into the target face. For instance, FaceShifter [9], SimSwap [10] and MobileFaceSwap [11] incorporate a source identity embedding module within the generator to preserve the source face's identity. As for target attribute-guided face swapping, it involves modifying the source face using attributes derived from the target face. This process initiates with the source face, which is then adjusted to align with the target face's attributes. For instance, FSGAN [12] applies the target face's features to direct the modifications of the source face, incorporating a hybrid network to seamlessly integrate the facial and background areas.

III. Proposed Method

A. Proposed Framework: DRFSI

As illustrated in Fig 1, DRFSI comprises two components, including Denoising Network and AttributeNet.

1) Denoising Network: DRFSI's backbone denoising network is based on the 64×64 conditional UNet model with temporal layers and inherits weights from the original SD UNet. As the input of the UNet, we follow the paradigm of diffusion-based image editing. First, we encode both the target and resulting facial images into latent space using a pre-trained autoencoder $\mathcal{E}_{img}$ to obtain Z_{result} and Z_{target}. Subsequently, a noisy version of the latent target facial image, $\widehat{Z}_{target}$, is generated as follows:

$$\begin{aligned} \widehat{Z}_{target} &= \sqrt{\overline{\alpha}_t} Z_{target} + \sqrt{1-\overline{\alpha}_t}\epsilon, \\ \overline{\alpha}_t &= \prod_{i=1}^{t}(1-\beta_i), \epsilon \sim \mathcal{N}(0, I), \end{aligned} \tag{1}$$

where β_i is derived from a predefined variance schedule β which governs the amount of noise injected at various time steps. Finally, we concatenate $\widehat{Z}_{target}$ with the latent resulting facial image Z_{result} and this concatenated feature map serves as the input for the denoising UNet.

2) Attribute Extraction: The objective of Face-swapping Deepfake is to transfer the identity of the source face onto the target face while preserving the attributes of the target face such as pose, expressions, and background. This means the result and target facial attribute information are highly relevant in Face-swapping Deepfake. Thus, we propose to utilizes the result-facial attribute information to aid in the restoration of target face.

The Information Bottleneck (IB) Principle has been effectively utilized to decouple the identity and attribute information in facial images [13]. We first average values of the same size feature maps in pre-trained IB model to obtain multi-scale attribute feature maps from result face:

$$\mathcal{A} = \left\{\mathcal{A}^{result}_{64\times64}, \mathcal{A}^{result}_{32\times32}, \mathcal{A}^{result}_{16\times16}\right\}. \tag{2}$$

Thus, when applying $\mathcal{A}$ to an encoded resulting facial image $\mathcal{Z}$, the multi-scaled facial attribute information $\mathcal{F}$ is highlighted

as follows:

$$\mathcal{F} = \mathcal{A} \odot \mathcal{Z}. \tag{3}$$

Subsequently, the highlighted attribute features $\mathcal{F}$ are integrated into the AttributeNet to facilitate the denoising restoration process via the control mechanism [8].

3) AttributeNet: As shown in Fig 1 ,the AttributeNet shares a similar structure with the encoder in the Denoising Network and serves to extract multi-scaled features from the result image. Unlike the backbone, however, AttributeNet does not employ prompt embeddings and time embeddings. Consequently, we have replaced the cross-attention layers with self-attention layers and removed the time-embedding layers. Subsequently, multi-scaled attribute feature maps $\mathcal{A}$ are extracted from the result face using a pre-trained Attribute Extractor[13]. Then, these maps are multiplied by the first feature map following each down-sampling operation within the encoder of AttributeNet. This method enables the extraction and retention of multi-scaled attribute information from the result face. Finally, the multi-scaled feature maps obtained from AttributeNet guide the denoising process in a manner analogous to that of ControlNet [8].

B. Loss Function

1) Diffusion Loss: The training objective of our diffusion-based model is similar to the conditional LDM.The Diffusion Loss can be defined as:

$$\mathcal{L}_D = E_{z_t,t,c,\epsilon}[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_l^l], \tag{4}$$

where $l = 1, 2$, z_t represents the noisy version of latent target facial image, c represents the condition, ϵ represents the added noise. As the diffusion loss is actually equivalent to the reconstructive loss, we make $l = 1$ to improve visual restoration quality which is different from the common diffusion loss.

2) Identity Loss: In the restoration of face-swapped image, ensuring identity coherence between the restored and original target faces is crucial. Therefore, we use identity loss to constrain the distance in feature space between the restored target face $\hat{I}_{target}$ and the original target face I_{tar}. We compute the identity loss as:

$$\mathcal{L}_{id} = 1 - cos\left\langle \mathcal{E}_{id}(I_{target}), \mathcal{E}_{id}(\hat{I}_{target}) \right\rangle, \tag{5}$$

where $\mathcal{E}_{id}$ represents a pre-trained face recognition model, $cos\left\langle \mathcal{E}_{id}(I_{target}), \mathcal{E}_{id}(\hat{I}_{tar}) \right\rangle$ represents the cosine similarity between the original target identity and the restored target identity, $\mathcal{D}$ is the decoder of the pre-trained autoencoder.

3) Contrastive Identity Loss: We designe Contrastive Identity Loss (CIL) based on the contrastive learning principles. Here we want to ensure that the anchor sample $\mathcal{E}_{id}(\hat{I}_{target})$ is closer to the positive sample $\mathcal{E}_{id}(I_{target})$ while being distant from the negative sample $\mathcal{E}_{id}(I_{source})$. Then, we can formulate our loss function in the form of triplet loss [14] as follows:

$$\begin{aligned}\mathcal{L}_{CIL} =& 2 - \max\Big\{\big(\cos\langle \mathcal{E}_{id}(\hat{I}_{target}), \mathcal{E}_{id}(I_{target})\rangle \\ & - \cos\langle \mathcal{E}_{id}(\hat{I}_{target}), \mathcal{E}_{id}(I_{source})\rangle - \lambda\big), 0\Big\},\end{aligned} \tag{6}$$

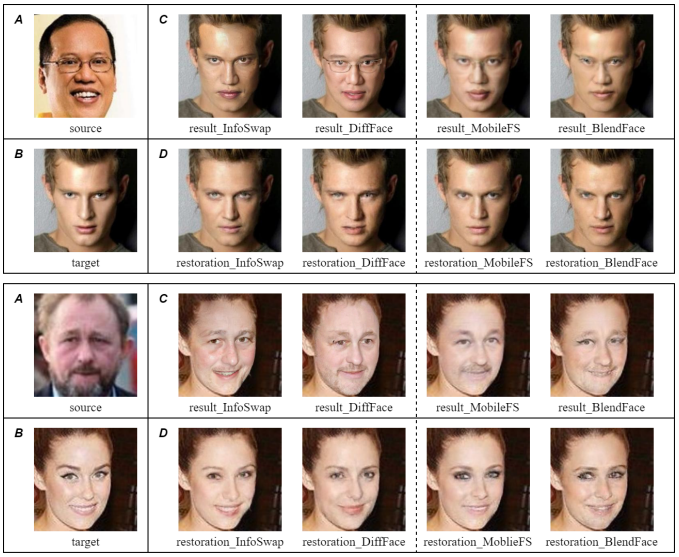


Fig. 2. **Qualitative results**. In the table above, Area ***A*** shows the source face. Area ***B*** shows the target face. Area ***C*** shows the result faces, generated by swapping the source face from ***A*** to the target face from ***B***. Correspondingly, Area ***D*** shows the target face restored by DRFSI. Specifically, the face-swapping algorithms on the left side of the dotted line were utilized for training, while those on the right side were only employed for test, indicating that our method generalizes well to unseen face-swapping deepfake algorithms.

where $\cos\langle \mathcal{E}_{id}(\hat{I}_{target}), \mathcal{E}_{id}(I_{target})\rangle$ is the cosine similarity of positive pair, and $\cos\langle \mathcal{E}_{id}(\hat{I}_{target}), \mathcal{E}_{id}(I_{source})\rangle$ is the cosine similarity of negative pair, λ is the enforced margin between positive and negative pairs.

In summary, the overall loss function for training DFDT can be written as:

$$\mathcal{L} = \mathcal{L}_D + \beta_1 \mathcal{L}_{id} + \beta_2 \mathcal{L}_{CIL}, \tag{7}$$

where β_1, β_2 are hyper parameters. We experimentally set them all to 10.

IV. Experiments

A. Experimental Settings

1) Implementation Details: In our experiments, the input images are uniformly resized to a size of 256×256 and then encoded into a $4 \times 64 \times 64$ latent space using a pre-trained autoencoder to improve the computational efficiency of the diffusion process. We train our diffusion model with batch size of 24 on 3 NVIDIA RTX4090 GPUs. We use Adam optimizer with the base learning rate of 1e-5 and the linear scaling rule.

2) Datasets: We constructed an extensive dataset based on the VGGFace2 [15], comprising 9,000 different identities, totally 100,000 source-target-result facial image pairs incorporating 4 diverse Face-swapping Deepfake algorithms (InfoSwap[13], DiffFace[16], BlendFace[17] and MobileFS[11]). In our experiment, we use two algorithms for train (InfoSwap and DiffFace), and two algorithms for generalizability test (BlendFace and MobileFS).

TABLE I
RETORATION ACCURACY (R-ACC / %) AND VERIFICATION ACCURACY (V-ACC / %) OF RESULT PAIR AND RESTORATION PAIR ON OUR TEST SET.

Methods	Result Pair		Restoration Pair	
	R-Acc	V-Acc	R-Acc	V-Acc
InfoSwap [13]	4.2	18.8	**92.8**	**84.9**
DiffFace [16]	7.2	28.8	**89.7**	**82.3**
MobileFS (test only) [11]	10.8	31.5	**88.0**	**81.4**
BlendFace (test only) [17]	7.6	35.3	**82.4**	**76.4**
Mean	7.5	28.6	**88.2**	**81.3**

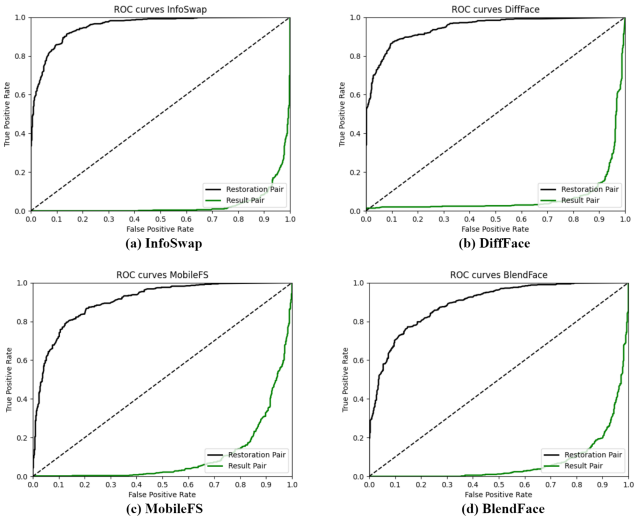


Fig. 3. The ROC curves of Result Pairs and Restoration Pairs on our test set.

B. Qualitative Results

In this section, we present the visual results of our DRFSI. As shown in Fig 2, we randomly select two cases from the test set, which two of them (left of the dotted line) used for train and another two (right of the dotted line) only used for *generalization test*. In the first case, we can find that the restored faces in area ***D*** have a certain traceability effect on the identity information of the target faces, such as eye, mouse and nose feature information, etc. At the same time, the attribute information of the target face, such as expression and demeanor, is retained. In addition, for the example where the result faces in area ***C*** have glasses (result_diffface, result_mobilefs), the target faces restored by our DTFT also correctly delete the glasses (recover_diffface, recover_mobilefs). In the second case, it's obvious that our traceability algorithm has effectively restored identity information such as the eye and mouth parts of the target face, as well as gender information.

C. Quantitative Results

1) Metrics: Given a source face, a target face and a restored face, the identity similarity between the restored face and the target face must exceed the similarity between the restored face and the source face. Then we can calculate the restoration accuracy (***R-Acc***) as follows:

$$Restoration\ Accuracy = \frac{pos}{N}, \tag{8}$$

where pos represents the number of samples that satisfy $sim\left\langle Id^{restore}, Id^{target}\right\rangle > sim\left\langle Id^{restore}, Id^{source}\right\rangle$, cosine similarity is used to evaluate the $sim\left\langle \cdot, \cdot \right\rangle$. N represents the total number of test samples.

Moreover, we evaluate our algorithm via *face verification*, which refers to assess whether two randomly selected facial images represent the same individual. For the given M sets of test facial image pairs, which include both positive and negative sample pairs, predicted labels are obtained by determining whether the identity similarity of each pair exceeds a specified threshold. Subsequently, face verification accuracy is calculated using the method for standard binary classification metrics.

2) Comparison to Result Pair: In this section, we quantitatively evaluate our method using the metrics introduced above. The comparison in Table I provides an intuitive demonstration of our DRFSI effectiveness. We define two sets of facial image pairs: the *Result Pair* consisting of source-target-result combinations, and the *Restoration Pair* consisting of source-target-restoration combinations. In Restoration Pair, the restoration and target images are considered positive sample pairs, whereas the result and source images are negative sample pairs. Conversely, in Result Pair, the result and target images are positive sample pairs, and the result and source images are negative sample pairs. Since the result and source images in the Result Pair share the same identity, similarly for the restoration and target images in the Restoration Pair, we expect much significantly lower Restoration Accuracy (R-Acc) and Verification Accuracy (V-Acc) in the Result Pair compared to the Restoration Pair. And as shown in Fig 3, the ROC curve for the Result Pair should approach the lower right corner, whereas the ROC curve for the Restoration Pair should align with the norm, approaching towards the upper left corner.

TABLE II
ABLATION EXPERIMENTS DEMONSTRATING THE EFFECT OF ATTRIBUTENET

Methods	Base		DRFSI	
	R-Acc	V-Acc	R-Acc	V-Acc
InfoSwap [13]	86.2	79.5	**92.8**	**84.9**
DiffFace [16]	86.3	78.8	**89.7**	**82.3**
MobileFS (test only) [11]	85.6	80.3	**88.0**	**81.4**
BlendFace (test only) [17]	81.2	75.0	**82.4**	**76.4**
Mean	84.8	78.4	**88.2**	**81.3**

D. Ablation Experiments

In this subsection, we conduct ablation experiments to evaluate the effectiveness of AttributeNet. We classify our framework into two configurations based on the inclusion or exclusion of the AttributeNet (AN): (1) **Base**, an unconditional diffusion model utilizing only a denoising network; (2) **DRFSI**, which combines AttributeNet with the Base model.

The experimental results presented in Table II demonstrate that the effectiveness of AN. Specifically, while an unconditional diffusion model can initially restore the target face, it falls in accurately tracing the target identity. AttributeNet leverages decoupled facial attribute maps to improve the restoration by ensuring the retention of critical target attributes through its effective spatial control. Integrating AN leads to an average improvement of 3.4% in R-Acc and 2.9% in V-Acc over Base.

V. Conclusions

In this paper, we study an important research problem, which aims to restore the original target face from a face-swapped deepfake image. Our research demonstrates that the synthetic face still retains traceable features of the original face, enabling its approximation and restoration. To address this problem, We have developed a powerful diffusion-based framework, named DRFSI. Extensive testing confirms the robust effectiveness and generalization capabilities of our framework. We hope that our efforts will inspire further research into the protection of facial privacy.

Acknowledgment

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010645; in part by the Key project of Shenzhen Science and Technology Plan under Grant 20220810180617001; in part by the Open Research Project Programme of the State Key Laboratory of Internet of Things for Smart City (University of Macau) under Grant SKLIoTSC(UM)-2021-2023/ORP/GA04/2022.

References

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[2] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[3] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6458–6467.

[4] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 044–15 054.

[5] Y. Li, L. Ye, H. Cao, W. Wang, and Z. Hua, "Cascaded adaptive graph representation learning for image copy-move forgery detection," *ACM Transactions on Multimedia Computing, Communications and Applications*,

[6] Y. Li, Y. He, C. Chen, *et al.*, "Image copy-move forgery detection via deep patchmatch and pairwise ranking learning," *IEEE Transactions on Image Processing*, pp. 1–16, 2024.

[7] Y. Tan, Y. Li, L. Zeng, J. Ye, W. Wang, and X. Li, "Multi-scale target-aware framework for constrained splicing detection and localization," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8790–8798.

[8] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[9] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5074–5083.

[10] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 2003–2011.

[11] Z. Xu, Z. Hong, C. Ding, *et al.*, "Mobilefaceswap: A lightweight framework for video face swapping," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 2973–2981.

[12] Y. Nirkin, Y. Keller, and T. Hassner, "Fsganv2: Improved subject agnostic face swapping and reenactment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 560–575, 2022.

[13] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, "Information bottleneck disentanglement for identity swapping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3404–3413.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[15] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *Proceedings of theIEEE international conference on automatic face & gesture recognition*, 2018, pp. 67–74.

[16] K. Kim, Y. Kim, S. Cho, *et al.*, "Diffface: Diffusion-based face swapping with facial guidance," *arXiv preprint arXiv:2212.13344*, 2022.

[17] K. Shiohara, X. Yang, and T. Taketomi, "Blendface: Re-designing identity encoders for face-swapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7634–7644.

A SEMI-SUPERVISED LOW-LIGHT IMAGE ENHANCEMENT WITH COLOR GUIDANCE

Yuxin Wang[1], Shuolin Yang[1], Qianxi Wu[1], Zhishuo Zhang[1], Yunxia Liu[1,2,*], Yang Yang[3], Yakui Dong[1,2], Cheng Fei[1,2], Junliang Liu[1,2], Lili Wang[1,2], Shuzhen Fan[1,2] and Yongfu Li[1,2]
[1]Center for Optics Research and Engineering (CORE), Shandong University, Qingdao, China
[2]Key Laboratory of Laser & Infrared system, Ministry of Education, Shandong University, Qingdao, China
[3]School of Information Science and Technology, Shandong University, Qingdao, China
[*]Corresponding Author
E-mail: yuxin_wang@mail.sdu.edu.cn, eyxliu@sdu.edu.cn

***Abstract*—Low-light image enhancement (LLIE) has found extensive applications in surveillance and related fields. Current deep learning-based algorithms have demonstrated significant advantages over traditional methods for LLIE. However, the issue of color distortion still remains a challenge. This paper proposes a semi-supervised curve estimation and restoration network that consists of two cascaded sub-networks with distinct functionalities. The first-level sub-network applies specific constraints to enhance different regions based on their brightness levels. The second-level sub-network is trained using superpixel color consistency loss and global histogram loss to effectively correct color distortion. Experimental results demonstrate that the proposed method achieves satisfactory outcomes in both visual quality and evaluation metrics.**

I. INTRODUCTION

Low-light image enhancement (LLIE) is one of the most important computer vision tasks. Traditional methods for LLIE typically categorized into two groups: histogram equalization-based methods [1][2][3], which aim to adjust the dynamic range of the image and improve contrast; and Retinex-based methods [4][5][6][7], which utilize various techniques to estimate the illumination component and derive the reflection component encapsulating pertinent information for image restoration.

Deep learning algorithms for LLIE can be broadly categorized into two types: supervised methods and unsupervised methods. The supervised methods rely on pairs of images for training the network. Wei et al. [8] addressed the decomposition, adjustment, and reconstruction of low-light images in a framework. Chen et al. [9] developed an end-to-end network, which operates on raw sensor data to simulate traditional image processing pipelines. However, a challenge with these methods is their heavy reliance on paired images, which can be difficult to obtain in large quantities. Semi-supervised and unsupervised methods have gradually gained prominence. Jiang et al. [10] introduced a generative adversarial network, utilizing global-local discriminators to guide the learning of global and local information. Li et al. proposed Zero-Reference Deep Curve Estimation (Zero-DCE) [11] and its accelerated and lightweight version, Zero-DCE++ [12]. They designed a curve capable of mapping low-light images to normal-light images. However, without guidance from paired images, these methods often exhibit color distortion and may amplify noise in certain scenarios.

There have been some efforts in reducing the problem of color distortion in LLIE. Zhang et al. [13][14] guided a network in recovering the reflective image by leveraging illumination information. Zhang et al. [15] introduced a pyramid color embedding module to fuse color information, thereby reducing color distortion. Fu et al. [16] devised a global-illumination attention branch for better color estimation and noise suppression. While these techniques succeed in mitigating color distortion to some extent during the enhancement process, they do not, however, engage in detailed color restoration for some regions.

Considering the fact that reference images are easy to obtain in surveillance applications, a semi-supervised LLIE algorithm is proposed for improved color restoration. To reduce the learning complexity, the overall architecture is consisted of two cascaded stages. The first stage is a curve estimation network inspired by Zero-DCE++ [12], with special attention paid to adaptive enhancement in varying brightness regions. In the second stage, a color restoration network based on the UNet [17] structure, guided by global and local color information to improve color fidelity.

The main contributions of our paper are summarized as follows:

- A two-stage semi-supervised low-light image enhancement algorithm is proposed for improved color restoration fidelity in surveillance application.
- Adaptive spatial consistency loss and graded exposure control loss are designed to enhance the dynamic range of extremely dark regions, facilitating varied exposure levels for different brightness areas.
- histogram calculated based on the reference image is employed to provide global color supervision. Simultaneously, we employ superpixel-based color

consistency loss to ensure locally supervised color alignment with semantic content.

- We conducted extensive tests on varying datasets, and experimental results indicate that proposed method significantly improves enhancing low-light images.

II. Related Work

A. Deep Learning Methods for Color Distortion Reduction

Currently, there have been many exploring efforts in mitigating the color distortion problem. Due to the strong dependence of the degradation distribution in the reflection components on the illumination components, Zhang et al. proposed Kinds [13] and Kinds++ [14] to recover the reflection component by using the illumination components. Zhang et al. [15] designed a pyramid color embedding module for matching color and content features. Fu et al. [16] employed an illumination-aware attention module to selectively emphasize semantic information. These algorithms effectively restore colors. However, they may encounter issues such as artifacts and ineffective color recovery in certain scenes. In response to this phenomenon, we intend to treat color restoration as a subsequent task after image enhancement.

B. Zero-DCE++

Zero-DCE++ [12] formulates light enhancement as estimating pixel-wise and high-order curves for dynamic range adjustment of a given image. For each pixel coordinate x, the image enhancement process is defined as:

$$LE_n(x) = LE_{n-1}(x) + A(x)LE_{n-1}(x)(1 - LE_{n-1}(x)), \quad (1)$$

where LE_n denotes the nth iteration enhanced image with LE_0 corresponds to the initial low-light image I, and $A(x)$ is a parameter map that controls the pixel-level magnitude adjustment that is related with exposure level.

However, the dynamic range of the enhanced image is limited, and color distortion may occur in some scenes. Based on this, the paper proposes a series of targeted loss functions.

III. The Proposed Method

In this section, we provide a comprehensive introduction of the proposed method, with the objective of achieving well-exposed images while preserving natural colors. As illustrated in Fig. 1, the framework is consisted of two cascaded networks.

A. Curve Estimation Network

Curve estimation network adopts DCE-Net [12] to estimate the light enhancement curve for low-light images. To achieve curve estimation and enhance regions of varying brightness to different extents, the following loss functions are employed to train the curve estimation network.

1) Adaptive spatial consistency loss: To maintain regional consistency during LLIE and prevent neighboring regions of the enhanced image in different brightness levels from being enhanced indiscriminately, an adaptive spatial consistency loss is designed as:

$$L_{spa} = \frac{\frac{1}{K}\sum_{i=1}^{K}\sum_{j\in\Omega(i)}(|Y_{CEi}-Y_{CEj}|-|I_i-I_j|)^2}{(1-\alpha I_i)+\epsilon}, \quad (2)$$

where K is the number of local regions with the size of 4×4, and $\Omega(i)$ is the four neighboring regions (top, down, left, right) around the region i. Y_{CE} and I respectively represent the average intensity values of local regions in the enhanced image processed by the curve restoration network and the input image. ϵ is a very small value to prevent the denominator from being zero. We empirically set α to 0.5 in our experiments.

For dark regions, the enhancement level of neighboring regions is greater than that of neighboring regions in bright regions. This enables darker regions in the enhanced image to have a larger dynamic range compared to brighter regions.

2) Graded exposure control loss: To control the exposure level and prevent overexposure or underexposure, a graded exposure control loss is designed as:

$$L_{exp} = \frac{1}{M}\sum_{k=1}^{M}|Y_{CEk} - E_i|, \quad (3)$$

where M represents the number of local regions with the size of 16×16, Y_{CE} denotes the average intensity values of local regions in the enhanced image processed by the curve restoration network. E_i represents the multi-level exposure levels, which in this paper are three levels: 0.55, 0.60, and 0.65.

By sorting the average intensity values of local regions in low-light images from low to high, they are divided into three levels, corresponding to three exposure levels. Through graded exposure control loss, it is possible not only to control the exposure levels but also to set varying exposure levels for regions with varying brightness.

3) Color constancy loss: According to [18], the values of each color channel in the image should not significantly exceed those of other channels. The color constancy loss [12] is formulated as follows:

$$L_{col} = \sum_{\forall(p,q)\in\varepsilon}(J^p - J^q)^2, \varepsilon = \{(R,G),(R,B),(G,B)\}, \quad (4)$$

where J^p denotes the average intensity value of p channel in the enhanced image, a pair of channels is represented as (p, q).

4) Illumination smoothness loss: In order to maintain the monotonicity relations between neighboring pixels, the illumination smoothness loss [12] is applied to the parameter mapping A.

$$L_{tvA} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c\in\xi}(|\nabla_x A_n^c| + |\nabla_y A_n^c|)^2, \xi = \{R,G,B\}, \quad (5)$$

where N is the number of iteration, the horizontal and vertical gradient operations are represented as ∇_x and ∇_y, respectively.

The total loss function of the curve estimation network is shown as follows:

$$L_{total1} = L_{spa} + \alpha L_{exp} + \beta L_{col} + \gamma L_{tvA}, \quad (6)$$

where α, β, and γ are used to balance the proportions of different loss functions.

B. Color Restoration Network

The curve estimation network effectively enhances the brightness of low-light images but may generate color

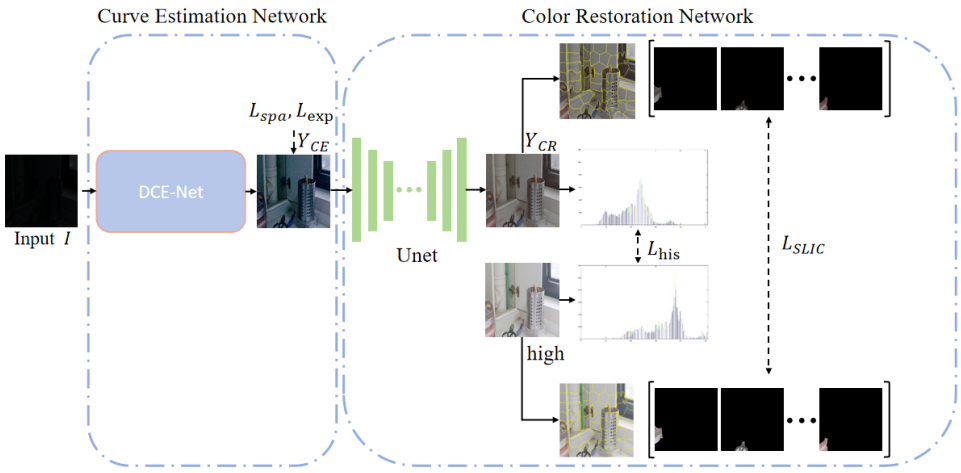


Fig. 1 The framework of the curve estimation and color restoration network.

distortion and amplify noise in certain scenes. The problem originates from the inherent inability to involve color information for unsupervised methods. However, in some applications, obtaining reference images, for example in the monitoring field, is convenient or desired.

To mitigate color distortion and noise, this paper proposes a color restoration network. The color restoration network adopts the UNet [17]. Given a low-light image I, the curve estimation network produces an enhanced image Y_{CE} for color restoration. Y_{CE} is paired with the corresponding reference image to form paired images used for training the color restoration network. The Y_{CR} obtained through the color restoration network is the final enhanced image.

The following loss functions are employed to train the color restoration network.

1) Global histogram loss: To ensure the final enhanced image closely matches the color of the reference image, a global histogram loss is designed.

$$L_{his} = \sum_{c\in\xi} |f_{his}(Y_{CR}^c) - f_{his}(H^c)|, \xi = \{R, G, B\}, \quad (7)$$

where f_{his} denotes the histogram of corresponding color channel, Y_{CR} and H respectively represent the final color restored image and the reference image.

2) VGG perceptual loss: To ensure consistency in human visual perception between enhanced image and reference image, constraints need to be applied at high-dimensional feature levels. Pretrained VGG network's intermediate features are commonly utilized to approximate the deep-level features between the network-generated image and the reference image.

$$L_{VGG} = \frac{1}{N}\sum_{n=1}^{N} |F_n(Y_{CR}) - F_n(H)|, \quad (8)$$

where F_n is the feature representation of the pre-trained VGG-19 network [19] at the nth layer. Y_{CR} and H respectively denote the final enhanced image and the reference image.

3)Superpixel color consistency loss: Although histograms provide global statistical information, it is also of vital importance to involve local guidance information in the color restoration process. Considering the fact that color information is more closely related with the content of an image, conventional loss calculation methods based on rectangular bounding boxes [10] generally cannot capture finer details of content-related guidance information. To this end, a superpixel color consistency loss is designed to address local color restoration. This paper obtains local regions through superpixel segmentation (see Fig. 2), where local regions are composed of neighboring pixels with similar features such as color, brightness, and texture. The Simple Linear Iterative Clustering (SLIC) [20] is employed for superpixel segmentation in this study. Compared to other superpixel segmentation methods, SLIC demonstrates favorable performance in terms of computational efficiency and preservation of superpixel boundaries.

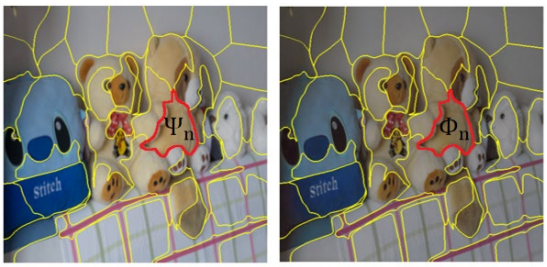


Fig. 2 SLIC superpixel segmentation schematic. From left to right, (a) reference image. (b) the final enhanced image.

For a given reference image H, denote Ψ ($\Psi = \{\Psi_1, \Psi_2, \ldots, \Psi_N\}$) as a series of N superpixel sets obtained through F_{SLIC} segmentation on H, the superpixel color consistency loss is defined as follows:

$$L_{SLIC} = \frac{1}{K}\sum_{n=1}^{N}\sum_{c\in\xi}\sum_{k=1}^{K}\left(\Phi_n^c(k) - \Psi_n^c(k)\right)^2, \quad \xi = \{R, G, B\}, \quad (9)$$

where Φ ($\Phi = \{\Phi_1, \Phi_2, \ldots, \Phi_N\}$) represents a series of superpixel sets in the final enhanced image, which are spatially aligned with Ψ, and K represents the number of pixels within each superpixel.

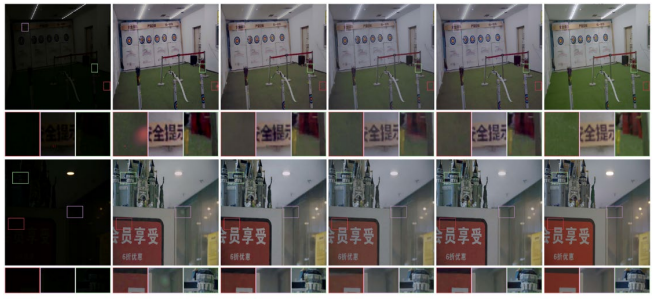
The overall loss function of the color restoration network is as follows:

Table 1. The quantitative results of the ablation study on L_{spa} and L_{exp}.

L_{exp}	L_{spa}	LOL-V2		SICE	
		PSNR	SSIM	PSNR	SSIM
		21.020	0.847	18.999	0.818
√		21.019	0.847	19.105	0.820
	√	21.017	0.846	19.139	0.821
√	√	21.077	0.847	19.320	0.823

Table 2. The quantitative results of the ablation study on L_{his} and L_{SLIC}.

L_{his}	L_{SLIC}	LOL-V2		SICE	
		PSNR	SSIM	PSNR	SSIM
		20.739	0.851	19.044	0.829
√		21.047	0.852	18.944	0.823
	√	20.212	0.827	19.272	0.824
√	√	21.077	0.847	19.320	0.823



(a) (b) (c) (d) (e) (f)

Fig. 3 Regarding the visual quality comparison of the ablation study on L_{his} and L_{SLIC}. From left to right, (a) input. (b) w/o L_{his}. (c) w/o L_{SLIC}. (d) w/o L_{his} and w/o L_{SLIC}. (e) proposed. (f) GT.

$$L_{total2} = L_{his} + \varphi L_{VGG} + \omega L_{SLIC}, \quad (10)$$

where φ and ω are used to balance the proportions of different loss functions.

IV. Experiments

A. Experimental Settings

We selected 739 images from the LOL-V2 dataset [21] and 360 images from the Part1 of SICE dataset [22], constructing a training dataset. In our effort to thoroughly examine the effectiveness of different methods in enhancing low-light images, including improving brightness, restoring color, and reducing noise, we randomly selected a subset of more challenging images, which includes low-light images with severe color and detail degradation, as well as ordinary low-light images, to constitute our test dataset. The test dataset comprises 50 pairs of images from the LOL-V2 dataset and 100 pairs of images from the second part of the SICE dataset. In this study, both the training and test images are resized to a resolution of $512 \times 512 \times 3$.

Low-light images from the LOL-V2 and SICE training sets are employed for the training of the curve estimation network. The color restoration network is trained using the paired LOL-V2 training sets. The proposal algorithm is trained on an NVIDIA GeForce GTX 1080 Ti GPU.

Table 3. Quantitative comparison of PSNR and SSIM on the LOL-V2 and SICE datasets.

Method	LOL		SICE	
	PSNR	SSIM	PSNR	SSIM
LIME[7]	11.760	0.456	14.516	0.713
RetinexNet[8]	15.731	0.524	18.410	0.736
EnlightenGAN[10]	14.993	0.616	16.176	0.720
Zero-DCE++[12]	14.678	0.519	17.174	0.753
KinD[13]	18.271	0.741	19.490	0.756
LE-GAN[16]	19.854	0.747	17.904	0.740
proposed	21.077	0.847	19.320	0.823

B. Ablation Study

In order to investigate the effectiveness of each loss function in the curve estimation and color restoration network, we conducted ablation studies.

The first series of experiments compares the effectiveness of the adaptive spatial consistency loss and the graded exposure control loss in the curve estimation network. Quantitative results of the ablation study on L_{spa} and L_{exp} are shown in Table 1, where not using L_{spa} and L_{exp} refers to not using the L_{spa} and L_{exp} proposed in this paper, but instead using the original L_{spa} and L_{exp} from Zero-DCE++ [12]. It can be obviously observed that the proposed and in this paper results in improvement in both peak signal-to-noise ratio (PSNR) and structural similarity (SSIM).

As shown in Table 2 and Figure 3, we further compare the effectiveness of the global histogram loss and the superpixel color consistency loss in the color restoration network. Fig. 3 shows the visual quality comparison of the ablation study on L_{his} and L_{SLIC}. It can be observed from Fig. 3(c) and (e) that the introduction of L_{his} enhances the saturation of the image, making it closer to the colors of the reference image. This indicates the importance of using global histogram loss for overall color restoration. However, using L_{his} alone may result in color distortion in local regions. Therefore, the addition of L_{SLIC} is necessary to achieve more precise color and detail restoration in local areas. Similarly, using L_{SLIC} alone may lead to the network overly focusing on local regions, resulting in the generation of strange artifacts. Only by simultaneously using L_{his} and L_{SLIC} (see Fig. 3(e)) can both the global and local regions of the image be effectively addressed, significantly enhancing PSNR.

C. Comparison with State-of-the-Arts

To evaluate the effectiveness of the proposed method, we carry out comparative experiments with LIME [7], RetinexNet [8], EnlightenGAN [10], Zero-DCE++ [12], KinD [13] and LE-GAN [16]. The evaluation is conducted on the LOL-V2 dataset and the second part of the SICE dataset, using the datasets and

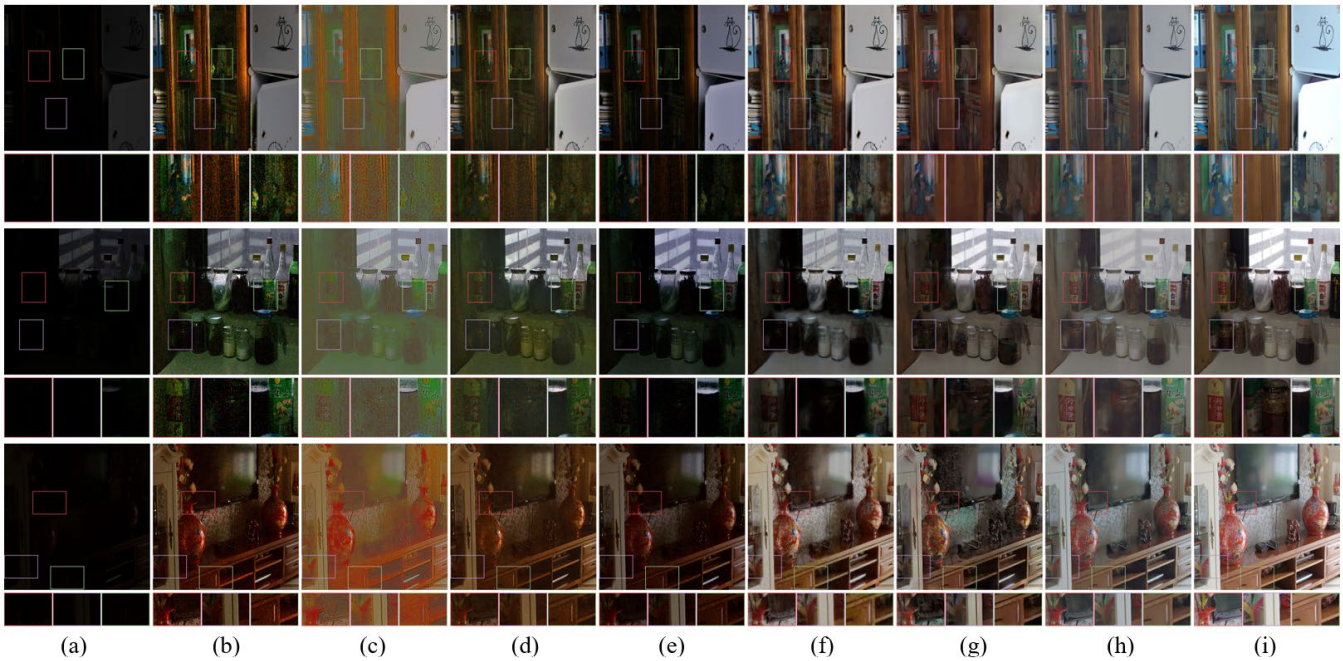


Fig. 4 Visual comparisons on low-light images sampled from the LOL-V2 datasets. From left to right, (a) input. (b) LIME[7]. (c) RetinexNet[8]. (d) EnlightenGAN[10]. (e) Zero-DCE++[12]. (f) KinD[13]. (g) LE-GAN[16]. (h) proposed. (i) GT.

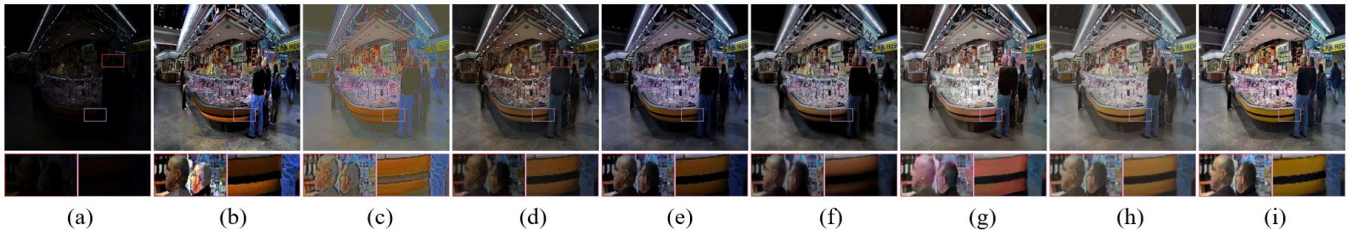


Fig. 5 Visual comparisons on low-light images sampled from the SICE datasets. From left to right, (a) input. (b) LIME[7]. (c) RetinexNet[8]. (d) EnlightenGAN[10]. (e) Zero-DCE++[12]. (f) KinD[13]. (g) LE-GAN[16]. (h) proposed. (i) GT.

parameter settings as provided in the original papers. Both PSNR and SSIM are employed for quantitative assess As shown in Table 3, LIME [7] refines the initial illumination component by applying a structure-aware prior, but its performance is relatively limited due to the lack of consideration for the color and detail degradation of the reflection component. RetinexNet [8] and EnlightenGAN [10] improve performance by incorporating denoising for the reflection component and employing a global-local discriminator, respectively. Zero-DCE++ [12], being a purely unsupervised method, exhibits performance that is not quite ideal compared to reference-image-guided approaches. By guiding the recovery of the degraded reflection component using the illumination component, KinD [13] achieves significantly better performance, particularly attaining the highest PSNR values on the SICE dataset. This success is attributed to its enhanced ability to recover the reflection component in low-light images with less severe degradation. LE-GAN [16] enhances performance by selectively emphasizing semantic information through an illumination-aware attention module. Our approach, employing a specific loss function conducive to the restoration of both global and local color information, has yielded excellent results in the LOL-V2 and SICE datasets. It is particularly noteworthy that our approach not only addresses the restoration of color information in low-light images but also emphasizes adaptive enhancement across different brightness regions. Consequently, on the SICE dataset, despite a slightly lower PSNR compared to KinD [13] , our approach still achieves an impressive SSIM value of 0.823, which represents the best outcome.

To demonstrate the improved visual effects of the proposed method, Fig. 4 and 5 depicts the qualitative evaluation results on LOL-V2 and SICE datasets, respectively. As shown in Fig. 4, RetinexNet [8], LIME [7] and EnlightenGAN [10] exhibits noticeable color distortion and noise. Zero-DCE++ [12] falls short in enhancement and fails to eliminate color distortion. KinD [13] effectively restores color but introduces artifacts, and its enhancement for extremely low-light images is inadequate, revealing a constrained dynamic range in dark regions. LE-GAN [16] effectively eliminates color distortion but may introduce artifacts or fail to robustly restore colors. Contrastingly, our method excels in not only enhancing low-

light images but also effectively reducing color distortion and noise. In Fig. 5, we show the enhancement results on the SICE dataset. It can be observed that our method produces results without significant noise and color distortion, exhibiting superior visual quality.

V. CONCLUSION

In this paper, a semi-supervised two-stage network is proposed for effectively enhance low-light images while restoring degraded color information. In the curve estimation stage, adaptive enhancement for different brightness regions is achieved through the utilization of the adaptive spatial consistency loss and the graded exposure control loss, facilitating preliminary image enhancement. The color restoration stage utilizes the global histogram loss and the superpixel color consistency loss to globally and locally restore the color information of the enhanced image. Qualitative and quantitative experimental results on a low-light dataset demonstrate the outstanding performance of our proposed method.

REFERENCES

[1] M. P. Stephen, E. P. Amburn, D. A. John, C. Robert, G. Ari, G. Trey, Bart, B. Z. John, and Z. Karel, "Adaptive Histogram Equalization and Its Variations," Computer Vision, Graphics, and Image Processing, vol. 39, no. 3, pp. 355-368, 1987.

[2] T. Arici, S. Dikbas, and Y. Altunbasak, "A Histogram Modification Framework and Its Application for Image Contrast Enhancement," IEEE Transactions on Image Processing, vol. 18, no. 9, pp. 1921-1935, 2009.

[3] C. Lee, C. Lee, and C.-S. Kim, "Contrast Enhancement Based on Layered Difference Representation of 2D Histograms," IEEE Transactions on Image Processing, vol. 22, no. 12, pp. 5372-5384, 2013.

[4] E. H. Land, "The Retinex Theory of Color Vision," Scientific American, vol. 237, no. 6, pp. 108-129, 1977.

[5] D. J. Jobson, Z. Rahman, and G. A. Woodell, "Properties and Performance of a Center/Surround Retinex," IEEE Transactions on Image Processing, vol. 6, no. 3, pp. 451-462, 1997.

[6] Z. Rahman, D. J. Jobson, and G. A. Woodell, "Multi-Scale Retinex for Color Image Enhancement," Proceedings of 3rd IEEE International Conference on Image Processing, Lausanne, Switzerland, 1996, pp. 1003-1006 vol.3.

[7] X. Guo, Y. Li, and H. Ling, "LIME: Low-Light Image Enhancement via Illumination Map Estimation," IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 982-993, 2017.

[8] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex Decomposition for Low-Light Enhancement," ArXiv, vol. abs/1808.04560, 2018.

[9] C. Chen, Q. Chen, J. Xu and V. Koltun, "Learning to See in the Dark," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 3291-3300.

[10] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep Light Enhancement Without Paired Supervision," IEEE Transactions on Image Processing, vol. 30, pp. 2340-2349, 2021.

[11] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2020, pp. 1777-1786.

[12] C. Li, C. Guo, and C. C. Loy, "Learning to Enhance Low-Light Image via Zero-Reference Deep Curve Estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 8, pp. 4225-4238, 2022.

[13] Y. Zhang, J. Zhang, and X. Guo, "Kindling the Darkness: A Practical Low-light Image Enhancer," in Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 2019, pp. 1632–1640.

[14] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond Brightening Low-light Images," International Journal of Computer Vision, vol. 129, no. 4, pp. 1013-1037, Apr, 2021.

[15] Z. Zhang, H. Zheng, R. Hong, M. Xu, S. Yan, and M. Wang, "Deep Color Consistent Network for Low-Light Image Enhancement," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022, pp. 1889-1898.

[16] Y. Fu, Y. Hong, L. Chen, and S. You, "LE-GAN: Unsupervised Low-Light Image Enhancement Network using Attention Module and Identity Invariant Loss," Knowledge-Based Systems, vol. 240, pp. 108010, 2022.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," ArXiv, vol. abs/1505.04597, 2015.

[18] G. Buchsbaum, "A Spatial Processor Model for Object Colour Perception," Journal of the Franklin Institute, vol. 310, no. 1, pp. 1-26, 1980.

[19] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," CoRR, vol. abs/1409.1556, 2014.

[20] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2274-2282, 2012.

[21] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse Gradient Regularized Deep Retinex Network for Robust Low-Light Image Enhancement," IEEE Transactions on Image Processing, vol. 30, pp. 2072-2086, 2021.

[22] J. Cai, S. Gu, and L. Zhang, "Learning a Deep Single Image Contrast Enhancer from Multi-Exposure Images," IEEE Transactions on Image Processing, vol. 27, no. 4, pp. 2049-2062, 2018.

Robust Watermarking via Dual Guidance

Yuhang Zhang[1], Yuanman Li[1*], Teng Sio Hong[2], Li Dong[3], Xia Li[1]
[1] Shenzhen University, Shenzhen, China
E-mail: yuanmanli@szu.edu.cn
[2] Faculty of Humanities and Social Sciences, Macao Polytechnic University, Macau
E-mail: p1308598@mpu.edu.mo
[3] Ningbo University, Ningbo, China
E-mail: dongli@nbu.edu.cn

***Abstract*—In recent years, digital watermarking methods based on deep learning have been extensively studied due to their crucial role in protecting copyright information and forensic tracing. Existing robust watermarking methods have attained notable imperceptibility and robustness through end-to-end training, as well as by employing techniques such as adversarial training and simulated attacks. However, these methods often neglect the watermark encoding process during the embedding stage. Typically, they use duplicated or upsampled watermarks for embedding, which disregards the characteristics of the cover image and limits the performance of the watermarking method. To address this limitation, we propose a robust watermarking via dual guidance. During the embedding stage, we utilize both global semantic features and local features of the cover image to guide the watermark encoding and embedding processes. The global semantic feature guides the watermark encoding, enabling an adaptive encoding process based on the cover image, thereby enhancing the imperceptibility and robustness of the watermarking method. Local features are used to generate embedding masks that adjust the strength of the watermark embedding. In the extraction stage, we introduce a multi-scale attention-based decoder to extract the watermark from the global features of the embedded image. Experimental results demonstrate that our method outperforms existing watermarking models.**

I. Introduction

Digital image watermarking has been extensively studied within the field of information security. As a prominent topic in security research, it involves embedding specific information into digital images to protect copyright and ensure data integrity. Digital image watermarking is classified into two main types: fragile watermarking and robust watermarking. Fragile watermarking is primarily used for detecting and identifying image tampering or modification, whereas robust watermarking is designed for copyright protection and leakage tracking. Robust watermarking possesses strong anti-interference capabilities, enabling effective detection and extraction even after various forms of image processing or attacks.

With the advancement of deep learning, robust watermarking methods have increasingly utilized neural networks for both watermark embedding and extraction. Zhu *et al.* [1] introduced the first deep learning approach for robust image watermarking, employing end-to-end trained neural networks. Since then, researchers have developed various deep watermarking frameworks, continually improving the imperceptibility and robustness of these methods. They have employed techniques such as embedding masks [2], adversarial examples [3], and distortion simulation [4].

However, in the embedding stage of existing methods, insufficient attention has been paid to the watermark encoding process. Typically, these methods either directly expand the watermark sequence [1], [5] or encode the watermark into specific patterns, such as QR codes or data matrices [6]. These approaches neglect the unique characteristics of the cover image and fail to adaptively encode the watermark based on the image, resulting in increased redundancy of the encoded watermark and compromised imperceptibility of the watermarking method.

To address this limitation, we propose a robust image watermarking model that employs dual guidance during both the watermark preprocessing and embedding stages. The attention mechanism is used to capture both global and local features of the cover image. Global features guide the preprocessing of watermark encoding, allowing for adaptive encoding based on the cover image's characteristics. This approach aims to reduce inconsistencies and redundancies during embedding. Simultaneously, local features generate embedding masks that guide the final watermark embedding, enhancing the visual quality of the watermarking method. The encoded watermark is embedded into the image using a U-Net, and the final watermark image is produced through residual connections and embedding mask guidance. At the extraction stage, we introduce a multi-scale attention-based decoder. This decoder, leveraging the attention mechanism, learns the global features of the image, facilitating effective extraction of the watermark that is closely integrated with the image's global semantic features, thereby improving the robustness of the watermarking method.

Our main contributions are summarized as follows:

- We propose a watermark embedding method that is guided by both global and local features of the image. The global feature is combined with the original watermark to generate an adaptive encoded watermark, while the local features are used to generate an embedding mask that guides the final embedding.
- In the extraction stage, we re-extract the global semantic feature of the image and extract the watermark, which is

*Corresponding author

highly integrated with the global feature, by employing a multi-scale attention-based decoder.

- We demonstrate that the proposed framework offers improved visual quality and greater robustness compared to current state-of-the-art methods. Additionally, our method effectively balances imperceptibility and robustness, making it adaptable to various conditions.

II. Related Work

A. Digital image watermarking

In the field of information-hiding technology, digital watermarking has garnered extensive research attention alongside steganography. Similar to steganography, digital watermarking embeds specific information by making subtle modifications to images. However, it emphasizes robustness. Robust watermarking methods are designed to recover the embedded watermark even if the image is modified or distorted. Schyndel *et al.* [7] introduced the concept of digital watermarking and proposed the first Least Significant Bit (LSB) watermarking method. Despite this, the LSB watermarking method has limitations in terms of robustness. To achieve improved robustness and imperceptibility, researchers have since explored embedding information in various transform domains of images, including the Discrete Cosine Transform (DCT) [8], the Fourier-Mellin domain [9] and histograms [10].

B. Deep watermarking

In recent years, deep learning has advanced rapidly, achieving notable success in various image processing tasks such as pattern recognition, image classification, and tamper detection[11]–[14]. It has also been increasingly integrated into digital watermarking, leading to the development of numerous watermarking network frameworks. Tancik M. *et al.* [5] proposed StegaStamp, which incorporates an image perturbation module into the training process, enabling digital watermarking methods to achieve robustness against real-world distortions such as printing and screen capturing. Luo *et al.* [3] introduced a distortion-aware watermarking framework that does not explicitly model image distortion during training. Instead, it uses adversarial training, incorporating distorted image samples generated by adversarial networks to improve the robustness of the watermarking method.

C. Watermark encoding

Watermark encoding is a crucial step in deep watermarking methods, as a watermark sequence or text typically cannot be directly embedded into an image by a neural network. Therefore, the watermark message or information must be transformed into a format that can be processed by the neural network and embedded into the image. Most existing methods directly duplicate the watermark message to match the size of the input image, concatenate it with the image, and then process it through the watermark encoder. Jia *et al.* [6] employs a data matrix to represent 96 bits of data and embeds this spatially redundant matrix into the image.

These encoding methods focus primarily on the watermark itself, often neglecting the cover image, which serves as the carrier for the encoded watermark. This lack of consideration for image characteristics can result in inconsistencies between the encoded watermark and the cover image, leading to the embedding of redundant or unnecessary information and damaging imperceptibility.

III. Proposed Method

As shown in Fig. 1, the proposed model is an end-to-end framework including an encoder, a noiser, and a decoder. The following subsections will describe these modules in detail.

A. Encoder

To address the limitation of existing methods that overlook the image's characteristics during the watermark encoding phase, we introduce guidance based on image's global features into the encoding process within the encoder. This approach generates encoded watermarks that are closely integrated with the global features of the image. By considering the characteristics of the image, these encoded watermarks reduce redundancy and enhance both imperceptibility and robustness.

Additionally, to further improve the method's imperceptibility, we incorporate embedding mask guidance during the watermark embedding phase. Local image features are used to generate embedding masks, which adaptively embed the watermark into the image with appropriate intensity based on the features of different image regions.

The final embedding process is performed using U-Net, a widely utilized generative model. Its U-shaped multi-scale structure effectively captures both shallow and deep features of images and encoded watermarks.

1) Dual guidance: In our method, we employ a transformer-based network as a preprocessor to provide dual guidance during the embedding stage. In the transformer blocks, images are divided into patches, with each token corresponding to a patch learning the local features of the image. Additionally, an extra learnable token aggregates the features from other tokens by the self-attention mechanism to generate the global semantic features of the image

$$Token_{local}, Token_{global} = Transformer(I_{co}) \quad (1)$$

where $Transformer(\cdot)$ represents the preprocessing process by transformer blocks.

We utilize global features to guide the watermark encoding process. The watermark sequence W is then concatenated with the global semantic feature $Token_{global}$ and then upsampled to generate an adaptively encoded watermark $W_{en} \in R^{1\times H\times W}$:

$$W_{en} = WatermarkEncoder(W, Token_{global}) \quad (2)$$

where $WatermarkEncoder(\cdot)$ represents the watermark encoding process by linear layers and reshape operation.

Furthermore, we utilize local features to generate an embedding mask that guides the final embedding of the watermark.

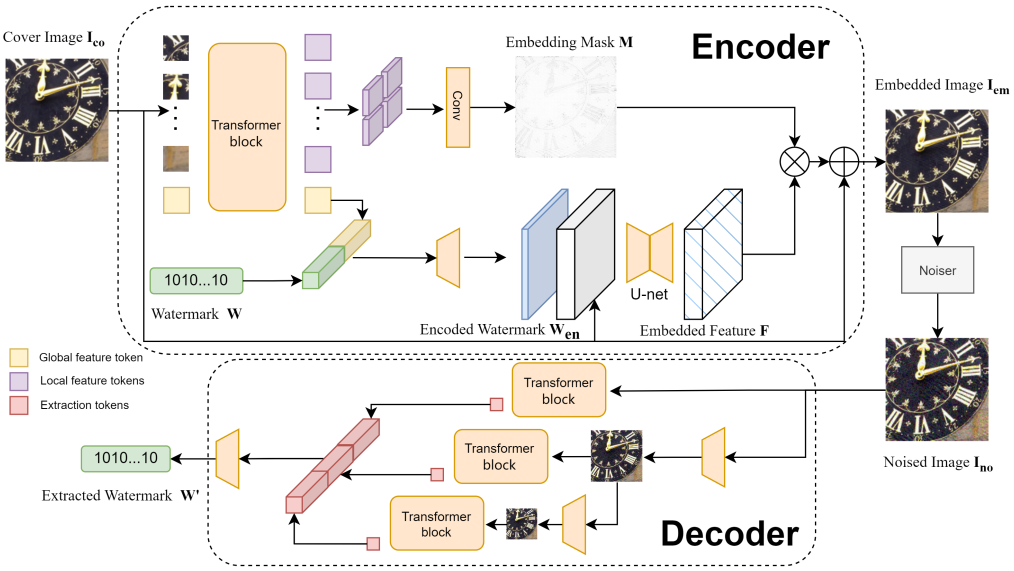


Fig. 1. The framework of the proposed method. The transformer block takes the cover image as input and extracts both the global feature and local features. The local feature tokens are reconstructed into an embedding mask; The watermark sequence is fused with the global feature to generate an encoded watermark; This encoded watermark is then combined with the cover image and fed into U-Net, which generates an embedded feature. The embedded feature, along with the embedding mask and the cover image, is merged via a residual connection to produce the final watermark image. The noise layer distorts the encoded image, producing a noised image. The decoder extracts the watermark from the noised image.

The local feature tokens are reshaped into patches and stitched together to match the image size. After passing through a convolution layer, the local feature tokens are reconstructed into the embedding mask $M \in R^{1\times H\times W}$:

$$M = Conv(RS(Token_{local})) \tag{3}$$

where $Conv(\cdot)$ represents the convolution layer, $RS(\cdot)$ represents the reshape and stitch operation.

2) Watermark embedding: In the watermark embedding process, a U-Net is employed to embed the encoded watermark into the cover image. The cover image I_{co} and the guided encoded watermark W_{en} are concatenated and used as inputs to the network, generating the embedded feature $F \in R^{3\times H\times W}$:

$$F = Unet(I_{co}, W_{en}) \tag{4}$$

where $Unet(\cdot)$ represents the embedding process of the U-Net.

The embedding mask M obtained from the previous stage is then multiplied with the embedded feature F and added to the cover image I_{co}. We introduce a strength factor s to adjust the embedding strength, setting $s = 1$ in the training phase. This setting means that the embedding strength factor remains unchanged during training. In subsequent experiments, we adjust s to balance the imperceptibility and robustness of our method. The embedded image I_{em} is generating as:

$$I_{em} = I_{co} + M \times F \times s. \tag{5}$$

B. Noiser

During training, noiser employs differentiable image processing operations to handle embedded images. With the assistance of this noise layer, the decoder learns to resist distortions effectively. We incorporated several common noise types into the network, including Cropout($p = 0.3$), Dropout($p = 0.3$), Gaussian Blur($\sigma = 2$), Random Noise, JPEG compression, and Color Jitter. Since real JPEG compression is a non-differentiable operation, we use a simulated JPEG method [15] as the noise layer, enabling the network to learn to handle real JPEG compression effectively.

The embedded image I_{em} passes through noise layers, generating a noised image I_{no}, which is then used for watermark extraction:

$$I_{no} = Noiser(I_{em}) \tag{6}$$

where $Noiser(\cdot)$ represents the noise adding process.

C. Decoder

At the embedding stage, the encoded watermark is generated with guidance from the global semantic features of the image, ensuring that it is closely aligned with these features. After processing through U-Net, the watermark is embedded within the global semantic features of the image, rather than being simply embedded into specific regions.

At the extraction stage, we introduce the transformer block and utilize the self-attention mechanism to learn the global semantic features of the watermark image. This allows us to extract watermark features that are highly integrated and correlated with the global features. To align with the multi-scale feature embedding process of U-Net during the embedding phase, we incorporated a multi-scale structure into the decoder. This involves using three transformer blocks with varying input sizes and patch sizes to handle features of different image scales. This multi-scale approach enhances the robustness of the watermark extraction process. The noised image I_{no} is

input into the decoder and the decoder outputs the extracted watermark W'

$$W' = Decoder(I_{no}) \tag{7}$$

where $Decoder(\cdot)$ represents the decoding process operated by the decoder.

D. Training strategy

We choose two loss functions to optimize the encoder :

$$L_{pix} = MSE(I_{co}, I_{em}) \tag{8}$$

$$L_{perp} = LPIPS(I_{co}, I_{em}) \tag{9}$$

where $MSE(\cdot)$ is the mean-square-error function and $LPIPS(\cdot)$ is the learned perceptual image patch similarity loss.

The loss function to optimize the decoder is defined as:

$$L_{dec} = MSE(W, W') \tag{10}$$

where $MSE(\cdot)$ is the mean-square-error function.

In our experiments, we observed that the training strategy significantly impacts the model's convergence. Directly training the entire network can result in poor convergence. Therefore, we train the model in the following two stages:

1) The first stage: We first train the network to enable it to extract watermarks. During this phase, we focus solely on optimizing the decoding process to ensure robustness against various forms of noise. The training at this stage is governed by the following loss function:

$$L_{stage_1} = \lambda_1 \times L_{dec} \tag{11}$$

where λ_1 represent the weights of L_{dec} .

2) The second stage: After the first stage, the network effectively recovers watermark information; however, the image quality produced by the encoder remains poor. In the second stage, we aim to balance imperceptibility and robustness, completing the training of the watermarking network. To achieve this, we incorporate both MSE loss and LPIPS loss into the loss function. The final loss function is as follows:

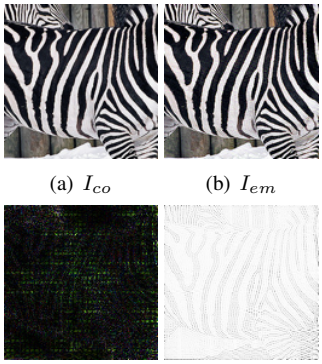
$$L_{stage_2} = \lambda_1 \times L_{dec} + \lambda_2 \times L_{pix} + \lambda_3 \times L_{perp} \tag{12}$$

where λ_1-λ_3 represent the weights of L_{dec}, L_{pix}, L_{perp}, respectively.

IV. Experimental results

A. Implementation

We use 76,000 images from the COCO train-2014 dataset for training and 1,000 images from the COCO test-2014 dataset for testing. The entire framework is implemented using PyTorch and executed on three NVIDIA RTX 3090 Ti GPUs. All images have a resolution of 128 × 128 pixels. Watermark sequences are sampled randomly and each watermark has a length of 30 bits.



(a) I_{co} (b) I_{em}

(c) R (d) M

Fig. 2. Visual quality of our methods (s=1). (a)cover image I_{co}; (b)embedded image I_{em}; (c)residual $R = |I_{em} - I_{co}| * 5$; (d) embedding mask M.

TABLE I
PERFORMANCE OF OUR METHOD (S=1) AND COMPARISON WITH[1],[16]

Method	PSNR	SSIM	Bit Accuracy under Different Noise(%)			
			JPEG	Cropout	Dropout	GaussianBlur
HiDDeN	33.55	-	63	94	93	96
TSDL	33.5	-	76.2	97.3	97.4	98.6
ours	**33.99**	0.959	**99.9**	**99.8**	**99.7**	**99.9**

B. Hyper-parameter Setting

The batch size is set to 384. In training, we use 10 epochs for the first stage and 490 epochs for the second stage, making a total of 500 epochs. In the first stage, the weight parameter of the loss function is set to $\lambda_1 = 1$. In the second stage, the weight parameters are set to $\lambda_1 = 3$, $\lambda_2 = 30$ and $\lambda_3 = 30$. For the optimizer, we employ Adam with an initial learning rate of 10^{-4} , which decays according to a cosine annealing schedule until it reaches 4×10^{-5}.

C. Evaluate Metrics

We used standard metrics to evaluate our watermarking method. Robustness is measured by Bit Accuracy, which is the ratio of identical bits between the input watermark W and the extracted watermark W'. The visual quality of the watermarking method is assessed using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) between the cover image I_{co} and the embedded image I_{em}.

D. Performance and Comparison with SOTA Methods

First, we present the result of our method with the strength factor $s = 1$ and compare it with two state-of-the-art methods: HiDDeN[1] and TSDL[16]. For HiDDeN and TSDL, we use the results published by their authors. We test our method using the following distortions: Cropout ($p = 0.3$), Dropout ($p = 0.3$), Gaussian Blur ($\sigma = 2$) and JPEG compression ($Q = 50$). JPEG compression is implemented by the PIL package.

Fig. 2 shows the visual result of our method with $s = 1$. Table I shows our method's performance in visual quality and robustness against common noises, compared to HiDDeN and TSDL. It demonstrates that our method has achieved better performance in terms of imperceptibility and robustness, and

TABLE II
PERFORMANCE COMPARISONS WITH [2]

Method	PSNR	SSIM	Bit Accuracy under Different Noise (%)									
			JPEG	Cropout	Dropout	GaussianBlur	MotionBlur	Bilateral	GaussianNoise	Hue	Contrast	Brightness
ARWGAN	37.87	0.981	51.2	97.5	**98.4**	99.7	91.7	78.2	80.8	64.6	98.2	97.8
ours(s=0.6)	**38.36**	**0.982**	**99.7**	**99.2**	97.9	**99.9**	**99.7**	**98.0**	**99.8**	**99.3**	**99.4**	**99.3**

TABLE III
PERFORMANCE COMPARISONS BETWEEN DIFFERENT EMBEDDING STRENGTH s

s	PSNR	SSIM	Bit Accuracy under Different Noise (%)									
			JPEG	Cropout	Dropout	GaussianBlur	MotionBlur	Bilateral	GaussianNoise	Hue	Contrast	Brightness
s=1	33.99	0.959	99.98	99.84	99.72	99.98	99.79	99.86	99.96	99.71	99.72	99.63
s=0.9	34.90	0.965	99.95	99.75	99.54	99.98	99.81	99.81	99.97	99.63	99.69	99.61
s=0.8	35.91	0.971	99.96	99.71	99.31	99.97	99.81	99.63	99.94	99.62	99.64	99.58
s=0.7	37.05	0.977	99.91	99.54	98.88	99.95	99.79	99.12	99.94	99.50	99.61	99.42
s=0.6	38.36	0.982	99.77	99.25	97.90	99.91	99.70	98.01	99.86	99.33	99.48	99.33

TABLE IV
PERFORMANCE COMPARISONS BETWEEN DIFFERENT NETWORK STRUCTURES

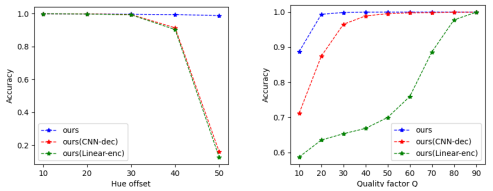
Method	PSNR	SSIM	Bit Accuracy under Different Noise (%)									
			JPEG	Cropout	Dropout	GaussianBlur	MotionBlur	Bilateral	GaussianNoise	Hue	Contrast	Brightness
ours	**33.99**	**0.959**	**99.98**	**99.84**	99.72	**99.98**	99.79	**99.86**	**99.96**	**99.71**	**99.72**	**99.63**
CNN-dec	33.72	0.944	99.54	97.32	99.71	99.90	99.90	99.86	99.86	99.49	99.35	99.27
Linear-enc	33.59	0.945	69.94	99.75	**99.90**	99.98	**99.92**	96.57	99.96	99.38	99.58	99.48

has a significant lead in image quality and accuracy of all noises. In the case of real JPEG compression, it is more than 23% higher than the two methods. In terms of robustness to other noises, our method has an advantage of 1% to 6%.

We also compare our model with the state-of-the-art model ARWGAN [2]. For ARWGAN, we use the pre-trained model provided by the authors and perform the same tests for comparison. To ensure a fair comparison, We adjust the embedding strength to $s = 0.6$ to approximate the PSNR of ARWGAN. We tested the robustness of both our method and ARWGAN against more noise types than in the previous comparison, including Motion Blur($angle = 30, kernel \quad size = 5$), Bilateral Filter($d = 9$), Gaussian Noise($\sigma = 0.05$), Hue($offset = 30$), Contrast($ratio = 1.5$), Brightness($offset = -80$). Table II shows that, under similar PSNR, our method demonstrates significant advantages in handling most types of noise. The only exception is with Dropout noise, where the accuracy of our method is slightly lower than that of ARWGAN by 0.5%. ARWGAN loses its robustness when facing real JPEG, while our method achieves robustness to real JPEG by training with simulated JPEG. When dealing with more noises, such as Hue, Brightness, Contrast, Motion Blur, and Bilateral Filtering, our method shows clear advantages and outperforms ARWGAN by 13.4% in accuracy on average. Especially for unseen noises such as Gaussian noise, Motion Blur, and Bilateral Filtering, our method has a 15% lead in accuracy on average. This means that our method has better versatility and can be better applied to situations with unknown noise.

E. Ablation Study

1) The Effect of Strength Factor: Depending on the requirements of different scenarios for watermarking methods, the embedding strength s can be adjusted to balance robustness and visual quality. Reducing the embedding strength s improves imperceptibility but may compromise some level of robustness. Table III shows the performance of our method under different s and indicates that our method achieves good performance under different image quality, which means that our method can be effectively applied to different scenarios.



(a) Accuracy under Hue (b) Accuracy under JPEG

Fig. 3. Comparison between different network structures

2) The Effect of the Attention-based Decoder: In this section, we trained a new network that employs a CNN architecture composed of multiple "conv-ReLU" blocks and a linear layer as the decoder, replacing the multi-scale attention-based decoder used in our method. This network, labeled 'CNN-dec' is compared with our proposed method.

Table IV shows that the multi-scale attention-based decoder helps achieve better imperceptibility and robustness against most types of noise compared to the CNN decoder, except for a slight decrease of approximately 0.11% under Motion Blur. Fig. 3 shows that our method is significantly more robust than the CNN decoder, particularly under strong hue offsets and low JPEG compression quality factors.

3) The Effect of Watermark Encoding Guided by the Global Feature: We train a new network that uses linear layers to encode and upsample the watermark, replacing our method's

global feature-guided encoding approach. This network, labeled 'Linear-enc' is compared with our method.

Table IV indicates that the performance of the 'Linear-enc' network under most noises is not as good as our method. This suggests that our guided watermark encoding approach improves both imperceptibility and robustness compared to methods that lack adaptive encoding.

What's more, as shown in Fig. 3, the 'Linear-enc' network has a more serious degradation than 'CNN-dec' under the two strong noises of Hue and JPEG compression, which means that the lack of guided encoding has a more serious impact on network performance. The incorporation of image feature-guided watermark encoding and the multi-scale attention-based decoder has led to noticeable improvements but guided encoding is more important for watermarking methods.

V. Conclusion

In this paper, we propose robust watermarking via dual guidance, which guides the watermark encoding and embedding process according to the features of the cover image. In our method, the global feature of the image guides the watermark encoding process to achieve adaptive encoding, while local features are reconstructed into embedding masks to guide the final embedding of the watermark. To effectively extract the watermark that is highly fused with the global features of the image, we employ a multi-scale attention-based decoder. Extensive experiments demonstrate that, compared to state-of-the-art methods, our proposed framework is not only robust to common noise but also resilient to various high-intensity noise. Furthermore, our watermarking method can adapt to different image quality requirements by adjusting the embedding strength while maintaining strong robustness.

Acknowledgment

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010645; in part by the Key project of Shenzhen Science and Technology Plan under Grant 20220810180617001; in part by the Open Research Project Programme of the State Key Laboratory of Internet of Things for Smart City (University of Macau) under Grant SKLIoTSC(UM)-2021-2023/ORP/GA04/2022.

References

[1] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 657–672.

[2] J. Huang, T. Luo, L. Li, G. Yang, H. Xu, and C.-C. Chang, "Arwgan: Attention-guided robust image watermarking model based on gan," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–17, 2023.

[3] X. Luo, R. Zhan, H. Chang, F. Yang, and P. Milanfar, "Distortion agnostic deep watermarking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 548–13 557.

[4] C. Qin, X. Li, Z. Zhang, F. Li, X. Zhang, and G. Feng, "Print-camera resistant image watermarking with deep noise simulation and constrained learning," *IEEE Transactions on Multimedia*, 2023.

[5] M. Tancik, B. Mildenhall, and R. Ng, "Stegastamp: Invisible hyperlinks in physical photographs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2117–2126.

[6] J. Jia, Z. Gao, D. Zhu, X. Min, G. Zhai, and X. Yang, "Learning invisible markers for hidden codes in offline-to-online photography," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2273–2282.

[7] R. G. Van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," in *Proceedings of 1st international conference on image processing*, IEEE, vol. 2, 1994, pp. 86–90.

[8] E. Koch and J. Zhao, "Towards robust and hidden image copyright labeling," in *IEEE Workshop on Nonlinear Signal and Image Processing*, vol. 1, 1995, pp. 123–132.

[9] J. J. Ò. Ruanaidh and T. Pun, "Rotation, scale and translation invariant spread spectrum digital image watermarking," *Signal processing*, vol. 66, no. 3, pp. 303–317, 1998.

[10] T. Zong, Y. Xiang, and I. Natgunanathan, "Histogram shape-based robust image watermarking method," in *2014 IEEE International Conference on Communications (ICC)*, IEEE, 2014, pp. 878–883.

[11] Y. Li, L. Ye, H. Cao, W. Wang, and Z. Hua, "Cascaded adaptive graph representation learning for image copy-move forgery detection," *ACM Transactions on Multimedia Computing, Communications and Applications*,

[12] Y. Li, Y. He, C. Chen, *et al.*, "Image copy-move forgery detection via deep patchmatch and pairwise ranking learning," *arXiv preprint arXiv:2404.17310*, 2024.

[13] Y. Tan, Y. Li, L. Zeng, J. Ye, W. Wang, and X. Li, "Multi-scale target-aware framework for constrained splicing detection and localization," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8790–8798.

[14] Y. Li, Y. He, C. Chen, *et al.*, "Image copy-move forgery detection via deep patchmatch and pairwise ranking learning," *IEEE Transactions on Image Processing*, pp. 1–16, 2024.

[15] R. Shin and D. Song, "Jpeg-resistant adversarial images," in *NIPS 2017 workshop on machine learning and computer security*, vol. 1, 2017, p. 8.

[16] Y. Liu, M. Guo, J. Zhang, Y. Zhu, and X. Xie, "A novel two-stage separable deep learning framework for practical blind watermarking," in *Proceedings of the 27th ACM International conference on multimedia*, 2019, pp. 1509–1517.

Long Audio File Speaker Diarization with Feasible End-to-End Models

Kai-Wei Huang* and Chia-Ping Chen†
* National Sun Yat-Sen University, Taiwan
E-mail: kaiwei890604@gmail.com
† National Sun Yat-Sen University, Taiwan
E-mail: cpchen@cse.nsysu.edu.tw

***Abstract*—End-to-End Neural diarization (EEND) systems have become increasingly popular in recent years. However, EEND systems that use permutation invariant training (PIT) losses lead to unpredictable speaker order, making the processing of long audio files challenging. In this paper, we replace the common Transformer encoder in EEND with a Longformer architecture to enhance the model's computational efficiency. We further refine the EEND model with a retention network (RetNet) architecture, achieving RNN-like characteristics without compromising accuracy. Evaluated on the Librimix dataset and our own recording, the Longformer encoder can feasibly diarize audio files 5 times longer than the baseline Transformer encoder. Moreover, the RetNet-EEND module even achieves 100 times longer than the baseline module. In practice, our implementation of RetNet-EEND-EDA model for speaker diarization can easily inference 5 hours long audio file without compromising recognition performance.**

I. Introduction

Speaker diarization is a task to separate different speakers in an audio recording. It is used in many scenarios such as meeting recordings and automatic subtitling of videos. Diarization is commonly used with Automatic Speech Recognition (ASR) to solve the problem of who spoke when and what was said. The traditional method treats diarization as a clustering problem [1], [2]. Voice Activity Detector (VAD) will be used to detect the active signal in the audio frames, which will be sent into a model to extract speaker embedding such as i-vector [1], [2], x-vector [3], [4], and d-vector [5], [6]. Speaker activities will be obtained by utilizing a clustering algorithm for speaker embedding. However, traditional diarization has several problems, such as its inability to handle overlap situations, which means that only one speaker can be active at the same time. Additionally, the clustering algorithm is not trainable so it can not be optimized to the best performance. Therefore, End-to-End Neural Diarization (EEND) [7], a neural network model based on an end-to-end architecture, is proposed as a solution.

EEND treats diarization as a multi-label classification problem by predicting the activity probability of different speakers at different times. Therefore, EEND can easily detect overlap situations, which is not possible with clustering methods. However, end-to-end models also have several limitations, such as handling an unknown number of speakers and the label permutation problem. These issues are not encountered in traditional clustering-based diarization methods. To address the unknown number of speakers, the EEND-EDA architecture introduces the Encoder Decoder Attractor (EDA) [8] framework. For the label permutation issue, EEND employs a permutation-free loss [7], [9] to dynamically determine the order of labels during training. The inconsistencies of the speaker's order and the self-attention mechanism inherent in the Transformer encoder [10], introduce significant challenges in processing long audio files. Consequently, in this paper, we replace the Transformer encoder in EEND to Longformer [11] architecture to reduce the computation. Additionally, we propose an improved RetNet-EEND architecture based on the retention network (RetNet) [12] to address the challenges of recognizing extremely long audio files.

The Longformer architecture enhances the ability to process longer audio files by reducing the complexity of self-attention calculations [10]. It significantly reduces the original computational complexity of self-attention from $O(n^2)$ to $O(n)$. Compared to the traditional Transformer, Longformer significantly reduces computational complexity. However, it still has limitations in processing audio lengths, which cannot meet our requirements for recognizing meeting audio files. Therefore, we propose the RetNet-EEND architecture, which is based on the retention network. The retention mechanism modifies the Transformer architecture by drawing inspiration from the characteristics of RNN models [13]. Traditionally, Transformers consider the tokens of the entire sequence length when computing attention. In contrast, the retention mechanism maintains only a single state for attention calculation. This approach allows for recurrent processing during inference. Through mathematical simplification, it still supports parallel computation during training. RetNet achieves a well-balanced trade-off in terms of computational speed, performance, and memory usage. Therefore, in this paper, we modify the Transformer encoder in the EEND-EDA model to incorporate RetNet. This modification enables the application of RetNet in diarization tasks, demonstrating its effectiveness and efficiency in handling such complex processes. In the RetNet encoder, to address the requirements of processing long conference audio files, we implement a Chunkwise Retention approach. This method involves segmenting the audio into chunks. Within each chunk, parallel computation is employed to calculate attention. Between chunks, a recurrent approach is used, retaining the state of each chunk. Finally, the final retention is

composed by cross chunk retention and inner chunk retention.

In this paper, the second section introduces the baseline architecture of EEND-EDA, along with the Longformer architecture that we aim to compare. We will also provide a detailed introduction to the RetNet-EEND architecture. In the third section, we will describe the datasets used in our experiments, parameter settings, and the performance metrics employed for comparison. The fourth section will present the experimental results in a tabular format and discuss our observations.

II. Model

In this section, we will introduce the baseline End-to-End Neural Diarization with Encoder-Decoder Attractor (EEND-EDA) architecture, Longformer-based encoder EEND, and the chunkwise forward RetNet-EEND architecture model.

A. Baseline Model

Our model is based on the End-to-End Diarization with Encoder-Decoder Attractor (EEND-EDA) architecture. EEND-EDA approaches diarization as a multi-label classification problem. The primary objective of EEND is to predict the speech activity of different speakers at various time points. When the probability at a given time point exceeds a certain threshold, EEND determines that someone is speaking at that moment. Consequently, the output of EEND is a probability distribution with dimension $S \times T$, where S represents the total number of speakers, and T is the temporal dimension of the predicted audio file.

A significant limitation of the traditional EEND model is the need to predetermine the number of speakers during the training phase. In order to address the issue of an unknown number of speakers, the Encoder Decoder Attractor (EDA) architecture was developed. The EDA module primarily utilizes a LSTM encoder-decoder architecture to predict the speaker's number S. The LSTM decoder, using the last cell state and hidden state of the encoder along with a zero vector, generates a variable number of attractors to form an attractor pool. These attractors are processed through linear and sigmoid functions to obtain their respective probabilities of existence. The EDA module will only utilize the top S attractors with none zero probabilities. Those attractors $A \in \mathbb{R}^{D_{\text{model}} \times S}$ are transposed and multiplied with the output of EEND Tranformer's encoder $e_t \in \mathbb{R}^{D_{\text{model}} \times T}$ to obtain the final speaker activity probability $p_t \in \mathbb{R}^{S \times T}$, where D_{model} is the hidden dimension in the model. The final calculation formula for p_t is as follows

$$p_t = \sigma\left(A^{\mathsf{T}} e_t\right) \in (0,1)^S \tag{1}$$

This approach effectively resolves the issue of an unknown number of speakers, which is a significant advancement over the traditional EEND model.

B. Longformer

In practical applications, due to the reliance of EEND on the Transformer architecture, the GPU's VRAM usage increases quadratically with the length of the audio file. Additionally, the permutation-free loss used in EEND results in an unpredictable order of speakers in the output, making it impractical to process the audio by segmentation. Long audio files segmented into different sections may cause inconsistencies in speaker output, leading to processing difficulties. Therefore, we attempted to incorporate the Longformer architecture into the EEND framework. Longformer simplifies the full attention mechanism in self-attention by employing three types of attention mechanisms: sliding window attention, dilated sliding window attention, and global + sliding window attention. In Figure 1, we show the difference between these three methods and full attention.

- Sliding Window Attention:
  This mechanism calculates attention only for tokens near the current token, significantly reducing the computational load compared to full attention, which considers all tokens.
- Dilated Sliding Window Attention:
  Similar to sliding window attention, this approach skips certain tokens to calculate attention, allowing for a broader scope of attention without significantly increasing computational complexity.
- Global + Sliding Window Attention:
  In this mechanism, Longformer selects specific tokens based on the task at hand to receive global attention. This allows certain key tokens to have a broader context, enhancing the model's overall understanding and performance.

By integrating these three types of attention, Longformer significantly reduces the computational complexity of self-attention without a substantial loss in performance. This makes it an efficient solution for processing longer audio files, addressing the limitations of the traditional Transformer-based EEND model in terms of VRAM usage. In this paper, the Longformer architecture is employed, utilizing the Sliding Window Attention and Dilated Sliding Attention mechanisms to construct local and global Attention, respectively. This replaces the Transformer architecture originally used in the EEND-EDA framework, reducing the system's overall complexity. We call this system as Longformer-EEND-EDA.

C. Retention Network

Although the computation is significantly reduced after incorporating Longformer, there still remains an upper limit to the length of audio files it could process. In this paper, we have adopted the retention network (RetNet) to replace the original Transformer architecture, which is inspired by its implementation in large language models (LLMs).

RetNet primarily draws from the architecture of RNNs. Given an input $X \in \mathbb{R}^{|x| \times D_{\text{model}}}$, where D_{model} is the hidden dimension in the model. During inference, it maintains a single state S_n and uses this state to calculate the output with attention. The output O_n is defined as follows

$$\begin{aligned} O_n &= Q_n S_n, & Q_n &\in \mathbb{R}^{1 \times D} \\ S_n &= A S_{n-1} + K_n^{\mathsf{T}} V_n, & A &\in \mathbb{R}^{D \times D}, K_n \in \mathbb{R}^{1 \times D} \end{aligned} \tag{2}$$

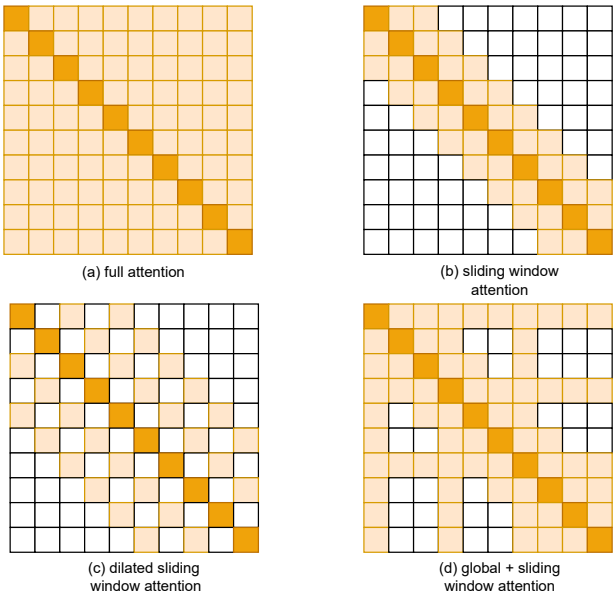


Fig. 1. Comparison between the full self-attention pattern and the attention patterns configured in Longformer.

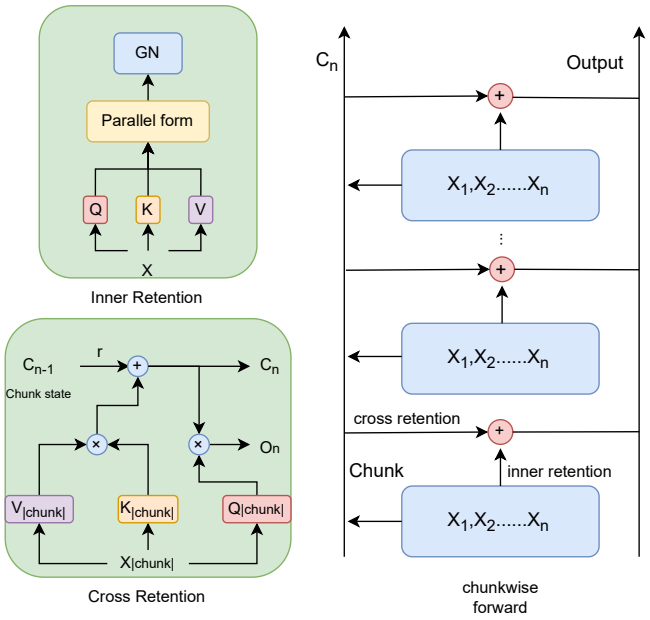


Fig. 2. RetNet chunkwise forward architecture within EEND. We segment the mel spectrogram into 5-second chunks before processing them through the RetNet encoder. Within each chunk, computations are performed in parallel form, while across chunks, we maintain a chunk state C_n to facilitate attention calculations between different chunks.

Q_n,V_n,K_n are simplified projection of X. By expanding the expression of O_n, we can obtain the recurrent representation as follows when S_0 is defined as zero

$$\sum_{m=1}^{n} Q_n A^{n-m} K_m^{\mathsf{T}} v_m \tag{3}$$

Next, diagonalize the matrix A to $p^{-1}\gamma e^{i\theta}p$ then absorbing p^{-1} and p into Q and K, and acknowledging the complex components $\gamma e^{i\theta}$ as Xpos [14] Θ, we can further simplify the recurrent form as follows

$$\text{Parallel Retention}(X) = (QK^{\mathsf{T}} \odot D)V \tag{4}$$

where D is a time exponential decay, and Q K are simply projection from X and dot product Θ. This approach eliminates the need to calculate the complexity of the entire sequence of tokens. During the training phase, the recurrent form can be simplified into a parallel form, significantly enhancing the training performance. This method combines the convenience of parallel training with the low-cost benefits of recurrent inference, effectively capturing the advantages of both approaches.

In the task of diarization, we employ a chunkwise forward representation in Figure 2 to handle extremely long audio files. And the full RetNet-EEND-EDA architecture with chunkwise forward is shown in Figure 3. The audio signal is segmented into various sizes of chunks. Within each chunk, the attention mechanism is calculated in a parallel manner, while between chunks, a recurrent approach is used, retaining only the state of previously computed chunks. The attention between different chunks is calculated using the chunk state, and this inter-chunk attention is then combined with the inner-chunk attention to form the final retention.

III. Experimental Setup

In this section, we will introduce the dataset and the configuration settings used in this experiment. Finally, we will introduce the performance metrics, which is used in our experiments.

A. Dataset

In our experiments, we use Librimix dataset [15], which is mixed by the Librispeech dataset and noise from WHAM! [16]. In this paper, we use Libri2mix only in our experiments, which contain just two speakers. Libri2mix consists of 2,484 speakers, including 1,450 males and 1,034 females. The Librimix dataset is mainly composed of 8khz conversational audio, 58 hours of data for training, and 11 hours of data for validation and testing. We also fine-tuned the EEND-EDA model with the Libri3Mix training set for the multi-speaker task. Our training set consists of a total of 40 hours, while the test and validation sets comprise 11 hours.

In addition to mixed datasets like LibriMix, we also used the ALi-meeting [17] dataset to fine-tune and test the performance on real audio files, and finally evaluated the Retnet-EEND-EDA's performance on long audio recordings. ALi-meeting is a Chinese diarization training set containing a total of 118 hours of annotated audio files, recorded using an 8-channel microphone array. However, in our training, we standardized the audio files to 8kHz and converted them to mono audio to reduce the computational load for end-to-end model training. When training the EEND-EDA-based model, the dataset must

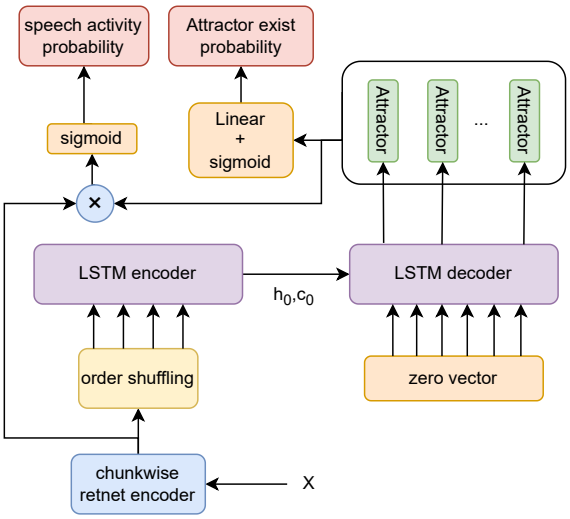


Fig. 3. The overall architecture of RetNet-EEND. The output from the RetNet encoder is sent into the EDA (Encoder Decoder Attractor) module to calculate the number of speakers. Finally, the attractors are multiplied with the original embeddings to obtain the probabilities of speaker activity.

have a uniform number of speakers to proceed. Therefore, we segmented the ALimeeting dataset, detecting 5-second segments within the training set and selecting 10 to 15-second clips containing two speakers for training. After processing, our training set totaled approximately 7 hours, with one hour each for the test and validation sets. Finally, in the data preparation phase, we use the Kaldi recognition toolkit [18] for preprocessing.

B. Configurations

This paper is primarily based on the EEND-EDA architecture from the ESPnet toolkit [19] on Pytorch framework [20]. In the baseline model, all parameter settings remain as original configuration. The Transformer encoder consists of four layers, each layer initially converting an 80-dimensional input into a 512-dimensional representation. In the multi-head attention mechanism, the number of heads is set to four, the final output size is configured to 256, and the dropout rate [21] is set at 0.1. In the EDA module, we maintain the use of an RNN architecture for the encoder-decoder, stacking only a single layer with 256 units. Regarding the Longformer, its overall structure is similar to the Transformer, consisting of four layers in total. For the RetNet, the encoder is also stacked in four layers. In terms of chunk segmentation, we divide 5-second audio clips into individual chunks for the model to process. All three architectures were trained on the Libri2Mix dataset, undergoing a total of 250 epochs with an early stop mechanism.

C. Performance Metrics

In diarization, we use the Diarization Error Rate (DER) [22] as a measure to evaluate performance. Since the results generated by EEND-EDA are discrete, we apply a median filter to smooth the output. In this experiment, we primarily employ an 11-frame median filter with a threshold of 0.5 and a collar tolerance of 0.0 seconds for evaluation purposes.

IV. Results

In this section, we will explore the Longformer-EEND-EDA and the RetNet-EEND-EDA architectures proposed in this paper. We will test these two models, which are improved for handling long audio files, on the test set of the Libri2Mix to evaluate their Diarization Error Rate (DER). These results will be compared with Transformer which is used full self-attention. Next, we will use our own recorded audio files to test the limits of recognition length for different architectures and present the final results in Table I and Table II.

In the baseline EEND-EDA model tested on the two-speaker Libri2Mix dataset, the DER was approximately 7.03. In the Libri3Mix fine-tuning tests, the EEND-EDA model achieved a DER of approximately 11.95. With a complete self-attention mechanism, the performance reveals commendably low error rates. However, this comes at the cost of significantly increased computational complexity. In our tests based on Nvidia GTX-1080Ti for inference, the EEND-EDA model was only capable of processing audio files up to three minutes in length. This duration is insufficient for typical audio recordings.

In contrast, the Longformer-EEND-EDA model, by simplifying the self-attention complexity from $O(n^2)$ to $O(n)$, effectively reduces the overall computational complexity. However, this simplification results in a slight decrease in performance, with the DER increasing from 7.03 for the Transformer to 10.32, marking an increase of 3.29. But the recognizable length of audio files saw a substantial improvement, extending from three minutes to forty-five minutes. In contrast, the Longformer model achieved a DER of approximately 11.75 in the Libri3Mix fine-tuning tests. For the three-speaker fine-tuning, the Longformer performed better compared to EEND-EDA, which we attribute to its relatively simplified model architecture. Without affecting the inference speed, the Longformer greatly extends the length of the audio files that can be recognized.

In the final phase of our research, we integrated the RetNet into the EEND model architecture, leveraging the characteristics of RNNs within RetNet to transition EEND towards a recurrent recognition approach. This modification enables EEND to process audio files of extremely long length. With the simplification of the recurrent form in RetNet, the model retains the advantages of the attention mechanism. Our experiments explored two distinct Xpos strategies: a global Xpos and a chunkwise Xpos calculation. With the global Xpos configuration, the model could understand global positional

relationships. However, this approach became impractical for excessively long audio files. Consequently, we refined the Xpos calculation method to only consider the positional encoding within each chunk in a chunkwise forward manner. This adjustment had a minimal impact on performance but significantly increased the length of audio files that could be processed. The global Xpos version encountered difficulties with audio files exceeding 15 minutes, whereas the revised chunkwise Xpos approach allowed us to process audio files up to three hours in length, with hardware utilization approximately at 60%. Since the model incorporating Global Xpos did not perform optimally on longer audio files, we chose not to proceed with testing on other datasets. We estimated that on GTX-1080Ti, it could handle audio files close to five hours in length, which we believe meets the vast majority of user needs. In GTX-1080Ti, EEND-EDA and Longformer architecture cost about 12 hours for training. The training of RetNet-EEND-EDA took nearly an entire day due to its recurrent mechanism.

In terms of recognition accuracy, the unique design of RetNet even surpassed the performance of traditional Transformer models. The DER achieved with global Xpos was 5.8, and with the transition to chunkwise Xpos, the DER slightly increased to 6.16. RetNet with chunkwise Xpos architecture successfully balances practicality and accuracy, indicating its substantial potential and utility for processing long audio files in diarization tasks. The Retnet-EEND model achieved a DER of approximately 10.75 in the Libri3Mix fine-tuning tests. Compared to the Transformer and Longformer models, the Retnet architecture demonstrated superior learning ability, reducing the error rate by about 1 relative to the previous systems.

In the ALi-meeting real recording test phase, we used a custom pre-processing program to extract segments with fixed two-speaker scenarios from the ALi-meeting dataset for fine-tuning. This significantly reduced the size of the overall dataset. However, in our tests, the Retnet-EEND-EDA model achieved a DER of 22.61 on our custom test set. This performance is notably better than the Transformer model's fine-tuning result of 25.75 on the ALi-meeting dataset. Furthermore, we tested the Retnet-EEND-EDA model on several long audio files from the original ALi-meeting test set, and the results were consistent with the fine-tuning performance. This demonstrates that the Retnet-EEND-EDA model surpasses the baseline Transformer model both in terms of recognition accuracy and the ability to handle longer audio files.

Although RetNet significantly improves accuracy and recognition length compared to Longformer and Transformer, its recognition speed is a drawback. Due to RetNet's recurrent design, it inherently cannot compete with the parallel computation capabilities of Longformer and Transformer in terms of speed. In our tests, recognizing a one-minute audio file with each chunk segmented into five seconds took approximately one second, whereas the Transformer and Longformer architectures required only about one-tenth of that time. We believe the RetNet architecture may be more suitable for non-real-time recognition scenarios.

TABLE I
PERFORMANCE OF OUR PROPOSED MODELS WITH DIFFERENT ARCHITECTURES. INCLUDING INFERENCE MAXIMUM LENGTH, AND TIME EACH MODEL COSTS WHEN INFERENCE 1 MINUTE AUDIO IN GTX 1080TI WITH 11GB VRAM.

Architecture	Length(↑)	Speed(↓)
Transformer	3 min	**0.11s/min**
Longformer	45 min	0.14s/min
RetNet global Xpos	15 min	1s/min
RetNet chunkwise Xpos	**300 min**	1s/min

TABLE II
DIARIZATION ERROR RATE (DER) PERFORMANCE OF OUR PROPOSED MODELS WITH DIFFERENT ARCHITECTURES IN DIFFERENT DATASETS. INCLUDING LIBRI2MIX, LIBRI3MIX, AND FINE-TUNING ON ALI-MEETING DATASET WITH PREPROCESSING

Architecture	Libr2mix(↓)	Libri3mix(↓)	ALi-meeting(↓)
Transformer	7.03	11.95	25.75
Longformer	10.32	11.75	27.34
RetNet global Xpos	**5.8**	-	-
RetNet chunkwise Xpos	6.16	**10.7**	**22.61**

V. CONCLUSIONS

In this paper, we propose two architectures, Longformer-EEND-EDA and RetNet-EEND-EDA, to mitigate the limitations of end-to-end diarization in recognizing long audio files. In the Longformer-EEND model, we leverage the Longformer's characteristic to reduce the computational load of self-attention within the Transformer framework, thereby decreasing the complexity of the EEND architecture. This adaptation enables the Longformer-EEND-EDA to extend the recognition length by 15 times compared to the baseline EEND-EDA model. However, due to the reduction in attention computation, the overall recognition accuracy decreased, and the recognition length still did not meet our requirements. Consequently, we explored incorporating the concept of the retention network into EEND, achieving promising results. After incorporating the retention network, due to RetNet's emulation of RNN characteristics and its unique attention mechanism calculations, the overall recognition performance improved by 100 times compared to the baseline EEND-EDA model. Moreover, the accuracy of RetNet-EEND-EDA slightly improved compared to the full attention EEND-EDA model. With these enhancements, RetNet-EEND-EDA significantly improves recognition length while maintaining a practical inference speed, alongside an increase in accuracy.

REFERENCES

[1] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2014, pp. 413–417.

[2] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[3] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 4930–4934.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 5329–5333.

[5] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," in *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 5239–5243.

[6] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 4879–4883.

[7] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[8] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," *arXiv preprint arXiv:2005.09921*, 2020.

[9] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 241–245.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[12] Y. Sun, L. Dong, S. Huang, *et al.*, "Retentive network: A successor to transformer for large language models," *arXiv preprint arXiv:2307.08621*, 2023.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] Y. Sun, L. Dong, B. Patra, *et al.*, "A length-extrapolatable transformer," *arXiv preprint arXiv:2212.10554*, 2022.

[15] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[16] G. Wichern, J. Antognini, M. Flynn, *et al.*, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[17] F. Yu, S. Zhang, Y. Fu, *et al.*, "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6167–6171.

[18] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[19] S. Watanabe, T. Hori, S. Karita, *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[20] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006, Bethesda, MD, USA, May 1-4, 2006, Revised Selected Papers 3*, Springer, 2006, pp. 309–322.

Rotation Invariant Spatio-Spectral Total Variation for Hyperspectral Image Denoising

Shingo Takemoto* and Shunsuke Ono*
* Institute of Science Tokyo, Japan
E-mail: takemoto.s.e908@m.isct.ac.jp; ono.s.5af2@m.isct.ac.jp
Tel/Fax: +80-45-924-5089

***Abstract*—We propose a novel regularization function, named Rotation-Invariant Spatio-Spectral Total Variation (RISSTV), for hyperspectral (HS) image denoising. Spatio-Spectral Total Variation (SSTV) and its extended methods, defined using the second-order spatio-spectral differences, are known as popular regularization approaches that capture the HS image-specific spatio-spectral piecewise smoothness for effectively removing noise from HS images. However, these methods lack rotation-invariance and tend to corrupt round structures and oblique edges on HS images. To address this issue, we assign rotation-invariance to TV and establish RISSTV, which can accurately recover the detailed structures in HS images. Furthermore, we formulate the HS image denoising problem as a convex optimization problem that includes RISSTV and develop an efficient algorithm based on a preconditioned primal-dual splitting method to solve this problem efficiently. Finally, we demonstrate the effectiveness of RISSTV compared with existing HS image regularization models through mixed noise removal experiments.**

I. Introduction

Hyperspectral (HS) images contain extensive spectral data across a broad range of wavelengths, spanning from near-infrared to ultraviolet. This spectral information can visualize phenomena and materials that are indistinguishable in traditional RGB images, making HS imaging crucial for applications in fields such as agriculture, astronomy, biotechnology, and mineralogy [1]–[4]. However, HS images are unavoidably contaminated by various types of noise during the measurement process, including thermal noise, quantization noise, and shot noise, which can significantly degrade the accuracy of subsequent processing like anomaly detection [5], [6], classification [7]–[9], and unmixing [10], [11]. Therefore, HS image denoising is an essential preprocessing task [12]–[15].

In this field of HS image denoising, the Spatio-Spectral Total Variation (SSTV) model [16] has emerged as a prominent regularization technique, capturing the inherent characteristics of HS images. SSTV leverages the ℓ_1-norm of the second-order spatio-spectral differences, with the first-order spatial differences are computed after spectral ones. Thanks to this formula, SSTV helps to effectively remove noise while maintaining the spatio-spectral consistency of HS images, and has been extensively employed in state-of-the-art HS image denoising methods [17]–[23]. One of the successful extensions of SSTV is Hybrid Spatio-Spectral Total Variation (HSSTV) [20] proposed by Takeyama *et al.*, which integrates the first-order spatial differences into SSTV. Directly promoting of spatial piecewise smoothness removes similar noise (called artifacts) between adjacent bands that SSTV tends to retain. However, SSTV and its extended methods, including HSSTV, can corrupt the round structures and oblique edges in HS images due to their lack of rotation-invariance. This raises the question: *Could we design SSTV-type functions with the rotation-invariance so that noise can be removed while preserving the more detailed structure of HS images?*

In this paper, we propose a denoising method for HS images using a newly introduced *Rotation Invariant Spatio-Spectral Total Variation* model (RISSTV). RISSTV is built upon a theoretical framework proposed by Condat [24] that significantly improves the rotation-invariance of TV for grayscale images. The rotation-invariance is an important property for restoring round structures and oblique edges because the TV value is independent of direction. The main contributions of this article are listed below:

- We design a novel regularization formulation, namely RISSTV. RISSTV is designed to combine two types of TVs, consisting of the second-order spatio-spectral differences and the first-order spatial differences, thus effectively removing noise and suppressing artifacts. Furthermore, since the two types of TVs are defined to be rotation-invariant, RISSTV can preserve the detailed structure on HS images, including round structures and oblique edges, while removing noise.
- We formulate the HS image denoising problem as a constrained convex optimization problem that includes RISSTV. To simplify parameter adjustment, we impose the terms that characterize the ara-fidelity and sparse noise as convex constraints instead of adding part of the objective function.
- To solve our optimization problem for HS image denoising, we develop an efficient algorithm based on a preconditioned primal-dual splitting method (P-PDS) [25] with an operator norm-based stepsize selection method [26]. Unlike the standard PDS [27], [28], our proposed algorithm can automatically determine the appropriate stepsizes based on the problem structure.

Finally, we demonstrate the effectiveness of the proposed method by comparing it with state-of-the-art HS image regularization models through HS image denoising experiments.

II. Preliminaries

A. Notations

Throughout this paper, we denote vectors and matrices by the boldface lowercase letters (e.g., $\mathbf{x}$) and boldface capital letters (e.g., $\mathbf{X}$), respectively. We treat an HS image, denoted by $\mathbf{u}$ with N_1 vertical pixels, N_2 horizontal pixels, and N_3 bands. We denotes the total number of cube data elements by $N = N_1N_2N_3$. For a matrix data $\mathbf{x} \in \mathbb{R}^{N_1N_2}$, the value of the location (n_1, n_2) in the domain $\{1, \dots, N_1\} \times \{1, \dots, N_2\}$ is denoted by $[\mathbf{X}]_{n_1,n_2}$. The ℓ_1-norm and the ℓ_2-norm of a vector $\mathbf{x} \in \mathbb{R}^N$ are defined as $\|\mathbf{x}\|_1 := \sum_{n=1}^N |x_n|$ and $\|\mathbf{x}\|_2 := \sqrt{\sum_{n=1}^N x_n^2}$, respectively, where x_n represents the n-th entry of $\mathbf{x}$. For an HS image $\mathbf{u} \in \mathbb{R}^N$, let $\mathbf{D}_v \in \mathbb{R}^{N\times N}$, $\mathbf{D}_h \in \mathbb{R}^{N\times N}$, and $\mathbf{D}_s \in \mathbb{R}^{N\times N}$ be the forward difference operators in the horizontal, vertical, and spectral directions, respectively, and the boundary condition is the Neumann boundary. Here, spatial difference operator is denoted by $\mathbf{D} := \begin{pmatrix}\mathbf{D}_v^\top & \mathbf{D}_h^\top\end{pmatrix}^\top \in \mathbb{R}^{2N\times N}$. Other notations will be introduced as needed.

B. Hybrid Spatio-Spectral Total Variation (HSSTV) [20]

For an HS image $\mathbf{u} \in \mathbb{R}^N$, HSSTV [20] is defined as

$$\mathrm{HSSTV}(\mathbf{u}) := \left\| \begin{matrix} \mathbf{D}\mathbf{D}_s\mathbf{u} \\ \omega\mathbf{D}\mathbf{u} \end{matrix} \right\|_{1,p}, \tag{1}$$

where $\|\cdot\|_{1,p}$ is the mixed $\ell_{1,p}$ norm, and $\omega \geq 0$ is a hyperparameter, p is assumed to be 1 or 2, i.e., the ℓ_1 norm ($\|\cdot\|_{1,1} = \|\cdot\|_1$) or the mixed $\ell_{1,2}$ norm, respectively. In HSSTV, the second-order spatio-spectral differences and the first-order spatial differences capture the spatio-spectral piecewise smoothness and directly the spatial piecewise smoothness, respectively. In addition, the first-order spatial differences suppress noise-like artifacts that are produced by only promoting the spatio-spectral piecewise smoothness.

C. Condat's rotation-invariant TV [24]

A new TV with rotation-invariance proposed by Condat [24] for a grayscale image $\mathbf{x} \in \mathbb{R}^{N_1N_2}$ is defined in the dual domain as follows: for all $n_1 = 1, \dots, N_1$, $n_2 = 1, \dots, N_2$,

$$\mathrm{TV_c}(\mathbf{x}) := \max_{\mathbf{y}\in\mathbb{R}^{2N_1N_2}} \langle \mathbf{D}_g\mathbf{x}, \mathbf{y}\rangle \text{ s.t. } \begin{cases} \|[\mathbf{L}_\updownarrow\mathbf{y}]_{n_1,n_2}\|_1 \leq 1, \\ \|[\mathbf{L}_\leftrightarrow\mathbf{y}]_{n_1,n_2}\|_1 \leq 1, \\ \|[\mathbf{L}_\bullet\mathbf{y}]_{n_1,n_2}\|_1 \leq 1, \end{cases} \tag{2}$$

where $\mathbf{D}_g \in \mathbb{R}^{2N_1N_2\times N_1N_2}$ is the spatial difference operator for the grayscale image and $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ is the dual image pair. For the definition of the three operators $\mathbf{L}_\updownarrow$, $\mathbf{L}_\leftrightarrow$, $\mathbf{L}_\bullet$, see Eqs. (9)-(14) in [24]. The dual images $[\mathbf{y}_1]_{n_1,n_2}$, $[\mathbf{y}_2]_{n_1,n_2}$, like $[\mathbf{D}_v\mathbf{x}]_{n_1,n_2}$, $[\mathbf{D}_h\mathbf{x}]_{n_1,n_2}$, can be located at the points $(n_1+\frac{1}{2}, n_2)$, $(n_1, n_2+\frac{1}{2})$, respectively. The three operators $\mathbf{L}_\updownarrow$, $\mathbf{L}_\leftrightarrow$, $\mathbf{L}_\bullet$ correct this half-pixel shift and interpolate bilinearly the dual image pair $\mathbf{y}$ onto the grids $(n_1+\frac{1}{2}, n_2)$, $(n_1, n_2+\frac{1}{2})$, (n_1, n_2), respectively. The dual image pair $\mathbf{y}$ satisfies the constraints on their respective grids, then Condat's TV has the rotation-invariant property. The dual formulation Eq. (2) can be rewritten into the equivalent primal formulation as

$$\begin{aligned} \mathrm{TV_c}(\mathbf{x}) = \min_{\mathbf{w}_\updownarrow,\mathbf{w}_\leftrightarrow,\mathbf{w}_\bullet\in\mathbb{R}^{2N_1N_2}} & \|\mathbf{w}_\updownarrow\|_{1,2} + \|\mathbf{w}_\leftrightarrow\|_{1,2} + \|\mathbf{w}_\bullet\|_{1,2} \\ \text{s.t. } & \mathbf{L}_\updownarrow^\top\mathbf{w}_\updownarrow + \mathbf{L}_\leftrightarrow^\top\mathbf{w}_\leftrightarrow + \mathbf{L}_\bullet^\top\mathbf{w}_\bullet = \mathbf{D}_g\mathbf{x}, \end{aligned} \tag{3}$$

where $\|\cdot\|_{1,2}$ is the mixed $\ell_{1,2}$ norm grouping the vertical and horizontal directions. The three vectors $\mathbf{w}_\updownarrow$, $\mathbf{w}_\leftrightarrow$, $\mathbf{w}_\bullet$ are viewed as gradients on the grids $(n_1+\frac{1}{2}, n_2)$, $(n_1, n_2+\frac{1}{2})$, (n_1, n_2), respectively. For a more compact form, let the linear operator $\mathbf{L}_g = \begin{pmatrix}\mathbf{L}_\updownarrow^\top & \mathbf{L}_\leftrightarrow^\top & \mathbf{L}_\bullet^\top\end{pmatrix}^\top \in \mathbb{R}^{6N_1N_2\times 2N_1N_2}$, and the vector $\mathbf{w} = \begin{pmatrix}\mathbf{w}_\updownarrow & \mathbf{w}_\leftrightarrow & \mathbf{w}_\bullet\end{pmatrix}^\top \in \mathbb{R}^{6N_1N_2\times 2N_1N_2}$. Then Eq. (3) can be written as

$$\mathrm{TV_c}(\mathbf{x}) = \min_{\mathbf{w}\in\mathbb{R}^{6N_1N_2}} \|\mathbf{w}\|_{1,2} \quad \text{s.t. } \mathbf{L}_g^\top\mathbf{w} = \mathbf{D}_g\mathbf{x}. \tag{4}$$

III. Proposed Method

A. Rotation Invariant Spatio-Spectral Total Variation

Combining the second-order spatio-spectral differences with the first-order spatial differences and having the rotation-invariant property, our RISSTV is defined as follows:

$$\begin{aligned} \mathrm{RISSTV}(\mathbf{u}) := \min_{\mathbf{w}_1,\mathbf{w}_2} & \omega\|\mathbf{w}_1\|_{1,2} + \|\mathbf{w}_2\|_{1,2} \\ \text{s.t. } & \begin{cases} \mathbf{L}^\top\mathbf{w}_1 = \mathbf{D}\mathbf{u}, \\ \mathbf{L}^\top\mathbf{w}_2 = \mathbf{D}\mathbf{D}_s\mathbf{u}, \end{cases} \end{aligned} \tag{5}$$

where $\omega \geq 0$, $\mathbf{w}_1$, $\mathbf{w}_2 \in \mathbb{R}^{6N}$ are auxiliary variables, $\mathbf{L} \in 2N \times 6N$ is linear operator formed by arranging N_3 diagonals of $\mathbf{L}_g$ defined in Eq. (4). Mainly, TV of the second-order differences related to the second term and second constraint in Eq. (5) captures the HS image-specific spatio-spectral piecewise smoothness. Associated with the first term and the first constraint, TV of the first-order spatial differences directly promotes the spatial piecewise smoothness and plays an important role in suppressing the noise-like artifacts produced by only imposing TV of the second-order spatio-spectral differences. These two types of TVs are an extension of Condat's idea and have rotation-invariance, so RISSTV can recover round structures and oblique edges in HS images more accurately than existing TV-based methods.

The parameter ω controls the relative importance of TV of the first-order spatial differences to one of the second-order spatio-spectral differences. If ω is larger, i.e., we make the direct spatial smoothness stronger on a restored HS image, RISSTV would cause over-smoothing of the detailed structures. Therefore, ω should be set to less than one. In fact, a good choice of ω is empirically around 0.01 to 0.05 for various HS images.

B. HS Image Denoising by RISSTV

An observed HS image $\mathbf{v} \in \mathbb{R}^N$ contaminated by mixed noise is modeled by

$$\mathbf{v} = \bar{\mathbf{u}} + \bar{\mathbf{s}} + \mathbf{n}, \tag{6}$$

where $\bar{\mathbf{u}}$ is a clean HS image, $\bar{\mathbf{s}}$ is a sparse noise that models outliers, and $\mathbf{n}$ represents random noise, respectively.

Based on the above observation model, we formulate the HS image denoising problem that handles RISSTV as a constrained convex optimization problem with the following form:

$$\min_{\mathbf{u},\mathbf{w}_1,\mathbf{w}_2,\mathbf{s}} \omega\|\mathbf{w}_1\|_{1,2} + \|\mathbf{w}_2\|_{1,2} \quad \text{s.t.} \begin{cases} \mathbf{u} \in R_{\underline{\mu},\bar{\mu}}, \\ \mathbf{s} \in B_{1,\eta}, \\ \mathbf{u}+\mathbf{s} \in B^{\mathbf{v}}_{2,\varepsilon}, \\ \mathbf{L}^\top \mathbf{w}_1 = \mathbf{D}\mathbf{u}, \\ \mathbf{L}^\top \mathbf{w}_2 = \mathbf{D}\mathbf{D}_s\mathbf{u}, \end{cases} \tag{7}$$

where

$$R_{\underline{\mu},\bar{\mu}} := \{\mathbf{x} \in \mathbb{R}^N |\, \underline{\mu} \le x_i \le \bar{\mu}\ (i = 1, \dots, N)\}, \tag{8}$$

$$B_{1,\eta} := \{\mathbf{x} \in \mathbb{R}^N |\, \|\mathbf{x}\|_1 \le \eta\}, \tag{9}$$

$$B^{\mathbf{v}}_{2,\varepsilon} := \{\mathbf{x} \in \mathbb{R}^N |\, \|\mathbf{x}-\mathbf{v}\|_2 \le \varepsilon\}. \tag{10}$$

The first constraint is a box constraint with $\underline{\mu} < \bar{\mu}$ which represents the dynamic range of $\mathbf{u}$. The second constraint serves as data-fidelity with the $\mathbf{v}$-centered ℓ_2-ball of the radius $\varepsilon > 0$. The third constraint characterizes sparse noise with the zero-centered ℓ_1-ball of the radius $\eta > 0$. By adding the constraints that characterize the data-fidelity and sparse noise instead of incorporating them directly into the objective function, ε and η can be determined independently of the other terms in the objective function and the constraints. This makes parameter adjustment easier, and this advantage is also addressed, for example, in [29]–[33].

C. Optimization

Using indicator functions[1] of $\{\mathbf{0}\}$, $B^{\mathbf{v}}_{2,\varepsilon}$, $B_{1,\eta}$, and $R_{\underline{\mu},\bar{\mu}}$, we rewrite Prob. (7) into an equivalent form:

$$\begin{aligned} \min_{\substack{\mathbf{u},\mathbf{w}_1,\mathbf{w}_2,\mathbf{s} \\ \mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3}} & \omega\|\mathbf{w}_1\|_{1,2} + \|\mathbf{w}_2\|_{1,2} + \iota_{R_{\underline{\mu},\bar{\mu}}}(\mathbf{u}) + \iota_{B_{1,\eta}}(\mathbf{s}) \\ & + \iota_{\{\mathbf{0}\}}(\mathbf{y}_1) + \iota_{\{\mathbf{0}\}}(\mathbf{y}_2) + \iota_{B^{\mathbf{v}}_{2,\varepsilon}}(\mathbf{y}_3) \\ \text{s.t.} & \begin{cases} \mathbf{y}_1 = \mathbf{D}\mathbf{u} - \mathbf{L}^\top \mathbf{w}_1, \\ \mathbf{y}_2 = \mathbf{D}\mathbf{D}_s\mathbf{u} - \mathbf{L}^\top \mathbf{w}_2, \\ \mathbf{y}_3 = \mathbf{u} + \mathbf{s}. \end{cases} \end{aligned} \tag{11}$$

Prob. (11) can be solved by P-PDS [25]. We show the detailed algorithm in Alg. 1. The proximity operators[2] of $\iota_{R_{\underline{\mu},\bar{\mu}}}$, $\iota_{\{\mathbf{0}\}}$, $\iota_{B^{\mathbf{v}}_{2,\varepsilon}}$, and $\|\cdot\|_{1,2}$ are calculated by

$$[\mathrm{prox}_{\gamma\iota_{R_{\underline{\mu},\bar{\mu}}}}(\mathbf{x})]_i = [P_{R_{\underline{\mu},\bar{\mu}}}(\mathbf{x})]_i = \begin{cases} \underline{\mu}, & \text{if } x_i < \underline{\mu}, \\ \bar{\mu}, & \text{if } x_i > \bar{\mu}, \\ x_i, & \text{otherwise}, \end{cases} \tag{12}$$

$$\mathrm{prox}_{\gamma\iota_{\{\mathbf{0}\}}}(\mathbf{x}) = \mathbf{0}, \tag{13}$$

$$\mathrm{prox}_{\gamma\iota_{B^{\mathbf{v}}_{2,\varepsilon}}}(\mathbf{x}) = P_{B^{\mathbf{v}}_{2,\varepsilon}}(\mathbf{x}) = \begin{cases} \mathbf{x}, & \text{if } \mathbf{x} \in B^{\mathbf{v}}_{2,\varepsilon}, \\ \mathbf{v} + \frac{\varepsilon(\mathbf{x}-\mathbf{v})}{\|\mathbf{x}-\mathbf{v}\|_2}, & \text{otherwise}, \end{cases} \tag{14}$$

$$[\mathrm{prox}_{\gamma\|\cdot\|_{1,2}}(\mathbf{x})]_i = \max\left\{1 - \frac{\gamma}{\sqrt{x^2_{dN+\tilde{i}} + x^2_{(d+1)N+\tilde{i}}}}, 0\right\} x_i, \tag{15}$$

where $\tilde{i} := ((i-1) \mod N)+1$, $d := (i-\tilde{i})/N$. The sum of $x^2_{dN+\tilde{i}}$ and $x^2_{(d+1)N+\tilde{i}}$ is intended to be the sum of elements in vertical and horizontal directions. The proximity operators of $\iota_{B_{1,\eta}}(\mathbf{s})$ can be efficiently computed by a fast ℓ_1-ball projection algorithm [34].

We set the stepsize parameters as $\gamma_{\mathbf{u}} = \frac{1}{41}$, $\gamma_{\mathbf{s}} = 1$, $\gamma_{\mathbf{w}_1} = \gamma_{\mathbf{w}_2} = \frac{1}{4}$, $\gamma_{\mathbf{y}_1} = \gamma_{\mathbf{y}_2} = \gamma_{\mathbf{y}_3} = \frac{1}{4}$ according to P-PDS [26].

Algorithm 1 P-PDS-based solver for (11)

Input: $\mathbf{u}^{(0)}, \mathbf{s}^{(0)}, \mathbf{w}_1^{(0)}, \mathbf{w}_2^{(0)}, \mathbf{y}_1^{(0)}, \mathbf{y}_2^{(0)}, \mathbf{y}_3^{(0)}$
Output: $\mathbf{u}^{(t)}$

while A stopping criterion is not satisfied **do**
  $\mathbf{u}^{(t+1)} \leftarrow P_{R_{\underline{\mu},\bar{\mu}}}\left(\mathbf{u}^{(t)} - \gamma_{\mathbf{u}}\left(\mathbf{D}^\top \mathbf{y}_1^{(t)} + \mathbf{D}_s^\top \mathbf{D}^\top \mathbf{y}_2^{(t)} + \mathbf{y}_3^{(t)}\right)\right)$
  $\mathbf{s}^{(t+1)} \leftarrow P_{B_{1,\eta}}\left(\mathbf{s}^{(t)} - \gamma_{\mathbf{s}}\mathbf{y}_3^{(t)}\right)$
  $\mathbf{w}_1^{(t+1)} \leftarrow \mathrm{prox}_{\gamma_{\mathbf{s}}\omega,\|\cdot\|_{1,2}}\left(\mathbf{w}_1^{(t)} + \gamma_{\mathbf{w}_1}\mathbf{L}\mathbf{y}_1^{(t)}\right)$
  $\mathbf{w}_2^{(t+1)} \leftarrow \mathrm{prox}_{\gamma_{\mathbf{s}},\|\cdot\|_{1,2}}\left(\mathbf{w}_2^{(t)} + \gamma_{\mathbf{w}_2}\mathbf{L}\mathbf{y}_2^{(t)}\right)$
  $\mathbf{u}' \leftarrow 2\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}$;
  $\mathbf{s}' \leftarrow 2\mathbf{s}^{(t+1)} - \mathbf{s}^{(t)}$;
  $\mathbf{w}_1' \leftarrow 2\mathbf{w}_1^{(t+1)} - \mathbf{w}_1^{(t)}$;
  $\mathbf{w}_2' \leftarrow 2\mathbf{w}_2^{(t+1)} - \mathbf{w}_2^{(t)}$;
  $\mathbf{y}_1^{(t+1)} \leftarrow \mathbf{y}_1^{(t)} + \gamma_{\mathbf{y}_1}\left(\mathbf{D}\mathbf{u}' - \mathbf{L}^\top \mathbf{w}_1'\right)$
  $\mathbf{y}_2^{(t+1)} \leftarrow \mathbf{y}_2^{(t)} + \gamma_{\mathbf{y}_2}\left(\mathbf{D}\mathbf{D}_s\mathbf{u}' - \mathbf{L}^\top \mathbf{w}_2'\right)$
  $\mathbf{y}_3' \leftarrow \mathbf{y}_3^{(t)} + \gamma_{\mathbf{y}_3}\left(\mathbf{u}' + \mathbf{s}'\right)$;
  $\mathbf{y}_3^{(t+1)} \leftarrow \mathbf{y}_3' - \gamma_{\mathbf{y}_3} P_{B^{\mathbf{v}}_{2,\varepsilon}}\left(\frac{1}{\gamma_{\mathbf{y}_3}}\mathbf{y}_3'\right)$;
  $t \leftarrow t+1$;
end while

IV. Experiments

To demonstrate the effectiveness of RISSTV, we conducted HS image denoising experiments where we compared RISSTV with three types of state-of-the-art methods. The first are TV-based methods: SSTV [16], HSSTV [20], and l_0-l_1HTV [21]. The second are tensor-based methods: Decomposable Nonlocal

[1]The indicator function of a closed convex set C is defined by $\iota_C(\mathbf{x}) := 0$, if $\mathbf{x} \in C$; ∞, otherwise.

[2]The proximity operator of index $\gamma > 0$ of a proper lower semicontinuous convex function f is defined by $\mathrm{prox}_{\gamma f}(\mathbf{x}) := \arg\min_{\mathbf{y}} f(\mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{y}\|_2^2$. The proximity operator of ι_C is the projection onto C, given by $\mathrm{prox}_{\iota_C}(\mathbf{x}) = \underset{\mathbf{y}\in C}{\mathrm{argmin}} \|\mathbf{y}-\mathbf{x}\|_2 := P_C(\mathbf{x})$.

TABLE I
MPSNRS AND MSSIMS OF ALL DENOISING RESULTS.

Image	$(\sigma, p_{\mathbf{s}})$		TDL [35]	ITSReg [36]	SSTV [19]	HSSTV1 [20]	HSSTV2 [20]	l_0-l_1HTV [21]	LRTDTV [18]	RISSTV
Jasper Ridge	(0.05, 0.05)	MPSNR	21.09	31.65	38.39	38.71	38.36	38.41	37.85	**38.82**
		MSSIM	0.7999	0.8577	0.9475	**0.9567**	0.9535	0.9463	0.9544	0.9540
	(0.1, 0.1)	MPSNR	19.85	27.86	34.03	34.22	33.94	34.29	34.63	**34.68**
		MSSIM	0.6917	0.7694	0.8923	0.9021	0.9005	0.8928	0.9086	**0.9149**
PaviaU	(0.05, 0.05)	MPSNR	22.03	31.38	38.98	39.20	39.07	39.08	34.30	**39.41**
		MSSIM	0.8159	0.8734	0.9585	**0.9638**	0.9616	0.9587	0.9006	0.9634
	(0.1, 0.1)	MPSNR	20.33	27.29	32.76	33.37	33.50	33.20	32.31	**34.04**
		MSSIM	0.7372	0.7969	0.8687	0.8847	0.8859	0.8807	0.8483	**0.9044**

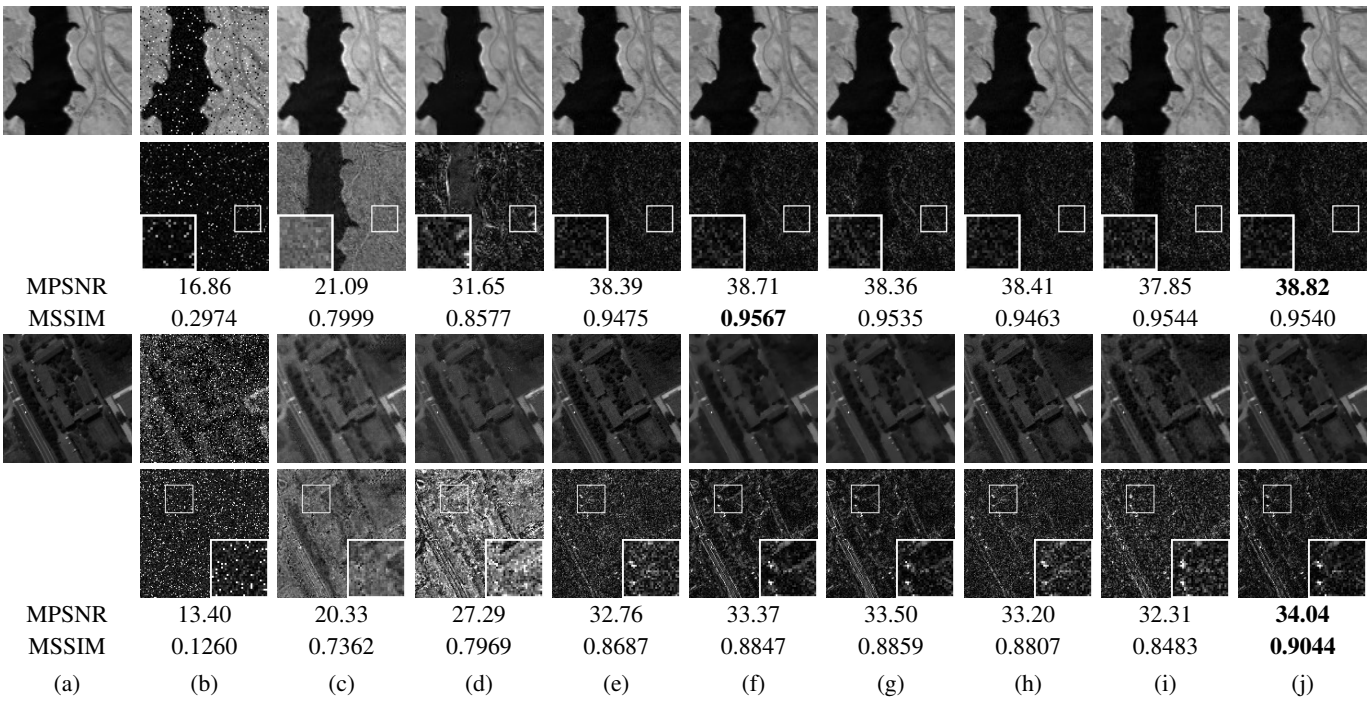


Fig. 1. Denoising results. The first and second rows are the results on Jasper Ridge with $\sigma = 0.05$, $p_{\mathbf{s}} = 0.05$ and the third and fourth rows are the results on PaviaU with $\sigma = 0.1$, $p_{\mathbf{s}} = 0.1$. (a) Ground-truth. (b) Observed noisy image. (c) TDL. (d) ITSReg. (e) SSTV. (f) HSSTV1. (g) HSSTV2. (h) l_0-l_1HTV. (i) LRTDTV. (j) RISSTV (ours).

Tensor Dictionary Learning (TDL) [35] and Intrinsic Tensor Sparsity Regularization (ITSReg) [36]. The third is a TV-tensor hybrid method, i.e., Total Variation-Regularized Low-Rank Tensor Decomposition (LRTDTV) [18]. Here, HSSTV with ℓ_1-norm and mixed $\ell_{1,2}$-norm are denoted by HSSTV1 and HSSTV2, respectively. We used *Jasper Ridge*[3] with size $100 \times 100 \times 198$ and *Pavia University*[4] (PaviaU) cropped to size $140 \times 140 \times 98$ as ground-truth HS images, and intensities were normalized to a range of $[0, 1]$. Noisy HSIs were generated based on the observation model (6). This experiment involved two noise cases. For the low-noise-intensity setting, we assign the standard deviation of the Gaussian noise σ to 0.05 and the ratio of salt-and-paper noise $p_{\mathbf{s}}$ to 0.05. For the high-noise-intensity setting, we assign σ to 0.1 and $p_{\mathbf{s}}$ to 0.1.

For a fair comparison, we replaced RISSTV in (7) with TV-based methods, i.e., SSTV, HSSTV1, HSSTV2, l_0-l_1HTV, respectively, and solved each problem by P-PDS. For TDL, ITSReg, and LRTDTV, we used the implementation published by the authors. In RISSTV, we fixed ω to 0.01 for all images and conditions. The radiuses η in (9) and ε in (10) were set to $0.93\frac{Np_{\mathbf{s}}}{2}$ and $0.93\sqrt{\sigma^2 N(1-p_{\mathbf{s}})}$, respectively. The algorithm's stopping criterion was defined as $\frac{\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|_2}{\|\mathbf{u}^{(t)}\|_2} < 1.0 \times 10^{-5}$. Quality was assessed using the mean peak signal-to-noise ratio (MPSNR) [dB]: $\frac{1}{N_3}\sum_{i=1}^{N_3} 10\log_{10}(N_1 N_2/\|\mathbf{u}_i - \bar{\mathbf{u}}_i\|_2^2)$, and the mean structural similarity index (MSSIM) [37]: $\frac{1}{N_3}\sum_{i=1}^{N_3} \mathrm{SSIM}(\mathbf{u}_i, \bar{\mathbf{u}}_i)$, where $\mathbf{u}_i$ is the i-th band of $\mathbf{u}$. where $\mathbf{u}_i$ and $\bar{\mathbf{u}}_i$ are the i-th band of the ground true HS image $\mathbf{u}$ and the estimated HS image $\bar{\mathbf{u}}$, respectively. Generally, higher MPSNR and MSSIM values are corresponding to better denoising performances.

Table I shows MPSNR [dB] and MSSIM of all the denoising results for each method. The best MPSNR and MSSIM values are highlighted in bold. Our RISSTV performed better than all other methods in most cases. Especially for MPSNR, our RISSTV is superior to all the other methods in all cases.

[3]https://rslab.ut.ac.ir/data

[4]https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

The first and second rows in Fig. 1 depict the denoising results at $\sigma = 0.05$ and $p_{\mathbf{s}} = 0.05$ in Jasper Ridge with the 99th band, and the third and fourth rows those at $\sigma = 0.1$ and $p_{\mathbf{s}} = 0.1$ in PaviaU with the 20th band. The images in the second and fourth rows show the absolute differences between the original images and each restored image in each result, multiplied by eight for visibility.[5]

In images (c) and (d) restored with the tensor-based methods TDL and ITSReg, respectively, edges and textures are restored to some extent, but the mean pixel values are significantly off due to spectral distortion. SSTV in (e) and existing TV-vased methods (HSSTV1, HSSTV2, l_0-l_1HTV, LRTDTV) in (f), (g), (h), and (i) exhibit improved structural preservation in comparison to TDL and ITSReg. However, in the case of higher noise intensity, noise and artifacts remain in the restored image by the methods using only the second-order spatio-spectral differences, i.e., SSTV, l_0-l_1HTV, and LRTDTV in (e), (h), and (i). The images in (f) and (g) restored by HSSTV1 and HSSTV2 can remove noise more effectively than these methods, but many edges and textures are present in the difference images. On the other hand, as shown in (j), RISSTV achieves the highest restoration performance. In addition, in the enlarged area of the difference images of RISSTV, edges and textures, oblique edges, do not appear much. These results suggest that RISSTV can remove noise while retaining edges and textures with high accuracy. This may be due to the fact that RISSTV has the rotation-invariant property.

V. Conclusion

We proposed a new HS image regularization model, named RISSTV, for HS image denoising. RISSTV is designed by combining the TVs of the spatial/spatio-spectral differences and making them rotation-invariant, leading to a powerful regularization model that fully recovers the spatio-spectral structures, including round structures and oblique edges, in the target HS image. The HS image denoising problem using RISSTV is formulated as a convex optimization problem and efficiently solved by an algorithm based on a preconditioned primal-dual splitting method. The experimental results on mixed noise removal show that RISSTV is sperior to several existing HS image regularization models.

Acknowledgment

This work was supported in part by JST PRESTO under Grant JPMJPR21C4 and JST AdCORP under Grant JPMJKB2307, in part by JSPS KAKENHI under Grant 22H03610, 22H00512, 23H01415 and 23K17461, and in part by Grant-in-Aid for JSPS Fellows under Grant Number JP24KJ1068.

[5]The difference images for the observed noisy images or TDL were multiplied by one or three, respectively, in order to avoid excessive white-out.

References

[1] M. Borengasser, W. S. Hungate, and R. Watkins, *Hyperspectral remote sensing: principles and applications*. CRC press, 2007.

[2] H. Grahn and P. Geladi, *Techniques and applications of hyperspectral image analysis*. John Wiley & Sons, 2007.

[3] P. S. Thenkabail, J. G. Lyon, and A. Huete, *Hyperspectral remote sensing of vegetation*. CRC press, 2016.

[4] B. Lu, P. D. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sens.*, vol. 12, no. 16, p. 2659, 2020.

[5] S. Matteoli, M. Diani, and J. Theiler, "An overview of background modeling for detection of targets and anomalies in hyperspectral remotely sensed imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2317–2336, 2014.

[6] H. Su, Z. Wu, H. Zhang, and Q. Du, "Hyperspectral anomaly detection: A survey," *IEEE Geos. Remote Sens. Mag.*, vol. 10, no. 1, pp. 64–90, 2022.

[7] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, 2017.

[8] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, 2019.

[9] N. Audebert, B. L. Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, 2019.

[10] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, *et al.*, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, 2012.

[11] W. Ma, J. M. Bioucas-Dias, T. Chan, *et al.*, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, 2014.

[12] H. Shen, X. Li, Q. Cheng, *et al.*, "Missing information reconstruction of remote sensing data: A technical review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 61–85, 2015.

[13] B. Rasti, P. Scheunders, P. Ghamisi, G. Licciardi, and J. Chanussot, "Noise reduction in hyperspectral imagery: Overview and application," *Remote Sens.*, vol. 10, no. 3, 2018.

[14] H. Shen, M. Jiang, J. Li, C. Zhou, Q. Yuan, and L. Zhang, "Coupling model- and data-driven methods for remote sensing image restoration and fusion: Improving physical interpretability," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 231–249, 2022.

[15] K. Naganuma and S. Ono, "A general destriping framework for remote sensing images using flatness constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[16] H. K. Aggarwal and A. Majumdar, "Hyperspectral image denoising using spatio-spectral total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 442–446, 2016.

[17] H. Fan, C. Li, Y. Guo, G. Kuang, and J. Ma, "Spatial–spectral total variation regularized low-rank tensor decomposition for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 6196–6213, 2018.

[18] Y. Wang, J. Peng, Q. Zhao, Y. Leung, X. Zhao, and D. Meng, "Hyperspectral image restoration via total variation regularized low-rank tensor decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1227–1243, 2018.

[19] T. Ince, "Hyperspectral image denoising using group low-rank and spatial-spectral total variation," *IEEE Access*, vol. 7, pp. 52 095–52 109, 2019.

[20] S. Takeyama, S. Ono, and I. Kumazawa, "A constrained convex optimization approach to hyperspectral image restoration with hybrid spatio-spectral regularization," *Remote Sens.*, vol. 12, no. 21, 2020.

[21] M. Wang, Q. Wang, J. Chanussot, and D. Hong, "l_0-l_1 hybrid total variation regularization and its applications on hyperspectral image mixed noise removal and compressed sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7695–7710, 2021.

[22] S. Takemoto, K. Naganuma, and S. Ono, "Graph spatio-spectral total variation model for hyperspectral image denoising," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[23] S. Takemoto and S. Ono, "Enhancing spatio-spectral regularization by structure tensor modeling for hyperspectral image denoising," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., (ICASSP)*, 2023, pp. 1–5.

[24] L. Condat, "Discrete total variation: New definition and minimization," *SIAM J. Imag. Sci.*, vol. 10, no. 3, pp. 1258–1290, 2017.

[25] T. Pock and A. Chambolle, "Diagonal preconditioning for first order primal-dual algorithms in convex optimization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 1762–1769.

[26] K. Naganuma and S. Ono, "Variable-wise diagonal preconditioning for primal-dual splitting: Design and applications," *IEEE Trans. Signal Process.*, vol. 71, pp. 3281–3295, 2023.

[27] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[28] L. Condat, "A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms," *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 460–479, 2013.

[29] M. V. Afonso, J. M. Bioucas-Dias, and M. Figueiredo, "An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, 2011.

[30] G. Chierchia, N. Pustelnik, J.-C. Pesquet, and B. Pesquet-Popescu, "Epigraphical projection and proximal tools for solving constrained convex optimization problems," *Signal Image Video Process.*, vol. 9, no. 8, pp. 1737–1749, 2015.

[31] S. Ono and I. Yamada, "Signal recovery with certain involved convex data-fidelity constraints," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6149–6163, 2015.

[32] S. Ono, "Primal-dual plug-and-play image restoration," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1108–1112, 2017.

[33] S. Ono, "Efficient constrained signal reconstruction by randomized epigraphical projection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., (ICASSP)*, 2019, pp. 4993–4997.

[34] L. Condat, "Fast projection onto the simplex and the ℓ_1 ball," *Math. Program.*, vol. 158, no. 1, pp. 575–585, 2016.

[35] Y. Peng, D. Meng, Z. Xu, C. Gao, Y. Yang, and B. Zhang, "Decomposable nonlocal tensor dictionary learning for multispectral image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 2949–2956.

[36] Q. Xie, Q. Zhao, D. Meng, *et al.*, "Multispectral images denoising by intrinsic tensor sparsity regularization," in *Proc. IEEE conf. comput. vis. pattern recognit. (CVPR)*, 2016, pp. 1692–1700.

[37] W. Zhou, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

Hyperspectral Unmixing with Row-Sparsity Enhancement: A Difference-of-Convex Approach

Kazuki Naganuma* and Shunsuke Ono*
* Institute of Science Tokyo, Tokyo
E-mail: naganuma.k.491a@m.isct.ac.jp; ono.s.5af2@m.isct.ac.jp

***Abstract*—This paper proposes a new hyperspectral unmixing method that enhances the row sparsity of an abundance matrix via a difference-of-convex (DC) approach. One promising unmixing technique is to promote the row sparsity by designing and solving optimization problems. However, existing methods are unable to both fully enhance the row sparsity and guarantee the convergence of their optimization algorithms, resulting in performance degradation. To address this limitation, we introduce a DC approach to unmixing. First, we design a DC function that promotes the row sparsity while preserving the energies of some row vectors of an abundance matrix. Then, incorporating the DC function, we formulate unmixing as a constrained DC optimization problem. In addition, we develop an algorithm to solve the optimization problem based on the proximal linearized DC algorithm and the preconditioned primal-dual splitting algorithm and guarantee its convergence to a critical point of the problem. Experiments on synthetic and real HS images show that our method is superior to existing methods.**

I. Introduction

Hyperspectral (HS) images are cube data consisting of rich spectral information and have various applications, such as ecology, mineralogy, and agriculture [1]–[5]. Due to the trade-off between spatial and spectral resolution, HS sensors do not have sufficient spatial resolutions, resulting in containing multiple components (called endmembers) in a pixel [6], which is referred to as a mixel. The process of decomposing the mixel into endmembers and their abundances is called unmixing. Unmixing has been actively studied in the remote sensing field because it is essential for HS image analysis [7]–[9].

One major category of unmixing tasks is non-blind unmixing, which estimates abundances using an endmember library consisting of all possible pre-collected spectral signatures that may appear in a target HS image. This indicates that the HS image often contains a smaller number of spectral signatures of the endmember library, i.e., many of the corresponding abundances are zeros. Therefore, the matrix created by arranging the abundances corresponding to each endmember in the row direction is exactly row sparse (we call the matrix *abundance matrix*).

To capture the exact row sparsity, non-blind unmixing methods often solve optimization problems that incorporate a function promoting the row sparsity[1]. Many of these methods adopt a mixed $\ell_{1,2}$-type norm [10]–[15]. Since a mixed $\ell_{1,2}$-type norm is convex, these methods formulate unmixing as a convex optimization problem and thus can obtain stable results. As a more advanced method, the authors in [16] employ a mixed $\ell_{0,2}$ pseudo norm, a nonconvex function that more appropriately models row sparsity.

However, the existing methods have the following limitations. The convex approach is weak in enhancing the row sparsity because minimizing the mixed $\ell_{1,2}$-type norm of an abundance matrix almost equally reduces all the ℓ_2 norms of the row vectors of the matrix. In the nonconvex approach, the obtained abundance matrix is not fully row sparse because the sequence generated by the optimization algorithm does not converge to a solution of the optimization problem involving a mixed $\ell_{0,2}$ pseudo norm. Therefore, to improve the unmixing performance, the following question arises: *Could we fully enhance the row sparsity while guaranteeing the convergence of the optimization algorithm?*

Based on the above discussion, we introduce a difference-of-convex (DC) approach to unmixing. The DC handles the difference of convex functions and convex functions, and thus provides the improvements of many signal processing performance [17]–[21], including the enhancement of (group) sparsity. In addition, a DC algorithm generates the sequence that converges to a critical point of a DC optimization problem. Our contributions are listed as follows.

- We design a new DC function that promotes row sparsity while preserving the energies of some row vectors of a matrix.
- We formulate unmixing as a constrained DC optimization problem that incorporates the DC function.
- Based on the proximal linearized DC algorithm (PLDC) [22] and the preconditioned primal-dual splitting algorithm (P-PDS) [23], [24], we develop an algorithm that generates the sequence converging to a critical point of the DC optimization problem.

Finally, we illustate the effectiveness of our method through unmixing experiments using both synthetic and real datasets.

II. Preliminaries

A. Notations

Vectors and matrices are respectively denoted by lowercase and capitalized boldface letters (e.g., $\mathbf{x}$ and $\mathbf{X}$), and the element at the i-th row and j-th column of matrix $\mathbf{X}$ is

[1]A deep neural network-based method has also been proposed [25]. However, as experimentally shown in Section IV, row sparsity-based methods are more suitable for non-blind unmixing.

denoted by $X_{i,j}$ or $[\mathbf{X}]_{i,j}$. For matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$, the inner product $\langle \mathbf{X}, \mathbf{Y}\rangle$, the ℓ_1-norm $\|\mathbf{X}\|_1$, the Frobenius norm $\|\mathbf{X}\|_F$, the mixed $\ell_{1,2}$-norm grouped by row $\|\mathbf{X}\|_{1,2}$ are defined by $\langle \mathbf{X}, \mathbf{Y}\rangle = \sum_{i,j} X_{i,j}Y_{i,j}$, $\|\mathbf{X}\|_1 = \sum_{i,j} |X_{i,j}|$, $\|\mathbf{X}\|_F = \sqrt{\langle \mathbf{X}, \mathbf{X}\rangle}$, and $\|\mathbf{X}\|_{1,2} = \sum_i \sqrt{\sum_j X_{i,j}^2}$, respectively. Let $\mathbf{G} : \mathbb{R}^{m_1\times n_1} \to \mathbb{R}^{m_2\times n_2}$ be a linear operator. A linear operator $\mathbf{G}^* : \mathbb{R}^{m_2\times n_2} \to \mathbb{R}^{m_1\times n_1}$ is called the adjoint operator of $\mathbf{G}$ if it satisfies $\langle \mathbf{G}(\mathbf{X}), \mathbf{Y}\rangle = \langle \mathbf{X}, \mathbf{G}^*(\mathbf{Y})\rangle$ for any $\mathbf{X} \in \mathbb{R}^{m_1\times n_1}$ and $\mathbf{Y} \in \mathbb{R}^{m_2\times n_2}$.

B. Weighted Sorted Mixed $\ell_{1,2}$ Norm [26]

For a matrix $\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_m\end{bmatrix}^\top \in \mathbb{R}^{m\times n}$, the weighted sorted mixed $\ell_{1,2}$ norm of $\mathbf{A}$ with the weight $\mathbf{w} \in \mathbb{R}^m$ is defined by

$$\phi_{\mathbf{w}}(\mathbf{A}) = \sum_{j=1}^{m} w_j \|\mathbf{a}_{\pi_{\mathbf{A}}(j)}\|_2, \tag{1}$$

where $\pi_{\mathbf{A}}(j)$ ($\forall j = 1, \ldots, m$) is the index corresponding to the j-th vector of the vectors obtained by sorting $\mathbf{a}_1, \ldots, \mathbf{a}_m$ in descending order in an ℓ_2 norm, i.e., $\|\mathbf{a}_{\pi_{\mathbf{A}}(1)}\|_2 \geq \cdots \geq \|\mathbf{a}_{\pi_{\mathbf{A}}(m)}\|_2$. Note that if the weight vector $\mathbf{w}$ satisfy $w_1 \geq \ldots \geq w_m$, $\phi_{\mathbf{w}}$ is convex.

C. Proximal Linearized Difference-of-Convex Algorithm [22]

Let $g, h : \mathbb{R}^{m\times n} \to (-\infty, \infty]$ be proper lower-semicontinuous convex functions and $g - h$ is bounded below. We consider *difference-of-convex* minimization (DC) problems of the form:

$$\min_{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}} g(\mathbf{X}) - h(\mathbf{X}). \tag{2}$$

First, we introduce the *proximity operator* of g with a parameter $\gamma > 0$ as follows:

$$\mathrm{prox}_{\gamma g}(\mathbf{X}) = \operatorname*{argmin}_{\mathbf{Y}} g(\mathbf{Y}) + \frac{1}{2\gamma}\|\mathbf{X} - \mathbf{Y}\|_F^2. \tag{3}$$

We also define the *subdifferential* of h by

$$\partial h(\mathbf{X}) = \{\mathbf{U} \in \mathbb{R}^{m\times n} \mid \forall \mathbf{Y}, \langle \mathbf{Y} - \mathbf{X}, \mathbf{U}\rangle + h(\mathbf{X}) \leq h(\mathbf{Y})\}. \tag{4}$$

Then, PLDC solves Prob. (2) by the following procedures: for a given $\mu > 0$, iterate

$$\left\lfloor \begin{array}{l} \text{Find } \mathbf{Z}^{(t+1)} \in \partial h(\mathbf{X}^{(t)}); \\ \mathbf{X}^{(t+1)} \leftarrow \mathrm{prox}_{\mu g}(\mathbf{X}^{(t)} + \mu\mathbf{Z}^{(t+1)}); \\ t = t + 1; \end{array} \right. \tag{5}$$

We summarize the theoretical results for the convergence of PLDC as follows:

Theorem II.1 (Convergence of the sequence generated by PLDC [22])**.** *Let $\{\mathbf{X}^t\}_{t\in\mathbb{N}}$ be generated by Algorithm* (5)*. Then, it is satisfied that $\lim_{t\to+\infty} \|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\|_F = 0$. Furthermore, suppose that $\{\mathbf{X}^{(t)}\}_{t\in\mathbb{N}}$ is bounded, and let $\mathbf{X}^*$ and $\mathbf{Z}^*$ be cluster-point of $\mathbf{X}^{(t)}$ and $\mathbf{Z}^{(t)}$, respectively. Then, every cluster-point of $\{\mathbf{X}^{(t)}\}_{t\in\mathbb{N}}$ is a critical point of the objective function in Prob.* (2)*, i.e., $\mathbf{Z}^* \in \partial g(\mathbf{X}^*) \cap \partial h(\mathbf{X}^*)$.*

Algorithm 1 An algorithm for solving Prob. (13)

Input: $\mathbf{V}$, $\mathbf{E}$, λ, μ, ε, and η
Output: $\mathbf{A}^{(k)}$, $\mathbf{S}^{(k)}$

Initialize $\mathbf{A}^{(0)}$, $\mathbf{S}^{(0)}$, $\mathbf{Y}_1^{(0)}$, $\mathbf{Y}_2^{(0)}$, $\mathbf{Y}_3^{(0)}$, $\mathbf{Y}_4^{(0)}$, and $\mathbf{Y}_5^{(0)}$;
Set $\gamma_1 = \frac{1}{10+\sigma_1(\mathbf{E})^2}$, $\gamma_2 = \frac{1}{2}$, and $\gamma_3 = \frac{1}{2}$;
while until $\frac{\|(\mathbf{A}^{(k+1)}+\mathbf{S}^{(k+1)})-(\mathbf{A}^{(k)}+\mathbf{S}^{(k)})\|_F}{\|\mathbf{A}^{(k)}+\mathbf{S}^{(k)}\|_F} \leq 10^{-5}$ **do**
  $\widetilde{\mathbf{A}} \leftarrow \mathbf{A}^{(k)} - \gamma_1(\mathbf{Y}_1^{(k)} + \mathbf{D}^*(\mathbf{Y}_2^{(k)}) + \mathbf{E}^\top\mathbf{Y}_3^{(k)} + \mathbf{Y}_4^{(k)})$;
  $\mathbf{A}^{(k+1)} \leftarrow \mathrm{prox}_{\gamma_1\iota_{[0,1]^{m\times n}}}(\widetilde{\mathbf{A}})$;
  $\mathbf{S}^{(k+1)} \leftarrow \mathrm{prox}_{\gamma_2\iota_{\mathcal{B}_{1,\eta}}}(\mathbf{S}^{(k)} - \gamma_2(\mathbf{Y}_3^{(k)} + \mathbf{Y}_5^{(k)}))$;
  $\widetilde{\mathbf{Y}}_1 \leftarrow \mathbf{Y}_1^{(k)} + \gamma_3(2\mathbf{A}^{(k+1)} - \mathbf{A}^{(k)})$;
  $\mathbf{Y}_1^{(k+1)} \leftarrow \widetilde{\mathbf{Y}}_1 - \gamma_3\mathrm{prox}_{\frac{1}{\gamma_3}\|\cdot\|_{1,2}}(\frac{\widetilde{\mathbf{Y}}_1}{\gamma_3})$;
  $\widetilde{\mathbf{Y}}_2 \leftarrow \mathbf{Y}_2^{(k)} + \gamma_3\mathbf{D}(2\mathbf{A}^{(k+1)} - \mathbf{A}^{(k)})$;
  $\mathbf{Y}_2^{(k+1)} \leftarrow \widetilde{\mathbf{Y}}_2 - \gamma_3\mathrm{prox}_{\frac{\lambda}{\gamma_3}\|\cdot\|_1}(\frac{\widetilde{\mathbf{Y}}_2}{\gamma_3})$;
  $\widetilde{\mathbf{Y}}_3 \leftarrow \mathbf{Y}_3^{(k)} + \gamma_3(\mathbf{E}(2\mathbf{A}^{(k+1)} - \mathbf{A}^{(k)}) + (2\mathbf{S}^{(k+1)} - \mathbf{S}^{(k)}))$;
  $\mathbf{Y}_3^{(k+1)} \leftarrow \widetilde{\mathbf{Y}}_3 - \gamma_3\mathrm{prox}_{\frac{1}{\gamma_3}\iota_{\mathcal{B}_{F,\varepsilon}^{\mathbf{V}}}}(\frac{\widetilde{\mathbf{Y}}_3}{\gamma_3})$;
  $\widetilde{\mathbf{Y}}_4 \leftarrow \mathbf{Y}_4^{(k)} + \gamma_3(2\mathbf{A}^{(k+1)} - \mathbf{A}^{(k)})$;
  $\mathbf{Y}_4^{(k+1)} \leftarrow \widetilde{\mathbf{Y}}_4 - \gamma_3\mathrm{prox}_{\frac{1}{2\mu_1\gamma_3}\|\cdot-\mathbf{A}'\|_2^2}(\frac{\widetilde{\mathbf{Y}}_4}{\gamma_3})$;
  $\widetilde{\mathbf{Y}}_5 \leftarrow \mathbf{Y}_5^{(k)} + 2(\mathbf{S}^{(k+1)} - \mathbf{S}^{(k)})$;
  $\mathbf{Y}_5^{(k+1)} \leftarrow \widetilde{\mathbf{Y}}_5 - \gamma_3\mathrm{prox}_{\frac{1}{2\mu_1\gamma_3}\|\cdot-\mathbf{S}'\|_2^2}(\frac{\widetilde{\mathbf{Y}}_5}{\gamma_3})$;
  $k \leftarrow k + 1$;

III. Proposed Method

A. Row-Sparsity Enhancement Weighted Sorted Function

Let a weight $\mathbf{w} \in \mathbb{R}^m$ be for $k \in \{1, \ldots, m\}$

$$w_j = \begin{cases} 1, & \text{if } j \leq k; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we define a new weighted sorted function with the weight $\mathbf{w}$ by

$$\Phi_{\mathbf{w}}(\mathbf{A}) = \|\mathbf{A}\|_{1,2} - \phi_{\mathbf{w}}(\mathbf{A}) = \sum_{j=k+1}^{m} w_j\|\mathbf{a}_{\pi_{\mathbf{A}}(j)}\|_2. \tag{7}$$

By minimizing Eq. (7), we can promote the row sparsity of $\mathbf{A}$ while preserving the energies of the top k vectors $\mathbf{a}_{\pi_{\mathbf{A}}(1)}, \ldots, \mathbf{a}_{\pi_{\mathbf{A}}(k)}$. Since the weight $\mathbf{w}$ in (6) satisfies $w_1 \geq \cdots \geq w_m$, $\phi_{\mathbf{w}}$ in (7) is convex. Therefore, $\Phi_{\mathbf{w}}$ is a DC function. In addition, $\Phi_{\mathbf{w}}$ is bounded below by 0, i.e., $\Phi_{\mathbf{w}}(\mathbf{A}) \geq 0$ for any $\mathbf{A} \in \mathbb{R}^{m\times n}$.

B. Problem Formulation

Let $\mathbf{E} \in \mathbb{R}^{l\times m}$, $\bar{\mathbf{A}} \in \mathbb{R}^{m\times n}$, $\bar{\mathbf{S}} \in \mathbb{R}^{l\times n}$, and $\mathbf{N} \in \mathbb{R}^{l\times n}$ be an endmember library, a true abundance matrix, sparse noise, and Gaussian noise, respectively. Then, the observed HS image $\mathbf{V}$ is modeled according to the linear mixing model as

$$\mathbf{V} = \mathbf{E}\bar{\mathbf{A}} + \bar{\mathbf{S}} + \mathbf{N}. \tag{8}$$

TABLE I
SREs AND RMSEs IN THE EXPERIMENTS.

Images	Noise σ / $p_{\mathbf{S}}$	Metrics	Methods							
			CLSUnSAL [10]	JSTV [11]	LGSU [13]	MdLRR [14]	RSSUn-TV [16]	UnDIP [25]	**Ours** ($k = 0.1m$)	**Ours** ($k = 0.2m$)
Synth	0.1 / 0	SRE	3.22	5.19	5.78	6.37	3.14	-0.11	**11.91**	11.90
		RMSE	0.0779	0.0621	0.0580	0.0543	0.0787	0.1143	**0.0286**	0.0287
	0.1 / 0.05	SRE	1.60	5.10	2.41	2.40	0.73	0.02	**11.58**	11.40
		RMSE	0.0939	0.0628	0.0855	0.0857	0.1039	0.1126	**0.0298**	0.0304
Real	0.1 / 0	SRE	8.27	3.67	6.91	7.02	6.63	0.24	**9.47**	**9.47**
		RMSE	0.0469	0.0797	0.0549	0.0542	0.0566	0.1183	**0.0409**	**0.0409**
	0.1 / 0.05	SRE	6.55	3.56	3.70	2.88	3.66	0.44	**8.14**	**8.14**
		RMSE	0.0572	0.0807	0.0794	0.0873	0.0798	0.1156	**0.0476**	**0.0476**

Based on Eq. (8), we formulate an unmixing problem as the following constrained DC minimization problem:

$$\min_{\mathbf{A},\mathbf{S}} \Phi_{\mathbf{w}}(\mathbf{A}) + \lambda\|\mathbf{D}(\mathbf{A})\|_1 \quad \text{s.t. } \mathbf{A} \in [0,1]^{m\times n},\ \mathbf{E}\mathbf{A}+\mathbf{S} \in \mathcal{B}^{\mathbf{V}}_{F,\varepsilon},\ \mathbf{S} \in \mathcal{B}_{1,\eta}, \tag{9}$$

where $\lambda > 0$ is a balancing parameter and $\mathbf{D}$ is a spatial difference operator. The first term promotes the row sparsity of $\mathbf{A}$. The second term captures the spatial piecewise smoothness of $\mathbf{A}$. The first constraint guarantees that abundances are within $[0, 1]$. Note that we do not explicitly adopt the abundance sum-to-one constraint because it tends to be a strong assumption for LMM-based unmixing in real-world situations due to spectral variabilities [7]. The second constraint serves as data-fidelity to the observed HS image $\mathbf{V}$ with the Frobenius norm ball $\mathcal{B}^{\mathbf{V}}_{F,\varepsilon} := \{\mathbf{X} \mid \|\mathbf{V}-\mathbf{X}\|_F \le \varepsilon\}$ with the center $\mathbf{V}$ and radius ε. The third constraint evaluates the sparsity of $\mathbf{S}$ with the ℓ_1-norm ball $\mathcal{B}_{1,\eta} := \{\mathbf{X} \mid \|\mathbf{X}\|_1 \le \eta\}$ with radius η. Using such constraints instead of data-fidelity and sparse terms makes it easy to adjust hyperparameters since the parameters ε and η can be determined based only on noise intensity[2].

C. Optimization Algorithm

Using the indicator functions[3] of $[0,1]^{m\times n}$, $\mathcal{B}^{\mathbf{V}}_{F,\varepsilon}$, and $\mathcal{B}_{1,\eta}$, we reformulate Prob. (9) as the following equivalent problem:

$$\min_{\mathbf{A},\mathbf{S}} \|\mathbf{A}\|_{1,2} + \lambda\|\mathbf{D}(\mathbf{A})\|_1 + \iota_{[0,1]^{m\times n}}(\mathbf{A}) + \iota_{\mathcal{B}^{\mathbf{V}}_{F,\varepsilon}}(\mathbf{E}\mathbf{A}+\mathbf{S}) + \iota_{\mathcal{B}_{1,\eta}}(\mathbf{S}) - \phi_{\mathbf{w}}(\mathbf{A}). \tag{10}$$

Since $\lambda\|\mathbf{D}(\mathbf{A})\|_1$, $\iota_{[0,1]^{m\times n}}(\mathbf{A})$, $\iota_{\mathcal{B}^{\mathbf{V}}_{F,\varepsilon}}$, and $\iota_{\mathcal{B}_{1,\eta}}$ are convex and $\|\mathbf{A}\|_{1,2} - \phi_{\mathbf{w}}(\mathbf{A})$ is bounded below, the objective function in Prob. (10) is bounded below. Therefore, Prob. (10), i.e., Prob. (9) can be solved by PLDC in (5).

In what follows, we derive specific computations of our PLDC-based algorithm for solving (9).

[2]Indeed, this kind of constrained formulation has played an important role in facilitating parameter setup of signal recovery problems [27]–[31].

[3]For a given nonempty closed convex set $C \subset \mathbb{R}^{m\times n}$, the indicator function of C is defined by $\iota_C(\mathbf{X}) := 0$, if $\mathbf{X} \in C$; ∞, otherwise.

1) Step 1 of our PLDC-based algorithm for solving Prob. (9)*:* First, we need to calculate one of the subgradients of $\phi_{\mathbf{w}}$ at $\mathbf{A}^{(t)}$, i.e., $\mathbf{Z}^{(t)}_{\mathbf{A}} \in \partial\phi_{\mathbf{w}}(\mathbf{A}^{(t)})$. The following proposition helps us to calculate $\mathbf{Z}^{(t)}_{\mathbf{A}}$.

Proposition III.1. *The subdifferential of weighted sorted mixed $\ell_{1,2}$ norm $\phi_{\mathbf{w}}$ is given by*

$$\partial\phi_{\mathbf{w}}(\mathbf{A}) = \left\{ [\mathbf{u}_1\ \cdots\ \mathbf{u}_m]^\top \middle| \forall j, \mathbf{u}_{\pi_{\mathbf{A}}(j)} \in \partial w_j\|\cdot\|_2(\mathbf{a}_{\pi_{\mathbf{A}}(j)}) \right\}, \tag{11}$$

where

$$\partial w_j\|\cdot\|_2(\mathbf{a}_{\pi_{\mathbf{A}}(j)}) = \begin{cases} \frac{w_j}{\|\mathbf{a}_{\pi_{\mathbf{A}}(j)}\|_2}\mathbf{a}_{\pi_{\mathbf{A}}(j)}, & \text{if } \mathbf{a}_{\pi_{\mathbf{A}}(j)} \neq \mathbf{0}; \\ \{w_j\mathbf{z} \mid \|\mathbf{z}\|_2 \le 1\}, & \text{if } \mathbf{a}_{\pi_{\mathbf{A}}(j)} = \mathbf{0}. \end{cases} \tag{12}$$

The following gives a rough sketch of the proof of Proposition III.1: $\partial\phi_{\mathbf{w}}(\mathbf{A}^{(t)})$ satisfies Eq. (4) from the rearrangement inequality [32, Theorem 6.1]. When $\mathbf{a}_{\pi_{\mathbf{A}}(j)} = \mathbf{0}$, we recommend selecting $\mathbf{z} = \mathbf{0}$.

2) Step 2 of our PLDC-based algorithm for solving Prob. (9)*::* Let $\mathbf{A}' = \mathbf{A}^{(t)} + \mu\mathbf{Z}^{(t)}_{\mathbf{A}}$ and $\mathbf{S}' = \mathbf{S}^{(t)}$. Then, $\mathbf{A}^{(t+1)}$ and $\mathbf{S}^{(t+1)}$ are given by solving the following optimization problem:

$$\min_{\mathbf{A},\mathbf{S}} \frac{1}{2\mu}\|\mathbf{A}-\mathbf{A}'\|_2^2 + \frac{1}{2\mu}\|\mathbf{S}-\mathbf{S}'\|_2^2 + \|\mathbf{A}\|_{1,2} + \lambda\|\mathbf{D}(\mathbf{A})\|_1 + \iota_{[0,1]^{m\times n}}(\mathbf{A}) + \iota_{\mathcal{B}^{\mathbf{V}}_{F,\varepsilon}}(\mathbf{E}\mathbf{A}+\mathbf{S}) + \iota_{\mathcal{B}_{1,\eta}}(\mathbf{S}). \tag{13}$$

Prob. (13) can be solved by the preconditioned primal-dual splitting method [23] and variable-wise diagonal preconditioning method [24]. The algorithm for solving Prob. (13) is summarized in Algorithm 1. The proximity operators in Algorithm 1 are calculated as shown in [33], [34].

Remark III.1 (Convergence of our PLDC-based algorithm). The proximity operators of $\iota_{[0,1]^{m\times n}}$ and $\iota_{\mathcal{B}_{1,\eta}}$ are the projections onto $[0,1]^{m\times n}$ and $\mathcal{B}_{1,\eta}$. This indicates that for any $t \in \mathbb{N}$, $\mathbf{A}^{(t)}$ and $\mathbf{S}^{(t)}$ satisfy $\mathbf{A}^{(t)} \in [0,1]^{m\times n}$ and $\|\mathbf{S}^{(t)}\|_1 \le \eta$, i.e., $\{\mathbf{A}^{(t)}, \mathbf{S}^{(t)}\}_{t\in\mathbb{N}}$ is bounded. Therefore, from Theorem II.1, $\{\mathbf{A}^{(t)}, \mathbf{S}^{(t)}\}_{t\in\mathbb{N}}$ is guaranteed to converge to a critical point of Prob. (9).

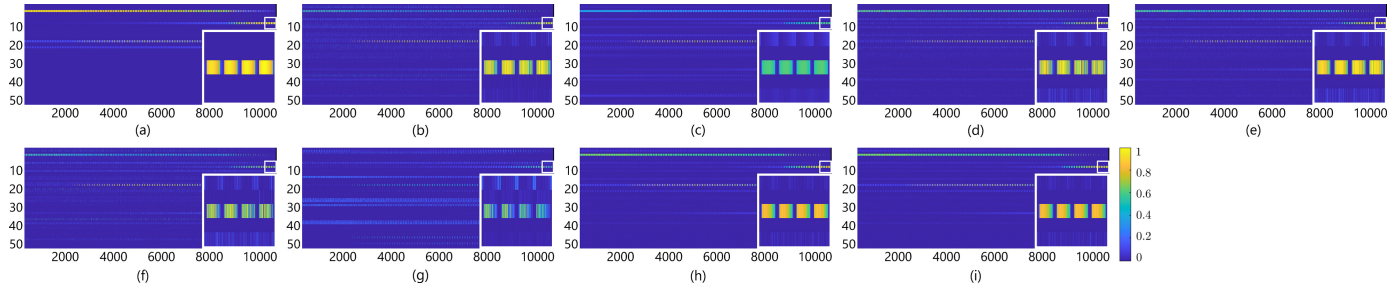


Fig. 1. Unmixing results of abundance map matrices for the experiments using the synthetic HS image with $\sigma = 0.1$ of Gaussian noise. Vertical and horizontal axes are the indices of endmembers and pixels, respectively. Expanded area shows the 8th through 12th endmembers. (a): Original abundance maps. (b): CLSUnSAL [10], (c): JSTV [11], (d): LGSU [13], (e): MdLRR [14], (f): RSSUn-TV [16], (g): UnDIP [25], (h): Ours ($k = 0.1m$), and (i): Ours ($k = 0.2m$).

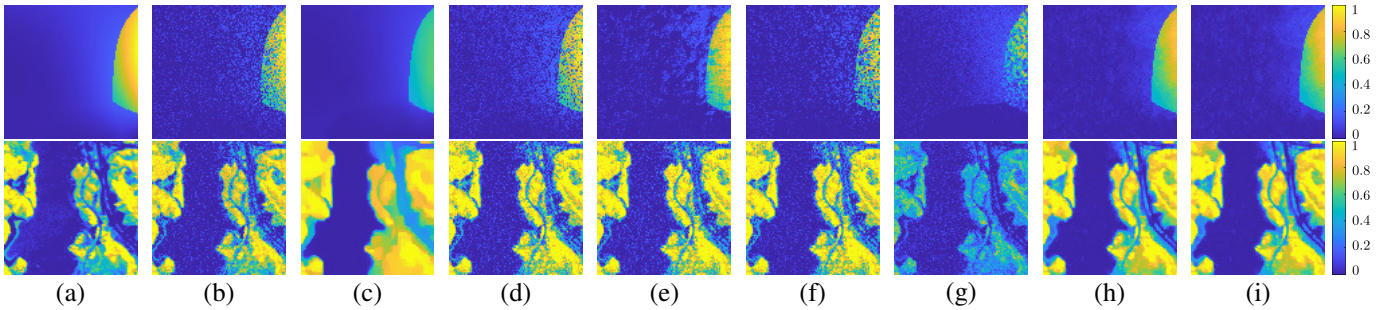


Fig. 2. Unmixing resutls of abundance maps. Top row shows the 22nd abundance maps of the experiments using the synthetic HS image with $\sigma = 0.1$ of Gaussian noise. Bottom row shows the 35th abundance maps of the experiments using the real HS image with $\sigma = 0.1$ of Gaussian noise and $p_{\mathbf{S}} = 0.05$ of sparse noise. (a): Original abundance maps. (b): CLSUnSAL [10], (c): JSTV [11], (d): LGSU [13], (e): MdLRR [14], (f): RSSUn-TV [16], (g): UnDIP [25], (h): Ours ($k = 0.1m$), and (i): Ours ($k = 0.2m$).

IV. Experiments

A. Experimental Setup

To illustrate the effectiveness of our method, we compare it with six methods, including state-of-the-art ones: four convex-based methods CLSUnSAL [10], JSTV [11], LGSU [13], and MdLRR [14], one nonconvex-based method RSSUn-TV [16], and one deep neural network-based method UnDIP [25]. The parameters of the existing methods were set to the values recommended in their references.

Experiments were performed using a synthetic image and a real image. For a synthetic image, we generate a ground-truth image of size $100 \times 100 \times 224$ using the hyperspectral data retrieval and analysis toolbox[4] and four spectral signatures from the U.S. Geological Survey Spectral Library[5]. For a real image, we use *Jasper Ridge* as a ground-truth image of size $100 \times 100 \times 187$ after removing several noisy bands and cropping the image. For both image experiments, by adding 46 spectral signatures from the USGS Spectral Library, we used $\mathbf{E}$ containing a total of $m = 50$ spectral signatures. The pixel values were normalized into $[0, 1]$. Contaminating these ground-truth videos by white Gaussian noise with 0.1 of the standard deviation σ and Salt & Pepper noise with 0 and 0.05 of the rate $p_{\mathbf{S}}$, we generated observed HS images.

For the PLDC-based algorithm, we set the stopping criterion as $\|\mathbf{A}^{(t+1)} - \mathbf{A}^{(t)}\|_F / \|\mathbf{A}^{(t)}\|_F \leq 10^{-5}$. We initialized $\mathbf{A}^{(0)}$ and $\mathbf{S}^{(0)}$ by zero matrices. The parameters λ, μ, ε, and η were set to 0.1, 10, $0.95\sigma\sqrt{(1-p_{\mathbf{S}})nl}$, and $0.45p_{\mathbf{S}}nl$, respectively. For the threshold k in Eq. (6), it is preferable to give the number of endmembers involved in a target HS image. However, in real-world situations, the number of endmembers is often unknown. Therefore, in this experiment, k was set in the following two ways based on the number of spectral signatures contained in the endmember libraries: $k = 0.1m = 5$ and $k = 0.2m = 10$. Note that the former is closer to the actual value of 4. For the quantitative evaluation of abundance maps, we used the signal reconstruction error (SRE): $\mathrm{SRE[dB]} := 10\log_{10}(\|\bar{\mathbf{A}}\|_F^2 / \|\bar{\mathbf{A}} - \hat{\mathbf{A}}\|_F^2)$, the root-mean-square error (RMSE): $\mathrm{RMSE} := \|\bar{\mathbf{A}} - \hat{\mathbf{A}}\|_F / \sqrt{mn}$.

B. Results and Discussion

Table I shows SREs and RMSEs of unmixing results. The best and second-best results are highlighted in bold and underlined. Our method was superior to all the existing methods. Compared to when $k = 5$, the performance did not drop as much when $k = 10$.

Fig. 1 shows the abundance matrices for the experiments using a synthetic HS image with $\sigma = 0.1$ of Gaussian noise. The abundance matrices of CLSUnSAL, RSSUnTV, LGSU, UnDIP, and MdLRR had slightly large values even when the corresponding endmembers did not actually exist in the ground truth HS images. The abundance matrix of JSTV had smaller values than the original abundance maps corresponding to the endmembers that exist in the ground truth HS images. These results indicate that all the existing methods did not accurately enhance the row sparsity. In contrast, our method enhanced the row sparsity of the abundance matrix.

[4]https://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Imagery_Synthesis_tools_for_MATLAB, accessed on Feb. 5, 2023

[5]https://www.usgs.gov/programs/usgs-library, accessed on Aug. 7, 2023

Fig. 2 depicts the abundance maps. All the existing methods did not estimate the abundance maps while removing noise. Specifically, the abundance maps of CLSUnSAL, RSSUnTV, LGSU, and MdLRR have residual noise. The abundance maps of JSTV were over-smooth. UnDIP does not employ a function that explicitly promotes the row sparsity; therefore, the intensities of the abundance maps are lower than the original abundance maps. In contrast, our method exactly estimates the abundance maps even when $k = 10$.

V. Conclusion

We proposed a new unmixing method via a DC approach. First, we designed a DC function that enhances the row sparsity of abundance maps and formulated unmixing as a constrained DC optimization problem that incorporates the DC function. The algorithm solving the optimization problem was developed based on PLDC and P-PDS. The experiments showed that our method can enhance the row sparsity of abundance maps and thus obtained superior results than existing methods.

Acknowledgment

This work was supported in part by JST ACT-X Grant Number JPMJAX23CJ, JST PRESTO under Grant JPMJPR21C4, and JST AdCORP under Grant JPMJKB2307, in part by JSPS KAKENHI under Grant 22H03610, 22H00512, 23H01415, and 23K17461, and in part by Grant-in-Aid for JSPS Fellows under Grant 23KJ0912.

References

[1] C-I Chang, *Hyperspectral imaging: techniques for spectral detection and classification*, vol. 1, Springer Science & Business Media, 2003.

[2] W.-K. Ma, J. M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C.-Y. Chi, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, 2014.

[3] P. S Thenkabail and J. G Lyon, *Hyperspectral remote sensing of vegetation*, CRC press, 2016.

[4] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, 2017.

[5] B. Lu, P. D. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sens.*, vol. 12, no. 16, pp. 2659, 2020.

[6] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 44–57, 2002.

[7] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, 2011.

[8] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Top. Appl. Earth. Obs. Remote Sens.*, vol. 5, no. 2, pp. 354–379, 2012.

[9] W.K. Ma, J. M. Bioucas-Dias, T.H. Chan, N. Gillis, P. Gader, A. Plaza, A. Ambikapathi, and C.Y. Chi, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, Jan. 2014.

[10] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Collaborative sparse regression for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 341–354, Jan. 2014.

[11] H. K. Aggarwal and A. Majumdar, "Hyperspectral unmixing in the presence of mixed noise using joint-sparsity and total variation," *IEEE J. Sel. Top. Appl. Earth. Obs. Remote Sens.*, vol. 9, no. 9, pp. 4257–4266, 2016.

[12] H Li, R Feng, L Wang, Y Zhong, and L Zhang, "Superpixel-based reweighted low-rank and total variation sparse unmixing for hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 629–647, 2020.

[13] X Shen, H Liu, X Zhang, K Qin, and X Zhou, "Superpixel-guided local sparsity prior for hyperspectral sparse regression unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, Art no. 6015105.

[14] L. Wu, J. Huang, and M.-S. Guo, "Multidimensional low-rank representation for sparse hyperspectral unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, Mar 2023, Art no. 5502805.

[15] K. Naganuma and S. Ono, "Toward robust hyperspectral unmixing: Mixed noise modeling and image-domain regularization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 8117–8138, 2024.

[16] J.J. Wang, T.Z. Huang, J. Huang, H.X. Dou, L.J. Deng, and X.L. Zhao, "Row-sparsity spectral unmixing via total variation," *IEEE J. Sel. Top. Appl. Earth. Obs. Remote Sens.*, vol. 12, no. 12, pp. 5009–5022, 2019.

[17] E. Esser, Y. Lou, and J. Xin, "A method for finding structured sparse solutions to nonnegative least squares problems with applications," *SIAM J. Imag. Sci.*, vol. 6, no. 4, pp. 2010–2046, 2013.

[18] Y. Lou, P. Yin, Q. He, and J. Xin, "Computing sparse representation in a highly coherent dictionary based on difference of L_1 and L_2," *J. Sci. Comput.*, vol. 64, pp. 178–196, 2015.

[19] W. Li, W. Bian, and K.-C. Toh, "Difference-of-convex algorithms for a class of sparse group ℓ_0 regularized optimization problems," *SIAM J. Optim.*, vol. 32, no. 3, pp. 1614–1641, 2022.

[20] Shunsuke Ono, Kazuki Naganuma, and Keitaro Yamashita, "Graph signal sampling under smoothness priors: A difference-of-convex approach," *arXiv preprint arXiv:2306.14634*, 2023.

[21] Koyo Sato, Kazuki Naganuma, and Shunsuke Ono, "Enhancing hyperspectral anomaly detection by difference-of-convex sparse anomaly modeling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 9921–9925.

[22] J. C. O Souza, P. R. Oliveira, and A. Soubeyran, "Global convergence of a proximal linearized algorithm for difference of convex functions," *Optim. Lett.*, vol. 10, no. 7, pp. 1529–1539, 2016.

[23] T Pock and A Chambolle, "Diagonal preconditioning for first order primal-dual algorithms in convex optimization," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1762–1769.

[24] Kazuki Naganuma and Shunsuke Ono, "Variable-wise diagonal preconditioning for primal-dual splitting: Design and applications," *IEEE Transactions on Signal Processing*, vol. 71, pp. 3281–3295, 2023.

[25] B. Rasti, B. Koirala, P. Scheunders, and P. Ghamisi, "UnDIP: Hyperspectral unmixing using deep image prior," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, Art no. 5504615.

[26] D. Brzyski, A. Gossmann, W. Su, and M. Bogdan, "Group slope – adaptive selection of groups of predictors," *J. Amer. Stat. Assoc.*, vol. 114, no. 525, pp. 419–433, 2019.

[27] M. Afonso, J.M. Bioucas-Dias, and M. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, 2011.

[28] G. Chierchia, N. Pustelnik, J.C. Pesquet, and B. Pesquet-Popescu, "Epigraphical projection and proximal tools for solving constrained convex optimization problems," *Signal, Image Video Process.*, vol. 9, no. 8, pp. 1737–1749, 2015.

[29] S. Ono and I. Yamada, "Signal recovery with certain involved convex data-fidelity constraints," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6149–6163, Nov. 2015.

[30] S. Ono, "Primal-dual plug-and-play image restoration," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1108–1112, Aug. 2017.

[31] S. Ono, "Efficient constrained signal reconstruction by randomized epigraphical projection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. IEEE, 2019, pp. 4993–4997.

[32] E. F Beckenbach and R. Bellman, *Inequalities*, vol. 30, Springer Science & Business Media, 2012.

[33] G. Chierchia, E. Chouzenoux, P. L. Combettes, and J.-C. Pesquet, "The proximity operator repository," *User's guide http://proximityoperator.net/download/guide.pdf (accessed October 3rd, 2021)*, 2020.

[34] L. Condat, "Fast projection onto the simplex and the l1 ball," *Math. Program.*, vol. 158, no. 1, pp. 575–585, Jul. 2016.

Region Aware Framework for Constrained Image Splicing Detection and Localization

Haokun Cao, Yuanman Li*, Xinyu Yang, and Xia Li
Shenzhen University, Shenzhen, China
yuanmanli@szu.edu.cn

***Abstract*—Constrained image splicing detection and localization (CISDL) is a newly proposed image forensics task. It provides two images to detect whether they have undergone splicing manipulation and localize the spliced regions on both images. Recent studies have made significant progress, but some limitations remain. Existing methods extract global features from images and compute feature correlation on them, introducing a large amount of background redundant information, which may interfere with the feature correlation results. Moreover, they calculate correlation matching on high-level coarse features, which results in the loss of many fine texture details, making it difficult to localize small manipulated regions. We propose a novel Region Aware framework that leverages a Region Aware module to perceive highly correlated regions in two images, reducing the impact of background information on feature correlation and final localization performance. In addition, we introduce the linear transformer for calculating feature correlation, which effectively models the relationship between features while addressing the high computational complexity issue of the vanilla transformer. As a result, our model can perform correlation matching on higher-resolution fine-grained features, enhancing its ability to detect and localize small splicing regions. The experimental results demonstrate that our method outperforms existing state-of-the-art methods on several benchmark datasets, proving its effectiveness.**

I. Introduction

With the development of image editing software and the growth of social media, digital images can be easily manipulated and disseminated. Unscrupulous individuals may propagate maliciously manipulated images, spreading rumors or influencing public opinion adversely. Therefore, researchers have begun studying image forensics techniques to verify image authenticity [1], [2].

Constrained image splicing detection and localization (CISDL) is a new image forensics task proposed by the Media Forensics Challenge. As shown in Fig. 1, given a probe image I_p and a potential donor image I_d, CISDL aims to find out if a region of I_p is spliced from I_d, and provides two masks M_p and M_d to indicate the splicing regions in I_p and I_d. CISDL provides new directions for the field of image forensics and plays an important role in image provenance.

Since CISDL was proposed, it has garnered significant attention from researchers. Wu *et al.* [3] proposed the first deep matching model DMVN to address CISDL specifically. Ye *et al.* [4] proposed a feature pyramid deep matching and localization network (FPLN), which fuses features of different sizes to obtain multi-scale information. Liu *et al.* [5] proposed a deep matching network named DMAV, which replaces pooling layers and deconvolution layers with atrous convolution to generate high-resolution features. Liu *et al.* [6] proposed an attention-aware correlation computation module to adjust the weight of the dot product correlation. Xu *et al.* [7] proposed a scale-adaptive correlation computation module and an attention-based mask generator to generate fine-grained masks. Tan *et al.* [8] integrates the feature extraction and feature correlation parts, and firstly introduces the Transformer attention mechanism to model interactions between feature patches.

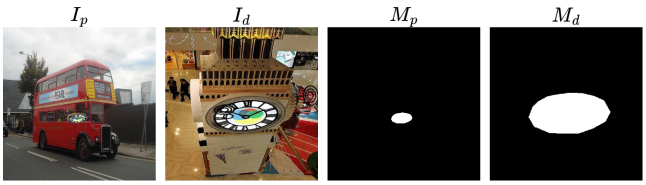


Fig. 1. Visual representation of constrained image splicing detection and localization (CISDL). I_p and I_d represent the probe image and the donor image, M_p and M_d are their corresponding masks. Spliced regions are shown in white pixels.

Despite previous methods have made considerable progress, several drawbacks remain: (1) Calculating correlation on global features introduces significant irrelevant background information redundancy, which interferes with the correlation results. (2) The feature correlation part just calculates the dot product correlation on coarse features, which cannot effectively model the relationship between features. Meanwhile, the high-level semantic coarse features lack detailed texture information, making it difficult to detect small splicing regions.

To address the problems mentioned above, we propose a novel Region Aware framework for CISDL. First, global features with different resolutions are extracted from the CNN backbone. Then, we introduce a Region Aware module to capture local features of highly correlated regions between I_p and I_d, substantially reducing the impact of background redundant information on the calculation of feature correlation. Furthermore, we employ Transformer [9] with self and cross-attention layers to complete the feature correlation matching. Transformer can effectively model the interdependencies between related image patches of I_p and I_d, while suppressing the feature expression of irrelevant regions. Linear attention

*Corresponding author

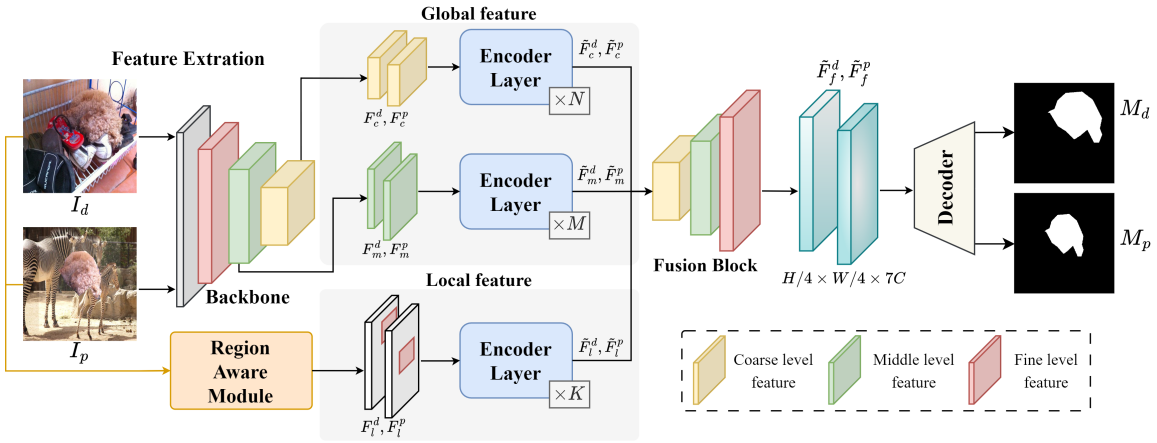


Fig. 2. Overview of our proposed Region Aware framework. Our method first extracts global features from the CNN backbone and obtains local features through the proposed Region Aware module. Next, the global and local features at different scales are encoded separately using Transformer encoder layers. Finally, the encoded features are fused and fed into the decoder to obtain the predicted mask.

[10] will be adopted on features with higher resolution to reduce computational complexity. Finally, we fuse the encoded global and local features at different scales and feed them into the decoder to obtain more precise prediction results.

Our main contributions can be summarized as follows:

- We propose a novel Region Aware CISDL framework based on Transformer. In this framework, we extract global features to effectively model the interdependence and correlation between image patches. Additionally, we incorporate Region Aware local features to reveal highly similar regions in probe and donor images, reducing the impact of background information redundancy on computing feature correlation.
- A Region Aware module based on the feature matching algorithm is developed to acquire Region Aware local features, which can identify highly similar regions, specifically the suspicious splicing regions.
- The experimental results demonstrate the superiority of our proposed approach over state-of-the-art methods on benchmark datasets.

II. PROPOSED METHOD

As shown in Fig. 2, our framework can be divided into four parts: Global Feature extraction, Local Feature extraction using the Region Aware module, Feature correlation through the Encoder layer, Feature Fusion and Decoder. Details of the proposed framework are described in the following subsections.

A. Global Feature Extraction

Given a pair of images $I_p, I_d \in \mathbb{R}^{H\times W\times 3}$, we first use the CNN backbone to extract coarse-level and middle-level global features from the images. We utilize $F_c^d, F_c^p \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 4C}$to denote the coarse-level global features at $1/16$ of the original image sizes, and $F_m^d, F_m^p \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C}$ to denote the middle-level global features at $1/8$ of the original image sizes.

B. Local Feature Extraction

Existing methods compute feature correlation on global features, which introduces irrelevant background information redundancy. To tackle this problem, we propose a Region Aware module to capture the highly correlated local features, identifying suspicious splicing regions between I_p and I_d. As demonstrated in Fig. 2, we utilize $F_l^d, F_l^p \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$ to denote the fine-level local features , which is at $1/4$ of the original image sizes. The following will introduce the Region Aware module.

1) Region Aware Module: As shown in Fig. 3 the Region Aware module takes image pair I_p and I_d as input, we first employ the state-of-the-art feature matching method SuperPoint [11] + LightGlue [12] to obtain correspondences between I_p and I_d. Then, we filter out some outliers by using RANSAC, ensuring the reliability of the matches.

$$\mathcal{M}_p, \mathcal{M}_d = RANSAC(Matcher(I_p, I_d)). \tag{1}$$

where $\mathcal{M}_p$ and $\mathcal{M}_d$ represents the set of matching points in I_p and I_d respectively, consisting of the coordinates (x, y) of every matching point. $Matcher$ represents the feature matching method SuperPoint [11] + LightGlue [12]. After that, we sample the pixels in a square region around all the matching points with side length l to generate the local images I_{d_local}, I_{p_local}.

$$\begin{aligned} I_{d_local} &= Sample(\mathcal{M}_d, l), \\ I_{p_local} &= Sample(\mathcal{M}_p, l). \end{aligned} \tag{2}$$

where $Sample(\cdot)$ retains all pixel values within a square area of length l centered on each matching point, setting all other pixel values to 0. The formula for the sampling side length l is as follows:

$$l = 2 \times \sqrt{\frac{(x_{max} - x_{min}) \times (y_{max} - y_{min})}{n}}. \tag{3}$$

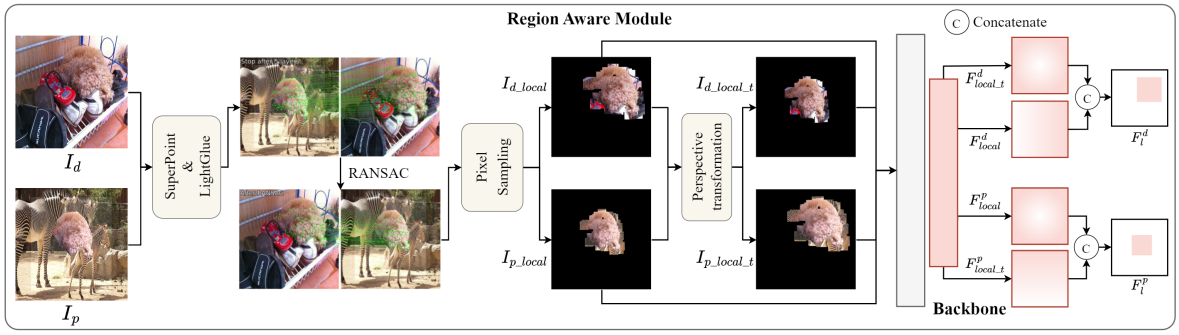


Fig. 3. The Region Aware Module. We first extract correspondences between I_p and I_d using the feature matching method. Next, RANSAC is employed to filter out outliers and sample pixels around the correspondences to get local images. Finally, we extract features from local images to obtain local features.

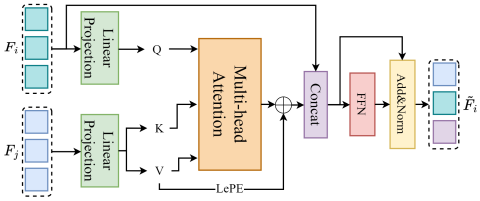


Fig. 4. The Attention Layers. If $i = j$, it represents self-attention layer. If $i \neq j$, it represents cross-attention layer. An Encoder layer consists of a self-attention layer and a cross-attention layer.

where x_{max}, y_{max}, x_{min}, y_{min} represent the maximum and minimum coordinates in the horizontal and vertical axes, and n represents the number of correspondences. In addition, we further utilize $\mathcal{M}_p$ and $\mathcal{M}_d$ to perform the perspective transformation on local images, which allows us to obtain scale, rotation, and other transformation relationships between the splicing regions:

$$\begin{aligned} I_{d_local_t} &= Warp(I_{d_local}, \mathcal{M}_d, \mathcal{M}_p), \\ I_{p_local_t} &= Warp(I_{p_local}, \mathcal{M}_d, \mathcal{M}_p). \end{aligned} \tag{4}$$

where $Warp(\cdot)$ represents the perspective transformation function. Finally, we feed both the local images I_{d_local}, I_{p_local} and the transformed images $I_{d_local_t}$, $I_{p_local_t}$ into the CNN backbone together, concatenating their outputs to obtain the local features F_l^d and F_l^p. Note that the backbone used in the Region Aware module is the same as the one used in global feature extraction.

C. Encoder Layer

After obtaining multi-scale global and local features, we utilize the Feature Encoder Layers to model the feature interdependence and correlation. Following [13], our Feature Encoder Layer consists of alternating self and cross-attention layers, which perform N, M, and K times on features at different levels respectively. The attention layer is shown in Fig. 4, we firstly flatten the features F^i and F^j into vectors F_i and F_j, then project them into the query vector Q, the key vector K and the value vector V as the following formula:

$$Q = W_Q F_i \quad K = W_K F_j \quad V = W_V F_j. \tag{5}$$

where W_Q, W_K, W_V represent linear projection matrix. And for self-attention layers, the input features F^i and F^j are the same. For cross attention layers, the input features F^i and F^j are different (either F^d and F^p or F^p and F^d). Then the Q, K, and V vectors are used to compute multi-head attention. To further highlight the feature representation of relevant regions, we adopt the LePE (Locally-Enhanced Positional Encoding) like [14]. For the coarse-level features $F_c \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 4C}$, we use vanilla dot-product attention to perform feature interaction:

$$VanillaAttention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{D}})V. \tag{6}$$

where D is the dimension of input features.

For the middle-level $F_m \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 2C}$ and the fine-level features $F_f \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$, due to their increase in spatial dimensions, using vanilla attention layers would lead to significant computational complexity. Thus we conduct linear attention [10] to perform feature interaction on these two levels of features:

$$LinearAttention(Q, K, V) = \phi(Q)(\phi(K)^T \phi(V)). \tag{7}$$

where the kernel function $\phi(\cdot) = elu(\cdot) + 1$. After the multi-head attention calculation, the features pass through two LayerNorm layers and an FFN layer to obtain the transformed vector $\tilde{F}_i$ as follows:

$$\begin{aligned} F_i^{'} &= LN(Concat(F_i, (Attention(F_i) + LePE))), \\ \tilde{F}_i &= LN(FFN(F_i^{'}) + F_i^{'}). \end{aligned} \tag{8}$$

where $Attention(\cdot)$ represents the attention layer, $LN(\cdot)$ represents the LayerNorm layer and $FFN(\cdot)$ represents the feed-forward network. The transformed vector $\tilde{F}_i$ is also the input vector of the next Encoder Layer.

D. Feature Fusion and Mask Decoder

After obtaining discriminative global and local features at different scales, we feed them into the feature fusion block like FPN [15] for multi-scale feature aggregation. Finally, the fused features are passed into a fully convolutional network decoder to generate the output prediction masks. Our decoder consists of several upsampling and convolutional layers to gradually

TABLE I
COMPARISON RESULTS ON THE SYNTHETIC SET

Method	Difficult			Normal			Easy		
	IoU	MCC	NMM	IoU	MCC	NMM	IoU	MCC	NMM
DMVN [3]	0.2772	0.3533	-0.4382	0.6818	0.7570	0.4042	0.8198	0.8544	0.6771
DMAV-adv [5]	0.5433	0.6584	0.1026	0.8317	0.8833	0.6877	0.9237	0.9411	0.8655
AttentionDM [6]	0.7228	0.8108	0.4793	0.8980	0.9320	0.8253	0.9602	0.9603	0.9388
SADM [7]	0.7759	0.8128	0.5129	0.9040	0.8288	0.8265	0.9621	0.9616	0.9410
MSTAF [8]	0.8394	0.8918	0.7064	0.9510	0.9700	0.9151	0.9788	0.9838	0.9646
Ours	**0.8760**	**0.9179**	**0.7748**	**0.9682**	**0.9799**	**0.9473**	**0.9875**	**0.9904**	**0.9813**

TABLE II
COMPARISON RESULTS ON THE MFC2018 DATASET

Method	Localization			Detection		
	IoU	MCC	NMM	Precision	Recall	F1-score
DMVN [3]	0.1204	0.1560	-0.7710	0.9099	0.3594	0.5153
DMAV-adv [5]	0.2071	0.2484	-0.6120	0.7162	0.5836	0.6431
MSTAF [8]	0.3576	0.3934	-0.3037	**0.8868**	0.6595	0.7565
Ours	**0.3897**	**0.4310**	**-0.2333**	0.7710	**0.7518**	**0.7613**

increase the resolution of feature maps by a factor of 2. The final predicted masks have the same size as the input images.

E. Loss Function

We use binary cross entropy (BCE) loss and compute the average loss of the two predicted masks.

$$Loss = -\frac{1}{HW}\sum_{i}^{HW} G^{(i)} \log M^{(i)} + (1 - G^{(i)}) \log (1 - M^{(i)}). \tag{9}$$

where $G^{(i)}$ denotes the pixels in the ground truth mask, with 0 representing the background and 1 indicating the spliced region. $M^{(i)}$ represents the pixels in predicted mask. HW represents the number of pixels in an image.

III. EXPERIMENT

A. Training Data and Implementation Details

We use the same synthetic dataset for training as MSTAF [8], and use the training subset of MS COCO dataset to generate training pairs according to the generation scripts provided by [5]. The generation process involves randomly segmenting objects through annotation from an image and then splicing it into another image. The spliced region will undergo five types of transformations: shift, rotation, scale, luminance, and deformation. The synthetic training set comprises 1 million image pairs. To preserve more detailed feature information, we resize all images to 512×512 before feeding them into the network. We implemented the proposed model in the PyTorch framework and trained our model on NVIDIA RTX 3090 GPU. The optimizer is AdamW with an initial learning rate of 1e-4. After the fifth epoch of training, the learning rate is reduced to 1e-5. We totally conduct 7-epoch training and the batch size is set to 8. The number of encoder layers for different level features is $N = 2$, $M = 3$, and $K = 1$ respectively, and the dimension C of features is set to 64.

B. Test Dataset and Evaluation Metric

We evaluate localization performance (pixel-level) and detection performance (image-level) on three datasets: (1) the Synthetic set. (2) the MFC2018 dataset. (3) the paired CASIA dataset. To ensure a fair comparison, we adopted IoU (Intersection over Union), NMM (Nimble Mask Metric), and MCC (Matthews Correlation Coefficient) to evaluate the localization performance, and Precision, Recall, and F1 as detection evaluation metrics like MSTAF [8]. The detection strategy is also kept consistent. We binarize the output masks to 0 and 1 according to a threshold of 0.5. If any pixel in both predicted masks equals 1, we consider the image pair as a positive sample.

TABLE III
COMPARISON RESULTS ON THE PAIRED CASIA DATASET

Method	Precision	Recall	F1-score
Chrislein *et al.* [16]	0.5164	0.8292	0.6364
Luo *et al.* [17]	0.9969	0.5353	0.6966
Ryu *et al.* [18]	0.9614	0.5895	0.7309
Cozzolino *et al.* [19]	0.9897	0.6334	0.7725
DMVN-loc [3]	0.9152	0.7918	0.8491
DMVN-det [3]	0.9415	0.7908	0.8596
DMAV-adv [5]	0.9657	0.8576	0.9085
AttentionDM [6]	0.9288	0.9204	0.9246
SADM [7]	0.9741	0.8758	0.9263
MSTAF [8]	0.9815	**0.9251**	**0.9524**
Ours	**1.0000**	0.8779	0.9350

The Synthetic set. In [5], they synthesized 9000 image pairs for testing using the same method as for the synthetic training set. The testing image pairs are synthesized from the testing subset of MS COCO, without overlapping with the training sets. The dataset is divided into three subsets: Hard, Normal, and Easy, based on the proportions of spliced pixels.

The MFC2018 dataset. The Media Forensics Challenge 2018 [20] dataset has 16,673 negative image pairs and 1327 positive image pairs, which is challenging for the CISDL task. According to [8], we also removed those extremely difficult samples (the proportion of spliced pixels was less than 1%), reducing the number of positive image pairs to 843. We evaluate the detection performance using all image pairs and the localization performance using 843 positive image pairs.

The paired CASIA dataset. The paired CASIA dataset contains 3642 positive image pairs and 5000 negative image pairs selected from CASIA v2.0 [21]. Since the paired CASIA dataset does not provide pixel-level ground truth, we can only

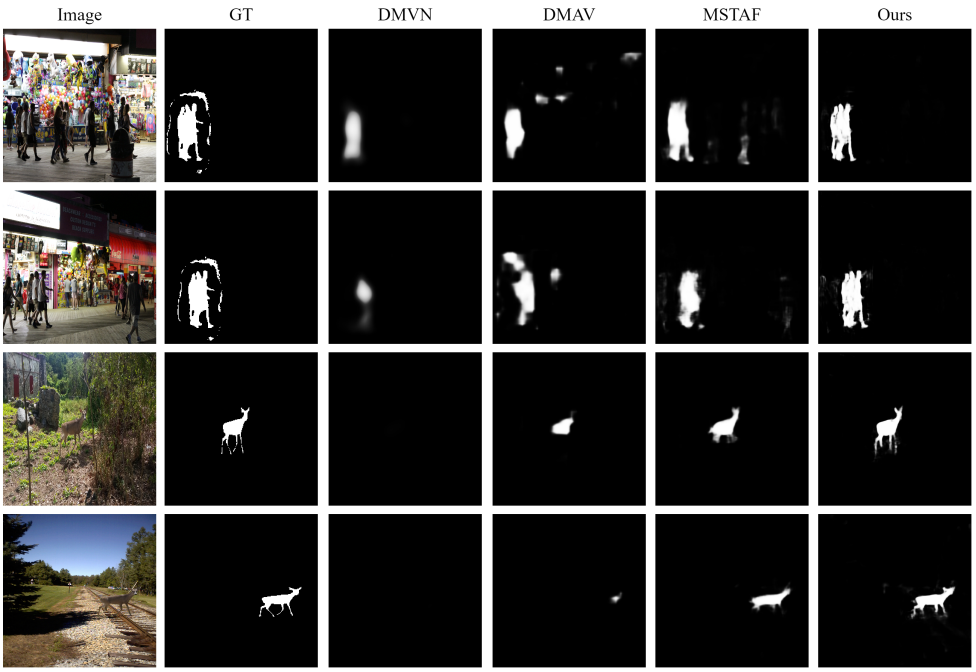


Fig. 5. Visualization of localization performance on MFC2018 dataset. GT is the ground truth mask.

TABLE IV
ABLATION STUDY ON THE SYNTHETIC SET

RA	Difficult			Normal			Easy		
	IoU	MCC	NMM	IoU	MCC	NMM	IoU	MCC	NMM
	0.8633	0.9109	0.7521	0.9533	0.9710	0.9228	0.9782	0.9831	0.9666
✓	**0.8760**	**0.9179**	**0.7748**	**0.9682**	**0.9799**	**0.9473**	**0.9875**	**0.9904**	**0.9813**

evaluate the detection performance on it.

C. Comparison Result

In the first experiment, We reported the localization performance on the Synthetic set and compared it with other CISDL methods. As shown in Tab. I, our proposed Region Aware framework outperforms state-of-the-art across all subsets and evaluation metrics. Especially in the difficult subset (where the spliced region occupies 1%-10% of the whole image pixels), our method has the most remarkable improvement.

Given that AttentionDM [6] and SADM [7] haven't made their code publicly available, we can only compare the localization performance and detection performance on the MFC2018 dataset with DMVN [3], DMAV [5], and MSTAF [8]. As shown in Tab. II, our approach still outperforms existing methods in both localization and detection performance. In terms of localization performance, our approach outperforms MSTAF in all three metrics, while in terms of detection performance, our recall and F1 scores are better than those of MSTAF. We also demonstrate the visualization results of the localization performance, as shown in Fig. 5. We can observe that MFC2018 is highly challenging, with splicing regions often being very small or subjected to rotation and scaling transformations. With the help of the proposed Region Aware module, our model effectively reduces the interference from background information and provides more accurate and complete predictions for small regions.

Tab. III shows the experiment result on the paired CASIA dataset. Since there's a lack of CISDL methods, [3] uses several copy-move methods [16]–[19] for comparison. We borrow their results and compare them with existing CISDL methods. As shown in the table, our approach achieves the best precision, with recall and F1 score performance second only to MSTAF [8].

D. Ablation

To validate the effectiveness of the Region Aware module in our proposed framework, we implement two different variants, and the comparison results are shown in Tab. IV. The first row is the result of our baseline model without the Region Aware module. We conduct these experiments using the same training strategy and evaluation metrics on the Synthetic set. As shown in Tab. IV, the introduction of the RA module led to improvements in localization across three different-sized splicing regions, indicating that the Region Aware module effectively guides the network to focus on highly correlated areas.

IV. CONCLUSION

In this work, we propose a novel Region Aware framework, which reduces the impact of background redundant informa-

tion on feature correlation. Furthermore, we introduce the linear transformer to address feature correlation computations on features with higher resolution, which alleviates the issue of losing texture details in small regions due to coarse features and effectively models the correlation between feature patches. The experimental results demonstrate that our proposed model outperforms existing state-of-the-art methods.

ACKNOWLEDGMENTS

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010645; in part by the Key project of Shenzhen Science and Technology Plan under Grant 20220810180617001; in part by the Open Research Project Programme of the State Key Laboratory of Internet of Things for Smart City (University of Macau) under Grant SKLIoTSC(UM)-2021-2023/ORP/GA04/2022.

REFERENCES

[1] Y. Li, L. Ye, H. Cao, W. Wang, and Z. Hua, "Cascaded adaptive graph representation learning for image copy-move forgery detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[2] Y. Li, Y. He, C. Chen, *et al.*, "Image copy-move forgery detection via deep patchmatch and pairwise ranking learning," *IEEE Transactions on Image Processing*, pp. 1–16, 2024.

[3] Y. Wu, W. Abd-Almageed, and P. Natarajan, "Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection," in *Proceedings of ACM international conference on Multimedia*, 2017, pp. 1480–1502.

[4] K. Ye, J. Dong, W. Wang, B. Peng, and T. Tan, "Feature pyramid deep matching and localization network for image forensics," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1796–1802.

[5] Y. Liu, X. Zhu, X. Zhao, and Y. Cao, "Adversarial learning for constrained image splicing detection and localization based on atrous convolution," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2551–2566, 2019.

[6] Y. Liu and X. Zhao, "Constrained image splicing detection and localization with attention-aware encoder-decoder and atrous convolution," *IEEE Access*, vol. 8, pp. 6729–6741, 2020.

[7] S. Xu, S. Lv, Y. Liu, C. Xia, and N. Gan, "Scale-adaptive deep matching network for constrained image splicing detection and localization," *Applied Sciences*, vol. 12, no. 13, p. 6480, 2022.

[8] Y. Tan, Y. Li, L. Zeng, J. Ye, W. Wang, and X. Li, "Multi-scale target-aware framework for constrained splicing detection and localization," in *Proceedings of ACM International Conference on Multimedia*, 2023, pp. 8790–8798.

[9] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *International conference on machine learning*, PMLR, 2020, pp. 5156–5165.

[11] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[12] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023, pp. 17 627–17 638.

[13] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[14] X. Dong, J. Bao, D. Chen, *et al.*, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022, pp. 12 124–12 134.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[16] V. Christlein, C. Riess, J. Jordan, C. Riess, and E. Angelopoulou, "An evaluation of popular copy-move forgery detection approaches," *IEEE Transactions on information forensics and security*, vol. 7, no. 6, pp. 1841–1854, 2012.

[17] W. Luo, J. Huang, and G. Qiu, "Robust detection of region-duplication forgery in digital image," in *International Conference on Pattern Recognition (ICPR)*, vol. 4, 2006, pp. 746–749.

[18] S.-J. Ryu, M.-J. Lee, and H.-K. Lee, "Detection of copy-rotate-move forgery using zernike moments," in *Information Hiding: International Conference*, 2010, pp. 51–65.

[19] D. Cozzolino, G. Poggi, and L. Verdoliva, "Efficient dense-field copy–move forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 11, pp. 2284–2297, 2015.

[20] *National institute of standards and technology (nist). (oct. 31, 2018). media forensics challenge 2018. [online]. https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018.*

[21] J. Dong, W. Wang, and T. Tan, "Casia image tampering detection evaluation database," in *IEEE China summit and international conference on signal and information processing*, 2013, pp. 422–426.

Analysis of Various Self-Supervised Learning Models for Automatic Pronunciation Assessment

Haeyoung Lee* Sunhee Kim*† and Minhwa Chung*‡
* Interdisciplinary Program in Cognitive Science, Seoul National University, Seoul, Republic of Korea
† Department of French Language Education, Seoul National University, Seoul, Republic of Korea
‡ Department of Linguistics, Seoul National University, Seoul, Republic of Korea
E-mail: {haeylee, sunhkim, mchung}@snu.ac.kr

***Abstract*—The development of automatic pronunciation assessment (APA) systems has been significantly improved by self-supervised learning (SSL) models. However, there remains a lack of systematic research on the effective utilization of SSL models and the explainability of their behavior in APA. This study aims to evaluate pronunciation with high accuracy using SSL models and to provide explanations for scoring outcomes. We fine-tune various SSL models using multiple strategies, employing extrinsic analysis to identify key factors influencing performance improvements, and intrinsic analysis via principal component analysis (PCA) to gain insights into the model's scoring patterns. The extrinsic analysis highlights the importance of strategic fine-tuning and the acoustic similarity between fine-tuning and pre-training datasets. The intrinsic analysis reveals that different SSL models focus on distinct pronunciation features, with Wav2Vec2.0 capturing more relevant information for APA. This study presents the first in-depth analysis of SSL models in APA, proposing a novel intrinsic analysis method based on feature distribution manifolds. We provide model-specific fine-tuning guidelines for APA and recommend appropriate SSL models based on specific pronunciation assessment goals. This research significantly contributes to the future development of explainable APA systems using SSL models.**

I. Introduction

Recent research in automatic pronunciation assessment (APA) has focused on improving performance using the Speechocean762 dataset [1]. Advancements have been made by incorporating word- and sentence-level features alongside traditional phoneme-level features, as well as by introducing multi-aspect scoring systems that evaluate fluency, prosody, and completeness through parallel structure models that predict all scores simultaneously [2]. Further improvements include hierarchical enhancements to the parallel model structure to effectively capture the linguistic hierarchy of pronunciation [3] and the implementation of L1-L2 aware word-level modeling that reflects the tonal characteristics of Chinese [4]. The recent integration of self-supervised learning (SSL) model features has significantly boosted performance [4]–[8]. In particular, the use of SSL models has had the greatest impact on performance improvements, as measured by the pearson correlation coefficient (PCC) [4], highlighting their substantial capabilities in APA. However, research fully exploiting the potential of SSL features in APA remains insufficient, with current approaches facing three primary limitations.

First, existing studies have primarily relied on automatic speech recognition (ASR) tasks, overlooking APA-specific characteristics. Traditional APA systems have utilized ASR results to predict pronunciation scores, and SSL-based research has largely followed this approach by fine-tuning models for speech or phoneme recognition tasks [5], [6], or by extracting general features without fine-tuning [8]. However, optimizing SSL features for APA requires score-based fine-tuning specific to APA tasks, rather than relying on ASR-oriented approaches.

Second, recent studies often employ the latest SSL models without comprehensive evaluation. Despite the different ways SSL models process speech data due to variations in training datasets and algorithms, some studies have used Wav2Vec2.0 XLSR without comparison [4] or conducted limited experiments with Wav2Vec2.0 and HuBERT [5]. A more comprehensive approach involving diverse pre-trained models and fine-tuning strategies is necessary to identify the optimal SSL model for APA.

Third, ensuring the reliability of APA systems requires not only accurate evaluation but also justification for assessment scores. Unlike traditional systems with manually designed features [9], the black-box nature of SSL models makes it difficult to understand their evaluation mechanisms. While some studies have analyzed layer-wise representations of ASR fine-tuned models [5], or extracted frame-level features to assess Wav2Vec2.0's phoneme discrimination on non-native datasets [7], these analyses remain relatively superficial. Moreover, combining SSL features with other features [4] complicates identifying performance-influencing factors.

In this study, we address these limitations by conducting diverse experiments to construct a high-accuracy APA model and identifying performance-enhancing features via extrinsic analysis. Motivated by recent NLP studies that analyze word embedding structures [10], [11], we employ principal component analysis (PCA) for intrinsic analysis. By examining the feature distribution manifolds generated from PCA on embedding vectors, our intrinsic analysis provides insights into the model's scoring mechanism.

II. Method

A. Score-based Fine-tuning of SSL models

This study proposes a score-based fine-tuning approach to optimize SSL models for APA tasks. Fig. 1 provides an overview of our method, illustrating (a) the fine-tuning process and (b) the subsequent extrinsic analysis and intrinsic analysis.

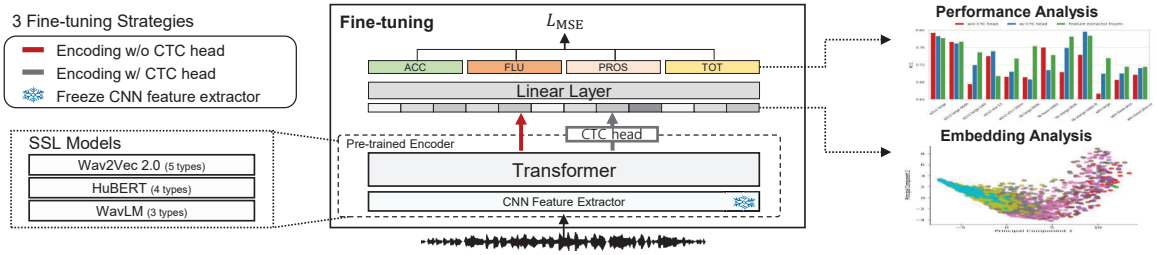


Fig. 1. Overview of the proposed method. (a) Fine-tuning process, (b) Extrinsic analysis and intrinsic analysis of the fine-tuned models.

The fine-tuning process involves adding a linear layer to the pre-trained SSL encoder to output logits for four pronunciation scores, with the model optimized using Mean Squared Error (MSE) loss between predicted and ground truth scores. This approach ensures that the SSL model is optimized for generating scores, thereby providing robust SSL feature-based interpretability.

1) SSL Pre-trained Models: We select Wav2Vec2.0 and HuBERT for their superior performance in various speech tasks [12], and include WavLM for its adaptability to non-ASR tasks. Wav2Vec2.0 utilizes contrastive loss for masked speech prediction, while HuBERT and WavLM employ cross-entropy loss for pseudo-label sequences. Our study evaluates 12 pre-trained models, each trained on distinct unlabeled datasets.

2) Fine-tuning Strategies: We implement three fine-tuning strategies to optimize the utilization of pre-trained knowledge: a) freezing the CNN feature extractor (FE) parameters to preserve the capabilities of the pre-trained models; b) adding a Connectionist Temporal Classification (CTC) head to the encoder output to encode information beneficial for ASR; and c) extracting general audio features without using a CTC head. The first strategy helps mitigate catastrophic forgetting [13] by freezing the CNN layers. This technique is particularly effective for large-scale datasets like those used in SSL models but requires further validation for APA tasks due to its varying effectiveness across different speech processing domains [14]–[16]. The second and third strategies involve the introduction of a CTC head, which automatically aligns speech and text in ASR tasks, enabling a comparison of its effectiveness against general acoustic features for APA. While previous studies have primarily focused on ASR-based phonetic features, prosodic features such as rhythm and intonation, which have been shown to be effective in predicting fluency and prosody scores [17], may also be beneficial and require evaluation.

B. Extrinsic Analysis and Intrinsic Analysis

This study employs extrinsic probing to compare the APA performance of SSL models under various conditions, aiming to identify the optimal adaptation settings for each model. However, models with similar PCC patterns may utilize different intrinsic evaluation factors. For instance, models highly correlated with fluency scores might prioritize prosodic or pitch information differently. Extrinsic analysis alone is insufficient to clarify these intrinsic mechanisms. Therefore, we propose an intrinsic probing methodology using PCA based on the manifold hypothesis [18]. PCA reduces the dimensionality of embedding vectors while preserving key variability. Focusing on the last hidden state, presumed to contain critical information for APA, as deep learning models typically make predictions in the final layer, we apply PCA to reduce high-dimensional embedding vectors. By categorizing these reduced features according to true pronunciation score labels and analyzing their distribution patterns, we gain insights into the inherent operational mechanisms of different model configurations. This dual approach provides a comprehensive understanding of SSL models in APA, revealing both performance outcomes and underlying evaluation factors.

III. Experiments

A. Dataset

This study employs Speechocean762 [1], a benchmark dataset for APA comprising 5,000 English utterances from 250 non-native Mandarin speakers. Each utterance was manually evaluated by five experts at the phoneme, word, and sentence levels. We focus on four utterance-level scores: accuracy, fluency, prosodic, and total, each rated on a 0-10 scale.

B. Experimental Setup

This study examines Wav2Vec2.0, HuBERT, and WavLM models, selecting 12 pre-trained variants based on diverse unlabeled datasets. These include Librispeech (LS) [19], consisting of 960 hours of clean, clearly pronounced speech, and Libri-light (LL) [19], which provides 60k hours of less refined public domain audiobook data. We also evaluate multilingual models (XLSR [20] and XLS-R [21]), and fine-tune ASR-pretrained models for the APA task to assess cross-task knowledge transfer. The pre-trained models, sourced from Hugging Face [22], are as follows, with the number of parameters in brackets:

- ***wav2vec2-large***: 960 hrs of LS [315.4M]
- ***wav2vec2-large-960h***: ASR fine-tuned of *wav2vec2-large* on 960 hrs of LS [315.4M]
- ***wav2vec2-large-lv60***: 53k hrs of LL [315.4M]
- ***wav2vec2-large-xlsr-53***: 56k hrs of 53 langs [315.4M]
- ***wav2vec2-xls-r-300m***: 436k hrs of 128 langs [315.4M]
- ***hubert-large-ll60k***: 60k hrs of LL [316.6M]
- ***hubert-base-ls960***: 960 hrs of LS [94.3M]

TABLE I
PCCS OF SSL FINE-TUNED MODELS. THE BEST MODEL FOR EACH ASPECT AMONG THE THREE STRATEGIES IS INDICATED IN BOLD, AND THE BEST MODEL FOR EACH STRATEGY IS MARKED WITH AN ASTERISK (*).

Finetuned Model	Accuracy			Fluency			Prosodic			Total		
	Train all		Freeze	Train all		Freeze	Train all		Freeze	Train all		Freeze
	w/o CTC	w/ CTC	FE	w/o CTC	w/ CTC	FE	w/o CTC	w/ CTC	FE	w/o CTC	w/ CTC	FE
w2v2-large	0.691	0.688	**0.694**	***0.794**	0.787	0.782	***0.786**	0.785	0.776	**0.728**	0.718	0.723
w2v2-large-960h	*0.706	**0.708**	0.702	0.773	0.770	**0.774**	0.773	0.771	**0.775**	***0.734**	0.729	0.727
w2v2-large-lv60	0.623	**0.666**	0.649	0.676	0.720	**0.749**	0.672	0.730	**0.742**	0.642	**0.686**	0.679
w2v2-xlsr-53	0.678	**0.691**	0.645	0.740	**0.752**	0.694	0.734	**0.751**	0.691	0.694	**0.706**	0.664
w2v2-xls-r-300m	0.633	0.649	**0.661**	0.693	0.705	**0.735**	0.681	0.692	**0.727**	0.647	0.663	**0.679**
hb-large-ll60k	0.620	0.616	**0.698**	0.692	0.687	**0.763**	0.683	0.681	**0.760**	0.633	0.633	**0.716**
hb-base-ls960	0.673	0.626	**0.674**	**0.760**	0.708	0.743	**0.759**	0.693	0.739	**0.704**	0.649	0.698
hb-xlarge-ll60k	0.631	0.686	**0.702**	0.704	0.759	**0.786**	0.693	0.761	**0.783**	0.646	0.705	**0.728**
hb-xlarge-ls960-ft	0.670	*0.719	***0.722**	0.743	***0.797**	*0.788	0.741	***0.788**	*0.784	0.693	*0.734	***0.745**
wlm-large	0.613	0.649	**0.656**	0.654	0.700	**0.736**	0.644	0.695	**0.726**	0.620	0.659	**0.680**
wlm-base-plus	0.603	0.636	**0.653**	0.686	0.701	**0.716**	0.681	0.696	**0.708**	0.632	0.653	**0.673**
wlm-base-plus-sv	0.649	0.641	**0.656**	0.697	0.713	**0.716**	0.687	0.698	**0.714**	0.667	0.664	**0.680**

- ***hubert-xlarge-ll60k***: 60k hrs of LL [962.5M]
- ***hubert-xlarge-ls960-ft***: ASR fine-tuned of *hubert-xlarge-ll60k* on 960 hrs of LS [962.5M]
- ***wavlm-large***: mix of 94k hrs data [315.4M]
- ***wavlm-base-plus, wavlm-base-plus-sv***: mix of 94k hrs data [94.3M] (sv for speaker verification)

Fine-tuning is conducted using the Hugging Face toolkit. For configurations employing a CTC head, we attach the CTC head to the transformer encoder output, encoding each sequence into logits representing probabilities of 32 alphabet tokens through linear transformation. For configurations without a CTC head, the SSL model's encoding approach is used, producing 768- or 1024-dimensional embeddings for base and larger models, respectively. Models are fine-tuned on 2,500 Speechocean762 training samples using MSE loss, with a batch size of 8, for 30 epochs, and the AdamW optimizer with a learning rate of 1e-5. Results are averaged over two runs with different random seeds. Hyperparameter optimization yields 1e-5 as the optimal learning rate. Notably, simultaneous prediction of all four pronunciation scores outperforms single-score prediction.

C. Evaluation Metric

PCC is used to measure the correlation between model-predicted scores and human-annotated scores. A PCC value close to 1 indicates a strong positive relationship between the model's predictions and the human ratings.

IV. RESULTS

A. Performance Comparison of Fine-tuned Models

Table I presents the APA performance (PCC) of 36 fine-tuned models. Wav2Vec2.0 and HuBERT exhibit higher PCC performance compared to WavLM. Specifically, within the Wav2Vec2.0 models, *w2v2-large-960h* excels in accuracy and total scores, while *w2v2-large* demonstrates superior performance in fluency and prosodic scores. Among the HuBERT models, *hb-xlarge-ls960-ft* achieves the best PCC performance.

Models trained on Librispeech outperform those trained on the larger but less refined Libri-light dataset. Notably, the Librispeech-trained models *w2v2-large* and *hb-base-ls960* outperform their Libri-light counterparts *w2v2-large-lv60* and *hb-large-ll60k*, even when the Librispeech models have smaller architectures. Conversely, multilingual models trained in diverse linguistic environments generally show lower performance.

TABLE II
EXPLAINED VARIANCE RATIOS OF PRINCIPAL COMPONENTS

	Pre-trained		Fine-tuned	
	PC1	PC2	PC1	PC2
w2v2-large	0.34	0.13	**0.74**	0.16
hb-large-ll60k	0.07	0.07	**0.85**	0.12
wlm-large	0.08	0.07	**0.83**	0.10

The efficacy of fine-tuning strategies varies across models. Freezing the CNN FE yields optimal results for most models. HuBERT, in particular, shows consistent performance improvement with increasing model size. Conversely, the Librispeech-trained *w2v2-large* and *hb-base-ls960* perform comparably or better when fully fine-tuned, outperforming their Libri-light-trained counterparts, *w2v2-large-lv60* and *hb-large-ll60k*.

CTC head integration generally enhances performance, though model-specific variations are observed. Notably, the impact of CTC head integration is significantly influenced by dataset differences. Wav2Vec2.0 and HuBERT exhibit contrasting patterns: multilingual Wav2Vec2.0 models and the Libri-light-trained *w2v2-large-lv60* benefit from the CTC head. In contrast, the Libri-light-trained *hb-large-ll60k* performs poorly regardless of CTC head integration, while the Librispeech-trained *hb-base-ls960* shows improved performance without it. This suggests that HuBERT may not fully leverage the benefits of the CTC head under certain conditions. Furthermore, the Librispeech-trained *w2v2-large* and the ASR-fine-tuned *w2v2-large-960h* demonstrate excellent performance irrespective of CTC head integration. Conversely, the HuBERT x-large model, known for its superior ASR performance, shows significant improvement when incorporating the CTC head.

B. PCA-based Analysis of SSL Model Embeddings

Table II presents the explained variance ratios of the first principal component (PC1) and second principal component (PC2) obtained from PCA performed on the final hidden states of three large-sized models, each encoded without a CTC head and labeled with correct score labels. The explained variance ratios of PC1 for the fine-tuned models are 0.74, 0.85, and 0.83,

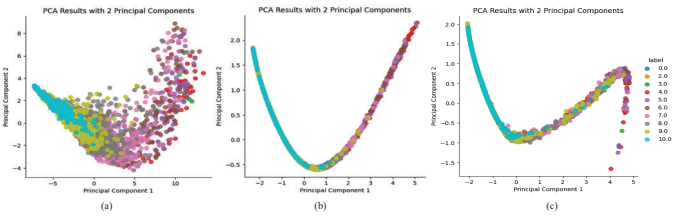


Fig. 2. 2D PCA visualization of (a) Wav2Vec2.0 (b) HuBERT (c) WavLM

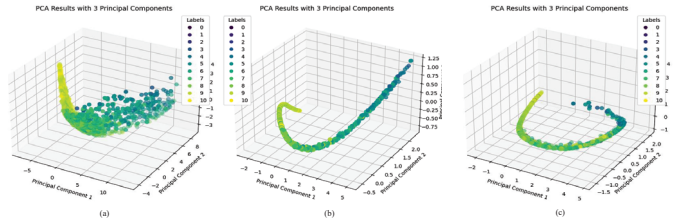


Fig. 3. 3D PCA visualization of (a) Wav2Vec2.0 (b) HuBERT (c) WavLM

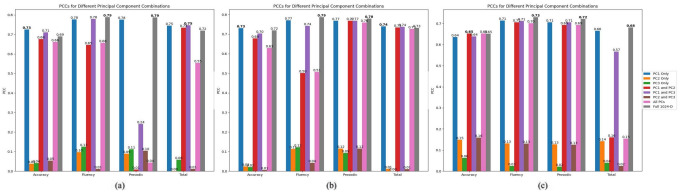


Fig. 4. PCCs for different PC combinations of (a) Wav2Vec2.0 (b) HuBERT (c) WavLM

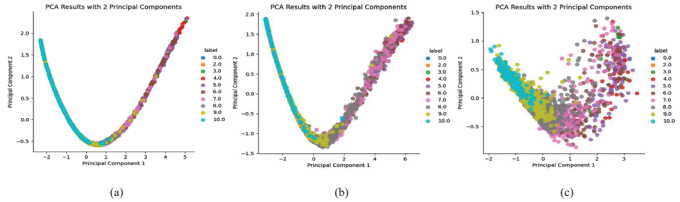


Fig. 5. PCA visualization of (a) *hb-large-ll60k* w/o CTC (b) *hb-large-ll60k* with frozen FE (c) *hb-xlarge-ls960-ft* w/ CTC

respectively, whereas for the non-fine-tuned models, they are 0.34, 0.07, and 0.08. This indicates that PC1 alone accounts for approximately 75% to 85% of the primary variability in the original data, suggesting that fine-tuning adjusts the model's feature vectors to be more suitable for pronunciation evaluation.

Fig. 2 visualizes the 2D PCA results for the fluency scores of the models in Table II, with PC1 and PC2 as axes. Fig. 3 adds PC3 for 3D PCA visualization. The manifold shapes vary by model type: (a) Wav2Vec2.0 forms a cone shape, (b) HuBERT forms a V shape, and (c) WavLM forms an S shape. While model type causes noticeable differences in manifold shapes, the shapes remain consistent across different fine-tuning strategies and datasets. In 2D, HuBERT and WavLM do not show a continuous score distribution, but in 3D, lower scores become more distinct with the addition of the z-axis.

The most notable model is Wav2Vec2.0, which demonstrates the largest variance in the mid-score range and greater variance in the lower score range compared to other models. Higher scores are sharply modeled, while lower scores are more dispersed, indicating a continuous distribution of scores.

Fig. 4 presents PCC results for various combinations of principal component (PC) vectors for each model. This analysis verifies whether the dimension-reduced PC vectors can accurately predict the actual pronunciation evaluation scores. Remarkably, all three models achieve similar or better performance using only PC1 compared to the original 1024-dimensional vectors. The HuBERT and WavLM models, in particular, show improved performance with multiple PCs. Additionally, the PC combinations affecting pronunciation evaluation vary across different aspects for each model.

Fig. 5 visualizes the fluency scores for different HuBERT models, categorized by score. The PCC performances for models (a), (b), and (c) are 0.692, 0.763, and 0.797, respectively. Notably, the higher-performing HuBERT models exhibit patterns similar to the manifold observed in Wav2Vec2.0 models.

V. Discussion

A. Extrinsic Analysis of Factors Contributing to Performance Improvement

Table III compares the performance of our best model, based on fluency and prosodic scores, with the latest APA baseline model. On the Speechocean762 dataset, our model achieves PCC values up to 0.048 and 0.037 higher in fluency and prosodic aspects, respectively. Although it does not surpass the state-of-the-art performance reported in Ref. [4], it is noteworthy that simply fine-tuning a pre-trained SSL model can yield performance comparable to the latest pronunciation assessment models.

a) Performance of ASR-adapted settings: ASR-adapted settings perform exceptionally well in APA tasks. Among the three models, Wav2Vec2.0 and HuBERT demonstrate the best performance when fine-tuned for ASR, while the noise-resistant WavLM model shows relatively lower performance. Additionally, models trained on the cleaner Librispeech dataset outperform those trained on noisier datasets.

b) Impact of Pre-Training Data: The type of data used for pre-training significantly influences APA performance. Models trained on the clean Librispeech dataset outperform those trained on the noisier Libri-light dataset, despite the quantitative disadvantage. Interestingly, the smaller *hb-base-ls960* model using Librispeech data outperforms the larger *hb-*

TABLE III
Performance Comparison with the Baseline Model

Model	Utterance Score (PCC) ↑			
	Accuracy	Fluency	Prosody	Total
Kim et al. [5]	-	0.780	0.770	-
GOPT [2]	0.714	0.753	0.760	0.742
HiPAMA [3]	0.730	0.749	0.751	0.754
3MH [4]	0.782	0.843	0.836	0.811
w2v2-large	0.691	**0.794**	**0.786**	0.728
hb-xlarge-ls960-ft	0.719	**0.797**	**0.788**	0.734
wlm-large	0.656	0.736	0.726	0.680

large-ll60k model, suggesting that dataset quality has a greater impact on performance than model size.

Moreover, models trained on the clean, English ASR dataset Librispeech outperform those trained on multilingual datasets, particularly for APA tasks involving non-native speakers. This is likely due to the acoustic similarity between the pre-training and fine-tuning data, as demonstrated in Ref. [23], emphasizing the importance of dataset relevance over volume and diversity. Overall, we find that the optimal tuning methods vary depending on the characteristics of the pre-trained models. Therefore, strategically tuning models while considering their inherent properties has a significant impact on APA performance. Below are the details of our discussion.

c) Role of low-level feature extraction: Maintaining the low-level feature extraction ability of pre-trained models is crucial for APA performance. Freezing the CNN FE, which is closely related to the unlabeled datasets used during pre-training, often leads to better performance. However, fine-tuning the entire model is more beneficial for models trained on Librispeech, aligning with findings in Ref. [13], which suggest that fine-tuning all layers with a low learning rate is effective when the source and target data are acoustically similar.

d) CTC head influence: Learning information between speech and text through a CTC head generally has a positive impact on APA performance. However, Wav2Vec2.0 and HuBERT exhibit contrasting improvements depending on the ASR environment. For Wav2Vec2.0, the CTC head is particularly beneficial in less favorable ASR conditions, such as when using the Libri-light or multilingual datasets, which have shown lower performance in English pronunciation assessment. In contrast, it has little impact in more favorable settings, such as the Librispeech dataset or ASR fine-tuned models. Conversely, HuBERT struggles even with the CTC head in challenging conditions. Notably, in the *hb-xlarge-ll60k* model, which demonstrates strong ASR performance [24], incorporating the CTC head significantly enhances performance. This suggests that while HuBERT benefits from the CTC head in favorable conditions, it performs poorly without it, even in optimal settings. In contrast, Wav2Vec2.0 models perform well without the CTC head in favorable ASR environments, likely due to their effective learning of phoneme-level information during pre-training, as supported by several studies [7], [25]. These findings highlight the different mechanisms by which the two models learn phonetic information, emphasizing that optimal tuning strategies should consider the specific characteristics of each pre-trained model to maximize APA performance.

B. Intrinsic Analysis of Feature Representation

The high variance explained by PC1 in Table II, along with the superior APA performance observed when using PC1 alone in Fig. 4, suggests that SSL models fine-tuned for APA tasks effectively capture key high-dimensional features critical for score prediction. Each SSL model evaluates scores based on these principal features.

As illustrated in Fig. 2 and Fig. 3, the varying manifold shapes in the feature space composed of PCs indicate different pronunciation assessment patterns across models. Specifically: (a) The conical shape of the Wav2Vec2.0 manifold suggests a primary reliance on PC1 for evaluation. (b) The V-shaped manifold of the HuBERT model implies the use of two main criteria. (c) The S-shaped manifold of the WavLM model indicates an assessment based on multiple criteria. Notably, unlike Wav2Vec2.0, HuBERT and WavLM demonstrate a clearer score distribution in three-dimensional space, as evidenced by the performance improvement with additional PCs in Fig. 4. This suggests that while Wav2Vec2.0 primarily relies on a single PC for evaluation, HuBERT and WavLM utilize a broader range of evaluation criteria.

As shown in Fig. 2(a), the Wav2Vec2.0 model uniquely captures a continuous score distribution and exhibits significant variance in the middle score range, unlike the other models. In the Speechocean762 dataset, the middle score range is predominant, and the model's ability to recognize various error patterns within this range highlights its superior performance. Furthermore, the data distribution in Wav2Vec2.0, with most data concentrated in the middle and high scores and scarcely any data in the low scores, aligns with the actual distribution of the Speechocean762 dataset. For real-world APA applications, it is crucial that models are calibrated to realistic scenarios, which Wav2Vec2.0 achieves most effectively. Additionally, Fig. 5 shows that the higher-performing HuBERT model exhibits a manifold similar to that of the Wav2Vec2.0, further emphasizing its excellence. In summary, Wav2Vec2.0 learns the most advantageous information for APA and can be considered the best model for this task.

Different fine-tuning strategies and datasets impact model performance but do not alter the intrinsic features the models learn. Fig. 2 and Fig. 3 show noticeable differences between models, with minimal changes due to dataset or fine-tuning variations. This suggests that while these factors affect performance, they do not fundamentally change the intrinsic features.

VI. Conclusion

This study explores the characteristics of SSL models that enhance APA performance through various experiments and explains their results. Additionally, we propose an intrinsic analysis methodology that visually confirms the distribution and variability of data via manifold shapes, reflecting the criteria and complexity used in pronunciation assessment. This approach helps in understanding the intrinsic operational mechanisms of each model. Future research should focus on detailed analysis of PC vectors to clarify the features they represent, as the exact scoring criteria of SSL models remain unclear. The diverse analytical results of this study offer crucial guidelines for utilizing SSL models in APA and enhance understanding of these previously opaque SSL models, contributing to the future development of explainable SSL model-based pronunciation assessment systems.

References

[1] J. Zhang, Z. Zhang, Y. Wang, *et al.*, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[2] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7262–7266.

[3] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5.

[4] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "A Hierarchical Context-aware Modeling Approach for Multi-aspect and Multi-granular Pronunciation Assessment," *arXiv preprint arXiv:2305.18146*, 2023.

[5] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," *arXiv preprint arXiv:2204.03863*, 2022.

[6] A. Zahran, A. Fahmy, K. Wassif, and H. Bayomi, "Fine-tuning Self-Supervised Learning Models for End-to-End Pronunciation Scoring," *IEEE Access*, 2023.

[7] W. Liu, K. Fu, X. Tian, *et al.*, "An ASR-free fluency scoring approach with self-supervised learning," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5.

[8] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 575–582.

[9] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech communication*, vol. 51, no. 10, pp. 883–895, 2009.

[10] R. Bommasani, K. Davis, and C. Cardie, "Interpreting pretrained contextualized representations via reductions to static embeddings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4758–4781.

[11] T. Musil, "Examining structure of word embeddings with PCA," in *Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22*, Springer, pp. 211–223.

[12] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[13] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," *Technologies*, vol. 11, no. 2, p. 40, 2023.

[14] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[15] Y. Gao, C. Chu, and T. Kawahara, "Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with ASR and gender pretraining," in *Proc. Interspeech*, 2023.

[16] H. Aronowitz, I. Gat, E. Morais, W. Zhu, and R. Hoory, "Towards A Common Speech Analysis Engine," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8162–8166.

[17] Y. E. Kheir, A. Ali, and S. A. Chowdhury, "Automatic Pronunciation Assessment–A Review," *arXiv preprint arXiv:2310.13974*, 2023.

[18] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[19] J. Kahn, M. Rivière, W. Zheng, *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673.

[20] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[21] A. Babu, C. Wang, A. Tjandra, *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[22] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45.

[23] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, *et al.*, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

Design of Spectrogram-Consistency Regularization Term Dependent on Observation in Independent Low-Rank Matrix Analysis for Blind Source Separation

Takaaki Kojima*, Norihiro Takamune*, Daichi Kitamura†, and Hiroshi Saruwatari*
* The University of Tokyo, Japan
† National Institute of Technology, Kagawa College, Japan

***Abstract*—Independent low-rank matrix analysis (ILRMA) is the state-of-the-art technique for blind source separation under the overdetermined condition. Recently, some methods that focus on the spectrogram consistency of the separated signals to improve the separation performance have been proposed. One of such methods introduces into ILRMA the regularization term that directly induces the separated signals to be consistent. Although in the conventional method, the regularization term is designed to be independent of the observed signals, we design the regularization term to be dependent on the observed signals. For the obtained cost function, we derive a new update rule on the basis of the majorization-minimization algorithm with the auxiliary function that is an extension of that derived in the conventional method. Finally, a numerical experiment is conducted to verify the separation performance.**

I. Introduction

Blind source separation (BSS) is a technique for estimating individual sources from an observed mixture without knowledge of the characteristics of sources or mixing conditions. Independent component analysis (ICA) [1] is a representative approach to solving BSS under the overdetermined condition (the number of microphones M is greater than or equal to that of sources N). In ICA, the demixing matrix is estimated under the assumption that individual sources are statistically independent and the mixing system is instantaneous. Since the sources are mixed by time-domain convolution for acoustic signals, ICA is often applied in the time-frequency domain via short-time Fourier transform (STFT) under the assumption that the mixing system can be approximated by instantaneous mixing in the time-frequency domain. This method is called frequency-domain ICA (FDICA) [2] and estimates frequency-wise demixing matrices by independently applying ICA to complex-valued signals in each frequency bin. ICA has arbitrariness to the permutation of the estimated signals; therefore, the permutations should be aligned for frequency bins in FDICA (permutation problem). Several permutation solvers have been studied to solve this problem [3]–[5]. In contrast, independent vector analysis (IVA) [6] and independent low-rank matrix analysis (ILRMA) [7] can estimate the frequency-wise permutations simultaneously with the demixing matrix. These techniques assume a more complex source model than FDICA and can therefore capture dependences between frequency bins and align the frequency-wise permutations. Among these methods, it has been reported that ILRMA tends to show a higher separation performance than FDICA and IVA [7]. However, it has been reported that there is room for performance improvement even in ILRMA [8].

The spectrogram consistency (or simply *consistency*) [9]–[11] has been studied as a new promising clue to align the frequency-wise permutations [12]–[14]. In some BSS methods in the time-frequency domain, such as FDICA, IVA, and ILRMA, assumptions on the spectrograms of the separated signals are made. However, these assumptions are not guaranteed to be satisfied for the final output in the time domain obtained as the inverse STFT (ISTFT) of the spectrograms of the separated signals, i.e., the spectrograms obtained as the STFT of the final output are often different from the spectrograms of the separated signals before ISTFT. This is because the operation of STFT after ISTFT is a projection operator on the image space of the STFT as a linear operator. The image space of the STFT is called the consistent subspace in this paper. For example, if the frequency-wise permutations are not aligned, the spectrogram is changed greatly by the projection and becomes blurred in both time and frequency directions as shown in Fig. 1. Then, the performance of BSS is expected to be improved by making assumptions on the projected spectrograms or by reducing the difference between the spectrograms before and after the projection. It has been reported that the performance of FDICA without a permutation solver can be improved by incorporating the projection operator into the cost function of FDICA [12].

As for ILRMA with the spectrogram consistency considered, consistent ILRMA (Consist. ILRMA) [13] and ILRMA with spectrogram-consistency regularization [14] have been proposed. In Consist. ILRMA [13], the update rules of ILRMA are modified to approximate the projected spectrograms instead of the spectrograms of the separated signals (before the

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 (for algorithm development) and Tateisi Science and Technology Foundation (for practical experiment).

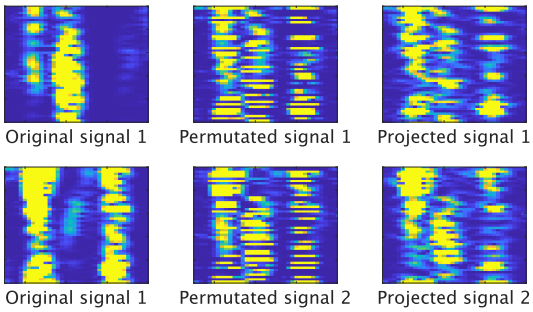


Fig. 1: Examples of the spread by projecting inconsistent spectrograms onto the consistent subspace. Original speech signals (left) were permutated randomly (middle) with frequency bins and projected onto the consistent subspace (right).

projection) by nonnegative matrix factorization (NMF) [15]. In [14], a regularization term that induces the spectrograms of the separated signals (before the projection) themselves to be close to the consistent subspace is designed for ILRMA. ILRMA with the spectrogram-consistency regularization has the advantage of being formulated as an optimization problem, whereas Consist. ILRMA is heuristic and unfortunately not formulated as an optimization problem. The regularization term proposed in [14] is designed to be independent of the observed signals; in other words, it is determined solely by the STFT conditions (e.g., the window length, window function, and hopsize). Therefore, neither the specific structure nor correlation of the observed signals can be fully utilized in [14]. In this paper, we design a new regularization term for ILRMA on the basis of the distances between the spectrograms of the separated signals and the projected spectrograms of the separated signals onto the consistent subspace. The proposed regularization term depends on the observed signals; thus, the separation performance is expected to be improved more than in [14]. For the new cost function, we unfortunately cannot use the same technique of [14] to design the auxiliary function for the majorization-minimization (MM) algorithm [16]. To address this problem, we extend the auxiliary-function-design methodology from [14] by utilizing the mathematical finding in this paper and design an auxiliary function on the basis of this finding. We derive a new update rule by applying vector-wise coordinate descent (VCD) [17] to the designed auxiliary function. Finally, we verify the separation performance of the proposed method by a simulation experiment.

II. PRELIMINARIES

In this paper, we consider a time-domain signal to be a real-valued signal of length L. The complex-valued spectrogram of a time-domain signal $\tilde{\boldsymbol{\psi}} \in \mathbb{R}^L$ is represented by the supervecter $\vec{\boldsymbol{\Psi}} \in \mathbb{C}^{IJ}$, where I and J are the window length and the number of time frames in STFT, respectively. Here, $(I(j-1)+i)$th element of $\vec{\boldsymbol{\Psi}}$ is the ith frequency bin of the spectrum in the jth frame, where $i = 1, \ldots, I$ and $j = 1, \ldots, J$ are the indices of the frequency bins and time frames, respectively. Hereafter, we assume the practical case where $IJ > L$ holds. Let τ be the hopsize of STFT, $\boldsymbol{h} \in \mathbb{R}^I$ be the window function. The length of the signal after zero-padding is denoted by $L' = L^{(\mathrm{pre})} + L + L^{(\mathrm{post})}$, where $L^{(\mathrm{pre})} (< I)$ and $L^{(\mathrm{post})} (< I)$ are the length of zero-padding at the beginning and end of the time signal, respectively. Note that $L' = \tau(J-1) + I$ holds. The $L' \times L$ matrix representing zero-padding is given by

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{O}_{L \times L^{(\mathrm{pre})}} & \boldsymbol{E}_L & \boldsymbol{O}_{L \times L^{(\mathrm{post})}} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{L' \times L}, \quad (1)$$

where $\boldsymbol{O}_{q \times q'}$ is the $q \times q'$ zero matrix and $\boldsymbol{E}_q$ is the $q \times q$ identity matrix for any nonnegative integer q, q'. Here, $\cdot^{\mathsf{T}}$ represents the transpose. That is, the signal after zero padding can be represented by $\boldsymbol{Z}\tilde{\boldsymbol{\psi}}$. The $(I(j-1)+1)$th to (Ij)th elements of $\vec{\boldsymbol{\psi}}$, which is the spectrum in the jth frame, are obtained by multiplying the $(\tau(j-1)+1)$th to $(\tau(j-1)+I)$th samples of the zero-padded signal $\boldsymbol{Z}\tilde{\boldsymbol{\psi}}$ by the window function $\boldsymbol{h}$ and applying discrete Fourier transform (DFT). Thus, the vector created by arranging the $(I(j-1)+1)$th to (Ij)th element of $\vec{\boldsymbol{X}}$ is represented by $\boldsymbol{F}\boldsymbol{\Delta}_j\boldsymbol{Z}\tilde{\boldsymbol{\psi}}$, where $\boldsymbol{F} \in \mathbb{C}^{I \times I}$ denotes the representation matrix of DFT and $\boldsymbol{\Delta}_j \in \mathbb{R}^{I \times L'}$ is defined as

$$\boldsymbol{\Delta}_j = \begin{bmatrix} \boldsymbol{O}_{I \times \tau(j-1)} & \mathrm{diag}\{\boldsymbol{h}\} & \boldsymbol{O}_{I \times (L' - \tau(j-1) - I)} \end{bmatrix}. \quad (2)$$

Here, $\mathrm{diag}\{\boldsymbol{h}\}$ represents the diagonal matrix whose diagonal elements are $\boldsymbol{h}$. Therefore, $\vec{\boldsymbol{\Psi}}$ is represented by $\vec{\boldsymbol{\Psi}} = \mathcal{F}\tilde{\boldsymbol{\psi}}$, where $\mathcal{F} \in \mathbb{C}^{IJ \times L}$ is defined as

$$\mathcal{F} = \begin{bmatrix} \boldsymbol{F}\boldsymbol{\Delta}_1\boldsymbol{Z} \\ \vdots \\ \boldsymbol{F}\boldsymbol{\Delta}_J\boldsymbol{Z} \end{bmatrix}. \quad (3)$$

In this paper, we consider that the STFT conditions satisfy the condition that $\mathcal{F}$ has full column rank so that ISTFT exists. We use the Moore–Penrose inverse $\cdot^{\dagger}$ of $\mathcal{F}$ as ISTFT, i.e., $\mathcal{F}^{\dagger}$ is a representation matrix of ISTFT. Note that $\mathcal{F}^{\dagger}\mathcal{F}$ is equal to $\boldsymbol{E}_L$ and $\mathcal{F}\mathcal{F}^{\dagger}$ is an orthogonal projection matrix onto the image space of $\mathcal{F}$. A spctrogram $\vec{\boldsymbol{\Psi}}' \in \mathbb{C}^{IJ}$ is called *consistent* when $\vec{\boldsymbol{\Psi}}'$ is in the image space of $\mathcal{F}$, i.e., $\vec{\boldsymbol{\Psi}}' = \mathcal{F}\mathcal{F}^{\dagger}\vec{\boldsymbol{\Psi}}'$ holds.

III. CONVENTIONAL METHODS

A. ILRMA [7]

Let $s_{ij,n}$, $x_{ij,m}$, and $y_{ij,n}$ be the (i,j)th bin in the spectrogram of the nth source, the mth observed, and nth separated signals, respectively, where $n = 1, \ldots, N$ and $m = 1, \ldots, M$ denote the indices of the source and observed signals, respectively. We define $\boldsymbol{s}_{ij} = (s_{ij,1} \cdots s_{ij,N})^{\mathsf{T}} \in \mathbb{C}^N$, $\boldsymbol{x}_{ij} = (x_{ij,1} \cdots x_{ij,M})^{\mathsf{T}} \in \mathbb{C}^M$, and $\boldsymbol{y}_{ij} = (y_{ij,1} \cdots y_{ij,N})^{\mathsf{T}} \in \mathbb{C}^N$.When the mixing system is time-invariant and the reverberation time is sufficiently shorter than the length of the STFT window, we can assume instantaneous mixing in the time-frequency domain as $\boldsymbol{x}_{ij} = \boldsymbol{A}_i\boldsymbol{s}_{ij}$, where $\boldsymbol{A}_i \in \mathbb{C}^{M \times N}$ is the mixing matrix. In this paper, we assume that $M = N$ and $\boldsymbol{A}_i$ is regular. Under these assumptions, there exists the demixing matrix $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1} \cdots \boldsymbol{w}_{i,N})^{\mathsf{H}} \simeq \boldsymbol{A}_i^{-1}$ that satisfies $\boldsymbol{y}_{ij} = \boldsymbol{W}_i\boldsymbol{x}_{ij}$. Let $w_{i,nm}$ be the mth elements of $\boldsymbol{w}_{i,n}$.

In ILRMA, $y_{ij,n}$ is assumed to be generated from a complex Gaussian distribution with a mean of zero and the variance $\rho_{ij,n}$ as follows:

$$p(y_{ij,n}) = \frac{1}{\pi\rho_{ij,n}} \exp\left(-\frac{|y_{ij,n}|^2}{\rho_{ij,n}}\right). \tag{4}$$

The variance $\rho_{ij,n}$ is assumed to have a low-rank structure and modeled by NMF [15] as

$$\rho_{ij,n} = \sum_{k=1}^{K} t_{ik,n} v_{kj,n}, \tag{5}$$

where $t_{ik,n} \geq 0$ and $v_{kj,n} \geq 0$ represent the basis and the activation, respectively, and $k = 1, \ldots, K$ denotes the index of the basis. The demixing matrix $\{\boldsymbol{W}_i\}$ and the NMF variables $\{t_{ik,n}\}$, $\{v_{kj,n}\}$ are obtained from maximum likelihood estimation by minimizing the following negative log-likelihood as a cost function:

$$\begin{aligned}\mathcal{L} =& \frac{1}{2}\sum_{i,j,n}\left(\frac{|\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2}{\sum_k t_{ik,n}v_{kj,n}} + \log\sum_k t_{ik,n}v_{kj,n}\right) \\ & - \frac{J}{2}\sum_i \log|\det \boldsymbol{W}_i|^2 + \text{const.},\end{aligned} \tag{6}$$

where const. denotes the terms independent of $\{\boldsymbol{W}_i\}$, $\{t_{ik,n}\}$ and $\{v_{kj,n}\}$. To optimize (6), the demixing matrix $\{\boldsymbol{W}_i\}$ and the NMF variables $\{t_{ik,n}\}$, $\{v_{kj,n}\}$ are updated alternately. The demixing matrix $\{\boldsymbol{W}_i\}$ is updated with iterative projection (IP) [18]. The NMF variables $\{t_{ik,n}\}$, $\{v_{kj,n}\}$ are updated on the basis of the MM algorithm [16]. All of these update rules ensure the monotonic nonincrease in the cost function (6).

B. ILRMA with spectrogram-consistency regularization independent of observation [14]

In [14], the regularization term based on spectrogram consistency is introduced to ILRMA. Let $\vec{\boldsymbol{Y}}_n$ be the supervectorized spectrogram of the nth separated signal whose $(I(j-1)+i)$th element is $y_{ij,n}$, and the degree to which $\vec{\boldsymbol{Y}}_n$ deviates from the consistent subspace is formulated as [14]

$$\begin{aligned}\mathcal{E}(\vec{\boldsymbol{Y}}_n) &= \|\vec{\boldsymbol{Y}}_n - \mathcal{F}\mathcal{F}^{\dagger}\vec{\boldsymbol{Y}}_n\|_2^2 \\ &= \left\|(\boldsymbol{E}_{IJ} - \mathcal{F}\mathcal{F}^{\dagger})\sum_m \mathcal{W}_{nm}\mathcal{F}\tilde{\boldsymbol{x}}_m\right\|_2^2,\end{aligned} \tag{7}$$

where $\mathcal{W}_{nm} \in \mathbb{C}^{IJ\times IJ}$ is a diagonal matrix whose $(I(j-1)+i)$th diagonal element is $w_{i,nm}^*$ and $\tilde{\boldsymbol{x}}_m \in \mathbb{R}^L$ is the mth observed signal in the time domain. Here, $\cdot^*$ denotes the complex conjugate. To induce consistency for arbitrary observed signals, the case where $\tilde{\boldsymbol{x}}_m$ is a white Gaussian signal without correlation between channels is considered, and the expected value $\sum_n \mathbb{E}[\mathcal{E}(\vec{\boldsymbol{Y}}_n)]$ is used as a regularization term in [14]. In this paper, the regularization used in [14] is called observation-independent consistency regularization (OICR).

The regularization term can be represented as the quadratic form of $\boldsymbol{w}_{i,n}$ as follows:

$$\sum_n \mathbb{E}\left[\mathcal{E}(\vec{\boldsymbol{Y}}_n)\right] = \sum_{n,i,i'} \boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{G}_{ii'}^{(\mathrm{ind})}\boldsymbol{w}_{i',n}, \tag{8}$$

where $\boldsymbol{G}_{ii'}^{(\mathrm{ind})} \in \mathbb{C}^{M\times M}$ is only determined by the STFT conditions. Since $\tilde{\boldsymbol{x}}_m$ is assumed to be independent and identically distributed with respect to m, $\boldsymbol{G}_{ii'}^{(\mathrm{ind})} = g_{ii'}^{(\mathrm{ind})}\boldsymbol{E}_M$ $(g_{ii'}^{(\mathrm{ind})} \in \mathbb{C})$ holds. The cost function of ILRMA with OICR is formulated as follows:

$$\mathcal{L}_{\mathrm{OICR}} = \mathcal{L} + \beta \sum_{n,i,i'} g_{ii'}^{(\mathrm{ind})}\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{w}_{i'n}, \tag{9}$$

where $\beta > 0$ is the weight of the regularization term in ILRMA with OICR. Since the added regularization term (8) is independent of the NMF variables, the update rules for the NMF variables $\{t_{ik,n}\}$ and $\{v_{kj,n}\}$ are the same as in ILRMA. For $\{\boldsymbol{W}_i\}$, an auxiliary function is designed for the MM algorithm [16] and updated using VCD [17] instead of IP (see details in [14]). The update rule of $\{\boldsymbol{W}_i\}$ also guarantees the monotonic nonincrease in the cost function (9).

IV. Proposed method

A. Motivation and approach

Since the regularization term in [14] is formulated to be independent of the to-be-separated observed signals, the regularization cannot be appropriate for specific observed signals. Therefore, if we can introduce a regularization term that depends on the specific structure of the to-be-separated observed signals, further improvement in separation performance is expected. In this paper, we propose the use of $\sum_n \mathcal{E}(\vec{\boldsymbol{Y}}_n)$ itself calculated with the to-be-separated observed signals as a regularization term for ILRMA. The regularization term in the conventional method has a simple structure because there is no correlation between the observed channels. Unfortunately, the proposed regularization term does not always have a simple structure because there exists a correlation between the observed channels. Thus, we cannot exactly follow the technique for designing the auxiliary function as in [14] for the newly obtained optimization problem. We find out the relationship between two matrices based on the positive semidefiniteness that allows us to extend the technique in [14], and we design an auxiliary function for the proposed cost function. On the basis of the MM algorithm, we derive an update rule by applying VCD [17] to the proposed auxiliary function.

B. Design of regularization term and update rules

By expanding (7), the proposed regularization term is also deformed into the following quadratic form of $\boldsymbol{w}_{i,n}$:

$$\sum_n \mathcal{E}(\vec{\boldsymbol{Y}}_n) = \sum_{n,i,i'} \boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{G}_{ii'}\boldsymbol{w}_{i',n}, \tag{10}$$

where $\boldsymbol{G}_{ii'} \in \mathbb{C}^{M\times M}$ is dependent on not only the STFT conditions but also the to-be-separated observed signals. The specific representation of $\boldsymbol{G}_{ii'}$ will be omitted for space limitation. The cost function of the proposed method is the weighted sum of the cost function of ILRMA (6) and the regularization term (10) as follows:

$$\mathcal{L}_{\mathrm{Prop}} = \mathcal{L} + \tilde{\gamma}\sum_{n,i,i'} \boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{G}_{ii'}\boldsymbol{w}_{i',n}, \tag{11}$$

where $\tilde{\gamma} \geq 0$ is the weight of the regularization term in the proposed method. Then, we derive update rules for the cost function (11). Since the regularization term (10) is independent of the NMF variables $\{t_{ik,n}\}$ and $\{v_{kj,n}\}$ as in [14], the update rules for the NMF variables are the same as in ILRMA. Therefore, it is sufficient to derive an update rule for $\{\boldsymbol{w}_{i,n}\}$.

The cost function (11) has a correlation term between frequencies and this makes it difficult to optimize $\{\boldsymbol{w}_{i,n}\}$ simultaneously. Therefore, we focus on one frequency bin i and treat the variables related to the other frequency bins as constants. That is, $\{\boldsymbol{w}_{i,n}\}$ is updated one by one for each frequency bin in a block coordinate descent manner. The cost function is rewritten with regard to the frequency i as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{Prop}} =& J\left[\sum_n \boldsymbol{w}_{i,n}^{\mathsf{H}} \frac{1}{J}\sum_j \frac{\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathsf{H}}}{\sum_k t_{ik,n}v_{kj,n}}\boldsymbol{w}_{i,n} - \log|\det \boldsymbol{W}_i|^2\right] \\ &+\tilde{\gamma}\sum_n \begin{bmatrix}\boldsymbol{w}_{i,n}\\ \boldsymbol{w}_{i,n}^*\end{bmatrix}^{\mathsf{H}} \begin{bmatrix}\boldsymbol{G}_{ii} & \boldsymbol{G}_{i\hat{i}}\\ \boldsymbol{G}_{\hat{i}i} & \boldsymbol{G}_{\hat{i}\hat{i}}\end{bmatrix}\begin{bmatrix}\boldsymbol{w}_{i,n}\\ \boldsymbol{w}_{i,n}^*\end{bmatrix} \\ &+2\tilde{\gamma}\sum_n \mathrm{Re}\left(\begin{bmatrix}\sum_{i'\neq i,\hat{i}} \boldsymbol{G}_{i'i}^{\mathsf{H}}\boldsymbol{w}_{i',n}\\ \sum_{i'\neq i,\hat{i}} \boldsymbol{G}_{i'\hat{i}}^{\mathsf{H}}\boldsymbol{w}_{i',n}\end{bmatrix}^{\mathsf{H}}\begin{bmatrix}\boldsymbol{w}_{i,n}\\ \boldsymbol{w}_{i,n}^*\end{bmatrix}\right) + \mathrm{const.},\end{aligned} \tag{12}$$

where $\mathrm{Re}\,(\cdot)$ denotes a real part of a complex number and const. denotes the terms independent of $\boldsymbol{w}_{i,n}$. The index $\hat{i}$ is the frequency index opposite of i across the Nyquist frequency, where $i + \hat{i} = I + 2$ $(i \neq 1)$. It is difficult to optimize (12) analytically because the terms $\boldsymbol{w}_{i,n}^{\mathsf{T}}\boldsymbol{G}_{\hat{i}i}\boldsymbol{w}_{i,n}$ and $\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{G}_{i\hat{i}}\boldsymbol{w}_{i,n}^*$ are included in the cost function. Note that this problem does not occur in the case $i = 1$ (corresponding to 0 Hz) or $i = I/2$, because $\boldsymbol{w}_{i,n}$ $(i = 1, I/2)$ is real-valued to satisfy the constraint that the separated signal is real-valued in the time domain. The demixing matrix related to $i = 1, I/2$ can be updated using VCD without auxiliary functions. We design the auxiliary function without the terms $\boldsymbol{w}_{i,n}^{\mathsf{T}}\boldsymbol{G}_{\hat{i}i}\boldsymbol{w}_{i,n}$ and $\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{G}_{i\hat{i}}\boldsymbol{w}_{i,n}^*$ in the case $i \neq 1, I/2$.

In this paper, we follow the strategy used in [14] to derive the auxiliary function on the basis of Claim IV.1 (Claim 1 in [14]) for (12).

Claim IV.1 (Claim 1 in [14])**.** For any integer $q \geq 1$, $\boldsymbol{w} \in \mathbb{C}^q$, $\boldsymbol{\alpha} \in \mathbb{C}^q$, and positive semidefinite matrix $\hat{\boldsymbol{G}} \in \mathbb{C}^{q\times q}$, we have

$$\begin{aligned}\boldsymbol{w}^{\mathsf{H}}\hat{\boldsymbol{G}}\boldsymbol{w} \leq& \boldsymbol{w}^{\mathsf{H}}\hat{\boldsymbol{G}}^{+}\boldsymbol{w} \\ &-2\mathrm{Re}\left(\boldsymbol{\alpha}^{\mathsf{H}}(\hat{\boldsymbol{G}}^{+}-\hat{\boldsymbol{G}})\boldsymbol{w}\right)+\boldsymbol{\alpha}^{\mathsf{H}}(\hat{\boldsymbol{G}}^{+}-\hat{\boldsymbol{G}}))\boldsymbol{\alpha},\end{aligned} \tag{13}$$

where $\hat{\boldsymbol{G}}^{+}$ is a positive semidefinite matrix satisfying $\hat{\boldsymbol{G}}^{+} - \hat{\boldsymbol{G}} \succeq \boldsymbol{O}$. Equality holds when $\boldsymbol{\alpha} = \boldsymbol{w}$.

Let $\hat{\boldsymbol{G}}^{+}$, $\hat{\boldsymbol{G}}$, $\boldsymbol{w}$, and $\boldsymbol{\alpha}$ in Claim IV.1 be set as

$$\hat{\boldsymbol{G}}^{+} = \begin{bmatrix}\boldsymbol{G}_{ii}^{+} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{G}_{\hat{i}\hat{i}}^{+},\end{bmatrix},\ \hat{\boldsymbol{G}} = \begin{bmatrix}\boldsymbol{G}_{ii} & \boldsymbol{G}_{i\hat{i}}\\ \boldsymbol{G}_{\hat{i}i} & \boldsymbol{G}_{\hat{i}\hat{i}}\end{bmatrix},$$

$\boldsymbol{w} = [\boldsymbol{w}_{i,n}^{\mathsf{T}}, \boldsymbol{w}_{i,n}^{*\mathsf{T}}]^{\mathsf{T}}$, and $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_{i,n}^{\mathsf{T}}, \boldsymbol{\alpha}_{i,n}^{*\mathsf{T}}]^{\mathsf{T}}$, where $\boldsymbol{G}_{ii}^{+}$ and $\boldsymbol{G}_{\hat{i}\hat{i}}^{+}$ are chosen so that $\hat{\boldsymbol{G}}^{+} - \hat{\boldsymbol{G}} \succeq \boldsymbol{O}$ is satisfied. By using (13), we can derive an auxiliary function without the terms $\boldsymbol{w}_{i,n}^{\mathsf{T}}\boldsymbol{G}_{\hat{i}i}\boldsymbol{w}_{i,n}$ and $\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{G}_{i\hat{i}}\boldsymbol{w}_{i,n}^*$ for the second term of (12).

In [14], since $\boldsymbol{G}_{ii'}^{(\mathrm{ind})}$ becomes the diagonal matrix due to no correlation with respect to m, the problem of finding the matrix $\hat{\boldsymbol{G}}^{+}$ is simplified to that of finding 2×2 matrices, and the following relation is used in [14]:

$$\begin{bmatrix} g_{ii}^{(\mathrm{ind})}+|g_{i\hat{i}}^{(\mathrm{ind})}| & 0\\ 0 & g_{\hat{i}\hat{i}}^{(\mathrm{ind})}+|g_{\hat{i}i}^{(\mathrm{ind})}|\end{bmatrix} - \begin{bmatrix} g_{ii}^{(\mathrm{ind})} & g_{i\hat{i}}^{(\mathrm{ind})}\\ g_{\hat{i}i}^{(\mathrm{ind})} & g_{\hat{i}\hat{i}}^{(\mathrm{ind})}\end{bmatrix} \succeq \boldsymbol{O}. \tag{14}$$

Since $\boldsymbol{G}_{ii'}$ has nondiagonal terms due to the correlation regarding m, we cannot, however, simply follow this derivation in [14]. For this problem, we find the following relation as

$$\begin{bmatrix}\boldsymbol{G}_{ii}+(\boldsymbol{G}_{i\hat{i}}\boldsymbol{G}_{\hat{i}i})^{\frac{1}{2}} & \boldsymbol{O}_{M\times M}\\ \boldsymbol{O}_{M\times M} & \boldsymbol{G}_{\hat{i}\hat{i}}+(\boldsymbol{G}_{\hat{i}i}\boldsymbol{G}_{i\hat{i}})^{\frac{1}{2}}\end{bmatrix} - \begin{bmatrix}\boldsymbol{G}_{ii} & \boldsymbol{G}_{i\hat{i}}\\ \boldsymbol{G}_{\hat{i}i} & \boldsymbol{G}_{\hat{i}\hat{i}}\end{bmatrix} \succeq \boldsymbol{O}, \tag{15}$$

which is an extension of (14). This fact can be proved by singular value decomposition, but we omit the details for space limitation.

By using Claim IV.1 and (15), we design the total auxiliary function for (12) as

$$\begin{aligned}\mathcal{L}_{\mathrm{Prop}}^{+} =& J\left[\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{D}_{i,n}^{+}\boldsymbol{w}_{i,n} + 2\mathrm{Re}\,(\boldsymbol{b}_{i,n}^{+\mathsf{H}}\boldsymbol{w}_{i,n}) - \log|\det \boldsymbol{W}_i|^2\right] \\ &+ \mathrm{const.},\end{aligned} \tag{16}$$

where const. denotes the terms independent of $\boldsymbol{w}_{i,n}$ and $\boldsymbol{D}_{i,n}^{+}$ and $\boldsymbol{b}_{i,n}^{+}$ are respectively defined as

$$\boldsymbol{D}_{i,n}^{+} = \frac{1}{J}\sum_j \frac{\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathsf{H}}}{\sum_k t_{ik,n}v_{kj,n}} + \frac{2\tilde{\gamma}}{J}\left(\boldsymbol{G}_{ii}+(\boldsymbol{G}_{i\hat{i}}\boldsymbol{G}_{\hat{i}i})^{\frac{1}{2}}\right), \tag{17}$$

$$\boldsymbol{b}_{i,n}^{+} = \frac{2\tilde{\gamma}}{J}\left[\sum_{i'\neq i,i}\boldsymbol{G}_{i'i}^{\mathsf{H}}\boldsymbol{w}_{i',n} - (\boldsymbol{G}_{i\hat{i}}\boldsymbol{G}_{\hat{i}i})^{\frac{1}{2}}\boldsymbol{\alpha}_{i,n} + \boldsymbol{G}_{\hat{i}i}^{\mathsf{H}}\boldsymbol{\alpha}_{i,n}^{*}\right]. \tag{18}$$

Here, $\boldsymbol{\alpha}_{i,n} \in \mathbb{C}^M$ is an auxiliary variable and the equality for the auxiliary function holds when $\boldsymbol{\alpha}_{i,n} = \boldsymbol{w}_{i,n}$. Since (16) is the sum of a quadratic form, a linear term, and a log-determinant term, an update rule using VCD [17] can be obtained as follows:

$$\boldsymbol{\zeta}_{i,n} \leftarrow (\boldsymbol{W}_i\boldsymbol{D}_{i,n}^{+})^{-1}\boldsymbol{e}_n, \tag{19}$$

$$\hat{\boldsymbol{\zeta}}_{i,n} \leftarrow (\boldsymbol{D}_{i,n}^{+})^{-1}\boldsymbol{b}_{i,n}^{+}, \tag{20}$$

$$\chi_{i,n} \leftarrow \boldsymbol{\zeta}_{i,n}^{\mathsf{H}}\boldsymbol{D}_{i,n}^{+}\boldsymbol{\zeta}_{i,n}, \tag{21}$$

$$\hat{\chi}_{i,n} \leftarrow \boldsymbol{\zeta}_{i,n}^{\mathsf{H}}\boldsymbol{D}_{i,n}^{+}\hat{\boldsymbol{\zeta}}_{i,n}, \tag{22}$$

$$\boldsymbol{w}_{i,n} \leftarrow \begin{cases}\frac{\boldsymbol{\zeta}_{i,n}}{\sqrt{\chi_{i,n}}} - \hat{\boldsymbol{\zeta}}_{i,n} & (\hat{\chi}_{i,n} = 0),\\ \frac{\hat{\chi}_{i,n}}{2\chi_{i,n}}\left(1-\sqrt{1+\frac{4\chi_{i,n}}{|\hat{\chi}_{i,n}|^2}}\right)\boldsymbol{\zeta}_{i,n} - \hat{\boldsymbol{\zeta}}_{i,n} & (\text{otherwise}),\end{cases} \tag{23}$$

where $\boldsymbol{e}_n \in \mathbb{C}^N$ is a vector whose nth component is one and the others are zero. This update rule guarantees the monotonic nonincrease in the cost function (11). All the update rules in

TABLE I: Averages of SDR improvements over 100 trials for each method

		ILRMA w/ OICR			Prop. w/o temp.			Prop. w/ temp.		
ILRMA	Consist. ILRMA	$\beta = 10^{-6}$	$\beta = 10^{-5}$	$\beta = 10^{-4}$	$\gamma = 10^{3}$	$\gamma = 10^{4}$	$\gamma = 10^{5}$	$\gamma_0 = 10^{3}$	$\gamma_0 = 10^{4}$	$\gamma_0 = 10^{5}$
9.6	11.0	9.6	9.5	8.7	10.5	10.8	5.3	10.4	**11.2**	8.7

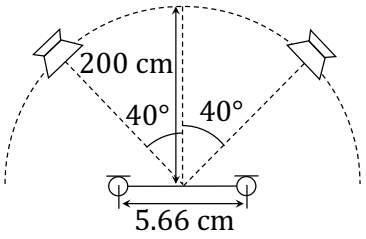
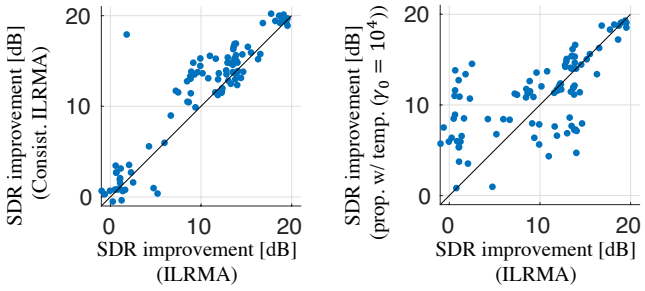


Fig. 2: Recording conditions of impulse response.



(a) (b)

Fig. 3: Pairplots for SDR improvements of ILRMA and (a) Consist. ILRMA or (b) prop. w/ temp. ($\gamma_0 = 10^4$).

the proposed method including those for the NMF variables $\{t_{ik,n}\}$ and $\{v_{kj,n}\}$ guarantee the monotonic nonincrease in the cost function (11).

C. Weight of regularization

In this paper, to prevent the effect of the regularization term from varying with the power and length of the observed signals, we normalize the regularization weight $\tilde{\gamma}$ as

$$\tilde{\gamma} = \gamma \cdot \frac{J}{\sum_{i,j} \|\boldsymbol{x}_{ij}\|_2^2} \tag{24}$$

and design $\gamma \geq 0$ instead.

In some regularization-based methods, the scheduling of the weight parameter is often performed. We use "tempering," in which the weight parameter is reduced with increasing iteration number, because the regularization to be consistent may conflict with the separation. In the tempering of γ, we set the the parameter γ at the cth iteration as $\max\{0,\ \gamma_0 - \gamma_0(c-1)/(C_0-1)\}$. Note that the parameter γ is γ_0 at the first iteration, decreases linearly thereafter, and is 0 after the C_0th iteration.

V. Simulation Experiment

A. Experimental conditions

To confirm the performance of the proposed method, a simulation experiment was conducted using two sources and two microphones ($M = N = 2$). Ten pairs of music signals from SiSEC2011 [19] were prepared as dry sources. They were convolved with the impulse responses E2A ($T_{60} = 300$ ms) in the RWCP database [20]. The sources and microphones were placed as shown in Fig. 2. Both the dry sources and the impulse responses were resampled at 16 kHz. The STFT was computed with a window length of 256 ms, a hopsize of 64 ms, and the Hamming window.

We compared ILRMA [7], Consist. ILRMA [13], ILRMA with OICR (ILRMA w/ OICR) [14], the proposed method with constant γ (prop. w/o temp.), and the proposed method with the tempering of γ (prop. w/ temp.). The number of bases for NMF K was 10. The demixing matrix was initialized to the identity matrix and the NMF variables were initialized with uniformly distributed random values over $(0, 1)$. Ten trials were performed using different random seeds, i.e., we conducted experiments under 100 conditions (10 pairs of sources $\times$ 10 random seeds). The number of iterations was 400. We used $\beta = 10^{-6},\ 10^{-5},\ 10^{-4}$ in ILRMA w/ OICR, $\gamma = 10^3,\ 10^4,\ 10^5$ in prop. w/o temp., and $\gamma_0 = 10^3,\ 10^4,\ 10^5$ and $C_0 = 50$ in prop. w/ temp. The source-to-distortion ratio (SDR) [21] improvement was used as a measure of the separation performance.

B. Results and discussion

Table I shows the average of SDR improvements. Prop. w/o temp. and prop. w/ temp. improved the separation performance compared with ILRMA except for $\gamma = 10^5$ in prop. w/o temp. and $\gamma_0 = 10^5$ in prop. w/ temp., in which the initial value of γ was very large. In particular, prop. w/ temp. ($\gamma_0 = 10^4$) showed an average SDR improvement of 0.2 dB higher than Consist. ILRMA. In this experiment, ILRMA w/ OICR showed a lower separation performance than ILRMA, but this may be due to differences between the evaluation data used in [14] and in this study. Fig. 3 shows the pairplots for the SDR improvements of ILRMA with Consist. ILRMA and prop. w/ temp. ($\gamma_0 = 10^4$). Prop. w/ temp. ($\gamma_0 = 10^4$) showed higher SDR improvements than Consist. ILRMA in many cases when the SDR improvements of ILRMA were lower than 5 dB. The proposed regularization term seems to prevent the ILRMA from converging to poor solutions.

VI. Conclusion

Several methods have been proposed with the spectrogram consistency considered in ILRMA. One of them is a

regularization-based method that induces the separated signals to be consistent. Although the regularization term independent of the observed signals is used in the conventional method, we designed the regularization term that depends on the observed signals. For the optimization of the new cost function with the new regularization term, we designed the auxiliary function that is an extension of that in the conventional method. By applying VCD to the auxiliary function, we derived the update rule that guarantees the monotonic nonincrease in the cost function. A numerical experiment confirmed that the proposed method with the tempering of the weight parameter of the regularization showed a higher performance than the conventional methods.

ACKNOWLEDGMENT

The authors would like to thank Rintaro Ikeshita and Tomohiro Nakatani (NTT) for useful discussions.

REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[3] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. ICASSP*, vol. 5, 2000, pp. 3140–3143.

[4] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. SAP*, vol. 12, no. 5, pp. 530–538, 2004.

[5] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," in *Proc. ISCAS*, 2007, pp. 3247–3250.

[6] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.

[7] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[8] D. Kitamura, N. Ono, and H. Saruwatari, "Experimental analysis of optimal window length for independent low-rank matrix analysis," in *Proc. EUSIPCO*, 2017, pp. 1170–1174.

[9] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. ASSP.*, vol. 32, no. 2, pp. 236–243, 1984.

[10] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE SP Letters*, vol. 20, no. 3, pp. 217–220, 2012.

[11] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin–Lim algorithm," in *Proc. WASPAA*, 2013, pp. 1–4.

[12] K. Yatabe, "Consistent ICA: Determined BSS meets spectrogram consistency," *IEEE SP Letters*, vol. 27, pp. 870–874, 2020.

[13] D. Kitamura and K. Yatabe, "Consistent independent low-rank matrix analysis for determined blind source separation," *EURASIP JASP*, vol. 2020, no. 46, pp. 1–35, 2020.

[14] S. Misawa, N. Takamune, K. Yatabe, D. Kitamura, and H. Saruwatari, "Blind source separation using independent low-rank matrix analysis with spectrogram-consistency regularization," in *Proc. APSIPA*, 2023, pp. 1035–1042.

[15] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[16] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *JCGS*, vol. 9, no. 1, pp. 60–77, 2000.

[17] N. Makishima, Y. Mitsui, N. Takamune, *et al.*, "Independent deeply learned matrix analysis with automatic selection of stable microphone-wise update and fast sourcewise update of demixing matrix," *Signal Processing*, vol. 178, pp. 1–12, 2021.

[18] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[19] S. Araki, F. Nesta, E. Vincent, *et al.*, "The 2011 signal separation evaluation campaign (SiSEC2011): - Audio source separation -," in *Proc. LVA/ICA*, 2012, pp. 414–422.

[20] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. LREC*, 2000, pp. 965–968.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

Teager Energy Cepstral Coefficients for Spoken Language Identification

Arth J. Shah*, Savita H. Yadav† and Hemant A. Patil*
* Dhirubhai Ambani Institute of Information and Communication Technology, Gujarat, India
* E-mail: {202101154, hemant_patil}@daiict.ac.in
† Uka Tarsadia University, Bardoli, Gujarat, India
† E-mail:savita22125@gmail.com

***Abstract*—Spoken Language Identification (SLID) is a key component in audio processing that facilitates the recognition and understanding of audio clips of spoken languages. Various applications, such as automatic speech recognition, multilingual voice assistants, and real-time translation services need SLID capabilities. Effective SLID enhances the transcription and processing of spoken language data and improves user experiences by providing personalized and relevant responses in the correct language. This study propose Teager Energy Cepstral Coefficients (TECC) features to capture the characteristics of spoken language. To evaluate how well TECC performs for SLID, we conducted a comparative analysis of its performance across various spectral features, namely, Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC). Two classifiers, deep residual networks (ResNet-50) and Time Delay Neural Networks (TDNN), were employed in this comparison. To maximize the performance of SLID system, we applied feature-level fusion and score-level Fusion techniques to advance the state-of-the-art. Additionally, latency analysis assesses time optimization to ensure the system operates efficiently. We obtained 98.25 % accuracy in this study for two languages, and 84.25 % on 10 different languages using TECC features.**

***Index Terms*—Spoken Language Identification, Teager Energy Cepstral Coefficients, ResNet-50, Feature-Level Fusion, Score-Level Fusion.**

I. Introduction

The task of automatically identifying language being used by a user from a sample of speech, regardless of the speaker's accent, gender, or identity is called as Spoken Language Identification (SLID) task. Solving the issue of SLID, has enormous potential in real-life cases, such as multilingual spoken translation, human-machine communication systems, spoken language retrieval, multilingual speech transcription system, and many other language-based tasks. Beside solving all these issues, also helps users in many international affairs, such as smooth communications between people of other countries. The proposed system has the potential to work seamlessly with automatic speech recognition, multilingual voice assistants, and real-time translation services, enhancing communication across different languages. The capability of proposed method enhances the transcription and processing of spoken language data, ensuring accurate interpretation. In our study, we employed a system that can identify 10 international languages from raw speech waveforms. We employ signal processing-based front-end feature extraction, and deep learning classifiers as back-end for SLID task. Numerous contemporary studies employ characteristics that provide insights into the structure of the vocal tract system, utilizing the acoustics of speech, which can be derived from Mel Frequency Cepstral Coefficients (MFCC) features [1], [2]. Such acoustic attributes encompass distinct phonemes for each language, thereby assisting in differentiating between multiple languages. Nevertheless, MFCC features are notoriously susceptible to noisy surroundings and variations in speaking styles.

In this paper, we propose use of Teager Energy Operator (TEO)-based features, namely, Teager Energy Cepstral Coefficients (TECC) features for SLID task. While the MFCC features align with human auditory perception and the LFCC features provide an equal frequency representation, TECC captures the vocal tract energy and the non-linear characteristics of speech signals. Deep learning classifiers, such as ResNet-50, and TDNN have been used in this study, in order to leverage the proposed methodology with recent technological advancement. Experimental results demonstrate that the combination of TECC features with the ResNet-50 classifier achieves better accuracy for LID as compared to the TDNN classifier (to be discussed in Section IV-A). Similar observations can be observed on MFCC, and LFCC features, with both the classifiers. Traditional classification methods have been used previously for many studies, such as in [3], [4], authors employ acoustic features with statistical methods, such as the Support Vector Machine (SVM), Hidden Markov Model (HMM), Universal Background Model (UBM), and Gaussian Mixture Model (GMM) for classification of acoustic features. Motivated by employment of residual-based features with TDNN classifier for SLID task [5], we employ TECC-based features with TDNN and ResNet-50 classifier.

In order to obtain optimal results by exploring variety of features, we also reported the results using two types of data fusion strategies (namely, feature-level, and score-level), which resulted in a further increase in accuracy. Although the proposed TECC feature vector does not perform at its optimal capacity, we successfully provide logical reasoning and analysis for the results obtained.

A. Related Works

Many existing studies in recent days have boost the research potential on SLID tasks. Some factor, that remains common in most of the studies on SLID tasks, till date are dataset

(VoxLingua107), and use of recent deep learning models. In [5], authors employ linear prediction residual-based features for SLID task for classification of Indian languages. For countries with multilingual and large population, such as India, SLID has key importance. For diversified countries, such as India, where there are more than 720 languages, and Papua New Guinea (840 languages), SLID plays a critical role. In [6], authors employ CNN-based method to identify 5 languages, with CNN classifier and spectrograms as features. However, study in [6] fails to demonstrate hard-link relationship between spectrograms and speech acoustics. In [7], authors use LSTM as a classifier and obtained 86 % accuracy. ECAPA-TDNN and MFCC30 have also been employed for SLID task, which reported an accuracy of 76.8 % for 8 language classes. Many other studies, were also reported for SLID task, with advancement in self-supervised learning based pre-trained models, such as whisper [8], wav2vec2.0 [9], HuBERT [10], and XLS-R [11]. Proposed methodology provides the following novelty:

- This study discussed how the energy of vocal tract system, and non-linearity in speech signal contribute to language-specific information for SLID task.
- Technical reasoning on energy difference of signals of different languages.
- As the SLI task should be robust against smaller speech segments, we show our model's latency period for the proposed TECC features and also for fair comparison with existing works [5].
- For improvement of results, we also did various types of fusion, namely, feature-level, and score-level.

II. Proposed Approach

In this Section, we explain the motivation for TEO, development of TEO, and computational detail's of TECC feature extraction.

A. Energy of Real Physical System

In [12], authors proposed that within a single pitch period, speech can be represented as a combination of signals, where the amplitude and frequency vary over time, known as AM-FM signals. This implies that Simple Harmonic Motion (SHM) can help to explain the concept of energy in speech wave. By applying Newton's second law of motion to the mass-spring system with spring constant C and mass N, it gives the dynamics as the 2nd order linear differential equation [13] :

$$\frac{d^2y}{dt^2} + \frac{C}{N}y = 0, \tag{1}$$

whose solution is recognized as Simple Harmonic Motion (SHM) :

$$y(t) = A\cos(\Omega_0 t + \phi), \tag{2}$$

where A is amplitude, and Ω_0 denotes the angular frequency (in radians) of oscillations. The previous solution can explained in following way. Any periodic function can be decomposed into Fourier series, which consists of an infinite sum of sine waves of varying frequencies. The general solution of such a 2nd order linear differential equation, where C and N are positive constant, form $e^{\pm j\sqrt{(C/N)}}$ which ultimately results the form $y(t) = A\cos(\Omega_0 t + \phi)$ [14].

The total energy consists the combined potential energy in the spring and the kinetic energy of the mass, i.e,

$$E = P.E. + K.E. = \frac{1}{2}Cx^2 + \frac{1}{2}N\left(\frac{dy}{dt}\right)^2. \tag{3}$$

Substituting $y(t) = A\cos(\Omega_0 t + \phi)$ and using trigonometry,

$$E = \frac{1}{2}NA^2\Omega^2, \tag{4}$$

$$\text{Or,} \quad E \propto A^2\Omega_0^2. \tag{5}$$

Eq. (2) is the true energy required to generate $y(t)$ in SHM.

B. Development of TEO

Kaiser proposed following algorithm for continuously estimating the energy level present in a signal, or the energy needed to generate the signal [3]. In discrete-time domain Eq. (2) can be expressed as:

$$y(n) = A\cos(\omega_0 n + \phi). \tag{6}$$

From (6), we can write:

$$y(n+1) = A\cos(\omega_0(n+1) + \phi), \tag{7}$$

$$y(n-1) = A\cos(\omega_0(n-1) + \phi). \tag{8}$$

Multiplying (7) and (8) and using trigonometry,

$$\begin{aligned} &y(n+1)y(n-1) \\ &= A^2\cos(\omega_0(n+1)+\phi)\cos(\omega_0(n-1)+\phi), \\ &= [A\cos(\omega_0 n + \phi)]^2 - A^2\sin^2\omega_0. \end{aligned} \tag{9}$$

Using Eq. (6) in Eq. (9),

$$A^2\sin^2\omega_0 = y^2(n) - y(n+1)y(n-1). \tag{10}$$

For low values of $\omega_0, \sin\omega_0 \approx \omega_0$, hence Eq. (10) can be written as

$$A^2\omega_0^2 \approx y^2(n) - y(n+1)y(n-1) = \psi\{y(n)\}, \tag{11}$$

where $\psi\{.\}$ is the TEO, which gives a running estimate of the energy of discrete-time signal $x(n)$. Fig. 1 represents functional block diagram of proposed SLID system using TECC features. TECC is known to capture energy of the vocal tract system and the non-linearities in the speech signal, which plays an crucial roles in extracting language-specific information. This energy is reflected in the formant frequencies, which are the resonant frequencies of the vocal tract system. Different languages utilize distinct sets of vowel and consonant sounds, which are characterized by their formant frequencies. The energy distribution across different frequency bands (spectral envelope) provides information about the articulation of phonemes. Different languages have unique spectral patterns due to their specific phoneme inventories, which may be key acoustic cue for SLID problem. The energy patterns also reflect prosodic features, such as stress, rhythm, and intonation, which

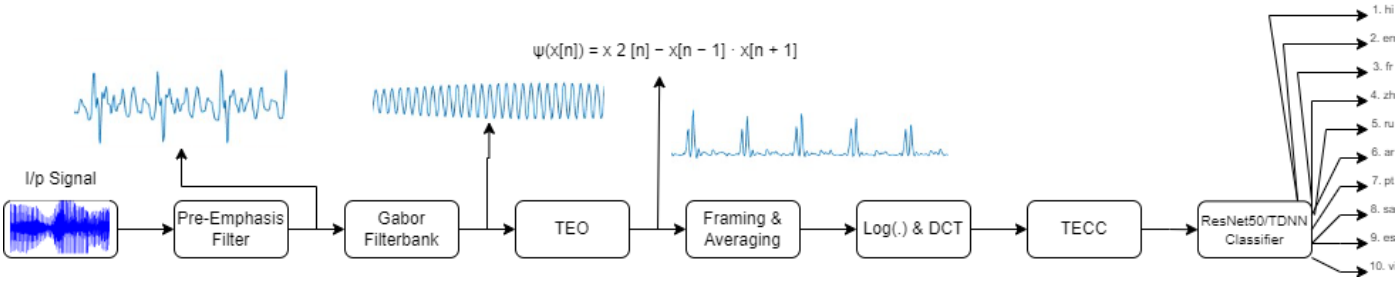


Fig. 1. Functional block diagram of proposed SLID system using TECC features, with ResNEt-50 / CNN classifier.

vary between languages and convey important linguistic information. The non-linear nature of speech production mechanism creates harmonics and overtones, which are crucial for the perception of timbre and the distinction of different phonetic elements. These patterns are also language-specific and can be used to identify phonemes and prosodic features. Hence, TECC works better with MFCC, and LFCC combined after data fusion for SLID task (to be discussed soon).

III. Experimental Setup

A. Dataset Used

We employed statistically meaningful VoxLingua107 dataset, which is a comprehensive multilingual speech corpus, designed for training language identification models. It comprises over 6,800 hours of speech data spanning 107 different languages. The recordings in this dataset are sourced from YouTube, ensuring a diverse range of accents, dialects, and speaking styles. The dataset is annotated at the segment-level, allowing for precise language identification. It includes a variety of speech contexts, such as news, conversations, and lectures. We selected 10 languages, namely (labels), Hindi (hi), English (en), French (fr), Mandarin Chinese (zh), Russian (ru), Arabic (ar), Portuguese (pt), Sanskrit (sa), Spanish (es), and Vietnamese (vi). 4000 random samples from each language were taken to conduct experiments, with average time spam of 9.4 seconds for each sample (1K samples ranging from each range {0-5, 5-10, 10-15, and 15-20 seconds}). We accounted around 11.3 hours of data for each of 9 languages (for Sanskrit, 9 hours data was taken), resulting into a total of 111.16 hours of vast data.

B. Classifiers Used

1) ResNet-50: ResNet contains four blocks within each block. The first block has three convolutional layers, followed by four, six, and three convolutional layers, respectively. Batch normalization and ReLU activation functions are applied after each convolutional layer. After the main blocks, there is a global average pooling layer that reduces the spatial dimensions of the feature maps. This is followed by a fully-connected layer with a softmax activation function, which produces the final output probabilities for different classes. This architecture is also known as *ResNet-50* [15].

2) Time Delay Neural Networks (TDNN): Motivated by previous works on SLID, authors in this study also decided to employ TDNN as classifier [5]. The TDNN architecture's ability to process sequential data and capture long temporal contexts proves advantageous for SLI tasks [16]. The incorporation of an attention mechanism facilitates the emphasis on more prosodically and linguistically significant frames within an utterance, enhancing the model's performance. The input shape to both the classifiers was constant, i.e., 20 * 750 (number coefficients*maximum number of frames). In SLID task, some frame-level features within an utterance exhibit more pronounced and unique prosodic characteristics. To tackle this, study reported in [17] has integrated attention mechanism to assign weights to these frame-level features. Specifically, [18] employs this attention mechanism on TDNN for SLID. With similar motivation, this study also employs an attention mechanism.

C. Cepstral Features Used

We employed MFCC and LFCC as baseline features for comparison with TECC. MFCC, and LFCC well known to capture human perspective-based characteristics, on Mel scale (human auditory scale), as well as linear scale, gives an perfect case, which can be compared with TECC, that captures energy of vocal tract system, and the non-linearity as in speech signal. MFCC has high resolution on low-frequency region, and has low resolution at high-frequency regions. LFCC on other hand, has linear frequency resolution across entire frequency axis. Such characteristics help in covering all necessary aspects for classification of speech signals based on a particular language. We employed static features for both MFCC and LFCC, of variable dimension (soon to be discussed in sub-Section IV-A), and for each feature sets, the frame length was kept to be 30 *ms*, and shift was 15 *ms*. Number of subband filters was taken as 40 (fixed for all the feature sets for fair comparison), and NFFT was taken to be 512.

IV. Experimental Results

A. Effect of Dimension of Feature Vector

The dimension of feature vector is the number of coefficients per audio frame (i.e., speech segment) obtained from speech signal after framing and windowing. As number of coefficients per frame increase, the extracted amount of features per unit file increases, resulting in more storage, space, and time

complexity (for both feature extraction and classification task). However, with more amount of data per frames, we can't get more clarity on data, that which data belongs to which class, thereby leading to cutting edge classification of classes. On the other hand, while number of coefficients increase than maximum number of required coefficients, then it starts *over-fitting* the model, leading to declassification of the model. For fair comparison with cepstral features, we also fine-tuned both comparative cepstral features. For MFCC, we obtained optimal feature dimension as 20, whereas for LFCC, and TECC, we obtained optimal feature dimension as 16. We obtained accuracy of 96.37 % for LFCC, and 76.37 % for TECC on ResNet-50 classifier as shown in Fig. 2. All the experiments of this sub-Section were performed using TDNN and ResNet-50 classifiers. Results of TDNN classifier were not as much good as for ResNet classifier, thereby the results of TDNN feature dimention can be found on[1]. We got results of TECC feature vector better on ResNet-50 classifier and hence, we selected ResNet as optimal classifier. Although we are getting much less accuracy on proposed TECC vector, we decided to explore data fusion of features to obtain possible complimentary observations.

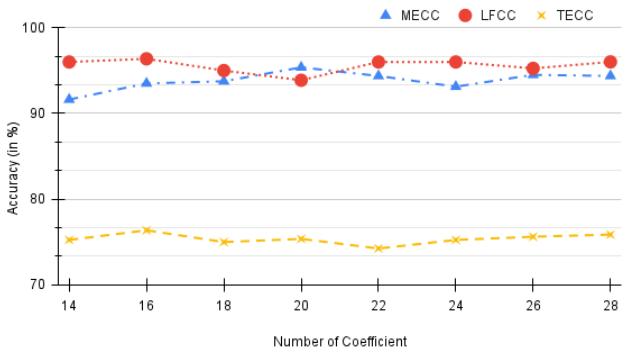


Fig. 2. Effect of dimension of feature vector using ResNet-50 classifier.

B. Effect of Number of Language Classes

With increase in number of languages, the complexity of data points increases, i.e., the probability of classifier classifying the data point correctly decreases. With a logistical observation, as the number of class increase, the number of clusters of data points formed increases, resulting in a more degraded classification performance. As the data points that have closely resembling phonemes, vowels may result into unequal classification of languages. Similar observations have been made around 17 years ago with two similar-sounding languages (namely, Hindi and Urdu) with TEO operator itself [19]. Fig. 3 denotes the decrease in accuracy with increase in number of languages. Similar observations can be made for both the comparative cepstral features, as well.

[1]"https://github.com/ARTHARKING55/TDNN_dimention/tree/main"

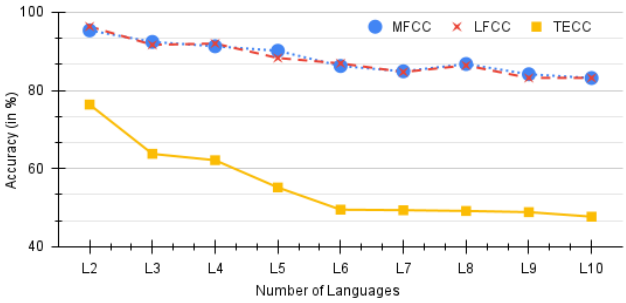


Fig. 3. Variation in accuracy with increase in number of language classes.

C. Score-Level Fusion

We examine score-level fusion in order to increase the accuracy of TECC features in this sub-Section. For comparison, we fused all three features, i.e., LFCC+MFCC, LFCC+TECC, and MFCC+TECC on both classifiers. As we obtained better results on ResNet-50 classifier and hence, we did all the experiments of this sub-Section on ResNet-50 classifier. Score-level fusion includes parameter, namely, α. Score-level fusion can be calculated by following formula :

$$Score_{TECC}(\alpha) + Score_{MFCC}(1-\alpha) = Score_{fusion}. \quad (12)$$

We obtained highest of 86.62 % of accuracy for 10 language classes, which is 3.42 % higher than individual optimal accuracy obtained for 10 languages. Fig. 4 denotes score-level fusion on MFCC and LFCC features. Fusion of all other features can be found at[1].

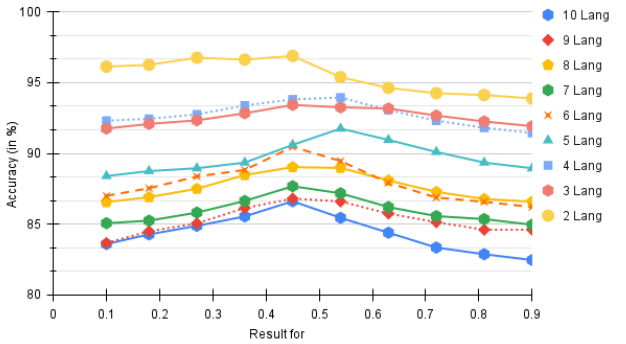


Fig. 4. Score-level fusion of MFCC and LFCC on ResNet-50 classifier.

D. Feature-Level Fusion

This sub-Section exploits feature-level fusion after exploring score-level fusion. Motivated by increase in accuracy at higher number of languages on score-level fusion, we explored feature-level fusion in order to examine the change in accuracy. Feature-level fusion can be also as stack fusion, as it is classification of two or more features, by stacking it one over other. We did double fusion as well as triple fusion in this study. In Fig. 5, an increase in accuracy of $\approx 4\%$ can be observed for two languages as compared to individual accuracies. We can further say from these observations that,

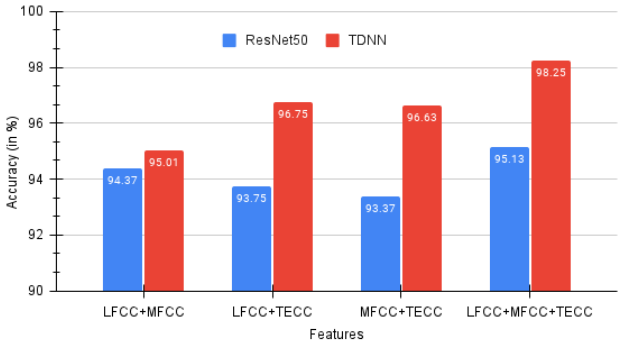


Fig. 5. Feature-level fusion of features on both TDNN and ResNet-50 classifier.

combined information of signal, i.e., energy of vocal tract system, nonlinearity in speech signal (TECC), human perception of audio information (MFCC), and linear frequency resolution-based human perception (LFCC), combine perform better than MFCC, LFCC, and TECC alone. Due to limited resources and time, we could not do feature-level fusion of more than two languages (Hindi and English), thereby leaving it to future tasks.

E. Analysis of latency Period

Latency Analysis (LA) of SLID leverages capability of system to its real-time development and usability. LA helps in finding out that in how much less time and storage, we can obtain promising results with acceptable performance. LA

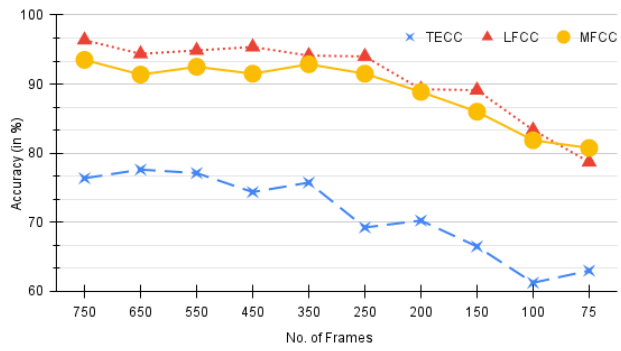


Fig. 6. Analysis of latency period using ResNet-50 classifier.

thereby helps in saving storage and time both. Moreover, from practical system deployment perspective, latency analysis is one of the key factor for evaluation of any systems' capability. LA have also been performed on SLID task previously in [5]. For fair analysis, we performed latency analysis on both cepstral features also and have reported them in Fig. 6. Observation of Fig. 6, denotes that even after selecting 350 frames instead of 750 frames, we see an accuracy decrease of only 0.62 %, whereas for MFCC and LFCC, accuracy decreased by 1.99 % and 2.25 %. This decrease in frames and minor decrease in results denote the superiority of proposed TECC-based model over the other models in realistic scenarios.

F. Comparison with Existing Studies

This sub-Section compares the proposed work with existing works on SLID task. Most of studies do not employ full dataset of VoxLingua107, due to limitations of resources and storage. We compared our results with various studies that have variable languages. Table I denotes the comparison of proposed methods with existing methods. For fair comparison, we compared our results with same number of language classes as in existing works. Key limitations in existing studies [20]–[28], is they employ either deep learning based methods, or are perceptually relevant for capturing language specific information, whereas we employ an method that captures acoustical information based on language specific characteristics.

TABLE I
COMPARISON WITH EXISTING WORKS AND PROPOSED WORK ON DIFFERENT NUMBER OF LANGUAGES.

Source	Number of Languages	Existing Results (in %)		Proposed Results (in %)	
		Accuracy	EER	Accuracy	EER
[20]	2	-	2.26	**98.25**	**1.75**
[20]	3	-	53.27	**92.67**	**5.5**
[21]	4	80.21	-	**92.88**	**4.75**
[22]	6	83	-	**87.54**	**7.47**
[23]	8	70.9	-	**86.72**	**7.58**
[24]	10	-	50	**84.25**	**8.75**
[25]	3	83.50%	-	**92.67**	**5.5**
[26]	5	88.41	7.24	**91.75**	**5.15**
[27]	10	-	11.35	**84.25**	**8.75**
[28]	6	-	8.4	**87.54**	**7.47**

V. SUMMARY AND CONCLUSIONS

In this study, we employed TECC feature vector, which captures vocal tract energy from the audio speech of spoken languages. TECC offers competitive performance compared to traditional MFCC and LFCC features for SLID tasks. ResNet-50 demonstrates superior effectiveness in utilizing TECC features, achieving higher accuracy than TDNN, For better real implementation of model proposed, we also performed various fusion techniques, in which it will enhance the classification accuracy of SLID. In order to compare our study to other existing approaches, in Section IV, we compared our results with existing optimal approaches for SLID, and found to be better.

REFERENCES

[1] D. Deshwal, P. Sangwan, and D. Kumar, "Feature extraction methods in language identification: A survey," *Wireless Personal Communications*, vol. 107, pp. 2071–2103, 2019.

[2] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[3] P. Heracleous, K. Takai, K. Yasuda, Y. Mohammad, and A. Yoneyama, "Comparative study on spoken language identification based on deep learning," in 26th *European Signal Processing Conference (EUSIPCO)*, 2018, Rome, Italy, pp. 2265–2269.

[4] D. Martinez, L. Burget, L. Ferrer, and N. Scheffer, "I-vector-based prosodic system for language identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, Kyoto, Japan, pp. 4861–4864.

[5] B. S. Hora, K. Parmar, S. Machhar, H. A. Patil, K. Praveen, and B. Radhakrishnan, "Exploring residual cepstral features for spoken language identification," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, Taipei, Taiwan, pp. 131–138.

[6] S. S. Sapkota, A. Shakya, and B. Joshi, "Spoken Language Identification Using Convolutional Neural Network In Nepalese Context," in 26th *Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, IEEE, 2023, Delhi, India, pp. 1–6.

[7] G. Satyanarayana, D. Varun, D. Sandeepika, and B. Uma, "Spoken language detection using deep learning,"

[8] M. Wang, Y. Li, J. Guo, *et al.*, "Whislu: End-to-end spoken language understanding with whisper," in *Proc. INTERSPEECH*, vol. 2023, Dublin, Ireland, 2023, pp. 770–774.

[9] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020, {Last Accessed Date : 2nd July, 2024}.

[10] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification, and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021, {Last Accessed Date : 2nd July, 2024}.

[11] S. Gupta, K. S. S. Motepalli, R. Kumar, V. Narasinga, S. G. Mirishkar, and A. K. Vuppala, "Enhancing language identification in indian context through exploiting learned features with wav2vec2. 0," in *International Conference on Speech and Computer, LNCS*, Springer, vol. 14339, 2023, pp. 503–512.

[12] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[13] V. Tomar and H. A. Patil, "On the development of variable length Teager energy operator (VTEO)," in *INTERSPEECH*, 2008, Brisbane, Australia, pp. 56–59.

[14] H. A. Patil, "Speaker recognition in indian languages: A feature based approach," *Indian Institute of Technology Kharagpur (IIT-K), Ph. D Thesis*, 2005.

[15] B. Koonce and B. Koonce, "ResNet-50," *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pp. 63–72, 2021.

[16] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015, Dresden, Germany, pp. 3214–3218.

[17] F. R. rahman Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, "Attention-based models for text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, Alberta, Canada, pp. 5359–5363.

[18] T. Mandava and A. K. Vuppala, "Attention based residual-time delay neural network for indian language identification," in *12th International Conference on Contemporary Computing*, 2019, Noida, India, pp. 1–5.

[19] H. A. Patil and T. Basu, "Cepstral domain Teager energy for identifying perceptually similar languages," in *Pattern Recognition and Machine Intelligence: Second International Conference, PReMI, Kolkata, India, December 18-22, 2007. LNCS, Springer*, pp. 455–462.

[20] H. Kim and J.-S. Park, "Automatic language identification using speech rhythm features for multi-lingual speech recognition," *Applied Sciences*, vol. 10, no. 7, p. 2225, 2020.

[21] G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, and M. Masud, "Spoken language identification using deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, p. 5 123 671, 2021.

[22] A. Draghici, J. Abeßer, and H. Lukashevich, "A study on spoken language identification using deep neural networks," in *Proceedings of the 15th International Audio Mostly Conference*, 2020, pp. 253–256.

[23] R. Zazo, A. Lozano-Diez, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez, "Language identification in short utterances using long short-term memory (lstm) recurrent neural networks," *PloS one*, vol. 11, no. 1, e0146917, 2016.

[24] Z. Ma and H. Yu, "Language identification with deep bottleneck features," *arXiv preprint arXiv:1809.08909, Sep 18*, 2018, {Last Accessed Date : 2nd July, 2024}.

[25] G. Montavon, "Deep learning for spoken language identification," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009, Vancouver, Canada, pp. 1–4.

[26] K. Parmar, B. S. Hora, S. Machhar, H. A. Patil, K. Praveen, and B. Radhakrishnan, "Spoken language identification using linear frequency residual cepstral coefficients," in *International Conference on Pattern Recognition and Machine Intelligence, LNCS*, Springer, 2023, pp. 724–733.

[27] H. S. Das and P. Roy, "Optimal prosodic feature extraction and classification in parametric excitation source information for Indian language identification using neural network based Q-learning algorithm," *International Journal of Speech Technology*, vol. 22, pp. 67–77,

[28] G. Gelly, J.-L. Gauvain, V. B. Le, and A. Messaoudi, "A divide-and-conquer approach for language identification based on recurrent neural networks.," in *INTERSPEECH*, San Francisco, USA, 2016, pp. 3231–3235.

Voice Liveness Detection Using Linear Frequency Residual Cepstral Coefficients

Arth J. Shah, Nandini V. Mandaviya, and Hemant A. Patil
Dhirubhai Ambani Institute of Information and Communication Technology, Gujarat, India
E-mail: {202101154, 202201487, hemant_patil}@daiict.ac.in

***Abstract*—This study deals with the problem of voice liveness detection (VLD) in order to increase biometric security. We propose Linear Frequency Residual Cepstral Coefficients (LFRCC) features, and Bi-directional Long Short-Term Memory (BiLSTM) classifier to ascertain whether the speech is live or not. We employed the POp COrpus (POCO) dataset, which is a standard dataset for the VLD problem. We compare our work with various other well known spectral features, such as Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC) using BiLSTM as pattern classifier. Our optimal results indicate 83.14 % accuracy; which is 20.85 % greater than the existing STFT-based approach. To investigate practical suitability of VLD system using the proposed LFRCC feature set, analysis of latency period, feature-fusion, and comparison with pre-trained models is also presented in this study.**

***Index Terms*—Voice Liveness Detection, Linear Prediction, Linear Frequency Residual Cepstral Coefficients, BiLSTM, Biometrics Security.**

I. Introduction

In contemporary times, discerning the authenticity of visual and auditory information has become increasingly challenging, owing to the rapid advances in technology. It often happens when a person tampers an audio or uses replayed audio to perform fraudulent activities. Such acts became critical to detect and deal with. Many times, such audios have created a huge fuss due to such fraudulent activities. Politicians and celebrities are among the top people, who become victims of such activities due to openly available footprints in the form of biometrics, such as audio and videos. To safeguard the reliability of audio, we need an optimal system that is unbiased in terms of classification. It is of utmost importance to take necessary actions to detect and stop such fraudulent activities. Motivated by such a challenging task, we propose a system that provides promising and reliable results for the Voice Liveness Detection (VLD) task. Before these, many attempts have been made for VLD, which unfortunately failed to provide promising results thereby creating a threat to reliability. Many previous studies also employ the most popular ASV Spoof dataset for VLD. However, in this study, we employ a popularly known POp COrpus (POCO) dataset for VLD. Alternatively, this study employs the POCO dataset and uses pop noise as a key to ascertain whether the voice is live or not. The study reported in [1] also used pop noise as a key for VLD, which resulted in a better classification of spoofed *vs.* genuine speech. In this study, we employ traditional Linear Prediction (LP)-based features in combination with the popular Deep Learning (DL) model, namely, Bi-directional Long Short-Term Memory (BiLSTM). Previously, LP has been used for various tasks, such as language identification [2], speaker recognition [3], replay attack detection [4], etc.

Linear Prediction Residual (LPR) being well-known for capturing Glottal Closure Instants (GCIs), motivated us to employ LP residual to analyze characteristics of the pop noise in the signal [3]. LPR is also renowned for capturing characteristics, such as stress, which remains unaffected by changes in a speaker's styles and environmental noise [3]. In this study, we did not directly use LPR as a feature, instead, we used LPR as an input instead of a speech signal to Linear Frequency Cepstral Coefficients (LFCC) feature extraction pipeline to extract linear frequency scale-based properties from the voice signal. This combination emerges to form a meaningful feature set, namely, Linear Frequency Residual Cepstral Coefficients (LFRCC), which extracts features similar to the LFCC feature set but takes LPR signal (i.e., excitation source) as input instead of raw audio, containing both vocal tract system and excitation source information.

A. Related work

Many studies have been using several features to solve VLD task previously. Among them, the first task to propose the POCO dataset [1], and to use it has been chosen as a baseline for this work. Studies state that many features and classifiers, such as Constant Q Transform (CQT)-based features and discriminative Support Vector Machine (SVM) classifier have been explored in [5]. Phase-based features (Group Delay (GD) function) have been exploited for VLD task [4], which did not obtain satisfactory results, even after exploiting both magnitude and phase-based features. In [6], [7], authors exploited the ResNet classifier with features, such as Constant Q Cepstral Coefficients (CQCC) and spectrogram, proposing skip connection methods for VLD tasks. Wavelet-based features and gammatone frequency-based features have also been explored in [8], [9], which proposed new optimal accuracy for the task. In [10], authors employ a recently proposed optimal system by combining CQCC-based features with ResNet and LCNN as pattern classifiers. The key idea here is to represent excitation source information via LPR in various subbands in the cepstral-domain. For comparison, we also employ Mel frequency scale-based features, namely, Mel

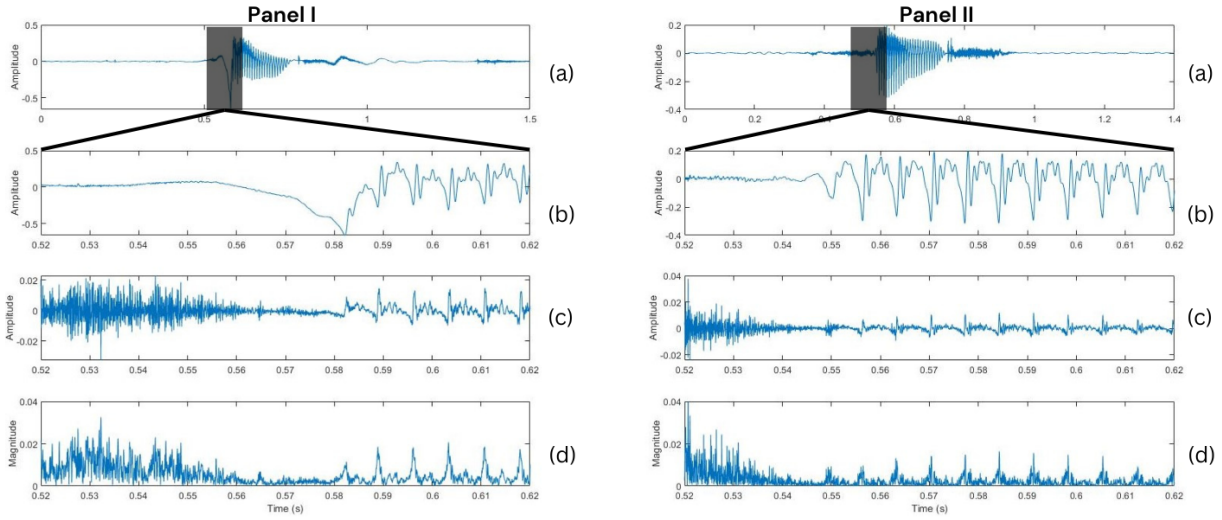


Fig. 1. Analysis of speech with pop noise (Panel I) and without pop noise (Panel II) : (a) time-domain representation of signal, (b) a selected range speech segment from (a), where pop noise is present, (c) LP residual of (b) (LP = 24), and (d) Hilbert envelope of LP residual shown in (c).

Frequency Cepstral Coefficients (MFCC) along with BiLSTM as a pattern classifier. We also compared the obtained results with large-scale pre-trained model to leverage this work with other works, as well as compared obtained results with existing works. This paper provides the following novelty :

- Significance of LFRCC feature set for VLD task.
- Experimental evaluation of proposed LFRCC feature set *w.r.t.* comparison with the existing MFCC and LFCC features and analysis of latency period.
- Score-level fusion of various cepstral features.

The rest of the paper is organized as follows: The computational details for LFRCC features are presented in Section II. Section III presents details of the experimental setup used, including the dataset, performance metrics, and classifiers. In Section IV, the results of the various experiments carried out using LFRCC are discussed. Finally, Section V reports the summary and conclusions along with potential future research directions.

II. Proposed Approach

In this study, we employ *pop noise* as a key acoustic feature for classifying if the speech is spoofed or genuine. Pop noise is a type of noise present in live human speech, when the airflow originating from the lungs excites the vocal tract system because it vibrates the vocal folds [9], [11]. As the glottal airflow passes through a cascade of digital 2nd order resonators (i.e., organ pipes), it reaches the mouth. Here, based on the pronunciation of the phoneme, the location and the force of the pop noise are determined, resulting in pop noise formation. The strength of pop noise depends on the distance between the speaker and the microphone. These sudden puffs arising from the glottal sound possess some excitation-based information, such as GCIs. These excitation-based features also have low energy, which decreases as distance increases. Analysing the knowledge related to VLD, the best possible way to perform an attack is by capturing the target speaker's voice samples from a far-distance and to use it for mounting attack. As the attacker is generally recording the speech from a distance, the energy of the speech reduces significantly due to the distance. According to the study mentioned in [1], pop noise is present in the lower frequency regions, ranging from $0-40$ Hz. Frequency is inversely proportional to time, therefore, as we increase time, lower frequency voice tends to fade out. The relation of time is with the distance, which can be further specified as the attacker records the audio from a far distance, due to which lower frequency points disappear. As a result, pop noise disappears on the attacker's recording, which results in the absence of GCI characteristics while feature extraction from residual speech.

A. Analysis of LP Residual

More used for analyzing and forecasting, the Linear Prediction (LP) technique examines and predicts the dynamics of a speech signal. The key concept of LP revolves around predicting the value of pth sample based on previous $p-1$ samples, where the selected number of previous $p-1$ samples is known as the order of linear prediction (O_{pred}). In the other words, LP is a technique to predict upcoming sample values based on a short number of previously received data and to reconstruct the signal at the receiver. If there exists a signal sig, then the pth sample of that signal is defined as $sig(p)$, based on the sum of previous samples, where c_k is the Linear Prediction Coefficient (LPC). In particular,

$$\hat{sig}(p) = -\sum_{k=1}^{p} c_k \cdot sig(p-k). \quad (1)$$

Thereby, the difference in the values between the original signal and the predicted signal can be mathematically called as LP error. However, this error in terms of signal processing

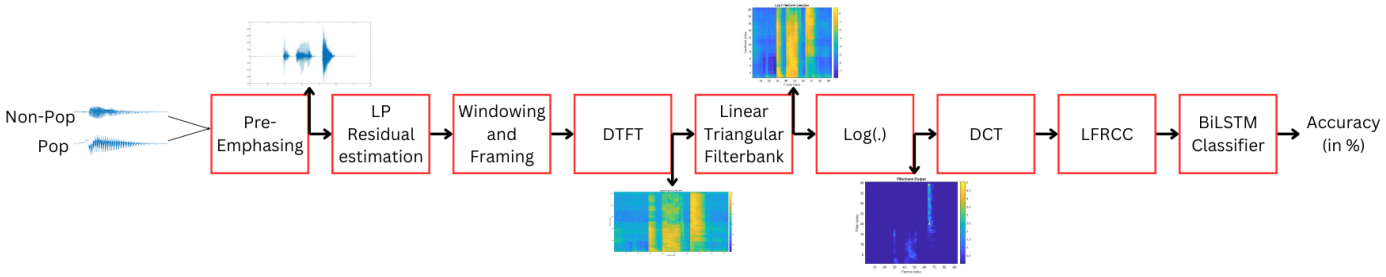


Fig. 2. Functional block diagram of performed LFRCC feature extraction followed by BiLSTM classifier.

is called residual, which can be estimated by,

$$res(p) = sig(p) - \hat{sig}(p) = sig(p) + \sum_{k=1}^{p} *c_k sig(p-k), \quad (2)$$

$$res_a(p) = res(p) + j * \hat{res}(p), \quad (3)$$

where $res(p)$ is LPR, $res_a(p)$ denotes analytic signal corresponding to $res(n)$, and $\hat{res}(p)$, can be given by $\hat{res}(p) = Hilbert\{res(p)\}$. Further Fourier transform of $\hat{res}(p)$, can be given as :

$$\hat{Res}(\omega) = F\{\hat{Res}(p)\} = \begin{cases} -jRes(\omega), & \omega > 0, \\ +jRes(\omega), & \omega < 0, \end{cases} \quad (4)$$

and Hilbert envelope of $res_a(p)$ is given by,

$$|res_a(p)| = \sqrt{res_a^2(p) + \hat{res_a}^2(p)}. \quad (5)$$

Moving on to spectrographic analysis, it can be observed in Fig. 1 that LP residual and Hilbert envelope show quasi-periodic pulses corresponding to GCIs. However, whenever there is pop noise, this quasi-periodic structure is broken into a noisy signal, including pop noise being a breath sound, and there is no glottal vibrations (including GCIs) during the production of the pop noise event along with natural speech production.

B. LFRCC

Further, we feed this LP residual signal as input to the LFCC feature vector to derive LFRCC, which is popularly known to capture GCIs and spectral envelopes, in which GCIs act as a clue to identify the presence of pop noise in audio signals. To obtain optimal predicted speech, we must obtain the optimal LP order. Inaccurate predictions, obtained from a non-optimal order of predictions may result in pop noise characteristics not getting captured. The optimal LP order in terms of existing linear mathematics, can be given as [12] :

$$\left(\frac{Fs}{1000}\right) + 2, \quad (6)$$

where Fs is the sampling frequency of audio samples. We obtained the optimal LP order as 24, mathematically as well as experimentally. We have used a frame duration of 25 *ms*, a frameshift of 13 *ms*, NFFT equal to 512 samples, and the dimension of feature vector equal to 20. Fig. 2 represents the flowchart of the proposed approach with the LFRCC feature vector and BiLSTM classifier.

III. Experimental Setup

A. Dataset Used

In a realistic scenario, attackers use spoofing attacks by acquiring audio samples of the selected person generally from a distance. They can record the speaker's voice from a distance to attack the VLD system. However, by analyzing the absence of pop noise in the spoofed samples, it is possible to differentiate genuine and spoofed speech. The researchers used the POCO dataset, which includes recordings of 66 subjects (34 females and 32 males) of different ages. The dataset contains English words, covering 44 phonemes, and is sampled at a sampling frequency (Fs) of 22050 Hz with *24*-bits depth [1]. The dataset is divided into three subsets: RC-A (recordings with a microphone, representing genuine speaker voice), RP-A (eavesdropping, simulating replay attacks without pop noise), and RC-B (recordings on microphone array). The RC-A subset was recorded in the presence of the live speaker using an Audio-Technica AT4040 microphone. RP-A subset simulates distant recorder audios for spoofing attacks. The dataset is further divided into training and evaluation subsets, ensuring an almost equal proportion of male and female speakers in each subset. More details on dataset employed are available on [11].

B. Other Spectral Features Used

1) MFCC: It is a feature extraction technique [13] commonly used in audio signal processing, particularly in speech and speaker recognition applications [14]. It transforms the short-term power spectrum of audio signals into a compact representation. Spectral information is captured by first windowing and then filtering the audio signal. Mel filterbank is employed to map frequencies to the Mel frequency scale. To decorrelate the resulting features, we take a logarithm of the filterbank energies, and then apply a Discrete Cosine Transform (DCT). DCT is taken into energy compaction, feature decorrelation, and dimensionality reduction of feature vector. Thus, we can retain the most significant coefficients.

2) LFCC: It is another approach widely used for feature extraction in tasks, such as speech recognition, anti-spoofing, and audio classification. Similar to MFCC, spectral information is captured by windowing and filtering the audio signal.

However, unlike MFCC which employs a Mel filterbank, LFCC employs a linear filterbank, distributing the subband filters evenly (i.e., linearly) across entire frequency range. After obtaining the filterbank energies, LFCC takes the logarithm of these energies. This step compresses the dynamic range and emphasizes lower energy components. Next, a DCT is applied to the subband energies and retains only the most significant coefficients. LFCC offers an alternative to MFCC, particularly in scenarios, where linearly-spaced frequency bins may be more appropriate or where a simpler feature extraction approach is desired.

C. Pattern Classifier Used

The BiLSTM classifier is a neural network architecture commonly used for sequence classification tasks. It employs bi-directional Long Short-Term Memory (BiLSTM) units, which process input patterns in both forward and backward directions [15]. By capturing content from both past and future tokens, BiLSTM enhances its ability to model long-range dependencies. A dropout layer is applied to prevent overfitting, and the final output is obtained by concatenating the hidden states from both directions and passing them through a fully-connected layer. The model is well-suited for tasks, such as sentiment analysis and named entity recognition. The learning rate was chosen to be 0.003, in combination with a batch size of 32. Each dropout layer is 25 % dropout, and the hidden size is 256. In this paper, the input size was chosen as a constant value of 20. We employ BiLSTM as a classifier, because it can capture both past and future dependencies in sequential data, enhancing context understanding. Its memory cells handle long-range dependencies, making it effective for tasks like speech, and audio classification. The bidirectional structure improves performance over standard LSTMs by considering the entire sequence.

D. Performance Metrics

Three key performance metrics, namely, Accuracy, F1-Score, and Equal Error Rate (EER), and 2 sub-performance metrics – Precision (P), and Recall (R), are employed to evaluate the efficiency of the VLD system. The performance of the system can be precisely measured using these metrics.

1) Accuracy : The ratio of correct predictions made by a model out of all the predictions made is known as the accuracy of the model.

2) Precision (P) : The ratio of true positive predictions to all positive predictions made by the model is known as the precision of the model [13].

3) Recall (R) : The ratio of true positive predictions to all actual positive instances in the dataset is known as the recall of the model [13].

4) F1-Score : It takes into consideration both precision and recall so that they have equal contributions to the final score. It is calculated as follows [13] :

$$F1 = (P * R)/(P + R). \tag{7}$$

5) Equal Error Rate (EER) : It represents the point, where the false acceptance rate (FAR) equals the false rejection rate (FRR) [16].

IV. Experimental Results

In this Section, the evaluation of proposed LFRCC features on various evaluation factors, such as fine-tuning of optimal LP order, comparison with the other cepstral features (such as, MFCC and LFCC), and analysis of latency period are discussed. Experiments were performed on the P100 GPU available on the Kaggle online platform.

A. LP Order Optimization

As discussed in sub-Section II-A, we need to optimize LP order (P) to make an optimal system using the LFRCC feature vector. This sub-Section presents accuracy obtained using different LP orders on the BiLSTM classifier. From Fig. 3, it can be observed that the optimal LP order is 24 for LFRCC features on the BiLSTM classifier. This indicates that exactly 24 previous samples are needed to predict the next sample for optimal results on the BiLSTM classifier, which may vary slightly. Further, as we decrease the number of samples used to predict samples, we get lower accuracy due to the lack of data used for prediction, however, we also get lower accuracy due to more than the required amount of data for prediction, resulting in misinformation captured from data. Optimal LP order helps us to estimate the number of samples we need to predict the next sample.

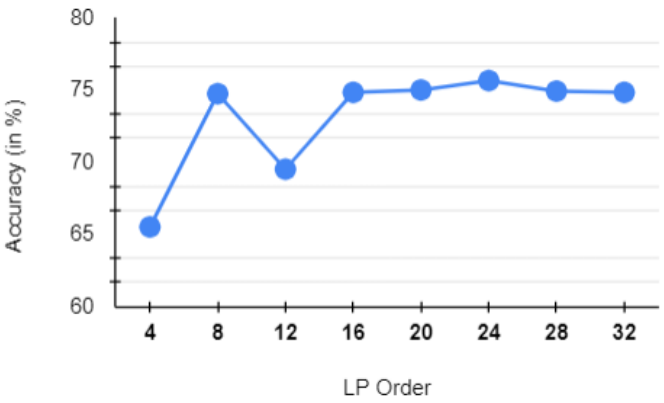


Fig. 3. Optimization of LP order on BiLSTM classifier.

B. Comparison with Other Cepstral Features

We compare our results with the other cepstral features. Results shown in Table I indicate that LFRCC outperforms MFCC and LFCC for all 5 performance metrics. Higher accuracy and lower EER on LFRCC feature vector than MFCC and LFCC indicate that LFRCC captures important information related to the excitation source, which LFCC and MFCC fail to capture. The high precision and low recall of all feature vector indicates low false negetive samples as compared to high number of false positive samples, thereby classifying speech more accurately. For a fair comparison, both the features, i.e., MFCC and LFCC have been employed earlier

TABLE I
COMPARISON WITH OTHER CEPSTRAL FEATURE VECTORS ON BILSTM CLASSIFIER

Feature	MFCC	LFCC	LFRCC (Order 24)
Accuracy	69.19	69.24	**75.67**
Precision	69.47	72.01	82.51
Recall	68.48	62.96	65.15
EER	30.81	35.67	**24.32**
F1-Score	69.19	69.12	75.39

and have been optimized in previous studies [4], [17], which we used with the same optimal comparison in this paper. Low EER on the LFRCC feature vector and high F1-score also indicate better performance of LFRCC over the other cepstral features. Further employing various different deep learning based classifier may result into decrease in EER of the system, thereby proposing optimal model in nearby future.

C. Analysis of Latency Period

In order to analyze the efficiency of the proposed LFRCC feature vector, we need to optimize the required number of frames and storage of data. Effectiveness of model lies not only on the accuracy of the model, but also on the effectiveness of model to train well on low amount of data and optimize the storage to create a model. To get an optimal model in less data, we explored the analysis of the latency period of all feature vectors as shown in Fig 4. It can be observed that when the amount of test data (in seconds) decreases, the accuracy of each feature vector decreases. Up to 0.5125 seconds, the accuracy remains constant, however, after that sudden drop in the accuracy of the LFRCC feature vector can be observed due to the insufficient amount of data available for training the model. On the other hand, it can be observed that MFCC and LFCC continue to perform poorly at low latency periods, which also helps us to prove the practical suitability of the proposed LFRCC feature vector over MFCC and LFCC even on low latency.

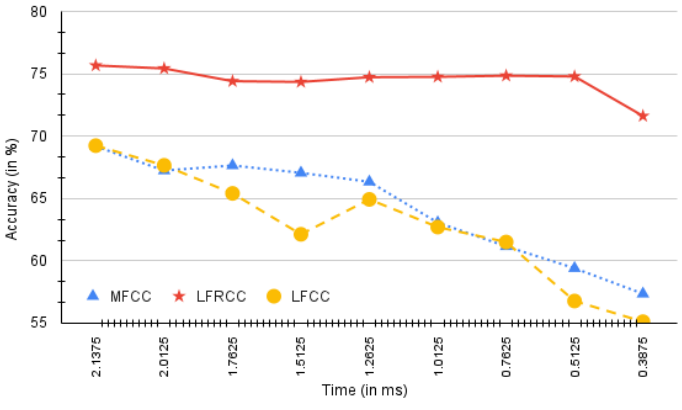


Fig. 4. Analysis of latency period on BiLSTM classifier.

D. Score-Level Fusion

We examine score-level fusion to increase the accuracy of LFRCC features in this sub-Section. For comparison, we fused all three features, i.e., LFCC+MFCC, LFCC+LFRCC, MFCC+LFRCC on both classifiers. As we obtained results

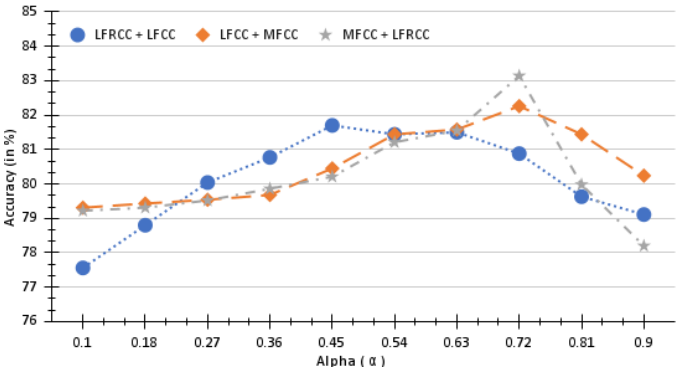


Fig. 5. Score-level fusion on BiLSTM classifier.

better on BiLSTM classifier and hence, we did all the experiments of this sub-Section on the BiLSTM classifier. Score-level fusion includes a parameter, namely, α. Score-level fusion can be calculated by :

$$(\alpha)Score_1 + (1-\alpha)Score_2 = Score_f, \tag{8}$$

where $score_f$ stands for the fused score. We obtained the highest of 83.14 % of accuracy on MFCC and LFRCC features, which is 1.33 % higher than individual optimal accuracy obtained on [10]. Fig. 5 shows score-level fusion on all three features.

E. Comparison with Existing Works

We obtained an optimal accuracy of 83.14 % (with MFCC + LFRCC + BiLSTM), which is 1.33 % less as compared to the proposed accuracy obtained by previous optimal accuracy [18]. Even after using simple feature changes (such as using residual signal) and feeding them to LFCC features, we can obtain a 20.85 % increase in accuracy as compared to baseline accuracy [1]. In [9], authors employ GFCC features with statistical classifier SVM, which resulted in 80.37 % accuracy, which is also less as compared to the proposed system. Table II represents the comparison of proposed study w.r.t. existing works, which include both types of classifiers (statistical and other acoustic models), and variety of features.

TABLE II
COMPARISON WITH EXISTING WORKS ON THE BASIS OF ACCURACY (IN %)

Source	Feature Set	Classifier	Accuracy (in %)
[1]	STFT	(Baseline)	62.29
[6]	CQCC	ResNet	78.45
[19]	MGDCC	GMM	63.67
	MGDCC	CNN	79.45
[7]	Spectrogram	ResNet	80.13
[9]	GFCC	SVM	80.37
[17]	MFCC	ResNet	81.81
Proposed	**MFCC + LFRCC**	**BiLSTM**	**83.14**

F. Comparison with SSL Models

In this sub-section, we compare our obtained results with pre-trained models, such as whisper and XLSR (all 4 versions), and wav2vec 2.0. These findings help us to prove superiority of model over existing pre-trained models, which have been

TABLE III
COMPARISON WITH PRE-TRAINED MODELS

Features	Accuracy (in %)	Feature	Accuracy (in %)
Whisper	73.86	XLSR-0.3B	76.71
XLSR53-large	68.71	XLSR-1B	80.03
Wav2vec2.0	73.94	XLSR-2B	80.51
[Proposed]	**83.14**	-	-

trained on large-scale data. Such findings indicate the ability of a model to process characteristics well and also help us to prove the superiority over the recently proposed pre-trained model. It can be observed from Table III that proposed features perform better than state-of-the-art deep learning pre-trained models.

V. SUMMARY AND CONCLUSIONS

In this study, we employed the LFRCC feature vector, which captures excitation source-based information via LP residual signal. We also optimized LP order to optimize the parameters of feature vector and pattern classifier. In addition, LFRCC proved better than LFCC and MFCC features. For better real-time, implementation of the model proposed, we also performed latency analysis, in which we gained optimal results for LFRCC feature vector than MFCC and LFCC. To compare our study to approaches, in sub-Section IV-E, we compared our results with existing optimal approaches for VLD. Future works include the variation of feature vector dimension, noise robustness, and feature-level fusion of the LFRCC feature vector with other feature sets.

VI. ACKNOWLEDGMENTS

This study has been funded by the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, under the project 'BHASHINI', (Grant ID: 11(1)2022-HCC.(TDIL)).

REFERENCES

[1] K. Akimoto, S. P. Liew, S. Mishima, R. Mizushima, and K. A. Lee, "POCO: A voice spoofing and liveness detection corpus based on pop noise," in *INTERSPEECH, Shanghai, China*, 2020, pp. 1081–1085.

[2] D. Nandi, D. Pati, and K. S. Rao, "Language identification using Hilbert envelope and phase information of linear prediction residual," in *International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Gurgaon, India, 2013, pp. 1–6.

[3] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, no. 10, pp. 782–796, 2008.

[4] M. Singh, "A survey on linear prediction residual based replay attack detection system," in *IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 2021, Indonesia, pp. 14–17.

[5] K. Khoria, A. T. Patil, and H. A. Patil, "Significance of Constant-Q Transform for voice liveness detection," in 29th *European Signal Processing Conference (EUSIPCO)*, 2021, Dublin, Ireland, pp. 126–130.

[6] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and Model Fusion for Automatic Spoofing Detection," in *INTERSPEECH*, 2017, Stockholm, Sweden, pp. 102–106.

[7] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *arXiv preprint arXiv:1907.00501*, 2019, {Last Accessed Date : $22^{nd} June, 2024$}.

[8] P. Gupta, S. Gupta, and H. A. Patil, "Voice liveness detection using bump wavelet with CNN," in 9th *International Conference on Pattern Recognition and Machine Intelligence (PReMI), Kolkata, India*, vol. 13102, 2021, LNCS, springer. Pp. 91–98.

[9] Q. Wang, X. Lin, M. Zhou, *et al.*, "Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *IEEE Conference on Computer Communications*, 2019, Paris, France, pp. 2062–2070.

[10] K. Khoria, A. T. Patil, and H. A. Patil, "On significance of constant-q transform for pop noise detection," *Computer Speech & Language*, vol. 77, pp. 101–421, 2023.

[11] P. Gupta and H. A. Patil, "Morse wavelet transform-based features for voice liveness detection," *Computer Speech & Language*, vol. 84, pp. 101–571, 2024.

[12] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America (JASA)*, vol. 50, no. 2B, pp. 637–655, 1971.

[13] J. Lee, T.-y. Kim, S. Beak, Y. Moon, and J. Jeong, "Real-time pose estimation based on ResNet-50 for rapid safety prevention and accident detection for field workers," *Electronics*, vol. 12, no. 16, p. 3513, 2023.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[15] M. Charola, S. Rathod, and H. A. Patil, "Robustness of whisper features for infant cry classification," in *International Conference on Speech and Computer*, SPECOM, LNCS, Springer, vol. 14339, 2023, pp. 421–433.

[16] A. Nolin-Lapalme, R. Avram, and H. Julie, "Privecg: Generating private ecg for end-to-end anonymization," in *Machine Learning for Healthcare Conference*, (PMLR), 2023, New Orleans, USA, pp. 509–528.

[17] P. Parasu, J. Epps, K. Sriskandaraja, and G. Suthokumar, "Investigating light-resnet architecture for spoofing detection under mismatched conditions," in *INTERSPEECH*, 2020, Shanghai, China, pp. 1111–1115.

[18] X. Sun, J. Fu, B. Wei, Z. Li, Y. Li, and N. Wang, "A self-attentional resnet-lightgbm model for iot-enabled voice liveness detection," *IEEE Internet of Things Journal*, vol. 10, no. 9, pp. 8257–8270, 2022.

[19] S. Singh, K. Khoria, and Hemant A. Patil, "Modified group delay cepstral coefficients for voice liveness detection," in *29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland*, 2021, pp. 146–150.

Adversarial Augmentation and Adaptation for Speech Recognition

Jen-Tzung Chien and Wei-Yu Sun
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

***Abstract*—It is crucial to conduct parameter efficient learning to adapt a large-scaled pre-trained backbone model to a downstream task where the desirable performance could be achieved for low-resource automatic speech recognition (ASR). However, the overfitting problem is prone to happen when the model adaptation is performed through fine-tuning individual parameters. The previous studies have explored different data augmentation methods to increase the size of training samples to enrich the data coverage and accordingly alleviate the overfitting issue in model training. In particular, this paper presents a new adversarial training for ASR based on a frozen pre-trained backbone model where the adversarial data augmentation is implemented so that a small amount of controllable parameters in adapters can be sufficiently estimated. In this study, the adversarial speech data are generated by adding the adversarial perturbations when training the adapters in an ASR model. The gradients of the intermediate adversarial examples are accumulated to calculate the augmented speech samples. The experiments on ASR using Common Voice and LibriSpeech datasets show the merit of the proposed adversarial augmentation and adaptation in terms of error rate and model size.**

I. Introduction

Automatic speech recognition (ASR) is built as an optimized computation model which is developed to automatically convert a human speech utterance into the corresponding word sequence. Different from speaker recognition, which attempts to identify the speaker who uttered the speech, ASR aims to recognize the lexical content from a speech signal. Traditionally, ASR was built with the acoustic model based on the hidden Markov model combined with Gaussian mixture model as well as the language model based on n-gram. Nowadays, the deep neural network model combining the acoustic model and language model in an end-to-end manner has been successfully developed to achieve a powerful performance in a sequence-to-sequence domain mapping from speech signal to text string where the massive trainable parameters are configured and the huge amount of training data is required. Importantly, the transformer [1] was exploited to build an end-to-end encoder-decoder framework which has shown dominant performance in a variety of tasks ranging from natural language processing (NLP) to computer vision (CV). Transformer was built according to the scheme of self attention and cross attention over individual and mutual sequences in source domain and target domain. More recently, different types of transformer have been designed to improve ASR by handling various issues in model size, memory capacity and representation efficiency [2], [3], [4]. However, a high-performance transformer rely on a large amount of neural network parameters as well as a huge collection of training utterances. A meaningful approach is to develop an ASR for a downstream task with a specific speaking style or under a different surrounding environment. Such an approach has been applied in many applications where a fine-tuning work was carried out by following a pre-trained backbone model using a limited amount of adaptation data [5].

In general, a large-scaled pre-trained transformer was estimated from a large amount of training data by using supervised, unsupervised and self-supervised learning methods. The data-driven transformer model could be sufficiently trained to cover various rules of syntax and grammar. Continual training is feasible to obtain desirable performance by using a pre-trained model. To customize the transformer to leverage knowledge transfer to fit a specific task using low-resource data in a target domain, it is straightforward to conduct fine-tuning for the encoder and decoder parameters in a pre-trained model. For the pre-trained model in ASR applications, the model is learned from audio signal and applied to calculate the audio features for a downstream task. The main-stream pre-trained ASR models such as wav2vec2.0 [6] and HuBERT [7] are seen as the variants of transformer which are similar in model architecture. In addition, the training procedure of ASR models is similar to that of pre-trained language models. The training objective aims to correctly predict or recover the masked input segments. Nevertheless, purely fine-tuning is easily overfitting in training process and likely dropping in ASR performance. It is meaningful to conduct data augmentation to mitigate the overfitting issue. A simple and general way to data augmentation for ASR could be implemented by applying the SpecAugment [8]. This method treated the Mel spectrum as an image sample. A kind of translational data enhancement was performed. Several consecutive rows and columns on the spectrum matrix were randomly crossed out. A resulting model was trained in a self-supervised way to enhance the capability of learning representation in time and frequency dimensions. There were no additional parameters and calculations. In addition, some other straightforward augmentation methods such as the pitch shift, time shift or noise addition to original audio speech data also worked well. This study deals with the overfitting issue in model construction for speech recognition based on a frozen large-scaled pre-trained backbone model where the small-scaled controllable adapter is learned from low-resource adaptation data [9]. A new adversarial augmentation and adaptation method is proposed to implement a parameter efficient learning approach to improve model robustness in presence of adversarial perturbations. The

training objective is developed to introduce the augmented data to ensure an extended and smoothed decision boundary to meet the local Lipschitz condition [10]. The experiments on various languages show the merit of the proposed method.

II. Adversarial Augmentation and Adaptation

Adversarial learning is known as a powerful machine learning paradigm which is feasible to build a robust model in presence of adversarial perturbation. This study presents an adversarial training algorithm to build a parameter efficient ASR model using adapter [11], [12].

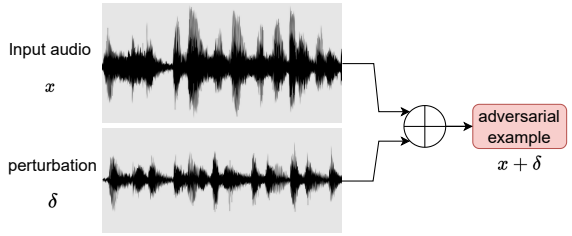


Fig. 1. Illustration of an adversarial example which is seen as a noisy speech signal with an additive perturbation noise.

A. Adversarial robustness

In general, deep learning is prone to build a model θ which is vulnerable to the adversarial example [13]. Such an example $x+\delta$ is basically seen as the addition of an imperceptible perturbation δ on the original example x which likely leads to a wrong prediction of output y. Figure 1 illustrates how an adversarial example of speech signal is observed. The adversarial example is considered as an augmented sample which is merged in training for the adapter parameter θ where a backbone ASR model ϕ is provided and frozen. The perturbation noise δ and the model parameter θ are jointly estimated from a dataset $\mathcal{D}=\{x,y\}$ by following

$$\min_{\theta}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{\delta_i}\left[\mathcal{L}\left(f_{\theta,\phi}\left(x+\delta_{i-1}\right),y\right)\right]\right. \\ \left.+\,\mathrm{KL}\left(f_{\theta,\phi}(x)\|f_{\theta,\phi}\left(x+\delta_N\right)\right)\right\}. \tag{1}$$

where δ_i denotes the perturbation at iteration i and totally N iterations are run. Correspondingly, the worst-case perturbation with the highest loss function $\mathcal{L}$ is estimated and then incorporated to train the optimal controlled model θ with the lowest loss in $\mathcal{L}$ where the classification outputs $f_{\theta,\phi}$ using original speech x and perturbed speech $x+\delta_N$ in the last iteration N are optimally near by with the smallest Kullback-Leibler (KL) divergence. Basically, this learning process is seen as a two-player game consisting of the players of finding the perturbation δ and then estimating the model θ. An adversarial robustness can be achieved accordingly [14], [15] by merging the adversarial examples $x+\delta$ in training for the controllable adapter θ. Basically, the worst-case perturbation may severely change the output of prediction. The proposed model is trained to improve the robustness of ASR by utilizing the perturbed speech data. Finding the perturbed speech corresponds to implement the generation step also called the adversarial attack. Alternatively, finding the estimated ASR adapter by minimizing the classification loss corresponds to carry out the discrimination step which is also called the adversarial defense. Adversarial training is seen as a process of fulfilling attack and defense for training a robust model.

B. Attack and defense

The key idea of adversarial training is to generate the adversarial examples that maximize the loss and simultaneously use these adversarial examples to train the model by minimizing the loss. Minimax optimization is comparable to fulfill an attack and defense process towards training a robust ASR model via a generation step and a discrimination step. The generation step for finding the perturbation is considered an attack while the discrimination step for training an ASR controllable model is acted as a defense against the attack.

In [16], the so-called fast gradient sign method (FGSM) was proposed as an approach to estimate the perturbation. Considering a supervised learning for a model f_θ from input sample x and its corresponding label y, FGSM is employed to solve Eq. (1) and estimate the perturbation by an iterative updating formula

$$\delta' \leftarrow \delta + \epsilon\cdot\mathrm{sign}(g^\delta) \tag{2}$$

where ϵ denotes the learning rate, δ' denotes the updated perturbation, and g^δ denotes the gradient

$$g^\delta = \nabla_\delta\mathcal{L}(f_{\theta,\phi}(x+\delta),y). \tag{3}$$

FGSM generates a perturbation δ for an addition with the input x. The estimation of perturbation is along the direction of the gradient towards the maximum value of loss function. In addition, an alternative approach to worst-case perturbation where the perturbation is scaled according to a normalized gradient without using the sign function, also known as a fast gradient method (FGM) [17] in a form of

$$\delta' \leftarrow \delta + \epsilon\cdot g^\delta/\|g^\delta\|_2. \tag{4}$$

FGM uses a one-step gradient ascent method to find an adversarial perturbation which may not be optimal. In [18], the projected gradient descent (PGD) was proposed to find an optimal perturbation through an iterative gradient ascent algorithm. In particular, if the perturbation is outside a volume as a norm ball with radius ϵ, i.e. $\|\delta\|_2\le\epsilon$, the perturbation is first projected back to the norm ball to ensure the perturbation is not too large. The solution to PGD is obtained and shown by

$$\delta' \leftarrow \mathrm{Proj}_{\|\delta\|_2\le\epsilon}(\delta+\alpha\cdot g^\delta/\|g^\delta\|_2) \tag{5}$$

where α is a step size. Using the perturbation δ_N in the last iteration N, the adversarial example is obtained as an attack by

$$x' = x+\nabla_\delta\mathcal{L}(f_{\theta,\phi}(x+\delta),y)\big|_{\delta\leftarrow\delta_N}. \tag{6}$$

On the other hand, the defense model is implemented by fulfilling a discrimination step towards estimating the controllable adapter θ for speech classifier in an ASR system. In the implementation, the gradient with respect to adapter θ consists of those for the first term g^θ and the second term in Eq. (1). The first gradient is calculated by initializing the gradient as $g^\theta \leftarrow 0$ and updating it iteratively in the inner loop as

$$(g^\theta)' \leftarrow g^\theta + \frac{1}{N}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\nabla_\theta \mathcal{L}(f_{\theta,\phi}(x+\delta), y)] \quad (7)$$

by N times to calculate the accumulated and normalized g_N^θ in the last iteration N. Then, the model parameter is iteratively updated in outer loop by

$$\theta' \leftarrow \theta - \beta \cdot (g_N^\theta + \nabla_\theta \mathrm{KL}(f_{\theta,\phi}(x) \| f_{\theta,\phi}(x+\delta_N))) \quad (8)$$

where β denotes the learning rate. The optimal adapter θ is trained by minimizing the cross-entropy loss for classification problem in an ASR system where KL divergence is also minimized to regularize the prediction consistency between adversarial sample and original sample.

III. Implementation Algorithm

This paper presents a new adversarial training algorithm to build a defense model for speech recognition where the generation step and discrimination step are implemented to increase the amount of the informative training samples and accordingly enhance the generalization and robustness of a learned model. This model is capable of improving the prediction performance for an ASR system consisting of a frozen pre-trained backbone model and a trainable adapter for a downstream task in a low-resource setting for domain adaptation [19]. The implementation algorithm is formulated and addressed.

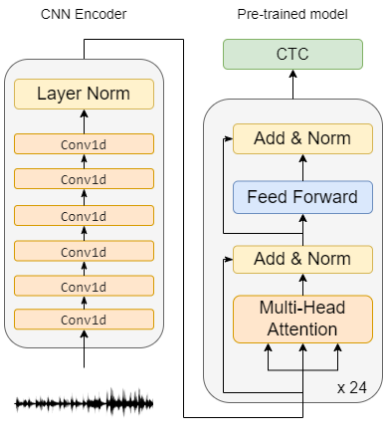


Fig. 2. Illustration of ASR pre-trained backbone model with parameter ϕ where speech signal x is transformed to find word string y. CNN encoder, transformer blocks and CTC classifier are stacked.

A. Model architecture

In this study, the pre-trained wav2vec2.0 [6] and HuBERT [7] were employed as the ASR backbone models. These two models had similar model structures which were trained in different strategies. Basically, wav2vec2.0 quantized the continuous audio data, and was learned to map them to discrete features and predict the masked features for training. The quantization module was used to discretize the output of the feature encoder. This model contained two sets of codebooks, and each codebook contained 320 variables. For each continuous output variable from the feature encoder, wav2vec2.0 identified a code in each set of codebooks, concatenated two codes together, and linearly mapped them to find the final quantization. On the other hand, HuBERT was trained by running two turns of the following procedure including using the k-means to find the clusters of Mel-frequency cepstral coefficient (MFCC) features and then predicting the masked features. After the first training turn was complete, the clustering model was reused to learn the detailed feature representation where the number of clusters in previous clustering model was increased from 100 to 500. The architecture of the pre-trained backbone model is shown in Figure 2. One-dimensional convolutional neural network (CNN) encoder with 7 layers dealt with the down-sampling of the input speech waveform x. The outputs of audio-encoded features were then fed forward to the pre-trained transformer encoder totally stacking 16 transformer blocks of multi-head attention and feedforward network. Then, the connectionist temporal classification (CTC) loss was calculated by using latent information from transformer encoder and then minimized to make prediction of word sequence y.

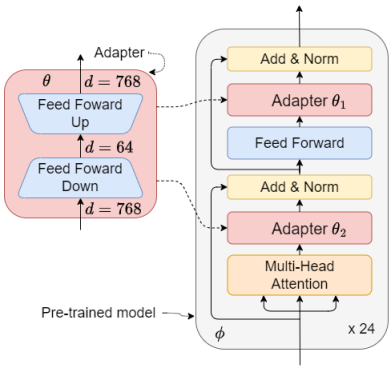


Fig. 3. Illustration of a stacked encoder where each frozen pre-trained transformer block ϕ is augmented with two controllable adapters $\{\theta_1, \theta_2\}$.

In model configuration, there were two adapters with parameter $\theta = \{\theta_1, \theta_2\}$ which were inserted in individual transformer blocks of a pre-trained backbone model ϕ. There were 24 blocks in an encoder. One adapter was connected with the multi-head attention and the other one was added to the feedforward layer as shown in Figure 3 where each adapter was shaped as a bottleneck structure with two feedforward layers with dimension from 768 to 64 and then back to 768. The parameters of the adapted transformer blocks were frozen except those parameters of adapters that were trainable. For each downstream task, only those adapter parameters θ were learned from relatively few training utterances. This adaptation strategy dramatically reduced the size of controllable parame-

ters and considerably increased the parameter efficiency for a downstream ASR task. The resulting model was compact for transfer learning where the numbers of training utterances and adjusted parameters were balanced.

B. Augmentation and adaptation

This paper presents an adversarial training method which is developed to generate the augmented examples in an adversarial attack step [20], [21]. The proposed adversarial algorithm eliminates the overhead cost of generating adversarial examples. A loopy optimization procedure with an inner loop and an outer loop is implemented as attack and defense, respectively. A robust ASR model with a pre-trained backbone model ϕ and a controllable adapter model θ is constructed. Finding the adversarial examples is implemented as the generation step for data augmentation in an inner loop via the gradient ascent algorithm while estimating the controllable adapter layers θ is performed as the discrimination step for classifier estimation in an outer loop via the gradient descent algorithm. In particular, the projection in PGD method is run by N times in inner loop to estimate the adversarial perturbation δ. The augmented data using the perturbation δ_N in the last iteration N is used to update the classifier or correspondingly the adapter θ in the outer loop as given in Eq. (6). A detailed implementation for this loopy procedure for adversarial augmentation and adaptation (AAA) is shown in Algorithm 1.

Algorithm 1 Adversarial augmentation & adaptation process

require: training samples $\mathcal{D} = \{x, y\}$, frozen pre-trained weights ϕ, adapter weights θ, perturbation δ, perturbation radius ϵ, adversarial step size α, learning rate β

  for $t = 1, \ldots, M$ **do** ▷ outer loop minimization

    initialize $g_0^\theta \leftarrow 0$, $g^\delta \leftarrow 0$, randomize δ_0

    for $i = 1, \ldots, N$ **do** ▷ inner loop maximization

      $g_i^\theta \leftarrow g_{i-1}^\theta + \frac{1}{N}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\nabla_\theta \mathcal{L}(f_{\theta,\phi}(x+\delta_{i-1}), y)]$

      $g^\delta \leftarrow \nabla_\delta \mathcal{L}(f_{\theta,\phi}(x+\delta_{i-1}), y)$

      $\delta_i \leftarrow \text{Proj}_{\|\delta\|_2 \le \epsilon}(\delta_{i-1} + \alpha \cdot g^\delta / \|g^\delta\|_2)$

    end for

    $\theta_{t+1} \leftarrow \theta_t - \beta \cdot (g_N^\theta + \nabla_\theta \text{KL}(f_{\theta,\phi}(x) \| f_{\theta,\phi}(x+\delta_N)))$

  end for

output: adapter weights θ_M

In this learning procedure, the gradient g^δ for perturbation δ is updated N iterations in the inner loop for maximization of cross-entropy loss function to cope with classification problem in ASR. The gradient g^θ for adapter θ is also updated N times to find g_N^θ and then combined with that for KL divergence between prediction outputs from original utterance x and perturbed utterance $x+\delta_N$. In the implementation, the gradient information g^δ is recycled to update the perturbation δ_i in every inner loop i, which means that the intermediate updatings of gradients for perturbation g^δ are reused. Therefore, we generate several augmented samples for data augmentation in one adversarial training in outer loop t. The loopy algorithm is performed to find adapter model θ_M which is obtained by running M iterations in outer loop t. The proposed method is inspired by the variants of the FREE algorithm in [22], [23] which were developed in computer vision and natural language processing. Considering this trick, this paper proposed a new AAA algorithm for speech recognition. The original sample x and the perturbed sample x' in the last iteration N as shown in Eq. (6) are both used when updating the adapter θ through

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f_{\theta,\phi}(x), y) + \mathcal{L}(f_{\theta,\phi}(x'), y)]. \tag{9}$$

The gradient with respect to θ is computed by using original sample x and adversarial sample x' with equal weighting.

C. Tradeoff between accuracy and robustness

In addition to FGM [17] and PGD [18], another adversarial training algorithm called the tradeoff-inspired adversarial defense via surrogate-loss minimization (TRADES) [24] was also implemented for comparison with the proposed AAA method. Using TRADES, the same setting as PGD in Eq. (6) was taken into account to generate the adversarial examples in accordance with maximizing the KL divergence between the logits of the prediction models by using original samples and adversarial samples in

$$\begin{aligned}\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[&\mathcal{L}(f_{\theta,\phi}(x), y) \\ &+ \lambda \cdot \max_{\|\delta\|_2 \le \epsilon} \text{KL}(f_{\theta,\phi}(x) \| f_{\theta,\phi}(x+\delta))\big].\end{aligned} \tag{10}$$

where $\lambda = 1$ is specified.

The attack and defense in adversarial training correspond to perform the generation and discrimination steps, which are adjustable to tradeoff between adversarial robustness and classification accuracy, respectively. The proposed AAA method carries out the adapter-based solution to adversarial robustness via Eq. (1) which comparably implements the tradeoff between two terms in Eq. (10) based on TRADES with a adjustable hyperparameter λ. The first term as a cross-entropy loss controls the accuracy while the second term as a KL divergence handles the robustness. Maximizing the KL divergence aims to generate the worst-case adversarial examples preserving the strongest protection to attack while minimizing the cross-entropy loss focuses on finding the best adapter to achieve the lowest word error rate (WER) for speech recognition.

Similar issue of balancing the tradeoff between accuracy and robustness was also mentioned in CV and NLP domains [10], [23], [25]. The second term of AAA in Eq. (1) is seen as a regularization condition for model generalization. This term encourages the prediction of adversarial example $x + \delta_N$ with the worst-case perturbation δ_N by using the learned adapter θ to get close to that of original example x. The resulting class boundary is likely fitted to the Lipschitz condition so that a smoothed boundary is determined to assure a better generalization [10]. This study presents a new approach to adversarial robustness in low-resource setting based on adversarial augmentation and adapter estimation. The KL term is optimized in an adversarial manner to get closer to the Lipschitz condition for a better model robustness or generalization.

IV. Experiments

This paper presents an adversarial augmentation and adaptation for speech recognition in two low-resource downstream tasks under two different pre-trained backbone models.

A. Experimental settings

As shown in Figure 2, a CTC classifier was stacked after an ASR pre-trained model based on 16 transformer blocks for word prediction where the audio embedding was calculated from a speech signal using CNN encoder. Word error rates (WERs) were measured to evaluate the ASR performance. Table I shows the settings of pre-trained backbone model based on wav2vec2.0 [6] and HuBERT [7]. The number of parameters in a backbone models is 315M while that of adapter is only 3.1M. AAA only adjusts the parameters of adapter instead of fine-tuning all backbone parameters as done in the previous methods. A parameter efficient learning is assured in using AAA method. Table II reports the settings of two speech datasets including Common Voice [26] and LibriSpeech (100 hours) [27] for Turkish and English, respectively. The number of samples or utterances is shown. In addition to the baseline result, this study also implemented the fine-tuning of the pre-trained model by using the other adversarial data augmentation methods where PGD [18], FGM [17] and TRADES [24] were included. Baseline method was implemented via a purely fine-tuning scheme. In the implementation, the perturbation δ was estimated and added to the audio signal as the augmented noisy speech signal. This strategy carried out an approach to noise-based data augmentation in a way of increasing the difficulty of correctly classifying the input audio signal into its corresponding features and word tokens based on the pre-trained model augmented with the controllable adapter.

TABLE I
SETTINGS OF THE PRE-TRAINED BACKBONE MODELS.

backbone	attention head	layer	hidden state	feature layer	dropout
wav2vec2.0	16	24	1024	7	0.1
HuBERT	16	24	1024	7	0.1

TABLE II
SETTINGS OF THE DATASETS FOR SPEECH RECOGNITION.

dataset	training set	testing set	language
Common Voice	3478	1647	Turkish
LibriSpeech	31242	2620	English

TABLE III
SETTINGS FOR THE ADVERSARIAL DATA AUGMENTATION.

method	ϵ	α	inner loop	norm type
baseline	N/A	N/A	N/A	N/A
PGD	1	0.3	3	ℓ_2
FGM	1	0.3	1	ℓ_2
TRADES	1	0.1	3	ℓ_2
AAA	1	0.1	3	ℓ_2

Furthermore, Table III lists the hyperparameter settings of using different methods for adversarial data augmentation including perturbation radius ϵ, adversarial step size α, number of inner-loop iterations N and the norm type which was fixed as ℓ_2 norm. In case of $N = 3$, the amount of augmented data were increased three times. In the experiments, the computation times in different settings were measured for two ASR tasks by using a single GPU, NVIDIA GeForce RTX 3090, with 24G RAM where the CPU was configured with an Intel®Core™i9-10900K CPU @ 3.70GHz. The batch sizes of speech utterances were 32 and 4 for Common Voice and LibriSpeech, respectively.

TABLE IV
WERS (%) AND TRAINING TIMES USING WAV2VEC2.0.

	CommonVoice		LibriSpeech	
	WER	training time	WER	training time
baseline	35.5	55min	4.25	14hr
PGD	35.2	2hr 6min	3.99	17hr 50min
FGM	34.5	1hr 22min	4.05	14hr 22min
TRADES	34.2	2hr 13min	4.18	18hr 20min
AAA	31.7	2hr 8min	3.94	18hr

TABLE V
WERS (%) AND TRAINING TIMES USING HUBERT.

	CommonVoice		LibriSpeech	
	WER	training time	WER	training time
baseline	51.3	59min	8.86	13hr 51min
PGD	48.4	2hr 8min	4.83	18hr 4min
FGM	51.8	1hr 19min	5.59	14hr 34min
TRADES	48.8	2hr 12min	5.02	18hr 25min
AAA	47.8	2hr 6min	4.77	17hr 57min

TABLE VI
AN EXAMPLE OF PREDICTED WORDS WHERE THE WORDS IN BLUE REPRESENT THE ERRORS. WAV2VEC2.0 WAS USED.

Label: he has grave doubts whether sir frederick leighton's work is really greek after all and can discover in it but little of rocky ithaca	
baseline	he has graved doubts whether sir frederick layton's work is ready greek after all and can discover in it but little of rocky ithica
AAA	he has grave doubts whether sir frederick laytons work is really greek after all and can discover in it but little of rocky ithaca

B. Experimental results

Tables IV and V compare the results of WER and training time by using wav2vec2.0 and HuBERT as the pre-trained backbone models, respectively. The number of outer-loop iterations M was varied and sufficient to assure convergence in using different methods. From the results, it is found that various adversarial data augmentation methods are beneficial in two ASR downstream tasks on the basis of pre-trained backbone models by using wav2vec2.0 and HuBERT. Wave2vec2.0 performs better than HuBERT in ASR using two different speech datasets. The adversarial augmentation and adaptation by using AAA consistently achieves lower

WERs than the adversarial data augmentation and fine-tuning by using PGD, FGM and TRADES by using different pre-trained backbone models under different speech datasets. In terms of computation time, the proposed AAA is more efficient than PGD and TRADES, but less efficient than FGM. FGM only ran one inner-loop iteration which is smaller than three inner-loop iterations in implementation of AAA. In addition, Table VI shows an example of comparing the predicted word sequences by using baseline and the proposed AAA where wav2vec2.0 was considered as the backbone model. The true label sequence is provided. It is obvious that AAA improves the ASR performance in this example.

V. Conclusions

This paper has presented an adversarial augmentation and adaptation algorithm to enhance the robustness of speech recognition in presence of perturbation noises. The overfitting issue in low-resource setting was handled through a frozen pre-trained backbone model combined with the learnable adapter. A parameter efficient learning algorithm was proposed to carry out an approach to ensure the adversarial robustness in speech recognition. The proposed method was illustrated as an attack and defense strategy through a procedure of data generation and model discrimination where the tradeoff between robustness and accuracy was adjustable for speech recognition, respectively. The relation of the proposed method to the other schemes to adversarial data augmentation was illustrated. A straightforward solution to an end-to-end data augmentation was exploited. In the experiments on a number of investigations over different languages and backbone models showed that the proposed method to adversarial robustness based on adapter performed better than the other conventional methods based on fine-tuning in terms of word error rates in speech recognition. The proposed method was more efficient in terms of parameter size than the other methods in most of settings. Future study on extending this method to the other kinds of downstream applications will be explored.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] J.-T. Chien and S.-E. Li, "Contrastive Disentangled Learning for Memory-Augmented Transformer," in *Proc. of Annual Conference of International Speech Communication Association*, 2023, pp. 2958–2962.

[3] J.-T. Chien and Y.-H. Huang, "Bayesian Transformer Using Disentangled Mask Attention," in *Proc. of Annual Conference of International Speech Communication Association*, 2022, pp. 1761–1765.

[4] C.-H. Leong, Y.-H. Huang, and J.-T. Chien, "Online Compressive Transformer for End-to-End Speech Recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2021, pp. 2082–2086.

[5] L.-J. Yang, C.-H. H. Yang, and J.-T. Chien, "Parameter-Efficient Learning for Text-to-Speech Accent Adaptation," in *Proc. of Annual Conference of International Speech Communication Association*, 2023, pp. 4354–4358.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[8] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[9] Y. Chen, X. Yang, H. Zhang, W. Zhang, D. Qu, and C. Chen, "Meta adversarial learning improves low-resource speech recognition," *Computer Speech & Language*, vol. 84, pp. 101576, 2024.

[10] Y.-Y. Yang, C. Rashtchian, H. Zhang, R. R. Salakhutdinov, and K. Chaudhuri, "A closer look at accuracy vs. robustness," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8588–8601, 2020.

[11] C.-Y. He and J.-T. Chien, "Learning adapters for code-switching speech recognition," in *Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2023, pp. 1–6.

[12] B. Aditya, M. Rohmatillah, L.-H. Tai, and J.-T. Chien, "Attention-guided adaptation for code-switching speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10256–10260.

[13] J.-T. Chien and Y.-A. Chen, "Towards a unified view of adversarial training: A contrastive perspective," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 5365–5369.

[14] C.-T. Chu, M. Rohmatillah, C.-H. Lee, and J.-T. Chien, "Augmentation strategy optimization for language understanding," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7952–7956.

[15] Z. Wang and J. H. L. Hansen, "Towards improving synthetic audio spoofing detection robustness via meta-learning and disentangled training with adversarial examples," *IEEE Access*, 2024.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[17] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," in *Proc. of International Conference on Learning Representations*, 2016.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. of International Conference on Learning Representations*, 2018.

[19] L. Li, M.-W. Mak, and J.-T. Chien, "Contrastive adversarial domain adaptation networks for speaker recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 2236–2245, 2022.

[20] H. Wang, Z. Jin, M. Geng, S. Hu, G. Li, T. Wang, H. Xu, and X. Liu, "Enhancing pre-trained ASR system fine-tuning for dysarthric speech recognition using adversarial data augmentation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 12311–12315.

[21] J.-T. Chien and W.-Y. Sun, "Adversarial augmentation for adapter learning," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–7.

[22] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[23] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu, "FreeLB: Enhanced adversarial training for natural language understanding," in *Proc. of International Conference on Learning Representations*, 2019.

[24] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. of International Conference on Machine Learning*, 2019, pp. 7472–7482.

[25] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2177–2190.

[26] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. of Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

Empathetic Response Generation via Regularized Q-Learning

Jen-Tzung Chien and Yi-Chien Wu
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

***Abstract*—This study presents a new offline reinforcement learning to improve the diversity of dialogue generation while the issue of limited training data is tackled. In particular, a regularized Q-function is introduced to generate the empathetic response where the distribution shift between the learned policy and the behavior policy is penalized and the resulting out-of-distribution response is mitigated. Importantly, an emotion aware reward function is designed to guide the agent to sufficiently learn various dialogue traits and emotions for empathetic response generation. In addition, a soft sampling method is implemented to draw the action to ensure the agent to reach those candidate responses which are marginal but appropriate. The flexibility of response generation is accordingly enhanced through the reward function by considering the informative dialogue features. The experiments on the emotional conversation generation show the merit of the proposed method.**

I. Introduction

Traditional response generation using rule-based or knowledge-based methods generally guarantee the high-quality responses but usually result in some errors when the user inputs are not matched with the pre-defined templates. With the rapid growth of deep learning from massive social media data, the conversational AI has been successfully developed to implement for deployment of various dialogue applications [1]. Large language models (LLMs) based on the generative pre-trained transformer (GPT) have been popularly developed to generate flexible and human-like expressions with diverse responses. However, purely considering the prediction of word sequence based on probability distribution using GPT may not really meet the human expectations. It is crucial to adjust the existing foundation model to build a large-scaled pre-trained backbone to meet a specific desired expectation. For example, how to attract users to keep interacting with the dialogue system is a critical demand whereas the empathy is seen as an important consideration as it significantly affects the user satisfaction.

To enhance the diversity of natural language generation in an empathetic conversation, the generation of sentences is treated as a sequential decision-making process where reinforcement learning (RL) can provide an useful solution [2]. The advantage of utilizing RL methods to build a dialogue agent [3], [4] rather than using supervised learning methods is due to the fact that RL methods more likely integrate the specific rewards which can be designed by customers to meet different tasks and thus better achieve true goals of the agents. Besides, by considering RL schemes, it is possible to incorporate long-term influences and dynamics from a generated response into an ongoing conversation. This paper presents an offline RL method that utilizes an emotion aware reward function to guide the agent to learn the desired features for response generation in a conversation. Particularly, the regularized Q-learning is developed to prevent the language model based on GPT2 from producing the out-of-distribution responses that may harm the overall conversation performance. The goal of this study is to enhance the empathy in dialogue generation and increase the user desire to interact with the system. Some analytical experiments on natural language generation are reported to illustrate the usefulness of this method.

II. Dialogue and Behavior Modeling

A. Learning for dialogue generation

Reinforcement learning (RL) is a powerful technique to simulate the action transitions [5], [6] which are implemented by designing or crafting the reward function where the interactive dialogue system considering the whole trajectory for selecting dialogue acts or generating conversational responses [7] can be flexibly constructed [8], [9]. In general, dialogue generation can be seen as a sequential decision-making process where the agent takes an action a_t based on the policy π_θ, transits to a new state s_{t+1}, and then receives a feedback r_t in a form of reward which is maximized along a trajectory with length T to improve its response performance. Hence, the learning objective is to maximize the expected cumulative future reward

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\gamma^t r_t\right] \tag{1}$$

where θ is the policy parameter and $\gamma \in (0, 1]$ is the discount factor. The goal of the whole procedure is to learn how to continuously choose an action that maximize $J(\theta)$.

Integrating RL techniques in a dialogue system involves learning the dialogues which are obtained incrementally through the tricks of self-play or human-in-the-loop. Typically, such a learning procedure is costly and time-consuming, unlike the supervised learning which trains a dialogue model from a data collection. The existing RL methods depend heavily on interacting with a learned model of a human [8], [10] to generate low-cost experience, but this may result in unintelligible language to generate incorrect response. There are twofold solutions to handle this issue. One is to impose the strong prior to ensure the generated response which is closely similar to that contained in the dataset [8], [11], and the other is to utilize the dialogue management or the template-based method to directly reuse the samples in the dataset [10], [12].

B. Learning with behavior regularization

Offline RL [13] is useful in a scenario where the frequent interactions with an environment are not allowed, such as the scenarios in healthcare tasks and self-driving cars. However, the training dataset $\mathcal{D}$ for offline RL follows the visitation distribution of the behavior policy π_β. This circumstance causes the demand to occasionally evaluate the state-action pairs that may not be well covered from the training data when computing the loss. In addition, the estimated Q-function tends to overestimate those out-of-distribution (OOD) behaviors [14], [15]. The shift in visitation distribution leads to a significant extrapolation error which will be propagated and accumulated across the decision-making chain process, and seriously harm the model performance for dialogue behavior. The solutions to address the distribution shift in offline RL mainly fall into two categories including policy regularization [14], [16], [17] and critic penalty [18], [19], [20].

1) Policy regularization: Policy regularization aims to guide the agent to learn an optimal policy π under specific constraints, such as minimizing the distance of the selected actions to those in the training dataset and keeping a state-action visitation similar to that appearing in the dataset so as to alleviate the extrapolation error in offline RL. One classic method, called the batch-constrained deep Q-learning (BCQ) [14], employed the conditional variational auto-encoder (VAE) $G_\omega(\cdot)$ to generate the plausible candidate actions with perturbations ξ_ϕ to increase the diversity of generation. The Q-function $Q(\cdot)$ measures the scores of n candidate actions a_i sampled from the generative model $\{a_i \sim G_\omega(s)\}_{i=1}^n$ and the one with the highest score is selected as the final output. This scheme effectively regularizes the learned policy $\pi_\theta(s)$ given by state s via

$$\pi_\theta(s) = \underset{a_i+\xi_\phi(s,a_i,\Phi)}{\operatorname{argmax}} Q\left(s, a_i + \xi_\phi\left(s, a_i, \Phi\right)\right) \tag{2}$$

to be as close as the behavioral policy π_β. For the application in natural language generation, the pre-trained language models such as those GPT variants [21], [22], [23] can provide a language prior $\mu(\cdot|s)$ that prevents RL model π from selecting inappropriate words. This language prior is useful in offline RL with limited data.

2) Critic penalty: Considering the conservative Q-learning (CQL) [18], the critic penalty is merged in the Q-function where the conservative Q-value is estimated to act as the lower bound of Q-value by adding the regularization term. The Q-network as a critic is updated to calculate the Q-value according to the minimax objective

$$\begin{aligned}\min_Q \max_\pi \ & \alpha \cdot \left(\mathbb{E}_{s\sim\mathcal{D},a\sim\pi}[Q(s,a)] - \mathbb{E}_{s\sim\mathcal{D},a\sim\widehat{\pi}_\beta}[Q(s,a)]\right) \\ & + \frac{1}{2}\mathbb{E}_{s,a,s'\sim\mathcal{D}}\left[\left(Q(s,a) - \widehat{\mathcal{B}}^{\pi_k}\widehat{Q}^k(s,a)\right)^2\right] + \mathcal{R}(\mu)\end{aligned} \tag{3}$$

where α is the learning rate, $\widehat{\pi}_\beta(a|s)$ is the empirical behavior policy, $\widehat{\mathcal{B}}^{\pi_k}$ is the empirical Bellman operator, and $\widehat{Q}^k$ is the updated Q-value in iteration k. In Eq. (3), the second term is seen as the temporal difference error [24] and the others are treated as the regularization terms. The term $\mathbb{E}_{s\sim\mathcal{D},a\sim\pi(a|s)}[Q(s,a)]$ is used as a penalty to minimize the Q-value for the action a chosen by the currently learned policy π at the state s contained in the dataset, and the term $-\mathbb{E}_{s\sim\mathcal{D},a\sim\widehat{\pi}_\beta(a|s)}[Q(s,a)]$ is utilized to encourage the Q-network to provide relatively high scores for those actions a that exist in the dataset with the state s. Such a regularization, known as the CQL regularizer, is added during the training process, which is helpful to reduce the chance of obtaining the overestimated Q-value. As for the final term $\mathcal{R}(\mu)$, it can be chosen as a particular regularizer according to the need of a task. For instance, the Kullback-Leiblier (KL) divergence is viewed as a popular choice of measuring the closeness between the current distribution $\mu = \pi$ and its prior distribution ρ, namely

$$\mathcal{R}(\mu) = -D_{\mathrm{KL}}(\mu\|\rho). \tag{4}$$

This regularization term is minimized to find a dialogue policy π which is as close as ρ.

III. Empathetic Response Generation

This study presents a new approach to empathetic response generation [25] where the Chinese sentiment analysis (cnsenti) library in [26], [27] was adopted to perform the sentiment analysis and the emotion detection from the user's text input and directly concatenate the detected emotion class to the input. The cnsenti library was also used to label the emotions for the replies in the training corpus to address the lack of high-quality corpus with Chinese training data for various emotion classes. This corpus was used in the experiments on RL dialogues by fine-tuning LLM based on GPT-2.

A. Emotion aware reward

A challenging issue for a successful dialogue system is to learn how to draw people to continue interacting with the system. This study takes into account the following six empathetic features $\{r_i\}_{i=1}^6$ [8], [11] and builds an emotion aware reward to learn an agent to generate the empathetic responses with diversity and safety.

- *length of the agent response:* r_1 is to encourage the agent to generate longer responses while avoiding very short responses that end the conversation easily.
- *consistency of user sentiment with agent emotion:* r_2 is to encourage the agent to provide the empathetic replies by considering the detected emotion of the user input and incorporating it into the generated response.
- *asking questions:* r_3 is to encourage asking the follow-up questions related to the inputs to show that the agents have actively listened to the user's concerns and are engaged in the conversation.
- *context consistency:* r_4 is to enhance the coherence of the generated reply and avoid the situation where the generated reply receives a high return but its content is ungrammatical or semantically incoherent by calculating

$$r_4 = \frac{1}{N_a}\log p_{\mathrm{s2s}}^{\mathrm{fw}}(a|s) + \frac{1}{N_s}\log p_{\mathrm{s2s}}^{\mathrm{bw}}(s|a) \tag{5}$$

where $p^{\text{fw}}_{\text{s2s}}(a|s)$ and $p^{\text{bw}}_{\text{s2s}}(s|a)$ denote the forward (fw) and backward (bw) probabilities for generating a response a given user input s and producing the previous sentence s given a response a, and N_a and N_s denote the lengths of a and s in a sequence-to-sequence (s2s) generation, respectively. Bidirectional context coherence is measured.

– *detection of offensive words:* r_5 is to measure the penalty when the generated response contains some inappropriate words which are predefined in a given list and received by a negative reward.

– *suppression of non-sense reply production:* r_6 is to measure how the sentences generated by the agent have the sensitivity or potential to continue the conversation by computing the negative log likelihood of a non-sense sentence s via

$$r_6 = -\frac{1}{N_{\mathbb{S}}} \sum_{s \in \mathbb{S}} \frac{1}{N_s} \log p_{\text{s2s}}(s|a). \tag{6}$$

In Eq. (6), $N_{\mathbb{S}}$ is the cardinality of a set of generated sentences $\mathbb{S}$, and N_s is the number of tokens in a non-sense response s.

Finally, the emotion aware reward function r_t of a response a at time t is a weighted sum of the rewards over empathetic features $r = \sum_{i=1}^{6} \lambda_i r_i$ by using the weights $\{\lambda_i\}_{i=1}^{6}$.

B. Regularized Q-function

This study presents the behavior regularization in dialogue generation by controlling the extrapolation errors due to the distribution shift in offline RL. A regularized Q-function is implemented to constrain the dialogue agent from choosing OOD actions. At the same time, the confident actions are up-weighted to allow the agent to focus on the successful emotion aware samples.

1) Extrapolation error control: In the implementation, the pre-trained language model is fine-tuned by using a specific dataset to obtain the probability $p(a|s)$ of generating the response a given a dialogue input s, which acts as a prior probability to avoid selecting those OOD actions [14]. The KL divergence between current policy q and prior policy p is added as the regularization term in the objective for Q-learning. The learning objective is formulated as

$$\mathbb{E}_{q(\tau)}[r(\tau)] - D_{\text{KL}}[q(\tau) \| p(\tau)] \tag{7}$$

where τ is a trajectory of dialogue actions, and the distributions of learned policy and prior policy are calculated by $q(\tau) = \prod_{t=1}^{T} \pi_\theta(a_t|s_t)$ and $p(\tau) = \prod_{t=1}^{T} p(a_t|s_t)$, respectively, and

$$D_{\text{KL}}[q\|p] = \sum_{\tau} q(\tau)(\log q(\tau) - \log p(\tau)). \tag{8}$$

Such a regularized Q-function guides the agent to prefer choosing the actions in the dataset and prevent selecting the OOD actions. To fulfill Eq. (7), the regularized Q-function (denoted as the RQ score) is yielded by

$$\begin{aligned} Q\left(s_t, a_t\right) = \mathbb{E}_\pi \Bigg[& \sum_{t'=t}^{T} r\left(s_{t'}, a_{t'}\right) + \log p\left(a_{t'}|s_{t'}\right) \\ & - \log \pi_\theta\left(a_{t'}|s_{t'}\right)\Bigg] \end{aligned} \tag{9}$$

which is maximized in RL to select the generated responses. In addition to reward maximization, the term $\mathbb{E}_\pi[\log p\left(a_{t'}|s_{t'}\right)]$ is maximized to encourage the agent to select the actions under the prior and prevent the unrealistic OOD actions via KL regularization, and the term $\mathbb{E}_\pi[-\pi_\theta\left(a_{t'}|s_{t'}\right)]$ is seen as an entropy function which is maximized to maintain the diversity of dialogue agent.

2) Importance weight approximation: Importance weight approximation is employed to up-weight the actions with high confidence which help the agent focus on successful dialogue trajectories. In the offline RL setting, we could not directly sample the trajectories from the learned policy π_θ to calculate the total expected reward. Accordingly, the behavior policy π_β is employed to sample the trajectories $\tau = \{s_t, a_t\} \sim \pi_\beta$ for learning from the fixed dataset in an offline manner. Importance sampling is used to address the difference between the sample policy π_β and the updated policy π_θ. The unbiased estimation of policy gradient given by the importance weight

$$w_t = \prod_{t'=1}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\pi_\beta(a_{t'}|s_{t'})} \tag{10}$$

can be obtained by

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\beta}\left[\sum_{t=1}^{T} w_t \nabla_\theta \log \pi_\theta(a_t|s_t) Q(s_t, a_t)\right] \tag{11}$$

which represents the gradient with the total expected reward regularized from the state s_t.

However, this process may be sensitive during optimization. The convergence time will be increased when the importance weight is multiplied over each action for numerous time steps [28]. To tackle this issue, the importance weight is approximated by $w_t \approx \frac{\pi_\theta(a_t|s_t)}{\pi_\beta(a_t|s_t)}$. This is equivalent to maximize the total expected reward under the on-policy action distribution of π_θ and the off-policy state distribution produced by π_β. Despite of being biased, this estimator has been demonstrated empirically to reduce the variance and perform well when π_β and π_θ are close. As for how to obtain the value of π_β, an intuitive way is to estimate π_β directly from the dataset. Here, a simple approximation $\pi_\beta(\tau) \approx 1/N$ is made when τ is included in the dataset, otherwise $\pi_\beta(\tau) = 0$. With this approximation, the denominator, i.e. π_β, can be neglected in the optimization since it is a constant. Then, the approximation to policy gradient $\nabla_\theta J(\theta)$ in Eq. (11) is further derived by

$$\begin{aligned} & \mathbb{E}_{\tau \sim \pi_\beta}\left[\sum_{t} \prod_{t'=1}^{t} \frac{\pi_\theta(a_{t'}|s_{t'})}{\pi_\beta(a_{t'}|s_{t'})} \nabla_\theta \log \pi_\theta(a_t|s_t) Q(s_t, a_t)\right] \\ & \approx \mathbb{E}_{\tau \sim \pi_\beta}\left[\sum_{t} \frac{\pi_\theta(a_t|s_t)}{\pi_\beta(a_t|s_t)} \nabla_\theta \log \pi_\theta(a_t|s_t) Q(s_t, a_t)\right] \\ & \approx \sum_{i=1}^{N} \sum_{t=1}^{T} \pi_\theta(a_t^i|s_t^i) \nabla_\theta \log \pi_\theta(a_t^i|s_t^i) Q(s_t^i, a_t^i) \end{aligned} \tag{12}$$

where i denotes the trajectory index, and $\pi_\theta(a_t^i|s_t^i)$ is the model confidence to up-weight the actions preferred by current policy.

C. Implementation procedure

The overall workflow of response generation and candidate selection is addressed. The whole process is divided into three stages. In the first stage, the cnsenti library [26] was employed as an emotion detector where the emotion class of an user input was estimated and concatenated with the input as an augmented user input to feed in language model based on GPT-2 [22]. This GPT-2 language model was fine-tuned by using the emotional conversation generation (ECG) [29] dataset which contained 1M samples of user inputs and their responses. The adapted GPT-2 was accordingly used to generate several candidate responses as the possible empathetic responses for subsequent selection. In the second stage, Q-function acted as a critic to provide the evaluation of RQ scores over individual candidate responses based on the emotion aware reward function. This implementation considered the double Q-network [30] where one Q-network was responsible for estimating the Q-value of each candidate response while the other Q-network was used as a target network to calculate the target Q-value. The updating process for double Q-network was based on the minimization of mean square error loss between the estimated Q-value and target Q-value. In particular, each Q-network was implemented by using BERT encoder [31]. In the third stage, the best candidate was selected according to the RQ scores. Importantly, the diversity of the selected response was assured since those suboptimal responses also had a chance to be selected during the RL procedure. Here, the selection was based on the distribution over Q-value

$$\pi_\theta(a|s) \propto \exp Q(s,a). \tag{13}$$

This implementation was carried out for evaluation the proposed method in the following experiments.

IV. Experiments

This study conducted the evaluation on empathetic response generation based on the GPT-2 (Chinese) which was fine-tuned by using ECG dataset in the first stage and then fine-tuned by using the proposed RL fine-tuning in the second stage.

A. Experimental settings

In the implementation, the dataset ECG [29] was used for GPT-2 fine-tuning in the first stage while the other dataset called the large-scale cleaned Chinese conversation (LCCC) [33] was used for RL-based fine-tuning where LCCC was rigorously filtered to eliminate various kinds of noises. There are 6.8M single-turn and multi-turn conversations in LCCC-base [33] dataset where their emotions were obtained through cnsenti [26]. The emotion classes include 喜歡 (like), 悲傷 (sad), 噁心 (disgusted), 憤怒 (angry), 開心 (happy), and 其它 (other). The simplified Chinese was converted to the traditional Chinese for all text data in the experiments. In the learning procedure, AdamW optimizer was used. The initial learning rate was 1.5e-4 and 5e-5, the batch size was 4 and 8, the number of parameters was 102M and 311M, and the training time was 4 and 60 hours in first and second stages, respectively, where GPU with GeForce RTX 4090 24GB was used and 90K text corpus was collected in RL training. For comparative study, the perplexity (PPL) was used to measure the performance for generation and prediction in the evaluation. BLEU-2 and BLEU-4 [34] were evaluated for Chinese response generation where the complete meanings of Chinese vocabulary with two and four words were evaluated, respectively. Inter-distinct unigram (Dist-1) and bigram (Dist-2) [35] were used to measure the diversity of the generated sentences. These metrics were used to conduct the automatic evaluation for diversity and emotion factor in the generated

TABLE I
RESPONSE GENERATIONS IN THE FIRST CASE STUDY BY USING DIFFERENT METHODS.

Type	Context	RQ Score
Input	明天要出去玩，興奮地睡不著覺 (Go out to play tomorrow, and too excited to sleep.)	
Pre-trained	哈哈，我也覺得興奮 (Ha! Ha! I also feel excited.)	–
CDial-GPT	哈哈，我也是 (Ha! Ha! Me too.)	–
RQL Cand 1	我也是興奮的睡不著 (I am also too excited to sleep.)	2.51
Cand 2	明天要是沒有下雨我也興奮的睡不著，明天要出去玩，晚安 (If there is no rain tomorrow, I am excited and sleepless. Go out to play tomorrow. Good night.)	1.44
Cand 3	你明天不是要上班麼 (You aren't going to work tomorrow, aren't you?)	1.58
Cand 4	興奮地睡不著，因為明天又是週末了 (So excited and could not sleep because weekend is coming again tomorrow.)	1.76
Cand 5	是啊，我今天也要去玩了，興奮得不行 (Yes! I also go playing today. I am super excited.)	0.54
Final reply	興奮地睡不著，因為明天又是週末了 (So excited and could not sleep because weekend is coming again tomorrow.)	1.76

TABLE II
RESPONSE GENERATIONS IN THE SECOND CASE STUDY BY USING DIFFERENT METHODS.

Type	Context	RQ Score
Input	我想念以前住在鄉下的日子，雖然不方便但至少一家人都在一起 (I miss the time when living in the suburb. Although it was inconvenient, at least all family members could live together.)	
Pre-trained	唉，這日子沒法過了！ (Oh! It was hard to live!)	-
CDial-GPT	想念我的床想念我的床 (Miss my bed, Miss my bed.)	-
RQL Cand 1	我也是，想念以前的日子 (Me too. Miss the old time.)	2.46
Cand 2	對啊，我們都要好 (Yes! We were all fine.)	1.14
Cand 3	不過我覺得我們這裡也是這樣，一個人在外面一定要小心啊，我現在一個人在外地也很害怕 (I feel that we here are the same. Be careful when living outside. I am now frightened when living alone and outside.)	1.98
Cand 4	是啊，我也想念以前一起玩的時光啊 (Yes! I also miss the old time when playing together.)	0.99
Cand 5	這是我第一次離開家，真是太想念了 (This is the first time that I left from home. Really miss a lot!)	1.08
Final reply	是啊，我也想念以前一起玩的時光啊 (Yes! I also miss the old time when playing together!)	0.99

TABLE III
EVALUATION OF DIFFERENT METHODS UNDER VARIOUS METRICS.

Model	Size	PPL (↓)	BLEU-2 (↑)	BLEU-4 (↑)	Dist-1 (↑)	Dist-2 (↑)
Transformer [32]	113M	22.10	6.72	3.14	0.0088	0.1397
GPT2-chitchat [33]	88M	–	2.28	0.54	0.0103	0.1625
Pre-trained GPT2 [33]	88M	21.27	5.96	2.71	0.0080	0.1172
CDial-GPT [33]	104M	22.76	5.69	2.50	0.0077	0.1087
RQL w/o IWA	108M	20.08	6.98	3.21	0.0100	0.1623
RQL w/o EEC	108M	19.31	7.15	**3.29**	0.0098	0.1585
RQL w IWA & EEC	108M	**18.01**	**7.20**	3.20	**0.0106**	**0.1695**

responses. The size of parameters was given. An empirical selection of λ_i was done. The multinomial sampling was used to select the final reply.

B. Experimental results

First of all, Table I shows the first case study on the comparison of evaluating the response generations by using the pre-trained model via GPT-2, the baseline model via CDial-GPT [33] and the proposed model via regularized Q-learning (hereafter is denoted by RQL). CDial-GPT is a Chinese GPT-2 which was purely trained on the LCCC dataset while the pre-trained model was fine-tuned by using the ECG dataset. In general, the pre-trained and baseline models generate the appropriate responses. But, both of them relatively lack richness and precision in the response generation when compared with that generated by the proposed RQL. Five candidate responses (ranked by GPT-2 in the first stage) and their RQ scores (calculated by the regularized Q-function in Eq. (9) in the second stage) are shown. The proposed RQL not only pays attention to the emotional word 興奮 (excited) in the input sentence but also considers the other information mentioned in the response such as 睡不著 (could not sleep) and 出去玩 (go out to play), making the response more diverse and informative. Additionally, the proposed RQL extends the information which is not mentioned in the user input like 上班 (go to work) and 週末 (weekend), making the response even more diverse. Notably, the final output is selected by the multinomial sampling which can increase the diversity of the generated response. It is noted that, next to the first candidate (Cand 1), the fourth candidate (Cand 4) receives higher regularized Q-value than those in the second, third and fifth candidates since the fourth response is more empathetic and accurate than the others. This fourth candidate captures the correct tokens 週末 (weekend) which are reached in the other candidates. It is noticed that some emotions are more explicit in guiding the model's response while the others are more implicit and related to the specific emotions and dialogue in the training data. In the second case study as shown in Table II, the related work CDial-GPT model focuses on the single term 想念 (miss), but generates less coherent response, while the proposed model provides more informative candidate replies. However, some responses may not be so appropriate, as seen in the third candidate. Here, the use of multinomial sampling in selecting the final output helps the model choose a response that may not have the top RQ score but is more suitable for the situation.

Table III shows the results of automatic evaluation by using the proposed RQL and the related methods which include the transformer decoder [32], the GPT2-chitchat [33], the pre-trained language model (Chinese variant of GPT-2) [33], and the CDial-GPT [33]. The best result in the comparison is shown by bold. Relative to the other methods, the proposed method conducts a two-stage implementation where the supervised fine-tuning and the regularized Q-learning are performed for domain adaptation and empathetic generation where the emotional domain and the diverse reply are concerned, respectively. There is no RL learning in previous works. For ablation study, the proposed RQL is also implemented by ignoring the modifications based on importance weight approximation (RQL w/o IWA) and extrapolation error control (RQL w/o EEC). There are several findings in RQL. First, the increase of model size is moderate. Notably, BLEU measures the matching between the generated result and the reference answer, and may not reflect the capability of generating diverse responses. Although the BLEU values are generally low for various methods, the proposed RQL is still better than the other methods. In addition, RQL with IWA & EEC obtains the highest Dist-1 and Dist-2 in the comparison. It is obvious that RQL with the schemes of importance weight approximation and extrapolation error control achieves the best results among different methods in most of evaluation metrics.

V. CONCLUSIONS

We have presented a regularized Q-learning approach with the custom-designed emotion aware reward function to train a dialogue agent to generate empathetic responses with specific emotions. The Kullback-Leibler control term, included in the Q-function, helped to penalize the agent for generating the OOD responses, reducing the extrapolation errors caused by the distribution shifts. Using the designed reward functions facilitated the agent to learn the desired features and improved the quality of dialogue responses. This study employed the soft sampling to enhance the diversity of the final output reply. The experiments on Chinese empathetic generation showed the merit of the proposed method. However, this approach may not be enough to handle multi-turn dialogue or integrate the previous conversation history in a long dialogue. Therefore, it will be a future direction to introduce the additional memory to handle the pieces of information that were mentioned in the previous turns. To enhance the empathy in the generated responses, we will also need more precise and appropriate reward functions to guide the empathetic agent to learn better.

REFERENCES

[1] Mahdin Rohmatillah and Jen-Tzung Chien, "Revise the NLU: A prompting strategy for robust dialogue system," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10956–10960.

[2] Jen-Tzung Chien, Wei-Lin Liao, and Issam El Naqa, "Exploring state transition uncertainty in variational reinforcement learning," in *Proc. of 2020 28th European Signal Processing Conference*, 2021, pp. 1527–1531.

[3] Mahdin Rohmatillah and Jen-Tzung Chien, "Hierarchical reinforcement learning with guidance for multi-domain dialogue policy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 748–761, 2023.

[4] M. Rohmatillah, B. G. Ngo, W. Sulaiman, P.-C. Chen, and J.-T. Chien, "Reliable dialogue system for facilitating student-counselor communication," in *Proc. of Annual Conference of International Speech Communication Association*, 2024, pp. 1003–1004.

[5] Dhawal Gupta, Yinlam Chow, Azamat Tulepbergenov, Mohammad Ghavamzadeh, and Craig Boutilier, "Offline reinforcement learning for mixture-of-expert dialogue management," *Advances in Neural Information Processing Systems*, vol. 36, pp. 5912–5935, 2024.

[6] Je-Wei Jang, Mahdin Rohmatillah, and Jen-Tzung Chien, "AVAST: Attentive variational state tracker in a reinforced navigator," in *Proc. of Conference of International Joint Conference on Natural Language Processing*, 2022, pp. 424–433.

[7] Tien-Ching Luo and Jen-Tzung Chien, "Variational dialogue generation with normalizing flows," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7778–7782.

[8] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao, "Deep reinforcement learning for dialogue generation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1192–1202.

[9] Verena Rieser and Oliver Lemon, *Reinforcement Learning for Adaptive Dialogue Systems: A Data-Driven Methodology for Dialogue Management and Natural Language Generation*, Springer Science & Business Media, 2011.

[10] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang, "Decoupling strategy and generation in negotiation dialogues," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2333–2343.

[11] Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard, "Human-centric dialog training via offline reinforcement learning," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 3985–4003.

[12] Jen-Tzung Chien and Shu-Hsiang Yang, "Model-based soft actor-critic," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2021, pp. 2028–2035.

[13] Mahdin Rohmatillah, Jen-Tzung Chien, et al., "Robust multi-domain multi-turn dialogue policy via student-teacher offline reinforcement learning," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 1, pp. e18, 2024.

[14] Scott Fujimoto, David Meger, and Doina Precup, "Off-policy deep reinforcement learning without exploration," in *Proc. of International Conference on Machine Learning*, 2019, pp. 2052–2062.

[15] Jen-Tzung Chien and Kuan Chen, "Learning outlier-aware representation with synthetic boundary samples," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, 2023, pp. 1–6.

[16] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine, "Stabilizing off-policy Q-learning via bootstrapping error reduction," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] Scott Fujimoto and Shixiang Shane Gu, "A minimalist approach to offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[18] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine, "Conservative Q-learning for offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1179–1191, 2020.

[19] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma, "MOPO: Model-based offline policy optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14129–14142, 2020.

[20] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn, "COMBO: Conservative offline model-based policy optimization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28954–28967, 2021.

[21] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving language understanding by generative pre-training," *OpenAI blog*, 2018.

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.

[23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[24] Jen-Tzung Chien, Yi-Chung Chiu, et al., "Bayesian multi-temporal-difference learning," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, pp. e34, 2022.

[25] Ge Li, Mingyao Wu, Chensheng Wang, and Zhuo Liu, "DQ-HGAN: A heterogeneous graph attention network based deep Q-learning for emotional support conversation generation," *Knowledge-Based Systems*, vol. 283, pp. 111201, 2024.

[26] Linhong Xu, Hongfei Lin, Yu Pan, Hui Ren, and Jianmei Chen, "Constructing the affective lexicon ontology," *Journal of China Society for Scientific and Technical Information*, vol. 27, pp. 180–185, 2008.

[27] Jen-Tzung Chien and Chih-Jung Tsai, "Amortized mixture prior for variational sequence generation," in *Proc. of International Joint Conference on Neural Networks*, 2020, pp. 1–6.

[28] Jen-Tzung Chien and Weiwei Lai, "Variational skill embeddings for meta reinforcement learning," in *Proc. of International Joint Conference on Neural Networks*, 2023, pp. 1–8.

[29] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proc of AAAI Conference on Artificial Intelligence*, 2018, vol. 32, pp. 730–738.

[30] Hado Van Hasselt, Arthur Guez, and David Silver, "Deep reinforcement learning with double Q-learning," in *Proc. of AAAI conference on Artificial Intelligence*, 2016, vol. 30, pp. 2094–2100.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. of International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[33] Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang, "A large-scale chinese short-text conversation dataset," in *Proc. of CCF International Conference on Natural Language Processing and Chinese Computing*, 2020, pp. 91–103.

[34] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 654–664.

[35] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan, "A diversity-promoting objective function for neural conversation models," in *Proc. of Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 110–119.

On Joint Dereverberation and Single Moving Source Separation with Online Source Steering

Yiting Zhang†, Kaien Mo†, Tetsuya Ueda†, Yichen Yang† and Shoji Makino†
† Waseda University, Japan

***Abstract*—To solve the Blind Source Separation (BSS) problems in time-variant and highly reverberant environments, an online joint optimization algorithm for Weighted Prediction Error (WPE)-based dereverberation and Independent Vector Analysis (IVA)-based source separation (online WPE-IVA) has been reported. On the other hand, Online Source Steering (OSS) was recently introduced for filter updating in online IVA with low computational cost for a target-source tracking task where only one speaker is moving. Inspired by this work, we integrate OSS into online WPE-IVA, aiming to reduce the computational cost of updating separation filters. Experimental results show that our proposed method reduces computational costs while achieving comparable separation performance to conventional online WPE-IVA in target-source tracking tasks in highly reverberant environments.**

I. INTRODUCTION

Blind Source Separation (BSS) is a group of methods aiming to extract clean source signals from mixtures without any prior information [1]. Short-time Fourier transform (STFT)-domain Independent Component Analysis (ICA) is a popular multi-channel BSS method, which estimates separation filters based on the statistical independence between the sources [2]. Since the STFT-domain ICA calculates the separation filters in each frequency bin independently, the output permutation of sources may not be aligned in different frequency bins. Independent Vector Analysis (IVA) is proposed to address the above permutation problem by leveraging the high-order dependencies among frequency bins [3], [4]. Furthermore, the majorization-minimization algorithm is used for IVA's separation filter updating, which is widely known as Auxiliary-function-based IVA (AuxIVA) [5].

IVA-based methods are initially used in offline processing, which will cause large latency and cannot adapt to time-variant environments. As a solution for these problems, online IVA-based methods are proposed [6]–[8]. Unlike the offline IVA which processes a long batch of data at once, the online IVA updates separation filters incrementally after every new frame is received. This makes it suitable for real-time processing under time-variant environments. Moreover, to accelerate computational efficiency, previous research proposed an algorithm, Online Source Steering (OSS) [9] for online IVA, which updates the separation filters under the task where a certain source moves and other sources are spatially stationary (target-source tracking task).

The methods discussed above suppose that the analysis window of STFT is longer than the length of room impulse responses [10]. This assumption is hard to be satisfied when reverberation is strong and the IVA-based methods suffer from significant performance degradation. To mitigate the impact of strong reverberation, Weighted Prediction Error (WPE)-based dereverberation is usually used before separation [11], [12]. Furthermore, an online algorithm that jointly optimizes WPE and IVA (online WPE-IVA) [13], [14] has been proposed to achieve highly accurate source separation in highly reverberant environments. However, the highly efficient implementation for online WPE-IVA has not been developed yet for the target-source tracking task.

This work aims to further promote the computational efficiency of existing online WPE-IVA [13], [14] for the target-source tracking task. We introduce OSS into updating separation filters efficiently for online WPE-IVA, which is named "WPE-IVA-OSS". Since matrix operation for updating separation filters is equally converted to the scalar operation like [9], the computational cost of WPE-IVA-OSS is lower than existing online WPE-IVA. Simulation experiments are carried out to validate the efficiency of our proposed method in highly reverberant environments.

II. MODELS

A. Separation model

Suppose that there are N speech sources in a reverberant environment and M microphones are used to capture the signals. The captured signals in the STFT domain can be expressed as

$$\boldsymbol{x}_{f,t} = \sum_{\tau=0}^{L_a-1} \boldsymbol{A}_{f,t,\tau} \boldsymbol{s}_{f,t-\tau}, \tag{1}$$

where $t = 1, \cdots, T$ and $f = 1, \cdots, F$ are the indexes of time frames and frequency bins, respectively. T denotes the total number of time frames and F denotes the total number of frequency bins. $\boldsymbol{s}_{f,t} \in \mathbb{C}^{N\times 1}$ and $\boldsymbol{x}_{f,t} \in \mathbb{C}^{M\times 1}$ are the vectors containing the source and microphone signals, respectively. $\boldsymbol{A}_{f,t,\tau} \in \mathbb{C}^{M\times N}$ is the convolutional mixing matrix at time lag τ, and L_a is the number of time lagged matrices. This paper assumes a determined case ($M = N$). Then, the separation rule can be expressed using a convolutional beamformer (CBF) [15], [16] as

$$\boldsymbol{y}_{f,t} = \boldsymbol{Q}_{f,t,0} \boldsymbol{x}_{f,t} + \sum_{\tau=D}^{L+D-1} \boldsymbol{Q}_{f,t,\tau} \boldsymbol{x}_{f,t-\tau}, \tag{2}$$

where $\boldsymbol{Q}_{f,t,0} \in \mathbb{C}^{N\times M}$ and $\boldsymbol{Q}_{f,t,\tau} \in \mathbb{C}^{N\times M}$ are the coefficient matrix of CBF, $\boldsymbol{y}_{f,t} \in \mathbb{C}^{N\times 1}$ is the vector containing the

separated signals, D is the prediction delay and L is the length of CBF filters. By the source-wise factorization of CBF [16], the nth separated signal in $\boldsymbol{y}_{f,t}$ can be rewritten into two equations:

$$\boldsymbol{z}_{n,f,t} = \boldsymbol{x}_{f,t} - \boldsymbol{G}_{n,f,t}^{\mathsf{H}} \overline{\boldsymbol{x}}_{f,t}, \tag{3}$$

$$y_{n,f,t} = \boldsymbol{w}_{n,f,t}^{\mathsf{H}} \boldsymbol{z}_{n,f,t}, \tag{4}$$

where $\overline{\boldsymbol{x}}_{f,t} = [\boldsymbol{x}_{f,t-D}^{\mathsf{T}} \ \cdots \ \boldsymbol{x}_{f,t-L-D+1}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{ML\times 1}$ represents a vector containing past microphone signals, $(\cdot)^{\mathsf{T}}$ and $(\cdot)^{\mathsf{H}}$ denote transpose and conjugate transpose, respectively. The first sub-filter $\boldsymbol{G}_{n,f,t} \in \mathbb{C}^{ML\times M}$ is a single-target dereverberation filter for the nth source. The output of this filter, $\boldsymbol{z}_{n,f,t}$, is the dereverberated signal of the nth source. The second sub-filter $\boldsymbol{w}_{n,f,t} \in \mathbb{C}^{M\times 1}$ is a separation filter extracting the nth source.

B. Probabilistic model

Same as WPE-IVA [16], we assume that each separated signal independently follows a zero-mean multivariate complex Gaussian distribution with a time-dependent variance $r_{n,t}$:

$$p(\boldsymbol{y}_{n,t}) = \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}_F, r_{n,t}\boldsymbol{I}_F), \tag{5}$$

where $\boldsymbol{y}_{n,t} = [y_{n,1,t} \ \cdots \ y_{n,F,t}]^{\mathsf{T}} \in \mathbb{C}^{F\times 1}$, $\boldsymbol{0}_F \in \mathbb{C}^{F\times 1}$ is a zero vector, and $\boldsymbol{I}_F$ denotes a $F \times F$ identity matrix. By introducing the forgetting factor $0 < \beta < 1$, the negative log-likelihood function for online processing can be derived:

$$\begin{aligned}\mathcal{L}(\Theta_t) =& \frac{1}{\sum_{t'\le t}\beta^{t-t'}} \sum_{n,f,t'\le t} \beta^{t-t'} \left(\log r_{n,t'} + \frac{|y_{n,f,t'}|^2}{r_{n,t'}}\right) \\ & - 2\sum_f \log|\det \boldsymbol{W}_{f,t}|,\end{aligned} \tag{6}$$

where $\Theta_t = \{\mathcal{R}_t, \mathcal{G}_t, \mathcal{W}_t\}$ is the parameter set to be estimated, $\mathcal{R}_t = \{r_{n,t}\}_n$, $\mathcal{G}_t = \{\boldsymbol{G}_{n,f,t}\}_{n,f}$, $\mathcal{W}_t = \{\boldsymbol{W}_{f,t}\}_f$, and $\boldsymbol{W}_{f,t} = [\boldsymbol{w}_{1,f,t} \ \cdots \ \boldsymbol{w}_{N,f,t}]^{\mathsf{H}}$ is the separation matrix.

The cost function (6) can be optimized iteratively at every time frame by using a recursive coordinate descent method [13], [17] as

$$\mathcal{R}_t \leftarrow \underset{\mathcal{R}_t}{\operatorname{argmin}} \ \mathcal{L}\left(\Theta_t; \mathcal{R}_{t-1}, \mathcal{G}_{t-1}, \mathcal{W}_{t-1}\right), \tag{7}$$

$$\mathcal{G}_t \leftarrow \underset{\mathcal{G}_t}{\operatorname{argmin}} \ \mathcal{L}\left(\Theta_t; \mathcal{R}_t, \mathcal{G}_{t-1}, \mathcal{W}_{t-1}\right), \tag{8}$$

$$\mathcal{W}_t \leftarrow \underset{\mathcal{W}_t}{\operatorname{argmin}} \ \mathcal{L}\left(\Theta_t; \mathcal{R}_t, \mathcal{G}_t, \mathcal{W}_{t-1}\right). \tag{9}$$

Based on the above update rule, online WPE-IVA [13], [14] have been proposed.

III. PROPOSED METHOD

In this section, we propose WPE-IVA-OSS by introducing the update rule of OSS [9] into online WPE-IVA [13], [14]. We consider the target-source tracking task [9], where only one target source is moving and other $N-1$ sources are stationary. Hereafter, we introduce each update rule for the target-source tracking task. Note that we briefly introduce the updates of $\mathcal{R}_t$ and $\mathcal{G}_t$ for a better reading experience because they are the same as [13].

A. Update of $\mathcal{R}_t$ [13]

After calculating $\boldsymbol{y}_{f,t}$ through (3) and (4) with the filters $\boldsymbol{G}_{n,f,t-1}$ and $\boldsymbol{W}_{f,t-1}$ updated in the previous time frame, $r_{n,t}$ can be updated by

$$r_{n,t} \leftarrow \frac{1}{F}\sum_f |y_{n,f,t}|^2. \tag{10}$$

B. Update of $\mathcal{G}_t$ [13]

According to [13], the update rule for $\boldsymbol{G}_{n,f,t}$ can be written:

$$\boldsymbol{k}_{n,f,t} = \frac{\boldsymbol{R}_{n,f,t-1}^{-1}\overline{\boldsymbol{x}}_{f,t}}{\beta r_{n,t} + \overline{\boldsymbol{x}}_{f,t}^{\mathsf{H}}\boldsymbol{R}_{n,f,t-1}^{-1}\overline{\boldsymbol{x}}_{f,t}}, \tag{11}$$

$$\boldsymbol{R}_{n,f,t}^{-1} = \frac{1}{\beta}\left(\boldsymbol{R}_{n,f,t-1}^{-1} - \boldsymbol{k}_{n,f,t}\overline{\boldsymbol{x}}_{f,t}^{\mathsf{H}}\boldsymbol{R}_{n,f,t-1}^{-1}\right), \tag{12}$$

$$\boldsymbol{G}_{n,f,t} = \boldsymbol{G}_{n,f,t-1} + \boldsymbol{k}_{n,f,t}\boldsymbol{z}_{n,f,t}^{\mathsf{H}}, \tag{13}$$

where $\boldsymbol{R}_{n,f,t} = \sum_{t'\le t}\beta^{t-t'}\overline{\boldsymbol{x}}_{f,t'}\overline{\boldsymbol{x}}_{f,t'}^{\mathsf{H}}/r_{n,t'}$ is a spatio-temporal covariance matrix and $\boldsymbol{k}_{n,f,t}$ is the Kalman gain.

C. Update of $\mathcal{W}_t$

If $\mathcal{G}_t$ and $\mathcal{R}_t$ are fixed, the cost function (6) is equivalent to the one used in online AuxIVA:

$$\mathcal{L}(\mathcal{W}_t) = \sum_{n,f}\boldsymbol{w}_{n,f,t}^{\mathsf{H}}\boldsymbol{V}_{n,f,t}\boldsymbol{w}_{n,f,t} - 2\sum_f \log|\det \boldsymbol{W}_{f,t}|, \tag{14}$$

where

$$\boldsymbol{V}_{n,f,t} = \alpha\boldsymbol{V}_{n,f,t-1} + (1-\alpha)\frac{\boldsymbol{z}_{n,f,t}\boldsymbol{z}_{n,f,t}^{\mathsf{H}}}{r_{n,t}} \tag{15}$$

is the recursive form of the spatial covariance matrix at every time frame, and α is the forgetting factor which is different from β in (6) for more practical parameter estimation [13], [18]. In WPE-IVA-OSS, we use the same rank-1 update for separation matrix $\boldsymbol{W}_{f,t}$ frame by frame [8], [9], [19]. Specifically, we let ith separation filter $\boldsymbol{w}_{i,f,t}$ corresponds to the moving speaker and update $\boldsymbol{W}_{f,t}$ using $\boldsymbol{W}_{f,t} \leftarrow \boldsymbol{\eta}_{f,t}\boldsymbol{w}_{i,f,t-1}^{\mathsf{H}}$ where $\boldsymbol{\eta}_{f,t}$ is the coefficients to be optimized. The update rule can be summarized with the following equations [8]:

$$u_{n,f,t} = \boldsymbol{w}_{i,f,t-1}^{\mathsf{H}}\boldsymbol{V}_{n,f,t}\boldsymbol{w}_{n,f,t-1}, \tag{16}$$

$$d_{n,f,t} = \boldsymbol{w}_{i,f,t-1}^{\mathsf{H}}\boldsymbol{V}_{n,f,t}\boldsymbol{w}_{i,f,t-1}, \tag{17}$$

$$\boldsymbol{w}_{n,f,t} = \begin{cases} d_{n,f,t}^{-1/2}\boldsymbol{w}_{i,f,t-1} & \text{if } n = i, \\ \boldsymbol{w}_{n,f,t-1} - \frac{u_{n,f,t}}{d_{n,f,t}}\boldsymbol{w}_{i,f,t-1} & \text{otherwise.} \end{cases} \tag{18}$$

Then, by using the same technique used for original OSS [9], we can obtain $u_{n,f,t}$ and $d_{n,f,t}$ as

$$u_{n,f,t} = \frac{(1-\alpha)\hat{y}_{i,f,t}y_{n,f,t}^*}{r_{n,t}}, \tag{19}$$

$$d_{n,f,t} = \alpha\frac{d_{n,f,t-1}}{d_{i,f,t-1}} + \frac{(1-\alpha)|\hat{y}_{i,f,t}|^2}{r_{n,t}}, \tag{20}$$

where $\hat{y}_{i,f,t} = \boldsymbol{w}_{i,f,t-1}^{\mathsf{H}}\boldsymbol{z}_{n,f,t}$ is the auxiliary estimated signal just for obtaining $u_{n,f,t}$ and $d_{n,f,t}$. By using the above rules into (18), this modification results in smaller computational complexity for updating $\mathcal{W}_t$ from $O(FN^3)$ (for online WPE-IVA with ISS [14]) to $O(FN^2)$ at every time frame t.

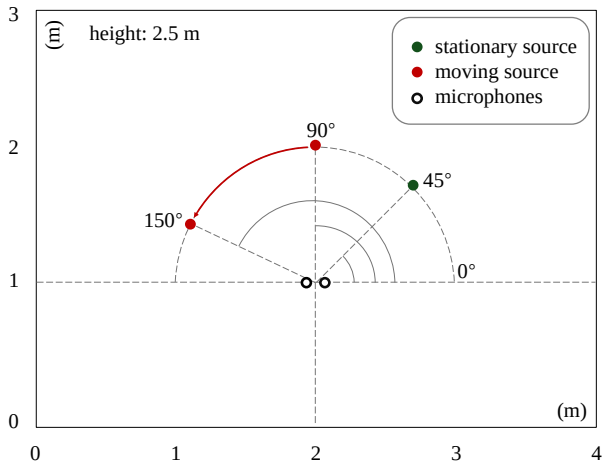


Fig. 1. Simulation layout. The stationary source is fixed at 45° for all 60 s and the moving source is fixed at 90° for the first 20 s. Then it started moving on an arc from 90° to 150° for the next 20 s and finally fixed at 150° for the last 20 s. The distance between the center of the microphone array and the source was 1 m.

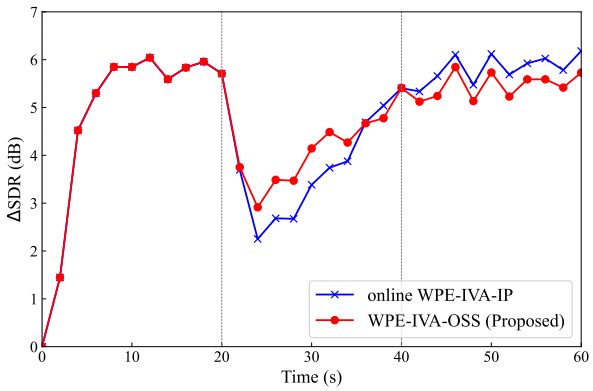


Fig. 2. SDR improvements (ΔSDR) measured every 2 s. Online WPE-GCIVA is used in the first 20 s, and WPE-IVA-OSS is applied in the last 40 s compared with online WPE-IVA-IP and online WPE-IVA-ISS. Since the SDR behavior of WPE-IVA-OSS is theoretically the same as online WPE-IVA-ISS, we omit drawing the latter.

TABLE I
TOTAL RUNTIME OF UPDATING THE SEPARATION FILTERS IN THE LAST 40 S

Method	Runtime
online WPE-IVA-IP [13]	2.4124 s
online WPE-IVA-ISS [14]	1.2775 s
WPE-IVA-OSS (Proposed)	0.4808 s

IV. EXPERIMENT

To evaluate the effectiveness of the proposed WPE-IVA-OSS, several target-source tracking experiments were conducted. We evaluate each method in terms of separation performance and computational efficiency.

A. Experimental Setup

20 mixture signals with a length of 60 s were generated with 2 source signals randomly selected from 2 speakers from the ATR Japanese Speech Database [20]. We concatenated the clean signals so that the signal length becomes 60 s. For mixing simulation, we used Matlab *signal_generator* based on the image method [21]. The position of sound sources and microphones is shown in Fig. 1, where the room boundary is controlled for a 600 ms room reverberation time T_{60}. A 2-element linear microphone array with an inter-element spacing of 2 cm was used. All the signals were sampled at 16 kHz. The STFT was computed with the Hann window of 64 ms (1024 samples) and a window shift length of 16 ms (256 samples). Other parameters α, β, D, L were set at 0.99, 0.96, 2, 10, respectively. The matrices $\boldsymbol{W}_{f,0}, \boldsymbol{G}_{n,f,0}, \boldsymbol{V}_{n,f,0}, \boldsymbol{R}^{-1}_{n,f,0}$ were initialized as $\boldsymbol{I}_M, \mathcal{O}_{ML\times M}, 10^{-5}\boldsymbol{I}_M, \boldsymbol{I}_{ML}$, where $\mathcal{O}_{ML\times M}$ is a zero matrix with the size as $ML \times M$.

Since the output permutation of separated signals is random when using online WPE-IVA [13], [14], we used Geometric Constraint (GC) [22] to detect which separated signal corresponds to the moving source. Specifically, we used online WPE-IVA with GC (online WPE-GCIVA) [23], in which we added the following penalty term derived from GC to the cost function in (6):

$$\mathcal{L}_{\mathrm{GC}}(\boldsymbol{W}_{f,t}) = \sum_{n} \sum_{\phi\in\Theta_n} \lambda_{f,\phi,t} |\boldsymbol{w}^{\mathsf{H}}_{n,f,t} \boldsymbol{d}_{\phi,f}|^2, \qquad (21)$$

where $\lambda_{f,\phi,t}$ is a weighting coefficient and $\Theta_n = \{\phi_1, \ldots, \phi_N\} \backslash \{\phi_n\}$ represents DOAs for all speakers excluding the nth speaker. $\boldsymbol{d}_{\phi,f}$ is the steering vector pointing to the ϕ direction. By applying and decreasing this penalty term, we can force the nth separation filter $\boldsymbol{w}_{n,f,t}$ to create a spatial null in the $\phi \in \Theta_n$ directions. We set angles for GC $\{\phi_1, \phi_2\} = \{45°, 90°\}$. We set $\lambda_{f,\phi,t} = \lambda_{f,\phi,0}\sigma^t$ where σ is the decreasing factor set to 0.8 and we set $\lambda_{f,\phi,0} = 8,000$. Since the detailed update rule has been discussed in [23] and is not directly related to our main discussion, we skip explaining it. In this paper, we used the above online WPE-GCIVA for the first 20 s. Since the correspondence between separated signals and sources was obtained in the first 20 s, there was no need to use GC for the last 40 s. After that, we used three different methods for the last 40 s: online WPE-IVA using Iterative Projection (online WPE-IVA-IP) [13], online WPE-IVA using Iterative Source Steering (online WPE-IVA-ISS) [14], and the proposed WPE-IVA-OSS. The improvement of signal-to-distortion ratio (ΔSDR) [24] was used as the metric to evaluate the separation performance.

B. Results

The SDR improvements of all the studied methods are shown in Fig. 2. In the period of 20 to 40 s when there was a source moving, WPE-IVA-OSS converged faster than online WPE-IVA-IP. However, after the convergence, WPE-IVA-OSS performed slightly worse than online WPE-IVA-IP. These results might be because WPE-IVA-OSS updates the separation matrix by rank-1 updating instead of the intact updating in online WPE-IVA-IP.

Table I shows the average separation filter updating time for the target-source tracking task for the last 40 s signal over 20 trials. To better illustrate the acceleration brought by the OSS-based optimization, the results in Table I only show the time cost for updating the separation filters instead of both the dereverberation and separation filters. Additionally, we did not consider the runtime of the first 20 s because the same method was used during the first 20 s. The results demonstrate that WPE-IVA-OSS spends a lower runtime than online WPE-IVA-IP [13] and online WPE-IVA-ISS [14].

V. CONCLUSION

This paper proposed a new algorithm named WPE-IVA-OSS for the target-source tracking task in highly reverberant environments. Thanks to the OSS algorithm, the separation matrix updating in WPE-IVA-OSS is simplified from matrix operations into scalar operations, and the computational complexity is reduced from $O(FN^3)$ to $O(FN^2)$. The experimental results show that our proposed method is computationally more efficient while achieving the same or at least comparable separation performance compared to the baseline methods.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 23H03423.

REFERENCES

[1] S. Makino, *Audio Source Separation*. Switzerland: Springer, 2018.
[2] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.
[3] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.
[4] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. ACSSC*, 2006, pp. 1393–1396.
[5] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE WASPAA*, 2011, pp. 189–192.
[6] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. Circuits Syst. I*, vol. 57, no. 7, pp. 1431–1438, 2010.
[7] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc HSCMA*, 2014, pp. 107–111.
[8] T. Nakashima and N. Ono, "Inverse-free online independent vector analysis with flexible iterative source steering," in *Proc. APSIPA ASC*, 2022, pp. 749–753.
[9] T. Nakashima, R. Ikeshita, N. Ono, S. Araki, and T. Nakatani, "Fast online source steering algorithm for tracking single moving source using online independent vector analysis," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.
[10] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sept. 2010.
[11] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sept. 2010.
[12] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.
[13] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online blind source separation based on joint optimization with blind dereverberation," in *Proc IEEE ICASSP*, 2021, pp. 506–510.
[14] K. Mo, X. Wang, Y. Yang, S. Makino, and J. Chen, "Low algorithmic delay implementation of convolutional beamformer for online joint source separation and dereverberation," *arXiv preprint arXiv:2406.09821*, 2024.
[15] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2267–2282, 2020.
[16] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Computationally efficient and versatile framework for joint optimization of blind speech separation and dereverberation," in *Proc. Interspeech*, 2020, pp. 91–95.
[17] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2010.
[18] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Blind and spatially-regularized online joint optimization of source separation, dereverberation, and noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1157–1172, 2024.
[19] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE ICASSP*, 2020, pp. 236–240.
[20] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.
[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
[22] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 10, no. 6, pp. 352–362, Dec. 2002.
[23] K. Mo, X. Wang, Y. Yang, T. Ueda, S. Makino, and J. Chen, "On joint dereverberation and source separation with geometrical constraints and iterative source steering," in *Proc. APSIPA ASC*, 2023, pp. 1138–1142.
[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.

Block Refinement Learning for Improving Early Exit in Autoregressive ASR

Naotaka Kawata*, Shota Orihashi*, Satoshi Suzuki*, Tomohiro Tanaka*, Mana Ihori*, Naoki Maikishima*, Taiga Yamane*, and Ryo Masumura*
* NTT Human Informatics Laboratories, NTT Corporation, Japan
E-mail: naotaka.kawata@ntt.com

***Abstract*—While the autoregressive transformer models of automatic speech recognition (ASR) are highly accurate, the inference time is long because of their sequential decoding. Early exit is a technique that aims to speed up the inference process by terminating it early on the basis of output from intermediate decoder layers. When an inference result with a high-confidence token is obtained in low intermediate layers, inference can be terminated at that point and the computational complexity is reduced compared to not terminating. In order to terminate inference early, it is necessary to improve the accuracy of the low intermediate layers. However, early exit often focuses on improving the accuracy of high intermediate layers because they determine the upper accuracy limit of ASR. As a result, it is difficult to terminate inference in low intermediate layers because the confidence of those layers is low. To solve this problem, we propose block refinement learning (BRL), which is a re-training method of existing early-exiting models. BRL trains the low intermediate layers while maintaining the overall accuracy of the model by considering gradients of not only low intermediate layers but also high intermediate layers. In this way, low intermediate layers can be trained while maintaining the accuracy of the high intermediate layers. We demonstrated the effectiveness of BRL on Japanese discourse ASR tasks.**

I. Introduction

Reducing inference time is important for automatic speech recognition (ASR). Previous studies have reduced the inference time by reducing the model parameters and computational complexity through the use of knowledge distillation [1]–[3], quantization [4]–[6], and pruning [7]–[9].

On the other hand, early exit [10], [11], which aims to speed up the inference process of ASR, has attracted attention in recent years due to its high affinity with ASR models that have multi-layer decoder. Early exit trains a model that has output layer associated with each of its intermediate decoder layers as well as the final decoder layer. In this paper, this model is referred to as the "early-exiting model". Also, the pair of a decoder layer and an output layer attached to it is referred to as the "intermediate layer". The early-exiting model can reduce the inference time by dynamically selecting inference paths for each input on the basis of the outputs of the intermediate layers. For instance, by terminating inference in low intermediate layers when inference result in low intermediate layers is high confidence and continuing inference until high intermediate layers when inference result in low intermediate layers is low confidence.

In training the early-exiting model, it often focuses on the accuracy of the high intermediate layers because it defines the upper limit of the ASR accuracy. However, the training policy that focuses on the accuracy of the high intermediate layers neglects the accuracy of low intermediate layers. As a result, the effect of reducing the inference time is smaller because it becomes hard for early exit to terminate inference in low intermediate layers. Given this background, a method is needed to improve the accuracy of low intermediate layers while maintaining the accuracy of high intermediate layers.

In this paper, we propose block refinement learning (BRL), which is a training method of the early-exiting model, to improve the ASR accuracy of low intermediate layers. An overview of BRL is shown in Figure 1. The key advantage of BRL is that it improves the ASR accuracy of low intermediate layers while maintaining the ASR accuracy of high intermediate layers in the early-exiting model. To this end, it trains the parameters of low intermediate layers, while keeping the parameters of all other layers fixed. At the same time, BRL trains low intermediate layers by exploiting not only the gradients for the corresponding attached output layers, but also the gradients for all output layers including the high intermediate ones. We demonstrated the effectiveness of BRL on Japanese discourse ASR tasks.

Our contributions are summarized as follows:

- We established a method to improve the ASR accuracy of low intermediate layers while maintaining the ASR accuracy of high intermediate layers in the early-exiting model.
- By our method, we experimentally demonstrated faster inference while maintaining ASR performance.

II. Related work

This section details previous studies on early exit. The previous studies used two different training methods to train the early-exiting models which has multiple output layers: two stage training [12]–[14], in which only the added output layers are trained, and joint training [15]–[17], in which the entire early-exiting model is trained in a single training. With two-stage training, only the added output layers are trained, while the parameters of the other component are fixed. Therefore, it is possible to maintain the performance of the final output layer. On the other hand, with joint training, all parameters included in the early-exiting model are trained in a single training. It has been pointed out that the accuracy of low intermediate layers is poor in two-stage training because it

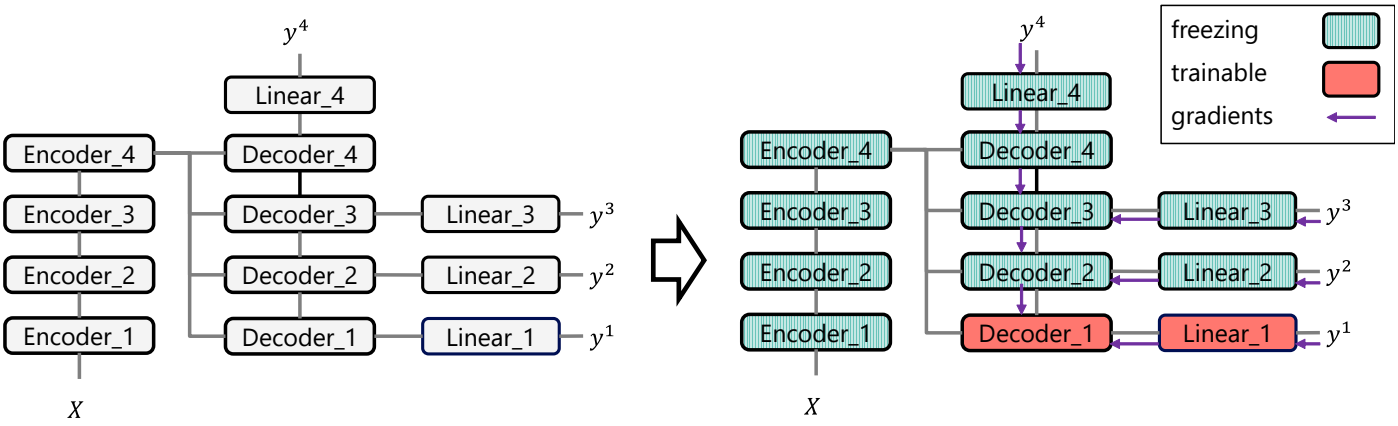


Fig. 1: Overview of the proposed method. The left side shows the early-exiting model, and the right side shows the training of BRL for the early-exiting model. As an example, this figure shows the early-exiting model with four encoder layers and four decoder layers, where the first layer is an intermediate layer trained by BRL.

trains only the added output layers [18]. Also, the accuracy of low intermediate layers is high in joint training because joint training trains all layers, not just the added ones. In this paper, we focus on joint training for its advantage of reducing the inference time by ensuring the high accuracy of the low intermediate layers.

In inference, the early-exiting model achieves a speedup by terminating the inference early based on exit criteria. The possible exit criteria include entropy-based [19], probability-based [16], [20], and patience-based methods [21]. We investigated the effectiveness of BRL by using probability-based and patience-based exit criteria as representative examples.

III. The Transformer-based Early-exiting Model

A. Transformer-based ASR model

This section details the ASR model based on transformer [22]. This model estimates a text $\boldsymbol{Y} = (y_1, \dots, y_T)$, that is the inference result from acoustic features $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$. y_t is the t-th token of the text, and T is the length of output sequence. $\boldsymbol{x}_n$ is the n-th frame of the features, and N is the frame length. The ASR model estimates the following probability P of generating $\boldsymbol{Y}$ for $\boldsymbol{X}$:

$$P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Theta}) = \prod_{t=1}^{T} P(y_t|y_{1:t-1}, \boldsymbol{X};\boldsymbol{\Theta}), \tag{1}$$

where $y_{1:t-1} = (y_1, \dots, y_{t-1})$ and $\boldsymbol{\Theta}$ denotes the ASR model parameter. The probability P obtained by the transformer-based ASR model is as follows:

$$\boldsymbol{h} = \mathrm{TransformerEnc}(\boldsymbol{X};\boldsymbol{\theta}_{\mathrm{enc}}), \tag{2}$$

$$\boldsymbol{e}_{t-1} = \mathrm{Embedding}(y_{t-1};\boldsymbol{\theta}_{\mathrm{emb}}), \tag{3}$$

$$\boldsymbol{c}^l_{t-1} = \begin{cases} \mathrm{TransformerDec}(\boldsymbol{h}, \boldsymbol{e}_{1:t-1};\boldsymbol{\theta}^l_{\mathrm{dec}}) & \text{if } l = 1, \\ \mathrm{TransformerDec}(\boldsymbol{h}, \boldsymbol{c}^{l-1}_{1:t-1};\boldsymbol{\theta}^l_{\mathrm{dec}}) & \text{otherwise}, \end{cases} \tag{4}$$

$$P(y_t|y_{1:t-1}, \boldsymbol{X};\boldsymbol{\Theta}) = \mathrm{Softmax}(\boldsymbol{c}^L_{t-1};\boldsymbol{\theta}^L_{\mathrm{linear}}), \tag{5}$$

where $\mathrm{TransformerEnc}(\cdot)$ is a transformer encoder that consists of an embedding layer, positional encoding layer, and multiple transformer encoder layers. The transformer encoder layer consists of a multi-head self-attention layer and a feed-forward layer. $\boldsymbol{\theta}_{\mathrm{enc}}$ denotes its parameters. $\mathrm{Embedding}(\cdot)$ is an embedding layer that consists of an embedding layer and a positional encoding layer. $\boldsymbol{\theta}_{\mathrm{emb}}$ denotes its parameters. $\mathrm{TransformerDec}(\cdot)$ is a transformer decoder that consists of a masked-multi-head attention layer, encoder-decoder multi-head-attention layer, and feed-forward layer. $\boldsymbol{\theta}^l_{\mathrm{dec}}$ denotes the parameter of the l-th transformer decoder, where $l \in [1, L]$. L is the number of decoder layers. $\mathrm{Softmax}(\cdot)$ is a softmax layer with a linear transformation. $\boldsymbol{\theta}^l_{\mathrm{linear}}$ denotes the parameter of the l-th softmax layer connected to the l-th transformer decoder. Note that the softmax layer is only connected to the last transformer decoder in the basic ASR model. The parameter $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{\mathrm{enc}}, \boldsymbol{\theta}_{\mathrm{emb}}, \boldsymbol{\theta}^1_{\mathrm{dec}}, \dots, \boldsymbol{\theta}^L_{\mathrm{dec}}, \boldsymbol{\theta}^L_{\mathrm{linear}}\}$ is optimized with a cross-entropy loss function $\mathcal{L}$, defined as

$$\mathcal{L}(\boldsymbol{\Theta}) = -\sum_{(\boldsymbol{X},\boldsymbol{Y})\in\mathcal{D}} \sum_{t=1}^{T} \log P(y_t|y_{1:t-1}, \boldsymbol{X};\boldsymbol{\Theta}). \tag{6}$$

$\mathcal{D}$ denotes training data containing multiple $\boldsymbol{X}$ and $\boldsymbol{Y}$ pair data.

B. The early-exiting model

This section details the transformer-based early-exiting ASR model. The model can obtain probabilities P^l from l-th intermediate decoder layer. The probabilities P^l are obtained as follows:

$$P^l(y_t|y_{1:t-1}, \boldsymbol{X};\boldsymbol{\Theta}_E) = \mathrm{Softmax}(\boldsymbol{c}^l_{t-1};\boldsymbol{\theta}^l_{\mathrm{linear}}). \tag{7}$$

In the early-exiting model, one $\mathrm{Softmax}(\cdot)$ is connected to each $\mathrm{TransformerDec}(\cdot)$. The parameter $\boldsymbol{\Theta}_E = \{\boldsymbol{\theta}_{\mathrm{enc}}, \boldsymbol{\theta}_{\mathrm{emb}}, \boldsymbol{\theta}^1_{\mathrm{dec}}, \dots, \boldsymbol{\theta}^L_{\mathrm{dec}}, \boldsymbol{\theta}^1_{\mathrm{linear}}, \dots, \boldsymbol{\theta}^L_{\mathrm{linear}}\}$ is optimized

Algorithm 1 Inference with probability-based exit criteria (Input: $\boldsymbol{X}$, Output: $\boldsymbol{Y}$)

$\boldsymbol{h} \leftarrow \mathrm{TransformerEnc}(\boldsymbol{X}; \boldsymbol{\theta}_{\mathrm{enc}})$
for $t = 1$ to T **do**
  $\boldsymbol{e}_{t-1} = \mathrm{Embedding}(y_{1:t-1}; \boldsymbol{\theta}_{\mathrm{emb}})$
  for $l = 1$ to L **do**
    if $l = 1$ **then**
      $\boldsymbol{c}^l_{t-1} \leftarrow \mathrm{TransformerDec}(\boldsymbol{h}, \boldsymbol{e}_{1:t-1}; \boldsymbol{\theta}^l_{\mathrm{dec}})$
    else
      $\boldsymbol{c}^l_{t-1} \leftarrow \mathrm{TransformerDec}(\boldsymbol{h}, \boldsymbol{c}^{l-1}_{1:t-1}; \boldsymbol{\theta}^l_{\mathrm{dec}})$
    end if
    $P^l(y_t|y_{1:t-1}, \boldsymbol{X}; \boldsymbol{\Theta}_{\mathrm{E}}) \leftarrow \mathrm{Softmax}(\boldsymbol{c}^l_{t-1}; \boldsymbol{\theta}^l_{\mathrm{linear}})$
    $s^l_t \leftarrow \mathrm{Score}(P^l(y_t|y_{1:t-1}, \boldsymbol{X}; \boldsymbol{\Theta}_{\mathrm{E}}))$
    if $s^l_t \geq \lambda^l$ **then**
      $y_t \leftarrow \arg\max_{y_t} P^l(y_t|y_{1:t-1}, \boldsymbol{X}; \boldsymbol{\Theta}_{\mathrm{E}})$
      break
    end if
  end for
end for

with a cross-entropy loss function $\mathcal{L}_E$, defined as

$$\mathcal{L}_E(\boldsymbol{\Theta}_E) = -\sum_{(\boldsymbol{X},\boldsymbol{Y})\in\mathcal{D}} \sum_{t=1}^{T} \sum_{l=1}^{L} w^l \log P^l(y_t|y_{1:t-1}, \boldsymbol{X}; \boldsymbol{\Theta}_E). \tag{8}$$

The weights w^l set in Schuster et al. [23] favor higher layers, as follows:

$$w^l = \frac{l}{\sum_{k=1}^{L} k}. \tag{9}$$

C. Inference of the early-exiting model

This section details two exit criteria of early exit, probability-based [20] and patience-based (PABEE) [21]. The probability-based exit criteria terminates inference when the early-exiting model outputs high-confidence token from its intermediate layers. Specifically, terminating inference occurs when the confidence score $s^l_t \in [0, 1]$ exceeds the thresholds $\lambda^l \in [0, 1]$. The confidence score $s^l_t = \mathrm{Score}(\cdot)$ is defined by the difference between the top two values of $\mathrm{Softmax}(\cdot)$. The thresholds λ^l are determined by setting arbitrary objective functions and solving an optimization problem. The flow of inference with the probability-based exit criteria is shown in Algorithm 1.

PABEE terminates inference when the early-exiting model outputs same token continuously from its intermediate layers. It uses a patience counter p to store the number of times that outputs do not change and the threshold $\hat{p} \in [1, L-1]$. Terminating inference occurs when p is equal to $\hat{p}$. The flow of inference with the patience-based exit criteria is shown in Algorithm 2.

Note that due to the structure of the transformer self-attention mechanism, all previous hidden states $\boldsymbol{c}^{l-1}_{1:t-1}$ are essential to the calculation of $\boldsymbol{c}^l_{t-1}$. Therefore, when terminating inference occurred in l-th layer in the estimation of t-th token, hidden states $\{c^{l+1}_{t-1}, ..., c^L_{t-1}\}$ are complemented by copying c^l_{t-1}.

Algorithm 2 Inference with patience-based exit criteria (Input: $\boldsymbol{X}$, Output: $\boldsymbol{Y}$)

$\boldsymbol{h} \leftarrow \mathrm{TransformerEnc}(\boldsymbol{X}; \boldsymbol{\theta}_{\mathrm{enc}})$
for $t = 1$ to T **do**
  $p \leftarrow 0$
  $\boldsymbol{e}_{t-1} = \mathrm{Embedding}(y_{1:t-1}; \boldsymbol{\theta}_{\mathrm{emb}})$
  for $l = 1$ to L **do**
    if $l = 1$ **then**
      $\boldsymbol{c}^l_{t-1} \leftarrow \mathrm{TransformerDec}(\boldsymbol{h}, \boldsymbol{e}_{1:t-1}; \boldsymbol{\theta}^l_{\mathrm{dec}})$
    else
      $\boldsymbol{c}^l_{t-1} \leftarrow \mathrm{TransformerDec}(\boldsymbol{h}, \boldsymbol{c}^{l-1}_{1:t-1}; \boldsymbol{\theta}^l_{\mathrm{dec}})$
    end if
    $P^l(y_t|y_{1:t-1}, \boldsymbol{X}; \boldsymbol{\Theta}_{\mathrm{E}}) \leftarrow \mathrm{Softmax}(\boldsymbol{c}^l_{t-1}; \boldsymbol{\theta}^l_{\mathrm{linear}})$
    if $\arg\max_{y_t} P^l(y_t|y_{1:t-1}, \boldsymbol{X}; \boldsymbol{\Theta}_{\mathrm{E}})$
    $= \arg\max_{y_t} P^{l-1}(y_t|y_{1:t-1}, \boldsymbol{X}; \boldsymbol{\Theta}_{\mathrm{E}})$ **then**
      $p \leftarrow p + 1$
    else
      $p \leftarrow 0$
    end if
    if $p = \hat{p}$ **then**
      $y_t \leftarrow \arg\max_{y_t} P^l(y_t|y_{1:t-1}, \boldsymbol{X}; \boldsymbol{\Theta}_{\mathrm{E}})$
      break
    end if
  end for
end for

IV. Proposed method

This section details BRL, an effective method of training the early-exiting model. The key advantage of BRL is that it improves the ASR accuracy of specified low intermediate layers, while maintaining the ASR accuracy of high intermediate layers in the early-exiting model. BRL has two features:

- Fixing the parameters of high intermediate layers and training the parameters of low intermediate layers.
- Calculating gradients from not only low intermediate layers, but also high intermediate layers.

As shown in Figure 1, the initial parameter of BRL is the existing early-exiting model. In BRL, only the weights $\{\boldsymbol{\theta}^1_{\mathrm{dec}}, \ldots, \boldsymbol{\theta}^j_{\mathrm{dec}}, \boldsymbol{\theta}^1_{\mathrm{linear}}, \ldots, \boldsymbol{\theta}^j_{\mathrm{linear}}\}$ in $\boldsymbol{\Theta}_E$ are updated using Equation (8). $j \in [1, L]$ denotes the number of specified low intermediate layers to be trained.

BRL sets uniform weights for w^l as follows:

$$w^l = \frac{1}{L}. \tag{10}$$

TABLE I: Dataset details.

	Data size (hours)	Number of utterances	Number of characters
Train	512.6	413,240	13,349,780
Valid	1.9	1,292	47,970
Test	1.3	1,385	32,089

BRL sets uniform weights w^l because BRL focuses on improving the ASR accuracy of low intermediate layers, while maintaining the ASR accuracy of high intermediate layers.

V. Experiment

A. Dataset

We evaluated the effectiveness of BRL on Japanese discourse ASR tasks by using the corpus of spontaneous Japanese (CSJ) [24]. We used CSJ as the training set (Train), validation set (Valid), and test set (Test). Valid and Test are Eval2 and Eval3 provided by CSJ, respectively. Details of these datasets are given in Table I. We used 80 log mel-scale filterbank coefficients as acoustic features and text segmented into character units. To exclude speech data that was too long, we removed from the dataset any data that contained text with more than 300 tokens or more than $1,500$ audio frames.

B. Settings

The early-exiting model is a transformer-based encoder-decoder model with intermediate layers. We used two early-exiting models with different numbers of decoder layers: EE-base and EE-deep. EE-base has 8 encoder layers and 6 decoder layers. Every intermediate decoder layer in EE-base has a corresponding output layer. EE-deep has 8 encoder layers and 12 decoder layers, i.e., more decoder layers than in EE-base. EE-deep has an output layer for each even-numbered intermediate decoder layer. EE-base and EE-deep were created under the following conditions: the dimensions of the input and output continuous representations were set to 512, the dimensions of the inner outputs in the position-wise feed-forward networks were set to $2,048$, and the number of heads in the multi-head attention was set to 4. To stabilize training, the dropout rate in the transformer encoder and decoder layers was set to 0.1. For the activation function, we used swish activation layer [25].

BRL was applied to EE-base and EE-deep to yield BRL-base and BRL-deep. BRL-base was trained from the first through the third of its intermediate layers ($j = 3$) by utilizing BRL. Also, BRL-deep was trained from the first through the sixth of its intermediate layers ($j = 6$). The number of intermediate layers to be trained, j, was chosen to maximize the sum of the ASR accuracies obtained from all the intermediate layers.

In the training, our models were optimized by using RAdam [26]. We set a mini-batch size of 64 and the learning rate of 0.0001. We used label smoothing [27] for regularization. For data augmentation, we applied SpecAugment with frequency masking and time masking [28], where the number of frequency masks and time-step masks were each set to 2, the frequency-masking width was randomly chosen from 0 to 27 frequency bins, and the time-masking width was randomly chosen from 0 to 100 frames. Early stopping was applied if no best model was found in the validation set after 10 epochs.

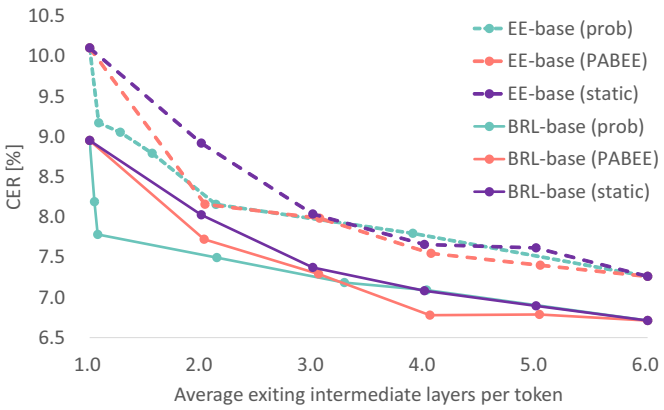


Fig. 2: CER and average exiting layer of models trained by BRL in EE-base.

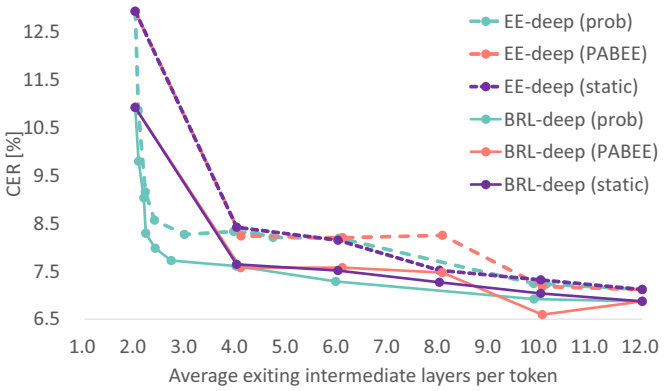


Fig. 3: CER and average exiting layer of models trained by BRL in EE-deep.

In the decoding, we used a beam search with a beam size of 4. In this experiment, we tested three exit criteria: probability-based (prob), patience-based (PABEE), and static (static). In prob, we set an optimal threshold λ^l that simultaneously minimized the character error rate (CER) and the average exiting intermediate layer per token in the validation set. To find the optimal threshold λ^l, we used NSGA-II [29], which is a multi-objective genetic algorithm for solving multi-objective optimization problems. We searched for the optimal λ^l in 300 epochs. In PABEE, we set a patience counter $\hat{p}$ from 1 to $L-1$. Static is the results of forced termination of the inference in each intermediate layer, which are presented to validate the training effect of BRL. We used the CPU, which was INTEL XEON Silver 4310, for the decoding and batch size was 1. The Intel oneAPI Math Kernel Library [30] was enabled for the computations, and 24 threads were used to perform parallel computations.

TABLE II: The impact of early exit on actual inference speed.

	↓ CER[%]	↑ Speedup
EE-base (static by final output)	7.26	×1.00
EE-base (prob)	7.80	×1.38
	8.16	×2.27
	9.17	×3.30
BRL-base (prob)	7.09	×1.39
	7.49	×2.56
	7.78	×3.43

The accuracy of the early-exiting model was measured by CER. Reduction in inference time was measured by the average exiting intermediate layers per token.

C. Results

Figure 2 shows the trade-off between inference speed and CER in EE-base. From a comparison of EE-base (static) and BRL-base (static), we can find that BRL improved the ASR accuracy of the individual low intermediate layers. From a comparison of EE-base (prob) and BRL-base (prob), we can also find that BRL-base (prob) achieved a comparable CER to that of EE-base (prob) with fewer average exiting intermediate layers. Furthermore, we can find the same trend even when the exit criteria were changed to PABEE. Figure 3 shows the trade-off between inference speed and CER in EE-deep. It confirms that BRL has a similar effect in EE-deep as in EE-base. These results indicate that BRL can improve the ASR accuracy to train a low intermediate layer while maintaining the ASR accuracy of high intermediate layers.

To show the actual effect of BRL, we evaluated it in terms of computation speed and CER. Table II shows the impact of early exit on the actual inference speed. EE-base (static by final output) shows the inference speed and CER when always exiting at the final decoder layer. Speedup is a ratio based on EE-base (static by final output). The effect of speedup can be compared by Speedup under close CER. By comparing EE-base (prob) and BRL-base (prob) in Table II, we can find that close CER to EE-base (static by final output) are achieved by Speedup of $\times 3.43$ in BRL-base (prob), while $\times 1.38$ in EE-base (prob). In other words, BRL-base (prob) achieves 2.49 times faster compared to EE-base (prob) at close CER.

Further, we investigated the effect of changing the number of trained low intermediate layers in BRL. We adjusted j from 1 to 6 in BRL. Figure 4 shows the result. Even when the number of intermediate layers to be trained changed, BRL remained effective under all experimental conditions. $j = 6$ is the result of re-training all decoder layers including the final one. $j = 6$ and EE-base (prob) have different settings for loss weights. The loss weights of $j = 6$ obey Equation (9) and the loss weights of EE-base (prob) obey Equation (10). The result that $j = 6$ exceeds the ASR performance of EE-base (prob) indicate that BRL can also be used for the early-exiting model with insufficient training in the intermediate layers.

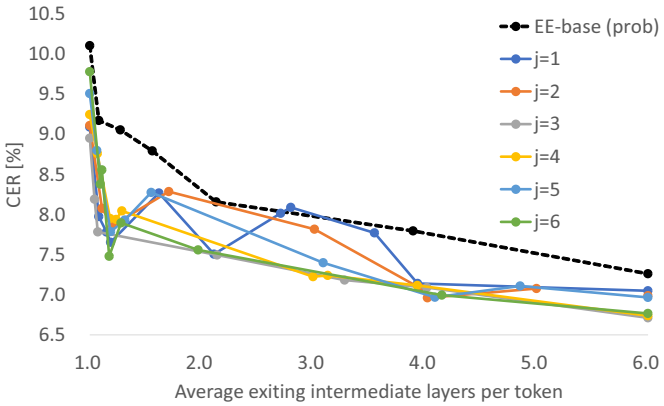


Fig. 4: All model used probability-based exit criteria. Result of varying the number of low intermediate layers. $j = 3$ indicates BRL-base (prob).

VI. Conclusion

We presented a block refinement learning (BRL) that achieves faster inference for the early-exiting model in automatic speech recognition (ASR). The key advantage of BRL is that it improves the ASR accuracy of low intermediate layers while maintaining the ASR accuracy of high intermediate layers in the existing early-exiting model. To this end, BRL trains the parameters of the low intermediate layers, while keeping the parameters of all other layers fixed. In addition, it exploits the gradients from not only the corresponding attached output layers, but also from all output layers including those associated with the high intermediate layers. We experimentally showed that BRL can improve the ASR accuracy of low intermediate layers while maintaining the ASR accuracy of high intermediate layers in the early-exiting model.

References

[1] M. A. Haidar, C. Xing, and M. Rezagholizadeh, "Transformer-based asr incorporating time-reduction layer and fine-tuning with self-knowledge distillation," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 2102–2106.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. Advances in Neural Information Processing Systems (NeurIPS) Workshop on Deep Learning*, 2015.

[3] L. Li, Y. Lin, D. Chen, S. Ren, P. Li, J. Zhou, and X. Sun, "Cascadebert: Accelerating inference of pre-trained language models via calibrated complete models cascade," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 475–486.

[4] S. Kim, A. Gholami, Z. Yao, N. Lee, P. Wang, A. Nrusimha, B. Zhai, T. Gao, M. W. Mahoney, and K. Keutzer, "Integer-only zero-shot quantization for efficient speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4288–4292.

[5] S. Shen, Z. Dong, J. Ye, L. Ma, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "Q-bert: Hessian based ultra low precision quantization of bert," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 8815–8821.

[6] W. Zhang, L. Hou, Y. Yin, L. Shang, X. Chen, X. Jiang, and Q. Liu, "TernaryBERT: Distillation-aware ultra-low bit BERT," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 509–521.

[7] C.-I. J. Lai, Y. Zhang, A. H. Liu, S. Chang, Y.-L. Liao, Y.-S. Chuang, K. Qian, S. Khurana, D. Cox, and J. Glass, "Parp: Prune, adjust and re-prune for self-supervised speech recognition," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[8] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[9] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 1135–1143.

[10] R. Tang, K. Kumar, J. Xin, P. Vyas, W. Li, G. Yang, Y. Mao, C. Murray, and J. Lin, "Temporal early exiting for streaming speech commands recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7567–7571.

[11] D. Berrebbi, B. Yan, and S. Watanabe, "Avoid overthinking in self-supervised models for speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[12] S. Scardapane, M. Scarpiniti, E. Baccarelli, and A. Uncini, "Why should we add early exits to neural networks?" *Cognitive Computation*, vol. 12, no. 5, pp. 954–966, Sep 2020.

[13] J. Xin, R. Tang, Y. Yu, and J. Lin, "BERxiT: Early exiting for BERT with better fine-tuning and extension to regression," in *Porc. European Chapter of the Association for Computational Linguistics (EACL)*, 2021, pp. 91–104.

[14] G. A. Wright, U. Cappellazzo, S. Zaiem, D. Raj, L. O. Yang, D. Falavigna, M. N. Ali, and A. Brutti, "Training early-exit architectures for automatic speech recognition: Fine-tuning pre-trained models or training from scratch," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7567–7571.

[15] S. Bae, J. Ko, H. Song, and S.-Y. Yun, "Fast and robust early-exiting framework for autoregressive language models with synchronized parallel decoding," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 5910–5924.

[16] S. Teerapittayanon, B. McDanel, and H. Kung, "Branchynet: Fast inference via early exiting from deep neural networks," in *Proc. International Conference on Pattern Recognition (ICPR)*, 2016, pp. 2464–2469.

[17] Y. Ji, J. Wang, J. Li, Q. Chen, W. Chen, and M. Zhang, "Early exit with disentangled representation and equiangular tight frame," in *Proc. Association for Computational Linguistics (ACL)*, 2023.

[18] J. Xin, R. Tang, J. Lee, Y. Yu, and J. Lin, "DeeBERT: Dynamic early exiting for accelerating BERT inference," in *Proc. Association for Computational Linguistics (ACL)*, 2020, pp. 2246–2251.

[19] S. Geng, P. Gao, Z. Fu, and Y. Zhang, "Romebert: Robust training of multi-exit bert," *arXiv preprint arXiv:2101.09755*, 2021.

[20] T. Schuster, A. Fisch, J. Gupta, M. Dehghani, D. Bahri, V. Q. Tran, Y. Tay, and D. Metzler, "Confident adaptive language modeling," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022, pp. 17 456–17 472.

[21] W. Zhou, C. Xu, T. Ge, J. McAuley, K. Xu, and F. Wei, "Bert loses patience: Fast and robust inference with early exit," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[23] Z. Liu, Z. Xu, H.-J. Wang, T. Darrell, and E. Shelhamer, "Anytime dense prediction with confidence adaptivity," in *Proc. International Conference on Learning Representations (ICLR)*, 2022.

[24] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. International Conference on Language Resources and Evaluation (LREC)*, 2000.

[25] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[26] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[28] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 2613–2617.

[29] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, pp. 182–197, 2002.

[30] M. Krainiuk, M. Goli, and V. R. Pascuzzi, "oneapi open-source math library interface," in *Proc. International Workshop on Performance, Portability and Productivity in HPC (P3HPC)*, 2021, pp. 22–32.

A Novel kind of WVD Associated with the Linear Canonical Transform

Jia-Yin Peng†, Jian-Yi Chen† and Bing-Zhao Li†,*
† Beijing institute of Technology, Beijing 100081, China
* E-mail: li_bingzhao@bit.edu.cn Tel: +86-13681493240

***Abstract*—Linear canonical transformation (LCT), as an extension of the Fourier transform and fractional Fourier transform, has emerged as a useful tool for the nonstationary signals processing. In this paper, a novel kind of Wigner-Ville distribution associated with the LCT (WVL) is presented. First, in order to eliminate the coupling between time variable t and lag variable τ, the WVL is defined based on the scaled parametric symmetric instantaneous autocorrelation function. Second, the WVL of multicomponent linear frequency modulated (LFM) signal is investigated, and the results show that WVL can mitigate the influence of some cross-terms. Finally, through simulation experiments and comparison with the classical WVD, it is verified that WVL has a better ability to suppress cross-terms and significantly improves the detection of multicomponent LFM signals.**

I. Introduction

Linear frequency modulated (LFM) [1] signal is a kind of nonstationary signal widely used in radar, sonar and communication systems, it has an important characteristic that each component can be uniquely determined by the centroid frequency and chirp rate (CFCR) and the spectral components of a LFM signal vary with time. Since traditional Fourier transform (FT) cannot meet the requirement of observation frequency changing with time, the LFM signals are often analyzed using time-frequency (TF) analysis methods. Common TF analysis methods include the short-time Fourier transform (STFT) [2], Wigner-Ville distribution (WVD) [3], short-time Wigner-Ville distribution (STWD) [4], wavelet transform (WT) [5], Wigner-Hough transform (WHT) [6], Radon-Wigner transform (RWT) [7], fractional Fourier transform (FRFT) [8] [9], linear canonical transform (LCT) [10] and time-fractional-frequency (TFF) [11]. Compared with other TF analysis methods, WVD has excellent TF resolution and energy concentration, making it more effective for the analysis of LFM signals. Despite this, when it is applied to multicomponent LFM signals, it is affected by cross-term interference. Therefore, reducing the impact of cross-term (interference) on auto-term (target) has become an important research focus.

In 2009, Lv et al., using the ideal of keystone formatting, proposed the keystone-Wigner transform (KWT) [12], which introduced a weight factor into the time-lag instantaneous autocorrelation function to decouple the time variable t and the lag variable τ. Along with this idea, Lv et al. improved the KWT and proposed the Lv's distribution (LVD) [13], which circumvented all the problems of the KWT, and enhanced the energy of the auto-term meanwhile reduced the energy of the cross-term in 2011. In 2017, the short time Fourier transform-multiple invariance-estimation of signal parameters via rotational invariance techniques (STFT-MI-ESPRIT) [2] was proposed by Cui et al., which improved the estimation precision greatly while maintaining computational complexity. In the same year, Wang et al. proposed the WHT [6], which used the double-deck weight to mitigate the impact of noise and disturbance. In [3] a new method to eliminate cross-terms in the WVD of multicomponent LFM signals was investigated. This method improved the signal to noise ratio (SNR) while maintaining high resolution by removing the cross-terms. In 2021, the definition of WVD associated with LCT (WVDL) [14] was proposed by Xin et al., which achieves smaller errors and lower computational costs compared with other WVD in LCT domain. [15] investigated the four-dimensional Wigner distributions associated with the linear canonical transform (WDLs) for the analytic signals in 2024.

This paper proposes a novel kind of WVD associated with the LCT (WVL). By using a scaling function, it decouples t and τ in the parametric symmetric instantaneous autocorrelation function (PSIAF) [13], eliminating the influence of some cross-terms thereby enhancing the accuracy of detecting multicomponent LFM signals. Then taking a 2D LCT [16] over the new scaled PSIAF time variable and lag variable respectively, can obtain WVL. Compared to the classical WVD, WVL offers more free parameters and greater flexibility.

The remainder of this paper is organized as follows: Section II reviews the relevant theories of LCT. In section III, presents the WVL from a new point of view and derive some properties and theorems. Section IV show the simulation results of the theorems and the effectiveness of the proposed tehniques. Section V is the conclution.

II. Preliminary

A. Two dimentional LCT

The 2D LCT of a signal $f(x, y)$ with parameter $A = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}$, $B = \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}$ is defined as follows[16]

$$
\begin{aligned}
&L_f^{A,B}[f(x,y)] = L_f^{A,B}(u,v)\\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)K_{A,B}(x,y,u,v)dxdy\\
&= \begin{cases}
\frac{1}{2\pi j}\sqrt{\frac{1}{b_1 b_2}}e^{j[(\frac{d_1u^2}{2b_1}+\frac{d_2v^2}{2b_2})]}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\\
\times e^{-j(\frac{ux}{b_1}+\frac{vy}{b_2})}e^{j(\frac{a_1x^2}{2b_1}+\frac{a_2y^2}{2b_2})}dxdy, b_1b_2\neq 0,\ |A|=|B|=1,\\
\sqrt{d_1d_2}e^{j[(\frac{c_1d_1u^2+c_2d_2v^2}{2})]}f(d_1u,d_2v), b_1^2+b_2^2=0,
\end{cases}
\end{aligned}
\tag{1}
$$

where $K_{A,B}(x,y,u,v) = K_A(x,u)K_B(y,v)$,

$$
\begin{aligned}
K_A(x,u) &= \sqrt{\frac{1}{2\pi j b_1}}e^{j\frac{d_1u^2}{2b_1}}e^{-j\frac{ux}{b_1}}e^{j\frac{a_1x^2}{2b_1}},\\
K_B(y,v) &= \sqrt{\frac{1}{2\pi j b_2}}e^{j\frac{d_2v^2}{2b_2}}e^{-j\frac{vy}{b_2}}e^{j\frac{a_2y^2}{2b_2}}\ .
\end{aligned}
$$

Some basic properties of 2D LCT have been studied, this paper uses the following two important properties:
(1) Let $h(x,y) = f(x,y)\times g(x,y)$. $L_h^{A,B}(u,v), L_g^{A,B}(u,v)$ denote 2D LCT of the $h(x,y), g(x,y)$ respectively, while $L_f(u,v)$ denotes 2D FT of the $f(x,y)$. The 2D LCT product theorem is defined as follows [17]

$$
\begin{aligned}
L_h^{A,B}(u,v) &= \frac{1}{4\pi^2}\frac{1}{|b_1b_2|}e^{j(\frac{d_1}{2b_1}u^2+\frac{d_2}{2b_2}v^2)}\\
&\times[(L_g^{A,B}(u,v)e^{-j(\frac{d_1}{2b_1}u^2+\frac{d_2}{2b_2}v^2)})*L_f(\frac{u}{b_1},\frac{v}{b_2})].
\end{aligned}
\tag{2}
$$

(2) Let $g(x,y) = e^{j(xu_0+yv_0)}f(x,y)$. $L_g^{A,B}(u,v), L_f^{A,B}(u,v)$ denote 2D LCT of the $g(x,y), f(x,y)$ respectively. The 2D LCT scaling properties is defined as follows [17]

$$
\begin{aligned}
L_g^{A,B}(u,v) &= e^{j(uu_0d_1+vv_0d_2)-\frac{j}{2}(u_0^2b_1d_1+v_0^2b_2d_2)}\\
&\times L_f^{A,B}(u-u_0b_1, v-v_0b_2),\ u_0,v_0\in R.
\end{aligned}
\tag{3}
$$

B. WVD and AF Associated with LCT

The WVD of a signal $f(t)$ associated with the LCT (WDL) with parameter $A = \begin{pmatrix} a & b\\ c & d\end{pmatrix}$ is defined as [18]

$$
\begin{aligned}
W_A^f[f(t)](t,u) &= WVD_{(a,b,c,d)}(t,u)\\
&= \int_{-\infty}^{+\infty} R_f(t,\tau)K_A(u,\tau)d\tau\\
&= \int_{-\infty}^{+\infty} f(t+\frac{\tau}{2})f^*(t-\frac{\tau}{2}))K_A(u,\tau)d\tau,
\end{aligned}
\tag{4}
$$

where $K_A(u,\tau) = \sqrt{\frac{1}{j2\pi b}}e^{j\frac{d}{2b}u^2}e^{-j\frac{u}{b}\tau}e^{j\frac{a}{2b}\tau^2}, |A| = 1$.

The AF of a signal $f(t)$ associated with the LCT (AFL) with parameter $A = \begin{pmatrix} a & b\\ c & d\end{pmatrix}$ is defined as [19]

$$
\begin{aligned}
AFL[f(t)](\tau,u) &= AF_{(a,b,c,d)}(\tau,u)\\
&= \int_{-\infty}^{+\infty} R_f(t,\tau)K_A(t,u)dt\\
&= \int_{-\infty}^{+\infty} f(t+\frac{\tau}{2})f^*(t-\frac{\tau}{2})K_A(t,u)dt,
\end{aligned}
\tag{5}
$$

where $K_A(t,u) = \sqrt{\frac{1}{j2\pi b}}e^{j\frac{d}{2b}u^2}e^{-j\frac{u}{b}t}e^{j\frac{a}{2b}t^2}$.

III. Main Results

Based on the aforementioned research of the TF analysis methods, we propose a novel kind of WVD associated with the LCT in this section.

A. Definition

Definition 1. *Suppose the kernel of the 2D LCT with parameter A, B is $K_{A,B}$, then the WVL of a signal $x(t)$ is defined as*

$$
\begin{aligned}
WVL_x^{A,B}(u,v) &= L_\Gamma^{A,B}(\Gamma[R_x^C(t,\tau)])\\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\Gamma[R_x^C(t,\tau)]K_{A,B}(t_n,\tau,u,v)dt_nd\tau\\
&= \sum_{i=0}^{K-1}WVL_{x_i}(u,v) + \sum_{i=0}^{K-2}\sum_{j=i+1}^{K-1}WVL_{x_ix_j}(u,v),
\end{aligned}
\tag{6}
$$

where $A = \begin{pmatrix} a_1 & b_1\\ c_1 & d_1\end{pmatrix}$, $B = \begin{pmatrix} a_2 & b_2\\ c_2 & d_2\end{pmatrix}$, and

$$
\begin{aligned}
&\Gamma[G(t,\tau)]\to G(\tfrac{t_n}{h(\tau+m)},\tau)\\
&R_x^C(t,\tau) = x(t+\tfrac{\tau+m}{2})x^*(t-\tfrac{\tau+m}{2}),
\end{aligned}
$$

where Γ is a scale operator and $R_x^C(t,\tau)$ is the PSIAF of $x(t)$.

B. Properties

Property 1. *Conjugation symmetry property*
Suppose the WVL of a signal $x(t)$ is denoted as $WVL_x^{A,B}(u,v)$, the WVL of $x^*(t)$ is

$$
WVL_{x^*}^{A,B}(u,v) = [WVL_x^{A^{-1},B^{-1}}(u,v)]^*.
\tag{7}
$$

Property 2. *Shifting property*
Suppose the WVL of a signal $x(t)$ is denoted as $WVL_x^{A,B}(u,v)$, the WVL of $\widetilde{x}(t) = x(t)e^{jwt}$ is

$$
\begin{aligned}
&WVL_{\widetilde{x}}^{A,B}(u,v)\\
&= e^{j(wa+d_2wv-\frac{b_2d_2w^2}{2})}WVL_x^{A,B}(u,v-wb_2).
\end{aligned}
\tag{8}
$$

C. Theorems

Consider an infinite-time multicomponent LFM signal with the constant amplitude A_i, the centroid frequency f_i and chirp rate γ_i as

$$
x(t) = \sum_{i=0}^{K-1}x_i(t) = \sum_{i=0}^{K-1}A_ie^{j2\pi f_it+j\pi\gamma_it^2},
\tag{9}
$$

where K is the number of signal components present in the signal. The PSIAF of $x(t)$ is defined by[13]

$$
\begin{aligned}
R_x^C(t,\tau) &= x(t+\frac{\tau+m}{2})x^*(t-\frac{\tau+m}{2})\\
&= \sum_{i=0}^{K-1}A_i^2e^{j2\pi f_i(\tau+m)}e^{j2\pi\gamma_i(\tau+m)t}\\
&+ \sum_{i=0}^{K-2}\sum_{j=i+1}^{K-1}(R_{x_ix_j}^C(t,\tau)+R_{x_jx_i}^C(t,\tau)),
\end{aligned}
\tag{10}
$$

where $R^C_{x_ix_j}(t,\tau) = A_iA_j e^{j\pi[\frac{\gamma_i-\gamma_j}{4}(\tau+m)^2+(f_i+f_j)(\tau+m)]}$
$\times e^{j\pi[(\gamma_i-\gamma_j)t^2+(2(f_i-f_j)+(\gamma_i+\gamma_j)(\tau+m))t]}$.
Because of the lag variable τ and time variable t couple with each other in the exponential phase terms, we use the scaling operator Γ of LVD [13] on the function $R^C_x(t,\tau)$. Then we can get $\Gamma[R^C_x(t,\tau)]$ as

$$\begin{aligned}\Gamma[R^C_x(t,\tau)] = &\sum_{i=0}^{K-1} A_i^2 e^{(j2\pi f_i(\tau+m)+j2\pi\frac{\gamma_i}{h}t_n)} \\ &+ \sum_{i=0}^{K-2}\sum_{j=i+1}^{K-1}\Gamma[R^C_{x_ix_j}(t_n,\tau)+R^C_{x_jx_i}(t_n,\tau)],\end{aligned} \tag{11}$$

where $\Gamma[R^C_{x_ix_j}(t_n,\tau)+R^C_{x_jx_i}(t_n,\tau)] = A_iA_jR^{C_1}_{x_ix_j}(t_n,\tau)$
$\times R^{C_2}_{x_ix_j}(t_n,\tau)$ and

$$\begin{aligned}&R^{C_1}_{x_ix_j}(t_n,\tau) = e^{j\pi[(\gamma_i+\gamma_j)\frac{t_n}{h}+(f_i+f_j)(\tau+m)]}, \\ &R^{C_2}_{x_ix_j}(t_n,\tau) \\ =&e^{j\pi(2(f_i-f_j)\frac{t_n}{h(\tau+m)}+(\gamma_i-\gamma_j)\frac{t_n^2}{h^2(\tau+m)^2}+\frac{\gamma_i-\gamma_j}{4}(\tau+m)^2)} \\ =&2cos(\pi[(\gamma_i-\gamma_j)\frac{t_n^2}{h^2(\tau+m)^2}+2(f_i-f_j)\frac{t_n}{h(\tau+m)} \\ &+\frac{(\tau+m)^2}{4}(\gamma_i-\gamma_j)]).\end{aligned} \tag{12}$$

As shown in [11], the scaled PSIAF of $x(t)$ can be expressed as two parts: auto-terms and cross-terms. Different from the existing methods [13], we take place of the kernel of 2D FT with the kernel of 2D LCT to get a novel kind of WVD associated with the LCT (WVL). The result is divided into two parts, auto-terms and cross-terms, leading to the following interesting theorems.

Theorem 1. *Each auto term in the WVL with parameter $a_1 = a_2 = 0$ can be expressed as*

$$\begin{aligned}WVL^{A,B}_{x_i}(u,v) = &\frac{A_i^2}{2\pi j}\sqrt{\frac{1}{b_1b_2}}e^{j(\frac{d_1u^2}{2b_1}+\frac{d_2v^2}{2b_2})}e^{j2\pi f_im} \\ &\times\delta(\frac{u}{2\pi b_1}-\frac{\gamma_i}{h})\delta(\frac{v}{2\pi b_2}-f_i).\end{aligned} \tag{13}$$

Proof: $WVL^{A,B}_{x_i}(u,v)$

$$\begin{aligned}=&\frac{A_i^2}{2\pi j}\sqrt{\frac{1}{b_1b_2}}e^{j(\frac{d_1u^2}{2b_1}+\frac{d_2v^2}{2b_2})}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}e^{j(\frac{a_1t_n^2}{2b_1}+\frac{a_2\tau^2}{2b_2})} \\ &\times e^{-j(\frac{ut_n}{b_1}+\frac{v\tau}{b_2})}e^{j2\pi f_i(\tau+m)+j2\pi\gamma_i\frac{t_n}{h}}dt_nd\tau \\ =&\frac{A_i^2}{2\pi j}\sqrt{\frac{1}{b_1b_2}}e^{j(\frac{d_1u^2}{2b_1}+\frac{d_2v^2}{2b_2})}e^{j2\pi f_0m}\delta(\frac{u}{2\pi b_1}-\frac{\gamma_i}{h})\delta(\frac{v}{2\pi b_2}-f_i).\end{aligned}$$

■

It is seen from Theorem 1 that the auto-term of WVL has the characteristics of energy accumulation.providing a solid foundation for further signal processing. Next, we explore the theorem 2 associated with the cross-terms.

Theorem 2. *Each cross term in the WVL with parameter $a_1 = a_2 = 0$, can be modeled as following cases and let $\widetilde{u} = \frac{uh}{2\pi b_1} - \frac{\gamma_i+\gamma_j}{2}, \widetilde{v} = \frac{v}{2\pi b_2} - \frac{f_i+f_j}{2}$.*

Case 1. For $\gamma_i = \gamma_j \neq \frac{uh}{2\pi b_1}, f_i \neq f_j$,

$$\begin{aligned}&WVL^{A,B}_{x_ix_j}(u,v) \\ =&\frac{A_iA_jh}{4j\pi^3|b_1b_2|}\sqrt{\frac{1}{b_1b_2}}e^{j(\frac{d_1}{2b_1}u^2+\frac{d_2}{2b_2}v^2)} \\ &\times e^{j\pi(f_i+f_j)m}e^{j2\pi m\widetilde{v}}\frac{f_i-f_j}{\widetilde{u}^2}cos(\frac{2\pi(f_i-f_j)}{\widetilde{u}}\widetilde{v}).\end{aligned} \tag{14}$$

Case 2. For $\gamma_i = \gamma_j = \frac{uh}{2\pi b_1}, f_i \neq f_j$,

$$WVL^{A,B}_{x_ix_j}(u,v) = 0. \tag{15}$$

Case 3. For $\gamma_i \neq \gamma_j,\ u = \pm|\frac{\pi(\gamma_i-\gamma_j)b_1}{h}|$,

$$\begin{aligned}&WVL^{A,B}_{x_ix_j}(u,v) \\ =&\frac{A_iA_j|h||\gamma_i-\gamma_j|^{\frac{3}{2}}}{j2\pi^3\sqrt{b_1b_2}[(\gamma_i-\gamma_j)\widetilde{v}-(f_i-f_j)\widetilde{u}]^2}e^{j2\pi m\widetilde{v}} \\ &\times e^{j\pi[\frac{(\gamma_i+\gamma_j)d_1}{h}u+(f_i+f_j)d_2v]}cos(\frac{\pi^2b_1d_1}{h^2}[2\widetilde{u}^2+\frac{(\gamma_i+\gamma_j)^2}{2}] \\ &+\pi^2b_2d_2[2\widetilde{v}^2+\frac{(f_i+f_j)^2}{2}]-\frac{\pi(f_i-f_j)^2}{\gamma_i-\gamma_j}-sgn(\frac{\gamma_i-\gamma_j}{2})\frac{\pi}{4}).\end{aligned} \tag{16}$$

Case 4. For $\gamma_i \neq \gamma_j, u \neq \pm|\frac{\pi(\gamma_i-\gamma_j)b_1}{h}|, \widetilde{u} \neq \pm|\frac{(\gamma_i-\gamma_j)}{2}|$,

$$\begin{aligned}&WVL^{A,B}_{x_ix_j}(u,v) \\ =&-\frac{8A_iA_jh[(\gamma_i-\gamma_j)\widetilde{v}-(f_i-f_j)\widetilde{u}]}{\pi\sqrt{b_1b_2}[(\gamma_i-\gamma_j)^2-4\widetilde{u}^2]^{\frac{3}{2}}}e^{j\pi[\frac{(\gamma_i+\gamma_j)d_1}{h}u+(f_i+f_j)d_2v]} \\ &\times sin(\frac{\pi^2b_1d_1}{h^2}[\frac{(\gamma_i+\gamma_j)^2}{2}-2\widetilde{u}^2]+\pi^2b_2d_2[\frac{(f_i+f_j)^2}{2}-2\widetilde{v}^2] \\ &+\pi\frac{(f_i-f_j)^2}{\gamma_i-\gamma_j}+\frac{4\pi[(\gamma_i-\gamma_j)\widetilde{v}-(f_i-f_j)\widetilde{u}]^2}{(\gamma_i-\gamma_j)[(\gamma_i-\gamma_j)^2-4\widetilde{u}^2]} \\ &-\frac{\pi}{4}sgn(\frac{\gamma_i-\gamma_j}{2})(1+sgn(\pi^2b_1^2[(\gamma_i-\gamma_j)-4\widetilde{u}^2]))).\end{aligned} \tag{17}$$

Case 5. For $\gamma_i \neq \gamma_j, u \neq \pm|\frac{\pi(\gamma_i-\gamma_j)b_1}{h}|, \widetilde{u} = \pm|\frac{(\gamma_i-\gamma_j)}{2}|$,

$$WVL^{A,B}_{x_ix_j}(u,v) = 0. \tag{18}$$

Proof: For $\gamma_i = \gamma_j$, using (2), we have

$$\begin{aligned}&WVL^{A,B}_{x_ix_j}(u,v) \\ &= L^{t_n,\tau}\{2A_iA_jR^{C_1}_{x_ix_j}(t_n,\tau)cos(\pi[2(f_i-f_j)\frac{t_n}{h(\tau+m)}])\} \\ &= \frac{1}{4\pi^2}\frac{1}{|b_1b_2|}e^{j(\frac{d_1}{2b_1}u^2+\frac{d_2}{2b_2}v^2)}[L^{t_n,\tau}\{2A_iA_jR^{C_1}_{x_ix_j}(t_n,\tau)\} \\ &\times e^{-j(\frac{d_1}{2b_1}u^2+\frac{d_2}{2b_2}v^2)}] * F^{t_n,\tau}(cos(\pi[2(f_i-f_j)\frac{t_n}{h(\tau+m)}])).\end{aligned} \tag{19}$$

First, taking 2D LCT into $2A_iA_jR^{C_1}_{x_ix_j}(t_n,\tau)$, we have

$$\begin{aligned}&L^{t_n,\tau}\{2A_iA_jR^{C_1}_{x_ix_j}(t_n,\tau)\} \\ =&\frac{A_iA_j}{\pi j}\sqrt{\frac{1}{b_1b_2}}e^{j[\frac{d_1u^2}{2b_1}+\frac{d_2v^2}{2b_2}]+j\pi(f_i+f_j)m} \\ &\times\delta(\frac{u}{2\pi b_1}-\frac{\gamma_i}{h})\delta(\frac{v}{2\pi b_2}-\frac{f_i+f_j}{2}).\end{aligned} \tag{20}$$

Second, taking 2D LCT into $cos(\pi[2(f_i - f_j)\frac{t_n}{h(\tau+m)}])$, we have

$$\begin{aligned}&F^{t_n,\tau}(cos(\pi[2(f_i - f_j)\frac{t_n}{h(\tau+m)}]))\\&=\begin{cases}\frac{f_i-f_j}{hu^2}e^{j2\pi mv}cos(\frac{2\pi(f_i-f_j)}{hu}v), f_i \neq f_j,\\ \delta(v)\delta(u), f_i = f_j.\end{cases}\end{aligned} \tag{21}$$

Then take (20) and (21) into (19). Because of $\widetilde{u} = \frac{uh}{2\pi b_1} - \frac{\gamma_i+\gamma_j}{2}, \widetilde{v} = \frac{v}{2\pi b_2} - \frac{f_i+f_j}{2}$, we can obtain (14) and (15).

For $\gamma_i \neq \gamma_j$, taking a LCT of $R^{C_2}_{x_ix_j}(t_n,\tau)$ with respect to t_n, we have

$$\begin{aligned}&L^{t_n}(R^{C_2}_{x_ix_j}(t_n,\tau))\\&=\sqrt{\frac{1}{j2\pi b}}e^{j\frac{d}{2b}u^2}e^{j\frac{\pi(\gamma_i-\gamma_j)}{4}(\tau+m)^2}\\&\times\int_{-\infty}^{+\infty}e^{j(\frac{\pi(\gamma_i-\gamma_j)}{h^2(\tau+m)^2}t_n^2+[\frac{2\pi(f_i-f_j)}{h(\tau+m)}-\frac{u}{b}]t_n)}dt_n.\end{aligned} \tag{22}$$

Using the principle of stationary phase (PSP) [20], first we let

$$\frac{d}{dt_n}(\text{phase}(R^{C_2}_{x_ix_j}(t_n,\tau)) - 2\pi\gamma t_n) = 0, \tag{23}$$

we have

$$t_n^* = (\frac{u}{b} - \frac{2\pi(f_i-f_j)}{h(\tau+m)})\frac{h^2(\tau+m)^2}{2\pi(\gamma_i-\gamma_j)}. \tag{24}$$

Second

$$\frac{d^2}{dt_n^2}(\text{phase}(R^{C_2}_{x_ix_j}(t_n^*,\tau)) - 2\pi\frac{u}{b}t_n^*) = \frac{2\pi(\gamma_i-\gamma_j)}{h^2(\tau+m)^2}. \tag{25}$$

Then

$$\begin{aligned}&L^{t_n}(R^{C_2}_{x_ix_j}(t_n,\tau))\\&=\sqrt{\frac{1}{j2\pi b}}e^{j\frac{d}{2b}u^2}\frac{|h(\tau+m)|}{\sqrt{\gamma_i-\gamma_j}}e^{jsgn(\frac{\gamma_i-\gamma_j}{2})\frac{\pi}{4}}\\&\times e^{j[-\frac{h^2(\tau+m)^2}{4\pi(\gamma_i-\gamma_j)b^2}u^2+\frac{h(\tau+m)(f_i-f_j)}{(\gamma_i-\gamma_j)b}u-\frac{\pi(f_i-f_j)^2}{\gamma_i-\gamma_j}+\frac{\pi(\gamma_i-\gamma_j)}{4}(\tau+m)^2]}.\end{aligned} \tag{26}$$

For $u = \pm|\frac{\pi(\gamma_i-\gamma_j)b_1}{h}|$, applying $\mathcal{F}(|(\tau+m)|) = -\frac{1}{2\pi^2f^2}$ [3], we have

$$\begin{aligned}&WVL^{'}_{x_ix_j}(u,v) = L^{\tau}(L^{t_n}(R^{C_2}_{x_ix_j}(t_n,\tau)))\\&=\frac{1}{j2\pi}\frac{1}{\sqrt{b_1b_2}}\frac{|h|}{\sqrt{|\gamma_i-\gamma_j|}}e^{-j(\frac{d_1}{2b_1}u^2+\frac{d_2}{2b_2}v^2)}e^{j\frac{\pi(f_i-f_j)^2}{\gamma_i-\gamma_j}}e^{jsgn(\frac{\gamma_i-\gamma_j}{2})\frac{\pi}{4}}\\&\times e^{j\frac{v}{b_2}m}\int_{-\infty}^{+\infty}|(\tau+m)|e^{-j2\pi(\frac{v}{2\pi b_2}-\frac{h(f_i-f_j)u}{2\pi(\gamma_i-\gamma_j)b_1})(\tau+m)]}d(\tau+m)\\&=\frac{1}{j2\pi}\frac{1}{\sqrt{b_1b_2}}e^{-j(\frac{d_1}{2b_1}u^2+\frac{d_2}{2b_2}v^2)}e^{j\frac{\pi(f_i-f_j)^2}{\gamma_i-\gamma_j}}e^{jsgn(\frac{\gamma_i-\gamma_j}{2})\frac{\pi}{4}}e^{j\frac{v}{b_2}m}\\&\times\frac{|h|}{\sqrt{|\gamma_i-\gamma_j|}}\frac{2}{(\frac{v}{b_2}-\frac{h(f_i-f_j)u}{(\gamma_i-\gamma_j)b_1})^2}.\end{aligned} \tag{27}$$

Denote * by the complex conjugate. Using the conjugation property of the LCT, we obtain

$$\begin{aligned}&L^{\tau}(L^{t_n}(R^{C_2^*}_{x_ix_j}(t_n,\tau))) = L^{\tau}(L^{t_n^*}(\tau,-u))\\&=WVL^{'*}_{x_ix_j}(-v,-u).\end{aligned} \tag{28}$$

Using (3), we have

$$\begin{aligned}&L^{t_n,\tau}(R^{C_1}_{x_ix_j}(t_n,\tau)R^{C_2}_{x_ix_j}(t_n,\tau))\\&=e^{j(u\frac{\pi(\gamma_i+\gamma_j)}{h}d_1+v\pi(f_i+f_j)d_2)-\frac{j}{2}[(\frac{\pi(\gamma_i+\gamma_j)}{h})^2b_1d_1+\pi^2(f_i+f_j)^2b_2d_2]}\\&\times L^{t_n,\tau}(R^{C_2}_{x_ix_j}(u-\frac{\pi(\gamma_i+\gamma_j)}{h}b_1, v-\pi(f_i+f_j)b_2)).\end{aligned} \tag{29}$$

Obviously

$$\begin{aligned}&L^{t_n,\tau}_{\Gamma}(A_iA_jR^{C_1}_{x_ix_j}(t_n,\tau)[R^{C_2}_{x_ix_j}(t_n,\tau)+R^{C_2^*}_{x_ix_j}(t_n,\tau)])\\&=A_iA_j[L^{t_n,\tau}(R^{C_1}_{x_ix_j}(t_n,\tau)R^{C_2}_{x_ix_j}(t_n,\tau))\\&+L^{t_n,\tau}(R^{C_1}_{x_ix_j}(t_n,\tau)R^{C_2^*}_{x_ix_j}(t_n,\tau))],\end{aligned} \tag{30}$$

substituting $\widetilde{u}$ and $\widetilde{v}$, we have

$$\begin{aligned}&WVL^{A,B}_{x_ix_j}(u,v)\\&=\frac{A_iA_j|h|}{j2\pi^3\sqrt{b_1b_2|(\gamma_i-\gamma_j)|}}\frac{1}{(\widetilde{v}-\frac{(f_i-f_j)\widetilde{u}}{\gamma_i-\gamma_j})^2}e^{j2\pi m\widetilde{v}}\\&\times e^{j\pi[\frac{(\gamma_i+\gamma_j)d_1}{h}u+(f_i+f_j)d_2v]}cos(\frac{2\pi^2b_1d_1}{h^2}\widetilde{u}^2+2\pi^2b_2d_2\widetilde{v}^2\\&-\frac{\pi(f_i-f_j)^2}{\gamma_i-\gamma_j}-sgn(\frac{\gamma_i-\gamma_j}{2})\frac{\pi}{4}\\&+\frac{\pi^2}{2}[\frac{(\gamma_i+\gamma_j)^2b_1d_1}{h^2}+(f_i+f_j)^2b_2d_2]).\end{aligned} \tag{31}$$

Simplify (31) we can obtain (16).

For $u \neq \pm|\frac{\pi(\gamma_i-\gamma_j)b_1}{h}|$, we have

$$\begin{aligned}&L^{\tau}(L^{t_n}(R^{C_2}_{x_ix_j}(t_n,\tau)))\\&=\frac{1}{j2\pi}\frac{1}{\sqrt{b_1b_2}}\frac{|h|}{\sqrt{|\gamma_i-\gamma_j|}}e^{-j(\frac{d_1}{2b_1}u^2+\frac{d_2}{2b_2}v^2)}\\&\times e^{-j\frac{\pi(f_i-f_j)^2}{\gamma_i-\gamma_j}}e^{jsgn(\frac{\gamma_i-\gamma_j}{2})\frac{\pi}{4}}\int_{-\infty}^{+\infty}|(\tau+m)|\\&\times e^{j[(\frac{\pi(\gamma_i-\gamma_j)}{4}-\frac{h^2u^2}{4\pi(\gamma_i-\gamma_j)b_1^2})(\tau+m)^2+(\frac{h(f_i-f_j)}{b_1(\gamma_i-\gamma_j)}-\frac{v}{b_2})(\tau+m)]}d(\tau+m).\end{aligned} \tag{32}$$

Similarly, using the PSP, we have

$$(\tau+m)^* = \frac{2\pi((\gamma_i-\gamma_j)b_1^2v-h(f_i-f_j)b_1b_2)}{b_2(\pi^2(\gamma_i-\gamma_j)^2b_1^2-h^2u^2)}, \tag{33}$$

then

$$\begin{aligned}&WVL^{'}_{x_ix_j}(u,v)\\&=\frac{2\pi b_1^{\frac{3}{2}}h[b_1(\gamma_i-\gamma_j)v-h(f_i-f_j)b_2u]}{jb_2^{\frac{3}{2}}[\pi^2(\gamma_i-\gamma_j)^2b_1^2-h^2u^2]^{\frac{3}{2}}}\\&\times e^{j(\frac{d_1}{2b_1}u^2+\frac{d_2}{2b_2}v^2)}e^{-j\pi\frac{(b_1(\gamma_i-\gamma_j)v-h(f_i-f_j)b_2u)^2}{b_2^2(\gamma_i-\gamma_j)(\pi^2(\gamma_i-\gamma_j)^2b_1^2-h^2u^2)}}\\&\times e^{j\frac{\pi}{4}sgn(\frac{\gamma_i-\gamma_j}{2})(1+sgn(\pi^2b_1^2(\gamma_i-\gamma_j)-h^2u^2))}e^{-j\frac{\pi(f_i-f_j)^2}{\gamma_i-\gamma_j}}.\end{aligned} \tag{34}$$

(17) and (18) are derived by substituting (34) into (29) and then applying the (28) to calculate (30). This completes the proof. ■

Theorem 2 shows that WVL can eliminate the effect of some cross terms. When $\gamma_i = \gamma_j = \frac{uh}{2\pi b_1}, f_i \neq f_j$ and $\gamma_i \neq \gamma_j, u \neq \pm|\frac{\pi(\gamma_i-\gamma_j)b_1}{h}|, \widetilde{u} = \pm|\frac{(\gamma_i-\gamma_j)}{2}|$, it is shown from (15) or (18) that the corresponding cross term turns out to be zero.

IV. Simulation

This section proforms the simulation to verify the derived results, and demonstrates the benefits of the WVL in multi-component LFM signal detection.

Firstly, consider a multicomponent LFM signal $x(t)$, which has three components, and the chirp rates are $\gamma_0 = 10, \gamma_1 = 10, \gamma_2 = -20$ respectively, the centroid frequencies are $f_0 = -15, f_1 = 5, f_2 = 5$ respectively. $A_0 = A_1 = A_2 = 1$, then

$$x(t) = e^{j2\pi(-15)t+j\pi 10t^2} + e^{j2\pi 5t+j\pi 10t^2} + e^{j2\pi 5t+j\pi(-20)t^2}. \tag{35}$$

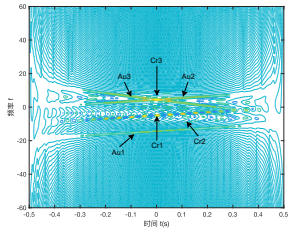


Fig. 1. Classical WVD

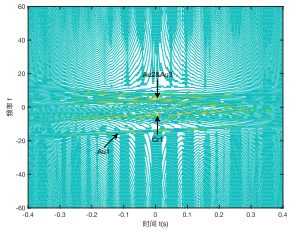


Fig. 2. LCT for the scaled PSIAF

Fig. 1 shows the result of the classical WVD of $x(t)$, Fig. 2 shows the result of the LCT with parameter [0,1/2,2,0] for the scaled PSIAF of $x(t)$. As can be seen from the Fig. 1 and Fig. 2, the instantaneous frequency of each component of the LFM signal varies with time, and the chirp frequency is shown by the slope, and the central frequency is the frequency value at $t = 0$. After taking the scaling operation for the PSIAF, the slope of the auto-terms become zero. Because of $\gamma_0 = \gamma_1$, the Cr1 still exists, however since $\gamma_0 \neq \gamma_2$ and $\gamma_1 \neq \gamma_2$, Cr2 and Cr3 oscillate according to the cosine function respectively, as can be clearly seen from formulas (12).

Secondly, by taking LCT with parameter [0,1/2,2,0] of the Fig. 2 with repect to τ, we can obtain the WVL's result, which is depicted in Fig. 3. When applying the WVL to the multicomponent LFM signal $x(t)$, it can be observed that the auto-terms are delta functions, indicating that WVL exhibits the characteristic of auto-terms energy accumulation. The energy of the cross-terms is negligible compared to the peaks of the auto-terms.

V. Conclution

Combining the advantages of the WVD and the LCT in the TF domain, this paper proposes the WVL and explores its conjugation symmetry and shifting properties. Especially in processing multicomponent LFM signals, WVL has the advantages of concentrating auto-term energy and eliminating some cross-terms. First, by scaling the PSIAF, the coupling between time variable and lag variable is eliminated. Then, by performing a 2D LCT on t and τ, each auto-term component

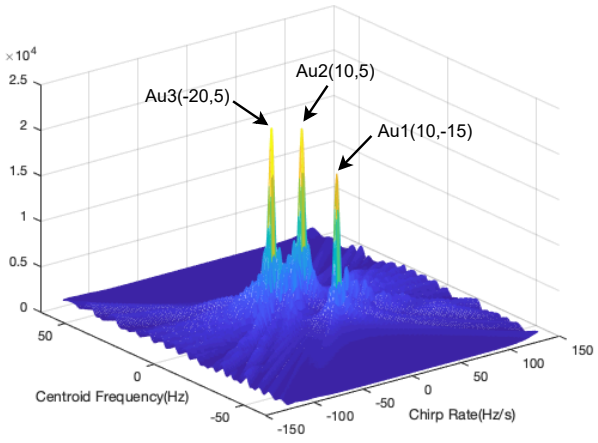


Fig. 3. WVL of multicomponent LFM signal

of the LFM signal can be easily identified in the WVL plane through peak detection. Additionally, the frequency parameters can be directly extracted from the coordinate values, and simulations have verified these conclusions. Therefore, the proposed WVL in this paper has important role for detecting and processing multicomponent LFM signals.

Acknowledgment

This work was supported by grants from the Natural Science Foundation of Beijing Municipality [No. 4242011]; the National Natural Science Foundation of China [No. 62171041].

References

[1] R. Dhawan, R. Parihar, and A. Choudhary, "Multiband dual-and cross-LFM waveform generation using a dual-drive Mach–Zehnder modulator," *Optics Communications*, vol. 530, p. 129 144, 2023.

[2] K. B. Cui, W. W. Wu, J. J. Huang, X. Chen, and N. C. Yuan, "DOA estimation of LFM signals based on stft and multiple invariance esprit," *AEU-International Journal of Electronics and Communications*, vol. 77, pp. 10–17, 2017.

[3] Y. S. Wu and X. K. Li, "Elimination of cross-terms in the Wigner–Ville distribution of multi-component LFM signals," *IET Signal Processing*, vol. 11, no. 6, pp. 657–662, 2017.

[4] J. Y. Chen and B. Z. Li, "The short-time wigner–ville distribution," *Signal Processing*, vol. 219, p. 109 402, 2024.

[5] Y. Guo, B. Z. Li, and L. D. Yang, "Novel fractional wavelet transform: Principles, MRA and application," *Digital Signal Processing*, vol. 110, p. 102 937, 2021.

[6] H. W. Wang, X. Y. Fan, Y. Chen, and Y. Z. Yang, "Wigner-Hough transform based on slice's entropy and its application to multi-LFM signal detection," *Journal of Systems Engineering and Electronics*, vol. 28, no. 4, pp. 634–642, 2017.

[7] J. F. Liu, Y. Wei, X. Y. Zeng, J. Lu, S. X. Zhang, and M. J. Wang, "A novel joint timing/frequency synchronization scheme based on Radon–Wigner transform of lfm signals in co-ofdm systems," *Optics Communications*, vol. 410, pp. 744–750, 2018.

[8] O. Aldimashki and A. Serbes, "LFM signal parameter estimation in the fractional fourier domains: Analytical models and a high-performance algorithm," *Signal Processing*, vol. 214, p. 109 224, 2024.

[9] R. Tao, B. Deng, and Y. Wang, "Fractional Fourier Transform and its Application (Second edition)," 2022.

[10] A. Y. Wu *et al.*, "A computationally efficient optimal Wigner distribution in LCT domains for detecting noisy LFM signals," *Mathematical Problems in Engineering*, vol. 2022, no. 1, p. 2 036 285, 2022.

[11] Y. Guo and L. D. Yang, "LFM signal optimization time-fractional-frequency analysis: Principles, method and application," *Digital Signal Processing*, vol. 126, p. 103 505, 2022.

[12] X. L. Lv, M. D. Xing, S. H. Zhang, and Z. Bao, "Keystone transformation of the Wigner–Ville distribution for analysis of multicomponent LFM signals," *Signal Processing*, vol. 89, no. 5, pp. 791–806, 2009.

[13] X. L. Lv, G. A. Bi, C. R. Wan, and M. D. Xing, "Lv's distribution: Principle, implementation, properties, and performance," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3576–3591, 2011.

[14] H. C. Xin and B. Z. Li, "On a new Wigner-Ville distribution associated with linear canonical transform," *EURASIP Journal on Advances in Signal Processing*, vol. 2021, no. 1, p. 56, 2021.

[15] J. Y. Chen and B. Z. Li, "Wigner distribution associated with linear canonical transform of generalized 2-d analytic signals," *Digital Signal Processing*, vol. 149, p. 104 481, 2024.

[16] J. J. Ding and S. C. Pei, "Eigenfunctions and self-imaging phenomena of the two-dimensional nonseparable linear canonical transform," *JOSA A*, vol. 28, no. 2, pp. 82–95, 2011.

[17] F. Qiang and B. Z. Li, "Convolution and correlation theorems for the two-dimensional linear canonical transform and its applications," *IET Signal Processing*, vol. 10, no. 2, pp. 125–132, 2016.

[18] R. F. Bai, B. Z. Li, and Q. Y. Cheng, "Wigner-ville distribution associated with the linear canonical transform," *Journal of Applied Mathematics*, vol. 2012, no. 1, p. 740 161, 2012.

[19] T. W. Che, B. Z. Li, and T. Z. Xu, "The ambiguity function associated with the linear canonical transform," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 1–14, 2012.

[20] A. X. Jiang and B. S. Wu, "A novel parameter estimation for hyperbolic frequency modulated signals using group delay," *Digital Signal Processing*, vol. 116, pp. 103–114, 2021.

2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

New Perspectives and Insights on Distortionless Microphone Array Beamforming

Fan Zhang*, Jacob Benesty†, Chao Pan*, and Jingdong Chen*,
* Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an, China
† INRS-EMT, University of Quebec, Canada

***Abstract*—This paper studies distortionless beamforming within a general signal model, where a target source is accompanied by multiple interference sources and ambient noise. In this context, the minimum variance distortionless response (MVDR) and linearly constrained minimum variance (LCMV) beamformers are commonly used techniques. However, the MVDR beamformer lacks control over residual interference sources, while the LCMV beamformer may result in insufficient noise attenuation and, in some cases, even amplify ambient noise at its output. To provide a flexible mechanism for balancing ambient noise reduction and interference suppression, we introduce a novel optimization criterion, which leads to the development of a general distortionless beamformer that includes both the MVDR and LCMV methods as special cases. The effectiveness of our proposed approach is demonstrated through various performance metrics and beampatterns.**

I. Introduction

In real-world environments, the signal recorded by a microphone often consists of the desired signal from the target source, along with one or more interference signals and pervasive ambient noise. To extract the desired signal and effectively eliminate the unwanted components, noise reduction techniques are commonly used. Among these techniques, beamforming with a microphone array, a method of spatial filtering, is frequently used due to its ability to reduce noise while minimizing distortion of the desired signal [1, 2].

Beamforming methods can be broadly classified based on whether they distort the desired signal. In the category where distortion of the desired signal occurs, representative algorithms include the Wiener beamformer (also known as the multichannel Wiener filter) [3], along with its parameterized variants [4–7], among others. This paper focuses on the category of distortionless beamforming, which aims to preserve the desired signal while suppressing unwanted components. Key algorithms in this category include fixed beamformers, such as the widely used delay-and-sum [3] and superdirective [8] beamformers, as well as adaptive methods like the minimum variance distortionless response (MVDR) beamformer [9] and the linearly constrained minimum variance (LCMV) beamformer [10]. The MVDR beamformer is designed to minimize the variance of the output interference-plus-noise while maintaining a distortionless constraint [11]. However, it often leaves some residual interference, especially in high ambient noise conditions. On the other hand, the LCMV beamformer aims to completely eliminate the effect of interference sources by imposing additional linear constraints and minimizing the residual noise variance [11–15]. Nonetheless, this approach may lead to less effective noise attenuation and, in some cases, amplify the ambient noise due to a reduced degree of freedom [16].

Several methods have been developed in the literature to balance background noise reduction and interference attenuation [16–19]. However, many of these methods focus primarily on scenarios involving a single interference source. For instance, the work in [16] introduced a parameterized distortionless beamformer, which is expressed as a weighted sum of the LCMV beamformer and a matched beamformer (an MVDR solution in the absence of interference). This approach allows for a tradeoff between the two functions by adjusting the weighting coefficient. In [17], an LCMV-like beamformer with specific constraints was developed to balance target signal distortion, noise reduction, and interference suppression. More recently, [19] presented a flexible beamformer that adjusts an "interference attenuation factor" to achieve a balance between interference suppression and ambient noise reduction. For scenarios involving multiple interference sources, the work [18] proposed a controllable LCMV beamformer with a variable number of constraints. This method separates interference sources into two groups: one group is completely nulled out while the other is suppressed along with the ambient noise. While this approach partially balances noise reduction and interference rejection, it does not offer full flexibility for individually controlling each interference source.

In this paper, we consider a general signal model where a desired signal source is present alongside multiple interference sources and ambient noise. We explore the problem of distortionless beamforming within this model and define key performance metrics clearly. Specifically, we introduce a novel optimization criterion featuring multiple design parameters, with each parameter associated with a particular interference source. Using this criterion, we develop a general distortionless beamformer that includes both the MVDR and LCMV beamformers as special cases. By adjusting the design parameters individually, our approach allows for varying levels of tradeoff between ambient noise reduction and interference suppression. The effectiveness and flexibility of this approach are demonstrated through simulations involving one target source and two interference sources.

II. Signal Model and Problem Formulation

Consider a two-dimensional (2-D) sensor array consisting of $M \geq 3$ omnidirectional microphones arranged on a 2-D plane,

This work was supported in part by the NSFC (No. 62171373).

which may or may not form a regular geometric pattern[1]. The positions of all microphones are assumed to be known precisely, and all sources (both desired and interference) lie on the same plane as the 2-D array. Angles on this plane, measured from the horizontal axis to a point, are referred to as azimuth angles (θ). We assume there is a single desired signal arriving at the array from the direction θ_0, while there are N interference sources coming from directions θ_n, where $n = 1, 2, \ldots, N$ and $\theta_0 \neq \theta_1 \neq \cdots \neq \theta_N$. Consequently, the frequency-domain observed signal vector of length M at temporal frequency f is given by [2]

$$\begin{aligned}\mathbf{y}(f) &= \begin{bmatrix} Y_1(f) & Y_2(f) & \cdots & Y_M(f) \end{bmatrix}^T \\ &= \mathbf{d}_{\theta_0}(f)X_0(f) + \sum_{n=1}^{N} \mathbf{d}_{\theta_n}(f)X_n(f) + \mathbf{v}(f), \end{aligned} \quad (1)$$

where $Y_m(f)$ is the mth ($m = 1, 2, \ldots, M$) microphone signal of the 2-D array, $\mathbf{d}_{\theta_0}(f)$ is the signal propagation vector at θ_0 (direction of the desired source), $X_0(f)$ is the (zero-mean) desired signal, $\mathbf{d}_{\theta_n}(f)$ is the signal propagation vector at θ_n (direction of the nth interference source), $X_n(f)$ is the (zero-mean) nth interference signal, $\mathbf{v}(f)$ is the (zero-mean) background noise signal vector defined similarly to $\mathbf{y}(f)$, and the superscript ^{T} is the transpose operator. We assume that all sources are mutually independent, and incoherent with $\mathbf{v}(f)$. In the rest, in order to simplify the notation, we drop the dependence on the frequency, f. We deduce that the covariance matrix of $\mathbf{y}$ is

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}} &= E\left(\mathbf{y}\mathbf{y}^H\right) \\ &= \phi_{X_0}\mathbf{d}_{\theta_0}\mathbf{d}_{\theta_0}^H + \sum_{n=1}^{N} \phi_{X_n}\mathbf{d}_{\theta_n}\mathbf{d}_{\theta_n}^H + \mathbf{\Phi}_{\mathbf{v}}, \end{aligned} \quad (2)$$

where $E(\cdot)$ denotes mathematical expectation, the superscript ^{H} is the conjugate-transpose operator, $\phi_{X_0} = E\left(|X_0|^2\right)$ is the variance of X_0, $\phi_{X_n} = E\left(|X_n|^2\right)$ is the variance of X_n, and $\mathbf{\Phi}_{\mathbf{v}} = E\left(\mathbf{v}\mathbf{v}^H\right)$ is the covariance matrix of $\mathbf{v}$. For a small and compact 2-D array, we can assume that the noise variance is uniform across all sensors, i.e., $\phi_V = \phi_{V_1} = \phi_{V_2} = \cdots = \phi_{V_M}$, with $\phi_{V_m} = E\left(|V_m|^2\right)$, $m = 1, 2, \ldots, M$; in this case, we can express (2) as

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}} &= \phi_{X_0}\mathbf{d}_{\theta_0}\mathbf{d}_{\theta_0}^H + \sum_{n=1}^{N} \phi_{X_n}\mathbf{d}_{\theta_n}\mathbf{d}_{\theta_n}^H + \phi_V\mathbf{\Gamma}_{\mathbf{v}} \\ &= \phi_{X_0}\mathbf{d}_{\theta_0}\mathbf{d}_{\theta_0}^H + \mathbf{D}\mathbf{\Phi}\mathbf{D}^H + \phi_V\mathbf{\Gamma}_{\mathbf{v}}, \end{aligned} \quad (3)$$

where $\mathbf{\Gamma}_{\mathbf{v}} = \mathbf{\Phi}_{\mathbf{v}}/\phi_V$ is the coherence matrix of the noise,

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_{\theta_1} & \mathbf{d}_{\theta_2} & \cdots & \mathbf{d}_{\theta_N} \end{bmatrix}$$

is a rectangular matrix of size $M \times N$, and

$$\mathbf{\Phi} = \mathrm{diag}\left(\phi_{X_1}, \phi_{X_2}, \ldots, \phi_{X_N}\right)$$

is a diagonal matrix of size $N \times N$. At this point, we do not specify the relationship between N and M; this will be discussed later.

From (3), we can determine the following measures: the input signal-to-noise ratio (SNR), input signal-to-interference-plus-noise ratio (SINR), input signal-to-interference ratio (SIR), and input noise-to-interference ratio (NIR). They are given by

$$\mathrm{iSNR} = \frac{\phi_{X_0}}{\phi_V}, \quad (4)$$

$$\mathrm{iSINR} = \frac{\phi_{X_0}}{\sum_{n=1}^{N}\phi_{X_n} + \phi_V}, \quad (5)$$

$$\mathrm{iSIR} = \frac{\phi_{X_0}}{\sum_{n=1}^{N}\phi_{X_n}}, \quad (6)$$

$$\mathrm{iNIR} = \frac{\phi_V}{\sum_{n=1}^{N}\phi_{X_n}}. \quad (7)$$

As a result, we have

$$\mathrm{iSINR} = \frac{\mathrm{iSIR}}{1 + \mathrm{iNIR}} = \frac{\mathrm{iSNR}}{1 + \mathrm{iNIR}^{-1}}$$

and

$$\mathrm{iSIR} = \mathrm{iNIR} \times \mathrm{iSNR}.$$

The objective of this paper is to give another perspective on distortionless beamforming based on the general signal model given in (1).

III. Beamforming and Performance Measures

Let $\mathbf{h}$ be a complex-valued linear filter of length M. The most effective method for performing beamforming involves applying $\mathbf{h}$ to the observed signal vector, $\mathbf{y}$, i.e.,

$$\begin{aligned} Z &= \mathbf{h}^H\mathbf{y} \\ &= X_{\mathrm{fd}} + X_{\mathrm{ri}} + V_{\mathrm{rn}}, \end{aligned} \quad (8)$$

where Z is the beamformer output, also an estimate of X_0, $X_{\mathrm{fd}} = X_0\mathbf{h}^H\mathbf{d}_{\theta_0}$ is the filtered desired signal, $X_{\mathrm{ri}} = \sum_{n=1}^{N} X_n\mathbf{h}^H\mathbf{d}_{\theta_n}$ is the residual interference, and $V_{\mathrm{rn}} = \mathbf{h}^H\mathbf{v}$ is the residual noise. Then, the variance of Z is

$$\begin{aligned} \phi_Z &= \mathbf{h}^H\mathbf{\Phi}_{\mathbf{y}}\mathbf{h} \\ &= \phi_{X_{\mathrm{fd}}} + \phi_{X_{\mathrm{ri}}} + \phi_{V_{\mathrm{rn}}}, \end{aligned} \quad (9)$$

where $\phi_{X_{\mathrm{fd}}} = \phi_{X_0}\left|\mathbf{h}^H\mathbf{d}_{\theta_0}\right|^2$ is the variance of the filtered desired signal, $\phi_{X_{\mathrm{ri}}} = \sum_{n=1}^{N}\phi_{X_n}\left|\mathbf{h}^H\mathbf{d}_{\theta_n}\right|^2 = \mathbf{h}^H\mathbf{D}\mathbf{\Phi}\mathbf{D}^H\mathbf{h}$ is the variance of the residual interference, and $\phi_{V_{\mathrm{rn}}} = \mathbf{h}^H\mathbf{\Phi}_{\mathbf{v}}\mathbf{h}$ is the variance of the residual noise. Also, from (8), we clearly see that the distortionless constraint is

$$\mathbf{h}^H\mathbf{d}_{\theta_0} = 1. \quad (10)$$

We emphasize that this study focuses solely on distortionless beamforming; therefore, (10) must always be satisfied. Now, from (9), we find that the output SNR is

$$\mathrm{oSNR}\left(\mathbf{h}\right) = \frac{\phi_{X_{\mathrm{fd}}}}{\phi_{V_{\mathrm{rn}}}} = \frac{\phi_{X_0}\left|\mathbf{h}^H\mathbf{d}_{\theta_0}\right|^2}{\mathbf{h}^H\mathbf{\Phi}_{\mathbf{v}}\mathbf{h}}. \quad (11)$$

[1]Extending the method developed in this paper to three-dimensional arrays is straightforward.

In the same way, the output SINR is

$$\mathrm{oSINR}\left(\mathbf{h}\right)=\frac{\phi_{X_{\mathrm{fd}}}}{\phi_{X_{\mathrm{ri}}}+\phi_{V_{\mathrm{rn}}}}=\frac{\phi_{X_0}\left|\mathbf{h}^H\mathbf{d}_{\theta_0}\right|^2}{\mathbf{h}^H\mathbf{D}\boldsymbol{\Phi}\mathbf{D}^H\mathbf{h}+\mathbf{h}^H\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{h}}. \quad (12)$$

We also deduce that the output SIR and output NIR are, respectively,

$$\mathrm{oSIR}\left(\mathbf{h}\right)=\frac{\phi_{X_0}\left|\mathbf{h}^H\mathbf{d}_{\theta_0}\right|^2}{\mathbf{h}^H\mathbf{D}\boldsymbol{\Phi}\mathbf{D}^H\mathbf{h}} \quad (13)$$

and

$$\mathrm{oNIR}\left(\mathbf{h}\right)=\frac{\mathbf{h}^H\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{h}}{\mathbf{h}^H\mathbf{D}\boldsymbol{\Phi}\mathbf{D}^H\mathbf{h}}. \quad (14)$$

We also have the following relationships:

$$\mathrm{oSINR}\left(\mathbf{h}\right)=\frac{\mathrm{oSIR}\left(\mathbf{h}\right)}{1+\mathrm{oNIR}\left(\mathbf{h}\right)}=\frac{\mathrm{oSNR}\left(\mathbf{h}\right)}{1+\mathrm{oNIR}^{-1}\left(\mathbf{h}\right)}$$

and

$$\mathrm{oSIR}\left(\mathbf{h}\right)=\mathrm{oNIR}\left(\mathbf{h}\right)\times\mathrm{oSNR}\left(\mathbf{h}\right).$$

Ideally, we aim to determine $\mathbf{h}$ such that $\mathrm{oSNR}\left(\mathbf{h}\right)\geq\mathrm{iSNR}$ and $\mathrm{oSIR}\left(\mathbf{h}\right)\geq\mathrm{iSIR}$, while achieving an optimal balance between output SNR and output SIR.

IV. Proposed Criterion and Beamformer

Let

$$\mathbf{Q}=\mathrm{diag}\left(Q_1,Q_2,\ldots,Q_N\right) \quad (15)$$

be a diagonal matrix of size $N\times N$, where $Q_n\geq 0,\ n=1,2,\ldots,N$. The proposed method for general distortionless beamforming is reformulated as the following constrained optimization problem:

$$\min_{\mathbf{h}}\mathbf{h}^H\left(\boldsymbol{\Phi}_{\mathbf{v}}+\mathbf{D}\boldsymbol{\Phi}\mathbf{Q}\mathbf{D}^H\right)\mathbf{h}\quad\text{s. t.}\quad\mathbf{h}^H\mathbf{d}_{\theta_0}=1. \quad (16)$$

The solution to the above optimization problem yields the following optimal distortionless beamformer:

$$\mathbf{h}_{\mathrm{o}}=\frac{\left(\boldsymbol{\Phi}_{\mathbf{v}}+\mathbf{D}\boldsymbol{\Phi}\mathbf{Q}\mathbf{D}^H\right)^{-1}\mathbf{d}_{\theta_0}}{\mathbf{d}_{\theta_0}^H\left(\boldsymbol{\Phi}_{\mathbf{v}}+\mathbf{D}\boldsymbol{\Phi}\mathbf{Q}\mathbf{D}^H\right)^{-1}\mathbf{d}_{\theta_0}}. \quad (17)$$

Using the Sherman-Morrison formula, the inverse of $\boldsymbol{\Phi}_{\mathbf{v}}+\mathbf{D}\boldsymbol{\Phi}\mathbf{Q}\mathbf{D}^H$ can be decomposed as

$$\begin{aligned}&\left(\boldsymbol{\Phi}_{\mathbf{v}}+\mathbf{D}\boldsymbol{\Phi}\mathbf{Q}\mathbf{D}^H\right)^{-1}\\&=\boldsymbol{\Phi}_{\mathbf{v}}^{-1}-\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{D}\boldsymbol{\Phi}\mathbf{Q}\left(\mathbf{I}_N+\mathbf{D}^H\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{D}\boldsymbol{\Phi}\mathbf{Q}\right)^{-1}\mathbf{D}^H\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\\&=\boldsymbol{\Phi}_{\mathbf{v}}^{-1}-\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{D}\boldsymbol{\Phi}\left(\mathbf{Q}^{-1}+\mathbf{D}^H\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{D}\boldsymbol{\Phi}\right)^{-1}\mathbf{D}^H\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\\&=\boldsymbol{\Phi}_{\mathbf{v}}^{-1}-\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{D}\boldsymbol{\Phi}\left(\boldsymbol{\Sigma}+\mathbf{D}^H\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{D}\boldsymbol{\Phi}\right)^{-1}\mathbf{D}^H\boldsymbol{\Phi}_{\mathbf{v}}^{-1}, \quad (18)\end{aligned}$$

where $\mathbf{I}_N$ is the $N\times N$ identity matrix and

$$\boldsymbol{\Sigma}=\mathrm{diag}\left(\varsigma_1,\ \varsigma_2,\ \ldots,\ \varsigma_N\right)$$

is a diagonal matrix of size $N\times N$ with $\varsigma_n=1/Q_n$, $0\leq\varsigma_n\leq 1$, and $n=1,2,\ldots,N$.

The decomposition in (18) is more intuitive because it reveals several important special cases. Indeed, for $\boldsymbol{\Sigma}=\mathbf{I}_N\ (\forall M,N)$, we have $\mathbf{h}_{\mathrm{o}}=\mathbf{h}_{\mathrm{MVDR}}$, which is the MVDR beamformer [3], and for $\boldsymbol{\Sigma}=\mathrm{diag}\left(0,\ 0,\ \ldots,\ 0\right)$, with $M>N$, we have $\mathbf{h}_{\mathrm{o}}=\mathbf{h}_{\mathrm{LCMV}}$, which is the LCMV beamformer [3]; indeed, one can verify in this case that $\mathbf{D}^H\left(\boldsymbol{\Phi}_{\mathbf{v}}+\mathbf{D}\boldsymbol{\Phi}\mathbf{Q}\mathbf{D}^H\right)^{-1}=\mathbf{0}$.

By individually adjusting the values of the ς_n's, we can achieve various trade-offs between background noise attenuation and interference suppression. This new beamformer offers significantly greater flexibility compared to the conventional MVDR and LCMV beamformers, as it encompasses both methods.

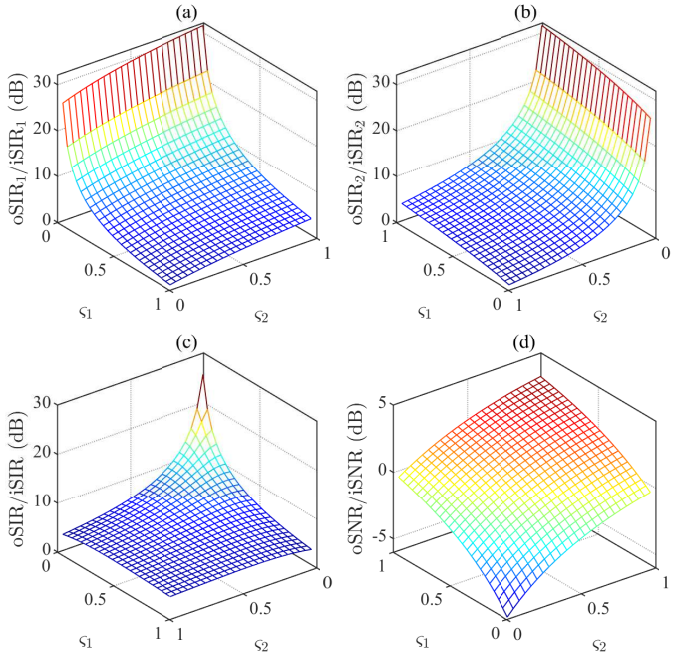


Fig. 1. Performance of the proposed beamformer as a function of ς_1 and ς_2: (a) gain in SIR_1, (b) gain in SIR_2, (c) gain in SIR, and (d) gain in SNR. Conditions: $f=1$ kHz, $\Delta_\theta=30°$, and $\mathrm{iSIR}_1=\mathrm{iSIR}_2=\mathrm{iSNR}=10$ dB.

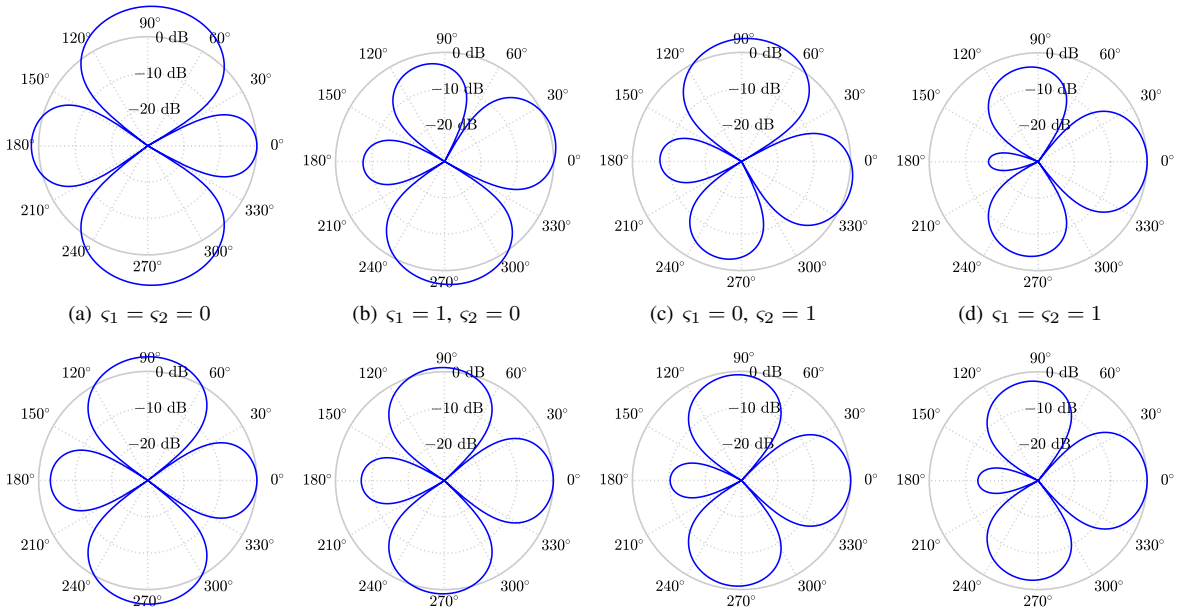
V. Simulations

In this section, we evaluate the performance of the developed distortionless beamformer through numerical simulations using various performance metrics. We examine the trade-offs between background noise reduction and interference suppression, and analyze the beampattern behavior, as illustrated in previous studies [3, 16, 20].

We consider a uniform circular array (UCA) with $M=8$ microphone sensors positioned in an anechoic environment. The UCA has a radius of 4 cm. There are one desired source and two interference sources in the far field of the array. The azimuth angle of the desired source is $\theta_0=0°$, while the azimuth angles of the two interference sources are $\theta_1=\theta_0+\Delta_\theta$ and $\theta_2=\theta_0-\Delta_\theta$, respectively, with $\Delta_\theta>0$. Taking the center of the UCA as the reference point, the steering vector at angle θ is given by [21]

$$\mathbf{d}_\theta(f)=\left[\ e^{\jmath 2\pi fr/c\cos(\theta-\psi_1)}\quad\ldots\quad e^{\jmath 2\pi fr/c\cos(\theta-\psi_M)}\ \right]^T, \quad (19)$$

where $\jmath$ is the imaginary unit with $\jmath^2=-1$, c is the speed of sound, which is typically 340 m/s, and $\psi_m=2\pi(m-1)/M$ is the angular position of the mth sensor. The noise field consists of spherically isotropic (diffuse) noise and spatially



(a) $\varsigma_1 = \varsigma_2 = 0$ (b) $\varsigma_1 = 1$, $\varsigma_2 = 0$ (c) $\varsigma_1 = 0$, $\varsigma_2 = 1$ (d) $\varsigma_1 = \varsigma_2 = 1$

(e) $\varsigma_1 = \varsigma_2 = 0.2$ (f) $\varsigma_1 = \varsigma_2 = 0.4$ (g) $\varsigma_1 = \varsigma_2 = 0.6$ (h) $\varsigma_1 = \varsigma_2 = 0.8$

Fig. 2. Beampattern examples of the proposed beamformer for different combinations of ς_1 and ς_2. Conditions: $f = 1$ kHz, $\Delta_\theta = 30°$, and $\text{iSIR}_1 = \text{iSIR}_2 = \text{iSNR} = 10$ dB.

white noise. In this case, the noise coherence matrix is

$$\mathbf{\Gamma_v}(f) = (1-\mu)\mathbf{\Gamma}_{\text{d}}(f) + \mu\mathbf{I}_M, \tag{20}$$

where μ is a real number that relates to the power ratio between the two components of noise, $\mathbf{\Gamma}_{\text{d}}(f)$ is the coherence matrix for the diffuse noise, and $\mathbf{I}_M$ is the $M \times M$ identity matrix. In our simulations, we set μ to 0.05.

To evaluate the performance of the beamformer on each individual interference source, we calculate the input and output SIRs for the nth interference source as follows:

$$\text{iSIR}_n(f) = \frac{\phi_{X_0}(f)}{\phi_{X_n}(f)} \tag{21}$$

$$\text{oSIR}_n(f) = \frac{\phi_{X_0}(f)\left|\mathbf{h}^H(f)\mathbf{d}_{\theta_0}(f)\right|^2}{\phi_{X_n}(f)\left|\mathbf{h}^H(f)\mathbf{d}_{\theta_n}(f)\right|^2}. \tag{22}$$

Figure 1 plots the performance of the proposed beamformer given in (17) and (18) as a function of the design parameters ς_1 and ς_2. The plot is generated with $\Delta_\theta = 30°$, $f = 1.0$ kHz, and $\text{iSIR}_1 = \text{iSIR}_2 = \text{iSNR} = 10$ dB. Figures 1(a) and (b) show that, for each $n \in \{1, 2\}$, the gain in SIR_n increases as the value of ς_n decreases. When $\varsigma_n = 0$, the beamformer is configured to completely eliminate the nth interference, resulting in a theoretical infinite gain in SIR_n. Figures 1(c) and (d) reveal that decreasing both values of ς_1 and ς simultaneously enhances the overall SIR gain, but this improvement comes at the cost of reduced SNR gain. It is important to note that negative SNR gains may occur with small values of ς_1 and ς, indicating an increase in background noise.

The LCMV and MVDR beamformers represent two specific cases of our proposed beamformer. The LCMV ($\varsigma_1 = \varsigma_2 = 0$) achieves infinite overall SIR gain but results in the lowest SNR gain. In contrast, the MVDR ($\varsigma_1 = \varsigma_2 = 1$) provides the maximum SNR gain but may not deliver sufficient SIR gain. Overall, by adjusting the values of ς_1 and ς_2, our proposed beamformer allows for various trade-offs between different performance gains.

We now analyze the beampattern behavior of the proposed beamformer to explain the observed variations in performance. Figure 2 illustrates the beampatterns for different combinations of ς_1 and ς_2.

Figure 2(a) shows that the LCMV ($\varsigma_1 = \varsigma_2 = 0$) creates perfect nulls at 30° and 330°. However, this results in a pattern issue where some sidelobes have heights significantly greater than unity (the response in the desired source direction), which can account for the amplification of background noise. Figures 2(b) and (c) demonstrate that when $\varsigma_n = 0$, the proposed beamformer achieves a perfect null in the direction of each corresponding interference source, effectively eliminating its impact. Figure 2(d) shows that the MVDR ($\varsigma_1 = \varsigma_2 = 1$) does not create perfect nulls for the interference sources but maintains well-behaved sidelobes with lower levels compared to the main beam. Figures 2(e)–(h) reveal that increasing both ς_1 and ς_2 simultaneously reduces the sidelobe levels but shifts the spatial nulls away from the interference directions.

VI. Conclusions

This paper investigated distortionless beamforming within a general signal model where multiple interference sources and background noise coexist with the desired signal. To offer highly flexible trade-offs between interference rejection and background noise reduction, we proposed a novel optimization criterion and developed a parameterized beamformer that includes the conventional MVDR and LCMV beamformers as special cases. The effectiveness of our proposed beamformer was demonstrated through numerical evaluations using different performance metrics and an analysis of beampattern behavior.

REFERENCES

[1] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[3] H. L. Van Trees, *Optimum Array Processing*. John Wiley & Sons, 2002.

[4] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, pp. 636–656, Jul. 2007.

[5] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 260–276, Feb. 2010.

[6] A. Adler, O. Schwartz, and S. Gannot, "A weighted multichannel Wiener filter and its decomposition to LCMV beam former and post-filter for source separation and noise reduction," in *Proc. IEEE ICSEE*, pp. 1–4, 2018.

[7] N. Pan, J. Benesty, and J. Chen, "A controlled noise reduction Wiener filter based on the quadratic eigenvalue problem," pp. 1990–1994, 2023.

[8] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 393–398, June 1986.

[9] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.

[10] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926–935, Aug. 1972.

[11] M. Souden, J. Benesty, and S. Affes, "Linear filtering for noise reduction and interference rejection," in *Proc. IEEE ICASSP*, pp. 89–92, 2010.

[12] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1053–1065, Mar. 2007.

[13] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, pp. 1071–1086, Aug. 2009.

[14] E. A. P. Habets, J. Benesty, S. Gannot, P. A. Naylor, and I. Cohen, "On the application of the LCMV beamformer to speech enhancement," in *Proc. IEEE WASPAA*, pp. 141–144, 2009.

[15] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multispeaker LCMV beamformer and postfilter for source separation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 940–951, May 2017.

[16] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, pp. 4925–4935, Sep. 2010.

[17] E. A. P. Habets, J. Benesty, and P. A. Naylor, "A speech distortion and interference rejection constraint beamformer," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 854–867, Mar. 2012.

[18] S. Karimian-Azari, J. Benesty, J. R. Jensen, and M. G. Christensen, "A broadband beamformer using controllable constraints and minimum variance," in *Proc. EUSIPCO*, pp. 666–670, 2014.

[19] C. Pan, J. Chen, and J. Benesty, "Microphone array beamforming with high flexible interference attenuation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1865–1876, 2022.

[20] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. John Wiley & Sons, 2018.

[21] J. Benesty, J. Chen, and I. Cohen, *Design of Circular Differential Microphone Arrays*. Berlin, Germany: Springer-Verlag, 2015.

How Accurate Can Large Vision Language Model Perform for Images with Compression Degradation?

Xiaohan Fang[a], Peilin Chen[a], Meng Wang[b], Shiqi Wang[*a,c]
[a]*Department of Computer Science, City University of Hong Kong*, Hong Kong SAR, China
[b]*School of Data Science, Lingnan University*, Hong Kong SAR, China
[c]*Shenzhen Research Institute, City University of Hong Kong*, Shenzhen, China

***Abstract*—The rapid evolution of Large Language Models (LLMs) has spurred the development of Large Vision Language Models (LVLMs), which demonstrate remarkable proficiency in various computer vision tasks through corresponding input prompts. These models perform impressively on diverse multimodal benchmarks, matching the effectiveness of conventional task-specific models. Nonetheless, many experiments with LVLMs often assume that the images are pristine, overlooking potential information loss during image transmission. To explore how image compression affects the semantic analysis capabilities of LVLMs in real-world scenarios, this paper presents a new image-text dataset named GPT-COMP. This dataset comprises 80,000 natural scene images, including 20,000 raw images from two public datasets. Each image is subjected to three different levels of compression distortion (QP = 32, 42, 52) using the latest Versatile Video Coding (VVC) Test Model. We leverage these variably compressed images in GPT-COMP to evaluate the state-of-the-art LVLM, GPT-4o, in vision understanding tasks. The text responses generated by GPT-4o are further organized into the GPT-COMP dataset and serve as the basis for evaluation. Specifically, the scene understanding capabilities regarding different compression levels are measured based on extracted semantic features and our proposed self-scoring strategy. This analysis sheds light on how image compression affects the semantic analysis capability of LLMs, offering valuable insights into the resilience of these models under realistic, suboptimal conditions.**

***Index Terms*—Large vision language models, GPT-4o, compression distortion, semantic evaluation**

I. Introduction

With the advent of Large Language Models (LLMs), researchers have developed their ability to understand multiple data types [1]. LLMs such as BEiT-3 [2], LLaVA [3], Sora [4], and GPT-4o [5] have demonstrated remarkable proficiency in handling multimodal tasks across text, images, and video. With the increase in training data, computational power, and model size, Large Vision Language Models (LVLMs) have outperformed dedicated deep-learning models in various vision tasks [6]–[10]. Recently, there have been many works focusing on investigating the visual interaction capabilities of LVLMs, such as image captioning [11]–[13], visual question answering [14]–[16], and image generation [17]–[20].

While most current research on LVLMs assumes pristine input images, real-world scenarios often involve visual content like images and videos being compressed to varying degrees due to storage and bandwidth limitations. Few studies have focused on the semantic analysis of LVLMs when dealing with compressed visual input. For instance, many online chat applications automatically compress images, and numerous video platforms adjust image quality based on available bandwidth to ensure smooth playback. Compression is nearly ubiquitous nowadays, and the resulting loss of detailed information can lead to misidentification of objects. This raises an important question: how do LVLMs perform when images suffer from compression distortion due to transmission or storage?

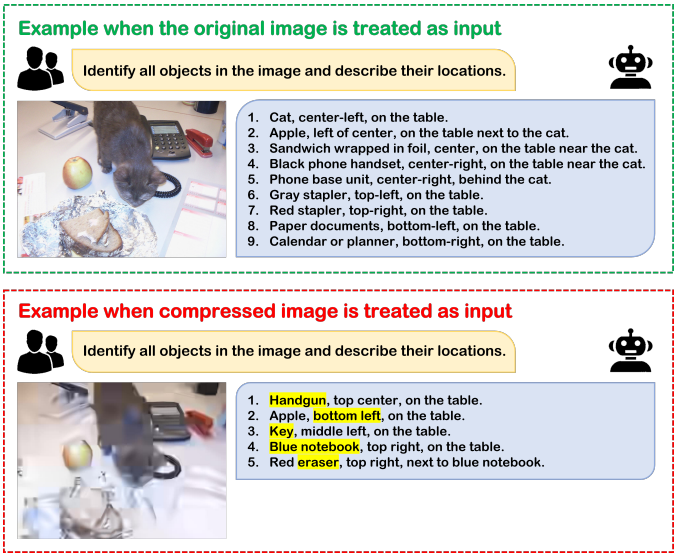


Fig. 1. Herein, GPT-4o is asked to conduct vision understanding. When the input image is pristine, GPT-4o can produce accurate answers. Whereas, when the input image is compressed and of lower quality, GPT-4o suffers from several types of hallucinations, including failing to recognize certain objects, misidentifying objects, and misidentifying the orientation of objects.

As illustrated in Fig. 1, GPT-4o is tasked with recognizing all objects in an input image and describing their orientation as accurately as possible. When the input image is uncompressed, GPT-4o can accurately identify almost all objects. However, as the level of compression increases, GPT-4o starts to make various errors. Certain objects, such as the tinfoil-wrapped sandwich at the bottom of the image and the gray stapler at the top left, are overlooked. The cat in the center of the image, along with the telephone and the red stapler on the right, are misidentified as other objects. Additionally, the position of the

*Corresponding author: Shiqi Wang.

apple, which is located to the left of the center, is incorrectly described by GPT-4o as being at the bottom left.

In this paper, we conduct a systematic study to explore the impact of compression on LLMs in visual understanding. A new image-text dataset, GPT-COMP, has been developed using 20,000 images sourced from two publicly available datasets: COCO [21] and GQA [22]. The Versatile Video Coding (VVC) [23] Test Model was employed to generate images of varying compression levels by applying three different Quantization Parameters (QPs). More importantly, a series of prompts were designed to task GPT-4o, and its responses under various conditions were collected within GPT-COMP. This dataset aims to provide abundant data for research on LVLM perception under suboptimal conditions. To quantitatively analyze the response texts of GPT-4o, we utilized four top-performing Semantic Textual Similarity (STS) models [24]–[27] from the Massive Text Embedding Benchmark (MTEB) [28] leaderboard. Overall, the contributions of this paper can be summarized as follows,

- We introduce GPT-COMP, a dataset featuring 20,000 images subjected to three different levels of coding distortion using VVC. This dataset includes the textual outcomes of vision understanding performed by GPT-4o, facilitating research on the impact of compression on the performance of large vision-language models.
- We leverage state-of-the-art text encoders to extract textual features and analyze the semantic similarity between LVLM's responses and the ground truth derived from uncompressed images. This allows quantitatively measuring how the model retains semantic accuracy in varying image quality.
- We explore a new evaluative approach by directing GPT-4o to self-score its responses against uncompressed reference images. This offers deeper insights into the model's perceptual and cognitive processing under compression, revealing how image compression impacts semantic understanding in a fine-grained manner.

II. Data Acquisition

To conduct a comprehensive analysis of GPT-4o's ability to recognize compressed images, we selected 20,000 raw images $\mathbf{I}_r$ from two public datasets, COCO and GQA, as our source images. These raw images were then compressed using the VVC Test Model (VTM) 23.3 at three different QP levels (i.e., 32, 42, and 52). This compression process introduced noise ε to the images, as described by the following equation:

$$\mathbf{I}^c = \mathbf{I}^r + \varepsilon, \tag{1}$$

where $\mathbf{I}^c$ represents the compressed images. Consequently, a total of 80,000 images depicting images under various coding conditions are obtained. For each image, GPT-4o was tasked with performing a visual understanding task and a corresponding self-scoring task. This process yielded a batch of textual responses for each image, which were then integrated with the 80,000 images to construct the final GPT-COMP dataset. The detailed process is outlined below.

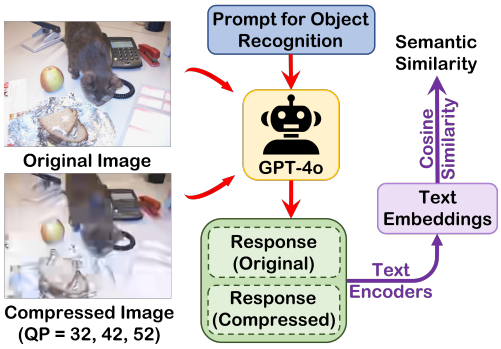


Fig. 2. The flow of our data collection process. The red lines represent our process for collecting visual understanding results from GPT-4o for images with different compression levels. The purple lines represent the process of quantifying the semantics of the recognition results utilizing a text encoder.

A. Visual Understanding

As shown in Fig. 2, to explore the effect of compression on the visual understanding ability of the GPT-4o, we ask it to output all the objects and their positions from the images that can be identified. For the convenience of subsequent analyses, we design a prompt p_1 such that GPT-4o can output texts in a specified format for each image, which is shown as follows,

- Please complete the task as required. List all the objects you can identify in the picture and describe the location of each one. The required format is: 1. Object1, Location. 2. Object2, Location... and so on. Objects can be described with up to two adjectives. For the location of each object, describe its position in the image and also its position in the scene of the image (e.g., red apple, right of center, in a bowl next to a vase). For coupled objects please describe the main object in terms of appendages. (use "Man in red suit" rather than "Red suit, worn by a man"). Please change lines after each object and its location is printed. No blank lines, no need to be concerned about the quality of the image, and no subjective statements in the text of your answer or anything outside the format.

An example of text response for the original and compressed images is shown in Fig. 1. For all 20,000 source images and the corresponding compressed images, we obtain response text from GPT-4o using the above process,

$$\mathbf{T}^r = \xi(\mathbf{I}^r, p_1), \mathbf{T}^c = \xi(\mathbf{I}^c, p_1), \tag{2}$$

where $\xi(\cdot)$ represents the processing operation of GPT-4o. The variables $\mathbf{T}^r$ and $\mathbf{T}^c$ represent the text responses for original and compressed images, respectively. The distribution of object numbers N in source images is presented in Table I. Notably, for more than half of the images, the number of objects N falls within the range of 5 to 10. We treat $\mathbf{T}^r$ as the ground truth in the following feature extraction and self-scoring.

B. Self-Scoring

In our implementation, each interaction with GPT-4o operates independently, ensuring that responses are unaffected

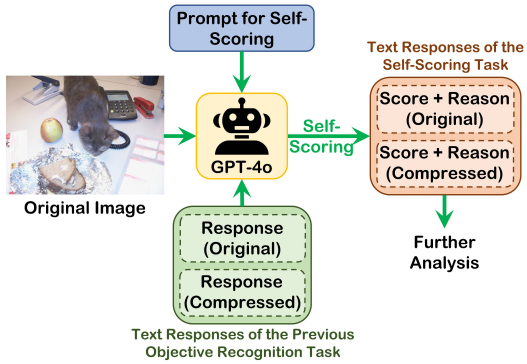


Fig. 3. Example of GPT-4o's self-scoring of object recognition results. The rationale corresponding to different scores from 0 to 10 is shown in this figure. Note that these results do not come from the same image.

TABLE I
STATISTICS OF THE NUMBER OF OBJECTS N IN GPT-4O'S RECOGNITION RESULT OF THE SOURCE IMAGES FROM COCO AND GQA, RESPECTIVELY.

Source	N<=5	5<N<=10	10<N<=15	15<N<=20	20<N
COCO	2305	5300	1926	417	52
GQA	1915	5439	2164	422	60

0	failed to identify any objects correctly.
1	incorrect identifications and locations like chocolate doughnut with sprinkles and cinnamon roll which are not present.
2	only correctly identified the pine tree and the snow-covered ground but incorrectly identified wooden benches and a street light.
3	identified fewer objects with limited detail and accuracy regarding their locations.
4	inaccurately identified sugar cube and piece of cake, didn't correctly mention the latte art, and had the wrong location for napkin.
5	many objects are identified with some appropriate locations, but there's confusion about certain items and their placements.
6	inaccurately identified the laptop or tablet, bookshelf placement, and the position of man and couch.
7	missed identifying objects like the plate and mistook the fork for a spoon.
8	incorrect identification and location for the bus driver; correct for all other objects.
9	accurate identification but slight difference in the color description of the flowers.
10	accurately identifies all objects and their locations.

Fig. 4. Example of GPT-4o's self-scoring of object recognition results. The rationale corresponding to different scores from 0 to 10 is shown in this figure. Note that these results may not come from the same image.

by prior sessions. This characteristic is pivotal as it allows us to implement a self-scoring mechanism where the model evaluates its own prior image recognition outputs. Specifically, for each quantization parameter level, GPT-4o is presented with 20,000 original images $\mathbf{I}^r$, alongside their corresponding textual responses $\mathbf{T}^c$, generated in prior visual understanding tasks. This setup enables the model to perform a self-assessment of its previous object recognition accuracy on a scale from 0 to 10. Specifically, each textual response $\mathbf{T}^c$ is re-input alongside $\mathbf{T}^r$, serving as a reference to anchor the model's evaluation. We further require GPT-4o to articulate not only the scores but also detailed justifications for each score, enhancing the transparency and reliability of the evaluation. The use of prompt p_2, detailed in Fig. 3, facilitates this comprehensive assessment. This self-scoring approach

TABLE II
SEMANTIC-BASED QUANTITATIVE PERFORMANCE OF GPT-4O IN THE VISUAL UNDERSTANDING TASK. THE TABLE DEMONSTRATES THE COSINE SIMILARITY VALUE ON THREE QPS WITH THE UNCOMPRESSED VERSION FOR THE IMAGES SOURCED FROM COCO AND GQA, RESPECTIVELY.

Source	QP	NV-Embed-v1	GTE-v1.5	Mxbai-v1	GIST-v0
COCO	32	0.8321	0.8787	0.9139	0.8913
	42	0.8096	0.8634	0.8995	0.8745
	52	0.6816	0.7790	0.7969	0.7616
GQA	32	0.8328	0.8744	0.9145	0.8904
	42	0.8113	0.8595	0.9005	0.8744
	52	0.7012	0.7841	0.8134	0.7774
Average	32	0.8324	0.8766	0.9142	0.8909
	42	0.8104	0.8614	0.9000	0.8744
	52	0.6914	0.7815	0.8051	0.7695

provides a unified framework for assessing the model's semantic analysis capabilities, especially under varying levels of image compression, offering insights into the adaptability and resilience of LVLMs in processing degraded images in real-world conditions.

- The two texts attempt to list all the objects in the image and describe their locations. Give each text a combined score based on: 1. the number of objects that are accurately identified. 2. The correctness of the corresponding locations of the objects. Score on a scale from 0 to 10, higher scores represent better performance. Also, give the reason for such scoring in one sentence.
- Please do not output anything else, but format it as the following example. NO blank lines. Do NOT output any explanatory text.
- X (X is an Arabic numeral, scoring for the text1), the reason for such scoring (in ONE sentence). Y (Y is an Arabic numeral, scoring for the text2), the reason for such scoring (in ONE sentence).

In this way, 60,000 scores are obtained (20,000 for each QP),

$$\mathbf{T}^s = \{S_r, \mathbf{R}\} = \xi(\mathbf{I}^r, \mathbf{T}^r, \mathbf{T}^c, p_2), \tag{3}$$

where $\mathbf{T}^s$ represents the self-scoring outputs by GPT-4o. S_r and $\mathbf{R}$ denote the scores and corresponding reasons, respectively. Some samples of $\mathbf{T}^s$ are shown in Fig. 4. The obtained $\mathbf{T}^s$ from the self-scoring task, along with the text responses $\mathbf{T}^r$ and $\mathbf{T}^c$ from the previous visual understanding task, are integrated with the images in a one-to-one correspondence to form the GPT-COMP dataset.

To summarize, the proposed GPT-COMP contains 20,000 source images $\mathbf{I}^r$, 60,000 compressed images $\mathbf{I}^c$, 20,000 $\mathbf{T}^r$ together with 60,000 $\mathbf{T}^c$ obtained from the visual understanding task, and 60,000 $\mathbf{T}^s$ obtained from the self-scoring task.

III. ANALYSES

A. Semantic Analyses

To analyze GPT-4o's performance in the visual understanding task quantitatively, we extract semantic feature embeddings from these textual responses using four state-of-the-art STS

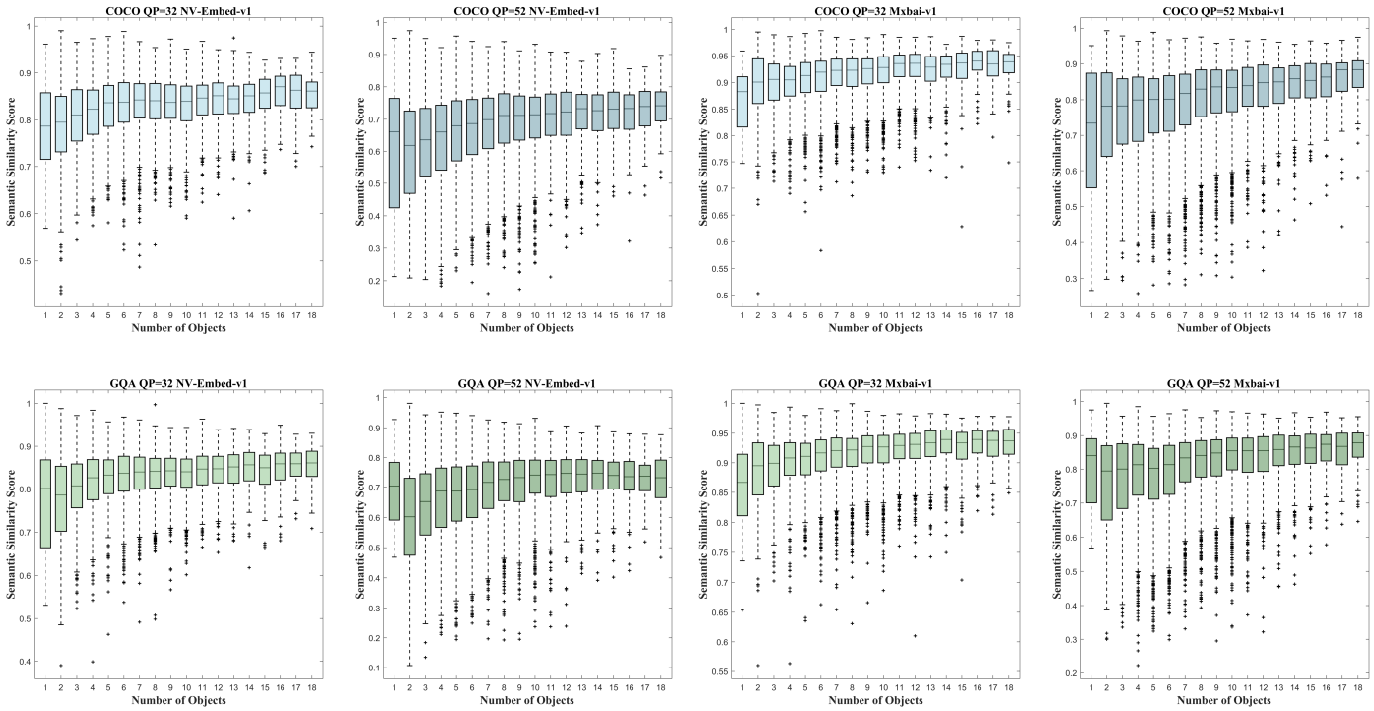


Fig. 5. Distribution of the number of objects in the image with the corresponding semantic similarity. The cases where the model NV-Embed-v1 and Mxbai-v1 with QPs of 32 and 52 are shown in the figure. The images containing more than 18 objects were not included in the graph due to the lack of samples.

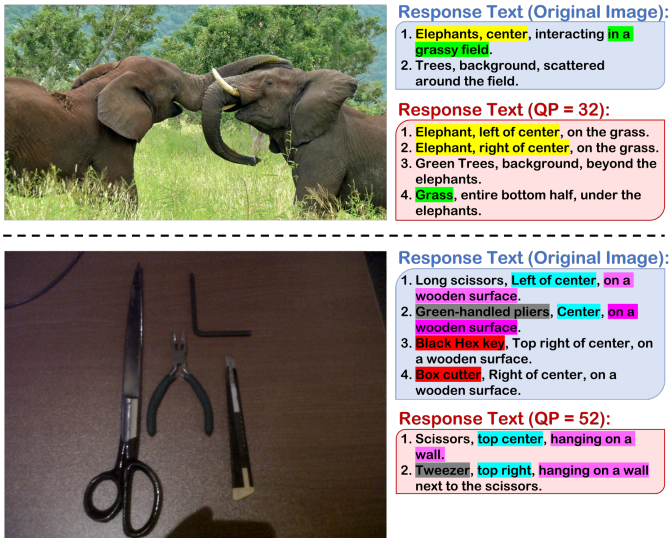


Fig. 6. Two representative images with objects number $N < 5$. The right side displays the response text, both for the original images ($\mathbf{T}^r$) and compressed images with specific QP values ($\mathbf{T}^c$). The various colors of highlights represent multiple types of differences between ($\mathbf{T}^c$) and ($\mathbf{T}^r$)

models on the MTEB leaderboard, including NV-Embed-v1 [24], GTE-v1.5 [25], Mxbai-v1 [26], and GIST-v0 [27], as shown by the purple line in Fig. 2. This process is performed on the text responses for both the original and compressed images, which are formulated as follows,

$$E^r = f(\mathbf{T}^r), E^c = f(\mathbf{T}^c), \tag{4}$$

where $f(\cdot)$ represents the text encoder of the STS models. E^r and E^c are the text embeddings corresponding to $\mathbf{T}^r$ and $\mathbf{T}^c$. In our experiments, E^r and E^c are one-dimensional vectors. We take the embeddings E^r from the original image as a reference and compute cosine similarity S_c with the feature E^c from its distorted version under three bitrate conditions, which is given by,

$$S_c = g(E^r, E^c) = \frac{\sum_{i=1}^{n} E_i^r E_i^c}{\sqrt{\sum_{i=1}^{n} {E_i^r}^2} \cdot \sqrt{\sum_{i=1}^{n} {E_i^c}^2}}, \tag{5}$$

where E_i^r and E_i^c are the ith components of vectors E^r and E^c, respectively. A larger value of S_c represents a greater semantic similarity between $\mathbf{T}^c$ and $\mathbf{T}^r$. The result is shown in Table II. It can be observed that the semantic similarity derived from the models is close for compressed images with QP = 32 and QP = 42, while a significant drop occurs when the QP is further increased to 52. We also organize the relationship between the number of objects (N) in the image and the semantic similarity, as shown in Fig. 5. As illustrated by the box plots, except for fluctuations due to the small sample size (e.g., $N = 1$), the average semantic similarity decreases when there are fewer objects in the image, and the distribution of semantic similarity becomes more dispersed. This phenomenon is further accentuated with higher QP values. Two representative images are displayed in Fig. 6. At a low compression distortion level (QP = 32, as in the upper part of Fig. 6), it can be observed that the content of the image is correctly described in the response of GPT-4o. However, compared with the text response in the original

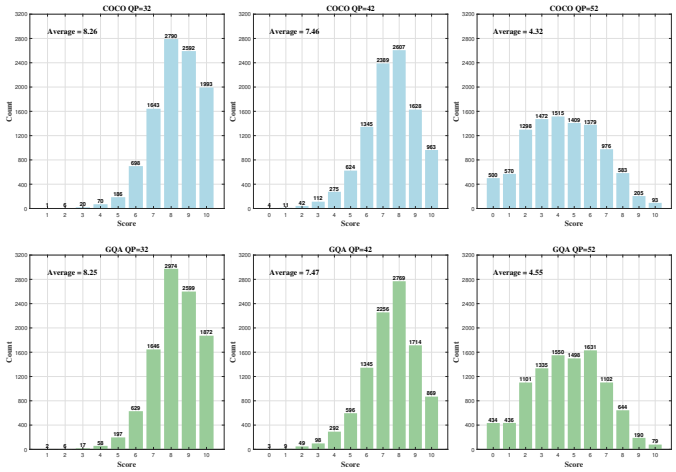


Fig. 7. Distribution of the scores S_r obtained in the self-scoring task for images from COCO and GQA at QP = 32, 42, and 52.

image ($\mathbf{T}^r$), the response at QP = 32 ($\mathbf{T}^c$) enumerates the two elephants separately (marked in yellow), and the grass in the background is treated as one of the objects (marked in green). In this case, although the descriptions of $\mathbf{T}^c$ and $\mathbf{T}^r$ are almost the same, there are two more objects listed in $\mathbf{T}^c$, which leads to a decrease in semantic similarity S_c. When the compression strength is increased, more hallucinations appear in the recognition of images by GPT-4o. The lower part of Fig. 6 illustrates an example of the situation when QP = 52. The first thing that can be observed is that $\mathbf{T}^c$ missed two objects, the hex key and the box cutter (marked in red). For the two remaining objects, the background is misidentified as a wall (marked in purple), while the orientation is incorrectly described in $\mathbf{T}^c$ (marked in cyan). Besides, the pliers are incorrectly identified as tweezers (marked in gray). Overall, as shown in Fig. 6, the semantic differences between $\mathbf{T}^c$ and $\mathbf{T}^r$ account for a significant proportion of the overall variance. We believe that both descriptive differences and recognition errors contribute to decreased semantic similarity. A smaller number of objects amplifies this impact on the similarity, resulting in a lower mean and more significant variance.

B. Self-Scoring Analyses

We compile the distributions of scores S_r given by GPT-4o in the self-scoring task on different QPs and present them in histograms as in Fig. 7. It can be observed that for images from the COCO and GQA datasets, the distribution of S_r at three different QPs is similar. In particular, the increase in QP leads to a decrease in S_r while broadening the distribution of S_r. It is also worth noting that in previous experiments, where semantics are converted to text embeddings, semantic similarity shows little differentiation in the QP = 32 and QP = 42 results. From Table II, the mean value of semantic similarity decreases by less than 0.03 for QP = 42 compared to QP = 32 across the 4 STS models. In contrast, self-scoring can circumvent the semantic differences caused by different descriptions (as in the upper part of Fig. 6) of the same thing on low compression levels, which leads to a more discriminating S_r at QP = 42 compared to QP = 32. Besides, we calculate the Spearman rank-order correlation coefficient (SRCC) and the Pearson linear correlation coefficient (PLCC) between S_r and the semantic similarity obtained via text embeddings. As shown in Table III, it can be observed that a larger QP leads to a stronger correlation between S_r and semantic similarity. This phenomenon occurs because, at a low compression distortion level, a significant portion of the semantic difference between $\mathbf{T}^c$ and $\mathbf{T}^r$ is attributable to variations in expression rather than in accuracy. Thus, both the quantized semantic similarity and the self-score S_r are predominantly distributed at a high value across the 20,000 images, as illustrated in Fig. 5 and Fig. 7, leading to a low correlation between S_r and semantic similarity.

TABLE III
THE CORRELATION BETWEEN SCORES OF TEXTUAL RESPONSES IN SELF-SCORING AND THE SEMANTIC SIMILARITY OBTAINED FROM THE STS MODEL, IN TERMS OF SRCC AND PLCC. HIGHER VALUES OF SRCC AND PLCC REPRESENT STRONGER CORRELATIONS. THE TABLE DEMONSTRATES THE SITUATION ON 3 QPS FOR THE IMAGES SOURCED FROM COCO AND GQA, RESPECTIVELY.

Source	QP	NV-Embed-V1		GTE-V1.5		Mxbai-V1		GIST-V0	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
COCO	32	0.1227	0.1614	0.1229	0.1508	0.064	0.1051	0.0973	0.1272
	42	0.2576	0.3003	0.2324	0.2658	0.2026	0.2611	0.2236	0.2754
	52	0.6031	0.6109	0.5790	0.6046	0.5863	0.6081	0.6013	0.6238
GQA	32	0.1042	0.1445	0.0938	0.1267	0.0631	0.1096	0.0831	0.1216
	42	0.2566	0.3149	0.2195	0.2589	0.1882	0.259	0.206	0.2654
	52	0.5943	0.6123	0.5401	0.5745	0.5619	0.5982	0.5696	0.6068
Average	32	0.1135	0.1530	0.1084	0.1388	0.0636	0.1074	0.0902	0.1244
	42	0.2571	0.3076	0.2260	0.2624	0.1954	0.2601	0.2148	0.2704
	52	0.5987	0.6116	0.5596	0.5896	0.5741	0.6032	0.5855	0.6153

IV. CONCLUSIONS

We present the GPT-COMP dataset, the inaugural dataset designed to explore the impact of compression on the comprehension abilities of LVLMs with a focus on visual perception. We employ specialized prompts to test the current advanced LVLM, GPT-4o, generating a collection of normalized text responses from 80,000 images. Additionally, four cutting-edge STS models are used to quantitatively assess GPT-4o's scene understanding capabilities under varying levels of image compression. A self-assess method is also developed, providing insights into how image compression affects visual comprehension in a fine-grained manner. This research is positioned to contribute to the ongoing advancement of studies on utilizing LVLMs under realistic, suboptimal conditions.

ACKNOWLEDGMENTS

This work was supported under the Shenzhen Science and Technology Program (Project No. JCYJ20220530140816037).

REFERENCES

[1] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu, "Multimodal large language models: A survey," in *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 2247–2256.

[2] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som *et al.*, "Image as a foreign language: BEiT pretraining for vision and vision-language tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 175–19 186.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[4] [Online]. Available: https://openai.com/index/sora/

[5] [Online]. Available: https://openai.com/index/hello-gpt-4o/

[6] M. Hall, L. Gustafson, A. Adcock, I. Misra, and C. Ross, "Vision-language models performing zero-shot tasks exhibit disparities between gender groups," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023, pp. 2770–2777.

[7] H. Chen, "Comparison of large language and vision models on representative downstream tasks," in *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 2023, pp. 307–311.

[8] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.

[9] H. Abukadah, M. Fereidouni, and A. Siddique, "Mapping natural language intents to user interfaces through vision-language models," in *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, 2024, pp. 237–244.

[10] B. Yang, L. He, K. Liu, and Z. Yan, "VIAssist: Adapting multi-modal large language models for users with visual impairments," in *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems Internet of Things (FMSys)*, 2024, pp. 32–37.

[11] S. Bianco, G. Ferrario, and P. Napoletano, "Image captioning using pretrained language models and image segmentation," in *2022 IEEE 12th International Conference on Consumer Electronics (ICCE-Berlin)*, 2022, pp. 1–6.

[12] N. Rotstein, D. Bensaïd, S. Brody, R. Ganz, and R. Kimmel, "Fusecap: Leveraging large language models for enriched fused image captions," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 5677–5688.

[13] G. G. Computer Science and P. S. Computer Division, "FashionVLM - fashion captioning using pretrained vision transformer and large language model," in *2024 International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2024, pp. 1–5.

[14] D. Koshti, A. Gupta, and M. Kalla, "Knowledge blended open domain visual question answering using transformer," in *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, 2023, pp. 823–828.

[15] J. Pan, Z. Lin, Y. Ge, X. Zhu, R. Zhang, Y. Wang, Y. Qiao, and H. Li, "Retrieving-to-answer: Zero-shot video question answering with frozen large language models," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023, pp. 272–283.

[16] J. Liu, C. Fang, L. Li, B. Li, D. Hu, and C. Ma, "Prompting large language models with fine-grained visual relations from scene graph for visual question answering," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 8125–8129.

[17] B. Liu, H. Zhang, J. Liu, and Q. Wang, "Acigs: An automated large-scale crops image generation system based on large visual language multi-modal models," in *2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2023, pp. 7–13.

[18] A. Singh, "A survey of ai text-to-image and ai text-to-video generators," in *2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC)*, 2023, pp. 32–36.

[19] J. Betker, "Dall·E 3," 2023. [Online]. Available: https://openai.com/dall-e-3

[20] [Online]. Available: https://www.midjourney.com/

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[22] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6700–6709.

[23] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC)," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1463–1493, 2021.

[24] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, and W. Ping, "NV-Embed: Improved techniques for training llms as generalist embedding models," 2024.

[25] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang, "Towards general text embeddings with multi-stage contrastive learning," *arXiv preprint arXiv:2308.03281*, 2023.

[26] D. K. J. L. Sean Lee, Aamir Shakir. (2024) Open source strikes bread - new fluffy embeddings model. [Online]. Available: https://www.mixedbread.ai/blog/mxbai-embed-large-v1

[27] A. V. Solatorio, "Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning," *arXiv preprint arXiv:2402.16829*, 2024. [Online]. Available: https://arxiv.org/abs/2402.16829

[28] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," *arXiv preprint arXiv:2210.07316*, 2022.

Effect of White Noise on Working Memory Using Event-Related Potentials

Seung-Won Lee[1,2], Jun-Seok Lee[1,2] and Han-Jeong Hwang[1,2*]

[1] Department of Electronics and Information Engineering, Korea University, Sejong, Korea

2 Interdisciplinary Graduate Program for Artificial Intelligence Smart Convergence Technology, Korea University, Sejong 30019, Republic of Korea

E-mail: superswl@korea.ac.kr, zck320@korea.ac.kr, *hwanghj@korea.ac.kr, Tel: +82-44-860-1762

***Abstract* — We investigated the effect of white noise on working memory performance using electroencephalography (EEG). Six healthy subjects participated in this study, and each subject performed a 2-back task under white noise and no sound conditions. After a series of EEG preprocessing, the amplitude of event-related potential (ERP) was calculated. The results indicated an increase in average working memory task accuracy and the mean P300 amplitude for target stimuli during the white noise condition compared to without it, suggesting that white noise may enhance working memory performance. However, due to the limited number of participants, statistical power was not attained. In our future studies, we will perform additional experiments with more subjects.**

I. Introduction

Working memory refers to the brain system used to temporarily store and manipulate information needed for complex cognitive tasks. Working memory capacity is closely related to learning and reading comprehension abilities, highlighting its importance. Recent studies have explored enhancing working memory capacity using external stimuli, such as transcranial electrical stimulation (tES) and sound-based auditory stimulation. However, while tES can cause discomfort and requires a separate electrical stimulator, auditory stimulation has the advantage of being accessible and easily integrated with wearable devices, as it can be provided anywhere with earphones or speakers.

White noise is a functional auditory stimuli known to enhance brain processing abilities, improving attention and cognitive performance [1]. Previous studies have primarily focused on inattentive children or those with attention deficit hyperactivity disorder (ADHD), showing cognitive improvement with strong white noise stimuli around 80 dB [2, 3]. Recently, it has been found that appropriate levels of white noise can positively affect working memory performance in healthy adults [4]. However, most studies have only reported behavioral performances, such as task accuracy or reaction time, without sufficient neurophysiological evidence.

Working memory is based on the continuous neural activity of a complex cerebral cortical network. These cortical dynamics can be observed by measuring neural oscillations using electroencephalography (EEG). Working memory is associated with EEG activity in the frontal and parietal regions, which can be simultaneously observed in the central area [5]. Therefore, this study aimed to neurophysiologically confirm the modulation of working memory capacity by white noise using central EEG. To this end, EEG was measured during working memory tasks with and without white noise, and changes in EEG patterns were analyzed under each condition.

II. Method

A. Data acquisition

Six healthy adults (2 males, 4 females, aged 21.67 ± 3.01 years) participated in this study. EEG was measured during cognitive tasks with and without white noise using Smartfones (mBrainTrain, Belgrade, Serbia), a wearable EEG measurement device in the form of a headset capable of simultaneous EEG measurement and white noise presentation (Fig. 1). EEG data were recorded at a sampling rate of 500 Hz, maintaining all electrode impedances below 15 kΩ during measurement. The study protocol was approved by the Institutional Review Board (IRB) of Korea University [KUIRB-2023-0139-01].

B. Auditory stimuli

We used white noise at 65 dB (mono track, 32-bit floating point with a sampling frequency of 44,100 Hz), which has shown to enhance visual working memory performance in healthy adults [4]. The white noise was generated using Audacity software (version 3.0.2 for Windows, generating white noise with a bandwidth of 86 Hz to 22.007 kHz). The generated white noise was delivered to participants through the Smartfones device.

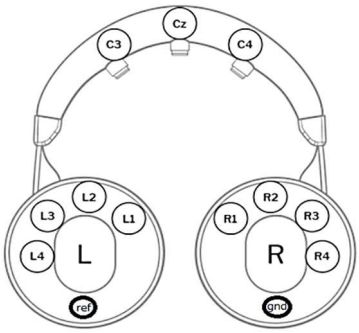


Fig. 1 EEG montage of the Smartfones EEG device by mBrainTrain (Belgrade, Serbia).

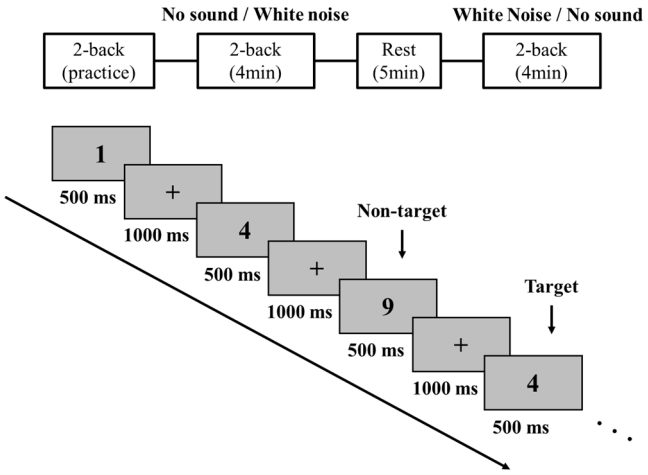


Fig. 2 Scheme of the experimental protocol used in this study.

C. *Experimental paradigm*

A 2-back task, widely used in neuroscience and electrophysiology studies, was used as the cognitive task in this study [6]. In the 2-back test, a series of letters or numbers is presented, and participants are asked to determine if the current stimulus corresponds to the one that appeared two steps earlier, assessing their working memory. The experiment was conducted in a shielded room to minimize external noise, and participants seated approximately 50 cm from the screen. Participants performed the 2-back task under both White Noise and No Sound conditions [4]. The visual stimuli consisted of random numbers (0-9) presented sequentially with an inter-stimulus interval of 1500 ms and a presentation time of 500 ms. Participants were instructed to press the right key on the keyboard if the current stimulus matched the one from two earlier (target stimulus) and the left key if it did not (non-target stimulus). They had to respond within 2 seconds, and missed or multiple responses were considered incorrect. The task consisted of 120 stimuli, with a target to non-target stimulus ratio of 3:7 [7]. Participants performed the task once under each condition, with a 5-minute break between two conditions. To exclude the learning and stimulation effects, half of the participants started with the White Noise condition, and the other half with the No Sound condition. Before the experiment, a brief practice session with 10 stimuli was conducted until 100% accuracy was achieved to ensure that participants were familiar with the 2-back task [8]. The performance of the 2-back task was evaluated based on accuracy and mean reaction time.

D. *EEG data analysis*

The preprocessing of EEG data was performed using the EEGLAB toolbox (version 2023.0) in MATLAB (MathWorks, Natick, MA, United States). To remove artifacts from raw EEG data, we performed a band-pass filtering between 0.5 and 30 Hz using basic finite impulse response (FIR) filter. Subsequently, an automatic noise removal algorithm, artifact subspace reconstruction (ASR), was applied to remove ocular and muscle artifacts, followed by independent component analysis (ICA) to manually remove any remaining artifacts. The artifact-free EEG data were segmented from -200 to 800 ms based on the stimulus onset and divided into target and non-target stimuli for event-related potential (ERP) analysis. Segments with amplitudes exceeding ±150 μV were excluded from the analysis. The P300 amplitude was calculated by averaging the power in the 300 – 500 ms window. Data from the ear electrodes were excluded from the analysis as they were not within the region of interest. Due to the limited number of participants, the statistical results were excluded from the analysis.

III. RESULTS

Table 1 shows the accuracy and reaction time for each of the six participants, as well as the average results. There was a trend towards higher average accuracy in the white noise condition, with all participants, except one subject (S1). While reaction times generally increased under the white noise condition, the differences were minimal

Table 1 Individual accuracy and reaction time

	Accuracy (%)		**Reaction time (ms)**	
	White noise	**No sound**	**White noise**	**No sound**
S1	91.53	93.22	272.81	252.35
S2	94.07	88.98	342.97	316.47
S3	90.68	87.29	293.26	250.93
S4	99.15	98.31	497.74	358.78
S5	96.61	91.53	166.51	221.41
S6	90.68	83.05	364.82	478.57
Mean	93.79	90.40	323.02	313.08
(Std)	(3.50)	(5.24)	(110.07)	(95.33)

Fig. 3 shows the average ERP of six participants for target and non-target stimuli in both conditions, with the P300 component appearing between 300 – 500 ms. Both conditions showed larger P300 amplitudes for target stimuli than for non-target stimuli, indicating successful task performance.

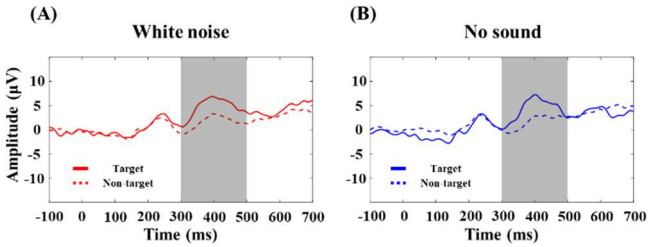


Fig. 3 (A) Average ERP of white noise condition, (B) Average ERP of no sound condition

Figure 4 presents the average target ERP under no sound and white noise conditions, with Table 2 quantifying the P300 values in the 300 – 500 ms window. The average P300 amplitude tended to be enhanced under the white noise condition. While not all participants demonstrated consistent neural responses, the majority exhibited a potentiated trend.

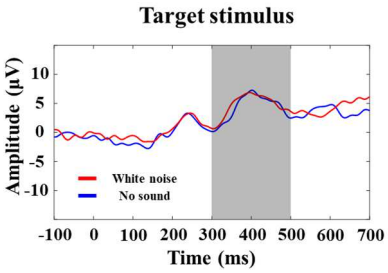


Fig. 4 Target stimulus average ERP of both condition

Table 2 Individual P300 amplitude

	Target ERP (μV)	
	White noise	**No sound**
S1	4.40	3.96
S2	5.52	2.75
S3	10.96	11.15
S4	0.36	-0.38
S5	4.06	3.97
S6	3.13	4.37
Mean (Std)	4.74 (3.51)	4.30 (3.78)

IV. Conclusions

This study aimed to neurophysiologically confirm the effect of white noise on working memory capacity using EEG in healthy adults. The P300 amplitude correlates with working memory capacity, with a higher P300 amplitude reflecting greater working memory capacity, potentially leading to more successful task execution [10]. In this study, both the average accuracy of the working memory task and the P300 amplitude tended to increase under the white noise condition, suggesting a potential positive effect of white noise on working memory. Although working memory performance exhibited improvements, there was a marginal increase in reaction times. Exposure to 65 dB of white noise may enhance task accuracy in working memory tasks but could also elevate stress levels [4]. The slight increase in reaction times could be a consequence of heightened stress. However, the effect was not consistent across all participants, and the limited number of participants precluded statistical significance. Future studies will recruit additional participants to obtain more conclusive results.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2023- 00258971) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

[1] Angwin, Anthony J., et al. “White noise facilitates new-word learning from context,” Brain and Language, vol.199, no.1, p 104699, 2019.

[2] Söderlund, Göran BW, et al. "The effects of background white noise on memory performance in inattentive school children," Behavioral and Brain Functions, vol.6, no.1, pp 1-10, 2010.

[3] Z. Wang, J. Zhang, X. Zhang, P. Chen, and B. Wang, Transformer Model for Functional Near-Infrared Spectroscopy Classification, IEEE Journal of Biomedical and Health Informatics 26 (2022), 2559-2569.

[4] Awada, Mohamad, et al. “Cognitive performance, creativity and stress levels of neurotypical young adults under different white noise levels," *Scientific Reports* vol.12, no.1, p 14566, 2022.

[5] PESONEN, Mirka; HÄMÄLÄINEN, Heikki; KRAUSE, Christina M. Brain oscillatory 4–30 Hz responses during a visual n-back memory task with varying memory load. Brain research, 2007, 1138: 171-177.

[6] CHEN, Yung-Nien; MITRA, Suvobrata; SCHLAGHECKEN, Friederike. Sub-processes of working memory in the N-back task:

An investigation using ERPs. Clinical Neurophysiology, 2008, 119.7: 1546-1559.

[7] DENG, Xin, et al. EEG analysis of working memory between sober state and intoxicated state. IEEE Access, 2021, 9: 145900-145911.

[8] SILVERMAN, Sarit; ASHKENAZI, Sarit. The differential relationship between visual and spatial working memory in children's mathematics performance. Trends in Neuroscience and Education, 2022, 29: 100188

[9] CHANG, Chi-Yuan, et al. Evaluation of artifact subspace reconstruction for automatic artifact components removal in multi-channel EEG recordings. IEEE Transactions on Biomedical Engineering, 2019, 67.4: 1114-1121.

[10] DONG, Shanshan, et al. Individual differences in working memory capacity are reflected in different ERP and EEG patterns to task difficulty. Brain research, 2015, 1616: 146-156.

Performance Optimization in the Cascade of VAD and ASR Systems: A Study on Evaluation and Alignment Strategies

Zhentao Lin*, Zihao Chen*§, Bi Zeng*, Leqi Chen* and Jia Cai†‡§
* Computer Science and Engineering, Guangdong University of Technology School
† China Electronic Product Reliability and Environmental Testing Research Institute
‡ Key Laboratory of MIIT for Intelligent Products Testing and Reliability
§ Corresponding Author

***Abstract*—Voice Activity Detection (VAD) is a critical pre-processing step in Automatic Speech Recognition (ASR) systems, tasked with distinguishing speech from non-speech segments in audio signals. The accuracy of this segmentation is essential, as errors can significantly impair ASR performance. This study integrates multiple VAD algorithms with ASR systems and analyzes their interaction by focusing on the Word Error Rate (WER) of ASR in relation to VAD's recall and precision. The research reveals a strong correlation between ASR's insertion and deletion errors and VAD's performance metrics. Insertions are mainly due to VAD misclassifying non-speech as speech, while deletions stem from misclassifying speech as non-speech. To address these issues, we introduce the Aligned Token Word Error Rate (atWER), a novel metric based on forced alignment from the Connectionist Temporal Classification (CTC) framework. Our experiments show that atWER accurately reflects the impact of VAD on ASR performance, effectively reducing discrepancies between VAD annotations and classification outcomes.**

I. Introduction

Voice Activity Detection (VAD) is a crucial initial step in Automatic Speech Recognition (ASR) systems, designed to accurately distinguish speech from non-speech elements within audio signals. The precision of this segmentation is vital, as non-speech interference—such as ambient noise or silence—can significantly impair ASR performance.

The accuracy of VAD is critical; incorrect classification of speech and non-speech can lead to a cascade of errors in ASR processing. Misclassifications may result in the loss of essential speech data or the inclusion of irrelevant noise, undermining the system's ability to accurately interpret and respond to spoken commands.

The correlation between VAD effectiveness and ASR error types is complex and significant [1]. The precision of the VAD mechanism significantly impacts the two predominant error categories in ASR systems, namely the occurrences of insertion and deletion errors.

- **Insertion Errors in ASR** occur when the system misinterprets non-speech sounds as speech, a failure of the VAD to filter out background noise. This leads to the ASR system processing unnecessary acoustic data, contaminating the output with false linguistic content and reducing reliability.
- **Deletion Errors in ASR** happen when the system misses actual speech segments due to the VAD misclassifying them as non-speech. This oversight can be due to stringent VAD detection criteria or inadequate representation of atypical speech in the training data, resulting in incomplete ASR outputs.

This study investigates the integration of VAD and ASR systems to enhance their performance. The methodology includes:

- A comparative temporal analysis of System VAD versus Oracle VAD using case studies, revealing a significant correlation between their metrics and temporal alignment.
- Experimental scrutiny was employed to examine the relationship between VAD and ASR metrics. The results indicated that when VAD and ASR systems are integrated sequentially, the composite performance metrics are predominantly determined by the VAD system's characteristics.
- Introduction of an algorithm for calculating Aligned Token Word Error Rate (atWER) to address error propagation challenges in the sequential deployment of VAD and ASR systems, mitigating misalignment between VAD annotations and classification outcomes.

II. Related Work

VAD, serving as a pivotal component in the pre-processing phase of ASR systems, has garnered escalating significance. The primary goal of VAD is to identify vocal segments within a continuous audio signal while reducing the impact of background noise and non-vocal sounds. This segmentation is vital for enhancing ASR system performance by minimizing incorrect interpretations and improving efficiency through reduced data processing.

Advancements in deep learning have propelled the development of VAD methods. Recurrent Neural Networks (RNNs) have been instrumental in capturing the temporal dynamics of audio signals [2]. Convolutional Neural Networks (CNNs) have also demonstrated effectiveness in feature extraction [3]. The Convolutional Recurrent Deep Neural Network (CRDNN) combines CNN and RNN features, utilizing convolutional layers for feature extraction and recurrent layers for time

series processing. This hybrid approach effectively captures relevant audio features to distinguish between speech and non-speech segments. The Feedback Sequential Memory Network (FSMN) leverages a feedforward architecture with memory units capable of storing and updating information, optimizing the handling of time series data [4].

Despite these advancements, challenges remain, particularly in the robustness of VAD systems when dealing with diverse accents, dialects, and noisy environments. The performance of VAD directly influences ASR outcomes. A recent study [5] introduces a multi-task learning framework that integrates VAD into the ASR system. This framework aims to improve ASR performance by using VAD alignment information during the training phase. [6] primarily introduces a novel method that applies Minimum Word Error training to a RNN to optimize VAD for improving the accuracy of speech recognition, especially on noisy data.

However, [7] point out that ASR systems are typically deployed in conjunction with a VAD system to operate ASR only on the voiced acoustic signals, thereby maintaining ASR performance by removing unnecessary non-speech parts from input audio signals during inference. However, if VAD fails to correctly split speech and non-speech segments, errors can propagate. Particularly in noisy environments or unknown acoustic domains, VAD is more prone to failure, which can trigger more significant insertion errors in ASR. [8] experimented with the hyperparameters of multi-channel VAD to mitigate the problems of insertion and deletion errors in ASR, but the approach continues to depend on the precision of the front-end System's VAD.

The problem of ASR accuracy being affected by VAD metrics such as Recall and accuracy is still resolved. Insertion errors in ASR can occur when VAD incorrectly classifies non-speech as speech, while deletion errors arise when speech is mistakenly categorized as non-speech.

III. Methods

A. System VAD and Oracle VAD

This section aims to delve into the intricacies of two significant approaches within the VAD domain: System VAD and Oracle VAD.

In System VAD, the decision-making process can be modeled as a binary classification problem, where the system must classify each frame of the audio signal as either speech or non-speech. A common approach is to define a decision function f_{vad} that takes a feature vector x_{vad} as input and outputs a binary decision y_{vad}:

$$y_{vad} = f_{vad}(x_{vad}, \theta) \tag{1}$$

Here, x_{vad} represents the feature vector extracted from the audio signal, which could include spectral, temporal, and statistical features. θ denotes the parameters of the VAD algorithm, which are typically learned from a training dataset.

Oracle VAD is an idealized model that represents the perfect detector with complete knowledge of the true speech segments. It serves as a theoretical benchmark to evaluate the performance of System VAD algorithms. The oracle detector can be mathematically represented as:

$$y_{oracle_vad} = 1[x_{vad} \in Speech] \tag{2}$$

Here, 1 is the indicator function that returns 1 if the condition is true (i.e., the feature vector x_{vad} belongs to a speech segment) and 0 otherwise.

B. VAD Metrics

The objective of this research is to delve into the effectiveness of VAD systems by examining four critical evaluation metrics: Accuracy (Acc), Precision (Pre), Recall, and F1-Score (F1). These metrics are essential for assessing the performance of VAD algorithms, providing a comprehensive view of their reliability and efficiency in detecting speech segments.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Pre = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 * \frac{Pre * Recall}{Pre + Recall} \tag{6}$$

where true positives (TP) is the number of segments correctly identified as speech, true negatives (TN) is the number of segments correctly identified as non-speech, false positives (FP) is the number of non-speech segments incorrectly identified as speech, and false negatives (FN) is the number of speech segments incorrectly identified as non-speech. Precision measures the proportion of true speech segments among all segments identified as speech by the system. Recall measures the proportion of true speech segments correctly identified out of all actual speech segments. The F1-Score provides a balanced measure, taking into account both precision and recall.

C. CTC-based and non-CTC-based ASR models

This section delves into the heart of ASR technology, examining two distinct approaches: CTC (Connectionist Temporal Classification)-based models and their non-CTC counterparts.

CTC-based ASR Models: The Connectionist Temporal Classification framework, introduced in [9], has been a pivotal breakthrough in the field of ASR. CTC-based models excel in handling variable-length input and output sequences, making them ideal for tasks where the alignment between the speech signal and the corresponding text is not predefined. This approach has been instrumental in the development of deep learning-based ASR systems, particularly those utilizing Transformer, RNN, and CNN. The CTC loss function, which allows for the direct alignment of input sequences with output labels, has been a key factor in the success of these models.

Non-CTC-based ASR Models: While CTC has set the standard for many ASR systems, alternative approaches have

emerged that challenge the status quo. Non-CTC-based models, such as those utilizing Attention Mechanisms and Transformer architectures, have gained prominence due to their ability to capture long-range dependencies and contextual information more effectively. These models often employ encoder-decoder frameworks that can directly predict the output sequence from the input sequence, offering a more interpretable and flexible approach to speech recognition.

D. ASR Metric

The WER (Word Error Rate) is an important metric for evaluating the performance of ASR systems. It measures the accuracy of recognition by quantifying the differences between the predicted text and the reference text. The formula for calculating WER is as follows:

$$WER = \frac{S + D + I}{N} \tag{7}$$

where S represents the number of substitution (SUB) errors, D represents the number of deletion (DEL) errors, I represents the number of insertion (INS) errors, and N represents the total number of words or characters in the reference text. The lower the WER value, the better the performance of the ASR system.

E. atWER

In this section, we will focus on introducing the atWER method, which mainly includes two steps: the alignment of the hypothesis and label, and the calculation of atWER.

We utilized the U2 Attention rescoring method [10] to recognize speech after VAD segmentation. Meanwhile, the CTC algorithm effectively addressed the alignment issue in sequence labeling by introducing a blank symbol and dynamic programming techniques.

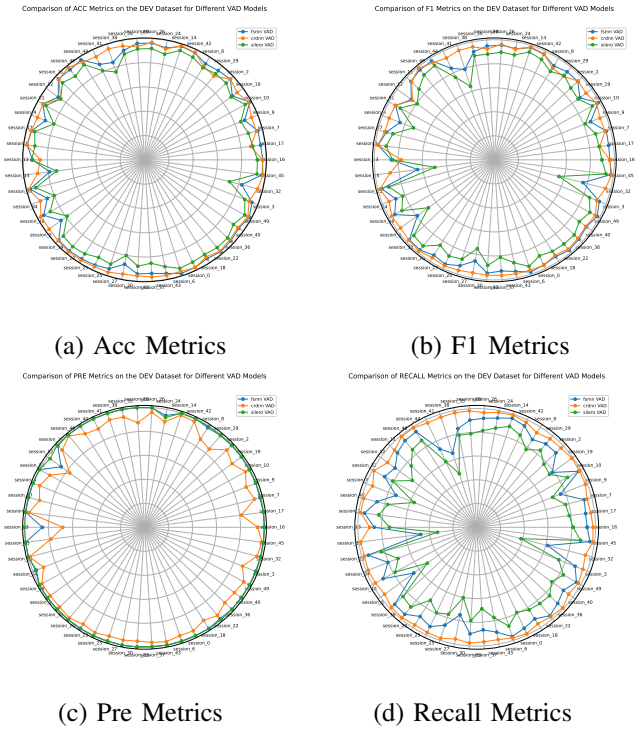
Specifically, the CTC algorithm initially procures the n-best candidate ensemble in a sequential stream, which is then subjected to the attention-based rescoring mechanism. This mechanism refines the n-best candidates through an apportioned scoring methodology, culminating in the derivation of a hypothesis that encapsulates sequence alignment data. Ultimately, the alignment outcomes at the character tier, predicated on the aforementioned hypothesis, are harmonized with the reference label of the Oracle VAD. The token is deemed credible when both the System VAD and the Oracle VAD are concurrently operational.The formula for calculating atWER is as follows:

$$atWER = \frac{S_c + D_c + I_c}{N_c} \tag{8}$$

where S_c represents the number of SUB errors among the credible tokens, D_c represents the number of DEL errors among the credible tokens, I_c represents the number of INS errors among the credible tokens, and N_c represents the total number of words or characters in the reference text within the credible tokens. The lower the atWER value, the better the performance of the ASR system.

VAD Model	Acc	F1	Recall	Pre
FSMN VAD	0.92	0.9	0.84	0.99
CRDNN VAD	0.73	0.71	0.73	0.7
Silero VAD	0.89	0.85	0.75	1.0

TABLE I
COMPARISON OF DIFFERENT SYSTEM VADS.



(a) Acc Metrics (b) F1 Metrics

(c) Pre Metrics (d) Recall Metrics

Fig. 1. Radar chart of the performance of different systems in different sessions on the DEV dataset.

IV. EXPERIMENT

This section delineates the specific contents of the experiments conducted in this study.

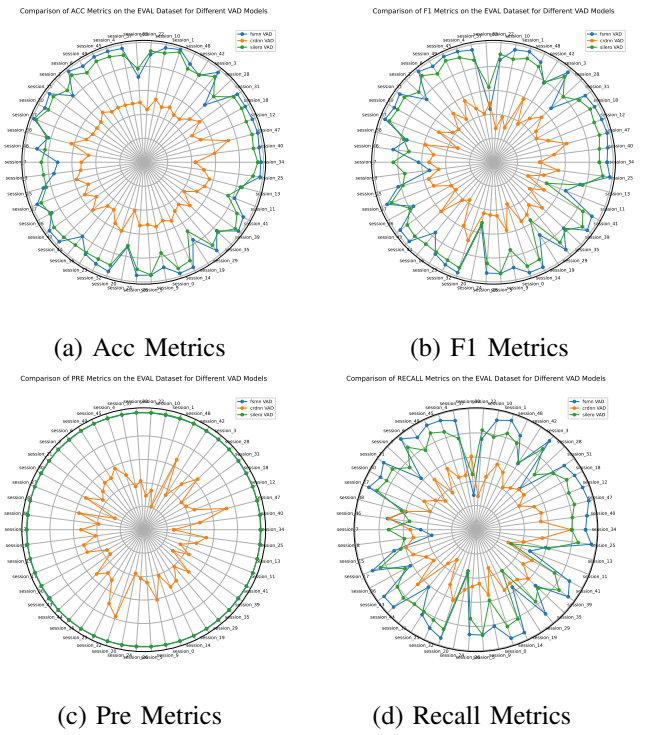
A. VAD System Evaluation

This section assesses the performance of three VAD models—the FSMN [4], CRDNN, and Silero [11] VAD models—on the LibriParty [1] dataset using the metrics introduced in Section III-E.

Initially, the experiment evaluated the performance of three VAD models using the Development (DEV) dataset, calculating metrics such as Accuracy, F1-score, Precision, and Recall. The results were compared using radar charts, as depicted in Figure 1. Figures 1.a and 1.b illustrate that the CRDNN VAD model outperforms the other two models in overall performance. However, Figure 1.c reveals a higher false positive rate for the CRDNN VAD model, indicating a propensity to misclassify non-speech segments as speech.

Subsequently, the experiment assessed the models' scores on the Evaluation (EVAL) dataset, as shown in Figure 2.

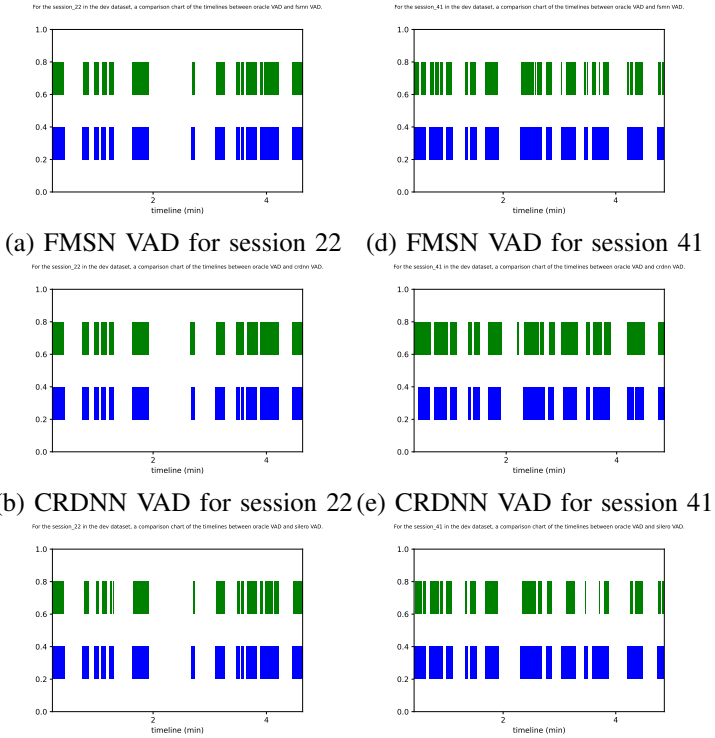
[1]https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriParty



(a) Acc Metrics (b) F1 Metrics

(c) Pre Metrics (d) Recall Metrics

Fig. 2. Radar chart of the performance of different systems in different sessions on the EVAL dataset.



(a) FMSN VAD for session 22 (d) FMSN VAD for session 41

(b) CRDNN VAD for session 22 (e) CRDNN VAD for session 41

(c) Silero VAD for session 22 (f) Silero VAD for session 41

Fig. 3. Sampled dev_session_22 and dev_session_41 from the DEV dataset, the alignment results of the Oracle VAD (blue) and the System VAD (green) on the timeline.

The findings diverged from the previous experiment, with the CRDNN VAD model significantly underperforming compared to the other models. Figures 2.a and 2.b demonstrate that the FSMN VAD model achieved the best performance, closely followed by the Silero VAD model. The latter's slightly lower performance is primarily attributed to its tendency to misclassify speech segments as non-speech, as illustrated in Figure 2.d.

The paper further illustrates the performance of the VAD models through visual representations of sample results in Figures 3 and 4. These figures delineate the manual segmentation outcomes in blue and the model predictions in green. Notably, in Figure 3.b and 3.e, the CRDNN VAD model is observed to misclassify non-speech frames as speech. Conversely, Figures 3.a, 3.c, 3.d, and 3.f indicate a tendency for both the FSMN and Silero VAD models to classify speech frames as non-speech, a behavior that is also evident in Figure 4. The sample analyses corroborate the findings from the radar chart assessments, reinforcing the consistency of the models' performance across different evaluation methods.

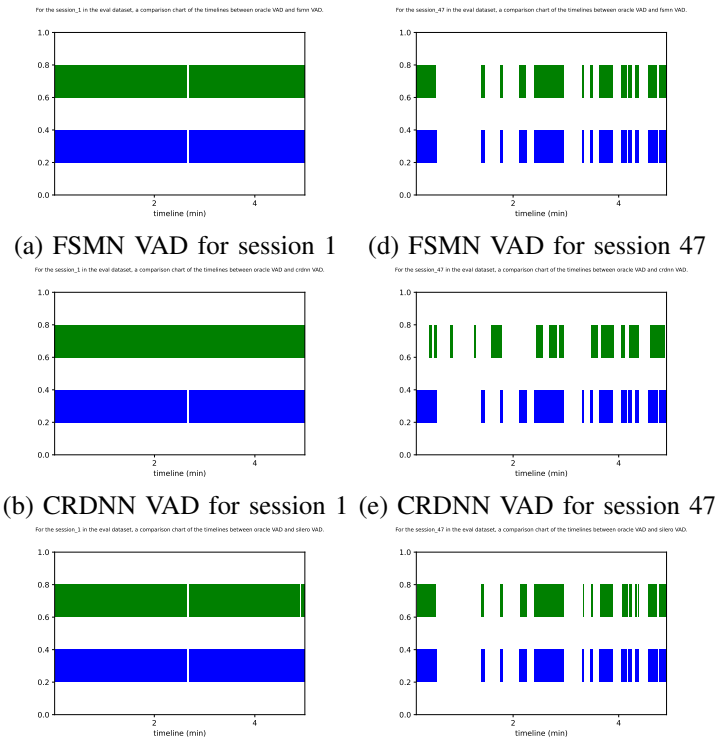
This experiment also systematically documents the comparative performance of the three VAD models across two datasets in Table I. The FSMN VAD model leads in performance, closely followed by the Silero VAD model, with the CRDNN VAD model ranking third. A minor performance discrepancy exists between the FSMN and Silero models, particularly in the Recall metric. This discrepancy aligns with our observations that the Silero VAD model is more prone to false negatives.

B. *VAD cascading ASR System Evaluation*

To assess the impact of cascading VAD systems on ASR systems, we integrated the three VAD systems from Section IV-A with the Paraformer [12] model and the SenseVoice [13] model, respectively, and measured the evaluation outcomes based on WER and atWER. The results are presented in Table II and III. In accordance with the principles of atWER, this experiment aligned the System VAD outcomes with the VAD labels, evaluating ASR results only under true positive conditions. It was observed that all three types of ASR errors were reduced, with the most significant decrease observed in deletion errors.

We argue that the atWER metric excludes false positive cases, where System VAD misclassifies non-speech as speech, from the evaluation. Consequently, the content erroneously identified as speech does not influence the atWER assessment. Additionally, atWER does not consider false negatives, where speech is misclassified as non-speech, resulting in omitted content in ASR recognition. We posit that these misclassifications are beyond the purview of ASR system enhancements and should not be included in ASR evaluation metrics. As a result, in our statistical analysis, we have excluded insertion and deletion errors associated with these scenarios.

Furthermore, we observed an unexpected increase in substitution errors, which seemingly conflicts with the anticipated effects of atWER. This anomaly can be explained by the



(a) FSMN VAD for session 1 (d) FSMN VAD for session 47

(b) CRDNN VAD for session 1 (e) CRDNN VAD for session 47

(c) Silero VAD for session 1 (f) Silero VAD for session 47

Fig. 4. Sampled eval_session_1 and eval_session_47 from the EVAL dataset, the alignment results of the Oracle VAD (blue) and the System VAD (green) on the timeline.

Exp	Evaluation Method	WER	SUB	INS	DEL
E1: FSMN VAD	WER	40.0	2121	198	6212
	atWER	38.88	2122	198	5821
E2: CRDNN VAD	WER	31.75	2250	304	4727
	atWER	30.24	2240	208	4379
E3: Silero VAD	WER	40.02	2695	202	6281
	atWER	38.83	2685	191	5861

TABLE II
WER RESULTS AND ATWER RESULTS OF THE VAD-CASCADED ASR SYSTEM EVALUATION ON THE DEV DATASET.

fact that atWER, by eliminating insertion errors, paradoxically generates new deletion errors. These deletions, in turn, can merge with existing errors, transforming them into a single substitution error.

V. CONCLUSIONS

This research presented highlights the significant correlation between the effectiveness of VAD and the types of errors encountered in ASR systems, specifically insertion and deletion errors. To address the challenges posed by insertion and deletion errors, this paper introduces the atWER metric calculation algorithm. This novel approach aims to resolve the alignment discrepancies between VAD labels and classification results, thereby enhancing the reliability and accuracy of both VAD and ASR systems. By improving the accuracy of speech detection and addressing the alignment issues between VAD and ASR, we can significantly enhance the user experience

Exp	Evaluation Method	WER	SUB	INS	DEL
E1: FSMN VAD	WER	34.5	2462	285	5165
	atWER	33.93	2444	215	5092
E2: CRDNN VAD	WER	67.45	1482	76	12828
	atWER	57.53	1359	56	7990
E3: Silero VAD	WER	42.38	2678	196	6164
	atWER	39.67	2681	196	5201

TABLE III
WER RESULTS AND ATWER RESULTS OF THE VAD-CASCADED ASR SYSTEM EVALUATION ON THE EVAL DATASET.

and the practical utility of speech recognition technologies in various applications.

ACKNOWLEDGMENT

This work was supported by Key Laboratory of MIIT for Intelligent Products Testing and Reliability 2023 Key Laboratory Open Project Fund(No.CEPREI2023-01).

REFERENCES

[1] S. Tong, N. Chen, Y. Qian, and K. Yu, "Evaluating vad for automatic speech recognition," in *2014 12th International Conference on Signal Processing (ICSP)*, IEEE, 2014, pp. 2308–2314.

[2] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[3] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approaches for vad," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 5695–5699.

[4] S. Zhang, M. Lei, Z. Yan, and L. Dai, "Deep-fsmn for large vocabulary continuous speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5869–5873.

[5] M. Li, Y. Xia, and F. Lin, "Incorporating vad into asr system by multi-task learning," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2022, pp. 160–164.

[6] G. Gelly and J.-L. Gauvain, "Minimum word error training of rnn-based voice activity detection.," in *INTERSPEECH*, 2015, pp. 2650–2654.

[7] S. Novitasari, T. Fukuda, and G. Kurata, "Improving asr robustness in noisy condition through vad integration.," in *INTERSPEECH*, 2022, pp. 3784–3788.

[8] O. Ichikawa, K. Nakano, T. Nakayama, and H. Shirouzu, "Multi-channel vad for transcription of group discussion.," in *Interspeech*, 2021, pp. 336–340.

[9] P. Paterlini-Brechot and N. L. Benali, "Circulating tumor cells (ctc) detection: Clinical impact and future directions," *Cancer letters*, vol. 253, no. 2, pp. 180–204, 2007.

[10] Z. Yao, D. Wu, X. Wang, *et al.*, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*, IEEE, Brno, Czech Republic, 2021.

[11] S. Team, "Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," *Retrieved March*, vol. 31, p. 2023, 2021.

[12] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in *INTERSPEECH*, 2022.

[13] T. SpeechTeam, "Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms," *arXiv preprint arXiv:2407.04051*, 2024.

A Discrete-Valued Signal Estimation by Nonconvex Enhancement of SOAV with cLiGME Model

Satoshi Shoji, Wataru Yata, Keita Kume and Isao Yamada
Dept. of Information and Communication Engineering, Institute of Science Tokyo
E-mail: {shoji, yata, kume, isao}@sp.ict.e.titech.ac.jp

***Abstract*—In this paper, for the discrete-valued signal estimation, we propose a regularized least squares model but with a nonconvex enhancement of the so-called SOAV convex regularizer. To design more contrastive regularizers whose minima correspond to desired discrete values, we propose a class of nonconvex functions with Generalized Moreau Enhancement (GME) of the weighted ℓ_1-norm. Promisingly, by tuning properly the design parameters of the proposed GME regularizers, (i) we can make the nonconvexly-regularized least squares model convex; and (ii) we can use an iterative algorithm for finding a global minimizer of the proposed model. We also propose a pair of simple technical improvements, of the proposed algorithm, called respectively a generalized superiorization and an iterative reweighting. Numerical experiments demonstrate the effectiveness of the proposed model and algorithms in a scenario of MIMO signal detection.**

I. Introduction

Many tasks in signal processing, including digital communication and discrete-valued control [1]–[7], have been formulated as the following discrete-valued estimation problem:

Problem 1 (A discrete-valued estimation problem)**.**

$$\text{Find } \mathbf{x}^\star \in \mathfrak{D} \text{ such that } \mathbf{y} = \mathbf{A}\mathbf{x}^\star + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathfrak{D} \coloneqq \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{L^N}\} = \mathfrak{A}^N \coloneqq \{a_1, a_2, \ldots, a_L\}^N \subset \mathbb{R}^N$, $\mathbf{y} \in \mathbb{R}^M$ is an observed vector, $\mathbf{A} \in \mathbb{R}^{M\times N}$ is a known matrix, and $\boldsymbol{\varepsilon} \in \mathbb{R}^M$ is noise (Note: the complex version of this problem can also be formulated as Problem 1 essentially via simple $\mathbb{C} \rightleftarrows \mathbb{R}^2$ translation (see Appendix A)).

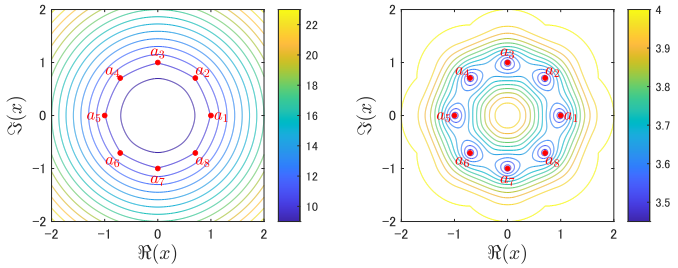
Problem 1 is a special instance of the mixed integer programming [8], but a direct application of naive solvers for the mixed integer programming leads to exponential computational complexity in N. From a practical viewpoint, continuous optimization approaches [9]–[14] have been utilized for Problem 1 as computationally efficient alternatives. For example, Problem 1 has been tackled with projection of a tentative estimate, say $\mathbf{x}^\diamond \in \mathbb{R}^N$, onto $\mathfrak{D}$ after solving a relaxed continuous optimization problem:

Scheme 1 (A scheme for (1) via regularized least squares)**.**
Step 1:

$$\text{Find } \mathbf{x}^\diamond \in \operatorname*{argmin}_{\mathbf{x}\in\widetilde{\mathfrak{D}}\subset\mathbb{R}^N} J_\Theta(\mathbf{x}) \coloneqq \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu\Theta(\mathbf{x}), \tag{2}$$

where the constraint set $\widetilde{\mathfrak{D}} \supset \mathfrak{D}$ is chosen usually as a connected subset (e.g., the convex hull of $\mathfrak{D}$) of $\mathbb{R}^N$, $\Theta : \mathbb{R}^N \to \mathbb{R}$ is a regularizer, and $\mu > 0$ is a regularization parameter.

This work was supported by JSPS Grants-in-Aid (19H04134,23KJ0945) and by JST SICORP (JPMJSC20C6).



(a) (b)

Fig. 1: Illustrations of the values of Θ in the case where $\mathfrak{D} = \mathfrak{A} \coloneqq \{a_l \coloneqq \exp[j(l-1)\pi/4] \mid l = 1, 2, \ldots, 8 =: L\} \subset \mathbb{C} \equiv \mathbb{R}^2$, $\omega_{l,1} = 1/8$ $(l = 1, 2, \ldots, 8)$, and $N = 1$. (a) $\Theta^{\langle 1\rangle}_{\text{SOAV}}(\mathbf{x})$, (b) a proposed nonconvex enhancement Θ_{GME} of $\Theta^{\langle 1\rangle}_{\text{SOAV}}$ by GME (5) with $\mathbf{B}^{\langle l\rangle} = \mathbf{I}$ $(l = 1, 2, \ldots, 8)$.

Step 2: With $\mathbf{x}^\diamond \in \widetilde{\mathfrak{D}} \subset \mathbb{R}^N$, compute the final estimate of $\mathbf{x}^\star \in \mathfrak{D}$ in (1) as

$$\mathbf{x}^\natural = P_{\mathfrak{D}}(\mathbf{x}^\diamond) \in \operatorname*{argmin}_{\mathbf{s}\in\mathfrak{D}} \|\mathbf{s} - \mathbf{x}^\diamond\|_2,$$

where $P_{\mathfrak{D}} : \mathbb{R}^N \to \mathfrak{D} : \mathbf{x} \mapsto P_{\mathfrak{D}}(\mathbf{x})$ is defined to choose randomly one of nearest vectors from $\mathbf{x}$.

In (2), the first term $\frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is called a data-fidelity term evaluating the consistency with the linear regression model in (1), while the second term $\mu\Theta : \mathbb{R}^N \to \mathbb{R}$ is called a regularization term designed strategically based on a prior knowledge on $\mathbf{x}^\star$.

For achieving an acceptable estimation of a signal in Problem 1, various prior knowledge, e.g., statistical properties, has been exploited for designing Θ in (2). For example, in a scenario of MIMO signal detection [15], the regularizer Θ in (2) has been found, e.g., as $\Theta = \|\cdot\|_p^p$ $(p = 1, 2)$ [9], [10] and $\Theta = \|\cdot\|_1 \circ D$ [11], where D is a first order difference operator.

For the relaxed continuous constrained optimization problem in (2), it would be desired for the regularizer Θ to penalize any point not in $\mathfrak{D}$. Along this regularization strategy, the so-called SOAV function [13], [14]

$$\Theta^{\langle 1\rangle}_{\text{SOAV}}(\mathbf{x}) \coloneqq \sum_{l=1}^{L} \|\mathbf{x} - a_l\mathbf{1}\|_{\boldsymbol{\omega}_l,1} \coloneqq \sum_{l=1}^{L}\sum_{n=1}^{N} \omega_{l,n}|x_n - a_l|, \tag{3}$$

has been used with weighting vectors $\boldsymbol{\omega}_l \coloneqq [\omega_{l,1}, \omega_{l,2}, \ldots, \omega_{l,N}]^\top \in \mathbb{R}^N_+$ $(l = 1, 2, \ldots, L)$ satisfying $\sum_{l=1}^{L} \omega_{l,n} = 1$ $(n = 1, 2, \ldots, N)$, where

$\|\mathbf{x}\|_{\boldsymbol{\omega}_l,1} \coloneqq \sum_{n=1}^{N} \omega_{l,n}|x_n|$ stands for the weighted ℓ_1-norm associated with $\boldsymbol{\omega}_l$.

Indeed, the weighted SOAV (W-SOAV) model [14] used (2) with $\Theta = \Theta_{\mathrm{SOAV}}^{\langle 1\rangle}$ and $\widetilde{D} \coloneqq \mathbb{R}^N$ for MIMO signal detection. Clearly, in this case, the model (2) is a convex model thanks to the convexity of $\Theta_{\mathrm{SOAV}}^{\langle 1\rangle}$, and therefore, a solution of (2) can be approximated iteratively by a convex optimization solver [14]. However, Fig. 1 (a) suggests that penalization by $\Theta_{\mathrm{SOAV}}^{\langle 1\rangle}$ is not contrastive enough for use as Θ in (2) because any point $\mathbf{s}_q \in \mathfrak{D} \subset \mathbb{C}^1$ is never unique minimizer of $\Theta_{\mathrm{SOAV}}^{\langle 1\rangle}$ over any neighborhood of $\mathbf{s}_q$. More contrastive regularizer than $\Theta_{\mathrm{SOAV}}^{\langle 1\rangle}$ has been certainly desired for use in (2) of Step 1 of Scheme 1 (see, e.g., Fig. 1 (b) and Fig. 2).

We are interested in the following natural questions:

(Q1) Can we design a class of functions that contains fairly contrastive regularizers Θ for use in (2) ?

(Q2) Can we choose any reasonable function Θ, from such a function class, which is tractable for minimization of J_Θ ?

(Q1) has been examined in special cases, e.g., $\Theta_{\mathrm{SOAV}}^{\langle p\rangle}(\mathbf{x}) \coloneqq \sum_{l=1}^{L}\|\mathbf{x} - a_l\mathbf{1}\|_{\boldsymbol{\omega}_l,p}$ $(0 \le p < 1)$ [16] as nonconvex variants of (3). However, any algorithm, of guaranteed to convergence to a global minimizer of $J_{\Theta_{\mathrm{SOAV}}^{\langle p\rangle}}$, has not yet been established mainly because of the severe nonconvexity of $J_{\Theta_{\mathrm{SOAV}}^{\langle p\rangle}}$. This situation tells us that computational tractability of (2) must be considered carefully from the beginning, i.e., (Q1) and (Q2) should be considered simultaneously.

In this paper, we present a positive answer to these questions

i) by designing a function class as

$$\Theta_{\mathrm{GME}}(\mathbf{x}) \coloneqq \sum_{l=1}^{L}(\|\cdot\|_{\boldsymbol{\omega}_l,1})_{\mathbf{B}^{\langle l\rangle}}(\mathbf{x} - a_l\mathbf{1}), \tag{4}$$

where $(\|\cdot\|_{\boldsymbol{\omega}_l,1})_{\mathbf{B}^{\langle l\rangle}}$ is a nonconvex enhancement (called *Generalized Moreau Enhancement (GME)*) of $\|\cdot\|_{\boldsymbol{\omega}_l,1}$ with a tunable matrix $\mathbf{B}^{\langle l\rangle}$ (see (5)) (Note: Θ_{GME} reproduces $\Theta_{\mathrm{SOAV}}^{\langle 1\rangle}$ with $\mathbf{B}^{\langle l\rangle} = \mathbf{O}$ (zero matrix));

ii) by exemplifying a fairly contrastive function (see Fig. 1 (b)) in the proposed class of Θ_{GME};

iii) by presenting a choice of $\mathbf{B}^{\langle l\rangle}$ $(l = 1, 2, \ldots, L)$ (see (9) and (10)) which achieves the overall convexity of $J_{\Theta_{\mathrm{GME}}}$;

iv) by proposing an iterative algorithm (see Algorithm 1), based on a relaxation of even symmetric condition [17, Problem 1] required for the so-called seed convex function in cLiGME model (see Section II), with guaranteed to convergence to a global minimizer of $J_{\Theta_{\mathrm{GME}}}$ over $\widetilde{\mathfrak{D}}$ under the overall convexity condition.

Indeed, via numerical experiments in a scenario of MIMO signal detection, we demonstrate the effectiveness of the proposed regularizer Θ_{GME} in the model (2).

We also propose a pair of simple technical improvements for the proposed iterative algorithm in Step 1 of Scheme 1 by exploiting adaptively the discrete information regarding $\mathfrak{D}$. More precisely, we propose (i) to use a strategic perturbation (14) to move the estimate closer to $\mathfrak{D}$ at each iteration $k \in \mathbb{N}$ (this idea is inspired by *superiorization* [18], [19]), and (ii) to update the weights $\omega_{l,n}$ $(l = 1, 2, \ldots, L; n = 1, 2, \ldots, N)$ in (4) adaptively (see (15)) by assigning larger weight to $\omega_{l,n}$ for (l, n) such that the distance between $a_l \in \mathfrak{A}$ and nth coordinate $x_n \in \mathbb{R}$ of the latest estimate $\mathbf{x} \in \mathbb{R}^N$ is smaller (this idea is inspired by iterative reweighting of SOAV [14]). Experimental results demonstrate that these simple techniques improve numerical performance of the proposed Algorithm 1.

Notation. $\mathbb{N}, \mathbb{R}, \mathbb{R}_+, \mathbb{R}_{++}$ and $\mathbb{C}$ denote respectively the set of all nonnegative integers, all real numbers, all nonnegative real numbers, all positive real numbers and all complex numbers (j stands for the imaginary unit, and $\Re(\cdot)$ and $\Im(\cdot)$ stand respectively for real and imaginary parts).

Let $\mathcal{H}, \mathcal{K}$ be finite dimensional real Hilbert spaces. The set of all *proper lower semicontinuous convex* functions[1] defined on $\mathcal{H}$ is denoted by $\Gamma_0(\mathcal{H})$. $f \in \Gamma_0(\mathcal{H})$ is said to be *prox-friendly* if $\mathrm{Prox}_{\gamma f} : \mathcal{H} \to \mathcal{H} : x \mapsto \mathrm{argmin}_{y\in\mathcal{H}}[f(y) + \frac{1}{2\gamma}\|y - x\|_{\mathcal{H}}^2]$ is available as a computable operator for any $\gamma \in \mathbb{R}_{++}$. A closed convex set $C \subset \mathcal{H}$ is said to be *simple* if the metric projection $P_C : \mathcal{H} \to \mathcal{H} : x \mapsto \mathrm{argmin}_{y\in\mathcal{H}}\|x - y\|_{\mathcal{H}}$ is available as a computable operator. The set of bounded linear operators from $\mathcal{H}$ to $\mathcal{K}$ is denoted by $\mathcal{B}(\mathcal{H}, \mathcal{K})$. For $\mathfrak{L} \in \mathcal{B}(\mathcal{H}, \mathcal{K})$, $\mathfrak{L}^* \in \mathcal{B}(\mathcal{K}, \mathcal{H})$ denotes the adjoint operator of $\mathfrak{L}$ (i.e., $(\forall (x, y) \in \mathcal{H} \times \mathcal{K})$ $\langle \mathfrak{L}x, y\rangle_{\mathcal{K}} = \langle x, \mathfrak{L}^* y\rangle_{\mathcal{H}}$).

For discussion in Euclidean space, we use boldface letters to express vectors and general font letters to represent scalars. For a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbf{X}^\top \in \mathbb{R}^{n\times m}$ denotes the transpose of $\mathbf{X}$. The symbols $\mathbf{I}$, $\mathbf{O}$ and $\mathbf{1}$ respectively stand for the identity matrix, the zero matrix and the all one vector.

II. Brief Introduction to cLiGME

Problem 2 (cLiGME model [20], [21])**.** Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}_l$ $(l = 1, 2, \ldots, L)$, $\widetilde{\mathcal{Z}}_l$ $(l = 1, 2, \ldots, L)$ and $\mathfrak{Z}$ be finite dimensional real Hilbert spaces. Suppose that (a) $A \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$, $y \in \mathcal{Y}$ and $\mu > 0$; (b) for each $l = 1, 2, \ldots, L$, $B^{\langle l\rangle} \in \mathcal{B}(\mathcal{Z}_l, \widetilde{\mathcal{Z}}_l)$, $\mathfrak{L}^{\langle l\rangle} \in \mathcal{B}(\mathcal{X}, \mathcal{Z}_l)$ and $\mu_l > 0$; (c) $C(\subset \mathfrak{Z})$ is a nonempty simple closed convex set and $\mathfrak{C} \in \mathcal{B}(\mathcal{X}, \mathfrak{Z})$; (d) $\Psi^{\langle l\rangle} \in \Gamma_0(\mathcal{Z}_l)$ is coercive, $\mathrm{dom}\,\Psi^{\langle l\rangle} = \mathcal{Z}_l$, even symmetry (i.e., $\Psi^{\langle l\rangle} \circ (-\mathrm{Id}) = \Psi^{\langle l\rangle}$), and prox-friendly. Then

i) with a tunable matrix $B^{\langle l\rangle} \in \mathcal{B}(\mathcal{Z}_l\widetilde{\mathcal{Z}}_l)$, the *Generalized Moreau Enhancement (GME)* of $\Psi^{\langle l\rangle}$ is defined by

$$\Psi_{B^{\langle l\rangle}}^{\langle l\rangle}(\cdot) \coloneqq \Psi^{\langle l\rangle}(\cdot) - \min_{v\in\mathcal{Z}_l}\Big[\Psi^{\langle l\rangle}(v) + \frac{1}{2}\|B^{\langle l\rangle}(\cdot - v)\|_{\widetilde{\mathcal{Z}}_l}^2\Big]; \tag{5}$$

ii) the *constrained LiGME (cLiGME)* model is given as

$$\text{Find } x^\diamond \in \underset{\mathfrak{C}x\in C}{\mathrm{argmin}}\frac{1}{2}\|y - Ax\|_{\mathcal{Y}}^2 + \mu\sum_{l=1}^{L}\mu_l\Psi_{B^{\langle l\rangle}}^{\langle l\rangle} \circ \mathfrak{L}^{\langle l\rangle}(x). \tag{6}$$

The regularizer $\Psi_{B^{\langle l\rangle}}^{\langle l\rangle}$ was proposed originally in [17] as an extension of the so-called *GMC penalty* in [22] mainly for the sparsity aware estimation. Furthermore, although $\Psi_{B^{\langle l\rangle}}^{\langle l\rangle}$ with $B^{\langle l\rangle} \neq O$ is nonconvex, the convexity of the cost function in (6) is achieved by a strategic tuning of GME matrices $B^{\langle l\rangle}$ $(l = 1, 2, \ldots, L)$ (see, e.g., [23]).

[1] A function $f : \mathcal{H} \to (-\infty, \infty]$ is (i) proper if $\mathrm{dom}(f) \coloneqq \{x \in \mathcal{H} \mid f(x) < \infty\} \neq \emptyset$, (ii) lower semicontinuous if $\{x \in \mathcal{H} \mid f(x) \le \alpha\}$ is closed for $\forall \alpha \in \mathbb{R}$, (iii) convex if $f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y)$ for $\forall x, y \in \mathcal{H}$, $0 < \theta < 1$.

III. Proposed regularizer and algorithm for discrete-valued signal estimation

A. Proposed GME regularizer and cLiGME algorithm

In this section, we propose a class of regularizers Θ_{GME} in (4), which is designed with the GME functions (5) of $\|\cdot\|_{\boldsymbol{\omega}_l,1}$ $(l=1,2,\dots,L)$. Indeed, as seen from Fig. 1 (b) (see also Fig. 2), the class of the proposed regularizers Θ_{GME} contains fairly contrastive functions. Moreover, if $\mathbf{B}^{\langle l\rangle}=\mathbf{O}$ $(l=1,2,\dots,L)$, then $\Theta_{\mathrm{GME}}=\Theta^{\langle 1\rangle}_{\mathrm{SOAV}}$ holds. By using Θ_{GME} (see (4)) in Step 1 of Scheme 1, we propose the following model:

$$\text{Find } \mathbf{x}^{\diamond}\in \underset{\mathbf{x}\in\widetilde{\mathfrak{D}}\subset\mathbb{R}^N}{\operatorname{argmin}} \underbrace{\frac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2+\mu\underbrace{\sum_{l=1}^{L}(\|\cdot\|_{\boldsymbol{\omega}_l,1})_{\mathbf{B}^{\langle l\rangle}}(\mathbf{x}-a_l\mathbf{1})}_{=\Theta_{\mathrm{GME}}(\mathbf{x})}}_{=J_{\Theta_{\mathrm{GME}}}(\mathbf{x})}, \tag{7}$$

where $\widetilde{\mathfrak{D}}\supset\mathfrak{D}$ is a closed convex set.

In order to solve (7), we consider the following problem which contains (7) as its special instance (see Remark 1).

Problem 3. Let $\mathcal{X}$, $\mathcal{Y}$ and $\widetilde{\mathcal{Z}}_l$ $(l=1,2,\dots,L)$ be finite dimensional real Hilbert spaces. Suppose that (a) $A\in\mathcal{B}(\mathcal{X},\mathcal{Y})$, $y\in\mathcal{Y}$ and $\mu>0$; (b) for each $l=1,2,\dots,L$, $B^{\langle l\rangle}\in\mathcal{B}(\mathcal{X},\widetilde{\mathcal{Z}}_l)$, $z^{\langle l\rangle}\in\mathcal{X}$ and $\mu_l>0$; (c) $C(\subset\mathcal{X})$ is a nonempty simple closed convex set; (d) $\Psi^{\langle l\rangle}\in\Gamma_0(\mathcal{X})$ is coercive, $\operatorname{dom}\Psi^{\langle l\rangle}=\mathcal{X}$, even symmetry, and prox-friendly. Then, consider

$$\text{Find } x^{\diamond}\in\underset{x\in C}{\operatorname{argmin}}\underbrace{\frac{1}{2}\|y-Ax\|_{\mathcal{Y}}^2+\mu\sum_{l=1}^{L}\mu_l\Psi^{\langle l\rangle}_{B^{\langle l\rangle}}(x-z^{\langle l\rangle})}_{=:J(x)}, \tag{8}$$

where $\Psi^{\langle l\rangle}_{B^{\langle l\rangle}}:\mathcal{X}\to\mathbb{R}$ is a GME function (5), with a tuning matrix $B^{\langle l\rangle}$, of a convex function $\Psi^{\langle l\rangle}$ $(l=1,2,\dots,L)$.

Remark 1. The model (7) is a special instance of Problem 3 with $\mathcal{X}=\mathbb{R}^N$, $\mathcal{Y}=\mathbb{R}^M$, $C=\widetilde{\mathfrak{D}}$, $\Psi^{\langle l\rangle}=\|\cdot\|_{\boldsymbol{\omega}_l,1}$ $(l=1,2,\dots,L)$, $z^{\langle l\rangle}=a_l\mathbf{1}$ $(l=1,2,\dots,L)$ and $\mu_l=1$ $(l=1,2,\dots,L)$.

By tuning properly the design parameters of the proposed GME regularizers, we can make the nonconvexly-regularized least squares model convex.

Fact 1 (Overall convexity condition [17] for (8))**.** *Consider Problem 3. Then J in* (8) *is convex if $(B^{\langle l\rangle})_{l=1}^L$ satisfy*

$$A^*A-\mu\sum_{l=1}^{L}\mu_l{B^{\langle l\rangle}}^*B^{\langle l\rangle}\ \textit{is positive semidefinite.} \tag{9}$$

For example, the following $B^{\langle l\rangle}$ [22] satisfy (9)*:*

$$B^{\langle l\rangle}:=\sqrt{\frac{\gamma_l}{\mu\mu_l}}A\in\mathcal{B}(\mathcal{X},\mathcal{Y})\quad(l=1,2,\dots,L), \tag{10}$$

where $\gamma_l\in\mathbb{R}_+$ $(l=1,2,\dots,L)$ are chosen to satisfy $\sum_{l=1}^L\gamma_l\le 1$.

In a special case where $z^{\langle l\rangle}=0\in\mathcal{X}$ $(l=1,2,\dots,L)$, the model (8) is reduced to the cLiGME model (6) with $\mathcal{Z}=\mathfrak{Z}=\mathcal{X}$, $\mathfrak{L}^{\langle l\rangle}=\mathrm{Id}$ $(l=1,2,\dots,L)$ and $\mathfrak{C}=\mathrm{Id}$. For this special case under the condition (9), we can find a global minimizer of (8), by cLiGME algorithm [20], [21].

In the following, we present an iterative algorithm applicable even to general cases $z^{\langle l\rangle}\in\mathcal{X}$ $(l=1,2,\dots,L)$. The proposed algorithm (12) in Theorem 1 is a variant of cLiGME algorithm [20], [21], [24] (see Remark 2), and Algorithm 1 illustrates a concrete expression of (12) (see also Remark 1).

Theorem 1 (A relaxation of cLiGME algorithm for Problem 3)**.** *Consider Problem 3 under the overall convexity condition* (9)*. Assume* $\operatorname{argmin}_{x\in C}J(x)\neq\emptyset$[2]*. Define the operator $T:\mathcal{H}:=\mathcal{X}\times(\mathcal{X})^L\times(\mathcal{X})^L\to\mathcal{H}:\big(x,(v^{\langle l\rangle})_{l=1}^L,(w^{\langle l\rangle})_{l=1}^L\big)\mapsto\big(\xi,(\zeta^{\langle l\rangle})_{l=1}^L,(\eta^{\langle l\rangle})_{l=1}^L\big)$ by*

$$\begin{cases}\xi:=P_C\big[\big(\mathrm{Id}-\frac{1}{\sigma}(A^*A-\mu\sum_{l=1}^L\mu_l{B^{\langle l\rangle}}^*B^{\langle l\rangle})\big)x\\ \qquad-\frac{\mu}{\sigma}\sum_{l=1}^L\big(\mu_l{B^{\langle l\rangle}}^*B^{\langle l\rangle}v^{\langle l\rangle}+w^{\langle l\rangle}\big)+\frac{1}{\sigma}A^*y\big],\\ \zeta^{\langle l\rangle}:=z^{\langle l\rangle}+\operatorname{Prox}_{\frac{\mu\mu_l}{\tau}\Psi^{\langle l\rangle}}\big[\frac{\mu\mu_l}{\tau}{B^{\langle l\rangle}}^*B^{\langle l\rangle}(2\xi-x)\\ \qquad+\big(\mathrm{Id}-\frac{\mu\mu_l}{\tau}{B^{\langle l\rangle}}^*B^{\langle l\rangle}\big)v^{\langle l\rangle}-z^{\langle l\rangle}\big],\\ \eta^{\langle l\rangle}:=\big(\mathrm{Id}-\operatorname{Prox}_{\mu_l\Psi^{\langle l\rangle}}\big)\big(2\xi-x+w^{\langle l\rangle}-z^{\langle l\rangle}\big),\end{cases}$$

where $(\sigma,\tau,\kappa)\in\mathbb{R}_{++}\times\mathbb{R}_{++}\times(1,\infty)$ is chosen to satisfy[3]

$$\begin{cases}(\sigma-\mu L)\mathrm{Id}-\frac{\kappa}{2}A^*A\ \textit{is positive definite},\\ \tau\ge(\frac{\kappa}{2}+\frac{2}{\kappa})\mu\max\{\mu_l\|B^{\langle l\rangle}\|_{\mathrm{OP}}^2\,|\,l=1,2,\dots,L\}.\end{cases} \tag{11}$$

Then, (a) T is an averaged nonexpansive operator[4] *by defining a proper inner product on $\mathcal{H}$ (see, e.g., [17], [21]), and (b) for any initial point $\mathfrak{u}_0\in\mathcal{H}$, the sequence $(\mathfrak{u}_k)_{k\in\mathbb{N}}\subset\mathcal{H}$ with $\mathfrak{u}_k:=\big(x_k,(v_k^{\langle l\rangle})_{l=1}^L,(w_k^{\langle l\rangle})_{l=1}^L\big)$ generated by the following* Picard-type fixed point iteration*:*

$$(k\in\mathbb{N})\quad\mathfrak{u}_{k+1}=T(\mathfrak{u}_k) \tag{12}$$

converges to a fixed point $\overline{\mathfrak{u}}=\big(\overline{x},(\overline{v}^{\langle l\rangle})_{l=1}^L,(\overline{w}^{\langle l\rangle})_{l=1}^L\big)\in\operatorname{Fix}(T):=\{\mathfrak{u}\in\mathcal{H}\mid T(\mathfrak{u})=\mathfrak{u}\}\subset\mathcal{H}$, where $\overline{x}\in\mathcal{X}$ enjoys the condition as a global minimizer $x^{\diamond}\in C$, in (8)*, of J over C.*

Remark 2. For Problem 3, the condition (9) and Theorem 1 are obtained by the following steps[5].

i) By defining $\widetilde{\Psi}^{\langle l\rangle}:=\Psi^{\langle l\rangle}(\cdot-z^{\langle l\rangle})$ $(l=1,2,\dots,L)$, we can check that $\widetilde{\Psi}^{\langle l\rangle}_{B^{\langle l\rangle}}=\Psi^{\langle l\rangle}_{B^{\langle l\rangle}}(\cdot-z^{\langle l\rangle})$.

ii) By using $\widetilde{\Psi}^{\langle l\rangle}$, we can express J in (8) equivalently as

$$J(x)=\frac{1}{2}\|y-Ax\|_{\mathcal{Y}}^2+\mu\sum_{l=1}^{L}\mu_l\widetilde{\Psi}^{\langle l\rangle}_{B^{\langle l\rangle}}(x). \tag{13}$$

---

[2] $\operatorname{argmin}_{x\in C}J(x)\neq\emptyset$ is guaranteed in many cases, e.g., if C is compact (not limited to this case).

[3] For example, choose $\kappa>1$ and compute (σ,τ) by

$$\begin{cases}\sigma:=\frac{\kappa}{2}\|A\|_{\mathrm{OP}}^2+\mu L+(\kappa-1),\\ \tau:=(\frac{\kappa}{2}+\frac{2}{\kappa})\mu\max\{\mu_l\|B^{\langle l\rangle}\|_{\mathrm{OP}}^2\mid l=1,2,\dots,L\}+(\kappa-1),\end{cases}$$

where $\|B^{\langle l\rangle}\|_{\mathrm{OP}}$ denotes the operator norm of $B^{\langle l\rangle}$ (i.e., $\|B^{\langle l\rangle}\|_{\mathrm{OP}}:=\sup_{x\in\mathcal{X},\|x\|_{\mathcal{X}}\le1}\|B^{\langle l\rangle}x\|_{\widetilde{\mathcal{Z}}_l}$).

[4] An operator $T:\mathcal{X}\to\mathcal{X}$ is said to be *nonexpansive* if $(\forall x,y\in\mathcal{X})\ \|T(x)-T(y)\|\le\|x-y\|$, in particular, ($\alpha$-) *averaged nonexpansive* if there exists $\alpha\in(0,1)$ and a nonexpansive operator $R:\mathcal{X}\to\mathcal{X}$ such that $T=(1-\alpha)\mathrm{Id}+\alpha R$.

[5] Even for Problem 2 with a replacement of the seed convex function $\Psi^{\langle l\rangle}$ by its shifted seed convex function $\widetilde{\Psi}^{\langle l\rangle}=\Psi^{\langle l\rangle}(\cdot-z^{\langle l\rangle})$, via an essentialy same discussion, we can derive the overall convexity condition of the objective function and an iterative algorithm with guaranteed convergence to a global minimizer.

Algorithm 1 A relaxation of cLiGME algorithm for (7)

1: Choose $\big(\mathbf{x}_0,(\mathbf{v}_0^{\langle l\rangle})_{l=1}^L,(\mathbf{w}_0^{\langle l\rangle})_{l=1}^L\big)\in\mathbb{R}^N\times(\mathbb{R}^N)^L\times(\mathbb{R}^N)^L$.
2: Choose $(\sigma,\tau,\kappa)\in\mathbb{R}_{++}\times\mathbb{R}_{++}\times(1,\infty)$ satisfying (11).
3: **for** $k=0,1,2,\ldots$ **do**
4: { Insert modification in Section III-B if necessary. }
5: $\mathbf{x}_{k+1}\leftarrow P_{\widetilde{\mathfrak{D}}}\big[\big(\mathbf{I}-\frac{1}{\sigma}(\mathbf{A}^\top\mathbf{A}-\mu\sum_{l=1}^L\mathbf{B}^{\langle l\rangle\top}\mathbf{B}^{\langle l\rangle})\big)\mathbf{x}_k$
$-\frac{\mu}{\sigma}\sum_{l=1}^L\big(\mathbf{B}^{\langle l\rangle\top}\mathbf{B}^{\langle l\rangle}\mathbf{v}_k^{\langle l\rangle}+\mathbf{w}_k^{\langle l\rangle}\big)+\frac{1}{\sigma}\mathbf{A}^\top\mathbf{y}\big]$
6: **for** $l=1,2,\ldots,L$ **do**
7: $\mathbf{v}_{k+1}^{\langle l\rangle}\leftarrow a_l\mathbf{1}+\mathrm{Prox}_{\frac{\mu}{\tau}\|\cdot\|_{\boldsymbol{\omega}_l,1}}\big[\frac{\mu}{\tau}\mathbf{B}^{\langle l\rangle\top}\mathbf{B}^{\langle l\rangle}(2\mathbf{x}_{k+1}-\mathbf{x}_k)$
$+\big(\mathbf{I}-\frac{\mu}{\tau}\mathbf{B}^{\langle l\rangle\top}\mathbf{B}^{\langle l\rangle}\big)\mathbf{v}_k^{\langle l\rangle}-a_l\mathbf{1}\big]$
8: $\mathbf{w}_{k+1}^{\langle l\rangle}\leftarrow\big(\mathrm{Id}-\mathrm{Prox}_{\|\cdot\|_{\boldsymbol{\omega}_l,1}}\big)\big(2\mathbf{x}_{k+1}-\mathbf{x}_k+\mathbf{w}_k^{\langle l\rangle}-a_l\mathbf{1}\big)$
9: **end for**
10: **end for**
11: **return** $\mathbf{x}_{k+1}$

iii) The overall convexity condition (9) for J in (8) is derived by applying [17, Proposition 1] to (13).
iv) Since $\widetilde{\Psi}^{\langle l\rangle}$ is not even symmetry in general, we cannot apply the cLiGME algorithm [20], [21] directly for (13). However, we can relax the even symmetric condition to $z^{\langle l\rangle}$-symmetric condition (i.e., $\widetilde{\Psi}^{\langle l\rangle}(z^{\langle l\rangle}+\cdot)=\widetilde{\Psi}^{\langle l\rangle}(z^{\langle l\rangle}-\cdot)$).

B. Two simple techniques for further improvement

For further improvement of Algorithm 1, we introduce two simple techniques in Algorithm 1 to exploit adaptively the discrete information regarding $\mathfrak{D}$.

1) Generalized superiorization of cLiGME algorithm: *Superiorization* [18], [19] is known as a technique for an iterative algorithm, e.g., Picard-type fixed point iteration, to reduce a certain objective cost by adding strategic bounded perturbations to updated estimate $\mathbf{x}_k$ $(k\in\mathbb{N})$.

We propose to incorporate a superiorization technique into Algorithm 1 in order to move the estimate closer to $\mathfrak{D}$ at each iteration. More precisely, we use a modification

$$\mathbf{x}_k\leftarrow\mathbf{x}_k+\beta_k\underbrace{(P_{\mathfrak{D}}-\mathrm{Id})(\mathbf{x}_k)}_{=:\mathbf{d}_k} \tag{14}$$

in line 4 of Algorithm 1, where $(\beta_k)_{k\in\mathbb{N}}\subset\mathbb{R}_+$, and $(\mathbf{d}_k)_{k\in\mathbb{N}}\subset\mathbb{R}^N$ is inspired by [19]. The global convergence guarantee is not violated even by the modification (14) if $(\beta_k)_{k\in\mathbb{N}}$ is summable and $(\mathbf{d}_k)_{k\in\mathbb{N}}$ is bounded (see Appendix B2). However, we dare to propose to use more general $(\beta_k)_{k\in\mathbb{N}}\subset\mathbb{R}_+$ which is not necessarily summable. We call such a modification *generalized superiorization*. As will be shown in numerical experiments (see Section IV), the proposed generalized superiorization is effective to guide the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to the discrete set $\mathfrak{D}$.

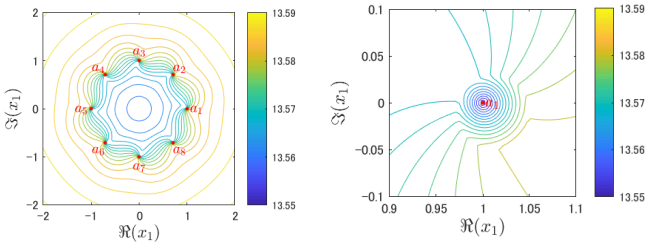
2) Iterative reweighting of cLiGME algorithm: The iterative reweighting technique, e.g., [25], has been used to enhance the effectiveness of the regularizer by updating the weights of the regularizer adaptively in an iterative algorithm. Iterative reweighting techniques are also used for Problem 1 [14], [26]. To utilize such a technique in Algorithm 1, we propose to set $\omega_{l,n}$ $(l=1,2,\ldots,L;\ n=1,2,\ldots,N)$ in the seed functions $\|\cdot\|_{\boldsymbol{\omega}_l,1}$ $(l=1,2,\ldots,L)$ adaptively by using the latest estimate $\mathbf{x}:=[x_1,x_2,\ldots,x_N]^\top$ as [26]

$$\omega_{l,n}=\frac{(|x_n-a_l|+\epsilon)^{-1}}{\sum_{l'=1}^L(|x_n-a_{l'}|+\epsilon)^{-1}}. \tag{15}$$

where $\epsilon>0$ is a small number. If $|x_n-a_l|$ is small, then the corresponding $\omega_{l,n}$ becomes large and x_n will be close to a_l. This iterative reweighting method can be realized by inserting

if k mod $K==0$ **then**
Update $\boldsymbol{\omega}_l=[\omega_{l,1},\omega_{l,2},\ldots,\omega_{l,N}]^\top$ $(l=1,2,\ldots,L)$ as (15) with $\mathbf{x}=\mathbf{x}_k$. (16)
end if

in line 4 of Algorithm 1, where $K\in\mathbb{N}\backslash\{0\}$ controls the frequency of reweighting.



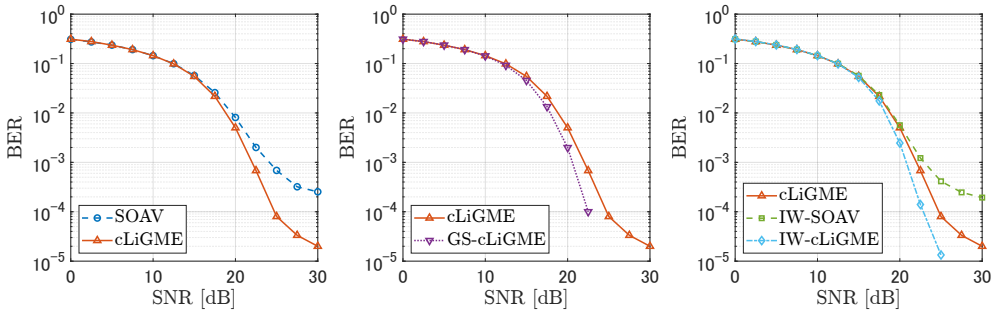
(a) Overall view (b) Enlarged view around a_1

Fig. 2: Illustrations of the values of $\Theta_{\mathrm{GME}}(\mathbf{x})$ in Eq.(4) with $\mathbf{B}^{\langle l\rangle}$ in (17) under the setting of Section IV. For visualization, we set $x_n=0$ $(n=2,3,\ldots,50)$.

IV. Numerical experiments

We conducted numerical experiments in a scenario of MIMO signal detection [12] with N-transmit antennas and M-receive antennas $(N=50,M=45)$ in 8PSK (phase shift keying) modulation with constellation set $\mathfrak{A}:=\{a_l:=\exp[j(l-1)\pi/4]\mid l=1,2,\ldots,8=:L\}\subset\mathbb{C}$. The task of this experiment is to estimate the transmitted signal $\mathbf{x}^\star\in\mathbb{C}^N$ from the received signal $\mathbf{y}=\mathbf{A}\mathbf{x}^\star+\boldsymbol{\varepsilon}\in\mathbb{C}^M$ with the channel matrix $\mathbf{A}\in\mathbb{C}^{M\times N}$ and a noise $\boldsymbol{\varepsilon}\in\mathbb{C}^M$. In this experiment, we chose randomly (i) $\mathbf{x}^\star\in\mathfrak{D}:=\mathfrak{A}^N$, (ii) $\mathbf{A}:=\sqrt{\mathbf{R}}\mathbf{G}\in\mathbb{C}^{M\times N}$, where each entry of $\mathbf{G}\in\mathbb{C}^{M\times N}$ was sampled from the complex gaussian distribution $\mathcal{CN}(0,1/M)$, and $\mathbf{R}\in\mathbb{R}^{M\times M}$ satisfies $(\mathbf{R})_{i,j}=0.5^{|i-j|}$ $(i=1,2,\ldots,M;\ j=1,2,\ldots,M)$, and (iii) each entry of $\boldsymbol{\varepsilon}\in\mathbb{C}^M$ was sampled from $\mathcal{CN}(0,\sigma_\varepsilon^2)$ with a variance $\sigma_\varepsilon^2>0$, which was chosen so that $10\log_{10}\frac{1}{\sigma_\varepsilon^2}$ achieved a given SNR (signal-to-noise ratio).

We consider to estimate $\mathbf{x}^\star\in\mathfrak{D}$ with Scheme 1 by employing the convex hull $\mathrm{conv}(\mathfrak{D})$ of $\mathfrak{D}$ as $\widetilde{\mathfrak{D}}$ in (2) via $\mathbb{C}\rightleftarrows\mathbb{R}^2$ translation (see Appendix A). In this experiment, we compared numerical performance of (i) the proposed cLiGME model (7), i.e., the model (2) with $\Theta=\Theta_{\mathrm{GME}}$ in (4), with that of (ii) the SOAV model [14], i.e., the model (2) with $\Theta=\Theta_{\mathrm{SOAV}}^{\langle 1\rangle}$ in (3).



(a) Basic cLiGME (b) Generalized superiorization (c) Iterative reweighting

Fig. 3: BER vs SNR (8PSK, $N = 50$, $M = 45$)

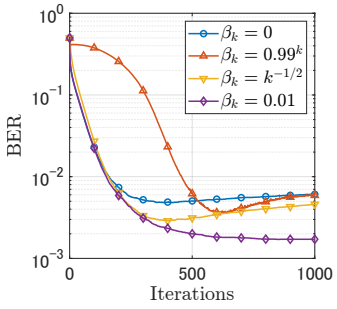


Fig. 4: BER vs iterations

For the cLiGME model, we used Algorithm 1 (denoted by 'cLiGME') by employing the following tuning matrices in (7)

$$\mathbf{B}^{\langle l\rangle} = \sqrt{0.99/\mu L}\widehat{\mathbf{A}} \quad (l = 1, 2, \ldots, L) \tag{17}$$

to achieve the overall convexity condition (9), where μ is a predetermined regularization parameter, and $\widehat{\mathbf{A}}$ is obtained via $\mathbb{C} \rightleftarrows \mathbb{R}^2$ translation (see (18)). Since SOAV model can be reduced to the cLiGME model (7) with $\mathbf{B}^{\langle l\rangle} = \mathbf{O}$ $(l = 1, 2, \ldots, L)$, we used Algorithm 1 (denoted by 'SOAV') with $\mathbf{B}^{\langle l\rangle} = \mathbf{O}$ $(l = 1, 2, \ldots, L)$ for the SOAV model. For both 'cLiGME' and 'SOAV', we employed the same (i) stepsize $(\sigma, \tau, \kappa) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times (1, \infty)$ as footnote 3 in Theorem 1 with $\kappa = 1.001$, and (ii) initial points $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{v}_0^{\langle l\rangle} = \mathbf{0}$ and $\mathbf{w}_0^{\langle l\rangle} = \mathbf{0}$ $(l = 1, 2, \ldots, L)$. Since $\widetilde{\mathfrak{D}}$ is compact, 'cLiGME' and 'SOAV' can find their global minimizers, respectively (see Theorem 1). Algorithm 1 were terminated when the iteration number k exceeded 500.

Before evaluating numerical performance, let us examine contrastiveness of Θ_{GME} used in these experiments. Fig. 2 shows the function values of Θ_{GME} in (4) designed with $\mathbf{B}^{\langle l\rangle}$ in (17) hence achieving the overall convexity condition (9). Each numerical value of $\Theta_{\text{GME}}(\mathbf{x})$ is computed with ISTA-type algorithm [27] (Note: the function value of Θ_{GME} is not required in the proposed Algorithm 1). As seen from Fig. 2 (b), we observe numerically that $\Theta_{\text{GME}}(\mathbf{x})$ is certainly contrastive as a regularizer for discrete-valued signal estimation because each constellation point in $\mathfrak{A}$ corresponds to a local minimizer of $\Theta_{\text{GME}}(\mathbf{x})$ as we expected.

As a performance metric, we adopted averaged BER (bit error rate) over 1,000 independent realizations of $(\mathbf{x}^\star, \mathbf{A}, \boldsymbol{\varepsilon})$. The parameter μ was chosen to achieve the lowest BER from the set $\{10^i \mid i = -10, -9, \ldots, 2\}$ at each SNR.

Fig. 3 (a) shows BER of 'SOAV' and 'cLiGME' at each SNR, where $\omega_{l,n} = 1/8$ $(l = 1, 2, \ldots, 8;\ n = 1, 2, \ldots, N)$ in (7) were fixed. From Fig. 3 (a), 'cLiGME' achieves lower BER than 'SOAV', which implies the effectiveness of the proposed contrastive nonconvex regularizer Θ_{GME} compared with the convex regularizer Θ_{SOAV}.

In the following, we verify the further performance improvements of 'cLiGME' by the proposed (i) generalized superiorization and (ii) iterative reweighting.

To examine the impact of choices of $(\beta_k)_{k\in\mathbb{N}}$ in generalized superiorization (14), we compared generalized superiorization of 'cLiGME' with (i) $\beta_k = 0$ (which reduces to the original 'cLiGME'), (ii) $\beta_k = 0.99^k$ ($(\beta_k)_{k\in\mathbb{N}}$ is summable), (iii) $\beta_k = k^{-1/2}$ ($(\beta_k)_{k\in\mathbb{N}}$ is nonsummable but $\beta_k \to 0$ $(k \to \infty)$), and (iv) $\beta_k = 0.01$. Fig. 4 shows history of BER achieved by generalized superiorization of 'cLiGME' with such $(\beta_k)_{k\in\mathbb{N}}$ in (14), where SNR $= 20$ dB, $\mu = 10^{-4}$ and $\omega_{l,n} = 1/8$ $(l = 1, 2, \ldots, 8;\ n = 1, 2, \ldots, N)$. From Fig. 4, $\beta_k = 0.01$ outperforms the others. Fig. 3 (b) shows BER, at each SNR, of 'cLiGME' and generalized superiorization of 'cLiGME' (denoted by 'GS-cLiGME') with $\beta_k = 0.01$. From Fig. 3 (b), we see 'GS-cLiGME' improves 'cLiGME'.

Fig. 3 (c) shows BER, at each SNR, of (i) 'cLiGME', (ii) iterative reweighting in (16) of 'cLiGME' (denoted by 'IW-cLiGME'), and (iii) iterative reweighting in (16) of 'SOAV' (denoted by 'IW-SOAV'), where the frequency period $K = 100$ in (16) was used (Note: the iterative reweighting of SOAV model was initially proposed [14], [26] with an ADMM-type algorithm). From Fig. 3 (c), 'IW-cLiGME' improves 'cLiGME', while even 'cLiGME' outperforms 'IW-SOAV'.

V. Conclusion

We proposed a class of fairly contrastive regularizers for discrete-valued estimation problems, and presented an iterative algorithm with guaranteed convergence to a global minimizer of the nonconvexly-regularized least squares model. We also proposed two simple techniques for performance improvements. The numerical experiments demonstrate that the proposed model and algorithm have a great potential for challenging discrete-valued signal estimation problem, and that two simple techniques successfully contribute to performance improvements of the proposed algorithm.

Appendix

A. $\mathbb{C} \rightleftarrows \mathbb{R}^2$ *translation*

Consider the complex version of Problem 1 where $\mathfrak{D}(\subset \mathbb{C}^N)$ is a finite set and $(\mathbf{x}^\star, \mathbf{y}, \mathbf{A}, \boldsymbol{\varepsilon}) \in \mathbb{C}^N \times \mathbb{C}^M \times \mathbb{C}^{M\times N} \times \mathbb{C}^M$. The $\mathbb{C} \rightleftarrows \mathbb{R}^2$ translation in this paper should be understood in the following sense:

$$\begin{aligned}
&\widehat{\mathfrak{D}} := \left\{ \begin{bmatrix} \Re(\mathbf{s}) \\ \Im(\mathbf{s}) \end{bmatrix} \in \mathbb{R}^{2N} \,\middle|\, \mathbf{s} \in \mathfrak{D} \right\}, \\
&\widehat{\mathbf{x}}^\star := \begin{bmatrix} \Re(\mathbf{x}^\star) \\ \Im(\mathbf{x}^\star) \end{bmatrix} \in \widehat{\mathfrak{D}} \subset \mathbb{R}^{2N}, \widehat{\mathbf{y}} := \begin{bmatrix} \Re(\mathbf{y}) \\ \Im(\mathbf{y}) \end{bmatrix} \in \mathbb{R}^{2M}, \\
&\widehat{\mathbf{A}} := \begin{bmatrix} \Re(\mathbf{A}) & -\Im(\mathbf{A}) \\ \Im(\mathbf{A}) & \Re(\mathbf{A}) \end{bmatrix} \in \mathbb{R}^{2M\times 2N}, \widehat{\boldsymbol{\varepsilon}} := \begin{bmatrix} \Re(\boldsymbol{\varepsilon}) \\ \Im(\boldsymbol{\varepsilon}) \end{bmatrix} \in \mathbb{R}^{2M}.
\end{aligned} \tag{18}$$

Clearly, via (18), we can translate $\mathbf{y} = \mathbf{A}\mathbf{x}^\star + \boldsymbol{\varepsilon}$ into $\widehat{\mathbf{y}} = \widehat{\mathbf{A}}\widehat{\mathbf{x}}^\star + \widehat{\boldsymbol{\varepsilon}}$, and can estimate $\widehat{\mathbf{x}}^\star$ by applying Algorithm 1 to the translated real model.

B. Bounded perturbation for Picard-type fixed point iteration

1) Picard iteration: Let $\mathcal{H}$ be a finite dimensional real Hilbert space. Suppose $T : \mathcal{H} \to \mathcal{H}$ is an averaged nonexpansive operator such that $\mathrm{Fix}(T) \coloneqq \{u \in \mathcal{H} \mid T(u) = u\} \neq \emptyset$. Then, a sequence $(u_k)_{k\in\mathbb{N}}$, generated by the so-called Picard iteration: $u_{k+1} = T(u_k)$ $(k \in \mathbb{N})$ with any initial point $u_0 \in \mathcal{H}$, is guaranteed to converge to a certain fixed point in $\mathrm{Fix}(T)$.

2) Bounded perturbation resilience of Picard iteration [18], [19]: Let $(\beta_k)_{k\in\mathbb{N}}$ be a summable sequence in $\mathbb{R}_+$ and $(d_k)_{k\in\mathbb{N}}$ be a bounded sequence in $\mathcal{H}$, where such a $(\beta_k d_k)_{k\in\mathbb{N}}$ is said to be a sequence of bounded perturbations. Then, with any initial point $u_0 \subset \mathcal{H}$, $(u_k)_{k\in\mathbb{N}}$ generated by

$$(\forall k \in \mathbb{N})\ u_{k+1} = T(u_k + \beta_k d_k)$$

also converges to a point $\overline{u} \in \mathrm{Fix}(T)$.

REFERENCES

[1] V. Bioglio, G. Coluccia, and E. Magli, "Sparse image recovery using compressed sensing over finite alphabets," in *ICIP*, 2014.

[2] A.-J. Van der Veen, S. Talwar, and A. Paulraj, "Blind estimation of multiple digital signals transmitted over fir channels," *IEEE Signal Processing Letters*, 1995.

[3] B. Knoop, F. Monsees, C. Bockelmann, D. Peters-Drolshagen, S. Paul, and A. Dekorsy, "Compressed sensing k-best detection for sparse multi-user communications," in *EUSIPCO*, 2014.

[4] A. Bemporad and M. Morari, "Control of systems integrating logic, dynamics, and constraints," *Automatica*, 1999.

[5] S. M. Fosson and M. Abuabiah, "Recovery of binary sparse signals from compressed linear measurements via polynomial optimization," *IEEE Signal Processing Letters*, 2019.

[6] B. Trotobas, A. Nafkha, and Y. Louët, "A review to massive mimo detection algorithms: Theory and implementation," *Advanced Radio Frequency Antennas for Modern Communication and Medical Systems*, 2020.

[7] P. Sarangi and P. Pal, "Measurement matrix design for sample-efficient binary compressed sensing," *IEEE Signal Processing Letters*, 2022.

[8] M. Toyoda and M. Tanaka, "Efficient iterative method for soav minimization problem with linear equality and box constraints and its linear convergence," *Journal of the Franklin Institute*, 2022.

[9] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Transactions on Communications*, 2011.

[10] M. Wu, C. Dick, J. R. Cavallaro, and C. Studer, "High-throughput data detection for massive mu-mimo-ofdm using coordinate descent," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2016.

[11] A. Kudeshia, A. K. Jagannatham, and L. Hanzo, "Total variation based joint detection and state estimation for wireless communication in smart grids," *IEEE Access*, 2019.

[12] J.-C. Chen, "Manifold optimization approach for data detection in massive multiuser mimo systems," *IEEE Transactions on Vehicular Technology*, 2018.

[13] M. Nagahara, "Discrete signal reconstruction by sum of absolute values," *IEEE Signal Processing Letters*, 2015.

[14] R. Hayakawa and K. Hayashi, "Convex optimization-based signal detection for massive overloaded mimo systems," *IEEE Transactions on Wireless Communications*, 2017.

[15] M. A. Albreem, W. Salah, A. Kumar, *et al.*, "Low complexity linear detectors for massive mimo: A comparative study," *IEEE Access*, 2021.

[16] R. Hayakawa and K. Hayashi, "Discrete-valued vector reconstruction by optimization with sum of sparse regularizers," in *EUSIPCO*, 2019.

[17] J. Abe, M. Yamagishi, and I. Yamada, "Linearly involved generalized Moreau enhanced models and their proximal splitting algorithm under overall convexity condition," *Inverse Problems*, 2020.

[18] Y. Censor, R. Davidi, and G. T. Herman, "Perturbation resilience and superiorization of iterative algorithms," *Inverse Problems*, 2010.

[19] J. Fink, R. L. G. Cavalcante, and S. Stańczak, "Superiorized adaptive projected subgradient method with application to mimo detection," *IEEE Transactions on Signal Processing*, 2023.

[20] W. Yata, M. Yamagishi, and I. Yamada, "A constrained LiGME model and its proximal splitting algorithm under overall convexity condition," *Journal of Applied and Numerical Optimization*, 2022.

[21] W. Yata and I. Yamada, "Imposing early and asymptotic constraints on ligme with application to bivariate nonconvex enhancement of fused lasso models," 2024. arXiv: 2309.14082.

[22] I. Selesnick, "Sparse regularization via convex analysis," *IEEE Transactions on Signal Processing*, 2017.

[23] Y. Chen, M. Yamagishi, and I. Yamada, "A unified design of generalized moreau enhancement matrix for sparsity aware ligme models," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 2023.

[24] D. Kitahara, R. Kato, H. Kuroda, and A. Hirabayashi, "Multi-contrast CSMRI using common edge structures with LiGME model," in *EUSIPCO*, 2021.

[25] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, 2008.

[26] R. Hayakawa and K. Hayashi, "Reconstruction of complex discrete-valued vector via convex optimization with sparse regularizers," *IEEE Access*, 2018.

[27] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, 2005.

U-Mamba-Net: A highly efficient Mamba-based U-net style network for noisy and reverberant speech separation

Shaoxiang Dang*, Tetsuya Matsumoto*, Yoshinori Takeuchi†, and Hiroaki Kudo*
* Graduate School of Informatics, Nagoya University, Nagoya, Japan
† School of Informatics, Daido University, Nagoya, Japan
E-mail: dang.shaoxiang.s0@s.mail.nagoya-u.ac.jp

***Abstract*—The topic of speech separation involves separating mixed speech with multiple overlapping speakers into several streams, with each stream containing speech from only one speaker. Many highly effective models have emerged and proliferated rapidly over time. However, the size and computational load of these models have also increased accordingly. This is a disaster for the community, as researchers need more time and computational resources to reproduce and compare existing models. In this paper, we propose U-mamba-net: a lightweight Mamba-based U-style model for speech separation in complex environments. Mamba is a state space sequence model that incorporates feature selection capabilities. U-style network is a fully convolutional neural network whose symmetric contracting and expansive paths are able to learn multi-resolution features. In our work, Mamba serves as a feature filter, alternating with U-Net. We test the proposed model on Libri2mix. The results show that U-Mamba-Net achieves improved performance with quite low computational cost.**

I. Introduction

Speech separation in complex environments is one of the most challenging branches of the speech processing community [1]–[3]. Noise, reverberation, and other factors severely interfere with the desired signal, making it even more difficult to capture long-term dependencies, which has long been valued by the community [4]–[6].

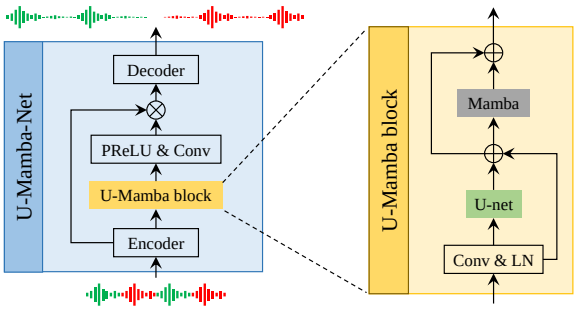
In the endeavor to capture long-term dependencies, recurrent neural networks (RNNs) are the first popular structure that demonstrates effectiveness on both time-frequency [4], [5], [7]–[10] and time-domain processing [3], [11]–[15]. In time-domain separation, RNNs are typically used in conjunction with dual-path (DP) structure [11]. The DP structure assists the network in modeling global information, complementing the processing of local information. Later on, self-attention mechanism like Transformer [16], as an alternative to RNNs, that allows parallel computing is introduced into the DP structure [14]. As for drawbacks, it adds extra computational overhead, and the quadratic scaling of the self-attention mechanism in transformers also make training time-consuming and resource-intensive, which has been a growing consensus and concern in this community.

The cascaded multi-task learning (CMTL) methods advocates solving complex speech separation task structurally [1]. Concretely, it decomposes a complex task into simpler, sequential sub-tasks. Through progressive supervision, cascaded multi-task structure can achieve further improvements in performance. However, due to the stacking of sub-modules, CMTL cannot effectively control the size of the model. Worse still, there may be gradient conflicts between modules, which significantly limits the model's capabilities [3].

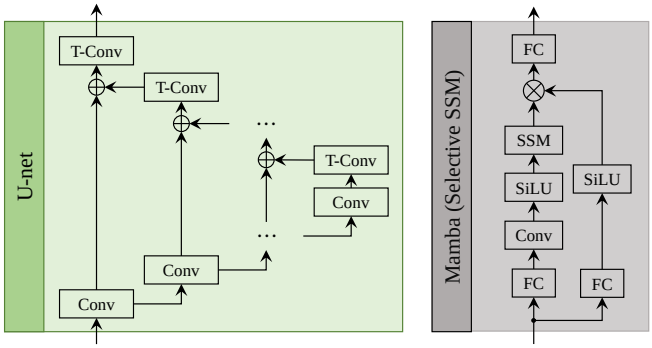
In this work, we shed light on U-net, which though was developed for biomedical image segmentation [17]. The U-net architecture consists entirely of convolutional neural layers, making it compact in size and low in computational complexity. Demucs [18] and SuDoRMRF [19] are successful attempts at using the U-Net architecture for music sound separation and audio source separation, respectively. To address its inability to learn global relationships, we add a Mamba module after each U-Net block. Mamba is a selective structured state space sequence model [20]. Concretely, Mamba features high-order polynomial projection operators (HiPPO) initialization and input-dependent structure while maintain linear computational complexity. HiPPO initialization assists in decomposing signals into basis functions [21]. We conduct experiments in the simulated noisy and reveberant version of open-source public dataset Libri2mix [22]. According to performance comparison of proposed model and previous models, we find that U-Mamba-Net is not only able to performance better cross various metrics, but also achieve so with low computational resource.

II. Proposed U-Mamba-Net

U-Mamba-Net consists of an encoder, U-Mamba blocks, and a decoder, similar to the prevailing structures used previously [15], [23]. The encoder is a 1-dimensional convolutional layer that maps waveform to time-frequency like representation. U-Mamba blocks aim to learn features with high representational capability. A convolutional layer following U-Mamba blocks assists estimating masks for all sources. After applying the estimated masks to the representations of the mixed speech, the features of all sources are decoded by the 1-dimensional transposed convolutional layer, resulting in the final separated speeches, The overview is illustrated in Fig. 1.



(a) U-Mamba-Net overview.



(b) Illustration of U-net (left) and Mamba (right).

Fig. 1. Overview of U-Mamba-Net.

A. U-Mamba blocks

Given the feature $\boldsymbol{X} \in \mathbb{R}^{F\times T}$ of a noisy and reverberant mixed speech, the goal of the B stacked U-mamba blocks is to estimate robust representation $\boldsymbol{M} \in \mathbb{R}^{F\times T}$. Then, a convolutional layer helps to generate masks equal to the number of sources S. As the core part of our proposed model, one U-Mamba block mainly owns a U-net module and a Mamba module. The U-net module is composed of L successive downsampling and upsampling layers, with each pair of sub-layers at corresponding depths connected by residual connections. Subsequently, the output of the U-net is fed into the Mamba module. The U-Mamba block outputs the estimated representations with another residual connection to the output of the U-net in this module. The detailed statement is presented in Algorithm 1.

B. Mamba

Mamba is an extension of state space models (SSMs) [24]. SSMs formulates a mapping from $\boldsymbol{x}(t) \in \mathbb{R}^F$ to $\boldsymbol{y}(t) \in \mathbb{R}^F$ via hidden state space $\boldsymbol{h}(t) \in \mathbb{R}^N$:

$$\boldsymbol{h}^{'}(t) = \boldsymbol{A}\boldsymbol{h}(t) + \boldsymbol{B}\boldsymbol{x}(t) \tag{1}$$
$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{h}(t) + \boldsymbol{D}\boldsymbol{x}(t) \tag{2}$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\boldsymbol{B} \in \mathbb{R}^{N\times F}$, $\boldsymbol{C} \in \mathbb{R}^{F\times N}$, and $\boldsymbol{D} \in \mathbb{R}^{F\times F}$ are state matrices. N is number of state space dimension.

Algorithm 1 An algorithm for b-th U-Mamba block

Input: $\boldsymbol{M}_b$, L
Output: $\boldsymbol{M}_{b+1}$
$\boldsymbol{D}^{(0)} \leftarrow \mathrm{PReLU}(\mathrm{LN}(\mathrm{BottleNeckConv}(\boldsymbol{M}_b)))$
▷ Bottle Neck Convolution
$l \leftarrow 1$
while $l \leq L$ **do** ▷ Downsampling
$\boldsymbol{D}^{(l)} \leftarrow \mathrm{LN}(\mathrm{Conv1D}(\boldsymbol{D}^{(l-1)}))$
$l \leftarrow l + 1$
end while
$l \leftarrow L$
$\boldsymbol{U}^{(L)} \leftarrow \boldsymbol{D}^{(L)}$
while $l \geq 1$ **do** ▷ Upsampling
$\boldsymbol{U}^{(l-1)} \leftarrow \mathrm{T\text{-}Conv1D}(\boldsymbol{U}^{(l)}) + \boldsymbol{D}^{(l-1)}$
▷ Residual connection
$l \leftarrow l - 1$
end while
$\boldsymbol{M}_b \leftarrow \mathrm{PReLU}(\boldsymbol{U}^{(0)})$
$\boldsymbol{M}_{b+1} \leftarrow \mathrm{Mamba}(\boldsymbol{M}_b) + \boldsymbol{M}_b$
▷ Mamba module and residual connection
return $\boldsymbol{M}_{b+1}$

Based on the continuous form above, its discrete form can be formulated by:

$$\boldsymbol{h}_t = \overline{\boldsymbol{A}}\boldsymbol{h}_{t-1} + \overline{\boldsymbol{B}}\boldsymbol{x}_t \tag{3}$$
$$\boldsymbol{y}_t = \overline{\boldsymbol{C}}\boldsymbol{h}_t + \overline{\boldsymbol{D}}\boldsymbol{x}_t \tag{4}$$

where $\overline{\boldsymbol{A}}$ and $\overline{\boldsymbol{B}}$ are discretized parameters, which are converted from $\boldsymbol{A}$, $\boldsymbol{B}$, and a step parameter Δ using bilinear method:

$$\overline{\boldsymbol{A}} = (\boldsymbol{I} - \frac{\Delta}{2}\boldsymbol{A})^{-1}(\boldsymbol{I} + \frac{\Delta}{2}\boldsymbol{A}) \tag{5}$$
$$\overline{\boldsymbol{B}} = (\boldsymbol{I} - \frac{\Delta}{2}\boldsymbol{A})^{-1}\boldsymbol{\Delta}\boldsymbol{B} \tag{6}$$
$$\overline{\boldsymbol{C}} = \boldsymbol{C} \tag{7}$$
$$\overline{\boldsymbol{D}} = \boldsymbol{D} \tag{8}$$

System in equations 3 and 4 can be rewritten as a convolutional form of structured kernel $\overline{\boldsymbol{K}}$ and input $\boldsymbol{x}$.

$$\boldsymbol{y} = \overline{\boldsymbol{K}} * \boldsymbol{x}, \overline{\boldsymbol{K}} = (\overline{\boldsymbol{C}}\overline{\boldsymbol{B}}, \overline{\boldsymbol{C}}\overline{\boldsymbol{A}}\overline{\boldsymbol{B}}, \cdots, \overline{\boldsymbol{C}}\overline{\boldsymbol{A}}^{T-1}\overline{\boldsymbol{B}}) \tag{9}$$

Structured SSMs (S4) feature a HiPPO initialization instead of random initialization, which is confirmed to easier decompose representation into orthogonal polynomials.

Mamba improves S4 with a selective mechanism [20]. Concretely, it adds input-dependency into SSM matrices via parallel scan algorithm, Furthermore, state variables are kept in SRAM, which holds a fast and efficient GPU memory hierarchy. Mamba is illustrated in the right part of Fig. 1b.

TABLE I
REVERBERATION CONFIGURATION.

Room	L (m)	$\mathcal{U}$(5,10)
	W (m)	$\mathcal{U}$(5,10)
	H (m)	$\mathcal{U}$(3,4)
T60	T (s)	$\mathcal{U}$(0.2,0.6)
Receiver	L (m)	$\frac{L_{room}}{2} + \mathcal{U}$(-0.2,0.2)
	W (m)	$\frac{W_{room}}{2} + \mathcal{U}$(-0.2,0.2)
	H (m)	$\mathcal{U}$(0.9,1.8)
Sources	H (m)	$\mathcal{U}$(0.9,1.8)
	Dist. (m)	$\mathcal{U}$(0.66,2)
	θ	$\mathcal{U}$(0,2π)

TABLE II
HYPERPARAMETERS OF U-MAMBA-NET.

Encoder / Decoder	channel / window / hop	F / - / -	128 / 41 / 20
U-net	Input / output channel	-	128
Mamba	Input / output channel	-	128
U-net	Down / upsampling depths	L	4
U-Mamba	Number of repeat core	B	16
Sources	-	S	2

III. EXPERIMENTS

A. Datasets

We use open-source mixture dataset Libri2mix to conduct experiments [22]. Source audios are from Librispeech, ambient noises are sampled from WHAM! [25]. As for simulation of indoor surrounding, we exploit Pyroomacoustics toolkit to generate reverberant version of clean source [26]. The simulated mixture is produced by adding each reverberant source and ambient noise together in time domain. The configuration of indoor reverebration is shown in Table I. T60 represents the reverberation intensity, indicating the time required for the sound energy to decay by 60 dB. As a result, our training, validation, and test datasets owns 13900, 3000, 3000 samples, respectively. All the samples are in 8 kHz.

B. Network

The basic hyperparameters that we use are exhibited in Table II. For the ablation studies, we indicate different parameter while keeping the same alphabet symbols. Additionally, we evaluate several different upsampling methods. Besides T-Conv1D, we also investigate the nearest neighbor (NN) and linear upsampling techniques.

C. Training phase

We train the proposed model on GeForce 4070Ti Super with a batch size of 4 and initial learning rate of 0.00015. Each sample is randomly cut into 3-s long. Notably, we report the computational load based on the input utterance with this length. We set the maximum epoch at 120. As for objective function, we adopt permutation-invariant scale-invariant single-to-noise ratio (SI-SNR) [15], shown as follows:

$$\mathcal{L} = -\max_{\pi \in \mathcal{P}} \frac{1}{I} \sum_{i} \text{SI-SNR}(\hat{s}_{\pi(i)}, s_i) \tag{10}$$

where π is the best permutation mapping set that allow overall SI-SNR to achieve maximum [27].

TABLE III
MAIN RESULTS.

Methods	CMTL	SI-SNRi	SDRi	SIRi	#Param (M)	GMACs
TasNet [32]	✗	5.70	5.05	10.84	23.2 M	27.8
SuDoRM-RF [19]	✗	2.90	3.36	6.14	2.6 M	3.6
SuDoRM-RF+ [19]	✗	5.33	6.05	11.02	2.7 M	3.0
Conv-TasNet [15]	✗	6.88	7.17	14.32	6.3 M	18.7
DPRNN [11]	✗	7.59	7.88	15.16	3.7 M	23.9
DPRNN	✓	8.08	8.62	16.39	5.6 M	40.2
U-Mamba-Net	✗	**8.50**	**8.62**	**17.67**	4.4 M	2.5

D. Evaluation phase

SI-SNR improvement (SI-SNRi) is an extension of SI-SNR, it checks how much information get excluded from mixture, we use SI-SNR improvement (SI-SNRi). We use signal-to-interference ratio improvement (SIRi) for separation effectiveness [28]. At the perceptual aspect, we use short-time objective intelligibility (STOI) [28] and perceptual evaluation of speech quality (PESQ) [29]. For denoising performance, we use predicted rating of speech distortion (CSIG), predicted rating of background distortion (CBAK), and predicted rating of overall quality (COVL) [30]. Additionally, we present the model size and computational load in terms of Giga Multiply-Add Operations per Second (GMACs) [31].

IV. EXPERIMENTAL RESULTS

A. Main results

We compare the results of U-Mamba-Net with several previous models in Table III. In the implementation of CMTL using DPRNN, we adopt an Enhancement Priority Pipeline (EPP). We can observe that, firstly, the proposed U-mamba-net achieves the best results cross major metrics. It is 0.92 dB better than DPRNN in SI-SNRi. Compared to CMTL, which uses DPRNN as sub-modules, U-mamba-net is still 0.42dB higher. Moreover, U-Mamba-Net model size is 20% smaller than DPRNN (CMTL). Most importantly, its computational efficiency (measured in GMACs) is only one-sixteenth of that of DPRNN (CMTL) (2.5 vs. 40.2) and one-ninth of that of DRRNN (2.5 vs. 23.9). This strongly demonstrates the proposed model's high efficiency and effectiveness.

B. Performance of perceptual quality and denoising effectiveness

TABLE IV
PERCEPTUAL AND DENOISING PERFORMANCE.

Methods	CMTL	STOI	PESQ	CSIG	CBAK	COVL
SuDoRM-RF+	✗	65.92	1.52	2.37	1.61	1.84
Conv-TasNet	✗	69.61	1.60	2.58	1.76	2.00
DPRNN	✗	71.86	1.67	2.65	1.83	2.07
DPRNN	✓	73.19	**1.75**	**2.75**	**2.16**	**2.18**
U-Mamba-Net	✗	**73.99**	1.70	2.56	1.89	2.05

In this section, we compare the U-Mamba-Net with previous models in terms of perceptual quality and denoising performance in Table IV. In terms of the STOI metric, U-Mamba-Net consistently maintains its advantage, just as the results and

TABLE V
ABLATION STUDIES.

F	R	L	Upsampling	SI-SNR	SI-SNRi	SDRi	SIRi	STOI (%)	PESQ	CSIG	CBAK	COVL	#Param (M)	GMACs
64	16	4	T-Conv1D	1.29	7.12	7.59	14.64	70.59	1.60	2.52	1.75	1.97	1.3 M	0.7
128	12	4	T-Conv1D	2.31	8.14	8.39	16.31	73.28	1.68	2.53	1.85	2.02	3.3 M	1.9
128	16	4	T-Conv1D	2.67	8.50	8.62	17.67	73.97	1.70	2.53	1.89	2.04	4.4 M	2.5
128	20	4	T-Conv1D	2.76	8.59	8.76	17.46	74.25	1.71	2.56	1.90	2.06	5.5 M	3.1
128	16	8	T-Conv1D	2.59	8.42	8.52	17.39	73.72	1.69	2.50	1.87	2.01	4.6 M	2.5
128	16	4	NN	2.70	8.52	8.66	17.44	73.99	1.70	2.56	1.89	2.05	4.4 M	2.5
128	16	4	Linear	2.70	8.53	8.70	17.54	74.06	1.71	2.55	1.89	2.05	4.4 M	2.5
192	16	4	T-Conv1D	3.02	8.85	8.90	18.15	74.87	1.74	2.64	1.93	2.11	9.7 M	5.3

conclusions in the previous section. It also shows superiority in PESQ, CSIG, CBAK, and COVL when compared to all models except DPRNN (CMTL). However, DPRNN (CMTL) outperforms U-Mamba-Net in these metrics, particularly in CBAK. This indicates that a single-task model's performance on specific sub-task is not as effective as that of a CMTL architecture incorporating additional processing modules. We believe the reason is that DPRNN (CMTL) can leverage additional information like noise-free mixture as intermediate label to supervised model in a step-by-step manner. This suggests that combining specialized modules in a multi-task framework can offer superior performance by addressing various aspects of the problem more comprehensively.

C. Ablation studies

To better understand the role of each parameter in the model, we design several ablation experiments. Their results are displayed in Table V. Firstly, the impact of feature dimensions on the model is most noticeable. As the feature dimension F increases from 64 to 128 to 192, the separation performance significantly improves like in SI-SNRi from 7.12 dB to 8.50 dB to 8.85 dB, but the model size also increases accordingly. Increasing the number of U-mamba-Net blocks R also enhances the model's performance, although the improvement is not as pronounced. With 20 blocks, the model reaches an SI-SNRi score of 8.59 dB. Increasing the depth L of the model has a negative effect on the test set. We speculate that for separation tasks, excessively low resolution may not be beneficial. Using NN and linear upsampling method though has a subtle improvement on the test set, we observe a decreased performance in the validation set during training. Therefore, upsampling methods are less critical than any other hyperparameters else.

D. Visualization

This section visualizes an example of separation results using DPRNN and U-Mamba-Net in Fig 2. First, from the spectrograms of the results, both DPRNN and U-Mamba-Net exhibit effective denoising and dereverberation, as evidenced by the clarity of the spectrograms of separations. Regarding the separation results, DPRNN exhibits more erroneous separations, where information originally belonging to one source is allocated to another source. This issue is highlighted by the red boxes in the figure. We believe this problem arises from the

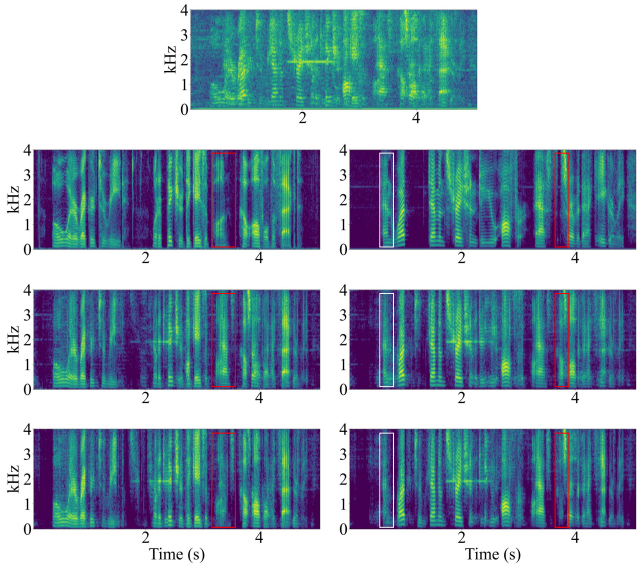


Fig. 2. Spectrogram of separation results. The sole spectrogram in the first row is noise and reverberant mixture. The following two spectrograms in the second row are ground truths. Third row are displaying the spectrograms of two separated results by DPRNN model. The last two are estimation of U-Mamba-Net. The red boxes highlight the places where DPRNN makes wrong separation, but U-Mamba does not. The white box outlines the place where U-Mamba-Net performs worse. Because the fundamental frequencies and harmonics of U-Mamba-Net are not as clear as those of the DPRNN.

model's less robust modeling of long-term dependencies. U-Mamba-Net also has some weaknesses. For instance, as shown in the white boxes, the generated speech spectrogram lines are not as clear as those produced by DPRNN. Even worse, when we listen to the separated speech, its perceptual quality remains relatively low, which is consistent with the STOI score indicated in the previous sections.

V. CONCLUSIONS

In this work, we proposed a lightweight U-Mamba-Net for noisy and reverberant speech separation. U-Mamba-Net not only demonstrates impressive separation capabilities but also maintains low computational load. In terms of SNR category metrics, it surpasses many previous models, demonstrating its overall robust separation capability. However, it still shows significant gaps in denoising and other specific metrics compared to CMTL.

ACKNOWLEDGMENT

This work is supported by a fellowship of the Nagoya University CIBoG WISE program from MEXT.

REFERENCES

[1] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 696–700.

[2] X. Shi, X. Li, and T. Toda, "Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation," in *Proc. Interspeech*, 2023, pp. 765–769.

[3] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, "A separation priority pipeline for single-channel speech separation in noisy environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 511–12 515.

[4] Y. Fu, Y. Liu, J. Li, *et al.*, "Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7417–7421.

[5] H. Shi, L. Wang, M. Ge, S. Li, and J. Dang, "Spectrograms fusion with minimum difference masks estimation for monaural speech dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7544–7548.

[6] Y. Liu, H. Sun, G. Chen, *et al.*, "Multi-level knowledge distillation for speech emotion recognition in noisy conditions," in *INTERSPEECH 2023*, 2023, pp. 1893–1897.

[7] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[8] Y. Hu, Y. Liu, S. Lv, *et al.*, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Interspeech 2020*, 2020, pp. 2472–2476.

[9] H. Shi, K. Shimada, M. Hirano, *et al.*, "Diffusion-based speech enhancement with joint generative and predictive decoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 951–12 955. DOI: 10.1109/ICASSP48485.2024.10448429.

[10] H. Shi, M. Mimura, and T. Kawahara, "Waveform-domain speech enhancement using spectrogram encoding for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3049–3060, 2024.

[11] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 46–50.

[12] Y. Liu, X. Liu, X. Miao, and J. Yamagishi, "Target speaker extraction with curriculum learning," in *Interspeech 2024*, 2024, pp. 4348–4352. DOI: 10.21437/Interspeech.2024-1375.

[13] S. Dang, T. Matsumoto, H. Kudo, and Y. Takeuchi, "A restriction training recipe for speech separation on sparsely mixed speech," in *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part VI 28*, Springer, 2021, pp. 736–743.

[14] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 21–25.

[15] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[16] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 2015, pp. 234–241.

[18] A. Défossez, N. Usunier, L. Bottou, and F. Bach, *Demucs: Deep extractor for music sources with extra unlabeled data remixed*, 2019. arXiv: 1909.01174 [cs.SD].

[19] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm -rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, Sep. 2020.

[20] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[21] A. Gu, I. Johnson, A. Timalsina, A. Rudra, and C. Ré, "How to train your hippo: State space models with generalized orthogonal basis projections," *arXiv preprint arXiv:2206.12037*, 2022.

[22] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, *Librimix: An open-source dataset for generalizable speech separation*, 2020. arXiv: 2005.11262 [eess.AS].

[23] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, "Using semi-supervised learning for monaural time-domain speech separation with a self-supervised learning-based si-snr estimator," in *Interspeech 2023*. ISCA, 2023.

[24] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[25] G. Wichern, J. Antognini, M. Flynn, *et al.*, "WHAM!: Extending speech separation to noisy environments," in *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, ISCA, 2019, pp. 1368–1372.

[26] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 351–355.

[27] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 626–630.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, vol. 2, 2001, pp. 749–752.

[30] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2007.

[31] V. Sovrasov. "Ptflops: A flops counting tool for neural networks in pytorch framework." (2024), [Online]. Available: https://github.com/sovrasov/flops-counter.pytorch.

[32] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 696–700.

StylebookTTS: Zero-Shot Text-to-Speech Leveraging Unsupervised Style Representation

Juhwan Yoon, Hyungseob Lim, Hyeonjin Cha and Hong-Goo Kang
Yonsei University, Seoul, South Korea
E-mail: {jhyoon10, hyungseob.lim, hcha}@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

***Abstract*—Zero-shot text-to-speech (ZS-TTS) is a TTS system capable of generating speech in voices it has not been explicitly trained on. While many recent ZS-TTS models effectively capture target speech styles using a single global style feature per speaker, they still face challenges in achieving high speaker similarity for voices that were not previously encountered. In this study, we propose StylebookTTS, a novel ZS-TTS framework that extracts and utilizes multiple target style embeddings based on the content. We begin by extracting style information from target speeches, leveraging linguistic content obtained through a self-supervised learning (SSL) model. The extracted style information is stored in a collection of embeddings called a stylebook, which represents styles in an unsupervised manner without the need for text transcriptions or speaker labeling. Simultaneously, the input text is transformed into content features using a transformer-based text-to-unit module, which links the text to the SSL representations of an utterance reading that text. The final target style is created by selecting embeddings from the stylebook that most closely align with the content features generated from the text. Finally, a diffusion-based decoder is employed to synthesize the mel-spectrogram by combining the final target style with the content features generated from the text. Experimental results demonstrate that StylebookTTS achieves greater speaker similarity compared to baseline models, while also being highly data-efficient, requiring significantly less paired text-audio data.**

I. Introduction

Text-to-speech (TTS) is a technology that converts written text into spoken language by passing it through various processing stages. In recent years, TTS systems have made substantial progress thanks to the advent of deep learning techniques [3]–[8], enabling their use in applications such as speech-based virtual assistants and accessibility tools for individuals with disabilities. Traditionally, TTS systems have been developed using text paired with the acoustic features of speakers encountered during training.

Recently, there has been increasing interest in generating voices for new speakers—who were not included in the training data—using only a few seconds of their speech samples. This approach is referred to as zero-shot multi-speaker TTS (ZS-TTS). Traditional ZS-TTS models [9]–[17] typically extract a global speaker embedding from the input speech using an external speaker encoder. This embedding is then used as a conditioning factor within the TTS framework. They demonstrate outstanding performance in synthesizing speech that closely resembles human-recorded utterances in terms of both naturalness and intelligibility. However, these networks continue to face challenges in bridging the similarity gap between observed and unobserved speakers during training. This issue arises because they rely on a single global vector to condition all the frames, which inadequately captures and transfers the target speaker's style to the generated speech from text. Although this issue can be mitigated by fine-tuning the pre-trained TTS network with a few seconds of speech samples from the desired speaker [12], this process is both time-consuming and memory-intensive for each new speaker.

In this work, we propose StylebookTTS, a zero-shot TTS framework that utilizes multiple target style representations during synthesis to enhance speaker similarity with the target speaker. StylebookTTS is built upon the stylebook framework [2], a speech analysis and synthesis method originally developed for any-to-any voice conversion. This framework captures diverse style representations based on linguistic content, without requiring text transcription or speaker labeling during training and inference. In the voice conversion framework, various style embeddings for different frames are extracted based on the source speech's content embeddings, which are derived from quantized self-supervised learning (SSL) features. These style and content embeddings are then integrated in a diffusion model [18] to produce the converted speech. To adapt the stylebook framework for TTS applications, we introduce a transformer-based [19] text-to-unit module that estimates the speech unit (i.e., quantized SSL feature) corresponding to the linguistic content of the speech from its text transcription. This module is trained independently and is then integrated with the pre-trained stylebook model to create the complete text-to-speech network. Our experiments demonstrate that StylebookTTS achieves enhanced speaker similarity with unseen target speakers and improved intelligibility, thanks to its use of fine-grained, content-dependent style embeddings.

In addition to improved speaker similarity, our proposed model efficiently utilizes small amounts of paired data. In contrast to previous models that need to learn both the pronunciation of sounds and speaker-specific style factors from paired text-audio data, our framework delegates the style adaptation to the pre-trained stylebook model. Instead, it focuses on establishing the relationship between text and its pronunciation, as captured by the quantized SSL features. Ultimately, StylebookTTS requires significantly less paired data to achieve near-optimal performance. These results indicate that this content-based target style modeling methodology can deliver excellent performance in speech synthesis tasks, even with limited paired

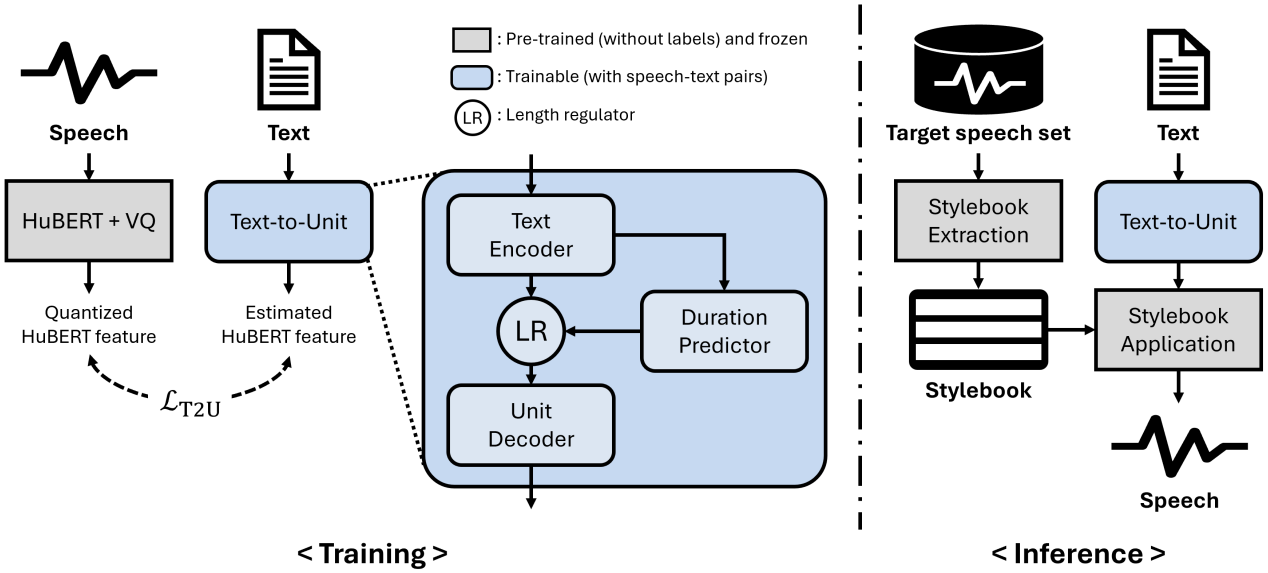


Fig. 1. Block diagram of the proposed text-to-speech framework utilizing pre-trained HuBERT [1] and a style extraction and application model [2]. During training, a text-to-unit module is trained to estimate the quantized HuBERT feature from a given text. After trainig, the HuBERT model originally included in the style application module is replaced with the trained text-to-unit module so that a speech can be generated from a text instead of a HuBERT feature.

data.

Our contributions are summarized as follows[1]:

- We extend the content-dependent stylebook network [2] to the TTS domain, achieving high speaker similarity and intelligibility with unobserved speakers, while maintaining comparable overall speech quality.
- We demonstrate that our proposed StylebookTTS can be efficiently optimized even with small amounts of paired data by relying on external stylebook module, which has been pre-trained in an unsupervised manner to handle the complex style modeling.

II. Related works

A. Zero-shot text-to-speech

Zero-shot text-to-speech aims to synthesize speech from a text transcript and reference audio (or target speech) of an unobserved speaker, ensuring that the characteristics of the generated speech closely match those of the target speaker. The goal is typically achieved by extracting style-related information from the target speaker's utterance and incorporating it as a conditional feature into the TTS model during synthesis. For style-related information, many previous works [9]–[12] have commonly used a speaker embedding vector extracted by a speaker verification model, which may be either pre-trained or jointly trained with a classification loss. However, compressing all style-related information into a single vector can lead to information loss, which limits the style transfer capability of the ZS-TTS system.

On the other hand, prompt-based approaches [13]–[17] allow the model to directly access the acoustic features of the target speech through the attention mechanism [19]. In contrast to approaches that rely on a global speaker embedding, prompt-based models can locally adapt the style at each time frame. This results in better speaker similarity, regardless of whether the model undergoes a discriminative learning process with speaker labels [13]–[17]. However, these models inevitably face increased computational costs as the length of the target speech grows, since the attention layer must reference the entire target utterance at each attention step.

Unlike the previous approaches, we employ stylebook [2]—a collection of style embeddings each implicitly linked to different linguistic content—as the source of style-related information for conditioning a ZS-TTS model. By adopting this approach, we achieve better speaker similarity compared to models that use global speaker embeddings, and we further enhance performance with longer target speech without increasing computational costs during inference.

B. Speech analysis and synthesis using stylebook

The stylebook [2] model is an autoencoder-like speech analysis and synthesis network originally designed for any-to-any voice conversion. In this framework, target speeches are first processed by the mel and style encoders, which transform them into style embeddings. Simultaneously, the target speeches are converted into HuBERT [1] representations and undergo vector quantization to extract features related solely to linguistic content, as done in [20] and [21]. The obtained style and contents embeddings are fed into a multi-head attention mechanism with a learned query set, resulting in a set of embeddings—referred to as the stylebook— for the given target speaker. Since the query set consists of embeddings representing different content-related features, the resulting stylebook forms a collection of embeddings that capture various speaking styles of the target speaker. Similarly,

[1]Audio samples are available at http://pineville17.github.io/StylebookTTS

source speeches are converted into content embeddings using a HuBERT network and vector quantization. The model then selects the appropriate style embeddings from the stylebook by comparing the source speech's content embeddings with the stylebook entries using attention mechanisms. The obtained embeddings are combined in a diffusion-based U-Net model [22] to generate the mel-spectrogram of the converted speech. Finally, a pre-trained HiFi-GAN [23] vocoder is used to synthesize speech from the obtained mel-spectrogram. During training, the same speech is used as both the source and the target. Here, the model is trained to effectively represent the missing information lost during the quantization of HuBERT features—typically, the detailed speaking style associated with specific linguistic content.

The main contribution of the stylebook model is its use of transposed dual attention to generate a stylebook that represents the speaking style of a target speaker and to adapt the style embedding at each time frame based on the content of the source speeches. The content-dependent style embeddings can be extracted without the need for text transcription or speaker labeling, as the model is trained to reconstruct the input itself following decomposition with HuBERT and vector quantization. Additionally, since the number of embeddings in the stylebook remains fixed regardless of the length of the target speech, the stylebook model can leverage longer target speeches without increasing computational costs once the stylebook is extracted.

III. Proposed Method

In this work, we propose StylebookTTS designed to extend the stylebook framework to a zero-shot text-to-speech (ZS-TTS) system. StylebookTTS consists of two distinct components: the stylebook network and the text-to-unit module. The stylebook network, which follows the architecture outlined in [2], handles both the extraction and application of the stylebook, as shown in Fig. 1.

A. Text-to-unit estimation

The core component of our proposed method is the text-to-unit (T2U) module, which estimates vector-quantized HuBERT representations from the input text. This enables the network to extract content embeddings from the input text, as vector-quantization of SSL features effectively captures content information from the input speech.

The architecture of the text-to-unit module is inspired by FastSpeech2 [4], a non-autoregressive TTS model known for its high generation speed and excellent speech quality. FastSpeech2 is designed to train pitch, energy, and duration predictors during training, and incorporates the estimated pitch, energy, and duration as conditions during inference, thereby providing style information. In our text-to-unit module, we omit the pitch and energy prediction components from FastSpeech2 to avoid confusion in capturing contextual information. We retain the duration predictor to accurately estimate phoneme durations, ensuring that the input text aligns well with the HuBERT units. During inference, the text is processed through the text encoder to generate phoneme embeddings. The length regulator adjusts the duration of the input sequence based on the predicted duration of each phoneme, indicating how many discrete HuBERT unit frames correspond to each phoneme. Subsequently, the unit decoder generates discrete HuBERT units from the length-adjusted sequence.

Once the quantized SSL features are estimated from the input text, they are fed into the content encoder and then processed through the multi-head attention layer during the stylebook application stage. This process facilitates the generation of speech that preserves the content of the given text while closely resembling the style of the target speaker.

B. Training Method

The training process for the text-to-unit module involves several key steps to ensure accurate prediction of the SSL features and high speech quality. During training, both speech and its corresponding text transcription are provided as inputs to the text-to-unit module. The target for estimation is the discrete HuBERT representation, which captures the content-related information from the input speech. A duration predictor, similar to the one used in FastSpeech2, estimates phoneme durations. This predictor utilizes the Montreal Forced Alignment (MFA) tool [24] to obtain target durations for each phoneme, ensuring precise alignment between the input text and the target HuBERT representations.

As the module generates outputs, two loss functions are utilized: duration loss and HuBERT loss. For the duration loss, mean-squared error (MSE) is computed between the target and predicted durations provided by the duration predictor, ensuring that the predicted durations closely match the target phoneme durations. For the HuBERT loss, L1 loss is used to optimize feature prediction accuracy by comparing the predicted and target feature representations, which helps maintain high fidelity in the synthesized speech. The overall loss function for the T2U network is the sum of these two loss components, combining their advantages to enhance performance, as follows:

$$\mathcal{L}_{\mathrm{dur}} = \frac{1}{n}\Sigma_{i=1}^{n}(d_{target,i} - d_{pred,i})^2, \tag{1}$$

$$\mathcal{L}_{\mathrm{hub}} = \Sigma_{i=1}^{n}|\mathrm{hvq}_{target,i} - \mathrm{hvq}_{pred,i}|, \tag{2}$$

$$\mathcal{L}_{\mathrm{T2U}} = \mathcal{L}_{\mathrm{dur}} + \mathcal{L}_{\mathrm{hub}}, \tag{3}$$

where d indicates the phoneme duration and hvq indicates vector quantized HuBERT representations and n indicates the number of data.

IV. Experiments

A. Experiment settings

1) ***Dataset****:* We conducted experiments using the LibriTTS dataset [25]. For training the T2U module, we randomly selected 115,183 samples from 900 speakers, totaling 186 hours, from the LibriTTS-train-clean-360 subset. For evaluating the proposed StylebookTTS network, we chose 27 speakers from the LibriTTS-test-clean subset and created target speeches with varying lengths, ranging from 5 to 180 seconds.

TABLE I
NATURALNESS EVALUATION RESULTS IN MOS RATINGS. SUBJECTIVE MOS VALUES WITH 95% CONFIDENCE INTERVALS ARE SHOWN.

	MOS ↑	
Target Length	5 sec	180 sec
MetaStyleSpeech	2.72±0.13	2.93±0.13
YourTTS	2.69±0.13	2.63±0.13
StylebookTTS	**3.57±0.12**	**4.00±0.09**
Ground Truth	4.55±0.09	

*2) **Model configuration**:* The model configuration of the stylebook framework follows that of [2], specifically using 128 embeddings in the stylebook, and is pre-trained on the same LibriTTS dataset. For the T2U module, the details are as follows. The text encoder consists of 4 feed-forward transformer (FFT) blocks, each with 2 attention heads and a hidden size of 256. The unit decoder is comprised of 6 FFT blocks, each with 2 attention heads and a hidden size of 256. The duration predictor consists of 2 convolution layers with a filter size of 256, a kernel size of 3, followed by a linear layer. The number of parameters of StylebookTTS including the T2U module and the stylebook framework is 77M.

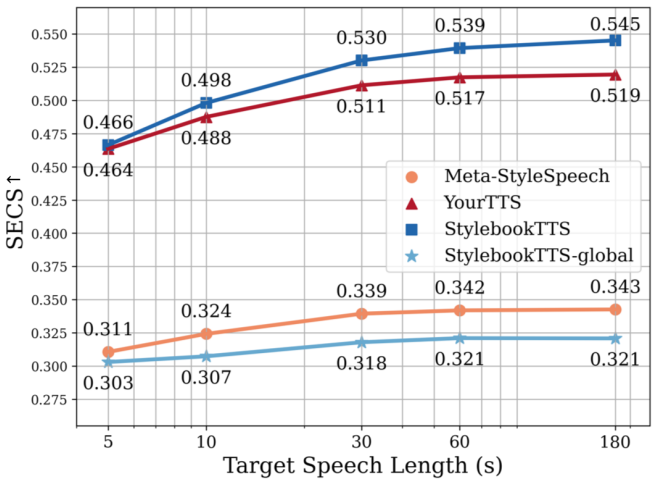
B. Evaluation Methods

We analyze and evaluate our proposed system using four metrics: mean opinion score (MOS) [26], ABX test preference score, speaker embedding cosine similarity (SECS), and character error rate (CER). The MOS test assesses the perceived quality of the generated speech through a subjective quality evaluation. The ABX test compares two synthesized speeches to determine which one exhibits higher speaker similarity to the real target speech. MOS and ABX results were obtained from evaluations involving 13 participants. SECS measures how closely the speaker features of the generated speech align with those of the target speech. We computed SECS by extracting speaker embeddings from both the synthesized and target speeches using the pre-trained ECAPA-TDNN [27] network[2] and calculated the cosine similarity between these embeddings. CER indicates the intelligibility of the generated speech. We computed CER by comparing the text of the source speech with the text of the converted speech using a HuBERT [1]-based automatic speech recognition (ASR) network[3].

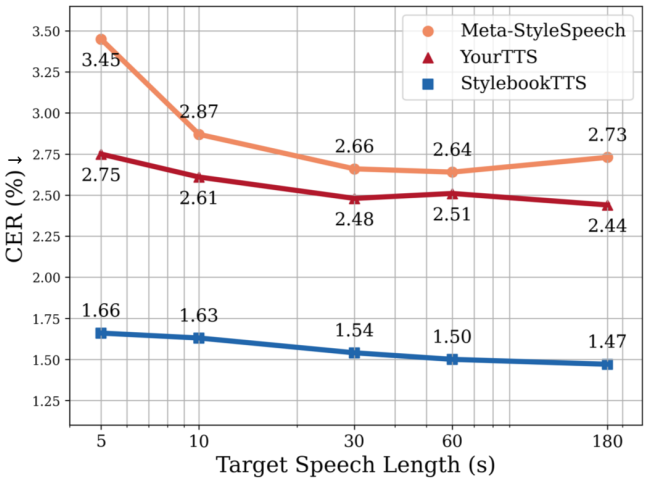
To evaluate performance based on the length of the target speech, we tested various target speeches with lengths of 5, 10, 30, 60, 120, 180, and 300 seconds. For comparison, we selected YourTTS [10] and Meta-StyleSpeech [11] as baseline models for non-finetuning zero-shot TTS frameworks and used their official implementations for synthesis.

C. Experiment results

*1) **Naturalness**:* Table I shows the MOS ratings for synthesized speeches generated by both the baseline and proposed networks for short (5 seconds) and long (180 seconds) input target speech lengths. The results show that our proposed StylebookTTS outperforms the baseline networks in terms of naturalness, regardless of the target speech length. Notably, as the target speech length increases, StylebookTTS generates speech that is nearly as natural as human utterances.



(a)



(b)

Fig. 2. Evaluation results of the proposed StyleBookTTS and the baseline models. (a) illustrates the SECS results and (b) shows CER (%) of various ZS-TTS networks.

*2) **Speaker similarity**:* StylebookTTS aims to enhance the speaker similarity of the synthesized voice to the original target voice by collecting and applying target style embeddings in a content-dependent manner. Fig. 2 (a) illustrates the SECS trends for synthetic speeches generated by the baseline and proposed models across various target lengths. In addition to the two baseline models and the proposed StylebookTTS, which uses multiple style embeddings (128 in this case) to form the stylebook, we also present results for another StylebookTTS network with a single style embed-

[2]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
[3]https://huggingface.co/facebook/hubert-large-ls960-ft

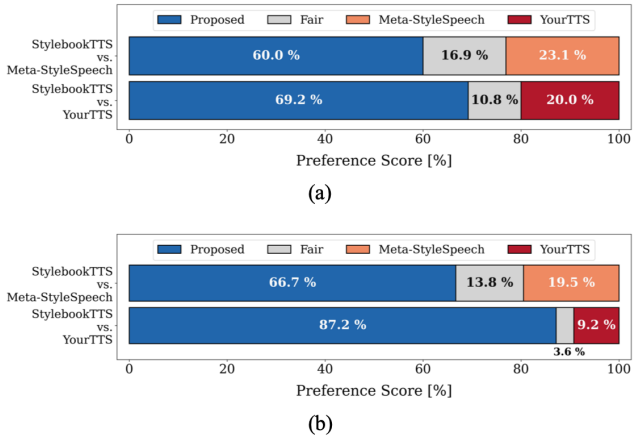


Fig. 3. ABX test result between the proposed StyleBookTTS and the baseline models. (a) illustrates the result when 5 seconds of target speech is introduced. (b) shows the result when 180 seconds of target speech is introduced.

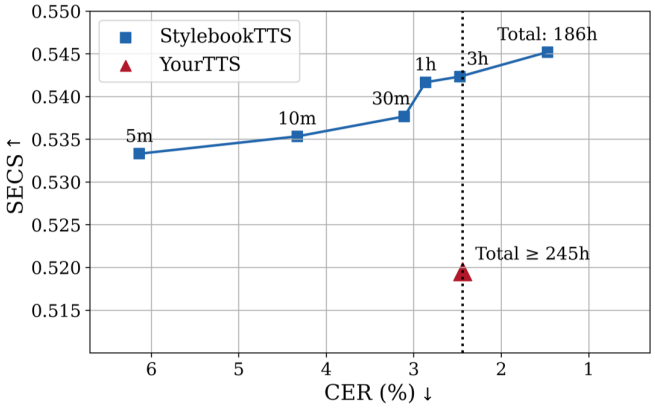


Fig. 4. Evaluation results (SECS and CER) regarding the length of the total speech paired with the text transcription used to train the text-to-unit module.

ding (StylebookTTS-global). The StylebookTTS-global can be viewed as a global style embedding within the stylebook framework. The figure demonstrates that, in terms of speaker similarity, StylebookTTS outperforms both the baseline networks and the StylebookTTS framework using a global style embedding to represent the target voice.

This trend is also evident in the ABX results. As shown in Fig. 3, StylebookTTS demonstrates superior speaker similarity compared to the baseline networks ($p < 0.01$ in all cases). When only 5 seconds of target voice are used, 60% of listeners preferred StylebookTTS over the baseline models for higher speaker similarity. This proportion increases significantly as the target speech length is extended to 180 seconds. Over 87% of listeners indicated that the proposed model generates speech that is more similar to the target speaker compared to YourTTS. These results demonstrate that our content-dependent style extraction approach enhances speaker similarity by extracting multiple content-dependent target styles and applying the most appropriate styles to match the content of the text.

*3) **Intelligibility**:* In addition to enhanced speaker similarity, our approach achieves remarkable intelligibility as shown in Fig. 2 (b). StylebookTTS achieves a CER of 1.66% with just 5 seconds of target speech, whereas Meta-StyleSpeech and YourTTS show CERs of 3.45% and 2.75%, respectively. As the target length increases, StylebookTTS continues to outperform other networks in terms of intelligibility.

*4) **Data efficiency**:* In Fig. 4, we illustrate the objective performance (SECS, CER) of our proposed model relative to the size of the paired data used for T2U training. For inference, 3 minutes of target speech were provided. Even with only 5 minutes of paired data, our proposed model achieves higher SECS compared to YourTTS trained on the full dataset. This demonstrates that StylebookTTS effectively utilizes the pre-trained stylebook model for style adaptation. Regarding intelligibility, StylebookTTS trained with 3 hours of paired data achieves performance comparable to YourTTS, which is trained on over 245 hours of text-audio paired data. These results indicate that our content-based target style modeling approach has the potential to deliver exceptional performance in various speech synthesis tasks, including multilingual and multi-modal synthesis, even with limited paired data.

V. Limitations and Future Works

Despite the improved performance in speaker similarity and intelligibility demonstrated by StylebookTTS, there are still some limitations to address. Since StylebookTTS models the target style based on the content of the target speech, its performance may be less reliable when the input target speech is short and lacks sufficient content information. This issue could be mitigated by conditionally combining global style embeddings with content-dependent style embeddings to address content scarcity. Another limitation is that StylebookTTS-generated speeches tend to have relatively monotonic prosody compared to ground truth speech, despite the high speaker similarity scores. Future research could address this by incorporating a separate variance adaptor to provide additional pitch information for more refined prosody control.

VI. Conclusion

In this work, we present StylebookTTS, a zero-shot text-to-speech (TTS) framework that synthesizes speech using multiple content-dependent target style representations derived from the given text. We developed a text-to-unit module that establishes the relationship between text and its pronunciation by utilizing quantized SSL features. The text-generated SSL features are then fed into the stylebook network for style application and synthesis. Our experimental results demonstrate that this approach achieves superior speaker similarity and intelligibility compared to other zero-shot TTS frameworks that rely on a single global style embedding. Furthermore, StylebookTTS outperforms baseline models even when trained with a relatively small amount of paired text and audio data. This work highlights the effectiveness of content-based target style modeling in speech synthesis applications with limited paired data, suggesting its potential for expansion into research areas such as multilingual or multimodal synthesis.

REFERENCES

[1] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[2] H. Lim, K. Byun, S. Moon, and E. Visser, "Stylebook: Content-dependent speaking style modeling for any-to-any voice conversion using only speech data," *arXiv preprint arXiv:2309.02730*, 2023.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017.

[4] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[5] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018.

[6] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI conference on artificial intelligence*, 2019.

[7] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. Advances in neural information processing systems (NeurIPS)*, 2019.

[8] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. International Conference on Machine Learning (ICML)*, 2021.

[9] E. Casanova, C. Shulby, E. Gölge, *et al.*, "Sc-glowtts: An efficient zero-shot multi-speaker text-to-speech model," in *Proc. Interspeech*, 2021.

[10] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *Proc. International Conference on Machine Learning (ICML)*, 2022.

[11] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *Proc. International Conference on Machine Learning (ICML)*, 2021, pp. 7748–7759.

[12] H. Kim, S. Kim, J. Yeom, and S. Yoon, "Unitspeech: Speaker-adaptive speech synthesis with untranscribed data," in *Proc. Interspeech*, 2023.

[13] Y. Zhou, C. Song, X. Li, *et al.*, "Content-dependent fine-grained speaker embedding for zero-shot speaker adaptation in text-to-speech synthesis," in *Proc. Interspeech*, 2022.

[14] C. Wang, S. Chen, Y. Wu, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[15] S. Kim, K. Shih, J. F. Santos, *et al.*, "P-flow: A fast and data-efficient zero-shot tts through speech prompting," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[16] C. Du, Y. Guo, F. Shen, *et al.*, "UniCATS: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding," in *Proc. AAAI Conference on Artificial Intelligence*, 2024.

[17] K. Shen, Z. Ju, X. Tan, *et al.*, "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *Proc. International Conference on Learning Representations (ICLR)*, 2024.

[18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[19] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proc. Advances in neural information processing systems (NIPS)*, 2017.

[20] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," in *Proc. Interspeech*, 2021.

[21] B. van Niekerk, M.-A. Carbonneau, J. Zaıdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[22] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[23] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," 2020.

[24] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Proc. Interspeech*, 2017.

[25] H. Zen, V. Dang, R. Clark, *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," in *Proc. Interspeech*, 2019.

[26] *Method for the subjective assessment of small impairments in audio systems*. Rec. ITU-R BS.1116-3, 2015.

[27] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020.

Knowledge Augmented Attention Gating Embedding for Link Prediction

Zewei Chen*, Shuhong Chen*, Chen Li†, Xianwei Zheng†, Minfan He† and Xutao Li*
* Key Lab of Digital Signal and Image Processing of Guangdong Province, Department of Electronic Engineering, Shantou University, Shantou, 515063, China
E-mail: 22zwchen@stu.edu.cn, 18shchen3@stu.edu.cn, lixt@stu.edu.cn
† School of Mathematics, Foshan University, Foshan, 52800, China
E-mail: fslichen@fosu.edu.cn, alexwzheng@fosu.edu.cn, heminfan1980@126.com

***Abstract*—Knowledge graphs (KGs) use graphical representations to illustrate relations between entities, forming a structured knowledge base. However, KGs often suffer from incompleteness, lacking some entities and their interactions. Current link prediction technologies based on KG embeddings primarily focus on learning entity features and relation patterns from existing KGs to address the incompleteness issue. Yet, these methods encounter the inherent sparsity issue in KGs—uneven distribution of entity connections and scarcity of direct links. This complexity challenges models in identifying intricate, potential associations, thereby compromising link prediction accuracy and overall model performance. In this paper, we introduce a innovative Knowledge Augmented Attention-based Gating Embedding (KAAGE) framework for link prediction. Our model augments knowledge by reversing relations and triples, then integrates both original and augmented structural data through gating units. Extensive experiments on three established benchmark datasets demonstrate the superiority of KAAGE over current models. Our implementation code is available at https://github.com/22zwChen/KAAGE**

I. Introduction

In recent years, KGs have played a critical role in numerous domains and caught the attention of various research communities, including information retrieval [1], [2], question answering systems [3], [4], and recommendation systems [5], [6]. These sophisticated repositories take the form of graphs, encapsulating real-world entities and the complex networks of relations that bind them together. Entities are represented as nodes, symbolizing tangible objects or conceptual abstractions found within reality, while relations serve as the interlinking fabric delineating the connections among these entities within the KG structure.

Despite their extensive coverage of relations, entities, and triples, current KGs still struggle with issues of pronounced data sparsity, diverse relational categories, and intricate hierarchies. These challenges have spurred a wave of advanced research aimed at improving knowledge graph completion. Early logic-based approaches, such as HL-MRFs [7] and CPRA [8], are limited by data sparsity and poor scalability. Distributed representation schemes were then proposed to translate entities and relations into points within a continuous vector space, where mathematical operations can be used to evaluate entity similarities, as demonstrated by RESCAL [9] and TransE [6]. With the remarkable ability to discern features, neural networks have become prominent in the field of Knowledge Graph Embedding (KGE) learning. Neural-network-driven methods, including DSKG [10] and KGQA [11], autonomously learn representations by applying non-linear transformations to remodel the feature distribution of input data from its original space to a distinct feature space. However, despite the significant progress made by these methodologies, most do not strategically employ effective data augmentation techniques to address the incompleteness of data. Additionally, they often handle knowledge facts in isolation, overlooking the structural context inherently woven into the fabric of KGs.

In response to these challenges, we advocate for the integration of Graph Data Augmentation to enrich both the initial and augmented structural information of KGs. Building on the principles of CompGCN [12], we allow the information in the directed edges of the knowledge graph to propagate bidirectionally. We then incorporate an efficient gating mechanism that facilitates the harmonious integration of both the original structural information and the augmented structural information. Our key contributions are summarized as follows:

- To enrich knowledge graph, we implement knowledge augmentation strategies, specifically including the consideration of inverse relations and triples, which are integrated with the knowledge graph to counteract its sparseness.
- A gating unit is incorporated to merge the original structural information with the augmented ones, to dynamically assessing how important each piece of information is.
- Massive experimental results indicate that KAAGE framework outperforms competitors on three standard benchmark datasets, particularly on the Kinship dataset.

The remainders of this paper are structured as follows: Section II delves into the core of the matter, detailing the key aspects and constituent parts. Section III provides an examination of the proposed framework, mapping out the essential methodologies that define its operation. Section IV is dedicated to the empirical substantiation of KAAGE to validate the framework's effectiveness. In Section V, we encapsulate the key contributions of this research and underscore their importance.

II. PRELIMINARIES

A. Problem Definition

Knowledge Graph. We define $\mathcal{E}$, $\mathcal{R}$ as the sets of entities e and relations r, respectively. A knowledge graph is denoted by $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$ where $\mathcal{T} \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is the set of triples. Each triple in $\mathcal{T}$ can be represented as (h, r, t) where $h \in \mathcal{E}$ is head entity, $t \in \mathcal{E}$ is tail entity and $r \in \mathcal{R}$ is relation.

Link Prediction. To achieve link prediction, we compute score for each triple (h, r, t) by our model. Specifically, our model predicts the missing one in an incomplete triple such as $(h, r, ?)$ by a pre-designed scoring function $\psi(h, r, t) \in \mathbb{R}$.

B. ConvE

ConvE, initially introduced by Dettmers et al. [13], revolutionized the approach to link prediction in KGs. This model leverages 2D convolutional operations and stacks of nonlinear transformations to accurately predict missing links within KGs. The scoring function for a triple within this framework can be formally expressed as follows:

$$\Psi(e_i, r_k, e_j) = f(\mathrm{vec}(f([\mathbf{h}_i; \mathbf{g}_k] * \omega))\mathbf{W})e_j \tag{1}$$

where e_i, e_j, $r_k \in \mathbb{R}^{k_h k_w \times 1}$ represent the embeddings of the head entity, tail entity, and relation, respectively. $\mathbf{h}_i, \mathbf{g}_k \in \mathbb{R}^{k_h \times k_w}$ are 2D reshaped representations of head entity and relation. A 2D convolution $*$ with kernel ω is applied to $[\mathbf{h}_i; \mathbf{g}_k]$, followed by activation f and linear transformation $\mathbf{W}$. Given its effectiveness, ConvE is adopted as the decoder in our model.

C. Gating unit

To balance diverse features for KG embedding, we incorporate an adaptive, learnable gating unit that dynamically merges them as follow:

$$\mathbf{g} = \sigma(\mathbf{W_{ge}}\mathbf{e_i} + \mathbf{W_{ga}}\mathbf{a} + \mathbf{b}), \tag{2}$$

$$\mathbf{u} = \tau(\mathbf{W_u}(\mathbf{e_i} \oplus \mathbf{a})), \tag{3}$$

$$\mathbf{e} = \mathbf{g} \odot \mathbf{u} + (\mathbf{1} - \mathbf{g}) \odot \mathbf{e_i}, \tag{4}$$

where $e_i \in \mathbb{R}^{d_1}$, $a \in \mathbb{R}^{d_2}$ are input embeddings. This unit employs linear transformations $\mathbf{W_u} \in \mathbb{R}^{d_1 \times (d_1 + d_2)}$, $\mathbf{W_{ge}} \in \mathbb{R}^{d_1 \times d_1}$, and $\mathbf{W_{ga}} \in \mathbb{R}^{d_1 \times d_2}$, along with a bias $\mathbf{b}$. σ is the *sigmoid* function, and τ is hyperbolic tangent function. Importantly, the gating unit handles inputs of differing dimensions and outputs $e \in \mathbb{R}^{d_1}$, which matches the dimension of e_i, defined as $e = gate(e_i, a)$.

III. METHODOLOGY

Figure 1 shows each attention head of our framework KAAGE and multi-head attention mechanism is applied to our experiments on partial datasets.

A. Knowledge Augmentation

To facilitate knowledge augmentation, we view each triple $t_{ij}^k = (e_i, r_k, e_j) \in \mathcal{T}$ as a fundamental unit of knowledge and enhance it by reversing its direction. Specifically, we transform each triple (e_i, r_k, e_j) into (e_j, r_k^{-1}, e_i), where $\mathcal{R}' = \{r^{-1} | r \in \mathcal{R}\}$, and define $\mathcal{T}'$ as set of inverse triples. We then propose the augmented knowledge graph $\mathcal{G}' = (\mathcal{E}, \mathcal{R}'', \mathcal{T}'')$, where $\mathcal{R}'' = \mathcal{R}' \cup \mathcal{R}$ and $\mathcal{T}'' = \mathcal{T}' \cup \mathcal{T}$.

B. Triple representation learning

We present a triple representation learning method to make use of relations. For each triple $t_{ij}^k = (e_i, r_k, e_j) \in \mathcal{T}''$, it can be formulated as follows:

$$a_{ijk} = \mathbf{W_a} \cdot (h_i + h_j + g_k) \tag{5}$$

where h_i, h_j, $g_k \in \mathbb{R}^d$ are the initial embeddings of entities e_i, e_j and relation r_k, respectively. $\mathbf{W_a} \in \mathbb{R}^{d \times d}$ is a linear transformation matrix.

After that, in order to learn and quantify the importance of triple t_{ij}^k, we introduce an attention vector $\mathbf{v_{att}}$ as below:

$$b_{ijk} = \mathrm{LeakyReLU}(\mathbf{v_{att}} \cdot a_{ijk}) \tag{6}$$

where LeakyReLU is the non-linear activation function.

C. representation learning of entities

Subsequently, we compute the attention value using the *softmax* function to derive the relative importance for entity e_i as follows:

$$\alpha_{ijk} = \frac{exp(b_{ijk})}{\sum_{(r_r, e_n) \in \mathcal{N}_{e(i)}} exp(b_{inr})}, (r_k, r_r \in \mathcal{R}) \tag{7}$$

where $\mathcal{N}_{e(i)}$ denotes the neighborhoods of e_i in $\mathcal{T}$. Similarly, we can calculate β_{ijk} by replacing $\mathcal{R}$ and $\mathcal{T}$ with $\mathcal{R}'$ and $\mathcal{T}'$ respectively.

Therefore, to update the representation of entity e_i, we further categorize and aggregate the representations of these triples based on their unique relation types. The specific definitions as follows:

$$h'_{i_O} = f_e\left(\mathbf{W}_\mathrm{O} \sum_{(r_k, e_j) \in \mathcal{N}_{e(i)}} \alpha_{ijk} a_{ijk}\right), r_k \in \mathcal{R} \tag{8}$$

$$h'_{i_I} = f_e\left(\mathbf{W}_\mathrm{I} \sum_{(r_k, e_j) \in \mathcal{N}_{e(i)}} \beta_{ijk} a_{ijk}\right), r_k \in \mathcal{R}' \tag{9}$$

where h'_{i_O} and h'_{i_I} represent the representations of entity e_i aggregated from the original and inverse relations, respectively. f_e denotes a non-linear activation function such as $tanh$. $\mathbf{W}_\mathrm{O}$, $\mathbf{W}_\mathrm{I} \in \mathbb{R}^{d' \times d}$ are the graph convolutional kernels. To fuse the aggregated representations and incorporate residual connection, we employ the gating unit mentioned above. This is done as follows:

$$h'_i = gate(\mathbf{W}_\mathrm{e} \cdot h_i, h'_{i_O} \| h'_{i_I}) \tag{10}$$

where h'_i indicates the final representation of entity e_i. $\mathbf{W}_\mathrm{e} \in \mathbb{R}^{d' \times d}$ is a linear transformation matrix.

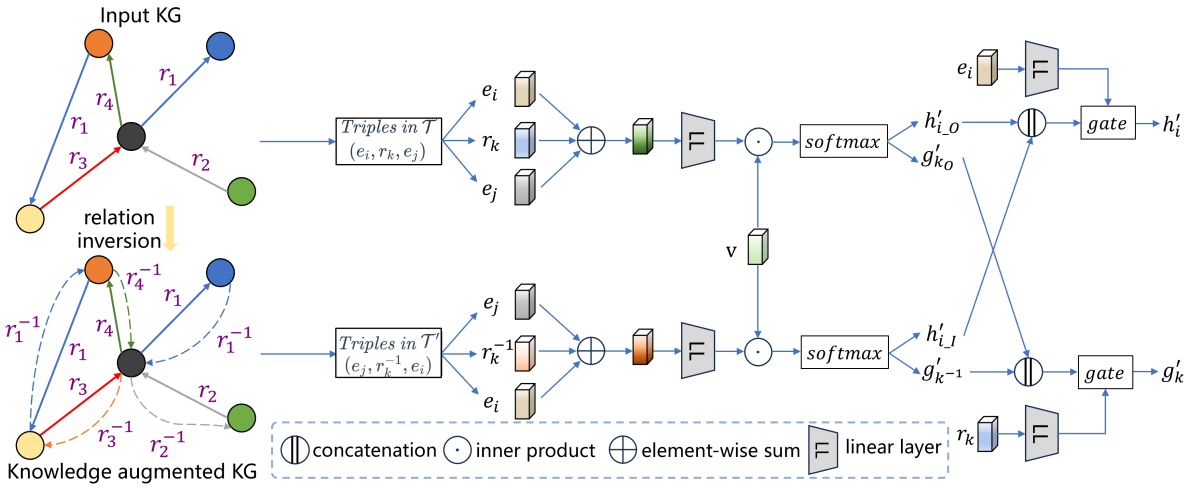


Fig. 1. Single attention head framework of our proposed model (KAAGE). e_i, r_k, e_j, r_k^{-1} are initial embeddings of entity i, relation k, entity j, inverse relation k^{-1}, respectively. "$softmax$" represents the softmax function and $\mathbf{v}$ is an attention vector. "$gate$" denotes the gating unit which fuses information and h'_i, g'_k are the final representations of entity i and relation k for each attention head.

Here, we adopt the multi-head attention mechanism, and the ultimate representation can be concatenated in the following formula:

$$h_i = \Big\Vert_{m=1}^{M} h'_{i_m} \tag{11}$$

where M denotes the total number of attention heads and h'_{i_m} signifies the final representation of e_i within the m-th attention head.

D. Representation learning of relations

Similarly, for each relation $r_k \in \mathcal{R}$, we compute its relative importance. The same process is followed for the inverse relation $r_k^{-1} \in \mathcal{R}'$:

$$\gamma_{ijk} = \frac{exp(b_{ijk})}{\sum_{(e_m,e_n)\in\mathcal{N}_{r(k)}} exp(b_{mnk})}, (k \in \mathcal{R}) \tag{12}$$

where $\mathcal{N}_{r(k)}$ denotes the set of related head-tail entity pairs of r_k. Also $\delta_{ijk^{-1}}$ can be obtained by replacing $\mathcal{N}_{r(k)}$ with $\mathcal{N}_{r(k^{-1})}$.

Then we proceed to aggregate the representations of the related triples based on these importance values:

$$g'_{k_O} = f_r\left(\mathbf{W}_{\mathrm{R}} \sum_{(e_i,e_j)\in\mathcal{N}_{r(k)}} \gamma_{ijk} a_{ijk}\right) \tag{13}$$

$$g'_{k^{-1}} = f_r\left(\mathbf{W}_{\mathrm{R}'} \sum_{(e_i,e_j)\in\mathcal{N}_{r(k^{-1})}} \delta_{ijk^{-1}} a_{ijk^{-1}}\right) \tag{14}$$

where f_r represents a non-linear activation function of relations such as tanh. $\mathbf{W}_{\mathrm{R}}$ and $\mathbf{W}_{\mathrm{R}'}$ are the graph convolutional kernels.

Analogous to the approaches used for entities, we integrate a residual connection into the final representation and fuse it with a gating unit to enhance the model's performance:

$$g'_k = gate(\mathbf{W}_{\mathrm{r}} \cdot g_k, g'_{k_O} \| g'_{k^{-1}}) \tag{15}$$

where $\mathbf{W}_{\mathrm{r}}$ is a linear transformation matrix.

Ultimately, the combined representation of the relation r_k with M attention heads can be concatenated as described below:

$$g_k = \Big\Vert_{m=1}^{M} g'_{k_m} \tag{16}$$

E. Training

Score function. In our model, we utilize ConvE as the decoder to determine the plausibility scores for all potential triples. Specifically, as detailed in Equation 1, the score for the triple t_{ij}^k can be calculated in the following formula:

$$\varphi(e_i, r_k, e_j) = f(vec(f([\overline{h_i}; \overline{g_k}] * w))\mathbf{W})h_j \tag{17}$$

Loss function. The challenge of link prediction in KG can be framed as a binary classification task, where the goal is to distinguish between the presence and absence of a link. To achieve this, we adopt the binary cross-entropy loss function to train our model effectively.

$$\mathcal{L} = -\frac{1}{N}\sum_{o=1}^{N}\left(y_{t_{io}^k} log(\hat{y}_{t_{io}^k}) + (1 - y_{t_{io}^k}) log(1 - \hat{y}_{t_{io}^k})\right) \tag{18}$$

where N is the number of candidates of tail entities. $y_{t_{io}^k} \in \{0, 1\}$ represents the real label of triple t_{io}^k. Employing sigmoid to the score, $\hat{y}_{t_{io}^k} = sigmoid(\varphi(e_i, r_k, e_o)) \in [0, 1]$ can be obtained.

IV. Experiments and Analysis

A. Datasets

In this study, we carefully selected three high-quality experimental datasets including **FB15K-237** [14], **WN18RR** [13], **Kinship** [11] to thoroughly assess and validate the efficacy and robustness of our proposed model. These three benchmark datasets encompass a range of relation and entity counts, with their fundamental statistical details neatly outlined in Table I.

TABLE I
DATASET STATISTICS

Dataset	Entities	Relations	Triples			
			Train	Valid	Test	All
FB15k-237	14 541	237	272 115	17 535	20 466	310 116
WN18RR	40 943	11	86 835	3034	3134	93 003
Kinship	104	25	8544	1068	1074	10 686

TABLE II
PARTIAL HYPERPARAMETER SETTINGS FOR EACH DATASET

Dataset	WN18RR	FB15k-237	Kinship
Learning rate	1e-4	1e-4	1e-4
Epochs	800	1000	1000
Batch size	128	128	128
Embedding dimension	200	300	300
GCN dropout rate	0.5	0.6	0.4
Attention heads	1	3	3

B. Evaluation metrics

In this paper, we evaluate our model using the rankings of triples. We assess performance of our model with simple metrics such as Mean Reciprocal Rank (MRR), Mean Rank (MR), and Hits@N, all based on how well the model ranks the correct triples.

C. Baselines and experimental settings

Here, we introduce a few baseline models tailored for the task of link prediction in KGs, to validate the effectiveness of our proposed model. These baselines are outlined as follows:

- TransE [6]: A translational model that represents entities and relations in vectors, minimizing the distance between them through translations to model relations.
- DistMult [15]: A tensor factorization model using bilinear scoring via matrix multiplication to represent entity-relation intricacies.
- ComplEx [16]: Extends DistMult to complex numbers, enabling capture of asymmetric and intricate relations between entities and relations.
- ConvE [13]: A CNN-based model applying convolutional operations to process relations and entities, inferring missing links through neural networks.
- WGCN [17]: A GCN variant modeling relation differences with weighted graph convolutions and learnable scalar weights for neighborhood messages.
- ComplexGCN [18]: Utilizes complex domain convolutions in GCNs to capture knowledge representations of entities and relations.
- MSHE [19]: a novel link prediction framework for knowledge graph embedding, leveraging multi-source and hierarchical neural networks to integrate complex knowledge.
- D-AEN [20]: A dual-attention embedding network for KGE, fusing neighborhood info with bidirectional and relation-specific attentions to propagate and update representations.

Our model implementation relies on the open-source *PyTorch* platform, executed on an Ubuntu 18.04 server powered by an *Intel Xeon Silver 4210R CPU* and an *NVIDIA GeForce RTX 3090 GPU*. The hyperparameters tailored for each dataset are detailed in Table II.

D. Results and discussions

Table III showcases the impressive link prediction capabilities of our KAAGE model across three renowned benchmark datasets. By benchmarking against various baseline models, we confirm that KAAGE excels, particularly on the Kinship dataset.

Specifically, by observing the performance of these models, we can conclude that: Our model performs best on most evaluation metrics of the three datasets. On the FB15k-237 dataset, ComplexGCN has the best Hits@1 and the second-best MRR because ComplexGCN maps entities and relations to the complex space, which has better representation capabilities for the multi-type and general-domain knowledge graph dataset such as FB15k-237. However, our KAAGE is still optimal on the other three metrics. Similarly, D-AEN, which reverses relations and triples, does not use a gating mechanism when fusing heterogeneous information, but uses simple addition, resulting in slightly inferior performance.

V. CONCLUSIONS

In this article, we introduce KAAGE, a novel GAT-based approach for link prediction. KAAGE augment knowledge by inversing relations and triples, leveraging attention mechanism to fuse triple representations, capturing rich semantics. It merges original and augmented knowledge via a gating unit, balancing structural information. Extensive experiments on three benchmarks reveal KAAGE's superiority, especially on Kinship.

ACKNOWLEDGMENT

This work is supported by the National Nature Science Foundation of China (Grant No: 61901116, Grant No: 61471229 and Grant No:62001115), the Guangdong Basic and Applied Basic Research Foundation (Grant No: 2023A1515140111, Grant No: 2019A1515010789, Grant No: 2021A1515012289 and Grant No: 2019A1515110136) and the Jihua Laboratory Scienctific project (Grant No: X210101UZ210).

REFERENCES

[1] C. Xiong, R. Power, and J. Callan, "Explicit semantic ranking for academic search via knowledge graph embedding," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17, International World Wide Web Conferences Steering Committee, Apr. 2017, pp. 1271–1279.

[2] Z. Liu, C. Xiong, M. Sun, and Z. Liu, "Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval," 2018.

TABLE III
HITS@1, HITS@3, HITS@10, MR, AND MRR RESULTS FOR LINK PREDICTION ON WN18RR, FB15K-237 AND KINSHIP. BOLD SCORES REPRESENT THE BEST RESULTS, AND UNDERLINED SCORES INDICATE THE SECOND-BEST RESULTS. "H@N" IS SHORT FOR "HITS@N"

	WN18RR					FB15k-237					Kinship				
Models	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10	MR	MRR
TransE	-	-	0.501	3384	0.226	-	-	0.465	357	0.294	0.009	0.643	0.841	3.8	0.309
DistMult	0.39	0.44	0.49	5110	0.43	0.155	0.263	0.419	254	0.241	0.367	0.581	0.867	5.26	0.516
ComplEx	0.41	0.46	0.51	5261	0.44	0.158	0.275	0.428	339	0.247	0.733	0.899	0.971	2.48	0.823
ConvE	0.39	0.43	0.48	5277	0.46	0.239	0.350	0.491	246	0.316	0.73	0.91	0.98	2	0.83
WGCN	0.43	0.48	0.54	-	0.47	0.26	0.39	0.54	-	0.35	-	-	-	-	-
ComplexGCN	0.245	0.371	0.524	-	0.338	**0.423**	0.468	0.516	-	0.455	-	-	-	-	-
MSHE	0.43	0.48	0.54	-	0.47	0.26	0.39	0.54	-	0.35	-	-	-	-	-
D-AEN	**0.443**	0.500	0.561	2248	0.484	0.337	0.471	0.611	164	0.429	0.968	0.984	**0.990**	**1.52**	0.977
KAAGE	0.442	**0.509**	**0.564**	**1969**	**0.487**	0.365	**0.500**	**0.634**	**145**	**0.456**	**0.980**	**0.986**	**0.990**	1.65	**0.983**

[3] X. Dong, E. Gabrilovich, G. Heitz, *et al.*, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 601–610.

[4] Q. Liu, Y. Li, H. Duan, Y. Liu, and Z. Qin, "A survey on knowledge graph construction techniques," *Journal of Computer Research and Development*, vol. 53, pp. 582–600, 2016, ISSN: 1000-1239.

[5] Z. Liu, M. Sun, Y. Lin, and R. Xie, "Progress in knowledge representation learning," *Journal of Computer Research and Development*, vol. 53, pp. 247–261, 2016, ISSN: 1000-1239.

[6] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in neural information processing systems*, vol. 26, 2013.

[7] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor, "Hinge-loss markov random fields and probabilistic soft logic," *Journal of Machine Learning Research*, vol. 18, no. 109, pp. 1–67, 2017.

[8] Q. Wang, J. Liu, Y. Luo, B. Wang, and C.-Y. Lin, "Knowledge base completion via coupled path ranking," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1308–1318.

[9] M. Nickel, V. Tresp, H.-P. Kriegel, *et al.*, "A three-way model for collective learning on multi-relational data.," in *Icml*, vol. 11, 2011, pp. 3 104 482–3 104 584.

[10] L. Guo, Q. Zhang, W. Ge, W. Hu, and Y. Qu, "Dskg: A deep sequential model for knowledge graph completion," in *Knowledge Graph and Semantic Computing. Knowledge Computing and Language Understanding: Third China Conference, CCKS 2018, Tianjin, China, August 14–17, 2018, Revised Selected Papers*, Springer, 2019, pp. 65–77.

[11] X. V. Lin, R. Socher, and C. Xiong, "Multi-hop knowledge graph reasoning with reward shaping," *arXiv preprint arXiv:1808.10568*, 2018.

[12] S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar, "Composition-based multi-relational graph convolutional networks," *arXiv preprint arXiv:1911.03082*, 2019.

[13] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2d knowledge graph embeddings," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[14] K. Toutanova and D. Chen, "Observed versus latent features for knowledge base and text inference," in *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, 2015, pp. 57–66.

[15] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," *arXiv preprint arXiv:1412.6575*, 2014.

[16] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *International conference on machine learning*, PMLR, 2016, pp. 2071–2080.

[17] C. Shang, Y. Tang, J. Huang, J. Bi, X. He, and B. Zhou, "End-to-end structure-aware convolutional networks for knowledge base completion," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 3060–3067.

[18] A. Zeb, S. Saif, J. Chen, A. U. Haq, Z. Gong, and D. Zhang, "Complex graph convolutional network for link prediction in knowledge graphs," *Expert Systems with Applications*, vol. 200, p. 116 796, 2022.

[19] D. Jiang, R. Wang, L. Xue, and J. Yang, "Multisource hierarchical neural network for knowledge graph embedding," *Expert Systems with Applications*, vol. 237, p. 121 446, 2024.

[20] H. Fang, Y. Wang, Z. Tian, and Y. Ye, "Learning knowledge graph embedding with a dual-attention embedding network," *Expert Systems with Applications*, vol. 212, p. 118 806, 2023.

2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Low-Complexity Adaptive Beamformer for Joint Reverberation and Noise Suppression

Fan Zhang*, Chao Pan*, Jingdong Chen*, and Jacob Benesty†
*Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an, China
†INRS-EMT, University of Quebec, Canada

***Abstract*—Reverberation and background noise can severely affect the quality and intelligibility of recorded speech, potentially impairing speech communication and human-machine interaction systems. The minimum variance distortionless response (MVDR) beamformer is commonly used to jointly suppress reverberation and noise, but it is computationally intensive due to the need for matrix inversion at each time frame and subband. In this paper, we introduce a beamforming method that combines an MVDR beamformer optimized for noise reduction with a fixed maximum directivity-factor beamformer for reverberation suppression. This hybrid beamformer offers a more efficient implementation than the traditional MVDR beamformer and adapts to varying levels of reverberation and noise by adjusting a single weighting factor. Simulation results demonstrate that the proposed beamformer achieves significantly faster processing speeds compared to the optimal MVDR beamformer, with only minor performance degradation in noise reduction and reverberation suppression.**

I. Introduction

In hands-free applications like speech communication or recognition systems, the speech signal captured by a distant microphone can be affected by pervasive background noise and reverberation due to multi-path propagation in the acoustic environment. These factors can significantly degrade speech quality, intelligibility, and the performance of speech communication and human-machine interaction systems.

When a microphone array is available, beamforming techniques [1–4] are commonly used to recover the desired signal from noisy and reverberant multichannel observations. A prominent method is the Wiener beamformer [4, 5], which seeks to minimize the mean-squared error between the desired speech component and the beamformer's output. Reverberation is often modeled as a diffuse sound field [5–11], leading the Wiener beamformer to be expressed as a function of the (time-varying) noise covariance matrix and the (time-varying) reverberation variance. This approach can be decomposed into two main components: 1) a minimum variance distortionless response (MVDR) beamformer, which aims to minimize the sum of the noise and reverberation variances in the output, while keeping the desired signal undistorted; and 2) a single-channel Wiener post-filter [12, 13] that further reduces residual noise and reverberation, albeit with some distortion to the desired signal. One challenge with the MVDR beamformer is that its implementation requires inverting the sum of two full-rank matrices for every time frame and subband, which can impose a significant computational burden, particularly with large arrays containing many microphones.

This work was supported in part by the NSFC (No. 62171373).

To tackle the computational challenges of the MVDR beamformer, we introduce a distortionless beamformer that combines a fixed maximum directivity-factor beamformer with an MVDR beamformer specifically optimized for noise reduction. Although the concept of using a weighted sum of different sub-beamformers has been employed in various applications, such as adaptive differential microphone arrays [14–16], robust superdirective beamforming [17, 18], and speech enhancement in underdetermined scenarios [19], it has not been previously applied to joint noise and reverberation suppression with a focus on computational efficiency. Simulation results demonstrate that the proposed combined beamformer provides significantly faster processing speeds, particularly with large microphone arrays, while only marginally sacrificing performance compared to the traditional MVDR beamformer.

II. Signal Model and Problem Formulation

Consider a reverberant environment where a speech signal from a distant source is captured by a uniform linear array (ULA) of M omnidirectional microphones. In the short-time Fourier transform (STFT) domain, the observed signal vector of length M can be represented as

$$\mathbf{y}(\ell) = \mathbf{d}_{\theta_S} S(\ell) + \mathbf{r}(\ell) + \mathbf{v}(\ell), \tag{1}$$

where ℓ is the time-frame index, $S(\ell)$ is the direct-path speech signal at the reference microphone, $\mathbf{d}_{\theta_S}$ is the signal propagation vector of the speech source, analogous to a typical steering vector, $\mathbf{r}(\ell)$ is the reverberation vector, and $\mathbf{v}(\ell)$ is the additive noise vector. Note that the STFT subband index is omitted in this model for simplicity, as the subsequent processing occurs on a subband basis, with each subband processed independently using the same algorithm. In the field of speech enhancement, it is common to assume that $S(\ell)$, $\mathbf{r}(\ell)$, and $\mathbf{v}(\ell)$ have zero mean, and are mtually uncorrelated [5]. Under these assumptions, the covariance matrix of $\mathbf{y}(\ell)$ can be expressed as

$$\mathbf{\Phi}_{\mathbf{y}}(\ell) = \phi_S(\ell)\mathbf{d}_{\theta_S}\mathbf{d}_{\theta_S}^H + \mathbf{\Phi}_{\mathbf{r}}(\ell) + \mathbf{\Phi}_{\mathbf{v}}(\ell), \tag{2}$$

where $\phi_S(\ell)$ is the variance of $S(\ell)$, the superscript ^{H} denotes the conjugate transpose, and $\mathbf{\Phi}_{\mathbf{r}}(\ell)$ and $\mathbf{\Phi}_{\mathbf{v}}(\ell)$ are the covariance matrices of $\mathbf{r}(\ell)$ and $\mathbf{v}(\ell)$, respectively. The reverberation is usually modeled as a spherically isotropic (diffuse) noise [20]. Hence,

$$\mathbf{\Phi}_{\mathbf{r}}(\ell) = \phi_R(\ell)\mathbf{\Gamma}_{\mathrm{d}}, \tag{3}$$

where $\phi_R(\ell)$ is the common variance of the elements in $\mathbf{r}(\ell)$, and $\mathbf{\Gamma}_{\mathrm{d}}$ is the coherence matrix for the diffuse field, which can be determined based on the array geometry information.

With the above signal model, beamforming with a filter $\mathbf{h}(\ell)$ is typically employed to estimate the desired signal $S(\ell)$ from the multichannel observations $\mathbf{y}(\ell)$, which can be written as

$$\begin{aligned} Z(\ell) &= \mathbf{h}^H(\ell)\mathbf{y}(\ell) \\ &= S_{\mathrm{o}}(\ell) + R_{\mathrm{o}}(\ell) + V_{\mathrm{o}}(\ell), \end{aligned} \tag{4}$$

where $Z(\ell)$ is the beamformer output, and

$$S_{\mathrm{o}}(\ell) \triangleq \mathbf{h}^H(t)\mathbf{d}_{\theta_S} S(\ell), \tag{5}$$

$$R_{\mathrm{o}}(\ell) \triangleq \mathbf{h}^H(\ell)\mathbf{r}(\ell), \tag{6}$$

$$V_{\mathrm{o}}(\ell) \triangleq \mathbf{h}^H(\ell)\mathbf{v}(\ell) \tag{7}$$

denote the components in the beamformer's output that correspond to the direct-path signal, reverberation, and noise, respectively.

From (4), we can define the narrowband output signal-to-reverberation ratio (SRR), output signal-to-noise ratio (SNR), and output signal-to-reverberation-plus-noise ratio (SRNR) as follows:

$$\mathrm{SRR}_{\mathrm{o}}\left[\mathbf{h}(\ell)\right] \triangleq \frac{\phi_{S_{\mathrm{o}}}(\ell)}{\phi_{R_{\mathrm{o}}}(\ell)}, \tag{8}$$

$$\mathrm{SNR}_{\mathrm{o}}\left[\mathbf{h}(\ell)\right] \triangleq \frac{\phi_{S_{\mathrm{o}}}(\ell)}{\phi_{V_{\mathrm{o}}}(\ell)}, \tag{9}$$

$$\mathrm{SRNR}_{\mathrm{o}}\left[\mathbf{h}(\ell)\right] \triangleq \frac{\phi_{S_{\mathrm{o}}}(\ell)}{\phi_{R_{\mathrm{o}}}(\ell) + \phi_{V_{\mathrm{o}}}(\ell)}, \tag{10}$$

where

$$\phi_{S_{\mathrm{o}}}(\ell) = \phi_S(\ell)\left|\mathbf{h}^H(\ell)\mathbf{d}_{\theta_S}\right|^2, \tag{11}$$

$$\phi_{R_{\mathrm{o}}}(\ell) = \phi_R(\ell)\mathbf{h}^H(\ell)\mathbf{\Gamma}_{\mathrm{d}}\mathbf{h}(\ell), \tag{12}$$

$$\phi_{V_{\mathrm{o}}}(\ell) = \mathbf{h}^H(\ell)\mathbf{\Phi}_{\mathbf{v}}(\ell)\mathbf{h}(\ell) \tag{13}$$

are the variances of $S_{\mathrm{o}}(\ell)$, $R_{\mathrm{o}}(\ell)$, and $V_{\mathrm{o}}(\ell)$, respectively.

The MVDR beamformer for joint speech dereverberation and noise reduction is given by [5]

$$\mathbf{h}_{\mathrm{MVDR}}(\ell) = \frac{\left[\phi_R(\ell)\mathbf{\Gamma}_{\mathrm{d}} + \mathbf{\Phi}_{\mathbf{v}}(\ell)\right]^{-1}\mathbf{d}_{\theta_S}}{\mathbf{d}_{\theta_S}^H\left[\phi_R(\ell)\mathbf{\Gamma}_{\mathrm{d}} + \mathbf{\Phi}_{\mathbf{v}}(\ell)\right]^{-1}\mathbf{d}_{\theta_S}}. \tag{14}$$

This beamformer is optimal in terms of maximizing the narrowband output SRNR defined in (10). However, its practical implementation involves inverting the matrix within the brackets for each time frame and subband, which has a computational complexity of $\mathcal{O}(M^3)$. This can impose a significant computational burden for real-time applications, particularly because no fast algorithms are currently available.

In the following section, we will present a combined beamformer that reduces computational complexity while effectively suppressing both reverberation and noise.

III. Combined Beamforming

In this section, we analyze the behavior of the MVDR beamformer in (14) under two extreme scenarios to identify two beamformers that individually maximize the narrowband output SNR and output SRR. By linearly combining these two beamformers, we create a new beamformer that depends on a real-valued weighting factor. We then discuss how to determine the optimal value for this weighting factor by minimizing the variance of the beamformer output.

A. Two Sub-Beamformers

We define the narrowband input reverberation-to-noise ratio (RNR) at frame ℓ as

$$\mathrm{RNR}(\ell) \triangleq \frac{\phi_R(\ell)}{\phi_V(\ell)}, \tag{15}$$

where $\phi_V(\ell)$ is the noise variance at frame ℓ. We now consider two extreme scenarios.

- When $\mathrm{RNR}(\ell) = 0$, meaning there is no reverberation, the MVDR beamformer in (14) degenerates to [21, 22]

$$\mathbf{h}_{\mathrm{MVDR,NR}}(\ell) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(\ell)\mathbf{d}_{\theta_S}}{\mathbf{d}_{\theta_S}^H\mathbf{\Phi}_{\mathbf{v}}^{-1}(\ell)\mathbf{d}_{\theta_S}}, \tag{16}$$

  which is the MVDR beamformer designed for noise reduction only. This beamformer maximizes the output SNR defined in (9).
- When $\mathrm{RNR}(\ell)$ approaches infinity, meaning there is no additive noise, the MVDR beamformer in (14) reduces to [21, 23]

$$\mathbf{h}_{\mathrm{MDF}} = \frac{\mathbf{\Gamma}_{\mathrm{d}}^{-1}\mathbf{d}_{\theta_S}}{\mathbf{d}_{\theta_S}^H\mathbf{\Gamma}_{\mathrm{d}}^{-1}\mathbf{d}_{\theta_S}}, \tag{17}$$

  which is the maximum-directivity-factor (MDF) beamformer. This beamformer maximizes the output SRR defined in (8).

The two beamformers in (16) and (17) will serve as sub-beamformers in constructing the combined beamformer.

B. Combined Beamformer

We construct a new beamformer by linearly combining the two beamformers in (16) and (17):

$$\mathbf{h}_{\alpha}(\ell) = \alpha(\ell)\mathbf{h}_{\mathrm{MVDR,NR}}(\ell) + \left[1 - \alpha(\ell)\right]\mathbf{h}_{\mathrm{MDF}}, \tag{18}$$

where $0 \leq \alpha(\ell) \leq 1$ is a real weighting factor used to balance between noise reduction and reverberation suppression. It is straightforward to verify that when $\alpha(\ell) = 0$, we obtain the MDF beamformer, while setting $\alpha(\ell) = 1$ yields the noise reduction MVDR beamformer in (16).

The MDF beamformer is a fixed beamformer with time-invariant filter coefficients. In contrast, the noise reduction MVDR beamformer involves computing the inverse of the noise covariance matrix. This inversion can be efficiently performed using the matrix inversion lemma [2], which has a complexity of $\mathcal{O}(M^2)$. Consequently, from a computational complexity standpoint, the combined beamformer $\mathbf{h}_{\alpha}(\ell)$ presents a valuable alternative to the MVDR beamformer in (14).

C. Determination of the Optimal Value of $\alpha(\ell)$

The goal of this section is to determine the optimal value of $\alpha(\ell)$ for the combined beamformer to maximize the output S-RNR, as defined in (10). Given that the combined beamformer is distortionless, this optimization problem can be equivalently framed as minimizing the variance of its output.

The output of the combined beamformer $\mathbf{h}_\alpha(\ell)$ can be expressed as

$$\begin{aligned} Z(\ell) &= \alpha(\ell) Z_{\mathrm{NR}}(\ell) + [1-\alpha(\ell)]\, Z_{\mathrm{DR}}(\ell) \\ &= Z_{\mathrm{DR}}(\ell) - \alpha(\ell) U(\ell), \end{aligned} \tag{19}$$

where

$$Z_{\mathrm{NR}}(\ell) = \mathbf{h}_{\mathrm{MVDR,NR}}^H(\ell)\mathbf{y}(\ell), \tag{20}$$

$$Z_{\mathrm{DR}}(\ell) = \mathbf{h}_{\mathrm{MDF}}^H\mathbf{y}(\ell) \tag{21}$$

are the outputs of the noise reduction MVDR beamformer in (16) and the MDF beamformer, respectively, and

$$\begin{aligned} U(\ell) &= Z_{\mathrm{DR}}(\ell) - Z_{\mathrm{NR}}(\ell) \\ &= \left[\mathbf{h}_{\mathrm{MDF}} - \mathbf{h}_{\mathrm{MVDR,NR}}(\ell)\right]^H \left[\mathbf{r}(\ell) + \mathbf{v}(\ell)\right], \end{aligned} \tag{22}$$

The term $U(\ell)$ does not include any component of the desired signal, and can thus be considered a reference signal used for noise cancellation in $Z_{\mathrm{DR}}(\ell)$ [24]. Leveraging the fact that $\alpha(\ell)$ is real-valued, the variance of $Z(\ell)$ in (19) can be expressed as

$$\phi_Z(\ell) = \alpha^2(\ell)\phi_U(\ell) - 2\alpha(\ell)\Re\left[\phi_{U,Z_{\mathrm{DR}}}(\ell)\right] + \phi_{Z_{\mathrm{DR}}}(\ell), \tag{23}$$

where $\phi_U(\ell)$ is the variance of $U(\ell)$, $\phi_{U,Z_{\mathrm{DR}}}(\ell)$ is the covariance between $U(\ell)$ and $Z_{\mathrm{DR}}(\ell)$, $\Re[\cdot]$ is the real part operator, and $\phi_{Z_{\mathrm{DR}}}(\ell)$ is the variance of $Z_{\mathrm{DR}}(\ell)$.

The variance $\phi_Z(\ell)$ in (23) is a quadratic function of $\alpha(\ell)$. Given the constraint that $\alpha(\ell)$ must lie between 0 and 1, the optimal value of $\alpha(\ell)$ can be determined by

$$\alpha_{\mathrm{c}}(\ell) = \begin{cases} 0 & \alpha_{\mathrm{uc}}(\ell) < 0 \\ 1 & \alpha_{\mathrm{uc}}(\ell) > 1 \\ \alpha_{\mathrm{uc}}(\ell) & \text{otherwise} \end{cases}, \tag{24}$$

where

$$\alpha_{\mathrm{uc}}(\ell) = \frac{\Re\left[\phi_{U,Z_{\mathrm{DR}}}(\ell)\right]}{\phi_U(\ell)}, \tag{25}$$

and the subscripts $_{\mathrm{c}}$ and $_{\mathrm{uc}}$ denote "constrained" and "unconstrained," respectively. In practice, the quantities $\phi_U(\ell)$ and $\phi_{U,Z_{\mathrm{DR}}}(\ell)$ are typically unknown and can be estimated recursively as follows:

$$\widehat{\phi}_U(\ell) = \lambda_r\widehat{\phi}_U(\ell-1) + (1-\lambda_r)\left|U(\ell)\right|^2, \tag{26}$$

$$\widehat{\phi}_{U,Z_{\mathrm{DR}}}(\ell) = \lambda_r\widehat{\phi}_{U,Z_{\mathrm{DR}}}(\ell-1) + (1-\lambda_r)U(\ell)Z_{\mathrm{DR}}^*(\ell), \tag{27}$$

where $0 < \lambda_r < 1$ is a smoothing factor, and the superscript * denotes complex conjugate.

IV. Simulations

In this section, we assess the performance of the combined beamformer in simulated reverberant and noisy acoustic environments.

A. Setup

We consider a room with dimensions of $5 \times 6 \times 4$ meters, where the reflection coefficients are uniformly set at 0.85. This configuration results in a reverberation time (T_{60}) of approximately 475 ms. A uniform linear array (ULA) with an inter-element spacing of 2 cm is placed in the room at a height of 1.4 m. A speech source is located directly in front of the ULA, at a distance of 1.5 m from the array's center. Additionally, an interference source is positioned at an angle of 180 relative to the ULA. Both sources are at the same height as the array. The impulse responses between each source and the sensors are simulated using the image-model method [25, 26].

The clean speech signals are obtained from the NOIZEUS database [27, 28], which are downsampled from their original higher sampling rate to 16 kHz. Given that the database contains only short samples, we compile 30 clean speech recordings from the same speaker and concatenate them into six longer segments by randomly combining five recordings each. The reverberant speech signals are created by convolving these source signals with the room impulse responses. Similarly, the interference signal is generated by convolving an air-conditioner noise recording, captured in an office environment, with the room impulse responses from the interference source position to the microphone sensors' positions. These reverberant speech and interference signals are then mixed, and white Gaussian noise is added to produce the final observation signals. The power ratio between the interference signal and the white Gaussian noise is set to 10 dB.

For evaluation, the impulse response is split into two components: the direct path and reflections. We derive the direct-path signal and reverberation by convolving the clean speech signal with these respective impulse response segments. This allows us to calculate broadband input SNR, SRR, and SRNR. The improvements in SNR, SRR, and SRNR, denoted as $\Delta\mathrm{SNR}$, $\Delta\mathrm{SRR}$, and $\Delta\mathrm{SRNR}$, are used to measure the extent of noise and reverberation attenuation.

To apply beamforming in the STFT domain, we first divide the observation signals into overlapping frames, each consisting of 256 points with a 75% overlap. A Hamming window of length 256 is then applied to each frame, and the resulting windowed frames are transformed into the STFT domain using the fast Fourier transform (FFT). Our proposed beamformer combines both the MDF beamformer and the noise reduction MVDR beamformer. To enhance robustness, we use a regularized MDF beamformer with a regularization parameter set to 10^{-4} [21]. For the noise reduction MVDR beamformer, we need to estimate the noise covariance matrix. We utilize an oracle estimator that assumes noise samples are available, providing an upper bound on performance. The inverse of the estimated noise covariance matrix is calculated using the matrix inversion lemma [2]:

$$\widehat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(\ell) = \frac{1}{\lambda_v}\left[\widehat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(\ell-1) - \frac{\widetilde{\mathbf{v}}(\ell)\widetilde{\mathbf{v}}^H(\ell)}{\mathbf{v}^H(\ell)\widetilde{\mathbf{v}}(\ell) + \widetilde{\lambda}_v}\right], \tag{28}$$

where $\widetilde{\mathbf{v}}(\ell) \triangleq \widehat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(\ell-1)\mathbf{v}(\ell)$ and $\widetilde{\lambda}_v \triangleq \lambda_v/(1-\lambda_v)$. In practice, the noise covariance matrix is typically updated when

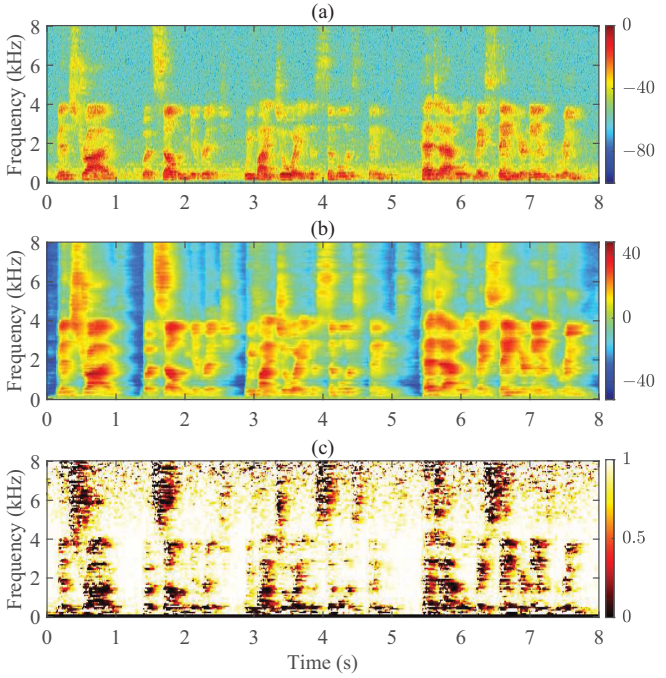


Fig. 1. Example of: (a) the spectrogram of the observed signal at sensor 1, (b) the narrowband input RNR, and (c) the estimate of α_c. The input SNR is 15 dB and $M = 6$.

speech is absent, which requires a voice activity detector or a more advanced speech presence probability estimator [29–33]. In our simulations, we empirically set the smoothing factors: λ_r and λ_v to 0.85 and 0.98, respectively.

For comparison purposes, we select the MVDR beamformer described in (14) as our baseline algorithm. Implementing this beamformer requires an estimate of the reverberation variance, denoted as $\phi_R(\ell)$. Several methods for estimating $\phi_R(\ell)$ have been proposed in the literature, such as those detailed in [5] and its references. To achieve optimal performance, we use an oracle estimator, which assumes that reverberation samples are available.

B. Simulation Results

The initial set of simulations is designed to evaluate the performance of the proposed beamformer. Figure 1(c) presents a plot of the estimated parameter α_c. Comparing this with the narrowband input RNR shown in Figure 1(b), we observe that α_c tends to be larger in regions with low RNR and smaller in regions with high RNR. Consequently, the proposed beamformer behaves similarly to the noise reduction MVDR beamformer in situations where noise predominates over reverberation, and approximates the MDF beamformer when reverberation is more significant compared to noise.

Figure 2 shows the performance of various beamformers as a function of input SNR with $M = 6$, based on averaging results from six speech source signals. The noise reduction MVDR beamformer provides the highest SNR gain, while the MDF beamformer achieves the most significant SRR improvement. Althoughthe hybrid beamformer performs slightly worse in SRNR improvement compared to the MVDR beamformer in

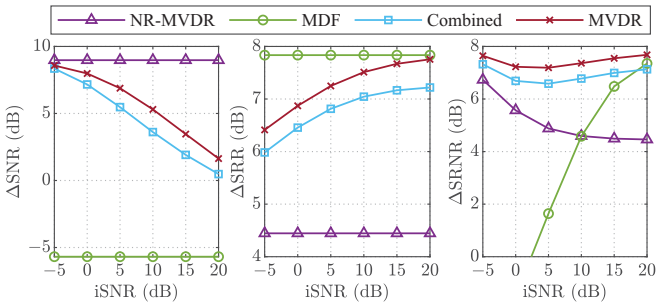


Fig. 2. Speech enhancement performance of various beamforming methods as a function of the input SNR. The number of microphones is $M = 6$.

TABLE I
REAL-TIME FACTORS OF TWO BEAMFORMERS FOR DIFFERENT VALUES OF M. THE INPUT SNR IS 15 DB.

M	2	4	8	16
Proposed	**0.22**	**0.24**	**0.30**	**0.75**
MVDR in (14)	0.34	0.40	0.53	1.39

(14), it generally outperforms both of its constituent sub-beamformers, which focus exclusively on either speech dereverberation or noise reduction.

The final set of simulations is conducted to compare the computational complexity of the proposed beamformer with that of the MVDR beamformer in (14). The real-time factor (RTF) [30, 34] is defined as the ratio between the processing time of an algorithm/system and the duration of the input signal, with a smaller RTF indicating a faster algorithm. For real-time applications, it is essential to maintain an RTF value below one. Table I shows the average RTF values for both beamformers across different numbers of microphones. The simulations were performed on a laptop PC (AMD Ryzen 5 4600H CPU at 3.00 GHz, 16.0 GB RAM). The results demonstrate that the proposed beamformer consistently achieves a lower RTF compared to the MVDR beamformer, with the difference becoming more significant as the number of microphones increases.

V. Conclusions

This paper presented a beamformer designed to reduce computational cost while simultaneously suppressing both reverberation and noise. The proposed beamformer is created by linearly combining a noise-reduction MVDR beamformer and a fixed MDF beamformer using an optimal real-valued weighting factor. We derived this optimal weighting factor by minimizing the variance of the beamformer output. By applying the matrix inversion lemma, the proposed beamformer achieves enhanced implementation efficiency. Speech enhancement simulations show that the proposed beamformer outperforms both the noise-reduction MVDR and MDF beamformers, and performs only slightly below the jointly optimized MVDR beamformer.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[2] H. L. Van Trees, *Optimum Array Processing*. John Wiley & Sons, 2002.

[3] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 945–958, May 2013.

[4] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 240–251, Feb. 2015.

[5] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1056–1071, June 2018.

[6] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2374–2384, Nov. 2011.

[7] S. Braun and E. A. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. EUSIPCO*, pp. 1–5, 2013.

[8] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. EUSIPCO*, pp. 61–65, 2014.

[9] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 1006–1018, June 2015.

[10] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," in *Proc. IEEE ICASSP*, pp. 151–155, 2016.

[11] O. Schwartz, S. Gannot, and E. A. Habets, "Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm," in *Proc. EUSIPCO*, pp. 1123–1127, 2016.

[12] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays* (M. Brandstein and D. Ward, eds.), ch. 3, pp. 39–60, Springer, 2001.

[13] Y. A. Huang, A. Luebs, J. Skoglund, and W. B. Kleijn, "Globally optimized least-squares post-filtering for microphone array speech enhancement," in *Proc. IEEE ICASSP*, pp. 380–384, 2016.

[14] G. W. Elko and A.-T. N. Pong, "A simple adaptive first-order differential microphone," in *Proc. IEEE WASPAA*, pp. 169–172, 1995.

[15] G. W. Elko and J. Meyer, "Second-order differential adaptive microphone array," in *Proc. IEEE ICASSP*, pp. 73–76, 2009.

[16] J. Jin, X. Luo, G. Huang, J. Chen, and J. Benesty, "Beamforming through online convex combination of differential beamformers," in *Proc. IEEE ICASSP*, pp. 8561–8565, 2024.

[17] R. Berkun, I. Cohen, and J. Benesty, "Combined beamformers for robust broadband regularized superdirective beamforming," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 877–886, May 2015.

[18] W. Meng, C. Zheng, and X. Li, "Fully automatic balance between directivity factor and white noise gain for large-scale microphone arrays in diffuse noise fields," in *Interspeech*, pp. 5433–5437, 2022.

[19] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3461–3475, 2021.

[20] F. Jacobsen and T. Roisin, "The coherence of reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 108, no. 1, pp. 204–210, 2000.

[21] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. John Wiley & Sons, 2018.

[22] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 67–79, Jan. 2014.

[23] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 393–398, June 1986.

[24] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propag.*, vol. 30, pp. 27–34, Jan. 1982.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, Apr. 1979.

[26] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Berlin, Germany: Springer-Verlag, 2006.

[27] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7, pp. 588–601, 2007.

[28] P. C. Loizou, *Speech Enhancement: Theory and Practice (2nd ed.)*. CRC Press, 2013.

[29] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2159–2169, Sep. 2011.

[30] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE ICASSP*, pp. 5210–5214, 2016.

[31] M. Taseska and E. A. P. Habets, "Nonstationary noise PSD matrix estimation for multichannel blind speech extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 2223–2236, Nov. 2017.

[32] J. M. Mart n-Doñas, J. Jensen, Z.-H. Tan, A. M. Gomez, and A. M. Peinado, "Online multichannel speech enhancement based on recursive EM and DNN-based speech presence estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 3080–3094, 2020.

[33] F. Zhang, C. Pan, J. Benesty, and J. Chen, "Directional gain based noise covariance matrix estimation for MVDR beamforming," in *Proc. IEEE ICASSP*, pp. 511–515, 2024.

[34] O. Cheng, W. Abdulla, and Z. Salcic, "Hardware-software codesign of automatic speech recognition system for embedded real-time applications," *IEEE Trans. Ind. Electron.*, vol. 58, pp. 850–859, Mar. 2011.

Graph Filter Transfer for Time-Varying Signal Estimation Between Two Networks

Tsutahiro Fukuhara, Junya Hara, Hiroshi Higashi, Yuichi Tanaka
Graduate School of Engineering, Osaka University, Japan
E-mail: t.fukuhara@msp-lab.org, {j.hara, higashi, ytanaka}@comm.eng.osaka-u.ac.jp

***Abstract*—This paper presents a filter transfer method for estimating time-varying signals, i.e., Kalman filtering between two different networks. In many sensor networks, observed signals are associated with nodes (i.e., sensors), and edges of the network represent the inter-node connectivity. For a large sensor network, measuring the signal values at all nodes requires huge resources, particularly in terms of energy consumption. To alleviate the issue, one may extract one cluster from the network and perform intra-cluster analysis based on the statistics in the cluster. The statistics are then utilized to estimate the signals from another cluster. This leads to the requirement for transferring a set of parameters in the Kalman filter from one cluster to another. In this paper, we propose a cooperative Kalman filter between two networks. The proposed Kalman filter alternately estimates signals in time between the two networks. We formulate a state-space model in the source cluster and transfer it to the target cluster on the basis of optimal transport. In the signal estimation experiments, we validate the effectiveness of the proposed method.**

I. Introduction

Sensor networks are used in various disciplines to analyze sensor data using the interrelationships among sensors [1], [2], where their nodes observe the signals of the corresponding sensors and their edges represent the inter-node connectivity. Their applications include bottleneck detection in traffic networks and leakage detection in infrastructure networks [3], [4].

A network often has a sub-network having similar statistics called a *community* or *cluster*. In many clustered sensor networks, sensor data in one cluster could impact on those in different clusters as well as those obtained at the previous time instances. Therefore, predictive control (PC) for networks among clusters is crucial and has been extensively studied in many application fields [5], [6].

Kalman filter is the most popular PC method for networks [7], [8]. It linearly tracks and estimates *time-varying* signals on a *static* network by minimizing the mean squared error (MSE) between estimated and original signals. The system of Kalman filter is modeled by a state-space model and its estimator is performed in prediction and update steps.

We often encounter large networks, however, observing all of their signals at every time instance may be costly in terms of storage burdens and energy consumption, which may also shorten the lifetime of sensors [9]. To alleviate this, one may extract one cluster from the network and perform an intra-cluster analysis based on the statistics in the cluster [10]. Since one cluster could affect another as previously mentioned, a transfer method for a set of parameters on the PC is required, especially for clusters having different sizes.

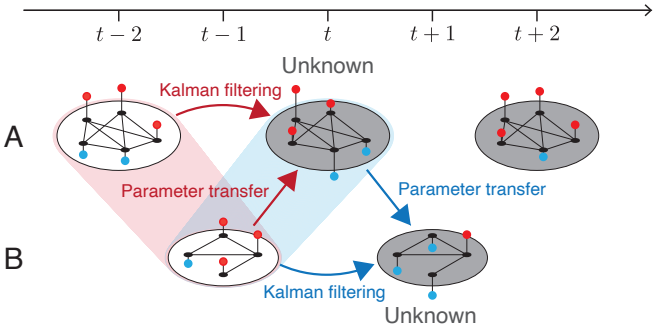


Fig. 1. Overview of a cooperative Kalman filter for the time-varying graph signals on two static graphs. Colored areas denote the set of the source and target at each time instance.

In this paper, we propose a cooperative Kalman filter for estimating time-varying signals between two networks, i.e., graphs. In the estimation, *graph signals*, defined as signals whose domain is the nodes in a graph [11], [12], are used to model the signals observed in a network. We illustrate the overview of the proposed method in Fig. 1. The proposed Kalman filter performs its estimation alternately in time between two graphs. Therefore, the two graphs work as the source and target alternately[1]. The set of parameters in the source is transferred to the target and then it is used in the Kalman estimator, which estimates the current target signal.

In the proposed method, first, we assume a *cyclic* graph wide sense stationarity (CGWSS) of time-varying graph signals for both of the source and target graphs. CGWSS is an extension of *static* GWSS [13], [14], [15] to the case where its power spectral density (PSD) changes periodically over time. Second, we formulate the state-space model in the source graph. The state equation is derived based on the optimal transport, which is a mathematical tool to determine an efficient mapping between two sets of random signals [16]. In our model, we utilize the optimal transport as the signal transition filter from the previous signal to the current signal within a graph. Finally, we transfer the state-space model of the source into the target graph by reflecting their statistics. In this transfer, we assume

This work is supported in part by JSPS KAKENHI under Grants 23K26110, 23K17461, and 24K21314, and JST AdCORP under Grant JPMJKB2307.

[1] While this paper focuses on estimation between two graphs for simplicity, it could be generalized to three or more graphs.

that the statistics of the two graphs essentially differ in their PSDs. To compensate for the gap of the PSDs, we utilize a transfer method based on Bayesian inference [17]. As a result, we obtain the state-space model of the target and derive the corresponding control laws.

Our experiments on synthetic data demonstrate that the proposed method effectively estimates time-varying graph signals on two graphs.

II. Preliminaries

In this section, we introduce preliminaries of random signals defined on a graph, i.e., graph signals. Firstly, we review the basics of graph signal processing (GSP). Second, we define the notion of a graph wide sense stationarity (GWSS).

A. Basics of Graph Signal Processing

A weighted undirected graph is denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which $\mathcal{V}$ and $\mathcal{E}$ are sets of nodes and edges, respectively. The number of nodes and edges are denoted by $N = |\mathcal{V}|$ and $E = |\mathcal{E}|$, respectively. We use a weighted adjacency matrix $\mathbf{W}$ for representing the connection between nodes, where its (m, n)-element $[\mathbf{W}]_{mn} \geq 0$ is the edge weight between the mth and nth nodes; $[\mathbf{W}]_{mn} = 0$ for unconnected nodes. The degree matrix $\mathbf{D}$ is a diagonal matrix whose element is defined as $[\mathbf{D}]_{mm} = \sum_n [\mathbf{W}]_{mn}$. Using $\mathbf{D}$ and $\mathbf{W}$, the graph Laplacian is given by $\mathbf{L} = \mathbf{D} - \mathbf{W}$. A graph signal $\mathbf{x} \in \mathbb{R}^N$ is defined as $\mathbf{x} : \mathcal{V} \to \mathbb{R}^N$ where $[\mathbf{x}]_n$ corresponds to the signal value at the nth node. Since $\mathbf{L}$ is a real symmetric matrix, it has orthogonal eigenvectors and can be diagonalized as $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U} = [\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{N-1}]$ is a matrix whose ith column is the eigenvector $\mathbf{u}_i$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{N-1})$ is their diagonal eigenvalue matrix. Without loss of generality, we can assume $0 = \lambda_0 \leq \lambda_1, \ldots, \lambda_{N-1} = \lambda_{\max}$ since $\mathbf{L}$ is a positive semidefinite matrix [11]. In GSP, λ_i is referred to as a *graph frequency*. Then, spectra of $\mathbf{x}$ in the graph frequency domain are defined as $\hat{\mathbf{x}} = \mathbf{U}^\top \mathbf{x}$: It is called *graph Fourier transform* [11].

B. Graph Wide Sense Stationarity

A graph signal $\mathbf{x}$ is a graph wide sense stationary (GWSS) process if the following two conditions are satisfied:

Definition 1 (GWSS [13])**.** *Let $\mathbf{x}$ be a random signal on graph $\mathcal{G}$. The signal $\mathbf{x}$ follows a GWSS process if and only if the following conditions are satisfied:*

$$\mathbb{E}[\mathbf{x}] = \mu = \mathrm{const}, \tag{1a}$$

$$\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] = \mathbf{\Sigma} = \mathbf{U}\mathrm{diag}(\mathbf{p})\mathbf{U}^\top, \tag{1b}$$

where $\mathbf{\Sigma}$ and $\mathbf{p}$ are referred to as the covariance matrix and power spectral density (PSD), respectively, and μ indicates the mean.

In this paper, we assume GWSS signals $\mathbf{x}$ are ergodic: The ensemble mean in Definition 1 is identical to the temporal mean. Later, we will consider a time-varying version of the GWSS process, whose PSD varies periodically over time.

III. Related Work

In this section, we review the preliminary studies on a transfer of Kalman filter. First, we introduce the standard (linear) Kalman filter, which is related to the main part of our framework. Then, we review optimal transport and a graph filter transfer method.

A. Kalman Filter

Kalman filter is an online algorithm for sequentially tracking and estimating dynamic signals from given observations [7], [8], [18]. In the following, we revisit the system of the standard Kalman filter and its control laws.

1) The System of Kalman Filter: Let the subscript t indicates a time instance. The system of Kalman filter is modeled by a state-space model. Generally, it is formulated as follows [18], [19]:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{u}_{t-1} + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(0, \sigma_v^2\mathbf{I}), \tag{2a}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}(0, \sigma_w^2\mathbf{I}), \tag{2b}$$

where $\mathbf{x}_t \in \mathbb{R}^N$ and $\mathbf{y}_t \in \mathbb{R}^M$ represent the current signal and its observation, respectively, and $\mathbf{u}_t \in \mathbb{R}^N$ indicates the control input. In (2a) and (2b), the matrices $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times N}$, and $\mathbf{C} \in \mathbb{R}^{M\times N}$ represent the signal transition, input, and observation matrices, respectively. The vector $\mathbf{v}_t \in \mathbb{R}^N$ represents system noise, and $\mathbf{w}_t \in \mathbb{R}^M$ denotes observation noise, conforming to white Gaussian noise with the standard deviations σ_v and σ_w, respectively. Note that (2a) describes the transition of the signals over time and (2b) represents the observation model at t.

In the Kalman filtering, the objective is to minimize the mean squared error (MSE) of the estimated signal for all t. The optimization problem is formalized as:

$$\min_{\tilde{\mathbf{x}}_t} \mathbb{E}[\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|_2^2], \tag{3}$$

where $\tilde{\mathbf{x}}_t$ represents the estimated signal at t.

In the following, we show the control laws of Kalman filter to estimate the current signal $\tilde{\mathbf{x}}_t$ in (3).

2) Control Laws of Kalman Filter: Hereafter, we denote the estimated signal conditioned on observations up to t, i.e., the prior estimation, by $\tilde{\boldsymbol{x}}_{t|t-1}$ and denote the posterior one by $\tilde{\boldsymbol{x}}_t$. The control laws are given as follows [18]:

Prediction step

i Calculating the prior signal estimation.

$$\tilde{\mathbf{x}}_{t|t-1} = \mathbf{A}\tilde{\mathbf{x}}_{t-1} + \mathbf{B}\mathbf{u}_{t-1}, \tag{4a}$$

ii Determining the error covariance matrix of the prior signal estimation.

$$\mathbf{P}_{t|t-1} = \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}^\top + \sigma_v\mathbf{I}, \tag{4b}$$

where $\mathbf{P}_{t-1}$ is the posterior error covariance at $t-1$.

iii Deriving the optimal filter (Kalman gain) in (3) from observations up to $t-1$.

$$\mathbf{K}_{t|t-1} = \mathbf{P}_{t|t-1}\mathbf{C}^\top(\mathbf{C}\mathbf{P}_{t|t-1}\mathbf{C}^\top + \sigma_w\mathbf{I})^{-1}, \tag{4c}$$

Update step

i Estimating the current signal using the Kalman gain.

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t|t-1} + \mathbf{K}_{t|t-1}(\mathbf{y}_t - \mathbf{C}\tilde{\mathbf{x}}_{t|t-1}), \quad (4d)$$

ii Updating the posterior error covariance matrix.

$$\mathbf{P}_t = (\mathbf{I} - \mathbf{K}_{t|t-1}\mathbf{C})\mathbf{P}_{t|t-1}. \quad (4e)$$

iii Returning to the prediction step with $t \leftarrow t+1$.

At the initial time instance, $t = 1$, $\mathbf{P}_{t-1}$ in (4b) is usually set to the scaled identity matrix, $\mathbf{P}_0 = \delta\mathbf{I}$, where the $\delta > 0$ is a scaling factor.

In our setting, $\mathbf{A}$ may not be known a priori and needs to be estimated based on statistics of $\mathbf{x}_t$. To this end, we introduce the optimal transport in the next subsection.

B. Optimal Transport

Optimal transport theory is a mathematical tool to determine the most efficient mapping between two sets of random signals having different probability distributions [20]. Let us consider the transport from an input signal $\mathbf{x}_1 \sim \alpha$ to a subsequent signal $\mathbf{x}_2 \sim \beta$, where α and β are two different probabilistic measures on $\mathbb{R}^N$. Then, the optimal transport seeks the assignment γ such that it minimizes some transport cost from $\mathbf{x}_1$ to $\mathbf{x}_2$. Given a transport cost function $c : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_+$, the optimal transport problem can be formulated as:

$$\inf_{\gamma \in \Pi(\alpha,\beta)} \mathop{\mathbb{E}}_{\mathbf{x}_1 \sim \alpha, T(\mathbf{x}_1) \sim \beta} c(\mathbf{x}_1, T(\mathbf{x}_1)) \quad \text{s.t. } T_{\#}\alpha = \beta, \quad (5)$$

where $\mathbf{x}_2 = T(\mathbf{x}_1)$ and Π represents a subset of joint distributions on $\mathbb{R}^N \times \mathbb{R}^N$. The mapping $T : \mathbb{R}^N \to \mathbb{R}^N$ is referred to as the optimal transport map, and $T_{\#}\alpha$ denotes the push-forward measure of α under T [16].

The cost function often employs the ℓ_p-norm. If the ℓ_p-norm is selected as the cost, the minimum value in (5) is referred to as the p-Wasserstein distance [16]. Obtaining the optimal transport map T analytically from (5) is generally challenging due to its non-uniqueness [16]. Nevertheless, if both of α and β are Gaussian distributions, i.e., $\alpha = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\beta = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, the unique solution is obtained by the 2-Wasserstein distance $W_2^2(\alpha, \beta)$[16]:

$$W_2^2(\alpha,\beta) = \\ \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\left(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{1/2}\right)^{1/2}\right). \quad (6)$$

Simultaneously, the optimal transport T is obtained [16] as

$$T(\mathbf{x}_1) = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_1^{-1/2}(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{1/2})^{1/2}\boldsymbol{\Sigma}_1^{-1/2}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \quad (7)$$

where we suppose that $\boldsymbol{\Sigma}_1$ is non-singular.

In our proposed method, we use T in (7) as the signal transition filter, which represents the optimal mapping from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$ in a graph, i.e., $\mathbf{A}$ in (2a). However, $\boldsymbol{\Sigma}_2$ in (7) is not known in our setting. Therefore, we need to find it from the known statistics in the other domain.

In the next subsection, we introduce a transfer learning method to find $\boldsymbol{\Sigma}_2$ from the statistics in the other graph.

C. Graph Filter Transfer

In this section, we review an existing study on graph filter transfer [17], which is mainly related to the proposed method. The graph filter transfer aims to transfer parameters of a graph filter for the source graph to the target one having different (but similar) statistics. Specifically, we introduce the method employed for probabilistic filters, such as the graph Wiener filter, Kalman filter, and Bayesian filter [17].

Hereafter, we denote the source and target graphs by src and trg, respectively. Let the subscript dom $\in \{\mathrm{src}, \mathrm{trg}\}$ represent one of the two graphs, and let $\mathbf{x}_{\mathrm{dom},t}$ and $\mathbf{y}_{\mathrm{dom},t}$ denote an *unknown* graph signal at time t and its corresponding *known* observation, defined in (2b). Now, we consider estimating $\mathbf{x}_{\mathrm{src},t}$ using a set of K historical observations, denoted by $\mathbf{Y}_{\mathrm{src},t} = [\mathbf{y}_{\mathrm{src},t}, \ldots, \mathbf{y}_{\mathrm{src},t-K+1}]$. We assume that the original graph signal $\mathbf{x}_{\mathrm{dom},t}$ satisfies the GWSS conditions in Definition 1. Under the assumption, the probabilistic filters are generally formulated through maximum a posteriori (MAP) estimation [14], [17], which is expressed as:

$$\tilde{\mathbf{x}}_{\mathrm{src},t} = \mathop{\arg\max}_{\mathbf{x}_{\mathrm{src},t}} \mathcal{P}(\mathbf{x}_{\mathrm{src},t} | \mathbf{Y}_{\mathrm{src},t}, \mathbf{p}_{\mathrm{src}}), \quad (8)$$

where $\mathcal{P}$ denotes a probability distribution function, and $\mathbf{p}_{\mathrm{src}}$ represents the PSD of $\mathbf{x}_{\mathrm{src},t}$. Since $\mathbf{p}_{\mathrm{src}}$ governs the behavior of the estimator in (8), transferring (8) to the target graph results in adjusting $\mathbf{p}_{\mathrm{src}}$ to adapt the target graph.

In general, the eigenvalue distributions of different graph variation operators are different. This discrepancy presents a challenge of estimating $\mathbf{p}_{\mathrm{trg}}$ directly from $\mathbf{p}_{\mathrm{src}}$. To address the issue, a graph filter transfer method in [17] is performed in the following three steps, as visualized in Fig. 2:

1) Approximating $\mathbf{p}_{\mathrm{src}}$ as a *continuous* ARMA graph filter [21] based on the least-squares (LS). The approximated PSD is denoted by $p_{\mathrm{ARMA}}(\lambda; \boldsymbol{\alpha}_{\mathrm{src}})$, where $\lambda \in [0, +\infty)$ and $\boldsymbol{\alpha}_{\mathrm{dom}}$ represents a sequence of parameters of the ARMA graph filter.
2) Adapting $\boldsymbol{\alpha}_{\mathrm{src}}$ in $p_{\mathrm{ARMA}}(\lambda; \boldsymbol{\alpha}_{\mathrm{src}})$ to the target graph using Bayesian inference. The adapted parameters in the target graph is represented as $\boldsymbol{\alpha}_{\mathrm{trg}}$.
3) Discretizing $p_{\mathrm{ARMA}}(\lambda; \boldsymbol{\alpha}_{\mathrm{trg}})$ according to the eigenvalues of $\mathbf{L}_{\mathrm{trg}}$ and estimating $\mathbf{p}_{\mathrm{trg}}$.

In the three steps, $\mathbf{p}_{\mathrm{trg}}$ is indirectly estimated from $\mathbf{p}_{\mathrm{src}}$ through the parameters of the ARMA filter. This approach can transfer a probabilistic filter from the source graph to the target graph having different eigenvalue distributions. Please see [17] for more details.

In the next section, we propose a cooperative Kalman filter for time-varying graph signals between two networks utilizing techniques introduced in this section.

IV. Cooperative Kalman Filter

In this section, we derive a graph filter transfer method using Kalman filter. Initially, we design a stochastic signal model for time-varying graph signals and then formulate a state-space model. We derive Kalman filter from the state-space model and perform the transfer learning across two different graphs.

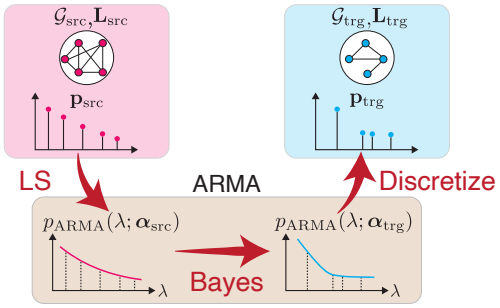


Fig. 2. Overview of a graph filter transfer method for the PSD estimation in the target graph in [17].

A. Signal Model

Here, we assume that $\mathbf{x}_{\text{src}}$ and $\mathbf{x}_{\text{trg}}$ satisfy the conditions defined below:

Definition 2 (Cyclic graph wide sense stationary (CGWSS))**.** *Let $\mathbf{x}$ be a random signal on the graph $\mathcal{G}$. The signal $\mathbf{x}$ is a cyclic graph wide sense stationary process if and only if the two following conditions are satisfied:*

$$\mathbb{E}\left[\mathbf{x}_{t\,(\mathrm{mod}\,P)}\right] = \mu_{t\,(\mathrm{mod}\,P)} = \text{const}, \tag{9a}$$

$$\mathbb{E}\left[\mathbf{x}_{t,(\mathrm{mod}\,P)}\mathbf{x}_{t\,(\mathrm{mod}\,P)}^{\top}\right] = \boldsymbol{\Sigma}_{t\,(\mathrm{mod}\,P)} = \mathbf{U}\mathrm{diag}(\mathbf{p}_{t\,(\mathrm{mod}\,P)})\mathbf{U}^{\top}, \tag{9b}$$

where P denotes the period of CGWSS, and $\boldsymbol{\Sigma}_t$ and $\mathbf{p}_t$ are the periodically varying covariance and PSD, respectively.

The PSD of CGWSS thus periodically changes over time, while that of the GWSS in Definition 1 is static[2].

In this paper, we assume two graphs are disconnected but the signals on them, $\mathbf{x}_{\text{src},t-1}$ and $\mathbf{x}_{\text{trg},t}$ have similar PSDs (see Fig. 1). Specifically, we define their similarity in terms of the graph filter kernels of the two PSDs, i.e., $p_{\text{dom},t}(\lambda_i)$, where $p_{\text{dom},t}(\lambda_i)$ can be obtained by smoothly interpolating $\mathbf{p}_{\text{dom},t}$ such that $p_{\text{dom},t}(\lambda_i) := [\mathbf{p}_{\text{dom},t}]_i$ (see Section III). Formally, we assume $\sup_\lambda (p_{\text{src},t-1}(\lambda) - p_{\text{trg},t}(\lambda))^2 < C$ where C is a small constant.

In the following, we formulate a state-space model based on Definition 2.

B. Control System

We consider the following state-space model similar to (2a) and (2b).

$$\mathbf{x}_{\text{dom},t} = T(\mathbf{x}_{\text{dom},t-2}) + \mathbf{B}\mathbf{u}_{\text{dom},t-2} + \mathbf{v}_{\text{dom}}, \quad \mathbf{v}_{\text{dom}} \sim \mathcal{N}(0, \sigma_v^2\mathbf{I}), \tag{10a}$$

$$\mathbf{y}_{\text{dom},t} = \mathbf{C}\mathbf{x}_{\text{dom},t} + \mathbf{w}_{\text{dom}}, \quad \mathbf{w}_{\text{dom}} \sim \mathcal{N}(0, \sigma_w^2\mathbf{I}). \tag{10b}$$

In (10a), we utilize the transition function T, which is not necessarily to be time-invariant instead of the time-invariant matrix $\mathbf{A}$ in (2a). In (10b), we employ the same model as (2b).

[2] When $P = 1$, CGWSS is identical to GWSS.

TABLE I
LIST OF SIGNALS. ✓ CORRESPONDS TO ACCESSIBLE (KNOWN) DATA. THE BLANK IS UNKNOWN DATA. THE COLORED CELL INDICATES THE SIGNAL WE ESTIMATE.

	$\mathbf{x}_{t-2}$	$\mathbf{x}_{t-1}$	$\mathbf{x}_t$	$\mathbf{y}_{t-2}$	$\mathbf{y}_{t-1}$	$\mathbf{y}_t$	$\mathbf{p}_{t-2}$	$\mathbf{p}_{t-1}$	$\mathbf{p}_t$
src		✓			✓			✓	
trg	✓			✓		✓	✓		

We summarize which variables are known or unknown in Table I. Our aim is to estimate the unknown signals $\mathbf{x}_{\text{trg},t}$ from known signals. Note that $\boldsymbol{\Sigma}_{\text{trg},t}$, which corresponds to $T(\cdot)$ in (10a) and $\boldsymbol{\Sigma}_2$ in (7), is unknown. Therefore, we first estimate $\mathbf{p}_{\text{trg},t}$ from $\mathbf{p}_{\text{src},t-1}$.

In the standard Kalman filter introduced in Section III-A, its estimator is derived in one cluster. In contrast, we perform the estimation alternately in time between two clusters (graphs). In the following, we derive the proposed Kalman filter.

C. Kalman Filter Transfer

We derive the cooperative Kalman filter between two graphs based on (10a) and (10b). For simplicity, we replace the subscripts $\cdot_{\text{trg},t-2}$ and $\cdot_{\text{trg},t}$ with $\cdot_1$ and $\cdot_2$, respectively. Following from Definition 2, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ can be jointly diagonalized. Therefore, the RHS in (7) in the target graph can be rewritten as follows:

$$T(\mathbf{x}_1) = \boldsymbol{\mu}_2 + \mathbf{Q}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \tag{11}$$

where $\mathbf{Q} = \mathbf{U}_{\text{trg}}\mathrm{diag}(\mathbf{p}_2)\mathrm{diag}(\mathbf{p}_1)^{-1}\mathbf{U}_{\text{trg}}^{\top}$.

To calculate the RHS in (11), the current PSD $\mathbf{p}_2$ ($= \mathbf{p}_{\text{trg},t}$) is required but unknown (see Table I). Therefore, we estimate $\mathbf{p}_2$ from $\mathbf{p}_{\text{src},t-1}$ by using the graph filter transfer method introduced in Section III-C. Consequently, the control laws of the proposed Kalman filter is described as follows:

Preprocessing step

i. Estimating $\mathbf{p}_2$ from $\mathbf{p}_{\text{src},t-1}$ by using the graph filter transfer method in Section III-C.
ii. Calculating the optimal transport map in (11) by using the estimated $\mathbf{p}_2$. We denote the transport map as $T_{\text{trg}|\text{src}}(\mathbf{x}_1)$.

Prediction step

i. Calculating the prior estimation of signals.

$$\tilde{\mathbf{x}}_{2|1} = T_{\text{trg}|\text{src}}(\tilde{\mathbf{x}}_1) + \mathbf{B}\mathbf{u}_1, \tag{12a}$$

ii. Determining the prior error covariance matrix.

$$\mathbf{P}_{2|1} = \mathbf{Q}\mathbf{P}_1\mathbf{Q}^{\top} + \sigma_v^2\mathbf{I}, \tag{12b}$$

iii. Deriving the Kalman gain.

$$\mathbf{K}_{2|1} = \mathbf{P}_{2|1}\mathbf{C}^{\top}(\mathbf{C}\mathbf{P}_{2|1}\mathbf{C}^{\top} + \sigma_w^2\mathbf{I})^{-1}, \tag{12c}$$

Update step

i. Estimating the current signals using the Kalman gain.

$$\tilde{\mathbf{x}}_2 = \tilde{\mathbf{x}}_{2|1} + \mathbf{K}_{2|1}(\mathbf{y}_2 - \mathbf{C}\tilde{\mathbf{x}}_{2|1}), \tag{12d}$$

ii. Updating the posterior error covariance matrix.

$$\mathbf{P}_2 = (\mathbf{I} - \mathbf{K}_{2|1}\mathbf{C})\mathbf{P}_{2|1}. \tag{12e}$$

iii. Swapping src with trg and returning to the preprocessing step with $t \leftarrow t+1$.

In the initial estimation for each graph, $\mathbf{P}_1$ in (12b) is set to $\mathbf{P}_1 = \delta\mathbf{I}$, similar to the setting in Section III-A.

V. Experiment

In this section, we perform signal estimation experiments for synthetic data.

1) Synthetic Graph Signals: We construct two different random sensor (RS) graphs[3] $\mathcal{G}_A$ and $\mathcal{G}_B$ with $N_A = 90$ and $N_B = 45$. The period of CGWSS signals is set to $P = 8$ and their PSDs are given by four different low-pass filters $\{\mathbf{p}_{\text{dom},p}\}_{p=0,\dots,P-1}$ as functions in $\mathbf{L}_{\text{dom}}$:

$$[\mathbf{p}_{\text{dom},p}]_i = \begin{cases} 1-\lambda_i/\lambda_{\max} & (p=0,4), \\ \exp(-\lambda_i/\lambda_{\max}) & (p=1,5), \\ 1/(1+\lambda_i) & (p=2,6), \\ \cos(\pi\lambda_i/2\lambda_{\max}) & (p=3,7). \end{cases} \tag{13}$$

Accordingly, we generate samples of the signals conforming to $\mathcal{N}(\mathbf{1}, \mathbf{U}_{\text{dom}}\text{diag}(\mathbf{p}_{\text{dom},p})\mathbf{U}_{\text{dom}}^\top)$, where $p = t(\text{mod}\, P)$ for $t = 1,\dots,T$. We denote the training and test datasets by $\mathbf{X}_{\text{dom}}^{\text{train}} \in \mathbb{R}^{N_{\text{dom}}\times T_{\text{train}}}$ and $\mathbf{X}_{\text{dom}}^{\text{test}} \in \mathbb{R}^{N_{\text{dom}}\times T_{\text{test}}}$, consisting of $T_{\text{train}} = 200$ and $T_{\text{test}} = 40$ samples, respectively.

Then, we consider a data update process for the sequential signal estimation experiment. Let $\mathbf{X}_{\text{dom},p}^{(l)} \in \mathbb{R}^{N_{\text{dom}}\times K}$ be a data slot at p, where $K = T_{\text{train}}/P$ and l indicates the lth cycle, i.e., it satisfies $t = lP + p$. We update the data slot at every time instance in a warm-start manner:

$$\mathbf{X}_{\text{src},p}^{(l+1)} = \begin{cases} \mathbf{X}_{\text{src},p}^{\text{train}} & \text{if } l = 0, \\ \left[\tilde{\mathbf{x}}_{\text{trg},t}, \left[\mathbf{X}_{\text{src},p}^{(l)}\right]_{:,1:K-1}\right] & \text{otherwise}, \end{cases} \tag{14}$$

where $\left[\mathbf{X}_{\text{src},p}^{(l)}\right]_{:,1:K-1}$ denotes the submatrix of $\mathbf{X}_{\text{src},p}^{(l)}$ whose columns from 1 to $K-1$. In (14), We divide $\mathbf{X}_{\text{dom}}^{\text{train}}$ into P periods and set it as the initial value, $\mathbf{X}_{\text{dom},p}^{\text{train}}$.

2) Experimental Setup: The observation matrix $\mathbf{C}$ in (10b) is set to a random sampling matrix with $M_A = 85$ and $M_B = 43$, respectively. For the initial estimation, we estimate $\mathbf{p}_{\text{src},0}$ and $\mathbf{p}_{\text{trg},1}$ $(= \mathbf{p}_1)$ in (11), from the training data by using a PSD estimation method [14], [22]. We set the scaling factor of the initial error covariance $\delta = 1$, and the initial estimation of signal $\tilde{\mathbf{x}}_1 = \left[\mathbf{X}_{\text{trg},p}^{\text{train}}\right]_{:,1}$ in (12a). We use a proportional control strategy for the control input [23]: $\mathbf{B} = \mathbf{I}$ and $\mathbf{u}_1 = \eta(\tilde{\mathbf{x}}_1 - \bar{\mathbf{x}}_{\text{trg},p})$, where η is a factor of proportionality and $\bar{\mathbf{x}}_{\text{trg},p}$ is a mean vector in the latest data slot. We empirically set $\eta = 5.0\times10^{-2}$.

Since there is no prior work on the cooperative Kalman filter to the best of our knowledge, we use the following well-known methods as baseline methods.

[3] Random sensor graphs are implemented by k nearest neighbor graphs whose nodes are randomly distributed in 2-D space $[0,1]\times[0,1]$ (See [22]).

TABLE II
Experimental results on synthetic dataset

σ_w	Average MSE(10^{-2})		
	Proposed	RRTK	TGWF
0.05	**0.33**	0.92	2.38
0.10	**1.00**	1.31	2.94
0.15	**2.26**	2.34	4.29

Ridge regression with Tikhonov regularization: The first baseline method is signal estimation based on Tikhonov regularization (RRTK) [12]. By using the space model in (10b), the estimated signal can be written as:

$$\begin{aligned}\tilde{\mathbf{x}}_{\text{trg},t}^{\text{RRTK}} &= \arg\min_{\mathbf{x}_{\text{trg,t}}} \|\mathbf{y}_{\text{trg},t} - \mathbf{C}\mathbf{x}_{\text{trg},t}\|_2^2 + \zeta\mathbf{x}_{\text{trg},t}^\top\mathbf{L}_{\text{trg}}\mathbf{x}_{\text{trg},t} \\ &= (\mathbf{C}^\top\mathbf{C} + \zeta\mathbf{L}_{\text{trg}})^{-1}\mathbf{C}^\top\mathbf{y}_{\text{trg},t},\end{aligned} \tag{15}$$

where ζ is a parameter, which controls the intensity of the regularization term and is set to $\zeta = 0.05$.

Transferred graph Wiener filter: The Second one is signal estimation based on a transferred graph Wiener filter (TGWF) proposed in [17]. The estimated signal can be written as:

$$\tilde{\mathbf{x}}_{\text{trg},t}^{\text{TGWF}} = \mathbf{H}\mathbf{y} + \mathbf{b}, \tag{16}$$

where $\mathbf{H} = \mathbf{\Sigma}\mathbf{C}^\top(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^\top)^{-1}$, $\mathbf{b} = (\mathbf{I} - \mathbf{H}\mathbf{C})\boldsymbol{\mu}$. The covariance matrix $\mathbf{\Sigma}$ is calculated from the estimated PSD $\mathbf{p}_{\text{trg},t}$ (see Definition 2 and Section III-C).

Both of the baseline methods only use the space equation in (10b). In contrast, the proposed method utilizes Kalman filter derived from the state-space model.

The regularization term in (15) imposes the smoothness of signals on the target graph and therefore RRTK does not rely on GWSS assumptions. In contrast, TGWF assumes GWSS, which requires the PSD for every time instance, as stated in Sec. III-C. Therefore, in this experiment, TGWF estimates the PSD using the method in [14] at each time instance and then performs signal estimation for the target graph.

To perform the two baseline methods in our time-varying setting, we repeat the baseline methods with alternately switching the source with the target at every time instance. We also set $\sigma_v = 0$ in (10a) to conduct the experiments for a fair comparison. We consider additive white Gaussian noise on observations with three different standard deviations σ_w.

3) Results: We evaluate the estimation performance of the proposed method with the MSE and compare it with the above alternative methods. Table II summarizes the average MSE in test datasets. We also visualize an example of absolute errors between the original and estimated signals in Fig. 3. Fig. 4 plots MSEs over time.

In Table II, the proposed method outperforms alternative methods for all σ_w. In Fig. 4, we observe that the proposed method shows consistent estimation performance for all t, while those of the other methods oscillate significantly over time. This is because the proposed method can compensate

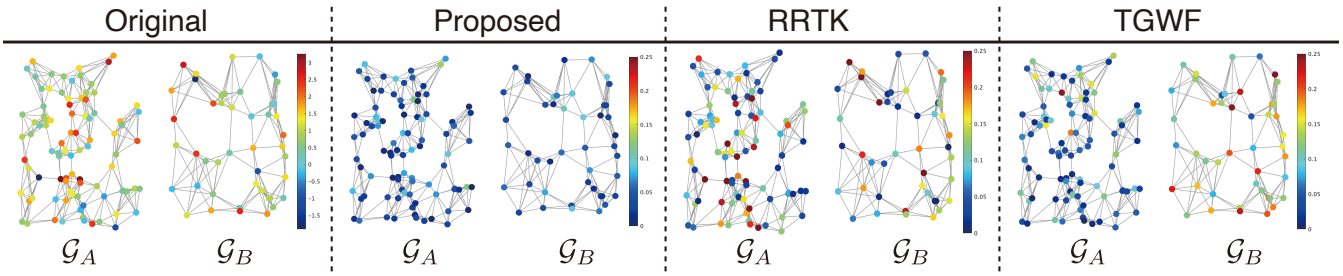


Fig. 3. Signal estimation experiments for noisy graph signals on two different RS graphs with $N_A = 90$, $N_B = 45$ nodes at time $t = 8$, $t = 7$ within T_{test}, respectively. The color of a node represents the absolute error between original and estimated signals.

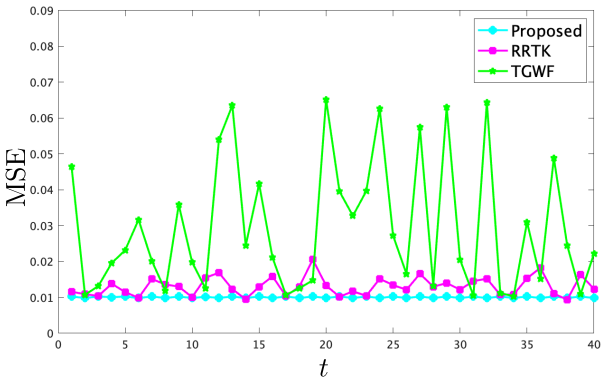


Fig. 4. Comparison of signal estimation MSEs for the RS graph ($\sigma_w = 0.10$). The MSEs for each of the methods is plotted as a horizontal line.

its own estimation using signals on the previous time instance and that on the different graph with similar statistics, while other methods perform the estimation independently at all time instances.

VI. Conclusion

In this paper, we propose a Kalman filter transfer method for estimating time-varying signals between two different graphs but having similar features. Initially, we assume that the time-varying graph signals conform to CGWSS for both graphs in our Kalman filter. We then formulate the control system based on the state-space mode and optimal transport. To alternately perform signal estimation over time in the source and target, we transfer a set of parameters in the Kalman filter from the source to the target. Our experiments demonstrate that the proposed method effectively estimates the time-varying graph signals.

References

[1] C.-Y. Chong and S. P. Kumar, "Sensor networks: evolution, opportunities, and challenges," *Proc. IEEE*, vol. 91, no. 8, pp. 1247–1256, 2003.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Comput. netw.*, vol. 38, no. 4, pp. 393–422, 2002.

[3] W.-H. Lee, S.-S. Tseng, J.-L. Shieh, and H.-H. Chen, "Discovering traffic bottlenecks in an urban network by spatiotemporal data mining on location-based services," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1047–1056, 2011.

[4] A. Martini, M. Troncossi, and A. Rivola, "Automatic leak detection in buried plastic pipes of water supply networks by means of vibration measurements," *Shock Vib.*, vol. 2015, no. 1, p. 165304, 2015.

[5] B. Kouvaritakis and M. Cannon, *Model Predictive Control: Classical, Robust and Stochastic*. Springer, 2016.

[6] A. Mesbah, "Stochastic model predictive control: An overview and perspectives for future research," *IEEE Control Syst. Mag.*, vol. 36, no. 6, pp. 30–44, 2016.

[7] G. Sagi, N. Shlezinger, and T. Routtenberg, "Extended kalman filter for graph signals in nonlinear dynamic systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 1–5, IEEE, 2023.

[8] M. Ramezani-Mayiami and B. Beferull-Lozano, "Joint graph learning and signal recovery via kalman filter for multivariate auto-regressive processes," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pp. 907–911, IEEE, 2018.

[9] C. Zhu, V. C. Leung, L. Shu, and E. C.-H. Ngai, "Green internet of things for smart world," *IEEE Access*, vol. 3, pp. 2151–2162, 2015.

[10] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Trans. Ind. Electron.*, vol. 66, no. 9, pp. 7316–7325, 2018.

[11] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, 2013.

[12] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[13] J. Hara, Y. Tanaka, and Y. C. Eldar, "Graph signal sampling under stochastic priors," *IEEE Trans. Signal Process.*, vol. 71, pp. 1421–1434, 2023.

[14] N. Perraudin and P. Vandergheynst, "Stationary signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3462–3477, 2017.

[15] A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Stationary graph processes and spectral estimation," *IEEE Trans. Signal Process.*, vol. 65, no. 22, pp. 5911–5926, 2017.

[16] J. Delon and A. Desolneux, "A wasserstein-type distance in the space of gaussian mixture models," *SIAM J. Imaging Sci.*, vol. 13, no. 2, pp. 936–970, 2020.

[17] K. Yamada, "Graph filter transfer via probability density ratio weighting," *arXiv:2210.14633*, 2022.

[18] G. Welch, G. Bishop, *et al.*, "An introduction to the kalman filter," *Proc. ACM SIGGRAPH*, 1995.

[19] S. Li, K. Y. Lim, and D. G. Fisher, "A state space formulation for model predictive control," *AIChE J.*, vol. 35, no. 2, pp. 241–249, 1989.

[20] C. Villani *et al.*, *Optimal transport: old and new*, vol. 338. Springer, 2009.

[21] E. Isufi, A. Loukas, A. Simonetto, and G. Leus, "Autoregressive moving average graph filtering," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 274–288, 2017.

[22] N. Perraudin, J. Paratte, D. Shuman, L. Martin, V. Kalofolias, P. Vandergheynst, and D. K. Hammond, "GSPBOX: A toolbox for signal processing on graphs," *arXiv:1408.5781*, 2014.

[23] K. Ogata, *Modern Control Engineering*. Prentice Hall PTR, 4th ed., 2001.

Multichannel Speech Enhancement Using Complex-Valued Graph Convolutional Networks and Triple-Path Attentive Recurrent Networks

Xingyu Shen and Wei-Ping Zhu
Department of Electrical and Computer Engineering, Concordia University, Canada
E-mail: shenxingyu97@outlook.com; weiping@ece.concordia.ca

***Abstract*—Multichannel speech enhancement has gained significant attention for its capability of improving speech quality and intelligibility in noisy environments. This paper presents a novel approach to multichannel speech enhancement utilizing complex-valued graph-in-graph convolutional networks (GiGCN) and triple-path attentive recurrent networks (TPARN). The proposed model leverages complex-valued operations to capture spatial dependencies and decoupled LSTM blocks to model temporal correlations. Meanwhile, the TPARN can effectively fuse the frequency, time, and spatial features for the reconstruction of the enhanced speech. Our experimental results based on the CHiME-3 and L3DAS22 datasets show that the proposed integrated model outperforms the state-of-the-art methods in terms of the PESQ, STOI and WER performance metrics.**

I. Introduction

Multichannel speech enhancement plays a critical role in improving speech quality and intelligibility across various real-world applications, including teleconferencing, hearing aids, and voice-controlled systems. In noisy environments, effective speech enhancement can maintain clear communication and interaction. Our speech signals captured by multiple microphones, where spatial, temporal, and spectral dependencies need to be accurately modeled to suppress noise and recover clean speech.

Traditional approaches to multichannel speech enhancement have often relied on beamforming techniques, which utilize spatial filtering to enhance speech signals captured by multiple microphones. Beamforming methods, such as minimum variance distortionless response (MVDR) [1][2] and generalized sidelobe canceler (GSC) [3], have demonstrated their effectiveness in many scenarios. However, these methods are limited by their reliance on accurate spatial information and may not fully exploit the temporal dynamics of speech signals.

With the advent of deep learning, there has been a significant shift towards using neural networks for speech enhancement. Convolutional neural networks (CNNs) and long short-term Memory (LSTM) networks have been widely adopted to model spatial and temporal dependencies in multichannel signals. For instance, recent works have employed CNNs to capture local spatial correlations [4] and LSTMs to model temporal dependencies in sequential data [5][6]. However, these methods often treat spatial and temporal features independently, potentially missing complex dependencies across these domains.

Graph convolutional networks (GCNs) have emerged as a powerful tool for modeling spatial relationships in various applications, including multichannel speech enhancement. GCNs effectively capture spatial dependencies across multiple channels by representing the microphones and their spatial relationships as a graph [7]. This approach allows for a more comprehensive understanding of spatial information, which can lead to more accurate enhancement of speech signals.

The use of complex-valued neural networks has shown great promise in directly processing the complex spectra obtained from short-time Fourier transform (STFT). Complex-valued operations can handle the phase information present in the STFT domain, which is crucial for accurate reconstruction of enhanced speech signal [8]. Moreover, attention mechanisms have been introduced to dynamically focus on important features across the time and frequency domains. These mechanisms can significantly improve the performance of speech enhancement models by selectively weighting those features that contribute most to the enhancement task [9].

Several state-of-the-art methods have advanced multichannel speech enhancement by combining various neural network architectures. For example, the use of spatial autoencoders to capture spatial dependencies while preserving the temporal structure has shown significant improvements in speech quality [10]. Additionally, decoupled spatial and temporal processing frameworks have demonstrated the potential to reduce computational complexity while maintaining high performance [11]. Dense frequency-time attentive networks have further enhanced the capability of speech enhancement models by integrating detailed frequency and temporal features [12].

Despite these advancements, there remains a need for a unified approach that can effectively leverage spatial, temporal, and frequency information. In this paper, we propose a novel method for multichannel speech enhancement that integrates complex-valued graph in graph convolutional networks (GiGCN) with a triple-path attentive recurrent network (TPARN). Our approach leverages the strengths of complex-valued operations to capture spatial dependencies directly from the STFT domain, while exploiting decoupled LSTM blocks to model both short-term and long-term temporal correlations. Moreover, a triple-path architecture is employed to extract the frequency, time, and spatial features effectively, allowing for a

comprehensive feature fusion.

The proposed model is extensively evaluated on two benchmark datasets, i.e., CHiME-3 [13] and L3DAS22 [14]. Our experimental results show that the proposed approach significantly improves the quality and intelligibility of the enhanced speech, as compared to the state-of-the-art methods in various noisy environments. These findings corroborate the effectiveness and robustness of our integrated model in handling complex acoustic scenarios for real-world applications.

II. Model Description

A. Method Overview

Our proposed multichannel speech enhancement model integrates the complex-valued graph in graph convolutional networks (GiGCN) with a triple-path attentive recurrent network (TPARN) to effectively capture the spatial, temporal, and frequency information. The architecture of the proposed integrated model is depicted in the upper half of Figure 1. The input noisy speech signals from multiple channels are first transformed into the frequency domain using the short-time Fourier transform (STFT). These transformed signals are then processed through a series of complex-valued GiGCN layers to capture spatial dependencies, resulting in spatial features denoted as $\mathbf{H}_{space}$. The enhanced features from the GiGCN block are subsequently passed through decoupled LSTM blocks, which separately model short-term and long-term temporal dependencies, producing the temporal features $\mathbf{H}_{time}$. Finally, the TPARN integrates these spatial and temporal features, along with frequency features $\mathbf{H}_{freq}$ derived from the input STFT, to form a cohesive representation. These integrated features are then fused and used to reconstruct the into enhanced speech signals.

B. Complex-Valued GiGCN Block

The complex-valued GiGCN block leverages the properties of complex-valued neural networks for processing the STFT coefficients, which are naturally represented in the complex domain. This allows for the direct manipulation of both amplitude and phase information, crucial for accurately reconstructing enhanced speech signals. Given a noisy speech signal captured by N microphones, the input signal $\mathbf{x}(t)$ can be represented as:

$$\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_N(t)] \tag{1}$$

where $x_i(t)$ is the signal from the i-th microphone. Applying the STFT to each channel, we obtain the complex-valued spectrogram:

$$\mathbf{X}(k,l) = [X_1(k,l), X_2(k,l), \ldots, X_N(k,l)] \tag{2}$$

where $X_i(k,l)$ represents the STFT of the i-th microphone signal at frequency bin k and time frame l. The complex-valued convolution operation is defined as:

$$(\mathbf{W} \star \mathbf{X})(k,l) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} \mathbf{W}(m,n)\mathbf{X}(k-m, l-n) \tag{3}$$

where M and N denote the size of the convolutional window, controlling the local feature aggregation. $\mathbf{W}(m,n)$ represents the complex-valued filter weight and $\mathbf{X}(k,l)$ is the input complex spectrogram.

To capture spatial dependencies among microphones in a multichannel setup, we model the system as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes $\mathcal{V}$ represent microphones and edges $\mathcal{E}$ represent the connections between them [15]. These connections are encoded in an adjacency matrix $\mathbf{A}$, whose element a_{ij} indicates the connection strength or similarity between microphones i and j.

In this setup, the graph convolution operation updates the feature vector $\mathbf{H}_i$ of each node by aggregating features from its neighboring nodes $j \in \mathcal{N}(i)$. This update at layer l is expressed as:

$$\mathbf{H}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \mathrm{a}_{ij} \mathbf{W}_{\text{GCN}}^{(l)} \mathbf{H}_j^{(l)}\right) \tag{4}$$

Here, $\mathbf{H}_i^{(l+1)}$ represents the updated feature vector for node i, $\mathbf{W}_{\text{GCN}}^{(l)}$ is the layer-specific learnable weight matrix, and σ is a non-linear activation function like ReLU.

GiGCNs extend this framework by introducing a nested graph structure, where each primary node (microphone) is further detailed as a secondary graph [16]. This nested approach is particularly advantageous for complex-valued data such as STFT coefficients, where both real and imaginary components are processed simultaneously. In contrast to GCNs, which primarily operate on real-valued data, GiGCNs can model more intricate relationships between nodes in the complex domain, capturing both magnitude and phase information, as given by:

$$\mathbf{H}_{i,j}^{(l+1)} = \sigma\left(\sum_{k \in \mathcal{N}(i)} \mathrm{a}_{ik}^{\text{nested}} \mathbf{W}_{\text{GiGCN}}^{(l)} \mathbf{H}_{k,j}^{(l)}\right) \tag{5}$$

where $\mathbf{H}_{i,j}^{(l)}$ represents the feature of the j-th component (either real or imaginary) of the i-th node at layer l. The nested adjacency matrix $\mathbf{A}^{\text{nested}}$ and weight matrix $\mathbf{W}_{\text{GiGCN}}^{(l)}$ facilitate the aggregation of features across the graph.

This dual-level graph structure allows the model to capture detailed spatial dependencies within and across nodes, enhancing its ability to process and improve multichannel audio signals. The nested structure ensures that both intra-node and inter-node interactions are considered, providing a more comprehensive understanding of the spatial relationships in the data.

The complex STFT features are first decomposed into their real and imaginary components. Each component undergoes feature extraction using multi-scale GiGCN layers, which

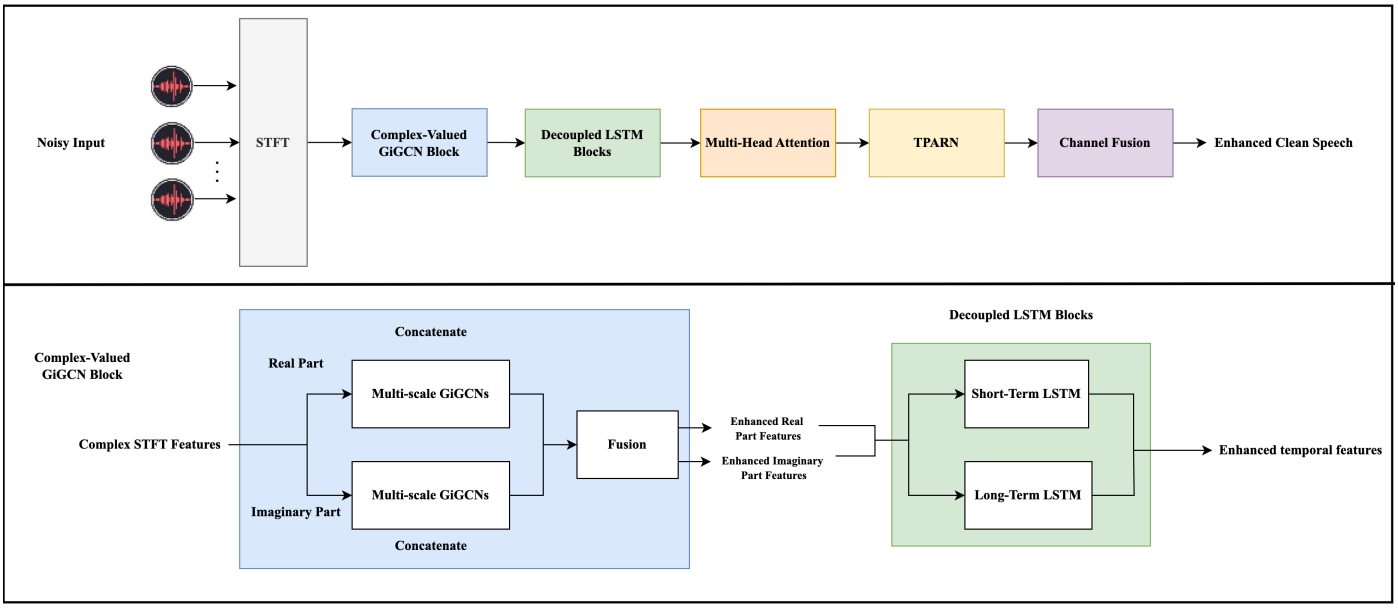


Fig. 1. Proposed multichannel speech enhancement framework

employ filters of varying sizes (3x3, 5x5, and 7x7). This multi-scale approach allows the capture of spatial features at different resolutions.

For both the real and imaginary parts, the features derived from the different filter sizes are concatenated, resulting in enhanced real and imaginary feature sets. These sets are then fused to produce a unified complex-valued feature representation, integrating information from both components. The detailed operation of the complex-valued GiGCN block is illustrated in the lower half of Figure 1.

This fused complex feature set effectively captures the intricate spatial details present in the STFT domain and serves as the input to the Decoupled LSTM blocks. These blocks are designed to further refine the features by modeling short-term and long-term temporal dependencies, thereby enhancing the model's ability to improve speech quality in noisy environments.

C. Decoupled LSTM Blocks

The decoupled LSTM blocks are specifically designed to separately model the short-term and long-term temporal correlations in the enhanced features obtained from the complex-valued GiGCN block. By decoupling the temporal modeling into distinct short-term and long-term LSTM networks, we can more precisely capture temporal dynamics across different time scales, which is crucial for enhancing speech signals. The input features $\mathbf{H}_{enhanced}$ from the complex-valued GiGCN block are processed through two separate networks: short-term LSTM (ST-LSTM) and long-term LSTM (LT-LSTM). The ST-LSTM focuses on capturing rapid temporal variations, which are essential for preserving fine-grained temporal details in the speech signal. Conversely, the LT-LSTM is designed to capture slower, more sustained temporal dependencies, which are important for maintaining the overall structure and coherence of the speech signal over longer periods.

The outputs from the ST-LSTM and LT-LSTM networks are concatenated to form the final temporal features $\mathbf{H}_{temporal}$. This comprehensive temporal representation ensures that the enhanced speech signal retains both detailed and structural temporal characteristics.

D. Triple-Path Attentive Recurrent Network (TPARN)

The TPARN [17] further combines the frequency, time, and spatial features through a unified triple-path architecture. The temporal features $\mathbf{H}_{temporal}$, obtained from the decoupled LSTM blocks, are combined with the spatial features $\mathbf{H}_{space}$ extracted from the GiGCN block and frequency features $\mathbf{H}_{freq}$ from the STFT domain. These three types of features are processed in parallel paths within the TPARN: the frequency path captures frequency-domain correlations, the time path handles temporal correlations, and the spatial path processes spatial correlations across different channels. These paths employ convolutional, recurrent, and graph-based layers, respectively. The outputs from these paths are integrated using a multi-head attention mechanism (MHA), which combines the distinct features into a unified representation. This combined output, denoted as $\mathbf{H}_{TPARN}$, is formulated as:

$$\mathbf{H}_{TPARN} = \text{MHA}(\mathbf{H}_{freq}, \mathbf{H}_{time}, \mathbf{H}_{space}) \tag{6}$$

E. Channel Fusion

The channel fusion module combines the enhanced features from multiple channels into a single, unified representation. This process is essential for consolidating the spatial information captured from different locations. The input to the channel fusion module, $\mathbf{H}_{TPARN}$, is processed through a fully

connected layer:

$$\mathbf{H}_{fused} = \text{ReLU}(\mathbf{W}_{fusion}\mathbf{H}_{TPARN} + \mathbf{b}_{fusion}) \quad (7)$$

where $\mathbf{W}_{fusion}$ and $\mathbf{b}_{fusion}$ are the weights and biases of the fully connected layer, respectively. The final output is transformed back to the time domain using ISTFT to obtain the enhanced clean speech.

F. Loss Function

The model is trained using a combination of time-domain and frequency-domain loss functions. The time-domain loss is calculated as the mean squared error (MSE) between the enhanced time-domain signal $\hat{\mathbf{s}}(t)$ and the ground truth clean signal $\mathbf{s}(t)$:

$$\mathcal{L}_{time} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\mathbf{s}}(t) - \mathbf{s}(t))^2 \quad (8)$$

The frequency-domain loss is calculated as the MSE between the STFTs of the enhanced and ground truth signals:

$$\mathcal{L}_{freq} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}|\hat{\mathbf{S}}(k,l) - \mathbf{S}(k,l)|^2 \quad (9)$$

The combined loss function is a weighted sum of the time-domain and frequency-domain losses:

$$\mathcal{L} = \alpha\mathcal{L}_{time} + \beta\mathcal{L}_{freq} \quad (10)$$

where α and β are the weighting coefficients that balance the contributions of the two losses. In our experiments, we set $\alpha = 0.5$ and $\beta = 0.5$ to equally weigh the contributions of both losses.

The above-proposed comprehensive approach leverages the strengths of each component to enhance speech signals effectively and its detailed architecture is shown in the lower portion of Figure 1.

III. Experiments

A. Datasets

To evaluate the performance of our proposed multichannel speech enhancement model, called integrated model below for comparison, we utilized two widely recognized datasets: L3DAS22 and CHiME-3. The L3DAS22 dataset, designed for 3D speech enhancement tasks, includes over 40,000 virtual 3D audio environments recorded using a 1st order Ambisonics microphone, providing four-channel recordings [14]. Clean utterances are sourced from the Librispeech [18] corpus, and noise signals are from the FSD50K dataset [19]. The CHiME-3 dataset comprises noisy recordings from six channels in real-world environments, including buses, cafes, pedestrian areas, and street junctions [13], making it suitable for robust automatic speech recognition and multichannel speech enhancement tasks.

B. Experimental Setup

The noisy speech signals are transformed into the frequency domain using the STFT, with a window size of 1024 samples, a hop size of 512 samples, and an FFT size of 1024 points. The resulting complex spectrograms are processed by our integrated model, comprising complex-valued GiGCN, decoupled LSTM blocks, TPARN, and channel fusion modules. The model is trained using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 16 for 100 epochs. Data augmentation involves adding various types of noise at different SNRs, ranging from 0 dB to 20 dB, ensuring robustness across a wide range of noisy conditions. The model is evaluated on separate test sets of unseen noisy speech recordings to ensure a fair performance assessment.

C. Experimental Results and Analysis

The evaluation metrics used in this study include perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI) [26], and Word Error Rate (WER). We have also used an overall metric to evaluate our model based on dataset L3DAS22, which is a combination of STOI and WER [14]:

$$\text{Metric} = \frac{\text{STOI} + (1 - \text{WER})}{2} \quad (11)$$

The WER is computed based on the transcription of the estimated target signal and that of the reference signal, both decoded by a pre-trained Wav2Vec2.0-based ASR model [27].

The experimental results on the CHIME-3 dataset demonstrate that our integrated model yields a superior performance compared to other models, with higher average PESQ and STOI scores across all scenarios (BUS, CAF, PED, STR), as shown in Table I. Our model achieves an average PESQ of 1.905 and STOI of 0.922, outperforming other state-of-the-art methods. The ablation studies show that each component of our model contributes to the overall performance, with the integrated model achieving the highest scores.

Based on the L3DAS22 dataset (Table II), our proposed model significantly outperforms other state-of-the-art models. Our integrated model achieves the highest STOI score of 0.941 and the lowest WER of 0.065, resulting in a superior overall metric of 0.938. The ablation studies show that each component of our model contributes to the overall performance. Specifically, using a single 3x3 or 5x5 GiGCN block slightly decreases performance, indicating the importance of the multi-scale approach. The removal of either short-term or long-term LSTM also leads to lower performance, highlighting the necessity of capturing both short-term and long-term dependencies. The simplified channel fusion method results in lower STOI and higher WER, demonstrating the effectiveness of our sophisticated fusion technique.

Overall, our proposed model demonstrates significant improvements in PESQ, STOI, and WER metrics over state-of-the-art methods on both datasets. The comprehensive ablation studies further validate the effectiveness of our design choices,

TABLE I
ABLATION EXPERIMENTS AND PERFORMANCE COMPARISON WITH DIFFERENT MODELS ON THE CHIME-3 DATASET.

Model	**PESQ**					**STOI**				
	BUS	CAF	PED	STR	Avg.	BUS	CAF	PED	STR	Avg.
Our Integrated Model	1.942	**1.863**	**1.899**	1.925	**1.905**	**0.938**	**0.910**	**0.927**	**0.935**	**0.922**
Single 3x3 GiGCN Block	1.781	1.757	1.787	1.803	1.782	0.919	0.902	0.911	0.931	0.916
Single 5x5 GiGCN Block	1.797	1.768	1.792	1.808	1.792	0.921	0.905	0.913	0.933	0.918
Only Short-Term LSTM	1.784	1.760	1.788	1.805	1.784	0.923	0.907	0.914	0.934	0.920
Only Long-Term LSTM	1.777	1.754	1.782	1.798	1.778	0.917	0.900	0.909	0.929	0.913
Only Frequency and Time Paths in TPARN	1.773	1.748	1.775	1.793	1.772	0.916	0.899	0.908	0.928	0.912
Simplified Channel Fusion (Averaging)	1.786	1.763	1.791	1.807	1.787	0.918	0.903	0.912	0.932	0.916
Noisy (channel 1)	1.272	1.153	1.176	1.263	1.216	0.903	0.808	0.807	0.877	0.849
CNN [20]	1.775	1.540	1.618	1.751	1.671	0.920	0.885	0.882	0.913	0.900
LSTM-IPD [21]	1.908	1.496	1.512	1.730	1.661	0.921	0.847	0.835	0.890	0.873
UNet-GCN [7]	1.655	1.448	1.495	1.645	1.560	0.918	0.875	0.871	0.909	0.893
STGCSEN [15]	**2.002**	1.743	1.747	**1.941**	1.858	0.923	0.891	0.876	0.914	0.901

TABLE II
ABLATION EXPERIMENTS AND PERFORMANCE COMPARISON WITH DIFFERENT MODELS ON THE L3DAS22 DATASET.

Model	**WER**	**STOI**	**Metric**
Our Integrated Model	**0.065**	**0.941**	**0.938**
Single 3x3 GiGCN Block	0.069	0.936	0.934
Single 5x5 GiGCN Block	0.070	0.935	0.932
Only Short-Term LSTM	0.068	0.937	0.935
Only long-Term LSTM	0.071	0.934	0.931
Only Frequency and Time Paths in TPARN	0.072	0.932	0.930
Simplified Channel Fusion (Averaging)	0.070	0.933	0.932
Spatial-DCCRN [22]	0.071	0.931	0.931
SEU Speech [23]	0.101	0.902	0.900
JLESS [24]	0.174	0.836	0.831
CCA SPEECH [25]	0.240	0.831	0.796

highlighting the importance of each component in achieving superior speech enhancement performance.

IV. CONCLUSIONS

In this work, we have presented a novel multichannel speech enhancement model integrating complex-valued GiGCNs, decoupled LSTM blocks, TPARN, and channel fusion. Our simulations based on L3DAS22 and CHiME-3 datasets demonstrated significant improvements over state-of-the-art methods in terms of PESQ and STOI metrics. Our ablation studies validated the contributions of each component, underscoring the importance of multi-scale approaches, temporal modeling, and feature integration. It is also shown that the proposed approach has a robust performance across various noisy environments, indicating a strong potential for practical applications in real-world scenarios.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[2] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[3] M. R. Bai and F.-J. Kung, "Speech enhancement by denoising and dereverberation using a generalized sidelobe canceller-based multichannel wiener filter," *Journal of the Audio Engineering Society*, vol. 70, no. 3, pp. 140–155, 2022.

[4] S. Chakrabarty and E. A. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[5] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2019, pp. 298–302.

[6] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 271–275.

[7] P. Tzirakis, A. Kumar, and J. Donley, "Multi-channel speech enhancement using graph neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 3415–3419.

[8] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[9] H. Chen, X. Peng, Q. Jiang, and Y. Guo, "Residual unet with attention mechanism for time-frequency domain speech enhancement," in *2022 41st Chinese Control Conference (CCC)*, IEEE, 2022, pp. 7007–7011.

[10] M. M. Halimeh and W. Kellermann, "Complex-valued spatial autoencoders for multichannel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 261–265.

[11] A. Pandey and B. Xu, "Decoupled spatial and temporal processing for resource efficient multichannel speech enhancement," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 206–12 210.

[12] D. Lee and J.-W. Choi, "Deft-an: Dense frequency-time attentive network for multichannel speech enhancement," *IEEE Signal Processing Letters*, vol. 30, pp. 155–159, 2023.

[13] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, 2015, pp. 504–511.

[14] E. Guizzo, C. Marinoni, M. Pennese, *et al.*, "L3das22 challenge: Learning 3d audio sources in a real office environment," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9186–9190.

[15] M. Hao, J. Yu, and L. Zhang, "Spatial-temporal graph convolution network for multichannel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6512–6516.

[16] S. Jia, S. Jiang, S. Zhang, M. Xu, and X. Jia, "Graph-in-graph convolutional network for hyperspectral image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 1157–1171, 2022.

[17] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "Tparn: Triple-path attentive recurrent network for time-domain multichannel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6497–6501.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[19] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[20] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[21] W. Rao, Y. Fu, Y. Hu, *et al.*, "Conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 679–686.

[22] L. Shubo, Y. Fu, J. Yukai, *et al.*, "Spatial-dccrn: Dccrn equipped with frame-level angle feature and hybrid filtering for multi-channel speech enhancement," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 436–443.

[23] J. Cheng, C. Pang, R. Liang, J. Fan, and L. Zhao, "Dual-path dilated convolutional recurrent network with group attention for multi-channel speech enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–2.

[24] J. Bai, S. Huang, H. Yin, Y. Jia, M. Wang, and J. Chen, "3d audio signal processing systems for speech enhancement and sound localization and detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–2.

[25] H. Wang, Y. Fu, J. Li, M. Ge, L. Wang, and X. Qian, "Stream attention based u-net for l3das23 challenge," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–2.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2010, pp. 4214–4217.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

Speech Dereverberation with Deconvolution Regularized by Denoising

Haonan Hu, Ziye Yang, Jie Chen, Lijun Zhang

Research and Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen, China
CIAIC, School of Marine Science and Technology, Northwestern Polytechnical University, China
{hnhu,yzy97}@mail.nwpu.edu.cn, dr.jie.chen@ieee.org, zhanglj7385@nwpu.edu.cn

***Abstract*—Deconvolution-based speech dereverberation continues to present challenges due to the difficulties in accurately acquiring Room Impulse Responses (RIRs) and the inherently ill-conditioned nature of deconvolution. Despite advancements in RIR measurement and estimation, substantial room for improvement remains in addressing the latter challenge. This paper proposes a novel prior-driven dereverberation framework utilizing Regularization by Denoising (RED) to incorporate data priors into the deconvolution process, thereby addressing this persistent challenge. Specifically, we formulate the dereverberation process via an optimization problem with the additional regularizer and the Half Quadratic Splitting (HQS) strategy is then utilized to solve the optimization problem. Experimental validation conducted on both the RIR simulation platform pyroomacoustics and the realistic acoustics platform SoundSpaces demonstrates the efficacy of our framework, even in the presence of environmental noise and RIR errors.**

***Index Terms*—Speech dereverberation, ill-conditioned deconvolution, deep priors, Regularization by Denoising, SoundSpaces**

I. INTRODUCTION

In enclosed spaces, speech signals propagate with inevitable energy attenuation and are subject to reflections off surfaces and objects, resulting in an acoustic phenomenon known as reverberation. This phenomenon can significantly degrade speech quality and intelligibility. Moreover, reverberation may adversely impact the performance of Automatic Speech Recognition (ASR) systems, particularly in scenarios compounded by additive noise. As a result, speech dereverberation technology has garnered considerable interest within the speech signal processing community.

Reverberation is physically modeled by convolving an anechoic speech utterance with an acoustic path, naturally leading to the use of deconvolution strategies. However, speech dereverberation has historically faced two primary challenges. Firstly, in real-world settings, the accurate acquisition and estimation of room impulse responses (RIRs), which is crucial for analyzing acoustic environments, present significant challenges [1], [2]. Secondly, even with available RIR data, the deconvolution operation for speech dereverberation remains an ill-conditioned inverse problem [3], [4], rendering inverse processing difficult, if not impossible. This difficulty is particularly pronounced in the presence of unavoidable environmental noise and minor RIR perturbations, which can significantly contribute to erroneous results. Recent advancements in the utilization of data-driven and statistical methods for RIR measurement and estimation [5]–[7] have contributed to progress in addressing the first challenge. However, there has been limited, if any, effort directed towards effectively tackling the second challenge—the ill-conditioned problem associated with speech deconvolution.

In fact, the ill-conditioned deconvolution problem encountered in speech dereverberation bears similarities to those encountered in image deconvolution/deblurring [8], which have been successfully addressed through the significant utilization of image priors. Among candidate strategies, the plug-and play method [9]–[15] shows its superior effectiveness in image and speech processing. This approach aims to incorporate data priors into the optimization iterations via a deep denoising algorithm, thereby facilitating the solution process of ill-conditioned deconvolution problems.

Inspired by this advance, our work aims to introduce a prior-driven deconvolution-based dereverberation framework to more effectively address the ill-conditioned problem under reverberant-noisy scenarios. Specifically, we formulate the speech dereverberation process through deconvolution using an optimization problem integrating an additional regularizer. This regularizer is not explicitly handcrafted; instead, we employ the Plug-and-Play (PnP) strategy, specifically the Regularization by Denoising (RED) strategy [16], to incorporate deep priors extracted from data. Additionally, we introduce the Half Quadratic Splitting (HQS) method to further solve the optimization problem. Extensive experiments on the widely employed RIR simulation platform, pyroomacoustics[17], and the highly realistic acoustics platform, SoundSpaces [18], [19], to evaluate the performance of the proposed framework.

Notation. Normal font letters x and X denote scalars, and boldface small letters $\mathbf{x}$ denote column vectors. Boldface capital letters $\mathbf{X}$ represent matrices. The operator $(\cdot)^\top$ denote matrix transpose and $(\cdot)^*$ denote conjugate, and $\langle\cdot\rangle$ represents the inner product.

II. PROBLEM FORMULATION

We consider the scenario where a single-channel microphone receives the reverberant speech interfered by the additive noise. The captured signal $y(t)$ can be modeled by the

The work was supported in part by Shenzhen Science and Technology Program JCYJ20230807145600001, TCL science and technology innovation fund, Shaanxi Key Industrial Innovation Chain Project 2022ZDLGY01-02.

following equation:

$$y(t) = h(t) * s(t) + n(t) \tag{1}$$

where t indexes discrete time and $s(t)$ represents the clean speech signal. $h(t)$ represents the time-invariant RIR, characterizing all reflections and attenuations along the propagation path from the source to the microphone. $n(t)$ represents the additive noise, and $*$ denotes the convolution operation. To facilitate the method presentation, (1) is often written in the following form as

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{2}$$

where $\mathbf{s}$ and $\mathbf{y}$ denotes the vector forms for clean speech and observed signals, and $\mathbf{H}$ is the convolutive matrix formed according to $h(t)$. Then the fidelity term of dereverberation process can be formulated by an inverse problem such that

$$\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2. \tag{3}$$

Given (3) is inherently ill-conditioned, it is essential to incorporate the regularization term to stabilize the solution process and enhance the quality of the solution of this inverse problem, which can be written as

$$\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 + \mathcal{R}(\mathbf{s}) \tag{4}$$

where $\mathcal{R}(\mathbf{s})$ represents the regularization term. Recently, considerable effort has been invested in meticulously designing the appropriate regularization term so as to integrate priors into the optimization problem, such as the sparsity prior [20] and the Complex Generalized Gaussian Prior [21]. Nevertheless, it is non-trivial to handcraft such effective regularizer for (4) with the efficient solution methodology.

III. THE PROPOSED FRAMEWORK

In this work, rather than elaborately handcrafting the regularizer, we propose to directly derive priors from speech data to facilitate the problem in (4). Specifically, we consider the regularization in the form as

$$\mathcal{R}(\mathbf{s}) = \frac{1}{2}\langle \mathbf{s}, \mathbf{s} - f(\mathbf{s}) \rangle, \tag{5}$$

where $f(\cdot)$ denotes an off-the-shelf denoiser. This form, called RED, measures the inner product between the desired speech and its denoised residual [16]. Different from the prototype PnP strategy, RED leverages a specific formulation and demonstrates advantageous derivative characteristics under mild assumptions. Integrating RED into (4), we can obtain the problem formulated as

$$\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 + \frac{\rho}{2}\mathbf{s}^\top(\mathbf{s} - f(\mathbf{s})) \tag{6}$$

with ρ being the regularization parameter.

To solve the problem in (6), we initially employ the variable splitting technique to introduce an auxiliary variable $\mathbf{z}$, leading to the following equivalent constrained optimization formulation

$$\begin{aligned} &\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 + \frac{\rho}{2}\mathbf{z}^\top(\mathbf{z} - f(\mathbf{z})) \\ &\text{s.t.} \quad \mathbf{z} = \mathbf{s}. \end{aligned} \tag{7}$$

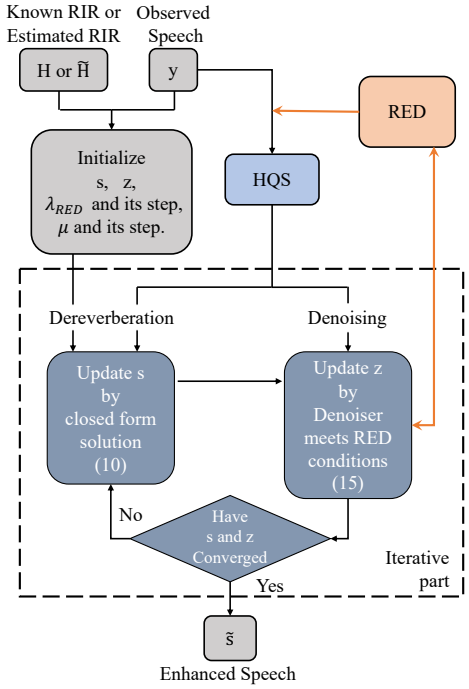


Fig. 1. Block diagram of the proposed framework.

Subsequently, we tackle (7) via the HQS algorithm [22], involving an additional quadratic penalty term

$$\mathcal{L}(\mathbf{s}, \mathbf{z}) = \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 + \frac{\lambda}{2}\|\mathbf{s} - \mathbf{z}\|^2 + \frac{\rho}{2}\mathbf{z}^\top(\mathbf{z} - f(\mathbf{z})), \tag{8}$$

where λ denotes the penalty parameter. The variables in (8) can then be optimized through the following alternating minimization problems with respect to $\mathbf{x}$ and $\mathbf{z}$ respectively, with $(\cdot)^{(n)}$ in the following content denoting the nth iteration.

1) Optimization with respect to $\mathbf{s}$*:* The optimization of (8) becomes:

$$\mathbf{s}^{(n+1)} = \arg\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 + \frac{\lambda}{2}\|\mathbf{s} - \mathbf{z}^{(n)}\|^2. \tag{9}$$

Given the assumption that $\mathbf{H}$ is available, the closed-form solution to (9) regarding $\mathbf{s}$ can be derived by

$$\mathbf{s}^{(n+1)} = \left(\mathbf{H}^\top\mathbf{H} + \frac{\lambda}{2}I\right)^{-1}\left(\mathbf{H}^\top\mathbf{y} + \frac{\lambda}{2}\mathbf{z}^{(\mathbf{n})}\right). \tag{10}$$

2) Optimization with respect to $\mathbf{z}$*:* The optimization problem (8) reduces to:

$$\mathbf{z}^{(n+1)} = \arg\min_{\mathbf{z}} \frac{\lambda}{2}\left\|\mathbf{s}^{(n+1)} - \mathbf{z}\right\|^2 + \frac{\rho}{2}\mathbf{z}^\top(\mathbf{z} - f(\mathbf{z})) \tag{11}$$

Setting the derivative of the above optimization w.r.t. $\mathbf{z}$ to zero, it becomes feasible to directly minimize it through an iterative scheme, leading to the following equation:

$$\lambda(\mathbf{z}^{(n+1)} - \mathbf{s}^{(n+1)}) + \rho(\mathbf{z}^{(n+1)} - f(\mathbf{z}^{(n+1)})) = 0. \tag{12}$$

The solution to this problem can be achieved via the fixed-point iteration:

$$\lambda(\mathbf{z}^{(n+1,i)} - \mathbf{s}^{(n+1)}) + \rho(\mathbf{z}^{(n+1,i)} - f(\mathbf{z}^{(n+1,i)})) = 0, \tag{13}$$

leading to

$$\mathbf{z}^{(n+1,i)} = \frac{\lambda}{\lambda+\rho}\mathbf{s}^{(n+1,i)} + \frac{\rho}{\lambda+\rho} f(\mathbf{z}^{(n+1,i)}). \tag{14}$$

For simplification, this can be rewritten as:

$$\mathbf{z}^{(n+1,i)} = \mu\mathbf{s}^{(n+1,i)} + (1-\mu) f(\mathbf{z}^{(n+1,i)}), \tag{15}$$

with $\mu = \frac{\lambda}{\lambda+\rho}$ and the inner iteration $i = 1, \cdots, I$, where μ serves as a balance parameter between the estimate provided by $\mathbf{s}^{(n+1)}$ and the denoiser $f(\mathbf{z})$.

The overall algorithm is illustrated in Fig. 1, where variables $\mathbf{s}$ and $\mathbf{z}$ are updated iteratively until convergence.

Remark 1: Though mathematically easy to follow, $\mathbf{H}^\top\mathbf{H}$ in matrix form results in inefficient computational processes. Computing in frequency domain presents to be more effective. In the frequency domain, the signal model in (1) writes

$$Y(f) = H(f)S(f) + N(f) \tag{16}$$

with f indicating the frequency bin, where $Y(f)$, $H(f)$, $S(f)$ and $N(f)$ are the counterparts of $y(t)$, $h(t)$, $s(t)$ and $n(t)$ in the frequency domain respectively. This way, the solution w.r.t. $S(f)$ in the frequency domain is given by:

$$S^{(n+1)}(f) = \left(|H(f)|^2 + \frac{\lambda}{2}\right)^{-1}\left(H^*(f)Y(f) + \frac{\lambda}{2}Z^{(n)}(f)\right), \tag{17}$$

which effectively reduces the computational burden.

Remark 2: Within the proposed framework, empirical evidence suggests that the adjustment of dynamic parameters can enhance the overall performance. Specifically, we implement dynamical values λ and μ, which are progressively increased from an initial value with the predefined step size.

IV. Experiments

In this section, we first introduce the simulation environments, followed by the details on the experimental settings, and finally discuss the results in several respects.

A. Experimental environments

Simulated scenarios based on pyroomacoustics: In the experiment, we simulated a scenario within a room configured to the dimension of 5 meters (m) × 4 m × 6 m, where a single microphone is placed at a randomly chosen position. Clean utterances, randomly selected from the Voice Bank corpus [23], were convolved with the RIRs generated via pyroomacoustics [17] to synthesize reverberant speech, where the reverberation times (T_{60}) were set to 190 millisecond (ms), 430 ms and 890 ms respectively. To test the method under varying levels of the additive noise, we considered white Gaussian noise (WGN) and added it to the reverberant speech with the signal-to-ratios (SNRs) setting to 0 dB, 10 dB, and 20 dB respectively.

Moreover, it is inevitable that various types of errors will manifest in the estimation of RIR, it is thus necessary to assess the efficacy of the proposed method under such conditions of uncertainty. Specifically, to simulate the scenario of RIR estimation errors, the accurate RIR ($h(t)$) generated by pyroomacoustics was replaced with an approximate representation, denoted as $\tilde{h}(t)$, given by

$$\tilde{h}(t) = h(t) + (\alpha|h(t)|)\, z(t). \tag{18}$$

The term $(\alpha|h(t)|)z(t)$ introduces a magnitude related perturbation to $h(t)$, with α a scaling constant $z(t)$ a random error [24].

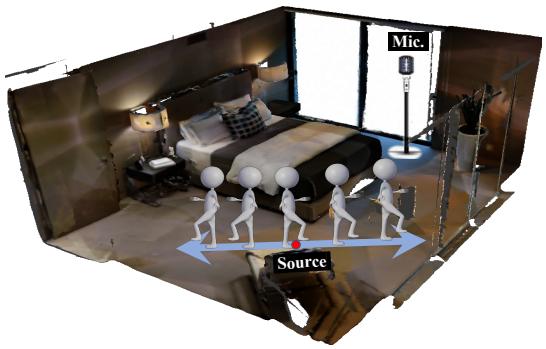


Fig. 2. An example of the experimental scenes simulated on SoundSpaces, showing positions of source and microphone. The red dot indicates the reference position of the sound source, and the source moves in the directions of the bidirectional arrow.

Simulated scenarios based on SoundSpaces: To assess the method in real-world scenarios with more realistic RIR error, we applied SoundSpaces [18], [19], a platform developed by Meta, to simulate another experimental environment. Due to the capability of SoundSpaces to closely mimic real world acoustic environments, it could precisely simulate the sound propagation and reflection across diverse environments, including reverberation and attenuation within rooms, thereby providing a practical and highly realistic testing ground. In addition, SoundSpaces enables the insertion of arbitrary sound sources across various publicly available 3D environments. Therefore, we could simulate the close-to-real-world scenario within closed rooms selected from the Matterport3D dataset [25], where the initial 3-dimensional coordinates of the microphone and source were randomly generated to set the reference position, thereby measuring the exact RIRs.

For this scenario, we also considered to establish the condition of RIR estimation error: First, we set a reference point of source (as mentioned before), serving as the focal point for all source movements. Then, we move the source along a random direction aligned with a straight line extending in both directions of the bidirectional arrow (as indicated by bidirectional arrow in Fig. 2). At each new position reached with a step size of 0.1 m, an RIR measurement was conducted. Afterwards, considering the scenario of reverberation with noise, clean speech signals were convolved with these measured RIRs, followed by the addition of WGN at SNR of 20 dB. During the evaluation, We employed the RIR corresponding to the focal point as $\tilde{\mathbf{H}}$, at each relocated position to derive (10) within the proposed framework.

TABLE I

THE RESULTS OF ALL COMPARISON METHODS IN THE SIMULATED SCENARIO AT DIFFERENT SNRS WITH VARIOUS T_{60}. THE TERM OF ITERATION (ITER.) DENOTES THE QUANTITY OF ITERATIONS REQUIRED FOR THE METHOD TO ACHIEVE CONVERGENCE. THE BEST RESULTS ARE IN BOLD.

		T_{60} →	190 ms				430 ms				890 ms			
SNR (dB)	Error(α) (%)	Methods↓	STOI ↑	PESQ ↑	F-SNR (dB)↑	Iter. ↓	STOI ↑	PESQ ↑	F-SNR (dB)↑	Iter. ↓	STOI ↑	PESQ ↑	F-SNR (dB)↑	Iter. ↓
0 dB	0%	Observed	0.631	1.035	3.772	-	0.548	1.028	3.482	-	0.554	1.048	3.480	-
		JE	0.752	1.074	5.965	81	0.634	1.042	3.400	250	0.728	1.099	6.052	222
		Proposed-s	0.763	1.099	6.318	84	0.648	1.039	3.996	238	0.755	1.156	7.368	270
		Proposed-d	**0.821**	**1.291**	**9.737**	**35**	**0.769**	**1.190**	**7.349**	**38**	**0.784**	**1.332**	**8.782**	**40**
10 dB	0%	Observed	0.718	1.078	5.721	-	0.607	1.034	4.911	-	0.611	1.100	4.496	-
		JE	0.862	1.625	10.636	**36**	0.801	1.154	7.250	105	0.827	1.698	10.796	124
		Proposed-s	0.868	1.683	11.065	**36**	0.811	1.208	8.061	112	0.828	1.721	11.869	103
		Proposed-d	**0.881**	**1.868**	**11.896**	**36**	**0.836**	**1.368**	**10.116**	**32**	**0.848**	**1.761**	**11.994**	**36**
20 dB	0%	Observed	0.772	1.313	7.263	-	0.659	1.095	6.205	-	0.635	1.209	5.149	-
		JE-0	0.897	2.104	11.392	**9**	0.879	1.902	12.321	40	0.873	2.109	13.483	**5**
		Proposed-s	**0.901**	2.206	11.430	10	**0.888**	2.049	12.932	**36**	0.881	2.482	14.962	24
		Proposed-d	0.900	**2.266**	**11.464**	10	**0.888**	**2.118**	**13.009**	**36**	**0.883**	**2.644**	**15.489**	8
	15%	JE	0.891	2.033	**11.050**	**7**	0.882	1.774	11.050	44	0.862	1.887	12.010	43
		Proposed-d	**0.896**	**2.188**	11.004	9	**0.888**	**2.068**	**11.759**	**36**	**0.876**	**2.229**	**12.455**	**8**

B. Experimental settings

Deep prior construction: Benefiting from the flexibility of the proposed framework, it permits the integration of any denoiser to incorporate deep speech priors. Therefore, we directly utilized a pre-trained denoiser, known as speech enhancement generative adversarial network (SEGAN) [26], to focus on the scope of this work. This denoiser, which is developed in an end-to-end manner within an adversarial framework, has been trained on an extensive noisy speech dataset, thereby deemed effective for speech denoising in complex noisy scenarios.

Method comparison and evaluation: For the method comparison, here we consider the joint enhancement framework [24] (denoted by `JE`), which utilizes the PnP strategy to simultaneously execute dereverberation and denoising within a unified framework. Furthermore, within the proposed method, different strategies for parameter selection are evaluated, including the static strategy (denoted by `Proposed-s`) and the dynamic strategy (denoted by `Proposed-d`).

To quantify the experimental results, we utilize the objective metrics including short-time objective intelligibility measure (STOI)[27], perceptual evaluation of speech quality (PESQ)[28] and frequency weighted segmental SNR (F-SNR)[29], to evaluate speech intelligibility and quality. Beside, the Word Error Rate (WER) is also employed to verify the influence of the proposed method on the backend ASR system, which was facilitated through the SpeechRecognition library[30].

Implementation details: Concerning the parameter settings within the framework, λ is set to 2.2 and μ is set to 0.28. For `Proposed-s`, the parameter settings remain static throughout the iterations. Conversely, for `Proposed-d`, these settings gradually increase from their respective base values with the step sizes of 0.28 for λ and 0.015 for μ. For all the above scenarios, the test speech signals are sampled at 16 kHz. Specifically, to simulate the scenario of RIR estimation errors, the accurate RIR $h(t)$ generated by pyroomacoustics is replaced with an approximate representation $\tilde{h}(t)$, perturbated by (18), where the scaling constant α is set from 0 to 0.5 (0% to 50%), and the random error $z(t)$ is modeled using a Gaussian distribution with zero mean and unit variance $z(t) \sim \mathcal{N}(0, 1)$. When performing the calculation of the "Optimization with respect to s" step in the frequency domain for more efficient computing, we utilize Fast Fourier Transform (FFT) with the following parameter settings: To capture the short-time characteristics of the speech signal, the frame length is set to 25 ms (400 samples at a sampling rate of 16 kHz) with a frame shift of 10 ms (160 samples) with Hamming window applied. The FFT length is chosen to be 512, which is a power of two and greater than the frame length, ensuring efficient computation. To ensure smooth transitions between frames, a 50% overlap is used, meaning each frame overlaps with the previous frame by 200 samples.

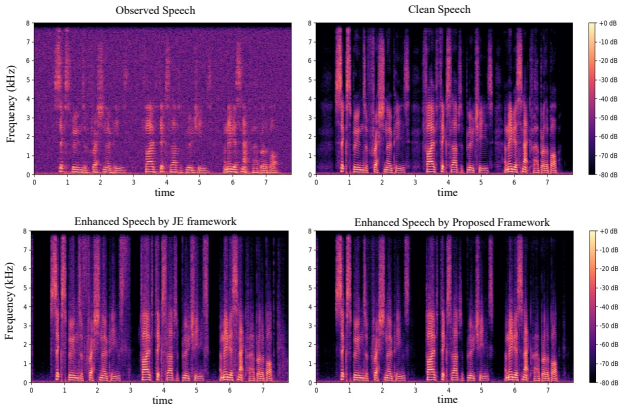


Fig. 3. Spectograms of observed (top left), clean (top right), `JE` (Bottom Left) and the proposed method (bottom right) in the simulated scenario.

C. Results discussion

From Table I, we find that the proposed method surpasses `JE` across all the scenarios, particularly under the condition

TABLE II
THE WER RESULTS OF PROPOSED-D AND JE IN THE SIMULATED SCENARIO AT SNR=20 DB.

	$T_{60}\rightarrow$	190 ms	430 ms	890 ms
Error (%)	Methods↓	WER (%) ↓		
0%	JE	27	25	41
25%	JE	30	29	48
0%	Proposed-d	24	20	32
25%	Proposed-d	24	25	32

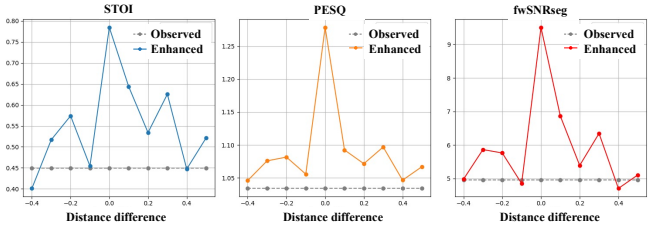


Fig. 4. Trend of F-SNR (left), PESQ (middle) and STOI (right) considering the RIR measurement errors on the SoundSpaces Platform.

characterized by high T_{60} and intense WGN. Additionally, the efficacy of the dynamic parameter adjustment strategy is validated, as `Proposed-d` yields almost all the best results. Meanwhile, it is evident from the values of `Iter.` that under a consistent convergence criterion, the proposed framework converges more rapidly than the comparison methods. Moreover, for visual comparison, we take the scenario of SNR = 20 dB and T_{60} = 430 ms as an example, and illustrate the spectrograms in Fig. 3. From the spectrograms, we can conclude that the inclusion of the RED strategy provides valuable priors for the deconvolution task, which helps to better preserve the naturalness and clarity of the speech signal during the dereverberation process.

In order to explore the impact of dereverberation methods on the performance of back-end ASR, we in addition present the WER results of `JE` and `Proposed-d` in Table II. From the findings, it is apparent that the integration of `Proposed-d` exhibits a notable enhancement in the efficacy of the ASR system, regardless of the presence or absence of errors in RIR estimation.

Fig. 4 presents the results under the scenarios simulated based on SoundSpaces. From these results, we can see that if the RIR measured at alternative locations is considered as the RIR estimation at the current location, which corresponds to the scenarios of RIR estimation with errors, our proposed framework still exhibits a capacity for dereverberation to a certain extent.

V. CONCLUSION

In this paper, we proposed a prior-driven dereverberation framework, providing an effective solution to the ill-conditioned deconvolution problem. Experiments were conducted on two respective platforms, including SoundSpaces, a realistic acoustic simulation platform. The results demonstrate that our framework is capable of solving ill-conditioned problem and boosting the speech dereverberation performance, even in the presence of environmental noise and RIR errors. Future work will focus on updating RIR estimates during iterations to further improve the performance.

REFERENCES

[1] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio engineering society*, vol. 50, pp. 249–262, 2002.

[2] N. Hahn and S. Spors, "Comparison of continuous measurement techniques for spatial room impulse responses," *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1638–1642, 2016.

[3] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, pp. 165–169, Jul. 1979.

[4] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *Journal of Sound and Vibration*, vol. 102, pp. 217–228, 1985.

[5] N. Mohanan, R. Velmurugan, and P. Rao, "Speech dereverberation using nmf with regularized room impulse response," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4955–4959, 2017.

[6] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 221–225, 2021.

[7] M. Pezzoli, D. Perini, A. Bernardini, F. Borra, F. Antonacci, and A. Sarti, "Deep prior approach for room impulse response reconstruction," *Sensors*, vol. 22, p. 2710, 2022.

[8] X. Wang, J. Chen, and C. Richard, "Tuning-free plug-and-play hyperspectral image deconvolution with deep priors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[9] J. Chen, M. Zhao, X. Wang, C. Richard, and S. Rahardja, "Integration of physics-based and data-driven models for hyperspectral image unmixing: A summary of current methods," *IEEE Signal Processing Magazine*, vol. 40, pp. 61–74, Mar. 2023.

[10] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Processing Magazine*, pp. 18–44, Mar. 2021.

[11] Z. Yang, W. Yang, K. Xie, and J. Chen, "Integrating data priors to weighted prediction error for speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3908–3923, 2024.

[12] Z. Yang, W. Yang, K. Xie, and J. Chen, "Speech dereverberation using weighted prediction error with prior learnt from data," *European Signal Processing Conference (EUSIPCO)*, pp. 356–360, 2023.

[13] Z. Yang, J. Chen, C. Richard, and J. Li, "Plug-and-play wpe guided by deep spectrum estimation for speech dereverberation," *IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2024.

[14] C. Chang, Z. Yang, and J. Chen, "Plug-and-play mvdr beamforming for speech separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1346–1350, 2024.

[15] K. Matsumoto and K. Yatabe, "Determined bss by combination of iva and dnn via proximal average," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[16] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (red)," *SIAM Journal on Imaging Sciences*, vol. 10, pp. 1804–1844, 2017.

[17] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 351–355, 2018.

[18] C. Chen, U. Jain, C. Schissler, *et al.*, "Soundspaces: Audio-visual navigation in 3d environments," *Computer Vision–ECCV*, pp. 17–36, Aug. 2020.

[19] C. Chen, C. Schissler, S. Garg, *et al.*, "Soundspaces 2.0: A simulation platform for visual-acoustic learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8896–8911, 2022.

[20] M. Witkowski and K. Kowalczyk, "Split bregman approach to linear prediction based dereverberation with enforced speech sparsity," *IEEE Signal Processing Letters*, vol. 28, pp. 942–946, 2021.

[21] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1509–1520, 2015.

[22] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, pp. 1–122, 2011.

[23] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," *International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Nov. 2013. DOI: 10.1109/icsda.2013.6709856.

[24] A. Raikar, S. Basu, L. Pandey, and R. Hegde, "Multi-channel joint dereverberation and denoising using deep priors," *IEEE India Council International Conference (INDICON)*, pp. 1–6, 2018.

[25] A. Chang, A. Dai, T. Funkhouser, *et al.*, "Matterport3d: Learning from rgb-d data in indoor environments," *2017 International Conference on 3D Vision (3DV)*, pp. 667–676, 2017.

[26] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 2125–2136, Sep. 2011. DOI: 10.1109/tasl.2011.2114881.

[28] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, pp. 749–752, 2001.

[29] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1462–1469, Jul. 2006. DOI: 10.1109/tsa.2005.858005.

[30] A. Zhang, "Speechrecognition: Speech recognition module for python, supporting several engines and apis, online and offline," *GitHub repository*, 2022.

Enhanced RefineDNet for Single Image Dehazing

Jingyu Ren* and Lei Yang†
* School of Computer Science and Engineering, Macau University of Science and Technology, Macau
E-mail: renjingyu0225@gmail.com Tel/Fax: +86-13971033951
† School of Computer Science and Engineering, Macau University of Science and Technology, Macau
E-mail: leiyang@must.edu.mo Tel/Fax: +853-88972045

***Abstract*—Haze-free images are the preconditions of many vision systems, and thus single image dehazing is of great significance in computer vision. RefineDNet [1] is a two-stage weakly supervised dehazing framework which combines the merits of prior-based and learning-based approaches together. However, under high image saturation or adverse weather conditions, its performance is limited. In this work, a modified RefineDNet is proposed by leveraging the saturation line prior (SLP) [2] method in the first stage and a plug-in based on a patch-based diffusion model [3] for image preprocessing to enhance the dehaze ability of the framework. Extensive experiments demonstrate that our modified RefineDNet framework achieves superior haze removal and produces visually pleasing results.**

I. Introduction

Hazy images are often one of the most encountered problems in natural images, mainly affecting the clarity of images. Solving the problem of hazy images has significant practical applications, for instance object detection, image recognition and autonomous driving etc. Hence, haze removal, especially single image dehazing, is vital and has been widely studied recently.

Existing dehazing methods mainly include the prior-based methods, the learning-based methods and their combinations. Among these methods, a two-stage weakly supervised dehazing framework namely RefineDNet, merges the merits of prior-based and learning-based approaches. Although RefineDNet combines the advantages of both prior-based and learning-based approaches, the traditional dark channel prior(DCP) [4][5] used in RefineDNet is prone to artifacts in some cases (e.g., low-contrast images and regions with large changes in color saturation), resulting in misclassification of features and inhibiting the model's generalization ability. Additionally, the defogging effect is poor when dealing with images containing water droplets and interference from rain and snow. Therefore, we aim to address these issues in our research.

In this paper, we will address the problem of image dehazing by using SLP approach and a plug-in based on the diffusion model [6][7][8]. This approach aims to solve the problems associated with the DCP, such as the formation of artifacts and the impact of water droplets and adverse weather conditions.

Compared to the DCP, SLP [2] exploits the linear relationship between saturation and inverse brightness in foggy images. And it allows for a more accurate estimation of transmittance, producing more precise dehazing results and avoiding artifact formation. The patch-based denoising diffusion models rely on the similarity and proximity between pixels in an image, using the pixel value information of local regions to correct noise or interference. These methods achieve size-agnostic image recovery by using a bootstrap denoising process with smooth noise estimation across overlapping patches in the inference process, which helps mitigate the effect of lens water droplets on the image.

The remainder of this paper is organized as follows. Section II reviews the related work. Section III describes the proposed framework with the modifications in detail. Section IV presents experimental results. Finally, Section V concludes this paper.

II. Related works

For traditional single-image dehazing, the main focus is generally on the estimation of atmospheric light [9] and the calculation of transmittance. Raanan Fattal proposed a basic framework for image dehazing [10] based on the estimation of transmitted light, providing a new perspective on image dehazing. With the development and popularity of supervised learning, Pádraig Cunningham's proposal [11] to define features that can classify other unlabeled data further enhanced image dehazing techniques, yielding better results. Since then, researchers have increasingly favored fusing these two frameworks for image dehazing.

In the domain of image dehazing, advancements have been achieved through prior-based, learning-based, and GAN-based methods. Prior-based methods utilize various image priors to estimate transmission maps, enhancing dehazing performance. Notable techniques include the DCP [4], which estimates transmission maps effectively, and the Color Ellipsoid Prior [12], which improves contrast in dehazed images. Learning-based methods leverage neural networks to address dehazing, with early models like DehazeNet [13] focusing on estimating transmission maps from paired images, and more recent approaches such as multi-scale networks [14] advancing performance further. GAN-based methods introduce adversarial training to enhance dehazing results, initially applied by [15] and refined by complex structures and unpaired data techniques [16]. These methods significantly improve image quality by addressing visibility and realism, marking a notable shift from paired data reliance to unpaired data solutions.

Building on these advancements, Shi-Yu Zhao proposed a weakly supervised refinement framework called RefineDNet [1], which fuses various methods for image dehazing. Zhao also proposed a novel perceptual fusion strategy for combining different dehazing results. Experimental results demonstrate

that RefineDNet has excellent dehazing ability and produces visually satisfactory results.

The DCP facing computational limitations. Improvements include methods based on saturation information [17][18] and advanced techniques like Boundary Constraint and Contextual Regularization [19] and the Color-Lines Model [20]. These methods address various dehazing challenges, including dense haze and color distortion, with additional solutions for complex scenarios [21][22][23][24][25].

Building on this, Pengyang Ling proposed a new saturation line prior-based dehazing method [2], which exploits the linear relationship between the inverse of saturation and luminance in localized image blocks to construct a saturation line and estimate transmittance, enhancing dehazing for images with large saturation variations.

Recent advancements in generative models, specifically diffusion-based [26][27][28][29] and score-matching models [30][31] focus on reversing the sequential corruption of data with additive noise to match a standard normal prior. These models, parameterized by deep neural networks, are unified through a continuous-time formulation based on stochastic differential equations [32], achieving remarkable image synthesis capabilities [28]. Concurrently, energy-based models estimate unnormalized probability densities and utilize Langevin sampling [33][34] differing mainly in their contrastive divergence training methods [35][36].

Based on these model Ozan Özdenizci and Robert Legenstein proposed a novel patch-based image restoration algorithm [3] based on the denoising diffusion model. This algorithm implements a bootstrapping denoising process by using smooth noise estimates on overlapping patches during the inference process, achieving state-of-the-art performance and demonstrating strong generalization capabilities for weather-specific and multi-weather image recovery.

III. Method and Model

In this section, the proposed RefineDNet with modification will be presented in detail. The overall framework will be introduced firstly. After that, the perceptual fusion is detailed. Finally, the loss function is described.

A. RefineDNet

RefineDNet is divided into two parts. The first part is the DCP and the second part is adversarial learning. The input part is a single image I_{real}, which will be used to estimate the atmospheric light, to generate a preliminary dehazing image J_{DCP}, and a transmission map T_{DCP}. R_{t} is structured as a U-Net with 8 down-sampled and 8 up-sampled convolutional layers, which serves to adapt the transmission map. R_{j} net, is a ResNet with 9 residual blocks, which is used to reconstruct the network on the initial defogged image to further refine the image. J_{DCP} and T_{DCP} are processed to obtain J_{ref} and T_{ref} images. J_{ref} map is the reconstructed dehazed image and T_{ref} map is the real transmission map. They are used to train the R_{t} network and update the R_{t} network by minimising the distance between I_{real} and I_{rec} to improve the realism and visibility of the dehazed image.

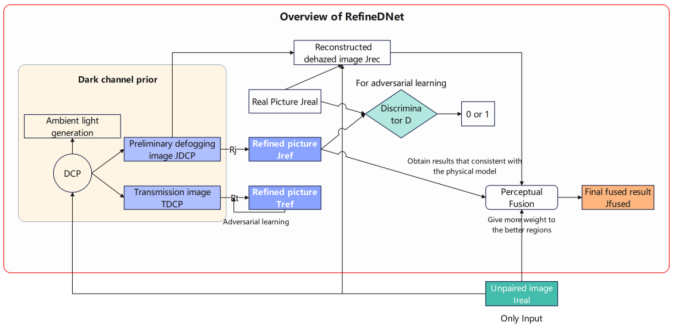


Fig. 1: For RefineNet, it is divided into two parts, the first part is the dark channel prior and the second part is adversarial learning. The DCP section is shown in the beige box and the Adversarial Learning section is shown in the remainder.

J_{rec} image is reconstructed image by estimating transmission and dark channel in which to ensure that the image suit to the physical model. In addition, the discriminator performs adversarial learning by receiving J_{ref} and clear samples J_{real} and enabling weakly supervised learning. Perceptual fusion ensures that the fusion process doesn't affect the visibility by giving more weights to better regions in the dehazed images J_{ref} and J_{rec} based on their reality. And then, the final dehazed image is generated. Figure 1 illustrates the overall algorithmic workflow.

The preliminary DCP results is from:

1) Dark Channel Extraction: Dark Channel Extraction [1]: For a given input RGB image I, we compute the channel minimum image, denoted as I_{min}. Next, we apply max pooling with a kernel size of 5×5 to the additive inverse of I_{min}, and then calculate the additive inverse of the pooling result to obtain the dark channel image I_{dark}. The selection of the dark channel can be formulated as follows:

$$I^{\text{dark}}(x) = -maxpool(-\min_{c \in \{R,G,B\}}(I^c(x))), \tag{1}$$

where I^c refers to one of the RGB channels of I.

2) Transmission Estimation: Transmission Estimation [1]: We get the dark channels at both sides of Koschmieder's law [37] as:

$$I^{\text{dark}}(x) = J^{\text{dark}}(x)t(x) + A(1 - t(x)), \tag{2}$$

where the dark channels of images I and J at pixel x, respectively. Based on the assumption of DCP, the pixels of most non-sky patches in natural images have intensity values near zero in at least one color channel, $J^{\text{dark}}(x) \to 0$. Then,

$$t(x) = 1 - \frac{I^{\text{dark}(x)}}{\Delta}. \tag{3}$$

If A is known, T_{DCP} can be obtained. In addition, we choose a guided filter to make T_{DCP} smooth. The guided filter is using an average pooling with the kernel size of 19×19 and the stride of 1.

B. Patch-Based Diffusion Models

For Patch-Based Diffusion Models, it learns a specific Markov chain [38], which is a random sampling over the full range of image sizes to learn the conditional inverse process. The Gaussian noise distribution [39] is then transformed into a data distribution. First, overlapping the patch size p patches are got from the image and back sampled based on the average estimated noise of neighboring overlapping, to ensure high realism. Then, for each sampling time step t, the additive noise is estimated for all overlapping patch locations d. The overlapping noise estimates are accumulated and normalized based on the number of estimates received per pixel, and an implicit sampling update is performed using the smoothed overall image noise estimate. This process is repeated until time step T, the value of T being defined by us.

1) Conditional Diffusion Models [9]: The conditional diffusion model is able to synthesise new pictures or modify the existing pictures based on the existing pictures according to the conditions or requirements provided by us. The core principle involves learning a conditional reverse process $p_{\boldsymbol{\theta}}(x_0 : \mathbf{T} \mid \tilde{\mathbf{x}})$ without altering the diffusion process $p_{\boldsymbol{\theta}}(x_1 : \mathbf{T} \mid \tilde{\mathbf{x}})$, ensuring that the sampled closely adheres to the data distribution conditioned on $\tilde{\mathbf{x}}$.

During the training phase, we sample $(x_0, \tilde{\mathbf{x}})$ from a paired data distribution $q(x_0, \tilde{\mathbf{x}})$, typically representing a clean image x_0 and its weather-degraded counterpart $\tilde{\mathbf{x}}$. Subsequently, we train a conditional diffusion model by providing $\tilde{\mathbf{x}}$ as input to the reverse process.

Note that conditioning the reverse process on $\tilde{\mathbf{X}}$ maintains its compatibility with implicit sampling. In this formulation one samples from $X_{t-1} \sim p_\theta \left(X_{t-1} \mid X_t, \hat{X} \right)$ with

$$\begin{aligned} \mathbf{x}_{t-1} =& \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta(\mathbf{x}_t, \tilde{\mathbf{x}}, t)}{\sqrt{\bar{\alpha}_t}} \right), \\ &+ \sqrt{1-\bar{\alpha}_{t-1}} \cdot \epsilon_\theta(\mathbf{x}_t, \tilde{\mathbf{x}}, t) \end{aligned} \tag{4}$$

which follows a deterministic reverse path towards x_0 with fidelity to the condition $\tilde{\mathbf{X}}$, starting from $X_t \sim \mathcal{N}(0, I)$.

2) Patch-Based Diffusive Image Restoration: The proposed method achieves high-quality image restoration through a series of structured steps. Initially, an image of arbitrary size is decomposed into overlapping $p \times p$ patches using a lattice-like parsing scheme with a step size of r. Given that overlapping regions may yield different restoration estimates, the method averages the noise estimates in these overlapping regions to ensure consistency between neighboring patches. Specifically, the backsampling process involves estimating the noise at all overlapping patch locations at each sampling time step t, accumulating and normalizing these noise estimates, and then using the smoothed whole-image noise estimates for implicit sampling updates. This approach provides higher fidelity in overlapping regions and avoids edge artifacts compared to simple post-sample averaging. Although a smaller step size r increases the computational burden, it significantly enhances the smoothness of the image. This method is analogous to some recent image editing techniques that steer the inverse process in the latent space to achieve high-quality image restoration results.

Algorithm 1 Patch-Based Diffusive Image Restoration [3]

Require: Weather-degraded image $\tilde{\mathbf{X}}$, conditional diffusion model $\epsilon_\theta(\mathbf{x}_t, \tilde{\mathbf{x}})$, number of implicit sampling steps S, dictionary of D overlapping patch locations.
Ensure: Restored image $\mathbf{X}_t$
0: $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$
0: **for** $i = S, \ldots, 1$ **do**
0: $t = (i-1)\frac{T}{S} + 1$
0: $t_{\text{next}} = (i-2) \cdot \frac{T}{S}$ if $i > 1$ else 0
0: $\hat{\boldsymbol{\Omega}}_t = 0$ and $M = 0$
0: **for** $d = 1, \ldots, D$ **do**
0: $\mathbf{x}_t^{(d)} = \text{Crop}(P_d \circ \mathbf{x}_t)$ and $\tilde{\mathbf{x}}^{(d)} = \text{Crop}(P_d \circ \tilde{\mathbf{X}})$
0: $\hat{\boldsymbol{\Omega}}_t = \hat{\boldsymbol{\Omega}}_t + P_d \cdot \epsilon_\theta(\mathbf{x}_t^{(d)}, \tilde{\mathbf{x}}^{(d)}, t)$
0: $M = M + P_d$
0: **end for**
0: $\hat{\boldsymbol{\Omega}}_t \leftarrow \hat{\boldsymbol{\Omega}}_t \oslash M$
0: $\mathbf{x}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t_{\text{next}}}} \left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t} \cdot \hat{\boldsymbol{\Omega}}_t}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t_{\text{next}}}} \cdot \hat{\boldsymbol{\Omega}}_t$
0: **end for**
0: **return** $\mathbf{x}_t$ =0

C. SLP

The Saturation Line Prior (SLP) [6] is based on the assumption that natural images can be modeled as a collage of projected surfaces, following the locally monochromatic surface assumption [40]. This assumption expresses the radiance of an image $J(x, y)$ as the product of the surface reflectance coefficients $R(x, y)$ and the shading factor $I(x, y)$, with $R(x, y)$ being constant within local patches. The saturation component $S_K(x, y)$ of an image $K(x, y)$ is defined based on the ratio of the minimum to maximum color components in each pixel.

After normalizing the atmospheric light [41], the haze-free image $J_N(x, y)$ can be expressed in terms of the transmission $t(x, y)$ and the saturation component $S_{J_N}(x, y)$. It asserts that under the locally monochromatic surface assumption, the saturation component of the haze-free image remains locally constant.

The saturation component of the normalized hazy image $I_N(x, y)$ is derived from $J_N(x, y)$ and the transmission $t(x, y)$. By introducing the brightness component $V_{I_N}(x, y)$ as the maximum color component, a linear relationship is established between the saturation component $S_{I_N}(x, y)$ and the inverse of the brightness component $V_{I_N}(x, y)$.

$$S_{I_N}(x, y) = k \cdot V_{I_N}^{-1}(x, y) + b \tag{5}$$

The formulation expresses the saturation line prior, where k and b represent the slope and intercept of the line, respectively. These parameters are determined based on the locally constant saturation component $\hat{S}_{J_N}$ and the locally constant transmission $\bar{t}$. With this prior, transmission estimation results in the construction of a saturation line, facilitating transmission

inference based on local pixel distribution characteristics in terms of saturation and brightness.

Due to this relationship, our implementation process is divided into two calculations of transmission. The first calculation divides the image into blocks with *patch_size* and calculates the transmission of the whole image. The second time we calculate the transmission of the center region, dropping the region with an edge of half *patch_size*. The reason for this step is that, for people who want to take a photo, he tends to place the subject in the area right in the middle of the image, so the center region should be the parts that remain when a light fog concentration. Therefore, in the first calculation of the transmission mean, the transmission of the edges is divided by 2, while the center area remains unchanged. This results in better dehazing of the edge areas compared to the center, and helps improve the entire image clarity. The second is to calculate the saturation value and calculate the transmission map twice. The transmission map we calculated us twice.

D. Loss function

The loss function of the modified RefineDNet includes three main components: GAN loss (L_G), reconstruction loss (L_{rec}), and identity loss (L_{idt}).

1) GAN loss: GAN loss (L_G) [1] is used to update the generator and discriminator through adversarial training. In this method, the discriminator (D) attempts to distinguish between real images (J_{real}) and generated images ($R_J(J_{\text{DCP}})$, while the generator (R_J) tries to deceive the discriminator by making the generated images as close to real images as possible. The specific formula is:

$$\begin{aligned} L_G(R_J, D) = E_{J_{\text{real}} \sim J_{\text{real}}}[\log D(J_{\text{real}})] \\ + E_{J_{\text{DCP}} \sim J_{\text{DCP}}}[\log(1 - D(R_J(J_{\text{DCP}})))]' \end{aligned} \tag{6}$$

where J_{real} represents the set of all possible real samples, and J_{DCP} represents the set of all possible input samples.

2) Reconstruction loss: Reconstruction loss (L_{rec}) [1] is used to regularize the reconstructed hazy image. It is defined as the distance between the input hazy image (I_{real}) and the reconstructed image (I_{rec}). The formula for the reconstruction loss is:

$$L_{\text{rec}} = \|I_{\text{real}} - I_{\text{rec}}\|. \tag{7}$$

3) Identity loss: Identity loss (L_{idt}) [1] is introduced to suppress the artifacts that may be generated by RefineDNet, encouraging the network to output images that are as close to real, clear images as possible. It is defined as the distance between the real sample (J_{real}) and the generated sample ($R_J(J_{\text{real}})$). The formula is:

$$L_{\text{idt}} = \|J_{\text{real}} - R_J(J_{\text{real}})\|. \tag{8}$$

Combining all these loss components, the overall objective function is formulated as:

$$R_J^*, R_T^* = \arg \min_{R_T, R_J} \max_D \lambda L_G + L_{\text{rec}} + L_{\text{idt}}, \tag{9}$$

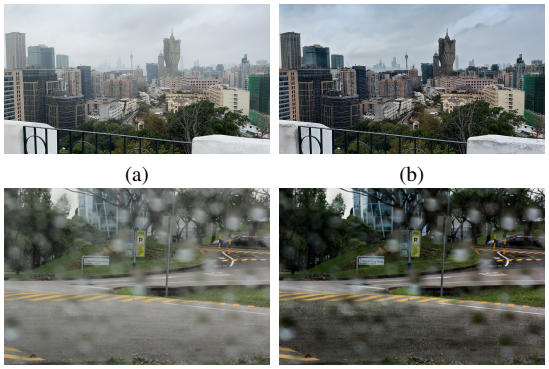


Fig. 2: Subfigure(a) and Subfigure(c) is the original figure with fog. Subfigure (b) presents the results obtained using the retrained RefineDNet model. As illustrated, RefineDNet effectively removes artifacts and enhances the deblurring effects produced by the DCP method. However, subfigure (d) demonstrates that RefineDNet is less effective when processing images containing water droplets.

where λ is a hyperparameter indicating the weight of the GAN loss, with a default value set to 0.02. This loss function design ensures that RefineDNet generates high-quality images while suppressing artifacts.

IV. Experimental result

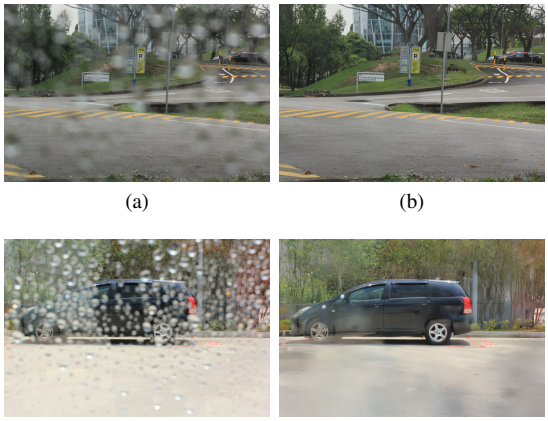
For the re-trained RefineDNet, we get the results shown in Fig. 2 subfigure(b) to demonstrate the dehazing effect of the modified RefineDNet. However, the RefineDNet has difficult in dehazing figure with droplets via subfigure(d).

For the patch-based diffusion model, it works better when dealing with images with a uniform distribution of water droplets, but we also found that its effectiveness decreases when there are more water droplets at the edges or when they are not uniformly distributed. This is because the random sampling method we used cannot average the unevenly distributed water droplet images, and this method will make the regions with more water droplets not be processed effectively. The results are presented in Figs. 3.

For the SLP, the results were better in the experiments with the following images, with significant improvements in both clarity and saturation. Compared to RefineDNet, the image shows a significant improvement in clarity. The results are presented in Fig. 4. Table 1 shows the effectiveness of our dehazing with quantitative data.

V. Conclusions

In this study, a modified RefineDNet for single image dehazing is proposed by integrating the Saturation Line Prior (SLP) and a patch-based diffusion model. The introduced SLP leverages the linear relationship between saturation and inverse brightness in hazy images, providing a more accurate estimation of transmittance and avoiding the artifact formation commonly associated with the DCP. Additionally, the



(a) (b) (c) (d)

Fig. 3: Denoising results based on the patch diffusion model in images with varying droplet distributions. Subfigure (a) and subfigure (c) depict the original images, with (a) containing evenly distributed droplets and (c) containing unevenly distributed droplets. Subfigure (b) and subfigure (d) present the corresponding denoising results.



(a) (b) (c)

Fig. 4: Comparison of defogging results using RefineDNet and a modified version of RefineDNet. The modified approach replaces the DCP method with the SLP method within RefineDNet, yielding improved defogging performance. Subfigure (a) shows the original image, subfigure (b) displays the result obtained by the original RefineDNet, and subfigure (c) illustrates the result achieved using the SLP method.

patch-based denoising diffusion model, which operates on the similarity and proximity between pixels, further enhances the dehazing process by ensuring smooth noise estimation across overlapping patches. This newly integrated framework offers a more robust and versatile dehazing solution.

TABLE I: Comparative Analysis of RefineDNet and SLP Based on Various Image Quality Metrics

Metrics	Comparison	
	(a) vs (b)	(a) vs (c)
SSIM	0.82226	0.680736
PSNR	15.971763	12.631061
Sharpness	75.621749	353.209707
Contrast	30.727260	52.482191
Exposure	144.964690	112.633122

ACKNOWLEDGMENT

The research of the authors was supported by the Science and Technology Development Fund, Macau SAR (Grant No. 0031/2022/A1).

REFERENCES

[1] S. Zhao, L. Zhang, Y. Shen, *et al.*, "Refinednet: A weakly supervised refinement framework for single image dehazing," *IEEE Transactions on Image Processing*, vol. 30, pp. 3391–3404, 2021.

[2] P. Ling, H. Chen, X. Tan, *et al.*, "Single image dehazing using saturation line prior," *IEEE Transactions on Image Processing*, 2023.

[3] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[4] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[5] S. Lee, S. Yun, J. Nam, *et al.*, "A review on dark channel prior based image dehazing algorithms," *EURASIP Journal on Image and Video Processing*, vol. 2016, pp. 1–23, 2016.

[6] F.-A. Croitoru, V. Hondru, R. T. Ionescu, *et al.*, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[7] T. Luhman and E. Luhman, "Improving diffusion model efficiency through patching," *arXiv preprint*, 2022. eprint: 2207.04316.

[8] Z. Wang, Y. Jiang, H. Zheng, *et al.*, "Patch diffusion: Faster and more data-efficient training of diffusion models," in *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[9] Z. Zhu, Y. Luo, H. Wei, *et al.*, "Atmospheric light estimation based remote sensing image dehazing," *Remote Sensing*, vol. 13, no. 13, p. 2432, 2021.

[10] R. Fattal, "Single image dehazing," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3, pp. 1–9, 2008.

[11] P. Cunningham, M. Cord, and S. J. Delany, "Supervised learning," in *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, Springer Berlin Heidelberg, 2008, pp. 21–49.

[12] T. M. Bui and W. Kim, "Single image dehazing using color ellipsoid prior," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 999–1009, Feb. 2018.

[13] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[14] A. Wang, W. Wang, J. Liu, and N. Gu, "Aipnet: Image-to-image single image dehazing with atmospheric illumination prior," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 381–393, Jan. 2019.

[15] H. Zhu, X. Peng, V. Chandrasekhar, *et al.*, "Dehazegan: When image dehazing meets differential programming," in *IJCAI*, 2018, pp. 1234–1240.

[16] X. Yang, Z. Xu, and J. Luo, "Towards perceptual image dehazing by physics-based disentanglement and adversarial training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 7485–7492.

[17] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3522–3533, 2015.

[18] S. E. Kim, T. H. Park, and I. K. Eom, "Fast single image dehazing using saturation based transmission map estimation," *IEEE Transactions on Image Processing*, vol. 29, pp. 1985–1998, 2020.

[19] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 617–624.

[20] R. Fattal, "Dehazing using color-lines," *ACM Transactions on Graphics*, vol. 34, no. 1, pp. 1–14, Dec. 2014.

[21] M. Ju, C. Ding, W. Ren, Y. Yang, D. Zhang, and Y. J. Guo, "Ide: Image dehazing and exposure using an enhanced atmospheric scattering model," *IEEE Transactions on Image Processing*, vol. 30, pp. 2180–2192, 2021.

[22] M. Ju, C. Ding, D. Zhang, and Y. J. Guo, "Bdpk: Bayesian dehazing using prior knowledge," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2349–2362, Aug. 2019.

[23] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2822–2837, Aug. 2021.

[24] C. Ancuti, C. O. Ancuti, C. D. Vleeschouwer, and A. C. Bovik, "Day and night-time dehazing by local airlight estimation," *IEEE Transactions on Image Processing*, vol. 29, pp. 6264–6275, 2020.

[25] W. Wang, A. Wang, and C. Liu, "Variational single nighttime image haze removal with a gray haze-line prior," *IEEE Transactions on Image Processing*, vol. 31, pp. 1349–1363, 2022.

[26] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 2256–2265.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 6840–6851.

[28] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021, pp. 8780–8794.

[29] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8162–8171.

[30] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, no. 4, pp. 695–709, 2005.

[31] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[32] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2020, arXiv preprint. arXiv: 2011.13456.

[33] Y. Du and I. Mordatch, "Implicit generation and generalization in energy-based models," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[34] Y. Song and D. P. Kingma, "How to train your energy-based models," 2021, arXiv preprint. arXiv: 2101.03288.

[35] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[36] T. Tieleman and G. Hinton, "Using fast weights to improve persistent contrastive divergence," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009, pp. 1033–1040.

[37] R. T. Tan, "Visibility in bad weather from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[38] J. R. Norris, *Markov Chains*. Cambridge University Press, 1998.

[39] S. O. Rice, "Distribution of the duration of fades in radio transmission: Gaussian noise model," *Bell System Technical Journal*, vol. 37, no. 3, pp. 581–635, 1958.

[40] S. Tominaga, "Surface identification using the dichromatic reflection model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 7, pp. 658–670, 1991.

[41] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *International Journal of Computer Vision*, vol. 48, pp. 233–254, 2002.

Domain Adaptation by Alternating Learning of Acoustic and Linguistic Information for Japanese Deaf and Hard-of-Hearing People

TAKAHASHI, Kaito*, WAKABAYASHI, Yukoh*, OHTA, Kengo†, KOBAYASHI, Akio‡ and KITAOKA, Norihide*
* Toyohashi University of Technology, Japan
† National Institute of Technology, Anan College, Japan
‡ Yamato University, Japan

***Abstract*—More than half of Japanese people with hearing impairments communicate using speech, however speech recognition systems trained using speech from individuals with normal hearing are unable to achieve sufficient recognition accuracy of speech from individuals with hearing impairments. Therefore, speech recognition systems adapted for individuals with hearing impairments are needed. In this study, we propose a learning method that retains both acoustic and linguistic information to achieve more accurate recognition of speech of the hearing-impaired. Our proposed method performs domain adaptation by alternately switching whether to train the Transformer encoder layer and decoder based on the input speech. By using this method, we can create a speech recognizer adapted to hearing-impaired speech acoustically, while preserving the linguistic information of a general training corpus, thereby improving recognition performance for hearing-impaired speech.**

I. Introduction

In recent years, the accuracy of automatic speech recognition (ASR) has improved, and it is now being utilized in various scenarios, such as smart speakers, voice assistants, and voice input. The ASR systems used in these applications are generally trained with the speech of hearing individuals, and can achieve high recognition accuracy for standard speech. However, it has been reported that such models have low recognition accuracy for the speech of hearing-impaired individuals [1]. Approximately 25% of hearing-impaired individuals are said to use sign language as a means of communication in their daily lives [2], but smooth communication through sign language requires both parties to understand it. Additionally, it has been reported that more than half of hearing-impaired individuals use speech to communicate, but their speech tends to be difficult to understand, thus the use of high-accuracy ASR would be very helpful. However, existing speech recognizers are unable to achieve sufficient recognition performance.

One obstacle to achieving sufficient recognition accuracy of the speech of the hearing-impaired is the lack of a corpus of their speech data. Another is the fact that their speech has different acoustic characteristics than that of hearing individuals, in terms of articulation, prosody, and phonation, which are factors that reduce speech recognition accuracy. Research has been conducted on speech recognition of the speech of individuals with articulation disorders, which has characteristics similar to the speech of hearing-impaired individuals [3], [4]. To overcome the problem of a lack of speech data for individuals with articulation disorders, methods have been proposed to adapt standard speech recognition models to recognize such speech [5], [6]. Furthermore, methods using self-supervised models pre-trained with a large amount of unlabeled data have also been proposed [7]. Self-supervised learning (SSL) has achieved high accuracy in various tasks such as speech recognition [8], [9], speech emotion recognition [10], and speaker identification [11]. Moreover, it has been reported that speech representations generated using SSL-based speech recognition models are robust to domain mismatches [7], [12], [13]. Pasad et al. [14] have shown that the layer-wise representations of wav2vec 2.0 [8], a framework that learns speech representations from audio only, follow an acoustic-linguistic hierarchy. They have also shown that the weights of the upper layers of a pre-trained wav2vec 2.0 framework are not suitable for ASR fine-tuning and ASR fine-tuning using wav2vec 2.0 can be improved by initializing the weights of the upper layers. It has also been reported that speech features captured by wav2vec 2.0 representation, particularly the speech features of XLSR-53 [15], are effective for improving recognition of the speech of individuals with articulation disorders [7]. However, the methods proposed in these studies are mostly for acoustic domain adaptation, and linguistic information also needs to be considered to achieve high speech recognition accuracy. John et al. [16] proposed constructing an ASR model which can recognize out-of-vocabulary words in the speech of individuals with articulation disorders by converting normal speech containing unknown words into the speech of individuals with articulation disorders, and using the converted speech as training data. However, this method relies on the accuracy of the voice conversion used for acoustic domain adaptation. In our previous research [17], we aimed to construct an ASR that retains acoustic information obtained through SSL and linguistic information obtained from a general large-scale corpus, proposing a learning method that includes the replacement of encoder layers. As a result, the recognition accuracy for speech from individuals with hearing

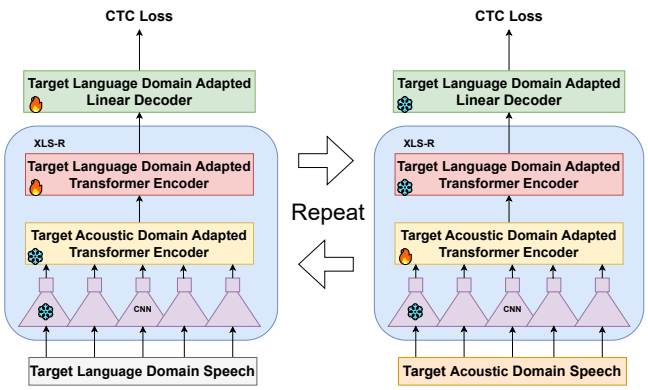


Fig. 1. Proposed alternating learning of acoustic and linguistic information.

impairments improved, and the ASR that retained SSL-based representations and linguistic information from a large-scale corpus achieved high recognition accuracy.

Therefore, in this study we also aim to construct an ASR that retains acoustic speech representations obtained through SSL, in addition to the linguistic information obtained from a general, large-scale speech corpus. Motivated by our previous research [17], we first considered replacing the encoder layers, but observed overfitting after fine-tuning the ASR with a small amount of data after layer replacement. To further improve the accuracy of recognition of speech from hearing-impaired individuals, we propose a learning method that suppresses overfitting while retaining the linguistic information obtained from a general, large-scale training corpus, and which also adapts the model to hearing-impaired speech using the available target domain speech. This process of acoustic and linguistic adaptation are repeatedly performed during model training. This method is an improvement of the learning method using the same architecture as the previous research [17]. The contributions of this paper are as follows:

- We demonstrate that the proposed method constructs a speech recognition model that retains both acoustic information from hearing-impaired speech and linguistic information from of standard speech.
- We show that a model trained using the proposed method outperforms the same model fine-tuned in a conventional manner, as well as other models with a larger number of parameters, in terms of recognition accuracy for hearing-impaired speech.

II. Proposed Method

Our previous work [17] proposed a method of replacing the encoder layers of an ASR model in order to adapt it to the processing hearing-impaired speech, using the acoustic information of speech from hearing-impaired individuals while retaining the linguistic information obtained from a general, large-scale speech corpus. In this method, fine-tuning the model with the speech of hearing-impaired individuals after replacing the encoder layers. However, we observed that fine-tuning with a small amount of speech data from hearing-impaired individuals caused overfitting of the linguistic information. To prevent this overfitting, in this paper we propose a learning method that alternately learns the acoustic information of hearing-impaired speech and the linguistic information of a general, large-scale speech corpus, after replacing the encoder layers. This method is inspired by the observation that the lower layers in the Transformer encoder of the ASR model process acoustic information while the upper layers process linguistic information. The proposed method uses as its initial state the model with replaced encoder layers described in our previous study [17], as described in Section II-B.

A. Alternating learning

A diagram of the proposed method is shown in Fig. 1. In the proposed method, Connectionist Temporal Classification (CTC) [18] fine-tuning is performed as follows:

1) When using the speech of hearing individuals as the target language domain, the lower layers of the Transformer encoder are frozen, and only the upper layers close to the output layer, and the decoder, are trained.
2) When using the speech of hearing-impaired individuals as the target acoustic domain, the upper layers of the Transformer encoder are frozen, and only the lower layers close to the input layer are trained.
3) Repeat steps 1) and 2).

Using this training method, the ASR can be domain-adapted to both the target language domain and the target acoustic domain.

B. Initialization of ASR model

We initialize the ASR model using the following procedure prior to the alternating learning, described in Section II-A. The model is initialized using the method shown in Fig. 2. This approach is based on a method proposed in our previous study [17] involving the following three steps:

1) Additional pre-training: We further pre-train an XLS-R model to adapt it to the the acoustic information of speech from hearing impaired individuals, using both large scale, standard Japanese speech data, and the speech data of hearing-impaired individuals, both of which will also be used during subsequent fine-tuning.

2) 1st fine-tuning: To learn linguistic information, we perform the first fine-tuning of the speech recognition model using a large amount of speech data from hearing individuals the target linguistic domain. We also add a single, fully-connected layer as the decoder and freeze the CNN encoder.

3) Replace layers: Assuming that the effect of acoustic domain adaptation from the additional pre-training has been forgotten due to the first fine-tuning, we construct a speech recognition model that retains both acoustic and linguistic information by replacing part of the encoder layers of the fine-tuned speech recognition model with the pre-trained XLS-R model's encoder layers. In this study, the initial state was set by replacing the lower half of the Transformer encoder layers.

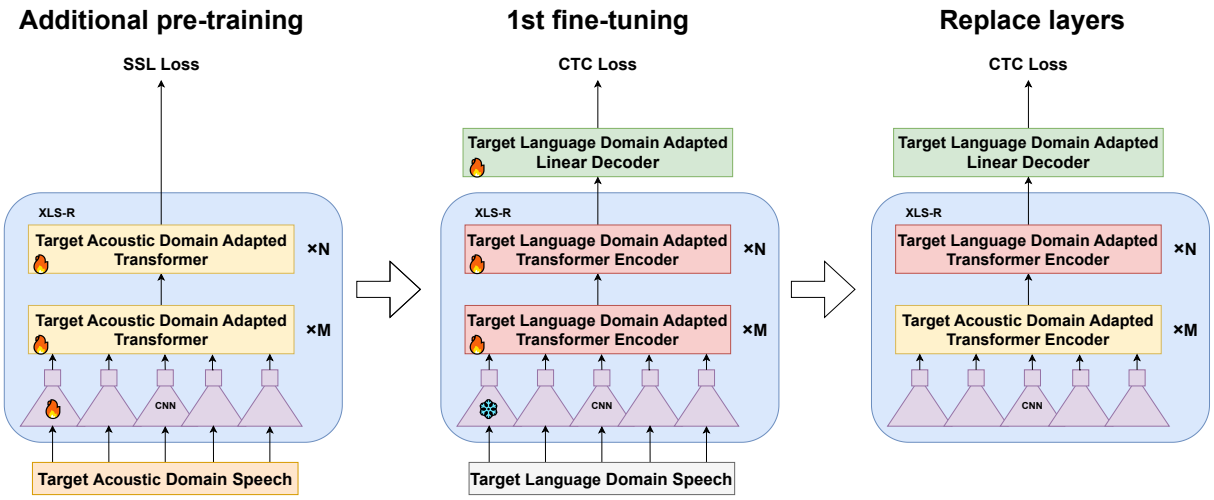


Fig. 2. Initialization of the ASR model before alternating learning. This process includes additional pre-training using SSL and Transformer, fine-tuning of the ASR model by combining the encoder and CTC decoder, and replacing some of the encoder layers.

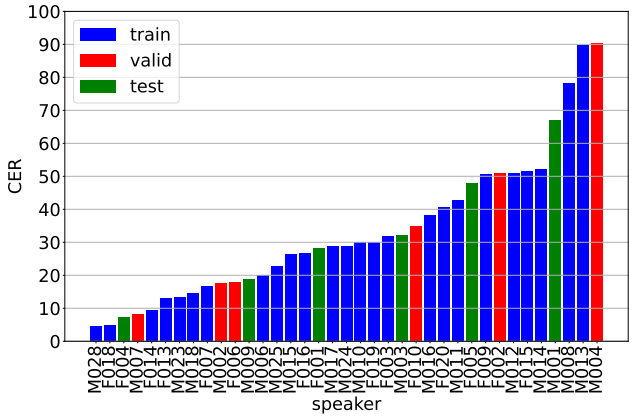


Fig. 3. Division at a splitting of DEAF corpus data into training, validation and testing sets, based on recognition results for each hearing-impaired speaker when using in an ASR model trained with standard on hearing individuals' speech.

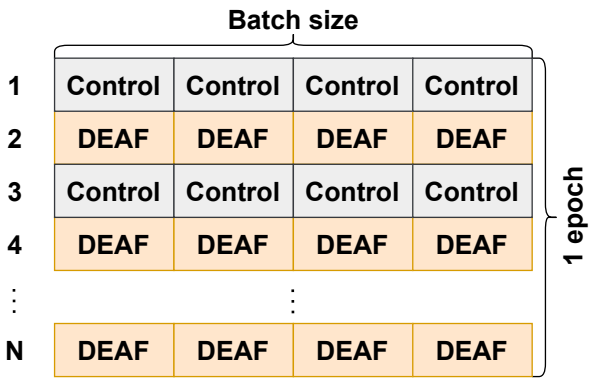


Fig. 4. Order of the training datasets during one epoch of alternating learning, where $1, 2, 3, 4, \ldots, N$ represent the iterations, "Control" and "DEAF" indicate data sampled from the corpus of standard Japanese speech and the corpus of hearing-impaired speech, respectively.

III. Experimental Setup

A. Hearing-impaired speech corpus

The corpus of speech from hearing-impaired individuals which was used in this study was recorded by Kobayashi et al. [1], and also includes additional speech data recorded subsequently by the same researchers. This corpus includes parts of the ATR phoneme-balanced 503 sentences used for the JNAS read speech corpus [19], as read by hearing-impaired individuals. Not all of the speakers read all of the selected ATR sentences, however all of the speakers read both the B and C sets of sentences. Figure 3 shows the Character Error Rate (CER) for the speech recognition of each hearing-impaired participant, sorted in ascending order, when using an ASR model trained only using speech from non-impaired individuals. Utterances with a high CER were difficult for even humans to understand as well. Based on these results, the corpus of hearing-impaired speech was proportionally divided according to the recognition difficulty of the data, into training, validation and test sets. During this division, care was taken to ensure there was no overlap between the speakers and the utterance content, and differences between the hearing-impaired speech and speech from hearing individuals were considered. In Fig. 3, the colors of the bars represent which of the datasets (training, validation or testing) a speaker's speech was assigned to. The validation set consists of the ATR 503 B set, the test set consists of the ATR 503 C set, and the training set consists of the remaining sets. The training set comprises approximately 16 hours of speech from 16 individuals (6 females and 10 males), while the validation and evaluation sets each comprise approximately 30 minutes of speech from

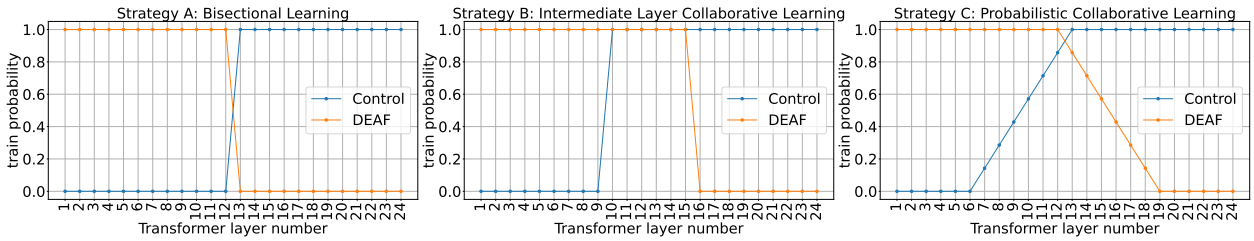


Fig. 5. The probability of learning for each layer of the Transformer encoder

3 females and 3 males. In this study, we refer to this corpus as the DEAF corpus.

B. Control speech corpus

We used speech from the JNAS corpus of read Japanese as our corpus of standard speech of hearing individuals because the DEAF corpus also consists of readings of the ATR phoneme-balanced 503 sentences included in the JNAS corpus. Therefore, as for the DEAF corpus, the validation set for the corpus of hearing individuals' speech consisted of the ATR 503 B set, the test set consisted of the ATR 503 C set, and the training set consisted of the remaining speech. As a result, the training set comprised approximately 80 hours of speech, the validation set approximately 2 hours, and the evaluation set approximately 3 hours. In addition, the large-scale, Japanese-language, Laboro TV Speech (LTV) corpus [20], which includes approximately 2,000 hours of general television speech, and 767 hours of news and report reading speech, was added to the standard/hearing speaker training set.

C. ASR Model

The XLS-R (0.3B) [21] model, which is based on SSL and includes a Transformer encoder with 24 layers, was used in this study as the ASR encoder. The XLS-R model is often used for downstream tasks after fine-tuning, including speech recognition. Performing SSL using on large-scale speech data allows high speech recognition accuracy, even when the amount of labeled data available for fine-tuning is limited. It has been reported that high recognition accuracy can also be achieved for dysarthric speech, a type of disordered speech, through the use of SSL representations [7]. During fine-tuning, a single full-connected layer was added as the decoder, and training was performed using CTC loss [18].

We set the baseline for ASR performance as the case where fine-tuning is performed on the DEAF corpus after initialization by layer replacement [17]. ASR accuracy when using Whisper Medium [22], which has approximately twice the parameters of XLS-R (0.3B), and ReazonSpeech v2.0 [23] were used for comparison. Whisper Medium is a model that, like XLS-R (0.3B), has a 24-layer Transformer encoder. ReazonSpeech v2.0 is an ASR model trained with a large-scale, Japanese speech dataset that achieves high recognition accuracy for standard speech. Each model was acoustically domain-adapted by fine-tuning only the encoder, using the DEAF corpus.

Algorithm 1 Learning Decision Based on Uniform Random Sampling and Threshold

```
Sample p ~ U(0, 1)
if p ≤ threshold then
    Train : Unfreeze layer
else
    Skip : Freeze layer
end if
```

D. Preparation of the dataset

Our proposed method uses speech from both hearing-impaired and unimpaired individuals, alternately, during model training, which requires equalizing the number of utterances from each corpus which are input during each epoch. However, the DEAF corpus has significantly fewer utterances than the Control corpus. To remove this bias, we first aligned the number of utterances in the DEAF and Control corpora during each epoch by copying the DEAF corpus. Next, we established a data sampling method to alternately input utterances from the Control corpus and the DEAF corpus into the model. The training dataset created from the Control and DEAF corpora is shown in Fig. 4. In each data batch, audio and its corresponding transcription text are sampled from the same corpus, and in the following batch, the same types of data is sampled from the other corpus. In other words, during each iteration data is sampled from the same corpus and input into the model (from either the DEAF or Control corpus), thereby performing alternating training of the model, i.e., the model is trained through alternative learning.

E. Training Strategy

In the Transformer encoder used for ASR, it is assumed that the lower layers process acoustic information, while the upper layers process linguistic information. Some evidences of this has been suggested in studies such as [14], but it is still unclear exactly how each layer processes information. Therefore, in this experiment, we designed three learning strategies and compared the resulting recognition accuracies of the three ASR models. Each of these learning strategies are illustrated in Fig. 5. These learning strategies determine the learning probability, for each layer of the Transformer encoder,

TABLE I
ASR ACCURACY WHEN USING EACH ENCODER TRAINING STRATEGY

Strategy	JNAS CER (%)	DEAF CER (%)
A	8.3	22.9
B	7.9	21.7
C	**7.7**	**21.3**

for both the Control and DEAF corpora. By comparing this learning probability to a threshold, Algorithm 1 determines probabilistically whether or not to train each layer of the Transformer encoder. The three learning strategies can be described as follow:

- **Strategy A: Bisectional Learning**
  We hypothesized that the lower layers of the Transformer encoder process acoustic information, while the upper layers process linguistic information, and divided the layers to be trained with each speech corpus accordingly. When the input is DEAF speech, the upper layers of the Transformer encoder are frozen, and when the input is Control speech, the lower layers are frozen. In other words, the lower 12 layers of the Transformer encoder are trained with the DEAF corpus, and the upper 12 layers are trained with the Control corpus.
- **Strategy B: Intermediate Layer Collaborative Learning**
  This is a modification of Strategy A, in which the intermediate layers of the Transformer are trained with both corpora. This is done because it is unclear whether the intermediate layers of the Transformer encoder process acoustic or linguistic information, and by modifying Strategy A, we can add a margin to the layers trained with each corpus. In other words, the lower layers of the Transformer encoder are trained with the DEAF corpus, the intermediate layers are trained with both the Control and DEAF corpora, and the upper layers are trained with the Control corpus.
- **Strategy C: Probabilistic Collaborative Learning**
  This is a modification of Strategy A where the intermediate layers are trained probabilistically. In Strategy B, we hypothesized that among the intermediate layers trained using both corpora, the closer a layer is to the input layer, the more likely it processes acoustic information, and the closer it is to the output layer, the more likely it processes linguistic information. Therefore, in Strategy C we chose to select which intermediate layers were to be trained with each corpus probabilistically, as an alternative to Strategy B.

IV. Experimental Results

A. Comparison of CER among training strategies

Table II shows CER with the three training strategies. Among the three Encoder training strategies, Strategy C achieved the highest recognition accuracy for both JNAS (unimpaired) and DEAF (impaired) speech. This indicates that Strategy C allowed the ASR model to learn the acoustic information from the DEAF corpus and the linguistic information from the control speech corpus more effectively than the other strategies. In other words, it is speculated that Strategy C may have matched the selection of layers for training in each corpus with the layers that process linguistic and acoustic information better than other strategies. Therefore this training strategy was used in our proposed method. Additionally, the use of Strategy B resulted in higher recognition accuracy than Strategy A for both JNAS and DEAF speech. These results suggest that having the Encoder layers trained with both corpora leads to better learning outcomes.

B. Comparison with other models

A comparison of speech recognition accuracy in terms of CER for both unimpaired and hearing-impaired speech, between the model trained using the proposed method, the same model trained using other methods, and other models with more parameters, is shown in Table II. The model trained using the proposed method achieved the lowest CERs for speech from either the JNAS or the DEAF corpora, compared to the same model trained with other methods and with the other models. Note however that the CER for JNAS speech for the XLS-R model trained with the JNAS and LTV datasets was 7.7%, which is the same as the model trained using the proposed method. This suggests that the proposed method was able to acquire the acoustic information of the DEAF corpus while suppressing forgetting of the linguistic information from the JNAS and LTV corpora during acoustic training. These results demonstrate that a speech recognition model constructed using the proposed method can retain both specialized acoustic information as well as standard linguistic information.

V. Conclusions

In this study, we proposed an alternating training method for ASR models that retains linguistic information obtained from a general, large-scale speech corpus while also being adapting to recognize the acoustic information of hearing-impaired speech, allowing accurate speech recognition of either type of speech. Our experimental evaluation confirmed that our proposed method achieved higher or equivalent speech recognition performance for standard Japanese speech, and higher recognition performance for hearing-impaired speech, than the other methods. Additionally, we tested three Encoder layer training strategies to investigate their effects on speech recognition accuracy, and found that training the intermediate layers probabilistically, using data from one or the other of the two corpora, resulted in the highest ASR accuracy.

In the future, we plan to confirm the dependence of corpus through additional experiments with other corpora since this study was conducted using a limited and specific corpus. Also, we optimize the layers to be trained to improve recognition accuracy further.

TABLE II
COMPARISON OF ASR PERFORMANCE WHEN USING EACH METHOD USING CHARACTER ERROR RATE (%), WITH JNAS (STANDARD) AND DEAF (HEARING-IMPAIRED) JAPANESE SPEECH ("$\rightleftarrows$" AND "+" INDICATE ALTERNATING TRAINING AND CONCATENATION OF THE DATASETS, RESPECTIVELY).

method	model	# Parameters (MB)	1st fine-tune	2nd fine-tune	JNAS CER (%)	DEAF CER (%)
Baseline	ReazonSpeech v2.0	619	DEAF	N/A	10.7	26.0
	Whisper Medium	769	DEAF	N/A	9.0	25.1
	XLS-R	319	JNAS + LTV	N/A	**7.7**	39.5
	XLS-R	319	JNAS + LTV	DEAF	8.3	23.0
	XLS-R w/ Replacement [17]	319	JNAS + LTV	DEAF	9.3	22.1
Proposed	XLS-R w/ Alternating learning	319	JNAS + LTV	JNAS + LTV $\rightleftarrows$ DEAF	**7.7**	**21.3**

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP23H00995.

REFERENCES

[1] A. Kobayashi, K. Yasu, H. Nishizaki, and N. Kitaoka, "Corpus Design and Automatic Speech Recognition for Deaf and Hard-of-Hearing People," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021, pp. 17–18. DOI: 10.1109/GCCE53005.2021.9621959.

[2] Japanese Ministry of Health, Labour and Welfare, *Survey on Difficulties in Living*. 2018. [Online]. Available: https://www.mhlw.go.jp/toukei/list/dl/seikatsu_chousa_c_h28.pdf.

[3] K. M. Baskar, T. Herzig, D. Nguyen, *et al.*, "Speaker adaptation for Wav2vec2 based dysarthric ASR," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 9, Incheon, KR: International Speech Communication Association, 2022, pp. 3403–3407. DOI: 10.21437/Interspeech.2022-10896. [Online]. Available: https://www.fit.vut.cz/research/publication/12854.

[4] S. Hu, X. Xie, Z. Jin, *et al.*, "Exploring Self-Supervised Pre-Trained ASR Models for Dysarthric and Elderly Speech Recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10097275.

[5] J. Shor, D. Emanuel, O. Lang, *et al.*, "Personalizing ASR for Dysarthric and Accented Speech with Limited Data," in *Interspeech 2019*, ISCA, Sep. 2019. DOI: 10.21437/interspeech.2019-1427. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1427.

[6] R. Takashima, T. Takiguchi, and Y. Ariki, "Two-Step Acoustic Model Adaptation for Dysarthric Speech Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6104–6108. DOI: 10.1109/ICASSP40776.2020.9053725.

[7] A. Hernandez, P. A. Pérez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, *"Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition"*, 2022. arXiv: 2204.01670 [cs.CL].

[8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, *"wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations"*, 2020. arXiv: 2006.11477 [cs.CL].

[9] S. Chen, C. Wang, Z. Chen, *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022, ISSN: 1941-0484. DOI: 10.1109/jstsp.2022.3188113. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2022.3188113.

[10] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404. DOI: 10.21437/Interspeech.2021-703.

[11] N. Vaessen and D. A. Van Leeuwen, "Fine-Tuning Wav2Vec2 for Speaker Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7967–7971. DOI: 10.1109/ICASSP43922.2022.9746952.

[12] W.-N. Hsu, A. Sriram, A. Baevski, *et al.*, *"Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training"*, 2021. arXiv: 2104.01027 [cs.SD].

[13] J. Zhao, G. Shi, G.-B. Wang, and W.-Q. Zhang, "Automatic Speech Recognition for Low-Resource Languages: The Thuee Systems for the IARPA Openasr20 Evaluation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 335–341. DOI: 10.1109/ASRU51503.2021.9688260.

[14] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-Wise Analysis of a Self-Supervised Speech Representation Model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 914–921. DOI: 10.1109/ASRU51503.2021.9688093.

[15] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, *"Unsupervised Cross-lingual Represen-*

tation Learning for Speech Recognition", 2020. arXiv: 2006.13979 [cs.CL].

[16] J. Harvill, D. Issa, M. Hasegawa-Johnson, and C. Yoo, "Synthesis of New Words for Improved Dysarthric Speech Recognition on an Expanded Vocabulary," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6428–6432. DOI: 10.1109/ICASSP39728.2021.9414869.

[17] K. Takahashi, Y. Wakabayashi, K. Ohta, A. Kobayashi, and N. Kitaoka, "Improving Speech Recognition for Japanese Deaf and Hard-of-Hearing People by Replacing Encoder Layers," in *ICAICTA 2024*, 2024.

[18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks," vol. 2006, Jan. 2006, pp. 369–376. DOI: 10.1145/1143844.1143891.

[19] T. A. S. of Japan, *"ASJ Japanese Newspaper Article Sentences Read Speech Corpus (JNAS)"*.

[20] S. Ando and H. Fujihara, *"Construction of a Large-scale Japanese ASR Corpus on TV Recordings"*, 2021. arXiv: 2103.14736 [cs.SD].

[21] A. Babu, C. Wang, A. Tjandra, *et al.*, *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*, 2021. arXiv: 2111.09296 [cs.CL].

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *"Robust Speech Recognition via Large-Scale Weak Supervision"*, 2022. arXiv: 2212.04356 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2212.04356.

[23] Y. Yin, D. Mori, and S. Fujimoto, "ReazonSpeech: A Free and Massive Corpus for Japanese ASR," *Reazon Holdings, Inc., Clear Code, Inc.*, Feb. 2024. [Online]. Available: https://research.reazon.jp/blog/2024-02-14-ReazonSpeech.html.

YOLO for High Resolution Images without Retraining

Daisuke MINAMI and Kiyoshi NISHIKAWA
Dept. of EECS, Tokyo Metropolitan University
E-mail: minami-daisuke@ed.tmu.ac.jp, kiyoshi@tmu.ac.jp

***Abstract*—In this paper, we consider the application of YOLO (You Only Look Once) to high-resolution images, i.e., so-called 4K, or 8K images. When those large images are applied to YOLO, in general, the original image will be resized before being processed by YOLO. In this case, the detection accuracy varies according to the size of the resized image. As an example, when a 4K image is resized to 320×320 pixels, YOLO would fail to detect objects that make up less than about 0.7% to 2% of the entire image. Or, it would fail to detect large objects that occupy about 2% to 10%, or more of the entire image, when 4K image is used without resizing. These values are confirmed in this paper through our experiments. Then, we propose a novel YOLO construction method that enables us to expand the area of the detectable objects when applied to high-resolution images. The advantage of the proposed method is that it does not require the retraining using high-resolution images. Instead, the proposed method uses images of multiple resolutions and combines the detection results. The effectiveness of the proposed method is confirmed by the experiments.**

I. Introduction

In this paper, we consider detectable size of objects in YOLO (You Only Look Once) and propose a novel configuration of YOLO for expanding the size for improving the detection accuracy. When YOLO is used for detecting objects from high-resolution images, such as so called 4K (images with approximately 4000 pixels on one side) or 8K (images with approximately 8000 pixels on one side) ones, the effectiveness of the proposed method increases. The advantage of the proposed method is that it achieves the expansion based on the existing model and the improvement can be achieved without retraining of the model.

YOLO is one of the standard object detection models based on CNN (convolutional neural network). Object detection technology based on CNN has been utilized in a wide variety of applications including autonomous driving systems [1], robot vision [2], etc. Various models for object detection have been proposed so far, and they can be categorized into two types, i.e., two-stage and one-stage models, based on the network architectures. The two-stage type performs region proposal and image classification in separate processes, while the one-stage type performs these two processes simultaneously. R-CNN [3], Faster R-CNN [4] are typical two-stage type detection models, and, on the other hand, YOLO [5] and SSD (Single Shot MultiBox Detector) [6] are well known one-stage type models. Of those models, YOLO is known for its high speed processing and detection accuracy, and, currently, it is regarded as one of the standards of detection models.

Most of conventional studies on object detection have focused on improving detection accuracy and increasing the processing speed. Thanks to those researches, real-time detection from video sequences, or processing on generic PC without GPUs have been realized and utilized in a wide variety of applications. In particular, YOLO has seen active research, with many researchers proposing improved versions, making it a representative model that enables fast and accurate object detection.

In contrast, this study aims to improve object detection from a different perspective than those existing studies. Namely, we examine the size of objects that can be detected using YOLO. We consider the relation between the number of convolutional layers in standard YOLO configuration and the range of the size of detectable objects. It is shown that we can select the optimum number of convolutional layers according to the size of each object. Based on the consideration, we propose the novel structure of YOLO which uses multiple images which are generated by resizing with different ratio the original image for detection. The effectiveness of the proposed configuration increases when we use the high-resolution images, such as 4K or 8K, as the input to YOLO. Conversely, less improvement would be obtained by using the proposed method to the standard resolution images, such as about 1000×1000 pixels in size. In this case, almost all of the objects in the image are in the detectable range of the standard YOLO.

The motivation of this study arises from the recent proliferation of digital images, where the performance of imaging devices such as cameras has improved, making it easy to obtain high-resolution images. Hence, this paper focuses on the object detection from those high-resolution images where importance of consideration on the detectable range of sizes of objects is high, and, based on the consideration, we propose modification of the structure of YOLO to extend the number of detectable objects.

II. Background

Here, we describe the object detection using YOLO version 7, or YOLOv7, which is the 7th version of YOLO. Note that, in this paper, we mainly use the version 7 of YOLO, and hence, in the following, we use the term 'YOLO' to indicate its version 7.

A. *Effects of the size of the image on the detectable objects*

When a high-resolution image is applied to YOLO, it is observed that the detection accuracy for specific objects varies depending on the size of the image processed by YOLO. We show examples of the characteristics in Fig. 1.

The difference in the conditions to obtain these figures is in the size of the input image when applied to YOLO. The size of the original image is 4000×4000 pixels, and we applied the image to YOLO after resizing it. For left of Fig. 1, the image is resized to 320×320 pixels, and for right Fig. 1 to 4000×4000 pixels.

Other conditions are same for both figures. Namely, we used the pre-trained model of YOLO obtained from [7]. The detection thresholds were set with a classification confidence of 0.6 or higher. The detected objects are enclosed in red rectangular regions.

Let us examine the effect of the differences in size of the input image for YOLO. Focusing on the person at the front, it can be seen that the person is detected when resized to 320×320 pixels, but not when resized to 4000×4000 pixels. On the other hand, focusing on the motorcycles and people enclosed in the orange box at the back right, it can be observed that three motorcycles and three people are detected when resized to 4000×4000 pixels, whereas only one motorcycle and one person are detected when resized to 320×320 pixels.

Thus, when we use a high-resolution image, it is confirmed that the detectable objects using YOLO varies according to size of the input image. More precisely, the range of the size of an object that can be detected is determined by the size of the input image which is obtained by resizing the original image.

This characteristic may not be considered as a serious problem when the size of the original image is around 1000×1000 pixels. In contrast, as the size of the original image become larger, such as 4K or 8K, it would be the characteristic to be considered. Because, in that case, the detection accuracy depends on the choice of size of the input image which is obtained by resizing the original image. In YOLO, a user can specify the size of the image for processing independently from the size of the original image, and the original image is resized accordingly based on the specification before detection. When the original image is high-resolution one, as shown in the above examples, the detectable objects may vary depending on the selected size for processing. Hence, when considering object detection from high-resolution images, setting the appropriate size for processing becomes crucial because they may contain objects of a wider range of sizes in the image.

B. *Object detection from a image of single resolution*

Next, we confirm that there are relation between the number of convolutional layers applied and the size of objects, that can be detected by extracting features through these convolutions.

YOLO features an architecture where convolution is performed with channel partitioning known as ELAN[8], and upsampling is conducted midway through convolution. Because

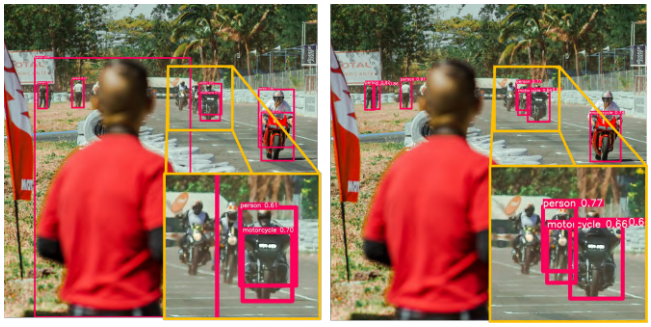


Fig. 1: Detection results using YOLO with input image sizes of 320×320 pixels (left) and 4000×4000 pixels (right).

of this processing, there is no straightforward expression to show the relationship between the number of convolutional layers and the detectable sizes of objects in YOLO. Instead, we show the relationship using the simpler structure of SSD300 in Table I. The relation was shown in [9]

We can say, as expected, for detecting the small objects, we need fewer convolution layers and conversely, more convolution layers are needed for large objects. These simple relations will be emphasized when a high-resolution image is applied to YOLO. Because YOLO has specific anchor box (AB) sizes prepared for each feature map used, it can only detect objects whose features match the size of the AB extracted by the specific feature map theoretically. Hence, our purpose in this paper is to propose a novel structure of YOLO that enables to detect both large and small objects from high-resolution images using existing YOLO model without retraining.

For the objects whose size is smaller than the specific limit, YOLO cannot extract enough features for detection. Conversely, the object whose size is larger than the specific limit, the number of convolutional layers used in YOLO is not enough for extracting features for those objects. These problems are not a problem for normal-sized images, but they are noticeable when high-resolution images are used as input.

Addition of convolutional layers may improve the detection ratio for large objects, however, this approach requires a significant amount of computational resources, and re-training of the network with the high-resolution images. More precise difinition of the size of the objects is given in the following consideration.

TABLE I: The relationship between the number of convolutional layers and the detectable sizes of objects in SSD300.

The number of convolutions	10	15	17	19	21	23
Min size	30	60	111	162	213	264
Max size	60	111	162	213	264	315

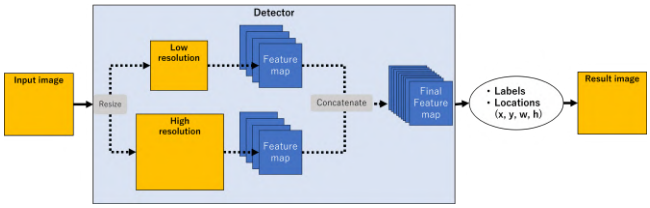


Fig. 2: The configuration of the conventional method.

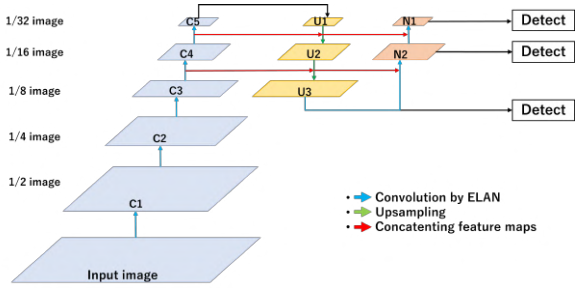


Fig. 3: Configuration of the standard YOLO.

III. Conventional Methods

A. Detection from Multi-Resolution Images

As described in the next section, the proposed method uses multi-resolution images, namely, two images with different resolutions as the inputs to the detection model. Here, we note that there are similar conventional methods which use multi-resolution images for object detection [10], [11].

In Fig. 2, we show a configuration of the method that involves resizing the image to multiple resolutions and performing convolutions in parallel. Features are extracted from multiple resolutions, and the extracted features are then combined for detection. This allows for detection across multiple resolutions[10], [11].

However, in contrast to the proposed method, it is difficult to apply them to the existing pre-trained object detection models such as YOLO or SSD. When we consider applying the methods to those pre-trained models, it is necessary to add a new layer to combine the feature maps and retrain the model. In doing so, the model becomes more complex, making optimization difficult, and training data of different resolution are also required.

B. Detection from High-Resolution Images Using SSD300

The authors[9] proposed a detection model based on SSD300, so called because it resizes the input image to 300×300 pixels, for high-resolution images without retraining. Specifically, the method enables detection of small objects (those occupying less than 1% of the image area) that were previously undetectable. The proposed method is an extension of the basic idea of [9] to YOLO for expanding the detectable range of the sizes of objects.

IV. Proposed Method

Here, we describe the proposed method. One of the advantages of the method is that it could achieve the improvement of detection accuracy using the standard YOLO model without retraining. The proposed method enables the improvement of accuracy of the existing pre-trained models by merging detection results from multiple images with different resolutions.

A. Preparation

Since the proposed method uses the images of several resolutions, we first define the "size of objects" as the proportion of the object area in the original image before resizing. Because the area proportion depends on the size of the original image, this study evaluates the conventional and the proposed methods under the assumption that the size of the input image is fixed as 4000×4000 pixels for consistency. We note that this assumption is used only in the analysis, and there is no restriction on the size of the image when the proposed method is used in actual applications.

B. Configuration of YOLO

In Fig. 3 we show the configuration of standard YOLO version 7 used in this paper. During detection, YOLO performs convolution and upsampling to the input image to reduce or enlarge the size of the image. It then detects objects based on the multiple feature maps obtained by this process. YOLO uses feature maps with resolutions of 1/8, 1/16, and 1/32 of the size of the input for detection. To extract feature maps at each resolution, YOLO employs an ELAN structure which enables fast processing by distributing the number of channels during convolution.

Note that there are several models of YOLO version 7. In the following, we also use YOLOv7-e6e (yolov7-e6e.pt on [7]), the largest model of YOLO. YOLOv7-e6e uses feature maps with resolutions of 1/8, 1/16, 1/32, and 1/64 for detection. Additionally, YOLOv7-e6e employs an E-ELAN [12] structure to extract feature maps at each resolution.

C. Relation between anchor boxes and detectable range

As shown in Section II-B, the detection accuracy for specific objects varies depending on the size of the input image. Because, YOLO predetermines the sizes of the anchor boxes (ABs), which are rectangular regions used to determine the bounding regions of objects during detection.

Table II shows the sizes of ABs used in YOLO. Three types of ABs are used for each feature map as shown in Section IV-B. After resizing the image, the size of objects that can be detected corresponds to the size of these ABs. However, in YOLO, the optimal ABs are used based on the training data, so the ideal detectable object sizes may differ from the actual detectable object sizes.

D. Relation between detectable range and the size of input image in YOLO

We experimentally investigated the detectable range of YOLO for each resolution. The resolutions used must be divisible by the value of the max stride used in YOLO. Namely,

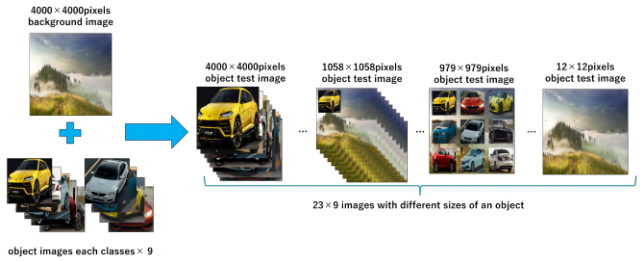


Fig. 4: Test images for investigating the detectable range of YOLO.

YOLO uses 32 as the max stride, and in YOLOv7-e6e uses 64. Therefore, in this study, we investigate the resolutions of 320, 960, 1600, 3200, and 4000 (4032).

Test images of four classes (Car, Cat, Horse, Sheep) are artificially generated as shown in Fig. 4, and, we evaluated whether each size of the object is detectable or not. The sizes of objects in the test images are ranging from 0.001% to 100% of the original image area.

Here, we regard the object as detectable if the following three conditions satisfy: (i) the predicted object class matches the ground truth label, (ii) the classification confidence is above 0.6, (iii) the IoU is above 0.75. Then, we calculate Precision and Recall, and the F-measure. If the F-measure is above 0.7, we regard that the object size is considered as detectable. Otherwise, it is classified as undetected.

The average values of these metrics for the four classes define the detectable range, as shown in Tables III and IV. In the tables, "Object size (%)" represents the percentage of the object's area in a 4000×4000 pixels background image, while "Object size (px)" denotes the actual number of pixels occupied by the object. Additionally, "Object size (After resize)" indicates the size of each object in pixels after the input image is resized to different resolutions, reflecting the sizes of ABs used by the pre-trained model. We used sizes 4000, 3200, 1600, and 960 for YOLO when calculating the average maximum value of Object size (After resizing). For YOLOv7-e6e, we used the sizes 4000, 3200, and 1600. Because, when considering that Object size (After resize) is the size of AB, unused sizes are presumed to be smaller than the maximum value of the size of AB.

Using these values, it is theoretically possible to determine the relation between the detectable object size and the resolution, and they are shown in equation (1). The equation indicates that within the original size of the image, the detectable object size becomes AB size after resizing to the Resize size.

TABLE II: Sizes of anchor boxes in YOLO.

Scale of feature maps	YOLO	YOLOv7-e6e
1/8	12×16, 19×36, 40×28	19×27, 44×40, 38×94
1/16	36×75, 76×55, 72×146	96×68, 86×152, 180×137
1/32	142×110, 192×243, 459×401	140×301, 303×264, 238×542
1/64		436×615, 739×380, 925×792

$$\text{Detectable Size} = \frac{\text{Original Size} \times \text{AB Size}}{\text{Resize Size}} \quad (1)$$

E. The Flow of The Proposed Method

From the results shown in IV-D, it is seen that by using resolutions of 320×320 pixels and 4000×4000 pixels (4032×4032 pixels for YOLOv7-e6e), we can theoretically achieves the maximum detectable range.

Based on this, we propose a novel configuration of YOLO for high resolution images by mixing the results of detection using the two images of the resolutions above. Flow of the proposed method is shown in Fig. 5 using YOLO. Here is an example using 320×320 and 4000×4000 resolutions for the proposed method, but we can actually use any combination of multiple resolutions for the proposed method. First, the input high-resolution image is resized to 320×320 pixels and to 4000×4000 pixels. Then, detection is performed on each of the two. According to Table III, the detector resized to 320×320 pixels is expected to detect objects ranging in size from 0.695% to 100%. Similarly, the other detector resized to 4000×4000 expected to detect objects ranging in size from 0.01375% to 2%. Afterward, the detection results from each resolution, specifically the five values representing the detected object's class name and bounding box information (class_name, x_center, y_center, width, height), are integrated. This approach is expected to detect objects ranging in size from 0.01375% to 100%.

V. Evaluation by Simulations

We conducted simulations to evaluate the effectiveness of the proposed method. Ten natural images are used as the original images, and they are 4000×4000 pixels in size and contain objects of various sizes. We compared the results on low-resolution images (320×320 pixels) with those on high-resolution images (4000×4000 pixels) and with the proposed method, and evaluated them.

TABLE III: Detectable range when YOLO is used.

Resize size (px)	Object size (%)	Object size (px)	Object size (After resize)
4000	0.01375 to 2	47 to 566	47 to 566
3200	0.01425 to 4	48 to 800	38 to 640
1600	0.375 to 17.5	77 to 1673	31 to 669
960	0.085 to 50	117 to 2828	28 to 679
320	0.695 to 100	333 to 4000	27 to 320
Average			**34.2 to 638.5**

TABLE IV: Detectable range when YOLOv7-e6e is used.

Resize size (px)	Object size (%)	Object size (px)	Object size (After resize)
4000	0.01275 to 10.75	45 to 1311	46 to 1322
3200	0.014 to 16.75	47 to 1637	38 to 1310
1600	0.06 to 67.75	98 to 3292	39 to 1317
960	0.1625 to 100	161 to 4000	39 to 960
320	1.935 to 100	556 to 4000	45 to 320
Average			**41.4 to 1316.3**

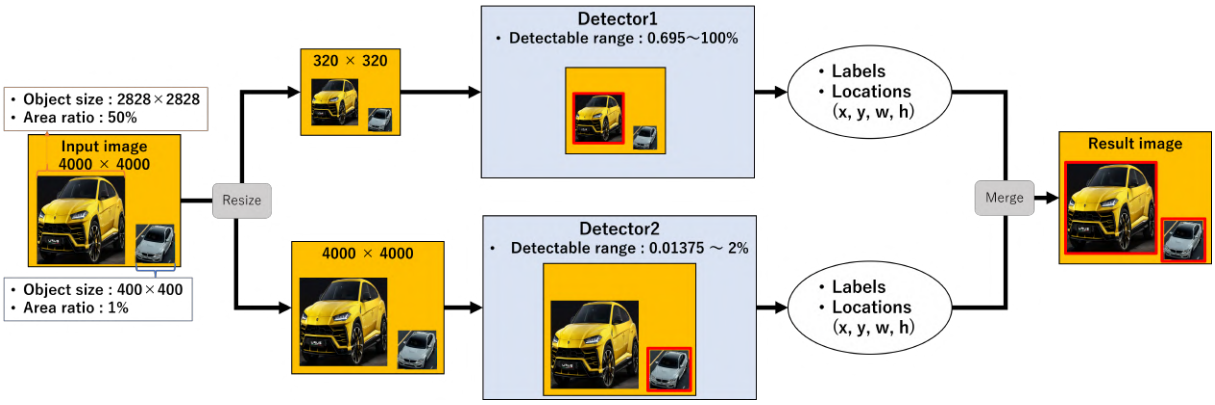


Fig. 5: Flow of the proposed method.

In order to evaluate the proposed method in a straightforward manner, we define evaluation criteria in terms of the size of objects in the image. In this study, since the resolution of the input image is varied when performing the detection, we express the 'object size' as a ratio (%) of the area of the object in the original image, not in pixels.

The MS COCO [13] object size metrics are used to define objects. Three types of MS COCO metrics are available ('small', 'medium', and 'large'), but they depend on the size of the original image because they are defined by pixel size. Therefore, we define the ratio of the area of the object based on the geometric mean of width and height in coco test2017 (525 pixels). In order to evaluate the detection for high-resolution images, we also define in more detail the ratio of the area of objects smaller than 'small' objects ('micro', 'very tiny', 'tiny') in Table V[14], [15].

In the proposed method, detections were performed using resolutions of 320×320 pixels, 1600×1600 pixels, and 4000×4000 pixels. Examples of the five detection results of the proposed method with each model are shown in Fig. 6. Additionally, Tables VI and VII show the total number of detections and the percentage (%) of that across ten natural images using low-resolution, high-resolution images and the proposed method with YOLO and YOLOv7-e6e.

We defined "detected" here using the ground truth we annotated. We evaluated the number of detections where the ground truth and the class of the detected object matched, with a confidence level greater than 0.6 and IoU greater than 0.5 and 0.75, respectively.

As mentioned in Section IV-D, it is clear that the size of detectable objects differs between low-resolution and high-resolution images. For YOLO, it is observed that the detection rate of objects larger than "very tiny" is higher for low-resolution images, whereas the detection rate of objects smaller than "micro" is higher for high-resolution images. Additionally, for YOLOv7-e6e, the detection rate of objects larger than "very tiny" is higher for low-resolution images, while the detection rate of objects smaller than "small" is higher for high-resolution images.

In addition, the total detection rate from a single resolution is 59% to 65% at best, but in both trained models, the proposed method is able to detect objects of any size evenly, improving the detection rate to over 70% to 80%.

TABLE V: The definition of object size.

Size name	Object size(%)
micro	0 to 0.38
very tiny	0.38 to 1.52
tiny	1.52 to 3.05
small	3.05 to 6.10
medium	6.10 to 18.29
large	18.29 to 100

TABLE VI: The number of detections by YOLO. Each categories of detection results are determined by each area ratios of the detection results.

Resize size	micro(25)	very tiny(17)	tiny(14)	small(13)	medium(8)	large(11)	total(88)
320 (IoU≥0.5)	5 (20%)	13 (76%)	11 (79%)	10 (77%)	7 (88%)	11 (100%)	57 (65%)
320 (IoU≥0.75)	4 (16%)	11 (65%)	11 (79%)	9 (69%)	7 (88%)	10 (91%)	52 (59%)
4000 (IoU≥0.5)	14 (56%)	10 (59%)	2 (14%)	0 (0%)	0 (0%)	0 (0%)	26 (30%)
4000 (IoU≥0.75)	13 (52%)	6 (35%)	2 (14%)	0 (0%)	0 (0%)	0 (0%)	21 (24%)
Proposed method (IoU≥0.5)	17 (68%)	18 (106%)	11 (79%)	13 (100%)	7 (88%)	11 (100%)	77 (88%)
Proposed method (IoU≥0.75)	16 (64%)	12 (71%)	11 (79%)	11 (85%)	7 (88%)	10 (91%)	67 (76%)

TABLE VII: The number of detections by YOLOv7-e6e. Each categories of detection results are determined by each area ratios of the detection results.

Resize size	micro(25)	very tiny(17)	tiny(14)	small(13)	medium(8)	large(11)	Total(88)
320 (IoU≥0.5)	3 (12%)	11 (65%)	7 (50%)	11 (85%)	8 (100%)	11 (100%)	51 (58%)
320 (IoU≥0.75)	2 (8%)	8 (47%)	7 (50%)	11 (85%)	8 (100%)	10 (91%)	46 (52%)
4032 (IoU≥0.5)	16 (64%)	20 (118%)	9 (64%)	12 (92%)	0 (0%)	0 (0%)	57 (65%)
4032 (IoU≥0.75)	14 (56%)	15 (88%)	9 (64%)	11 (85%)	0 (0%)	0 (0%)	49 (56%)
Proposed method (IoU≥0.5)	17 (68%)	18 (106%)	11 (79%)	12 (92%)	8 (100%)	11 (100%)	77 (88%)
Proposed method (IoU≥0.75)	16 (64%)	12 (71%)	11 (79%)	11 (85%)	8 (100%)	10 (91%)	68 (77%)

VI. Conclusions

In this paper, we proposed a method to expand the sizes of detectable objects using YOLOv7 on high-resolution images without retraining. In the proposed method, the original image, from which the objects are detected, is resized to different

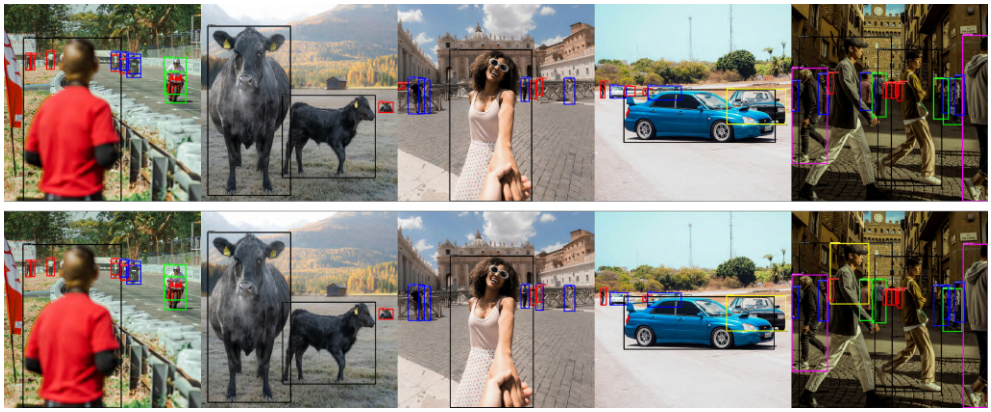


Fig. 6: Ten images processed by YOLO and YOLOv7-e6e using the proposed method. The top figure shows the objects detected by YOLO, while the bottom figure shows the objects detected by YOLOv7-e6e. The colors of the bounding boxes (BB) in the images are changed according to the sizes of the detected objects: (micro: red, very tiny: blue, tiny: green, small: yellow, medium: pink, large: black).

resolutions, and the detection results from multiple resolutions are merged to increase the number of detectable objects.

The effectiveness of the proposed method is demonstrated through simulations by comparing the results with those obtained using standard YOLO with a single resolution. We confirmed that the method improved the detection rate from 70% to over 80% and that objects of various sizes could be detected as expected.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22K12170.

References

[1] A. Balasubramaniam and S. Pasricha, *Object detection in autonomous vehicles: Status and open challenges*, 2022. arXiv: 2201.07706.

[2] S. Hoshino and K. Niimura, "Robot vision system for human detection and action recognition," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 24, no. 3, pp. 346–356, 2020.

[3] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[4] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015.

[5] R. Joseph, D. Santosh, G. Ross, and F. Ali, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[6] L. W. et al, "SSD: single shot multibox detector," in *Computer Vision – ECCV 2016*, L. Bastian, M. Jiri, S. Nicu, and W. Max, Eds., 2016, pp. 21–37.

[7] *GITHUB: wongkinyiu/yolov7*. [Online]. Available: https://github.com/github-linguist/linguist.

[8] C.-Y. Wang, H.-Y. M. Liao, and I.-H. Yeh, *Designing network design strategies through gradient path analysis*, 2022. arXiv: 2211.04800.

[9] I. Kei and N. Kiyoshi, "Detection method from 4k images using ssd300 without retraining," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 877–883.

[10] L. Guosheng, S. Chunhua, van den Hengel Anton, and R. Ian, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[11] F. Clement, C. Camille, N. Laurent, and L. Yann, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[12] W. Chien-Yao, B. Alexey, and L. H.-Y. Mark, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 7464–7475.

[13] L. T.-Y. et al, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, F. David, P. Tomas, S. Bernt, and T. Tinne, Eds., 2014, pp. 740–755.

[14] K. Aleksandra, M. Karol, and B. Dominik, "Where to look for tiny objects? roi prediction for tiny object detection in high resolution images," in *2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2022, pp. 721–726.

[15] W. Jinwang, Y. Wen, G. Haowen, Z. Ruixiang, and X. Gui-Song, "Tiny object detection in aerial images," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 3791–3798.

Anomalous Machine Sound Detection Based on Time Domain Gammatone Spectrogram Feature and IDNN Model

Primanda Adyatma Hafiz[1], Candy Olivia Mawalim[2], Dessi Puji Lestari[1], Sakriani Sakti[3], and Masashi Unoki[2]

[1] Bandung Institute of Technology, Bandung, Indonesia
E-mail: {13520022@std, dessipuji@staff}.stei.itb.ac.id

[2] Japan Advanced Institute of Science and Technology, Nomi, Japan
E-mail: {candylim, unoki}@jaist.ac.jp

[3] Nara Institute of Science and Technology, Nara, Japan
E-mail: ssakti@is.naist.jp

***Abstract*—Anomalous sound detection (ASD) systems distinguish normal and abnormal machinery conditions on the basis of sound. While most ASD systems rely on the log Mel spectrogram, it lacks sufficient frequency resolution and performs suboptimally for rapidly changing sounds. Alternatively, the Gammatone spectrogram, extracted using a time domain Gammatone filterbank, offers enhanced spectrogram resolution. This study investigates the effectiveness of the time domain Gammatone spectrogram for ASD. To optimally learn the time domain Gammatone spectrogram features, an Interpolation Deep Neural Network (IDNN) model was proposed as the detection model. This model detects nonstationary sound frames highly reliably. An evaluation was conducted using MIMII dataset with area under receiver operating characteristic curve (ROC AUC) as the metric. Experimental results showed that our proposed method achieved ROC AUC of 92.5%, outperforming the log Mel spectrogram feature by 5.9 percentage points.**

***Index Terms*—Anomalous sound detection, time domain Gammatone filterbank, IDNN, spectrogram**

I. Introduction

In anomalous sound detection (ASD), unsupervised learning methods are often used to assess machine conditions. With this approach, a model can distinguish a machine's typical patterns using solely data of machine sound's in normal conditions for training [1]. Consequently, detecting anomalies in machines does not necessarily require data of machine sounds in anomalous or defective conditions, which is difficult to obtain [2].

Most studies on ASD utilize deterministic acoustic features, primarily spectrogram-based features [3] due to their rich representations of time, frequency, and intensity, which make them ideal for neural network-based models [4]. The log Mel spectrogram, which is extracted in the frequency domain, has been used in various studies on ASD and is mainly modeled using various neural network-based models [5]. However, the log Mel spectrogram has limitations in classifying machine conditions for some machines [6]. Additionally, using short-term Fourier transform (STFT) to extract spectrogram in the frequency domain results in suboptimal resolution due to the inherent weaknesses of the STFT method [7][8].

Meanwhile, the Gammatone spectrogram, another variation of the spectrogram, closely mimics the human auditory system's stimulus response, with its parameters also derived from psychoacoustic experiments [9]. Therefore, the Gammatone spectrogram better simulates a human experts' ability to distinguish machine sounds. While most spectrograms are frequency domain-based [10], the Gammatone spectrogram can be directly calculated from the time domain representation. This is achieved by using a time domain Gammatone filterbank to transform the raw waveform directly [11].

Modeling spectrograms requires a model capable of processing 2-dimensional data. Employing a standard autoencoder (AE) model alone often yields suboptimal results due to challenges in accurately reconstructing nonstationary frames [6]. This issue can be addressed by utilizing an Interpolation Deep Neural Network (IDNN) model, which predicts only the central frame rather than all frames. By predicting only the central frame, IDNN has been proven to work better especially for modelling nonstationary sound [12].

Another study reported a UNet model can outperform the IDNN model in terms of ASD for some machines [13]. UNet architecture has the potential to be used in IDNN since it has some advantages over the IDNN model that can potentially result in a complementary effect.

This paper aims to propose a novel unsupervised ASD method employing a time domain Gammatone spectrogram. The IDNN model, enhanced by an overlapping frame technique, is applied to enhance model performance. Additionally, the IDNN model is also tested with both AE and UNet architectures to determine the most optimal design. The study findings offer insights into the effectiveness of various spectrograms, particularly when modeled using the IDNN approach for ASD. This paper is part of the ASEAN IVO 2023 project, "Spoof Detection for Automatic Speaker Verification," which aims to enhance the identification of irregularities in audio

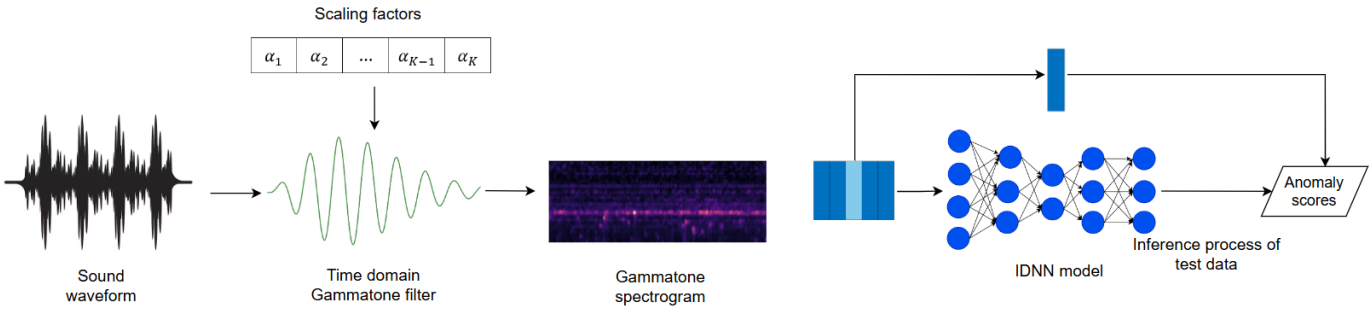


Fig. 1: Proposed method using time domain Gammatone spectrogram and IDNN model for anomalous sound detection

patterns by effectively distinguishing between normal and anomalous machine sounds.

II. Anomalous Sound Detection

ASD involves determining whether the sound emitted from a machine is normal or anomalous. One challenge in this task is the scarcity of anomalous data for training, prompting the adoption of unsupervised learning approaches, which can be constructed using only normal training data [14].

Deterministic acoustic features are commonly favored in ASD systems due to their potential for better generalization, especially when data quantity is limited [15]. Spectrograms, which offer rich acoustic information from sound, are widely utilized in ASD [5]. One type of spectrogram is the Gammatone spectrogram derived from a Gammatone filterbank. Studies in automatic speech recognition have highlighted the effectiveness of Gammatone filterbanks over Mel filterbanks in terms of noise robustness [16]. Although one method [17] employs Gammatone spectrograms extracted in the frequency domain, the use of STFT may potentially reduce resolution.

Qi et al. reported spectrograms derived directly from the time domain offer better representation than those derived through intermediate processes like STFT in the frequency domain. This is caused by frequencies being unnecessarily approximated during STFT calculation, which may result in the loss of some information [7]. Additionally, STFT also lacks resolution for representing sound with fast temporal changes, so it is not particularly suitable for nonstationary sound [8]. Therefore, with spectrograms extracted from the frequency domain, anomalous sounds are potentially harder to detect in valves than in other machines due to nonstationary and sparse characteristics in the sound signal [18].

Regarding the model, unsupervised ASD typically involves training solely on normal data and treating anomalous data as outliers. A basic implementation is the standard AE model, which learns to reconstruct only normal input data [19]. Consequently, the reconstruction of anomalous data may not be as accurate. Some studies have employed self-supervised learning techniques, learning machine label IDs and flagging data as anomalous if there is a significant deviation in predicted IDs [20]. However, self-supervised techniques often perform less unstable across different machines than unsupervised models [21] and rely on data augmentation methods.

III. Proposed Method

The proposed method comprises two primary processes: extracting the time domain Gammatone spectrogram and training the IDNN model for ASD, as illustrated in Fig. 1. During training, the model exclusively utilizes normal data, resulting in higher reconstruction error values for anomalous data, effectively representing the anomaly score.

A. Time Domain Gammatone Spectrogram Feature

A time domain Gammatone spectrogram is a time-frequency domain feature that is calculated using the time domain Gammatone filterbank. Time domain Gammatone filterbank is a well-known set of filters used to simulate the response of the basilar membrane [9]. The time domain Gammatone filterbank uses a gamma function on multiple centers of frequency (f_c) to directly transform the raw waveform.

The impulse response of the time domain Gammatone filterbank at a center frequency f_c is defined as

$$g(t) = t^{n-1} e^{-2\pi b \mathrm{ERB}(\mathrm{f_c})\mathrm{t}} e^{j2\pi f_c t} \tag{1}$$

where $t \geq 0$ is the time in seconds, n is the filter order, and b is the bandwidth coefficient. The time domain Gammatone spectrogram uses Equal Rectangular Bandwith Scale (ERB) to define the nonlinear spacing of the Gammatone frequency band [22]. ERB has finer frequency resolution than mel scale at lower frequency [23]. Therefore, the time domain Gammatone spectrogram can prevent the loss of information at low frequency which in most cases is important for machine sounds. ERB at frequency center f_c is defined as

$$\mathrm{ERB}(\mathrm{f_c}) = 24.7 + 0.108\mathrm{f_c} \tag{2}$$

The time domain Gammatone filterbank is implemented with K time domain infinite impulse response (IIR) Gammatone filters $g^{(k)}(t)$, each representing different center frequencies. By focusing on a specific frequency band, each IIR filter is able to provide accurate frequency magnitude response without the trade-off between time and frequency resolution that comes with windowed methods like STFT [24]. The output of the filterbank $X_k(t)$ from an input signal $x(t)$ can be defined as

$$X_k(t) = x(t) * g^{(k)}(t) \tag{3}$$

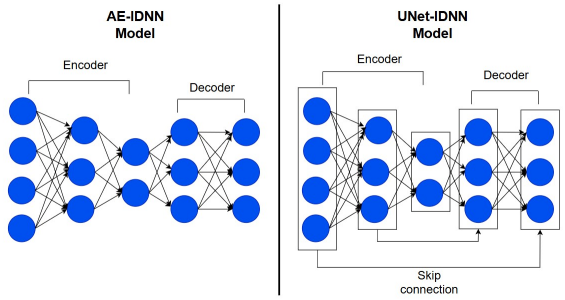


Fig. 2: Architectures of AE-IDNN and UNet-IDNN model

The time domain Gammatone filterbank can be implemented using a wavelet transform where the mother wavelet is $\psi(t) = g(t)$ [25]. Then, with $\alpha > 1$, the k-th filter $g^{(k)}(t)$ can be defined by scaling $\psi(t)$ with a factor a_k of t, as

$$g^{(k)}(t) = \psi(\alpha_k t) \tag{4}$$

$$\alpha_k = \alpha^{\frac{2k}{K-1}-1} \tag{5}$$

The process of extracting and modeling the time domain Gammatone spectrogram is shown in Fig 1.

B. Interpolation Deep Neural Network (IDNN) Model with Autoencoder and UNet Architecture

The IDNN model is a neural network-based model that is an improvement over the conventional approach of using a standard AE model to reconstruct the spectrogram input. IDNN overcomes the difficulty of detecting edge frames in standard AE, which is especially difficult for nonstationary sounds [12].

Instead of reconstructing all input, IDNN uses the center frame prediction method. This means that IDNN constructs the center frame on the basis of the input of left and right frames. Therefore, IDNN receives inputs of $[x_1, ..., x_{\frac{n+1}{2}-1}, x_{\frac{n+1}{2}+1}, ..., x_n]$ frames and predicts the center frame ($x_{\frac{n+1}{2}}$). The loss function of IDNN is defined as

$$L(x_{\frac{n+1}{2}} | D(E([x_1, ..., x_{\frac{n+1}{2}-1}, x_{\frac{n+1}{2}+1}, ..., x_n])) \tag{6}$$

where L defines the algorithm of IDNN's loss function.

IDNN performs better at predicting nonstationary frames using the center frame prediction method [12]. Thus, to increase the quantity of training data, instead of predicting only $\frac{N}{K}$ center frames with N as the total number of frames and K as the number of frames predicted each time, the proposed method uses $N - K + 1$ center frames with some frames overlapping in the data. Therefore, the i_{th} processed data are $[x_i, x_{i+1}, ..., x_{i+K-1}]$ frames instead of $[x_{(i-1)K}, x_{(i-1)K+1}, ..., x_{iK-1}]$ frames. The method is better able to handle the limited quantity of data and the rare occurence of anomalous conditions.

IDNN is implemented with AE architecture (AE-IDNN), which is its standard architecture [12] and UNet architecture (UNet-IDNN), which has the potential to complement the use of the center frame prediction method [13]. With AE architecture, IDNN consists of an encoder and decoder, where the decoder aims to predict the center frame. Meanwhile, with UNet architecture, IDNN consists of an encoder, a decoder, and additional skip connections. With the skip connection, each layer is connected with not only its successive layers because the skip connection creates direct links between layers of the encoder and corresponding layers of the decoder, thus providing more information to the model [26]. The architectures of the AE-IDNN and UNet-IDNN model are depicted in Fig. 2

TABLE I: Normal and anomalous data distribution in MIMII dataset

Machine		Normal data	Anomalous data
Slider	ID 00	1068	356
	ID 02	1068	267
	ID 04	534	178
	ID 06	534	89
Fan	ID 00	1011	407
	ID 02	1016	359
	ID 04	1033	348
	ID 06	1015	361
Valve	ID 00	991	119
	ID 02	708	120
	ID 04	1000	120
	ID 06	992	120
Pump	ID 00	1006	143
	ID 02	1005	111
	ID 04	702	100
	ID 06	1036	102
Total		14719	3400

IV. Experiment

A. Datasets

We used real machinery sounds data from the malfunctioning industrial machine investigation and inspection (MIMII) dataset for evaluation [18] at a 6 dB signal-to-noise ratio (SNR) because the recorder is assumed to be located closer to the target machine. This dataset consists of sounds from four different actual machines including sliders, fans, valves, and pumps augmented with real factory noise. Generally, each machine can be divided into two main categories by its emitted sounds' characteristics including stationary sound (fans and pumps) and nonstationary sound (valves and sliders) [12][18][27]. The MIMII dataset is already used in the DCASE challenge for ASD task [14]. It also contains actual machinery sounds in various anomalous conditions.

Each machine type consists of four machine IDs, with the distribution shown in Table I. For each machine ID, the model is trained separately using the unsupervised method. The inference process uses all anomalous data and normal data with the same amount as anomalous data, whereas the training process uses the remaining normal data.

B. Evaluation Metrics

The area under the receiver operating characteristic curve (ROC AUC) is used as the evaluation metric. It is calculated

TABLE II: Performance comparison of log Mel spectrogram and time domain Gammatone spectrogram feature across different models

Machine		Log Mel Spectrogram				Time Domain Gammatone Spectrogram			
		AE	UNet	AE-IDNN	UNet-IDNN	AE	UNet	AE-IDNN	UNet-IDNN
Slider	ID 00	0.991	0.989	0.958	0.992	1.000	0.997	0.997	0.997
	ID 02	0.893	0.947	0.916	0.984	0.690	0.715	0.822	0.839
	ID 04	0.744	0.799	0.786	0.875	0.906	0.940	0.984	0.954
	ID 06	0.656	0.653	0.668	0.739	0.976	0.997	1.000	0.999
	Average	0.821	0.847	0.832	0.898	0.893	0.912	0.951	0.947
Fan	ID 00	0.763	0.809	0.716	0.765	0.803	0.844	0.855	0.884
	ID 02	0.937	0.947	0.913	0.903	0.911	0.876	0.939	0.917
	ID 04	0.923	0.920	0.891	0.943	0.996	0.994	0.987	0.991
	ID 06	0.981	0.988	0.975	0.973	0.996	1.000	0.995	1.000
	Average	0.901	0.916	0.874	0.896	0.927	0.929	0.944	0.948
Valve	ID 00	0.550	0.656	0.702	0.847	0.450	0.463	0.922	0.947
	ID 02	0.622	0.682	0.704	0.887	0.929	0.957	1.000	1.000
	ID 04	0.602	0.708	0.665	0.932	0.698	0.845	0.946	0.916
	ID 06	0.666	0.696	0.732	0.900	0.523	0.639	0.854	0.836
	Average	0.610	0.686	0.701	0.892	0.650	0.729	0.931	0.925
Pump	ID 00	0.874	0.813	0.804	0.695	0.839	0.827	0.804	0.813
	ID 02	0.503	0.560	0.498	0.531	0.704	0.681	0.715	0.725
	ID 04	0.997	1.000	0.999	1.000	0.960	0.997	0.997	0.998
	ID 06	0.933	0.894	0.907	0.890	0.974	0.971	0.946	0.979
	Average	0.827	0.817	0.802	0.779	0.869	0.869	0.866	0.879
All	Average	0.790	0.816	0.802	0.866	0.835	0.859	0.923	0.925

by measuring the entire two-dimensional area underneath the entire ROC curve. The AUC is a metric used to determine the ability of a model to distinguish between classes [28] and is calculated using the following formula

$$\mathrm{AUC} = \frac{1}{\mathrm{N_-N_+}} \sum_{\mathrm{i}=1}^{\mathrm{N_-}} \sum_{\mathrm{j}=1}^{\mathrm{N_+}} \mathcal{H}(\mathcal{A}_\theta(\mathrm{x_j^+}) - \mathcal{A}_\theta(\mathrm{x_i^-})) \qquad (7)$$

N_- and N_+ subsequently represent the number of anomalous and normal data, x_i and x_j subsequently represent the current anomalous and normal data being processed, $\mathcal{H}(x)$ represents a function that return 1 if $x > 0$ and 0, otherwise, and $\mathcal{A}_\theta(x)$ represents the anomaly score of the data.

C. Experimental Configurations

The Gammatone spectrogram was extracted using the time domain IIR Gammatone filterbank with $K = 64$ and $\alpha = 10$. For the mother wavelet $\psi(t)$, we use an 4th order Gammatone filterbank ($n = 4$) with $b = 1.019$ and $f_c = 600$ Hz. The time domain Gammatone spectrogram was also downsampled to reduce the temporal dimension to 310 frames for each 10-second sound data. Five frames were concatenated to produce a 320-dimensional input vector that was later fed into the model.

We compare the results of our system with a log Mel spectrogram feature that was calculated using librosa library implementation with $n_{\mathrm{FFT}} = 1024$, hop length $= 512$, and mel bands $= 64$. The spectrogram also used the concatenation of 5 frames to produce a 320-dimensional input vector that was later fed into the model.

For both spectrograms, IDNN received an input of $[x_1, x_2, x_4, x_5]$ frames and predicted the center frame (x_3) while minimizing the loss function of the mean squared error (MSE) loss. The hidden layers AE-IDNN consisted of $[64, 32, 16, 32, 64]$ neurons. Meanwhile, the hidden layers of UNet-IDNN consisted of $[64, 32, 16, 64, 128]$ neurons. UNet-IDNN also employed a batch normalization layer to complement the use of skip connections [29]. At the training stage, the time domain Gammatone spectrogram and log Mel spectrogram were trained at 200 epochs except for the log Mel spectrogram with AE-IDNN which was trained at 400 epochs because we observed that it could not provide an optimal solution at 200 epochs.

For comparison, we also used the standard AE model [18] and UNet model [13] to evaluate the generalizability of the proposed features. Both spectrograms were trained at 200 epochs for each model, and the reconstruction error was calculated using MSE.

V. Results

The results in Table II show that time domain Gammatone spectrogram performed better than the log Mel spectrogram in all comparative models. Moreover, the time domain Gammatone spectrogram was also able to improve on the poor performance of the log Mel spectrogram on some machines.

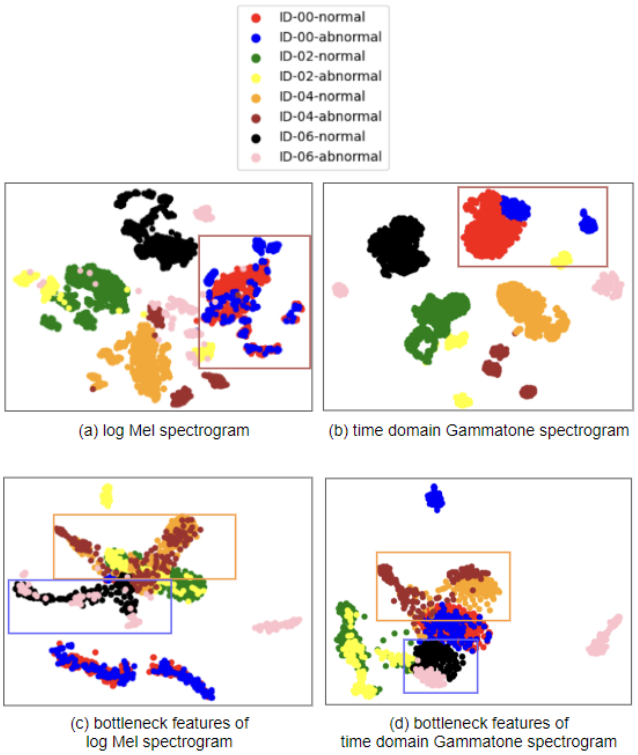


Fig. 3: The t-SNE visualization of the log Mel spectrogram, time domain Gammatone spectrogram, and bottleneck features of both spectrograms in UNet-IDNN model for the machine type fan. Different colors represent difference machine IDs and conditions. The normal and anomalous clusters are highlighted with colored contours for some machines that show a significant difference in the degree of separation between both features.

Hence, the time domain Gammatone spectrogram better distinguished normal and anomalous conditions than the log Mel spectrogram.

Even though using a conventional modeling approach with standard AE can already produce good results for the time domain Gammatone spectrogram, employing IDNN model further enhanced the results, especially on sliders and valves. The improvement of IDNN on sliders and valves was mostly caused by the characteristics of sliders and valves, which have more nonstationary characteristics than other machines such as fans and pumps. The most significant improvement lies in the valves with a more than 20-percentage point (pp) increase in AUC score compared to baseline model AE and UNet.

Additionally, UNet-IDNN performed slightly better than AE-IDNN with the time domain Gammatone spectrogram. Meanwhile, in the log Mel spectrogram, UNet-IDNN significantly improved over AE-IDNN. Thus, this shows the effectiveness of the skip connections component on IDNN for ASD.

Moreover, the use of the time domain Gammatone spectrogram yielded optimal results across both AE-IDNN and UNet-IDNN. Meanwhile, the log Mel spectrogram exhibited optimal result only in UNet-IDNN. This demonstrates that the advantages of the time domain Gammatone spectrogram complemented the use of IDNN model which also led to more consistently optimal results across both IDNN architectures.

To further improve the performance of the UNet-IDNN model, we also employed hyperparameter tuning with a grid search method to find the most optimal structure and other parameters of the model [30]. However, hyperparameter tuning did not successfully improve the overall results.

VI. Discussion

When each feature was modelled using its best model which is UNet-IDNN, the time domain Gammatone spectrogram improved over log Mel spectrogram by 5.9 pp overall. The results also demonstrate the time domain Gammatone spectrogram's ability to model both stationary and nonstationary sound as the overall performance is better than log Mel spectrogram for every machie type in UNet-IDNN. The improvement in nonstationary machines, particularly valves, indicates the potential of the time domain Gammatone spectrogram for representing sounds with rapidly changing characteristics.

In addition to non-stationary sound, the time domain Gammatone spectrogram functions well in stationary machines such as fans. The effectiveness of this approach was evaluated via t-distributed stochastic neighbor embedding (t-SNE) cluster visualization of the latent features of the log Mel spectrogram and time domain Gammatone spectrogram as shown in Fig. 3(a) and 3(b). T-SNE distribution is chosen because of its ability to distinguish nonlinearly separable data which is similar to neural network model capability [31]. The figures show the ability of the time domain Gammatone spectrogram to distinguish normal (red) and anomalous (blue) conditions in fan ID 00 compared to the log Mel spectrogram.

Moreover, the time domain Gammatone spectrogram's performance is also complemented by the use of IDNN models. Fig. 3(c) and 3(d) show the t-SNE distribution of bottleneck features produced by the UNet-IDNN model for fans. In comparison to the log Mel spectrogram, bottleneck features of time domain Gammatone spectrogram can better separate normal and anomalous data. Hence, this indicates the compatibility between the time domain Gammatone spectrogram and neural network-based models.

However, the separation between normal and anomalous data is more distinguishable in the t-SNE distribution of the original features than that of the bottleneck features, as shown in Fig. 3. This indicates the inability of bottleneck features to solely represent spectrogram features, highlighting the importance of input from skip connections in UNet-IDNN model.

Although the time domain Gammatone spectrogram performs optimally, it has limitations especially in terms of computational complexity. Compared to STFT's [32], extracting the time domain Gammatone spectrogram requires more computational resources due to the computation of an IIR filter

over multiple frequency bands. Therefore, this makes the time domain Gammatone spectrogram unsuited to real time ASD.

VII. Conclusion

This study proposed an unsupervised learning approach for anomalous sound detection using the time domain Gammatone spectrogram feature and modeled using Interpolation Deep Neural Network (IDNN) based models. Our proposed method demonstrated a significant improvement of 5.9 pp over the baseline feature, the log Mel spectrogram. Moreover, the time domain Gammatone spectrogram performed better overall than the log Mel spectrogram in all comparative models. Utilizing the IDNN model further enhanced the effectiveness of the time domain Gammatone spectrogram feature, particularly IDNN with UNet architecture. This study underscores the superiority of the time domain Gammatone spectrogram in representing machine sounds compared to the log Mel spectrogram feature, and highlights the effectiveness of the IDNN model for spectrogram-based features.

Acknowledgment

This work was supported by the JST Sakura Science Exchange Program and JSPS KAKENHI grant (23H04344, 23K18491, 22K21304). The ASEAN IVO (http://www.nict.go.jp/en/asean_ivo/index.html) project, 'Spoof Detection for Automatic Speaker Verification', was involved in the production of the contents of this publication and financially supported by NICT (http://www.nict.go.jp/en/index.html).

References

[1] J. Zipfel, F. Verworner, M. Fischer, U. Wieland, M. Kraus, and P. Zschech, "Anomaly detection for industrial quality assurance: A comparative evaluation of unsupervised deep learning models," *Computers & Industrial Engineering*, vol. 177, p. 109045, 2023.

[2] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PloS one*, vol. 11, no. 4, p. e0152173, 2016.

[3] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212–224, 2018.

[4] S. Abbasi, M. Famouri, M. J. Shafiee, and A. Wong, "Outliernets: Highly compact deep autoencoder network architectures for on-device acoustic anomaly detection," *Sensors*, vol. 21, no. 14, p. 4805, 2021.

[5] E. C. Nunes, "Anomalous sound detection with machine learning: A systematic review," *arXiv preprint arXiv:2102.07820*, 2021.

[6] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda *et al.*, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2006.05822*, 2020.

[7] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on gammatone filters for robust speech recognition," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2013, pp. 305–308.

[8] I. Daubechies, *Ten lectures on wavelets*. SIAM, 1992.

[9] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, 1995.

[10] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous sound detection using deep audio representation and a blstm network for audio surveillance of roads," *Ieee Access*, vol. 6, pp. 58 043–58 055, 2018.

[11] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[12] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 271–275.

[13] T. V. Hoang, H. C. Nguyen, and G. N. Pham, "Unsupervised detection of anomalous sound for machine condition monitoring using different auto-encoder methods," Tech. Rep, Tech. Rep., 2020.

[14] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," *arXiv preprint arXiv:2206.05876*, 2022.

[15] D. Ditter and T. Gerkmann, "A multi-phase gammatone filterbank for speech separation via tasnet," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 36–40.

[16] L. Deng and Y. Gao, "Gammachirp filter banks applied in roust speaker recognition based on gmm-ubm classifier." *Int. Arab J. Inf. Technol.*, vol. 17, no. 2, pp. 170–177, 2020.

[17] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, and M. Cobos, "Anomalous sound detection using unsupervised and semi-supervised autoencoders and gammatone audio representation," *arXiv preprint arXiv:2006.15321*, 2020.

[18] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," *arXiv preprint arXiv:1909.09347*, 2019.

[19] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pp. 353–374, 2023.

[20] W. Junjie, W. Jiajun, C. Shengbing, S. Yong, and L. Mengyuan, "Anomaly sound detection system based on multi-dimensional attention module," 2023.

[21] J. Guan, Y. Liu, Q. Kong, F. Xiao, Q. Zhu, J. Tian, and W. Wang, "Transformer-based autoencoder with id constraint for unsupervised anomalous sound detection," *EURASIP journal on audio, speech, and music processing*, vol. 2023, no. 1, p. 42, 2023.

[22] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[23] W. Lambamo, R. Srinivasagan, and W. Jifara, "Analyzing noise robustness of cochleogram and mel spectrogram features in deep learning based speaker recognition," *applied sciences*, vol. 13, no. 1, p. 569, 2022.

[24] R. G. Lyons, *Understanding digital signal processing, 3/E*. Pearson Education India, 1997.

[25] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, vol. 27, no. 3-4, pp. 261–279, 1999.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[27] B. A. Tama, M. Vania, I. Kim, and S. Lim, "An efficientnet-based weighted ensemble model for industrial machine malfunction detection using acoustic signals," *IEEE Access*, vol. 10, pp. 34 625–34 636, 2022.

[28] C. X. Ling, J. Huang, H. Zhang *et al.*, "Auc: a statistically consistent and more discriminating measure than accuracy," in *Ijcai*, vol. 3, 2003, pp. 519–524.

[29] A. Labatie, "Characterizing well-behaved vs. pathological deep neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3611–3621.

[30] S. Mezzah and A. Tari, "Practical hyperparameters tuning of convolutional neural networks for eeg emotional features classification," *Intelligent Systems with Applications*, vol. 18, p. 200212, 2023.

[31] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[32] W. T. Cochran, J. W. Cooley, D. L. Favin, H. D. Helms, R. A. Kaenel, W. W. Lang, G. C. Maling, D. E. Nelson, C. M. Rader, and P. D. Welch, "What is the fast fourier transform?" *Proceedings of the IEEE*, vol. 55, no. 10, pp. 1664–1674, 1967.

Speech emotion recognition based on crossmodal transformer and attention weight correction

Ryusei Terui* and Takeshi Yamada*
* University of Tsukuba, Japan
E-mail: s2320626@u.tsukuba.ac.jp

***Abstract*—In recent years, speech emotion recognition (SER) methods that use both acoustic features and text features derived through automatic speech recognition (ASR) have become mainstream. Furthermore, the crossmodal integration of acoustic and text features using a crossmodal transformer encoder has been proposed and succeeded in improving the SER accuracy. However, these methods have a problem in that ASR errors occur frequently, especially for speech that contains emotion, which affects the SER accuracy. To solve this problem, a method of correcting the self-attention weights based on the word-level confidence measure (CM), which indicates the reliability of ASR results, has been proposed. In this paper, we propose a method that combines the crossmodal transformer encoder and the attention weight correction with CM to further improve the SER accuracy. The network of the proposed method includes two different attention mechanisms: scaled dot-product attention and self-attention. In this paper, we applied the attention weight correction to each attention mechanism and verified their effectiveness. Results of experiments using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset revealed that the attention weight correction for self-attention can achieve higher SER improvement and the SER accuracy of the proposed methods is equal to or higher than that of state-of-the-art SER methods.**

I. Introduction

Emotion recognition is a key to realizing natural and smooth human-to-human and human-to-machine communication. Previously, many studies of the recognition of emotions from various modalities such as speech, facial expressions, gestures, and brain waves were conducted[1] [2] [3]. In particular, speech is a basic modality used on a daily basis and there are many situations where only speech can be used. Therefore, there are high expectations for speech emotion recognition (SER) technology. This technology makes it possible, for example, to provide responses that consider the client's emotions in call centers and responses that show empathy for users in speech assistant applications.

In previous SER methods, emotions have been classified using acoustic features based on frequency spectrograms such as the Mel-frequency cepstrum coefficient (MFCC), which is widely used in automatic speech recognition (ASR) and speaker recognition [1]. In recent years, methods that apply ASR to input speech to obtain the utterance text and classify emotions using both acoustic features and text features have become mainstream. Furthermore, the crossmodal integration of acoustic and text features using a crossmodal transformer encoder has been proposed and succeeded in improving the SER accuracy [4].

However, these methods have a problem in that ASR errors occur frequently, especially for speech that contains emotion, which affects the SER accuracy. To solve this problem, Feng *et al.* proposed a method of multi-task learning of ASR and SER to make ASR robust to emotional speech [5]. However, the computational and time costs are very high. On the other hand, Santoso *et al.* regarded ASR errors to be emotional cues and proposed a method of correcting the self-attention weights based on the word-level confidence measure [6], which indicates the reliability of ASR results, and showed its effectiveness [7].

In this paper, we propose a method that combines the crossmodal transformer encoder and the attention weight correction to further improve the SER accuracy. The network of our proposed method consists of a crossmodal transformer encoder for the crossmodal integration of acoustic and text features, a self-attention mechanism for emphasizing time frames important for classification, and a dense layer for classification. It includes two different attention mechanisms: scaled dot-product attention and self-attention. In this paper, we apply the attention weight correction to each attention mechanism and verify these effectiveness.

II. Conventional SER methods

A. Method using crossmodal transformer encoder

A method using a crossmodal transformer encoder was proposed to realize the crossmodal integration of acoustic features and text features. A transformer encoder is the encoder part of the transformer model [8]. It is composed of a multihead attention layer, a feed forward layer, and residual connections to these layers. Generally, the multihead attention layer of the singlemodal transformer encoder uses the same modality features for the query Q, the key K, and the value V. On the other hand, that of the crossmodal transformer encoder uses different modality features for K and V from Q. It has been reported that the SER accuracy is improved by performing the crossmodal integration of acoustic and text features instead of extracting them independently [4].

Fig. 1 shows an example of an SER method using a crossmodal transformer encoder. This is used as the base method in this paper. Fig. 2 also shows an overview of the crossmodal transformer encoder in the acoustic feature

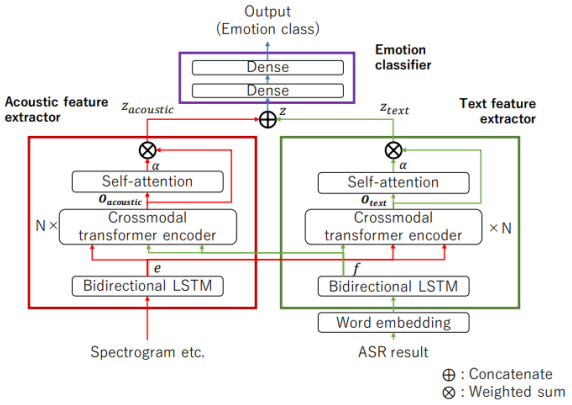


Fig. 1. Overview of the base method.

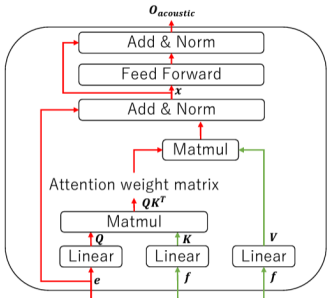


Fig. 2. Overview of the crossmodal transformer encoder.

extractor. In this figure, for simplicity, singlehead attention is used, but multihead attention is actually used. In the acoustic feature extractor of the base method, the acoustic features are fed to a bidirectional long short-term memory (BLSTM) to extract the feature $\boldsymbol{e}_k$, which is defined for each time frame index k as

$$\boldsymbol{e}_k = \boldsymbol{g}_k \oplus \boldsymbol{h}_k, \tag{1}$$

where $\boldsymbol{g}_k$, $\boldsymbol{h}_k$, and $\oplus$ represent the forward and backward hidden states of the BLSTM and a concatenation operation, respectively. In the text feature extractor, the text features are also fed to a BLSTM in the same manner as in the acoustic feature extractor to obtain $\boldsymbol{f}_l$ for the lth word. Then, in the acoustic feature extractor, $\boldsymbol{e}$ and $\boldsymbol{f}$ are fed to the crossmodal transformer encoder, and $\boldsymbol{e}$ is transformed to $\boldsymbol{Q}$ and $\boldsymbol{f}$ is transformed to $\boldsymbol{K}$ and $\boldsymbol{V}$ to apply the scaled dot-product attention as follows.

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{2}$$

where d_k represents the dimension of $\boldsymbol{Q}$. This results in the crossmodal integration of acoustic and text features. We add $\boldsymbol{e}$ to the output of the scaled dot-product attention and normalize it to obtain $\boldsymbol{x}$, which is fed into the feed forward network layer defined as

$$\mathrm{FFN}(\boldsymbol{x}) = \max(0, \boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2, \tag{3}$$

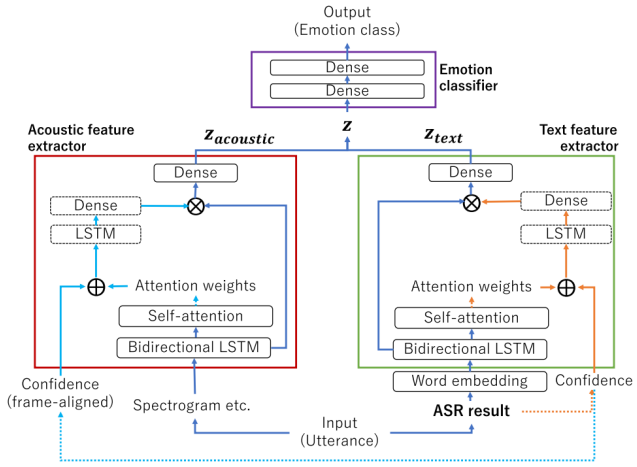


Fig. 3. Overview of the SER method with the attention weight correction.

where $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, $\boldsymbol{b}_1$, and $\boldsymbol{b}_2$ are trainable parameters. We add $\boldsymbol{x}$ to the output of the feed forward network layer and normalize it to obtain $\boldsymbol{O}_{acoustic}$, which is then fed into the self-attention mechanism to emphasize the important time frames.

$$\alpha_k = \mathrm{softmax}(\boldsymbol{y}_k \tanh(\boldsymbol{Y}\boldsymbol{O}_{acoustic,k}^T)), \tag{4}$$

where α_k is the attention weight for the kth time frame, and $\boldsymbol{y}_k$ and $\boldsymbol{Y}$ are trainable parameters. The weighted sum $\boldsymbol{z}_{acoustic}$ from the attention weight α_k and $\boldsymbol{O}_{acoustic}$ is defined as

$$\boldsymbol{z}_{acoustic} = \sum_{k=1}^{T} \alpha_k \boldsymbol{O}_{acoustic,k}, \tag{5}$$

We apply the same processing as in the acoustic feature extractor to the text features to obtain $\boldsymbol{z}_{text}$ in the text feature extractor. Finally, these two features are concatenated to obtain z and classified with a fully connected network to obtain the emotion class. Although it is superior in terms of the crossmodal integration of acoustic and text features, this method is vulnerable to ASR errors.

B. Method using attention weight correction

Fig. 3 shows an overview of the SER method with attention weight correction [7]. This method takes acoustic features such as a frequency spectrogram and the ASR text as input, and uses BLSTM to extract acoustic features and text features independently as in Eq. (1). Then, a feature vector that emphasizes important time frames (or words) is obtained using the self-attention mechanism. This is similar to Eqs. (4) and (5). The two features obtained in this manner are concatenated and fed into the fully connected layer to obtain the emotion class.

The advantage of this method is the attention weight correction. When ASR errors occur, unexpected attention weights may be added, which may lead to a decrease in SER accuracy. To solve this problem, a method of correcting the attention weights using a confidence measure (CM) output together with the utterance text from ASR was introduced. The CM is an indicator of the reliability of ASR results and is expressed as

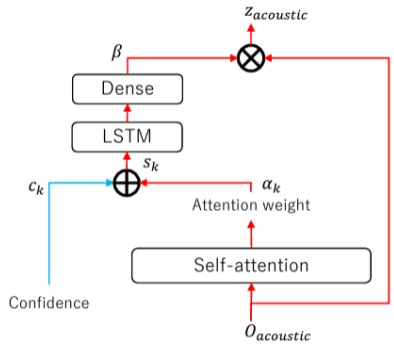


Fig. 4. Overview of the self-attention weight correction (proposed method 1).

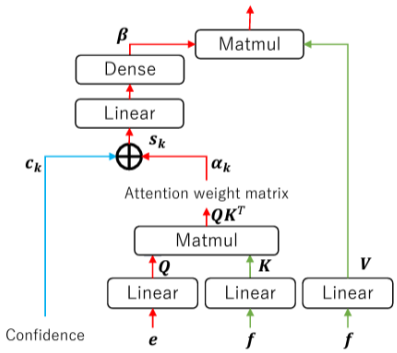


Fig. 5. Overview of the scaled dot-product attention weight correction (proposed method 2).

a value from 0 to 1, and the closer this value is to 1, the higher the reliability. Words with low reliability have a high possibility of ASR errors and are therefore likely to include significant emotional features, so the attention weights of text features are corrected to be small and those of acoustic features are corrected to be large. The details of the attention weight correction method will be described later.

III. Proposed Method

In this section, we describe the proposed method in detail. The network of the proposed method is the same as that of the base method using the crossmodal transformer encoder in Fig. 1. The proposed method includes two different attention mechanisms, scaled dot-product attention and self-attention, but it is not clear how to apply attention weight correction to which attention mechanism. Therefore, in this paper, we compare the following two attention weight correction methods:

1) Self-attention weight correction (proposed method 1): Fig. 4 shows an overview of this method. This method is represented as

$$s_k = \mathrm{LSTM}(\alpha_k \oplus c_k), \tag{6}$$

$$\beta_1, ..., \beta_T = \mathrm{softmax}(s_1, ..., s_T), \tag{7}$$

where α_k is the attention weight for the kth time frame, c_k is the corresponding CM, and β_k is the corrected attention weight. This is the same method as the attention weight correction described in Sect. II.B.

2) Scaled dot-product attention weight correction (proposed method 2): Fig. 5 shows an overview of this method. It shows the network structure of the crossmodal transformer encoder in Fig. 1, and attention weight correction is applied to the attention weight matrix. First, the confidence matrix is obtained by calculating the cross product of CM vectors of acoustic features and text features. The corrected attention weight matrix is obtained as follows:

$$\boldsymbol{s_k} = \mathrm{linear}(\boldsymbol{\alpha_k} \oplus \boldsymbol{c_k}), \tag{8}$$

$$\boldsymbol{\beta_1}, ..., \boldsymbol{\beta_T} = \mathrm{softmax}(\boldsymbol{s_1}, ..., \boldsymbol{s_T}), \tag{9}$$

where $\boldsymbol{\alpha_k}$ is the kth column of the attention weight matrix, $\boldsymbol{c_k}$ is the column of the confidence matrix, and $\boldsymbol{\beta_k}$ is the column of the corrected attention weight matrix. This is a newly proposed method for scaled dot-product attention.

The proposed method 1 performs crossmodal integration while ignoring the presence of ASR errors, and then modifies the attention weights in the final self-attention step. On the other hand, the proposed method 2 performs crossmodal integration while considering the presence of ASR errors. In the next section, we compare the effectiveness of each method.

IV. Experiments

In this section, we first verify the effectiveness of crossmodal integration of acoustic features and text features. We then compare the effectiveness of the proposed method 1 with that of the proposed method 2, and compare their performance with that of state-of-the-art methods.

A. Experimental setting

In this study, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [9], which is widely used as one of the benchmarks in SER. This dataset consists of English speech data of 1 to 19 seconds from 10 speakers, 5 males and 5 females, and each utterance is labeled with an emotion class. In this paper, we use four emotion classes, namely, happy, sad, neutral, and angry, following past studies, whose numbers of data are 1689, 1084, 1708, and 1103, respectively.

The acoustic features are 33-dimensional vectors consisting of 20-dimensional MFCC, 12-dimensional constant Q-transform (CQT), and 1-dimensional fundamental frequency (F0). The text features are 768-dimensional word-embedding vectors obtained by pretrained BERT [10]. The utterance text is obtained using ASR pretrained with the LibriSpeech [11] dataset and the Kaldi speech recognition toolkit [12].

In this experiment, we use the following two evaluation measures: unweighted accuracy (UA), which is the classification accuracy for the entire test data, and weighted accuracy (WA),

TABLE I
SER PERFORMANCE OF THE BASE METHOD AND THE PROPOSED METHODS.

method	crossmodal integration	UA (%)	WA (%)
base method	w/o	73.4	75.7
base method	w	74.7	76.9
proposed method 1	w	75.4	77.5
proposed method 2	w	75.3	77.2

which considers the imbalance in the number of data among emotion classes in the test data. UA and WA are defined as

$$\mathrm{UA} = \frac{\sum_{i=1}^{N} t_{ii}}{\sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij}}, \tag{10}$$

$$\mathrm{WA} = \frac{1}{N} \sum_{i=1}^{N} \frac{t_{ii}}{\sum_{j=1}^{N} t_{ij}}, \tag{11}$$

where N is the number of emotion classes and t_{ij} is the number of data for class i that was classified as class j. In this experiment, we conducted fivefold cross validation and each evaluation value was the average of 5 validations. We used Adam [13] as the optimizer with a learning rate of 1.0e-4 and a weight decay of 1.0e-9. The loss function used for model training was the cross-entropy loss and the dropout and the batch size were set to 0.4 and 40, respectively. The results were taken from the best WA out of 100 epochs. Finally, the number of layers in the crossmodal transformer encoder was set to 3, and the number of heads in the multihead attention layer was set to 8.

B. Results

Table I shows the experimental results for each method. First, we confirm the effectiveness of crossmodal integration. In the table, the base method without crossmodal integration is the case where the same features are input to the Q, K, and V of the base method, and corresponds to extracting acoustic and text features independently. From the table, both the UA and WA of the base method are higher than those of the base method without crossmodal integration, confirming the effectiveness of crossmodal integration.

Then, we verify the effectiveness of the two types of attention weight correction. Compared with that of the base method, the WA values of the proposed methods 1 and 2 improved by 0.6% and 0.3%, respectively. This shows that the effect of ASR errors could be reduced by correcting the attention weights, considering the CM. In addition, the WA of the proposed method 1 is 0.3% higher than that of the proposed method 2. This implies that correcting the self-attention weight is more effective. One possible reason for this is that the process of self-attention weight correction is simpler and therefore easier to train.

Finally, Table II shows the UA and WA values of the proposed and state-of-the-art SER methods. These methods use acoustic and text features and parts of these methods adopt the crossmodal integration. We can see that the WA values of the proposed methods 1 and 2 are equal to or higher than those of the state-of-the-art SER methods.

TABLE II
SER PERFORMANCE OF THE STATE-OF-THE-ART AND PROPOSED METHODS.

method	crossmodal integration	UA (%)	WA (%)
Feng *et al.* [5]	w/o	69.7	68.1
Santoso *et al.* [7]	w/o	76.8	76.6
Chu *et al.* [4]	w	75.1	76.3
Zhang *et al.* [14]	w	76.4	77.1
Priyasad *et al.* [15]	w	76.8	77.3
Proposed method 1	w	75.4	77.5
Proposed method 2	w	75.3	77.2

From the above, it was confirmed that the combination of the crossmodal transformer encoder and the attention weight correction successfully improves the SER performance.

V. CONCLUSIONS

In this paper, we proposed a method that combines the crossmodal transformer encoder and the attention weight correction to further improve the SER accuracy. Since the network of the proposed method includes two different attention mechanisms, scaled dot-product attention and self-attention, we applied the attention weight correction to each of them and compared their effectiveness. The experimental results showed that the attention weight correction for self-attention can achieve higher SER accuracy and that the SER accuracy of the proposed methods is equal to or higher than that of state-of-the-art SER methods.

REFERENCES

[1] Z. Liao and S. Shen, "Speech emotion recognition based on swin-transformer," *Journal of Physics: Conference Series*, ser. 2508 012056, 2023.

[2] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[3] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "EEG-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[4] I.-H. Chu, Z. C. Xinlu, Y. M. Han, J. Xiao, and P. Chang, "Self-supervised cross-modal pretraining for speech emotion recognition and sentiment analysis," in *EMNLP*, 2022, pp. 5105–5114.

[5] H. Feng, S. Ueno, and T. Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word asr," in *INTERSPEECH*, 2020, pp. 501–505.

[6] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech communication*, vol. 45, no. 4, pp. 455–470, 2005.

[7] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto, and S. Makino, "Speech emotion recognition based on self-attention weight correction for acoustic and text features," *IEEE Access*, vol. 10, pp. 115 732–115 743, 2022.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture dataset," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, vol. 1, 2019, pp. 4171–4186.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio book," in *ICASSP*, 2015, pp. 5206–5210.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *ASRU*, 2011, pp. 1–4.

[13] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015, pp. 1–15.

[14] Z. Zhang, D. Liu, S. Liu, A. Wang, J. Gao, and Y. Li, "Turbo your multi-modal classification with contrastive learning," in *INTERSPEECH*, 2023, pp. 1848–1852.

[15] D. Prisayad, T. Fernando, S. Sridharan, S. Denman, and C. Fookes, "Dual memory fusion for multimodal speech emotion recognition," in *INTERSPEECH*, 2023, pp. 4543–4547.

Adaptive Time-Varying Graph Learning for Traffic Flow Data Based on Anomaly Moment Detection

Shuhong Chen*, Zewei Chen*, Chen Li†, Xianwei Zheng†, Minfan He† and Xutao Li*
* Key Lab of Digital Signal and Image Processing of Guangdong Province, Department of Electronic Engineering, Shantou University, Shantou, 515063, China
E-mail: 18shchen3@stu.edu.cn, 22zwchen@stu.edu.cn, lixt@stu.edu.cn
† School of Mathematics and Big Data, Foshan University, Foshan, 52800, China
E-mail: fslichen@fosu.edu.cn, alexwzheng@fosu.edu.cn, heminfan1980@126.com

***Abstract*—This paper proposes an adaptive time-varying graph structure construction strategy for traffic flow data based on anomaly moment detection, which is capable of identifying anomaly moments within traffic flow signals and adaptively dividing the temporal dimension to construct corresponding graph representations for distinct traffic patterns. Specifically, we leverage a threshold determination method that incorporates the smoothness and energy of temporal difference traffic flow signals to identify anomaly moments in the traffic flow. Upon detecting these anomaly moments, we utilize them as segmentation boundaries along the timeline, adaptively partitioning the continuous time series into multiple dynamic and relatively stable subsequences. Within each subsequence, we apply a time-varying graph learning method to independently construct a graph structure that reflects the traffic flow characteristics of that specific period, thereby achieving an adaptive time-varying property in the traffic flow graph structure. Experimental results demonstrate that the proposed method exhibits reliable performance in detecting anomaly moments in traffic flow, and the adaptive time-varying graph structures accurately reflect the varying characteristics of traffic flow across different time periods.**

I. Introduction

With the acceleration of urbanization and the increasing complexity of transportation, traffic problems such as congestion and frequent accidents have become more prominent [1]. If these traffic abnormalities are not taken care of promptly, they will have adverse impacts on urban economy, environment, and social stability [2]. Against this backdrop, numerous methods for modeling and anomaly detection of traffic flow data have emerged [3][4]. Aiming to effectively identify anomaly situations in urban road networks, such as the typical phenomenon of sharp increases in traffic flow during peak hours, we are dedicated to delving into the dynamic evolution of traffic flow data. By capturing its temporal variation characteristics, we aim to uncover the pattern changes and underlying differences exhibited by the traffic network structure at different times, i.e., the time-varying graph structural properties. Ultimately, this endeavor serves to provide a scientific basis for optimizing traffic management strategies and thereby enhancing road traffic safety [5].

The traffic network is a typical complex network structure. To comprehensively describe its intricate spatial topological relationships and consider the spatio-temporal correlation between flow data on traffic networks, many research methods have chosen the graph structure as an effective modeling tool for road networks [6][7].

Currently, many traditional traffic flow analysis methods rely primarily on static graphs [4], which involve modeling and analysis on traffic networks with fixed graph structures. However, these methods have certain limitations, as they cannot fully capture the dynamic evolution of traffic networks over time. Given the time-varying characteristic of the transportation system, traffic flow exhibits significant variations across different time periods. Compared to static graphs, time-varying graphs can better reflect the spatio-temporal characteristics of actual traffic networks.

In time-varying graphs, each time point is treated as a slice of the graph, where the attributes of the nodes or the edge weights dynamically adjust over time [8]. This dynamic modeling approach can better capture fluctuations, periodic changes, and long-term trends in traffic flow, facilitating a more accurate understanding of dynamic patterns of traffic flow and improving the flexibility and adaptability of anomaly detection methods. Recent research on time-varying graph learning is mostly based on the assumption of temporal difference smoothness of graph signals [8][9][10][11].

In this paper, we propose an innovative strategy for time-varying graph learning. Traditional methods of time-varying graph learning tend to rely on predefined time windows, constructing time-varying graphs at fixed time intervals and performing corresponding graph learning and analysis within each time window [8]. However, this fixed time interval strategy may fall short in handling the immediacy and complexity of dynamic changes in traffic flow. Additionally, fixed time intervals often lead to the generation of numerous temporal subsequences containing redundant information, thereby increasing unnecessary computational burden.

To overcome the limitations of traditional time-varying structural analysis methods and meet the need for more flexible anomaly detection while conserving computational resources, this paper proposes an adaptive time-varying graph learning method based on anomaly moments. This method employs a thresholding approach grounded in the smoothness and energy of temporal difference traffic flow signals to detect these anomaly moments, subsequently enabling an adaptive partitioning of the graph sequences. Within each partitioned

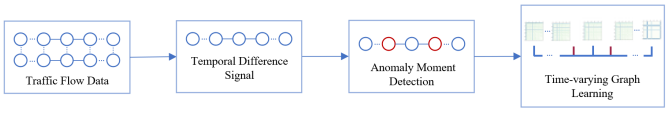


Fig. 1. The overall framework proposed in this article, achieving the task of detecting anomaly moments and conducts time-varying graph learning from input traffic flow data. Our framework comprises temporal difference processing, anomaly moment detection, and time-varying graph learning. Specifically, we first apply temporal difference processing to the input traffic flow data to obtain temporal difference traffic flow signals. Then, the threshold determination method based on the smoothness and energy of temporal difference traffic flow signals is used to identify the anomaly moments in the traffic flow. Finally, we divide the timeline into multiple subsequences with these anomaly moments as segmentation boundaries. Time-varying graph learning is then conducted within each subsequence, where the graph structure remains consistent during each subsequence period, reflecting the interactive characteristics of the traffic network within that specific time frame. Note that the detected anomaly moments are marked with red circles and vertical lines.

sequence, time-varying graph learning is conducted, enabling a more accurate and rapid reflection of the spatio-temporal characteristics of traffic flow. Compared to conventional methods, the proposed approach delves deeper into the dynamic properties of traffic networks, allowing for a more precise analysis of structural differences manifested across different traffic periods (e.g., peak and off-peak hours). It enhances the model's adaptability to data variations and achieves notable optimization in the utilization of computational resources.

II. Proposed Method

In our approach, we leverage graphs to model complex relationships between traffic network sensors. Fig. 1 illustrates the overall pipeline of our method. In what follows, we detail our problem and each module of the pipeline.

A. Problem Specification

The adaptive time-varying graph learning method proposed in this article is based on anomaly momnet detection.

In general, anomaly detection is defined as the process of identifying instances that deviate from regular behavioral patterns within observed datasets. Regarding the types of traffic flow anomalies, there are two most prevalent categories: (1) **Recurrent traffic flow anomalies**, which occur daily and are generally attributed to factors such as routine commuting activities. (2) **Irregular traffic flow anomalies**, which are typically triggered by unexpected events like traffic accidents or special societal occurrences.

This paper focuses on recurrent flow anomalies, investigating the problem of detecting anomalous moments in traffic flow and the methodology for constructing adaptive time-varying graph structures based on such detections.

Let the traffic flow data at any time stamp t be denoted as $\mathbf{x}_t \in R^N$, where N represents the number of traffic flow sensors in the road traffic network, and $x_{i,t} \in R$ indicates the traffic volume captured by the i-th sensor at time t. We handle a series of continuously observed traffic flow data, covering H time steps, denoted as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_H] \in R^{N\times H}$.

Our primary objective is to detect anomaly moments within the traffic flow. Specifically, we aim to design an algorithm that outputs a binary label vector $\hat{\mathbf{y}} \in R^H$, where each element $\hat{\mathbf{y}}_t \in \{0, 1\}$ the presence or absence of an anomaly at the corresponding time step t.

Subsequently, to analyze the transportation network more flexibly, we adaptively divide the timeline to obtain the unique graph structures of traffic flow in different time stages.

B. Temporal Difference Graph Signals

Given the inherent stability of traffic flow over short periods, we introduce time-difference graph signals to capture the dynamic variations in traffic flow with greater precision.

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_H]$ represent the time-varying graph signal, where $\mathbf{x}_i \in R^N$ denotes the graph signal at time i. To capture changes in traffic flow between adjacent time points, we introduce the time difference operator $\mathbf{D}_h \in R^{M\times(M-1)}$ [12], defined as follows:

$$\mathbf{D}_h = \begin{bmatrix} -1 & & \\ 1 & -1 & \\ & 1 & \ddots \\ & & \ddots \ -1 \\ & & 1 \end{bmatrix} \in R^{M\times(M-1)}. \tag{1}$$

Utilizing this operator, we can compute the time-difference graph signal:

$$\mathbf{X}\mathbf{D}_h = [\mathbf{x}_2 - \mathbf{x}_1, \mathbf{x}_3 - \mathbf{x}_2, \ldots, \mathbf{x}_H - \mathbf{x}_{H-1}]. \tag{2}$$

By capturing the differences between different time points, the time-difference graph signals can accurately reveal the instantaneous fluctuations in traffic flow. Subsequently, we will utilize the smoothness and energy features of the time-difference graph signals to detect momentary anomaly in traffic flow data, pinpointing potential anomalous time points. Specifically, the smoothness of the time-difference graph signals reflects the continuity and gradualness of changes between adjacent time points, while the energy measures the intensity of the signal variations at these points.

a) Smoothness: To quantify the smoothness of the graph signals, we introduce the graph Laplacian quadratic form $S_2(\mathbf{x})$ [13], defined as follows:

$$S_2(\mathbf{x}) = \sum_{(i,j)\in\mathcal{E}} \mathbf{A}_{i,j}[x_j - x_i]^2 = \mathbf{x}^T\mathbf{L}\mathbf{x}. \tag{3}$$

When subtle variations occur in graph signals, i.e., the graph signals exhibit smoothness properties, the value of the graph Laplacian quadratic form $S_2(\mathbf{x})$ will be relatively small.

For time-varying graph signals, we further extend the application of the graph Laplacian quadratic form. By computing the Laplacian quadratic form for the graph signal $\mathbf{x}_i$ at each time instant and summing them up, the Laplacian quadratic form $S_2(\mathbf{X})$ [12] for time-varying graph signals can be expressed as:

$$S_2(\mathbf{X}) = \sum_{i=1}^{M} \mathbf{x}_i^T \mathbf{L} \mathbf{x}_i = tr(\mathbf{X}^T \mathbf{L} \mathbf{X}). \tag{4}$$

Accordingly, we obtain the smoothness metric expression for time-lagged graph signals as:

$$S_2(\mathbf{X}\mathbf{D}_h) = tr((\mathbf{X}\mathbf{D}_h)^T \mathbf{L} \mathbf{X}\mathbf{D}_h). \tag{5}$$

In particular, compared to $S_2(\mathbf{X})$ in real-world datasets, the time-varying graph signal $S_2(\mathbf{X}\mathbf{D}_h)$ exhibits superior smoothing properties. Specifically, even if the original time-varying graph signal $\mathbf{x}_i$ does not exhibit smoothness on the graph, its difference signal $\mathbf{x}_i - \mathbf{x}_{i-1}$ often exhibits smooth characteristics on the graph. In this paper, we utilize the correlation matrix between nodes as the initial graph structure.

b) Energy: We determine the energy by calculating the sum of squared Frobenius norms of each time-difference graph signal, mathematically expressed as follows:

$$E(\mathbf{X}\mathbf{D}_h) = \sum_{j=1}^{H-1} \|(\mathbf{X}\mathbf{D}_h)_j\|_F^2, \tag{6}$$

where H represents the time step, $(\mathbf{X}\mathbf{D}_h)_j$ denotes the time-difference graph signal at the j-th time step, and $|\cdot|_F^2$ stands for the squared Frobenius norm. A smooth graph signal typically exhibits small differences between adjacent time points, resulting in a lower energy in its time-difference representation.

C. Anomaly Moment Detection

To capture and identify anomaly moments associated with spatio-temporal dynamics from traffic flow data, we will adopt a threshold algorithm to recognize such moments by setting appropriate thresholds after obtaining the smoothness and energy features of the time-difference graph signals. The specific algorithm is as follows:

$$\hat{\mathbf{y}} = mS_2(\mathbf{X}\mathbf{D}_h) > S \parallel nE(\mathbf{X}\mathbf{D}_h) > E, \tag{7}$$

where m and n are coefficients for smoothness and energy, respectively, reflecting the sensitivity of these two characteristics to the detection of anomaly moments. S and E are the corresponding thresholds. To ensure the accuracy of threshold selection, we will set these thresholds based on the statistical characteristics of the signals and domain-specific expertise. In this paper, we set the thresholds based on prior knowledge.

D. Time-varying Graph Learning

Our time-varying graph learning method adaptively partitions the timeline by incorporating anomaly moments. Specifically, based on the detection results of abnormal moments, we re-divide the entire time period to extract continuous subsequences:

$$\hat{\mathbf{X}} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T] = g(\mathbf{X}, \hat{\mathbf{y}}). \tag{8}$$

Here the duration of each subsequence is not fixed. These subsequences will be further used for subsequent time-varying graph structure learning to delve deeper into the association between anomaly moments and network topology.

The next step aims to learn the correlation patterns among sensors within different time segments for reconstructed traffic flow data, specifically the representation of graph structure (adjacency matrix). We have segmented the traffic flow data into non-overlapping time segments, denoted as $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$, where the observation data within each time segment t is represented as $\mathbf{X}_t = \left[\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_K^{(t)}\right] \in R^{N \times K}$ with a length of K. It should be noted that the value of K can vary for different time segments $\mathbf{X}_t$. Each time segment is considered as a static slice of the traffic network, indicating that the graph structure corresponding to observations within the same time segment should remain unchanged. Here, nodes represent traffic flow sensors, and edges reflect the connection relationships among these nodes.

Time-varying graph learning methods require the learned graphs to maintain smoothness across adjacent time slots. To achieve this, we incorporate prior knowledge of temporal graph variations and transform the time-varying graph learning problem into solving the following optimization problem.

$$\min_{\mathbf{A}_t \in \mathcal{A}_m} \sum_{t=1}^{T} \frac{1}{2} \|\mathbf{A}_t \circ \mathbf{Z}_t\|_1 + f(\mathbf{A}_t) + \eta \sum_{t=2}^{T} R(\mathbf{A}_t - \mathbf{A}_{t-1}). \tag{9}$$

Here, the valid set of weighted adjacency matrices $\mathcal{A}_m$ is defined as follows:

$$\mathcal{A}_m = \{\mathbf{A} \in R_+^{N \times N} \mid \mathbf{A} = \mathbf{A}^T, A_{ii} = 0\}. \tag{10}$$

Additionally, $\mathbf{Z}_t$ represents the pairwise distance matrix, defined as:

$$(\mathbf{Z}_t)_{ij} = \sum_{k=1}^{K} \left\| \left(\mathbf{x}_k^{(t)}\right)_i - \left(\mathbf{x}_k^{(t)}\right)_j \right\|^2. \tag{11}$$

In equation Eq. (9), the third term represents the regularization term characterizing temporal variations in the edges of the graph, with the parameter η controlling the regularization strength. The first term is equivalent to the smoothness term in the time-varying graph (Eq. (4)), which takes the form of $\frac{1}{2}\|\mathbf{A}_t \circ \mathbf{Z}_t\|_1 = \frac{1}{2} tr(\mathbf{A}_t \mathbf{Z}_t) = tr(\mathbf{X}_t^T \mathbf{L}_t \mathbf{X}_t)$, quantifying the smoothness of signals in the time-varying graph. The function $f(\cdot)$ in the second term serves as a regularization method to avoid obtaining trivial solutions.

For the regularization of the weighted adjacency matrix, we set it as follows:

$$f(\mathbf{A}_t) = -\alpha \mathbf{1}^T \log(\mathbf{A}_t \mathbf{1}) + \beta \|\mathbf{A}_t\|_F^2, \tag{12}$$

where α and β are parameters. The first term in the above regularization ensures that the degree of each node is positive while not preventing edge weights from becoming zero. The second term controls the sparsity of the resulting graph; a lower value leads to a sparser graph. For specific optimization algorithms, please refer to reference [14].

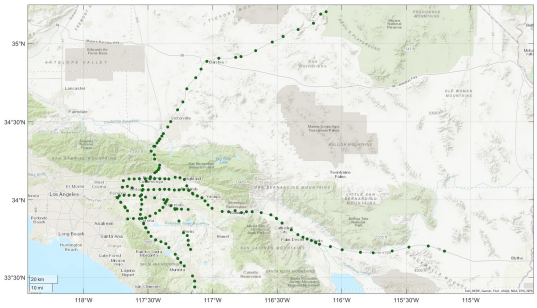


Fig. 2. Sensor maps of PeMS08.

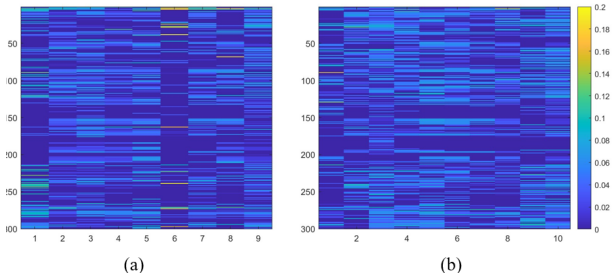


Fig. 3. The time-varying graph learning results for two weekdays. (a) January 3, (b) January 4.

III. Experiments

A. Dataset

The experiment presented in this paper relies on the Caltrans Performance Measurement System (PeMS)[1] to collect real-time and historical traffic data from the highway network across California. The primary data source is vehicle detection stations (VDS) distributed across highways, with sensors recording traffic data for corresponding road segments every five minutes. In this work, we focus on the PeMS08 region, selecting 271 VDS sensors as depicted in Fig. 2. Each data record encapsulates the total traffic volume for the corresponding road segment within a five-minute interval.

B. Anomaly Moment Detection Result Analysis

To evaluate the rationality and effectiveness of our proposed anomaly moment detection method, we conducted a series of small-scale validation experiments.

Firstly, by applying our anomaly moment detection method to the traffic flow data on January 1, 2023 (weekend), we successfully identified five anomaly time points. Specifically, these anomaly moments were: 12:15, 13:35, 15:05, 16:20, and 18:45. The detection results at these time points indicated that the primary variations in weekend traffic flow concentrated in the afternoon and evening hours, which could be intimately linked to people's leisure activities during weekends, such as shopping, dining out, and visiting attractions. These activities typically peak in the afternoon and evening.

Subsequently, we further selected the datasets from January 3, 2023, and January 4, 2023, two weekdays, as experimental samples. In traffic flow data on January 3, we identified eight specific anomaly moments: 07:20, 09:40, 11:15, 12:45, 13:35, 14:50, 17:10, and 18:20. In the traffic flow data on January 4, we identified nine anomaly moments: 07:55, 09:20, 11:10, 12:01, 13:05, 14:35, 17:50, 20:30 and 21:25. These anomaly moments correspond to significant changes in traffic flow, often corresponding to the start or end of peak hours.

As most people's working hours concentrate between 9:00 a.m. and 5:00 p.m., the morning peak usually starts around 7:00 a.m. and lasts until around 9:00 a.m. During this period, people drive or use public transportation to go to work or school, causing significant congestion on roads and public transportation systems. The evening peak starts at 5:00 p.m. and continues into the evening, when most people return home after work or school, again resulting in high use of roads and public transportation. The coincidence between the anomaly moments detected by our method and the actual periods of people's travel activities effectively validated the accuracy of our framework.

C. Adaptive Time-varying Graph Structures Result Analysis

Based on the identified anomaly moments, we further divided the traffic flow data into consecutive subsequences. For example, the traffic flow data on January 1 was divided into five subsequences, while the data on January 3 and January 4 were divided into nine and ten subsequences, respectively. Subsequently, we performed time-varying graph structure learning on these subsequences. Fig. 3 details the time-varying graph learning results for the two weekdays, with each graph exhibiting the first 300 edges of each subsequence's graph structure. Fig. 4 further compares the visualization results of the graph structures of the morning peak and a non-peak period subsequence on the two weekdays, corresponding to the second segment in Fig. 3(a) and the second and first segments in Fig. 3(b).

Observing the results, we can find from Fig. 3 that, although the graph structures learned by each subsequence differ, they maintain a certain degree of smoothness, with only subtle changes between adjacent time periods. In Fig. 4, we can observe that compared to the morning peak period on the same day, the graph structures of nonpeak periods, as shown in Fig. 4(b) and Fig. 4 (c), exhibit significant differences, even though they belong to adjacent time periods. However, the graph structures of the peak periods in the morning on different days of the week, as shown in Fig. 4(a) and Fig. 4 (b), show more prominent similarities. This phenomenon reveals the inherent long-term correlation in traffic flow data.

IV. Conclusions

This paper delves into the realm of urban traffic flow anomaly moment detection and time-varying graph structured modeling, proposing an innovative adaptive time-varying graph learning method based on anomaly detection. The graph structure constructed through this method not only captures

[1] http://pems.dot.ca.gov/

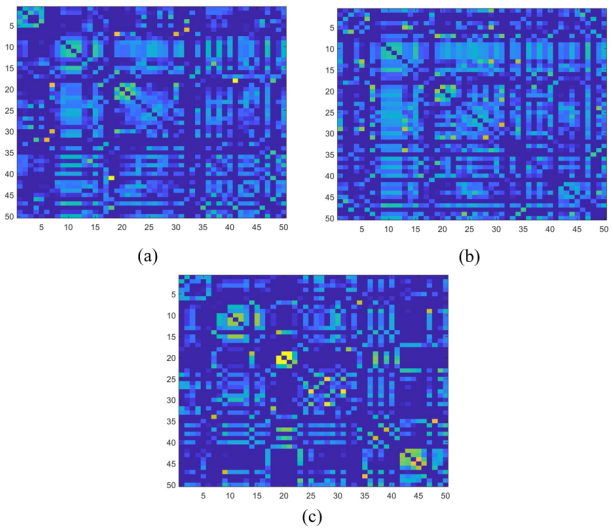


Fig. 4. Visualization results of graph structure. (a) The morning peak on January 3, (b) The morning peak on January 4 (4) The non-peak on January 4.

the spatial interrelationships within the traffic network but also integrates the dynamic characteristics of the temporal dimension. Experimental results demonstrate that our approach not only verifies the accuracy of the anomaly moment detection mechanism but also reveals significant differences in graph structures between peak and off-peak hours of traffic flow, thereby opening up new avenues for the analysis and modeling of traffic flow data.

This research endeavor holds potential for further expansion. Firstly, given the distinct graph-structural characteristics exhibited by traffic flow, we could yield typical graph structures for peak and off-peak traffic periods, providing strong support for related traffic flow research. Secondly, future research could delve deeper into node-level anomaly detection, beyond moment anomalies, aiming to identify and locate specific nodes exhibiting anomalous behaviors within given time periods.

Acknowledgment

This work is supported by the National Nature Science Foundation of China (Grant No: 61901116, Grant No: 61471229 and Grant No:62001115), the Guangdong Basic and Applied Basic Research Foundation (Grant No: 2023A1515140111, Grant No: 2019A1515010789, Grant No: 2021A1515012289 and Grant No: 2019A1515110136) and the Jihua Laboratory Scientific Project (Grant No: X210101UZ210).

References

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, pp. 1–55, 2014.

[2] Y. Tan, H. Xu, and X. Zhang, "Sustainable urbanization in china: A comprehensive literature review," *Cities*, vol. 55, pp. 82–93, 2016.

[3] L. Deng, D. Lian, Z. Huang, and E. Chen, "Graph convolutional adversarial networks for spatiotemporal anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2416–2428, 2022.

[4] W. Yu, M. Huang, S. Wu, and Y. Zhang, "Ensembled masked graph autoencoders for link anomaly detection in a road network considering spatiotemporal features," *Information Sciences*, vol. 622, pp. 456–475, 2023.

[5] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Deep learning on traffic prediction: Methods, analysis, and future directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4927–4943, 2021.

[6] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," *arXiv preprint arXiv:1709.04875*, 2017.

[7] W. Yu, "Discovering frequent movement paths from taxi trajectory data using spatially embedded networks and association rules," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 855–866, 2018.

[8] K. Yamada, Y. Tanaka, and A. Ortega, "Time-varying graph learning with constraints on graph temporal variation," *arXiv preprint arXiv:2001.03346*, 2020.

[9] V. Kalofolias, A. Loukas, D. Thanou, and P. Frossard, "Learning time varying graphs," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Ieee, 2017, pp. 2826–2830.

[10] D. Hallac, Y. Park, S. Boyd, and J. Leskovec, "Network inference via the time-varying graphical lasso," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 205–213.

[11] X. Zhang and Q. Wang, "Time-varying graph learning under structured temporal priors," in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 2141–2145.

[12] K. Qiu, X. Mao, X. Shen, X. Wang, T. Li, and Y. Gu, "Time-varying graph signal reconstruction," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 6, pp. 870–883, 2017.

[13] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE signal processing magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[14] K. Yamada, Y. Tanaka, and A. Ortega, "Time-varying graph learning based on sparseness of temporal variation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 5411–5415.

Unsupervised Anomalous Sound Detection Using Timbral and Human Voice Disorder-Related Acoustic Features

Malik Akbar Hashemi Rafsanjani[1], Candy Olivia Mawalim[2], Dessi Puji Lestari[1], Sakriani Sakti[3], Masashi Unoki[2]
[1] Bandung Institute of Technology, Bandung, Indonesia
E-mail: {13520105@std, dessipuji@staff}.stei.itb.ac.id
[2] Japan Advanced Institute of Science and Technology, Nomi, Japan
E-mail: {candylim, unoki}@jaist.ac.jp
[3] Nara Institute of Science and Technology, Nara, Japan
E-mail: ssakti@is.naist.jp

***Abstract*—Anomalous sound detection (ASD) crucially prevents industrial accidents by distinguishing normal and abnormal machine sounds. Previous research utilizing timbral and short-term features attained a notable F1 score of 0.920. However, relying solely on supervised learning models is impractical due to the difficulty of acquiring anomaly data. This study focuses on developing an unsupervised learning model for ASD, emphasizing prominent timbral features. We also investigate human voice disorder (HVD)-related features, which are potentially linked to human perception of anomalous sounds in machines. We conducted a comparative analysis using 5-fold cross-validation to evaluate our proposed method, with the area under receiver operating characteristic (ROC AUC) as the metric. The proposed ASD method using timbral and HVD-related features significantly improved AUC by 10.87% compared to the baseline system in the DCASE Challenge 2020.**

I. Introduction

Anomalous sound detection (ASD) is the task of recognizing whether the sound emitted from a machine is normal or not [1]. ASD is crucial for early detection, preventing industrial accidents and mechanical failures [1], [2]. These objectives are crucial since industrial machines are complex and prone to malfunctions.

Supervised learning requires labeled data, but accurate anomaly labels are often difficult to obtain [3]. As a result, many ASD methods use unsupervised learning, which trains only on normal data. This approach can handle large amounts of unlabeled data from continuous monitoring systems [4].

ASD studies often use spectrograms and Mel frequency cepstral coefficients (MFCCs) as features [2]. Yet, recent research has explored timbral features, which are associated with human perception of sound differences [5]. Using Support Vector Machines (SVMs), this research achieved an average accuracy of 0.984 using only 7 features. However, this study uses supervised learning which may not be ideal for ASD.

This paper aims to address two issues in earlier studies. First, the performance of utilizing timbral features in unsupervised learning is unknown. Second, research on utilizing human perception-related features, particularly those associated with human voice disorder (HVD), including pathological features, for ASD remains underexplored.

Human perception is highly attuned to subtle variations in sound quality associated with HVD, such as asthenia, roughness, and strain. These variations are often accompanied by changes in specific acoustic features, such as shimmer, jitter, and fundamental frequency (F0) [6]. Since anomalous machine sounds also produce changes in acoustic features, similar features, which capture analogous variations in industrial machine sounds, could be valuable for ASD.

We propose an unsupervised ASD approach using a Gaussian Mixture Model (GMM) with timbral and HVD-related features. GMM is a widely used model for unsupervised anomaly detection and is chosen for its ability to model complex distributions. We hypothesize that both timbral and HVD features could significantly enhance ASD performance.

This paper is part of the ASEAN IVO 2023 project, 'Spoof Detection for Automatic Speaker Verification', which focuses on distinguishing between normal and spoofed sounds. In alignment with this goal, the paper examines methods for improving the detection of normal versus anomalous machine sounds. Furthermore, the project's objective to explore pathological features for spoof detection is reflected here, as pathological features are also investigated for anomalous sound detection.

II. Related Work

A. Anomalous Sound Detection

ASD is the task of identifying whether a sound is normal or anomalous [2]. ASD can be applied in security, network monitoring, and machine maintenance [7]. Unsupervised techniques are very suitable for use in the context of anomaly detection as they only require normal data for training. These approaches are advantageous given the challenges in obtaining accurate labels for anomalous sounds [3]. Additionally, unsupervised methods can handle large amounts of unlabeled data from continuous monitoring systems [4].

ASD studies often use MIMII and ToyADMOS datasets [2], which contain anomalous and normal sounds of machine operations [8], [9]. Inspectors, who manage machine maintenance, rely on their senses, especially hearing, to detect anomalies [5]. This approach drives the usage of timbral

features, which refer to the attribute of sound that allows humans to distinguish among different sound sources [10]. This human-centric approach aligns the technology with practical inspection experience, making it more intuitive and effective.

B. Human Voice Disorder-Related Features

HVD-related features are crucial in various fields, including healthcare and speech therapy. These features encompass a wide range of characteristics related to voice quality and may offer unique insights into sound abnormalities in ASD. The GRBAS scale is a well-known HVD measurement tool [11]. It rates five parameters—grade, roughness, breathiness, asthenia, and strain—on a scale from 0 (normal) to 3 (most severe), correlating well with features like shimmer and jitter.

Studies have shown that shimmer and jitter are closely linked to speech impairments from Parkinson's disease [12], and the standard deviation of F0 and asthenia are negatively correlated [6]. Additionally, jitter local and breathiness are significantly correlated [13].

Acoustic features each represent specific audio properties: jitter indicates fundamental frequency perturbation [12], shimmer shows amplitude perturbation [12], HNR measures the harmonic-to-noise ratio [14], and fundamental frequency (F0) is the smallest true period in the audio [15]. Details of these features and their relation to HVD are summarized in Table I. The following equations express the derivation of HVD-related acoustic features [14], [16], [17]:

$$\text{Shimmer(local)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100, \quad (1)$$

$$\text{Shimmer(local_dB)} = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log(\frac{A_{i+1}}{A_i})|, \quad (2)$$

$$\text{Shimmer(apqL)} = \frac{\frac{100}{N-L+1}\sum_{i=1+\frac{L-1}{2}}^{N-\frac{L-1}{2}}\left|A_i - \frac{1}{L}\sum_{k=i-\frac{L-1}{2}}^{i+\frac{L-1}{2}} A_k\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i}, \quad (3)$$

$$\text{Shimmer(dda)} = 3 \times \text{Shimmer(apq3)}, \quad (4)$$

$$\text{Jitter(local)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100, \quad (5)$$

$$\text{Jitter(local_absolute)} = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|, \quad (6)$$

$$\text{Jitter(ppqL)} = \frac{\frac{100}{N-L+1}\sum_{i=1+\frac{L-1}{2}}^{N-\frac{L-1}{2}}\left|T_i - \frac{1}{L}\sum_{k=i-\frac{L-1}{2}}^{i+\frac{L-1}{2}} T_k\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i}, \quad (7)$$

$$\text{Jitter(rap)} = \text{Jitter(ppq3)}, \quad (8)$$

$$\text{Jitter(ddp)} = 3 \times \text{Jitter(rap)}, \quad (9)$$

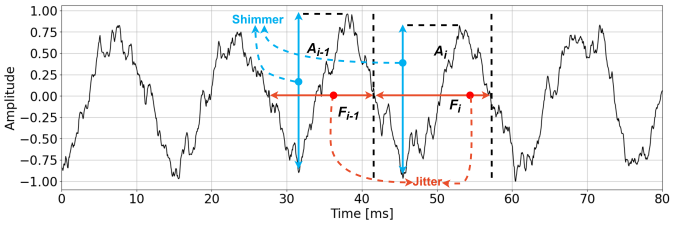


Fig. 1: Jitter and shimmer illustration

TABLE I: Several acoustic features and their relation to HVD

Features	Relation to the HVD
Shimmers	Grade[6], [18], [13], Roughness [6], [13], Breathiness [6], [13], Asthenia [18], [13], laryngeal diseases [19], hoarseness and lesions of larynx [20]
Jitters	Grade[18], Breathiness [6], [18], [13], Asthenia [6], [18], [13], laryngeal diseases [19], hoarseness and lesions of larynx [20]
Fundamental Frequency (F0)	Asthenia [6], Strain [6], background noise-related hearing impairment [21]
Harmonic-to-Noise Ratio (HNR)	Grade[6], [18], Breathiness [6], Asthenia [6], [18], Strain [18]

where A_i, T_i, and N are the i-th amplitude, fundamental period of the input signal of the sound, and the number of samples, respectively. Those features are visualized in Fig. 1.

Additional features include the directional perturbation factor (DPF) and the phonatory frequency range (PFR). DPF indicates vocal fold vibratory perturbation by measuring waveform period changes [22]. PFR is the maximum and minimum F0 that the speaker can produce [23], and is related to the superior laryngeal nerve paresis or paralysis[24]. We also use pitch period entropy (PPE), which is suitable for predicting Parkinson's disease by detecting dysphonia [25].

Another potential HVD-related feature is the harmonic structure of sustained vowels, which has been recognized as an essential feature for voice pathology identification [14]. In that research, the mean and standard deviation of the harmonic structure of the sustained vowels H_p and $RelH_p$, which represent the inverse of the sum of the absolute value of the mean and standard deviation of H_p, respectively, are used as features to classify amyotrophic lateral sclerosis (ALS) patients. Lastly, the pathological vibrato index (PVI) is useful for detecting bulbar dysfunction in ALS patients [26].

III. Proposed Method

The proposed method involves three main steps: feature extraction, model training and optimization, and model testing, as shown in Fig. 2. The data is split into train, validation, and test data. During the training step, the model only uses normal data, leading to higher anomaly scores for anomalous data.

A. Features Extraction

We use the Troparion library [26], a pathological voice analysis tool, to extract HVD-related acoustical features. It uses the instantaneous robust algorithm for pitch tracking (IRAPT) algorithm as a pitch estimation technique. This technique is immune to any pitch modulations and provides more accurate instantaneous pitch values due to time-warping [27].

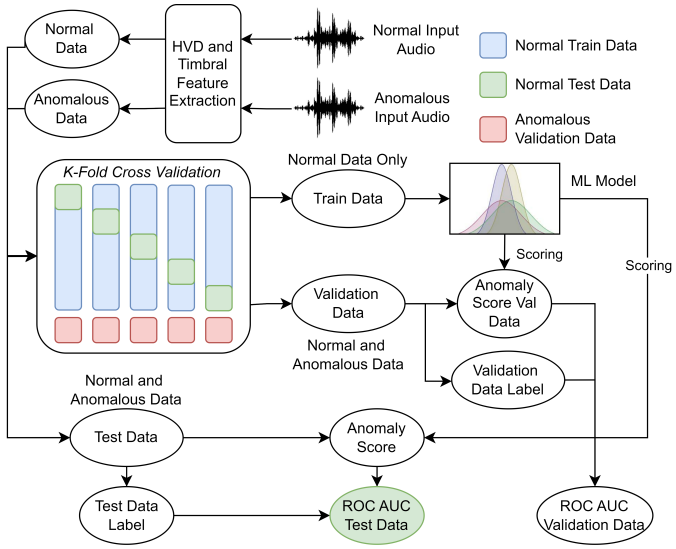


Fig. 2: Proposed method using HVD-related and timbral features for unsupervised anomalous sound detection (ASD)

Additionally, we also implement several features, such as Jitter(ddp), Shimmer(dda), Jitter(local absolute), Shimmer(local dB), and fundamental frequency, given formulas from these studies [14], [16], [17]. In addition, we use the Parselmouth library[1] and Signal Analysis library[2] to calculate two variants of HNR features.

Timbral features are extracted using a model from the Audio Commons project[3] [28]. Using this library, we can extract eight timbral features: hardness, depth, brightness, roughness, warmth, sharpness, booming, and reverberation.

B. Unsupervised Detection Model

We chose a GMM for its effectiveness in probability-based anomaly detection [29]. The GMM, defined by a weighted sum of Gaussian components [30], models complex distributions and identifies data points that deviate from expected patterns. It assumes normal data aligns with the distributions, while anomalies do not fit any clusters. We also trained using a one-class SVM, local outlier factor, and isolation forest, but GMM proved the most suitable, yielding the highest ROC AUC.

C. Model Training

The dataset, which consists of normal and anomaly audios, was extracted to obtain the feature values. These features were divided into training, validation, and test sets. The test set included half of the anomalous data, with an equal amount of normal data. Models were trained solely on normal data and tested on both normal and anomalous data.

The remaining data was split into training and validation sets using K-fold cross-validation ($K = 5$), as suggested by Rodriguez et al. [31], to minimize bias. This method comprehensively evaluates the model's generalizability. The validation set was used for parameter optimization through feature scaling, selection, and hyperparameter tuning.

Three feature groups were evaluated: timbral features, HVD-related acoustic features (HVD features), and their combination. Combining timbral and HVD features involved appending them together. By comparing HVD features with timbral features, we can analyze the capabilities of HVD features for ASD, compared to timbral features, which can be used for ASD. By comparing the combinations of these two groups of features, we can also analyze whether those two groups of features can complement each other.

IV. Experiments

A. Dataset

The Malfunctioning Industrial Machine Investigation and Inspection (MIMII) dataset [8] is a publicly available collection of audio recordings from industrial machines under normal and faulty conditions. This research used the 6 dB SNR subset, chosen for its minimal noise as we are not focused on noise handling. The dataset includes sounds from various machines like valves, pumps, fans, and sliders, with anomalous sounds such as contamination, leakage, rotation imbalance, and rail damage recorded to emulate real scenarios.

B. Evaluation Metrics

The model's performance was evaluated using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The ROC curve depicts the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various thresholds. ROC AUC is a global measure of a test's ability to discriminate whether a specific condition is present or not [32]. In this context, the AUC provides a single metric summarizing the model's ability to discriminate between normal and anomalous sounds, with higher AUC values indicating better performance. The ROC AUC metric formula can be expressed as

$$\mathrm{AUC}_{m,n,d} = \frac{1}{N_d^- N_n^+} \sum_{i=1}^{N_d^-} \sum_{j=1}^{N_n^+} H(A_\theta(x_\mathrm{j}^+) - A_\theta(x_\mathrm{i}^-)) \quad (10)$$

where m represents the index of a machine type, n represents the index of a section, and $d = \{source, target\}$ represents a domain. The N_d^- is the number of normal test clips in domain d, N_n^- is the number of normal test clips in section n, and N_n^+ is the number of anomalous test clips in section n. Here, $\{x_i^-\}_{i=1}^{N_d^-}$ are normal test clips in domain d in section n and $\{x_j^+\}_{j=1}^{N_n^+}$ are anomalous test clips in section n in machine type m. $H(x)$ is the hard-threshold function that returns 1 when $x > 0$ and 0 otherwise [33].

C. Model Optimization

We used feature scaling, feature selection, and hyperparameter tuning for the optimization because the features extracted are numerical data and contain no NaN values. We evaluated the models' performance using validation data to determine the best technique. For feature scaling, we tried four techniques, with standard scaling producing the best ROC AUC.

[1]https://parselmouth.readthedocs.io/en/stable/
[2]https://pypi.org/project/Signal_Analysis/
[3]https://github.com/AudioCommons/timbral_models

TABLE II: Selected features after feature selection

Machine Type	Timbral Features	HVD Features
Fan	boominess, roughness, warmth	Jitter PPQ5, Jitter absolute, Jitter local, PPE, PPF, Praat.HNR, RelHp, Shimmer local dB, Signal_Analysis.HNR, mean Hp, stdev F0
Pump	brightness, reverb, sharpness, warmth	DPF, Jitter RAP, Jitter local, PPE, PPF, Praat.HNR, RelHp, Shimmer APQ55, Shimmer local dB, mean F0, stdev F0, stdev Hp
Slider	brightness, depth, reverb, roughness, warmth	Jitter PPQ55, Praat.HNR, RelHp, Shimmer APQ11, Shimmer APQ5, Shimmer APQ55, Shimmer local dB, Signal_Analysis.HNR
Valve	brightness, hardness, roughness, sharpness, warmth	Jitter PPQ55, PPE, PPF, PVI, Praat.HNR, Shimmer APQ11, Shimmer APQ3, Shimmer APQ5, Shimmer APQ55, Shimmer local dB, Signal_Analysis.HNR, mean F0, stdev F0

To reduce redundant features and identify beneficial ones, we used feature selection. We utilized the sequential feature search (SFS) technique due to its low computational requirements. We determined the granularity of the feature selection on the machine type level, to incorporate shared characteristics of each machine type and prevent overfitting on the machine ID or model level. Selected features after feature selection are shown in Table II.

For hyperparameter tuning, we used grid search on various parameters, including the number of components, covariance type, convergence threshold, max iteration, and parameter initializer. We found the best combination for the timbral features are 3 mixture components, covariance type full, convergence threshold 0.001, max iteration 100, and K-means as a parameter initializer. We used 5 mixture components, covariance type full, convergence threshold 0.01, max iteration 300, and randomly selected data points as the parameter initializer for HVD features. We used the same parameter as timbral features for the combination of both features.

D. Baseline Model

The ASD model by Morita et al. [34] was used as a baseline model because it uses GMM and the MIMII dataset, the same model and dataset that we use. This ASD model extracts log-mel spectrogram features from the audio input and reduces the dimension using PCA. The output of PCA is aggregated using the mean or variance of the output and then is used as the model input. We chose the GMM and variance methods to be replicated as the model produces better ROC AUC than other combinations.

V. Results

A. Proposed Features Produce Excellent ROC AUC

The overall results are shown in Table III. The results show the performance evaluation of different machine types (fan, pump, slider, and valve) based on three different feature sets: timbral features, HVD features, and combined features. The replicated baseline results are also shown. Both feature sets performed well individually, but their effectiveness varied across different machine types and IDs. On average, these two features were comparable for valves, but HVD features excelled for sliders, while timbral features excelled for fans and pumps. The aggregate average performances across all machine types—0.921 for timbral and 0.912 for HVD—highlight a marginal and slight advantage for the timbral features in detecting anomalous sound.

TABLE III: ROC AUCs for each machine type with GMM and using timbral and HVD features, compared to baseline

Machine Type	Machine ID	Replicated Baseline [34]	Timbral Features	HVD Features	Combined Features
Fan	00	0.768	**0.903**	0.769	0.879
Fan	02	0.854	0.978	0.973	**0.981**
Fan	04	0.754	0.947	0.967	**0.981**
Fan	06	0.847	**0.999**	0.994	0.994
Fan	Avg	0.806	0.957	0.926	**0.959**
Pump	00	0.951	**0.967**	0.960	0.963
Pump	02	**0.917**	0.839	0.891	0.852
Pump	04	0.758	0.948	0.938	**0.978**
Pump	06	0.825	**0.990**	0.847	0.950
Pump	Avg	0.863	**0.936**	0.909	0.936
Slider	00	0.999	0.977	0.994	**1.000**
Slider	02	0.930	0.960	0.983	**0.998**
Slider	04	0.759	0.908	**0.969**	0.962
Slider	06	0.840	0.660	0.680	**0.707**
Slider	Avg	0.882	0.876	0.907	**0.917**
Valve	00	0.933	**1.000**	0.942	**1.000**
Valve	02	0.786	**0.980**	0.912	0.976
Valve	04	0.840	**0.968**	0.957	0.964
Valve	06	0.775	0.717	0.814	**0.820**
Valve	Avg	0.833	0.916	0.906	**0.940**
All	Avg	0.846	0.921	0.912	**0.938**

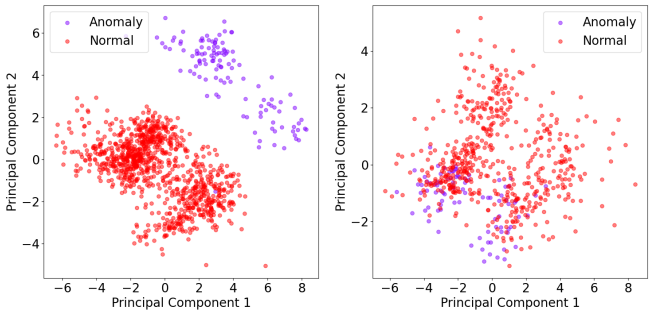
Moreover, the data indicate a notable variance in the performance of individual machine IDs within each machine type, implying that the effectiveness of the feature sets might be influenced by specific machine characteristics. The valve type, for example, performed optimally with timbral features for ID 00 but performed significantly worse for valve ID 06, possibly due to inherent differences in the machine IDs, as different IDs may come from different product models.

Overall, the fact that combined features were better than timbral features and HVD features for every average score of machine types and the aggregated averages across all machine types (0.921 for timbral, 0.912 for HVD, and 0.938 for combined features) demonstrate that while each feature set individually offers substantial discriminative power, the combined features consistently yielded better performance. This result suggests that the integration of multiple feature sets significantly enhanced the anomaly detection capability.

We also compare the result with a recent unsupervised ASD study. A newly proposed unsupervised method, Denoising Sparse Wavelet Network (DeSpaWN), achieves a very good performance for ASD [35]. DeSpaWN is a deep-learning architecture designed to perform denoising and feature extraction from high-frequency time-series signals. It produces an AUC of 0.946 with the 6 dB MIMII dataset. This study's proposed method gives almost similar results using the combined features, which is 0.938. Although only GMM, a shallow learning model, is used, the proposed method produces a very comparable result to that of deep learning. This finding suggests the proposed method gives excellent results, comparable with state-of-the-art methods, while only using shallow learning.

TABLE IV: Average F1 score for each machine type

Machine Type	Timbral Features	HVD Features	Combined Features
Fan	0.901	0.875	**0.905**
Pump	**0.871**	0.825	0.868
Slider	0.835	0.856	**0.889**
Valve	**0.875**	0.812	0.865
All	0.870	0.842	**0.882**



(a) Valve ID 00 (b) Slider ID 06

Fig. 3: Plot of PCA results on the best and the worst performing machine ID using combined features.

B. Operational Prediction Performance

To evaluate the effectiveness of the proposed feature in real-world situations, particularly in predicting whether data is normal or anomalous, the F1 score of the models was tested. This involved finding the optimal threshold for classifying anomaly scores using validation data. The granularity of the threshold was specific to each machine type and model due to their shared characteristics.

The results of the average F1 score of each machine type using the three groups of features are shown in Table IV. Although the F1 score is lower than the ROC AUC score due to the exact threshold from the validation data, the results are still very good. The models produced an F1 score of 0.870 using timbral features, 0.842 using HVD features, and 0.882 using combined features. These results have the same patterns as the ROC AUC results: the best result was achieved by the combined features and the timbral features result was better than the HVD features result. From these findings, we can conclude that the proposed approach can perform very well in real-world situations.

C. PCA Visualization

To understand the varying performance observed across different machine IDs, we compared the visualization of the PCA of scaled combined features data with the standard scaler. We visualize the best (valve ID 00, AUC: 1.000) and the worst (slider ID 06, AUC: 0.707) machine ID results in Fig 3. For valve ID 00, there is a clear separation between the normal (red) and anomalous (purple) data, whereas for slider ID 06, a significant overlap is observed between the two. These results suggest that the varying performance stems from the model's difficulty in distinguishing anomalies in certain machine IDs, where normal and anomalous data exhibit greater characteristic overlap compared to others.

TABLE V: Ablation study results

Feature Group	Relative ROC AUC Score	Features
Shimmer	-0.050	Shimmer.APQ11, Shimmer.APQ3, Shimmer.APQ5, Shimmer.APQ55, Shimmer.local, Shimmer.DDA, Shimmer.local_db
Harmonicity	-0.033	Praat.HNR, Signal_Analysis,HNR
Pathology	-0.024	DPF, mean Hp, stdev Hp, RelHp, PPE, PPF, PVI
F0	-0.003	mean F0, stdev F0
Jitter	-0.003	Jitter PPQ55, Jitter PPQ5, Jitter RAP, Jitter local, Jitter DDP, Jitter absolute

D. Ablation Study

We conducted an ablation study to assess the importance of each newly proposed HVD feature. By grouping related features, removing each group, and calculating the ROC AUC, we determined how critical each group was. The results when no features were excluded were then subtracted from the results. The more negative the result, the more important the group of features is. We conducted the ablation study using the HVD features, and the results are presented in Table V. The findings suggest that the most crucial feature groups are Shimmer, Harmonicity, and Pathology.

Removing the shimmer group resulted in a -0.050 decrease in ROC AUC, indicating that amplitude perturbation can play a crucial role in ASD. Despite the harmonicity group having only two features, its removal caused a -0.033 decrease in ROC AUC, showing the importance of the HNR feature in classifying anomalous sound. In contrast, removing F0 (fundamental frequency) and Jitter, which represents perturbation in fundamental frequency, groups only resulted in a minimal decrease of ROC AUC, indicating modest contributions to ASD performance. These findings suggest that the amplitude and harmonicity aspects are more useful than the frequency aspect for classifying machine sound anomalies. In addition, the pathology group that contains a lot of features that directly correlate to HVD also produced a high decrease in ROC AUC when removed, strengthening our hypothesis that features related to HVD enhance and are beneficial to ASD performance.

VI. Conclusion and Future Work

This study explored human voice disorder (HVD)-related features for unsupervised anomalous sound detection (ASD). Our experimental results showed the feasibility of using HVD-related features for ASD, with an average area under the curve (AUC) of 0.912, as these features are comparable to the timbral features that can be used to detect anomalous sound, with an average AUC of 0.921. The results also highlight the enhanced detection capability achieved by combining HVD-related features with timbral features with an average AUC of 0.938. This work contributes to the development of more robust and efficient unsupervised ASD methods. Future work could investigate the application of different algorithms for estimating the fundamental frequency (F0), such as the YIN and SWIPE algorithms. Utilizing these alternative algorithms may yield more robust HVD-related acoustic features and uncover new aspects of sound anomalies that were not captured by the IRAPT algorithm.

ACKNOWLEDGMENTS

This work was supported by the JST Sakura Science Exchange Program and JSPS KAKENHI grant (23H04344, 23K18491, 22K21304). The ASEAN IVO (http://www.nict.go.jp/en/asean ivo/index.html) project, 'Spoof Detection for Automatic Speaker Verification', was involved in the production of the contents of this publication and financially supported by NICT (http://www.nict.go.jp/en/index.html).

REFERENCES

[1] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma," in *25th European Signal Processing Conference, EUSIPCO 2017, Kos, Greece, August 28 - September 2, 2017*. IEEE, 2017, pp. 698–702.

[2] E. C. Nunes, "Anomalous sound detection with machine learning: A systematic review," *CoRR*, vol. abs/2102.07820, 2021.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, 2009.

[4] G. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2168–2187, 2022.

[5] Y. Ota and M. Unoki, "Anomalous sound detection for industrial machines using acoustical features related to timbral metrics," *IEEE Access*, vol. 11, pp. 70 884–70 897, 2023.

[6] S. V. Freitas, P. M. Pestana, V. Almeida, and A. Ferreira, "Integrating voice evaluation: correlation between acoustic and audio-perceptual measures," *Journal of Voice*, vol. 29, no. 3, pp. 390–e1, 2015.

[7] E. D. Fiore, A. Ferraro, A. Galli, V. Moscato, and G. Sperlì, "An anomalous sound detection methodology for predictive maintenance," *Expert Syst. Appl.*, vol. 209, p. 118324, 2022.

[8] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," pp. 209–213, 2019.

[9] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2019, New Paltz, NY, USA, October 20-23, 2019*. IEEE, 2019, pp. 313–317.

[10] K. Patil, D. Pressnitzer, S. A. Shamma, and M. Elhilali, "Music in our ears: The biological bases of musical timbre perception," *PLoS Comput. Biol.*, vol. 8, no. 11, 2012.

[11] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldán, "Automatic assessment of voice quality according to the GRBAS scale," in *28th International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2006, New York City, NY, USA, August 30 - September 3, 2006, Main Volume*. IEEE, 2006, pp. 2478–2481.

[12] S. S. Upadhya, A. Cheeran, and J. Nirmal, "Statistical comparison of Jitter and Shimmer voice features for healthy and Parkinson affected persons," in *2017 second international conference on electrical, computer and communication technologies (ICECCT)*. IEEE, 2017, pp. 1–6.

[13] B. Sabir, B. Touri, and M. Moussetad, "Correlation between acoustic measures, voice handicap index and GRBAS scales scores among Moroccan students." *Current Pediatric Research*, vol. 21, pp. 343–353, 04 2017.

[14] M. Vashkevich and Y. Rushkevich, "Classification of ALS patients based on acoustic analysis of sustained vowel phonations," *Biomed. Signal Process. Control.*, vol. 65, p. 102350, 2021.

[15] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[16] J. P. Teixeira and A. Gonçalves, "Algorithm for jitter and shimmer measurement in pathologic voices," *Procedia Computer Science*, vol. 100, pp. 271–279, 2016.

[17] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[18] R. Fujiki and S. Thibeault, "Examining relationships between GRBAS ratings and acoustic, aerodynamic and patient-reported voice measures in adults with voice disorders," *Journal of Voice*, vol. 37, 03 2021.

[19] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Automated speech analysis applied to laryngeal disease categorization," *Comput. Methods Programs Biomed.*, vol. 91, no. 1, pp. 36–47, 2008.

[20] H. Lathadevi and S. P. Guggarigoudar, "Objective acoustic analysis and comparison of normal and abnormal voices." *Journal of Clinical & Diagnostic Research*, vol. 12, no. 12, 2018.

[21] C. A. Brown and S. P. Bacon, "Fundamental frequency and speech intelligibility in background noise," *Hearing research*, vol. 266, no. 1-2, pp. 52–59, 2010.

[22] M. B. Higgins and J. H. Saxman, "A comparison of intrasubject variation across sessions of three vocal frequency perturbation indices," *The Journal of the Acoustical Society of America*, vol. 86, no. 3, pp. 911–916, 1989.

[23] G. J. Cler, V. S. McKenna, K. L. Dahl, and C. E. Stepp, "Longitudinal case study of transgender voice changes under testosterone hormone therapy," *Journal of Voice*, vol. 34, no. 5, pp. 748–762, 2020.

[24] C. A. Eckley, R. T. Sataloff, M. Hawkshaw, J. R. Spiegel, and S. Mandel, "Voice range in superior laryngealnerve paresis and paralysis," *Journal of Voice*, vol. 12, no. 3, pp. 340–348, 1998.

[25] M. A. Little, P. E. McSharry, E. J. Hunter, J. L. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, 2009.

[26] M. Vashkevich, A. Petrovsky, and Y. Rushkevich, "Bulbar ALS detection based on analysis of voice perturbation and vibrato," in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications, SPA 2019, Poznan, Poland, September 18-20, 2019*. IEEE, 2019, pp. 267–272.

[27] E. Azarov, M. Vashkevich, and A. A. Petrovsky, "Instantaneous pitch estimation based on RAPT framework," in *Proceedings of the 20th European Signal Processing Conference, EUSIPCO 2012, Bucharest, Romania, August 27-31, 2012*. IEEE, 2012, pp. 2787–2791.

[28] A. Pearce, B. Tim, and M. Russell, "Release of timbral characterisation tools for semantically annotating non-musical content," 2020.

[29] R. Blanco, P. Malagón, S. Briongos, and J. M. Moya, "Anomaly detection using gaussian mixture probability model to implement intrusion detection system," in *Hybrid Artificial Intelligent Systems: 14th International Conference, HAIS 2019, León, Spain, September 4–6, 2019, Proceedings 14*. Springer, 2019, pp. 648–659.

[30] D. A. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics, Second Edition*, S. Z. Li and A. K. Jain, Eds. Springer US, 2015, pp. 827–832.

[31] J. D. Rodríguez, A. P. Martínez, and J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 569–575, 2010.

[32] Z. H. Hoo, J. Candlish, and D. Teare, "What is an ROC curve?" pp. 357–359, 2017.

[33] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *CoRR*, vol. abs/2305.07828, 2023.

[34] K. Morita, T. Yano, and K. Q. Tran, "Anomalous sound detection by using local outlier factor and gaussian mixture model," in *Proceedings of the 5th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), Tokyo, Japan*, 2020, pp. 2–4.

[35] G. Michau, G. Frusque, and O. Fink, "Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series," *Proceedings of the National Academy of Sciences*, vol. 119, no. 8, Feb. 2022. [Online]. Available: http://dx.doi.org/10.1073/pnas.2106598119

Tsnake: A Time-Embedded Recurrent Contour-Based Instance Segmentation Model

Chen-Jui Hsu[1], Jian-Jiun Ding [2*], and Chun-Jen Shih[3]
[1] National Taiwan University, Taiwan
E-mail: r11942085@ntu.edu.tw Tel: +886-233663662
[2*] National Taiwan University, Taiwan
E-mail: jjding@ntu.edu.tw Tel: +886-233669652
[3] National Taiwan University, Taiwan
E-mail: f10942045@ntu.edu.tw Tel: +886-233663662

***Abstract* — Instance segmentation is a critical task in computer vision. It is widely applied in autonomous driving and medical image analysis. Contour-based methods have been studied for their elegant structures, compact representation, and low inference time. In existing work, contour-based methods apply iterative reformation to fit the object boundary, however, the modules are independent in each loop, which enlarges the model sizes. In this paper, we introduce the TSnake. It is a novel time-embedded driven recurrent contour-based instance segmentation model. At each iteration, the refinement module adjusts the contour with the help of semantic-rich features and the time step. Moreover, a dilated module is applied to well adopt the local contextual information. We also modify the loss function to well balance the trade-off between over-smoothing and over-sensitivity of the contour. The proposed TSnake method surpasses the contour-based state-of-the-art instance segmentation and can be performed in real time.**

I. Introduction

Instance segmentation, which aims to detect and delineate each distinct object instance in an image, is a fundamental yet challenging task in computer vision. It plays a crucial role in various applications, including autonomous driving robotics, medical imaging, and augmented reality.

Most approaches to instance segmentation can be broadly categorized into two main types: pixel-based and contour-based. Pixel-based methods predict the masks of the objects directly [1-5]. Some of them perform a two-stage pipeline, in which the first stage detects the bounding boxes of objects and the latter predicts the masks, to achieve higher accuracy. These designs cause intensive computation and slow inference time. Meanwhile, some directly predict masks from the backbone to reach real-time speed.

As for contour-based models [6-11], they predict several points or coefficients and form them into a polygon to approach the boundary of the objects. For example, Xie *et al.* [6] proposed PolarMask which predicts 36 segments started from the object center uniformly to generate a polygon, and Xu *et al.* [7] applied Chebyshev polynomials to fit the shape. Contour-based methods only predict distinct points or coefficients, thus the computation costs are usually less than pixel-based methods.

Inspired by the active contour model [11], in [8-10], they proposed iterative refinement pipelines to perform instance segmentation. They all apply a module with several layers of circular convolution, or "snake", to refine the local segment to approach the boundary of the object. These recurrent contour-based models treat object segmentation as a regression problem. However, existing contour-based methods still face several challenges, including over-smoothed or over-sensitive contours, redundant information during iteration and limited receptive fields, etc.

In this paper, we introduce TSnake, a recurrent contour-based instance segmentation model to address these limitations. Our contributions includes:

1. A time-embedding controlled snake module that enables iterative refinement without additional information while maintaining weight sharing is proposed.
2. A dilated module that expands the receptive field is proposed. It can well apply local contextual information.
3. A weighted dynamic matching loss is proposed. It well balances the trade-off between over-smoothing and over-sensitivity of the contour.

Our experiments on the SBD and Cityscapes datasets demonstrate that the proposed TSnake achieves state-of-the-art performance among contour-based instance segmentation methods while maintaining real-time efficiency.

II. Related Work

A. Pixel-based Instance Segmentation

Pixel-based instance segmentation methods can be classified into: two-stage methods and one-stage methods. Two-stage methods basically follow the pipeline of the Mask RCNN [1]. In the first stage, they predict several bounding boxes that probably contain objects. The second stage is for classification. In [2, 12], they applied cascade prediction series to improve masks iteratively. Shen *et al.* [3] applied the DCT to reduce the loss from resizing.

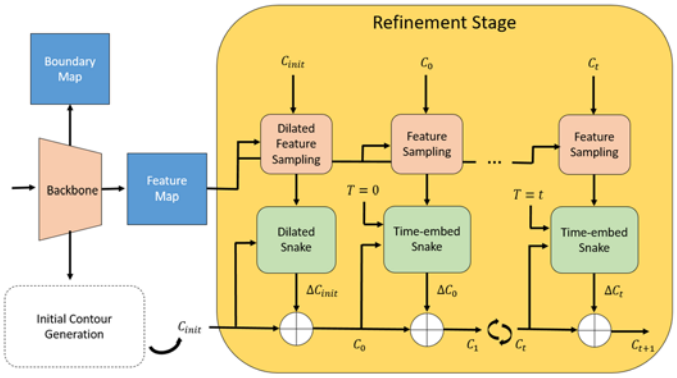


Fig. 1 The overall structure of TSnake. For an image passed to the model, initial contour generation will predict C_{init} and pass it to the dilated snake to generate C_0. C_0 will be deformed for T times by the weight sharing time-embed snake.

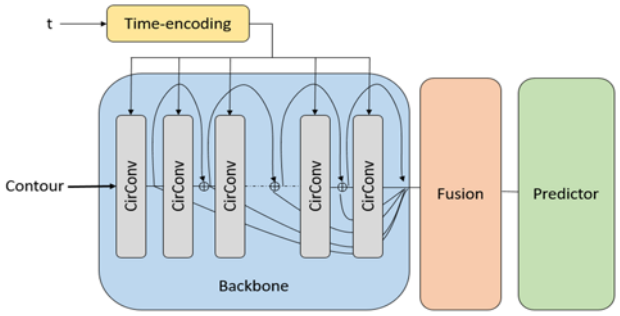


Fig. 2 Structure of the time-embed snake. The sequence t will be encoded by sinusoid functions and added to all circular convolution in the snake backbone.

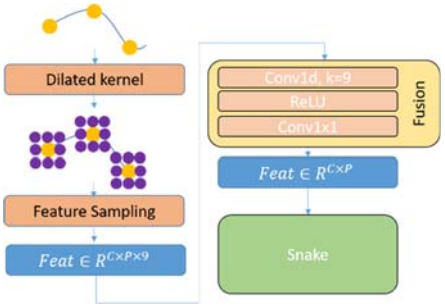


Fig. 3 Pipeline of the dilated snake. The dilated kernel fetches the samples around the points in the coordinate system of the origin image.

By contrast, one-stage methods directly predict masks from the backbone, making them usually have low inference time. Bolya *et al.* [4] predicted plenty of protomasks to present each mask of an object. Wang *et al.* [5] separated the origin image into N^2 grids and predicted a mask for each grid. These methods generally offer faster inference times but may struggle with fine-grained details and high memory usage.

B. Contour-based Instance Segmentation

Contour-based methods focus on object boundaries rather than full masks. They transform instance segmentation into a regression problem, predicting points or coefficients to form polygons. Xu *et al.* [7] applied Chebyshev polynomials and Wang *et al.* [13] predicted the coefficients of the Fourier series to generate results. In another way, directly predicting the coordinates of the points makes the predicted contour much fitter. Xie *et al.* [6] divided 360 degrees into 36 pieces uniformly to predict surrounding points. Peng *et al.* [8] proposed Deep Snake to iteratively refine the contour. DANCE [14] followed Deep Snake and added an attention mechanism to fetch more edge information.

Also, Zhang *et al.* [9] modified initial contour generation of [8] to learning-based to achieve better performance, and Feng *et al.* [10] decreased the model weights by an RNN-like weight-sharing refinement stage. However, most of them do not share the weights for iteration. Also, the contours of the above recurrent instance segmentation models are sometimes too over-smooth to enclose the extreme points, or too sensitive to noise which makes the contour jagged.

III. Proposed Method

A. Time-Embedded Snake

We address two issues from previous work. The first one is the contours are over-smooth or over-sensitive, and the second is the weight-sharing module is needed while the information from the previous state is unnecessary.

We propose TSnake, a time-embedding controlled weight-sharing recurrent refinement model that makes a balance between too smooth and too sharp. The overall pipeline is shown in Fig. 1. Our TSnake consists of three main components: a backbone network for extraction, an initial contour generation module inherent from [9], and a refinement stage which is the core innovation of Tsnake. When the image passes into the backbone, the initial contour generation produces an initial contour based on the backbone features, and the contour will be refined to a fit boundary at the refinement stage. Below we will introduce the components of Tsnake and the proposed loss.

Though a weight-sharing scheme has been proposed in [10] in the GRU style, we assume that the features from the previous iteration are not necessary. However, the model still needs some information to tell it which iteration is to do corresponding refinement. Inspired by the diffusion model [15], the diffusion model is a generative model that denoise the Gaussian noise to a clear image with several loops. At each loop the model denoise different levels of noise. The time encoding is added to tell it which iteration is now. We treat the ground truth contour as a clear image, and the initial contour will be like a noisy image. We plugin time embedding into the snake module at the circular convolution layers, which is shown in Fig. 2. We adopt sinusoid functions for time encoding as the same as [15]. The contour deformation can be written as in (1):

$$x^{(t+1)} = snake_T\left(\left[F\left(x^{(t)}\right);\bar{x}^{(t)}\right], t_{enc}\left(t\right)\right) + x^{(t)} \qquad (1)$$

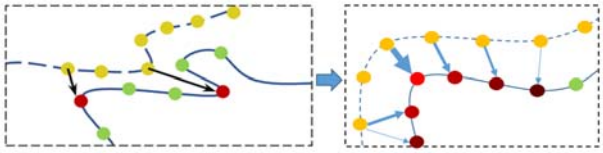


Fig. 4 The representation of the weighted DML. The yellow points are predicted points, red ones are keypoint ground truths, and green ones are the normal ground truth. The left subfigure is the original DML [9] and right one is the proposed weighted DML. The origin DML only attracts one point to the keypoint, which makes the contour sensitive. Our approach builds the corner with much more points to reduce the sensitivity.

where $x^{(t)}$ is the contour coordinates at the tth interation, F is the features from the backbone with 64 channels, and $\bar{x}$ is the translated coordinates that subtract the min value of each dimension. The features from the backbone and the coordinates will then be concatenated together and fed to the module. $snake_T$ is the time-embedded module. $t_{enc}()$ is a function to encode the time step. The iteration loop is set to 8 in our setting.

B. Dilated Snake

We have observed that if the initial contour located at a relatively smooth area, points will be easily misled to the wrong side. Intuitively, we take more features from the neighborhood for a point to decide where is the correct side. In order to not increasing the calculation too much, we employ dilated convolution [16] with kernel size 3. The structure is shown in Fig. 3. Note that the dilated points are in the coordinate system of the origin image. By the dilated kernel, each point will get 9 feature vectors. We fuse these 9 vectors to one vector by Conv1d and pass it to the original snake module. Dilated snake is not only used for deformation with nearby but also to increase the receptive field so that 8N more features can be updated and propagated back to the backbone when training.

C. Weighted DML

In previous works, the contours are usually over-smooth, but E2EC [9] is the one that makes the contour jagged. This is caused by the dynamic matching loss (DML) [9]. The jagged problem occurred because DML only attracts one point to the keypoint. Thus, we assign the adjacent points of keypoints as new keypoints but with different weights. Those new keypoints can attract predicted points as well but with weak forces. The weights are reweighted with Gaussian distribution, where the original keypoints are with the weight 1. The Gaussian kernel of distribution is set as:

$$\exp((x-\text{int}(\mu/2))^2/2\sigma^2). \tag{2}$$

Eq. (2) is sampled in the discrete form and μ and σ are set as 7 and 0.5, respectively. The loss of our weighted DML is:

$$x_i^* = arg\min_x \left\| pred_i^{in} - gt_x^{ipt} \right\|_2 \tag{3}$$

$$L_1 = \frac{1}{N}\sum_{i=1}^{N} SL_1\left(pred_i^{out}, gt_{x_i^*}^{ipt}\right) \tag{4}$$

$$y_i^* = arg\min_y \left\| pred_y^{in} - gt_i \right\|_2 \tag{5}$$

$$L_2 = \frac{1}{sum(K)}\sum_{i=1}^{N} K[i]\times SL_1\left(pred_{y_i^*}^{out}, gt_i\right) \tag{6}$$

$$L_{WDML}\left(pred^{in}, pred^{out}, gt\right) = \frac{L_1+L_2}{2} \tag{7}$$

where SL_1 means the smoothed L_1 loss, *pred* and *gt* means the predicted result and the ground truth, respectively. The part that attract predicted points to the nearest contour remains the same in [9], which is shown in (3) and (4).

D. Loss Function

The loss function is a combination of detection, initial contour, and refinement. Also, a boundary map is adopted to predict to enhance the performance. The losses are shown as follows:

$$L = L_{det} + L_{bd} + L_{init} + L_{refine}, \tag{8}$$

$$\begin{aligned} L_{refine} = &\sum_{k=0}^{4}\lambda^{K-k}\times SL_1\left(x^{(k)}, x\right) \\ &+\sum_{k=5}^{8}\lambda^{K-k}\times wdml\left(x^{(k-1)}, x^{(k)}, x\right)+\sum_{k=0}^{8}\lambda^{K-k}\times shape\left(x^{(k)}, x\right) \end{aligned} \tag{9}$$

where L_{det} is the detection loss, L_{bd} is the boundary loss that applies focal loss [17], L_{init} is the initial contour loss applying the smooth-L_1 loss, λ is a constant less than 1, which we set as 0.8 in experiments, and *wdml* means the weighted DML. L_{refine} is the refinement loss that consists of 3 parts. Also, contours at each iteration are applied to shape loss in [10]. The first four iterations and the contour from the dilated snake are applied smooth-L_1 loss and our weighted DML is used for the latter four iterations.

IV. Experiments and Results

A. Datasets and Metrics

Datasets. The SBD and Cityscapes datasets [18] were used in experiments. The SBD dataset has 20 object classes and 11355 images in total from the PASCAL VOC [19] dataset and is split into a training set for 5623 images and a validation set for 5732 images. Cityscapes has 8 classes that usually appear on the road, and it contains 2975 training images and 500 validation images.

Metrics. We evaluate the mask quality by standard AP. For SBD, we report the performance based on metrics 2010 VOC AP_{vol}, AP_{50}, and AP_{70}. AP_{vol} is the average of AP with nine thresholds from 0.1 to 0.9. For Cityscapes, the results are evaluated in terms of AP metric averaged by 8 categories.

TABLE I: THE RESULT ON THE SBD DATASET WHERE * MEANS THAT MOSAIC AUGMENTATION IS APPLIED.

Name	AP_{vol}	AP_{50}	AP_{70}
ESE-20 [7]	35.3	40.7	12.1
Deep Snake [8]	54.4	62.1	48.3
DANCE [14]	56.2	63.6	50.4
E2EC [9]	59.2	65.8	54.5
TSnake	60.2	66.8	55.5
TSnake*	**60.5**	**67.2**	**56.1**

TABLE II: THE RESULT ON THE CITYSCAPES DATASET WHERE * MEANS THAT MULTI-COMPONENT DETECTION IS NOT APPLIED.

Name	AP
DANCE* [14]	36.7
Deep Snake [8]	37.4
E2EC [9]	39.0
TSnake	**39.5**

TABLE III: ABLATION STUDY OF THE TIME-EMBED SNAKE AND THE DILATED SNAKE. THE BOLD STYLE MEANS THE FIRST PLACE IN EACH EXPERIMENT.

Exp.	Method	AP_{vol}	AP_{50}	AP_{70}
Time - Embed Snake	w/o	58.9	65.9	54.8
	All layers	59.6	66.6	54.8
	backbone	**60.2**	**66.8**	**55.3**
Dilated Snake	w/o	59.9	66.2	55.1
	w/ Snake	59.9	66.5	55.0
	w/ Dilated	**60.2**	**66.8**	**55.3**
	Kernel=3	60.0	66.4	**55.4**
	Kernel=5	**60.2**	**66.8**	55.3

TABLE IV: ABLATION STUDY ON LOSS FUNCTIONS.

Experiment	Method	AP_{vol}	AP_{50}	AP_{70}
Loss Function	smooth L_1 loss	59.0	65.9	54.5
	DML	59.3	66.0	55.0
	Weighted DML	**60.2**	**66.8**	**55.3**
Weighted DML	After the 2nd	58.5	65.2	53.8
	After the 4th	**60.2**	**66.8**	**55.3**
	After the 6th	58.7	65.2	53.8

B. Implement Details

Fixed setting. We trained our models on Intel i9-9900K, one Nvidia RTX 4080, with Python 3.7, PyTorch 1.11, and cuda 11.3 version.

Backbone. For fair comparison to previous works, the detection backbone is CenterNet [20] with DLA34 [21] +DCNv2 [22].

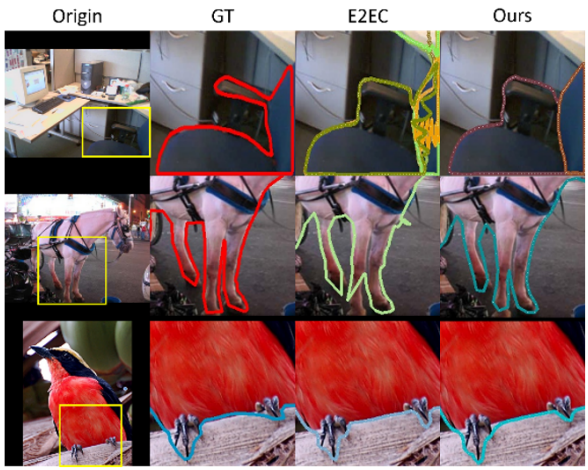


Fig. 5 From left to right are original images, ground truths, the results of E2EC [9] and our proposed method, respectively. Our method can provide contours that enclose the extreme points and not over-smoother than others.

Training setting. For SBD, TSnake was trained for 200 epochs with Adam optimizer. The learning rate starts from 1e-4 and decays based on StepLR at epochs 60, 100, 130, 155, 180. The batch size is set as 20. Mosaic augmentation [23] is adopted for the first 150 epochs, and the ratio of mosaic and original images is 3:2. The training and testing images are all 512x512.

For Cityscapes, multi-component detection [8] is applied for the separated objects. The optimizer and scheduler are the same as those in SBD but the learning rate decays at epochs 70, 110, 150, and 190.

C. Results

SBD. Our result of SBD is shown in Table I. TSnake outperformed the state-of-the-art (SOTA) contour-based method. It is worth mentioning that if we remove the dilated snake and refine the contours by time-embed snake only, our approach achieves 60.3 AP_{vol}, which yields a gap of 0.3 than SOTA. Meanwhile, our approach gets a balance between too smooth and too sensitive compared to and E2EC [9], represented in Fig. Some segmentation results are visualized in Fig. 4.

Cityscapes. Our result of Cityscapes is shown in Table II. By adopting multi-component detection, our TSnake gets 39.6 AP in validation set, which is the second place on the board compared to other contour-based methods. It yields 0.6 AP to the E2EC method [9]. Some results are shown in Fig. 5.

D. Ablation Study

In this section, we are going to verify the importance of each main component in our proposed method, including the time-embed snake, the dilated snake, and the weighted dynamic matching loss. Following [8], all the models will be trained with the SBD dataset [24] and do not conduct mosaic augmentation in ablations.

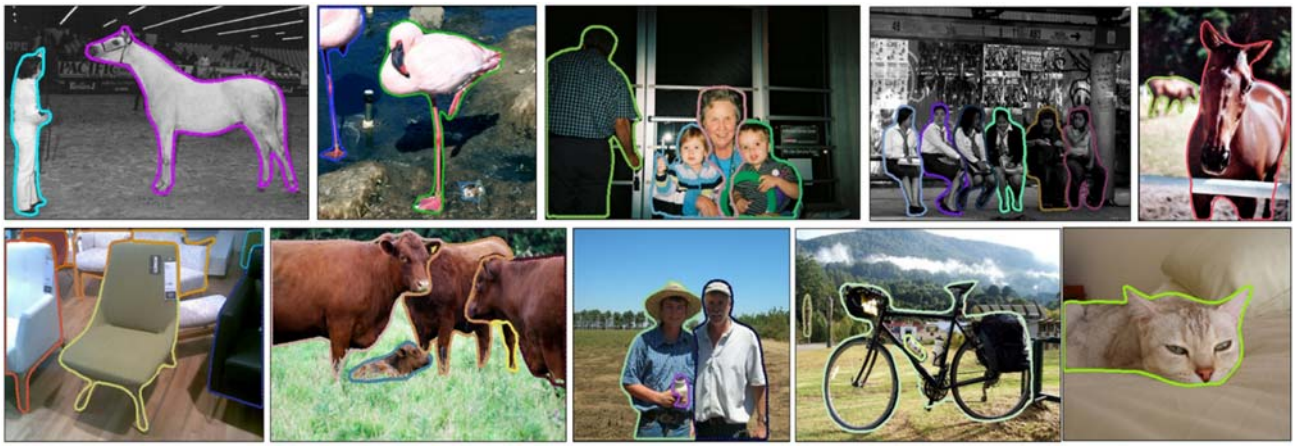


Fig. 6 The results on the SBD dataset.



Fig. 7 The results on the Cityscapes dataset.

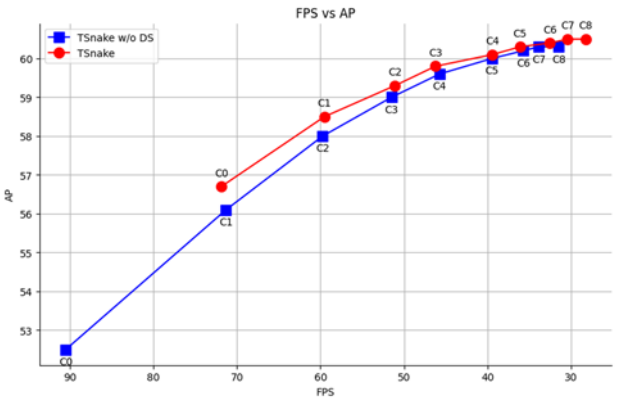


Fig. 8 Performance vs FPS in each iteration. The red and blue lines are the results of the TSnake with and without using the dilated snake, respectively. The performance converges at the 5th iteration, but still gets about 30 FPS. It shows that the proposed algorithm can be performed in real time.

Time-embed snake. We developed three settings to verify time embedding, shown in Table III. The first one is the origin snake module from Deep Snake [8], the second one will be added time embedding in every layer, and the third one is our proposed setting. Without time embedding, the origin snake got the worst performance with 1.3 AP_{vol} drop contrast to ours, which shows that time embedding really correctly controls the model in every iteration.

Dilated snake. Three cases are tested: (1) training without a dilated snake, (2) training by replacing a dilated snake with a snake module [8], and (3) training and testing with a dilated snake. The dilation of the kernel is set to 5, which is the same as our setting. The result is in Table III. Our dilated snake module got the best performance, yielding 0.3, 0.6 and 0.2 in AP_{vol}, AP_{50} and AP_{70} respectively compared to no addition one.

Loss function. Two experiments are conducted. The first is the importance of the weighted DML, and the second is which iteration is better to apply it. The results are shown in Table IV. In the first experiment, we replaced weighted DML by DML [9] and the smooth-L_1 loss. It shows that our proposed weighted DML improves by 0.9 AP_{vol} than the original DML, and yields 1.2 AP than the smooth-L_1 loss.

For the second experiment, we applied the weighted DML after the 2nd, 4th, and 6th iterations. As shown in Table IV, weighted DML started after the 4th iteration performs the best, yielding 1.0 AP_{vol} and 1.5 AP_{vol} to the 2nd and 6th iterations, respectively.

Iterative deformation. To show that the model is iteratively refined, we show the performance vs frame per second at each iteration, which is shown in Figs. 6 and 7.

Without the module, the performance significantly lagged behind the original in the first loops, however, by the 5th iteration, the gap narrowed to only 0.3 AP_{vol}, while the FPS increased by 3.5. An example of contour refinement at each iteration is shown in Fig. 8.

VI. Conclusion

In this work, we introduce TSnake, a time-embedded driven recurrent contour-based instance segmentation. TSnake applies time embedding to tell it which iteration is and deform the contours correspondingly. Next, the dilated snake shows that enriching receptive fields helps better reformation. Also, weighted DML reduces the sensitivity to noise, making it attractive to corners but not jagged. Our extensive experiments on the SBD and Cityscapes datasets demonstrate that TSnake achieves state-of-the-art performance while maintaining real-time efficiency.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE Int. Conf. Computer Vision*, pp. 2961-2969, 2017.

[2] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, issue 5, pp. 1483-1498, 2019.

[3] X. Shen, J. Yang, C. Wei, B. Deng, J. Huang, X. S. Hua, X. Cheng, and K. Liang, "DCT-Mask: Discrete cosine transform mask representation for instance segmentation," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 8720-8929, 2021.

[4] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *IEEE Int. Conf. Computer Vision*, pp. 9157-9166, 2019.

[5] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting objects by locations," in *European Conf. Computer Vision*, pp. 649-665, 2020.

[6] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and Ping Luo, "PolarMask: Single shot instance segmentation with polar representation," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 12193-12202, 2020.

[7] W. Xu, H. Wang, F. Qi, and C. Lu, "Explicit shape encoding for real-time instance segmentation," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 5168-5177, 2019.

[8] S. Peng, W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou, "Deep snake for real-time instance segmentation," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 8533-8542, 2020.

[9] T. Zhang, S. Wei, and S. Ji, "E2EC: An end-to-end contour-based method for high-quality high-speed instance segmentation," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 4443-4452, 2022.

[10] H. Feng, K. Zhou, W. Zhou; Y. Yin, J. Deng, Q. Sun, and H. Li, "Recurrent generic contour-based instance segmentation with progressive learning," *IEEE Trans. Circ. Syst. Video Tech.*, doi: 10.1109/TCSVT.2024.3383829. keywords: Apr. 2024.

[11] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[12] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid task cascade for instance segmentation," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 4974-4983, 2019.

[13] X. Wang, F. Zhang, and Y. Huang, "Fourier contour embedding deep learning for arbitrary-shaped target detection," in *Advanced Fiber Laser Conference*, vol. 12595, pp. 303-307, 2023.

[14] Z. Liu, J. H. Liew, X. Chen, and J. Feng, "DANCE : A Deep attentive contour model for efficient instance segmentation," in *IEEE Winter Conf. Applications of Computer Vision*, pp. 345–354, 2021.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840-6851, 2020.

[16] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv*: 1511.07122, 2015.

[17] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE Int. Conf. Computer Vision*, pp. 2980-2988, 2017.

[18] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016.

[19] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[20] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv*: 1904.07850, 2019.

[21] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 2403–2412, 2018.

[22] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results,". in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 9308-9316, 2019.

[23] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv*: 2004.10934, 2020.

[24] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Int. Conf. Computer Vision*, pp. 991–998, 2011.

GENERATING PHONETIC TRANSCRIPTIONS FOR KOREAN ENGLISH L2 LEARNERS USING MULTIPLE SELF-SUPERVISED-MODEL-BASED ASR SYSTEMS AND ROVER METHOD

Jong In Kim* and Sunhee Kim† Minhwa Chung*‡
* Interdisciplinary Program in Cognitive Science, Seoul National University, Republic of Korea
E-mail: prows12@gmail.com Tel/Fax: +82-028809309
† Department of French Language Education, Seoul National University, Republic of Korea
E-mail: sunhkim@snu.ac.kr Tel/Fax: +82-028807693
‡ Department of Linguistics, Seoul National University, Republic of Korea
E-mail: mchung@snu.ac.kr Tel/Fax: +82-028809195

***Abstract*—In Computer-Assisted Pronunciation Training (CAPT), accurate phonetic transcriptions are essential for identifying mispronunciations in non-native speech corpora. Publicly available corpora for Korean English L2 learners often lack these transcriptions due to their limited size and availability. To address the shortage of accurate phonetic transcriptions, we propose a method that combines multiple Self-Supervised Learning (SSL)-based phone recognition systems with Recognizer Output Voting Error Reduction (ROVER). We trained SSL-based phone recognizers (Data2vec, Hubert, Wav2vec) on the Librispeech and CommonVoice datasets and used them to decode the L2arctic corpus. By applying ROVER, we achieved 85.5% accuracy in phone transcription compared to manual tagging. Additionally, an error analysis of 140 beginner-level sentences from the Korean Spoken English Corpus (NIA144) identified common pronunciation errors among Korean English speakers.**

I. INTRODUCTION

In recent years, interest in learning a second language (L2) has significantly increased. This growing enthusiasm for L2 acquisition is driven by its numerous benefits, including improvements in career prospects, educational and academic advancements, and the expansion of social opportunities. In foreign language learning, correct pronunciation is essential as it enhances communication, accurately conveys the speaker's intentions, and improves both listening skills and overall comprehensibility with native speakers.

Computer-Assisted Pronunciation Training (CAPT) plays a crucial role in second language (L2) acquisition. It provides explicit and real-time feedback, which motivates learners to practice pronunciation and supports self-directed learning. Additionally, CAPT allows learners to repeatedly practice specific mispronunciations at their own pace. Typically, a CAPT system includes Mispronunciation Detection and Diagnosis (MDD) and Pronunciation Scoring. MDD identifies and diagnoses specific pronunciation errors, delivering explicit feedback to learners. Pronunciation Scoring assesses the learner's pronunciation against native speaker benchmarks, offering quantitative evaluations that guide improvement.

In MDD, identifying patterns of pronunciation errors is crucial. This process provides personalized feedback to learners and monitors their pronunciation progress, thereby ensuring effective learning outcomes. The error patterns at the phone level can vary based on the learner's proficiency level and are influenced by the specific interactions between the learner's first language (L1) and the target second language (L2).

Accurate pronunciation assessment depends on identifying phone error patterns, yet comprehensive phoneme error datasets are severely lacking. As shown in Table 1, the challenge of limited resources for phonetic annotation in non-native English is evident. For instance, the TIMIT dataset, widely used for native English, includes only 5,232 phone annotations. Datasets for non-native English, such as L2Arctic [1] and Speechocean762 [2], provide even fewer annotations, with just 3,619 and 5,000 phone annotations, respectively. Furthermore, essential datasets for L2 Korean, like NIA144(Topic-Adaptive English Speaking Assessment Data for Korean Speakers), NIA037(English Speech Corpus of Korean Learners for Educational Use), which are large-scale Korean English speech corpora released by AIHub, a part of the National Information Society Agency (NIA) in Korea, lack phonetic transcriptions entirely. This scarcity of annotated data significantly impedes the development of effective Mispronunciation Detection and Diagnosis (MDD) systems

Manual phone transcription is challenging due to pronunciation variations caused by stress, accent, and prosody, as well as the complexity of different phone systems based on the speaker's L1 and L2 background. It requires extensive linguistic knowledge and is often time-consuming and costly, typically necessitating the expertise of phonetics professionals[3], [4]. These factors make accurate phone tagging difficult.

Dataset	L1/L2	Hrs	Sentences	Phone Label
TIMIT	L1	4	5,232	5,232
L2arctic	L2 Various Countries	11.2	11,026	3,619
Speechocean762	L2 Mandarin	6	50,00	5,000
Commonvoice v6	L1	2,182	1,821,529	x
Librispeech	L1	1,000	292,367	x
NIA012	L2 Korean	5,000	7,276,761	x
NIA037-SPK	L2 Korean	1016.81	34,340	x
NIA144	L2 Korean	300	66,889	x

TABLE I
A COMPARISON OF BASIC STATISTICS, FOCUSING ON THE DISTINCTIONS BETWEEN NONNATIVE AND NATIVE ENGLISH DATASETS; NIA012: KOREAN CHILDREN'S ENGLISH SPEECH DATA(AIHUB), NIA037-SPK: ENGLISH SPEAKING EVALUATION DATASET FOR KOREANS(AIHUB), NIA144: TOPIC-BASED ENGLISH SPEAKING EVALUATION DATASET FOR KOREANS(AIHUB)

A study indicates that the correlation for phoneme tagging is 0.88[5]. In contrast, automatic phone transcription offers several advantages. It is cost-effective, reducing the financial burden of manual annotation. It ensures consistent annotation across datasets, minimizing human error and variability. Additionally, automatic transcription systems can adapt to various L1-L2 combinations, making them versatile for different linguistic backgrounds. They also handle large corpora efficiently, a task that is impractical with manual tagging due to the time and labor involved. However, Automatic Phonetic Transcription (APT) is still not perfect due to the complexity and accuracy of ASR, human speech variability, limitations in current technology, and the need for high precision in certain applications. Given these challenges in automatic phonetic transcription, our study proposes a methodology to address these issues.

To tackle the challenges of automatic labeling, researchers have developed various techniques. Most studies in this area have focused on automatic tagging for speech recognition. For example, [6] introduced the momentum pseudo-labeling method, which integrates online and offline models. Online models predict pseudo-labels guided by offline models, iteratively improving performance. Similarly, [7] applied pseudo-labeling to 60 languages using multilingual concepts, fine-tuning models with pseudo-labels across different languages. To address noisy data in pseudo-labeling, [8] modified the training objective to detect incorrect labels, reducing errors. In pronunciation assessment, [9] proposed a zero-shot approach using the Hubert model, incorporating a transformer encoder, k-means clustering, and a scoring module to evaluate pronunciation. While sentence-level automatic tagging has been extensively studied, more thorough research is needed for phone-level automatic transcription.

To address this challenge, we propose an efficient method for automatically generating phonetic transcriptions for large non-native English corpora. Our approach utilizes multiple SSL-based phone recognizer systems in conjunction with the Recognizer Output Voting Error Reduction (ROVER) technique. ROVER enhances transcription robustness by reducing errors and improving accuracy. This approach enables the analysis of phone-level error patterns in non-native speech corpora.

The remainder of this paper describes our methods (Section 2), outlines the experimental conditions (Section 3), and discusses the results (Sections 4 and 5)

II. PROPOSED METHOD

A. Step1. Constructing the three types of SSL-based phone recognizer

In initial phase, we trained phone recognizer to obtain actual phone sequences. We assume that pronunciations by native English speakers adhere to standard pronunciation. Using available English corpora such as LibriSpeech and CommonVoice, we use audio transcriptions. We then apply a Grapheme-to-Phoneme (G2P) conversion to map written forms (graphemes) to their phonetic representations. Specifically, we train three type of SSL-based phone recognizers, including Data2vec[10], Hubert[11], and Wav2vec 2.0[12].

B. Step2. Actual Phonetic Transcription via Recognizer Output Voting Error Reduction (ROVER)

In the second phase, we derive the final phone sequence using Recognizer Output Voting Error Reduction (ROVER)[13]. ROVER combines outputs from multiple phone recognizers to produce a composite result, reducing the overall error rate compared to individual systems. In our experiment, ROVER was applied to the outputs of three phone recognizers to obtain high accuracy. The three phone recognizers used are Wav2vec-xlsr53, Hubert-large, Data2vec-large.

Given phone sequences $P_H ubert, P_D ata2vec, P_W av2vec$ from three phone reocgnizer systems:

$$P_H ubert = [p_{Hubert,1}, p_{Hubert,2}, \ldots, p_{Hubert,T_1}]$$

$$P_D ata2vec = [p_{Data2vec,1}, p_{Data2vec,2}, \ldots, p_{Data2vec,T_1}]$$

$$P_W av2vec = [p_{Wav2vec,1}, p_{Wav2vec,2}, \ldots, p_{Wav2vec,T_1}]$$

In alignment, each phone sequence is aligned to group matching phones at each time t. Algorithms such as dynamic programming with a word transition network (WTN) are used to perform this alignment.

$$Hubert = \text{Align}(P_1, P_2, \ldots, P_n) \rightarrow \{A_1, A_2, \ldots, A_T\}$$

$$Data2vec = \text{Align}(P_1, P_2, \ldots, P_n) \rightarrow \{A_1, A_2, \ldots, A_T\}$$

$$Wav2vec = \text{Align}(P_1, P_2, \ldots, P_n) \rightarrow \{A_1, A_2, \ldots, A_T\}$$

where A_t is the set of aligned phones at time t.

Subsequently, the voting module analyzes this data to generate the final transcription.

In Voting Process, at each time t the most frequent phone in the alignment set A_t. is selected. The voting metric follows the majority rule.

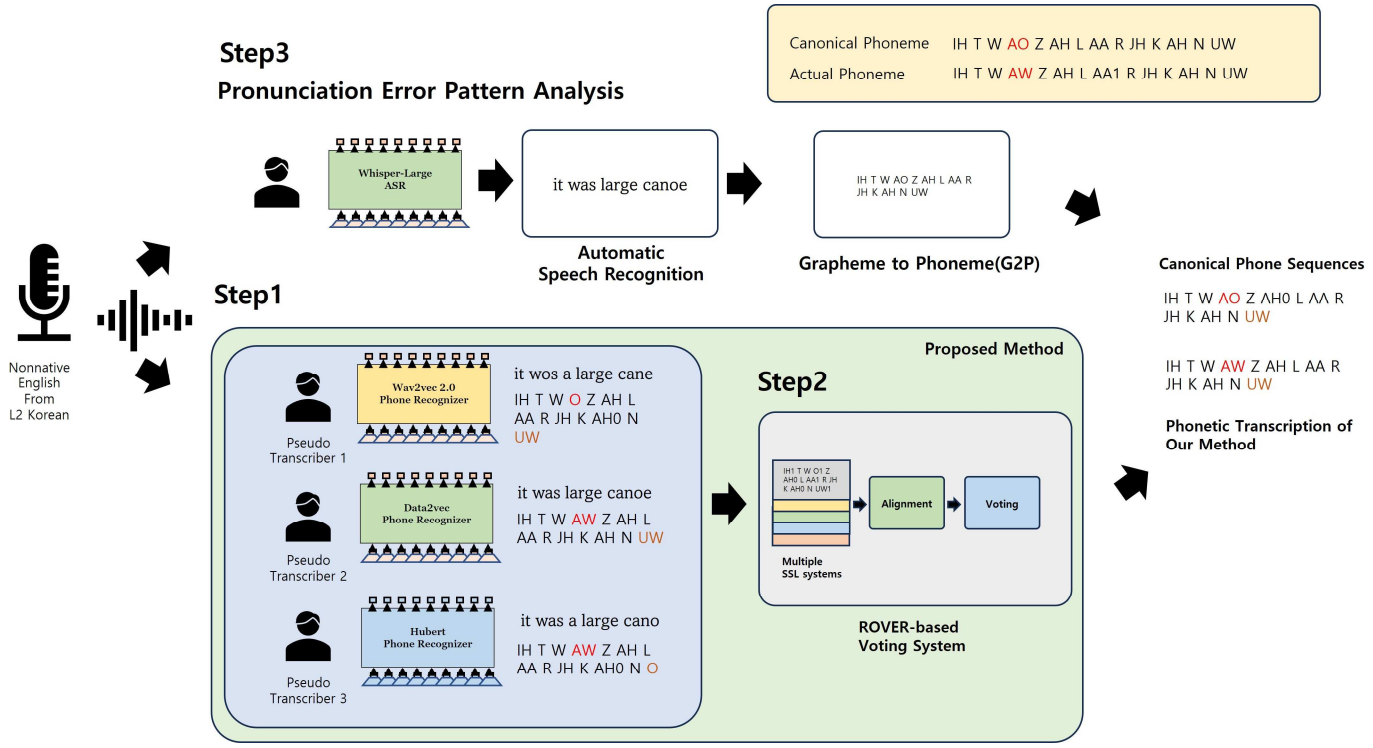


Fig. 1. The procedure of our ROVER-based phonetic transcription methodology. In Steps 1 and 2, we detail our proposed method for generating actual phone transcriptions. In Step 3, we analyze and present the pronunciation error patterns by comparing these actual phone transcriptions with the canonical transcription.

$$W_t = \arg\max_{p \in A_t} \sum_{i=1}^{n} \delta(p, P_i[t])$$

where $\delta(p, P_i[t])$ is a function that returns 1 if phone p is present at time t in the output of recognizer P_i, and 0 otherwise.

Combining the outputs of the three phone recognizers at each time t to generate the final phone sequence.

$$\text{Output} = \{P_1, P_2, \ldots, P_T\}$$

By combining the outputs of these modules, we unify shared characteristics among three phone recognizers, leading to a representation of phone annotations at each time t and generating the final phone transcription sequence

C. Step3. Phone-level Error Pattern Analysis

We adapt our automatic phonetic transcription method for phone-level error analysis. we use officially available Whisper-based ASR systems to convert raw audio into text, focusing on creating canonical phone sequences. Next, we apply a grapheme-to-phone (G2P) model to convert the decoded text into canonical phone sequences for comparing with the actual phone sequences generated by our method. We compare two sets of results in our analysis. By comparing the standard sequences with those generated by our method, we can identify error patterns. Our method assumes the availability of only raw audio files from L2 English speakers, with no textual information.

III. Experiment

A. Data

To train the phone recognizer, we use native English corpora, specifically CommonVoice[14] and LibriSpeech[15]. These corpora are chosen for their diverse range of spoken English styles, which enhances recognizer accuracy. The training set includes 1,951,560 utterances from CommonVoice and 562,480 utterances from LibriSpeech, totaling 2,514,042 utterances. CommonVoice v13 provides 1,366 hours of speech, while LibriSpeech offers 961 hours, amounting to a combined total of 2,327 hours of read speech.

We convert the transcribed text into ARPA-style phone representations using the Montreal Forced Aligner with the English US ARPA v2.00 G2P models. We use 70 ARPA phone types, which are categorized into two groups: one group excludes stress markers (e.g., AA0, AA1, AA2 grouped as AA), and the other group includes stress markers with two-group (specifically, (AA0, AA1) into AA, and AA2).

B. SSL-Based Phone Recognizers

We use the fairseq toolkits to build SSL-based phone recognizers: Data2vec[10], Hubert[11], and Wav2vec 2.0[12]. We apply the default training settings provided by fairseq. We train three type of SSL based phone recognizers: Wav2vec-xlsr53, Hubert-large, Data2vec-large. For decoding, we use the flashlight toolkit to perform Viterbi decoding without a language model.

phoneme recognizer	Validation set (PER)	Testset (PER)
Data2vec-large	5.395	7.373
Hubert-large	6.968	9.915
Wav2vec xlsr-53	5.846	7.83

TABLE II
ASSESSING THE PERFORMANCE OF PHONE RECOGNIZER SYSTEM WITH SSL (PER : PHONE ERROR RATE)

	ROVER(Recognizer Output Voting Error Reduction)
Data2vec-large	85.4%/85.2%
Hubert-large	85.0%/84.7%
Wav2vec-xlsr53	84.9%/84.6%
ROVER Accuracy	**85.5%/85.3%**

TABLE III
COMPARISON OF OUR METHOD AND L2 ARCTIC HUMAN ANNOTATION ON KOREAN SPOKEN ENGLISH, BOTH WITH AND WITHOUT STRESS MARKER(ACCURACY)

	ROVER(D2vlarge+Hubert-large+Wavxlsr53)
Arabic	84.2%
Hindi	83.8%
Mandarin	81.0%
Spanish	80.3%
Vietnamese	78.1%
Korean	85.5%

TABLE IV
PERFORMING AN EXTENSIVE ANALYSIS OF ROVER(DATAVEC-LARGE + HUBERT-LARGE + WAV2VEC-XLSR53) TO EVALUATE THE ALIGNMENT ACCURACY OF OUR PHONETIC TRANSCRIPTION METHOD IN COMPARISON TO THE MANUAL ANNOTATION IN L2ARCTIC, ENCOMPASSING MULTIPLE COUNTRIES(ACCURACY)

C. ROVER-based phone transcription

To use ROVER, segmental boundaries (time alignment) are required. We obtain these boundaries by performing forced alignment with the Wav2Vec acoustic model. We then use the SCTK toolkit to integrate phone recognizers through the ROVER method. we use three recognizers to obtain final phone sequences in the ROVER system.

D. Evaluation

We verify the performance of Automatic Phonetic Transcription. We compare the manual transcription of an existing publicly available L2 Arctic dataset with our proposed automatic transcription, calculating alignment accuracy.

E. Analyzing Error Patterns

The L2Arctic corpus [1] is a collection of nonnative English speech data designed for mispronunciation detection. It includes recordings from 24 nonnative English speakers, with a balanced gender distribution and 150 utterances per participant. We use the L2Arctic dataset to assess the performance of our phonetic transcription method by comparing the phone transcriptions generated by our system with the manual annotations for Korean speakers in the dataset.

To further investigate our methodology, we used the NIA144 dataset to analyze and compare phone-based error patterns with our methodology. The NIA144 corpus includes 450 hours of recordings from L2 Korean speakers, categorized into five proficiency levels. We selected 140 cases from the lowest proficiency level (Level 1) to analyze the resulting error patterns. If the error patterns identified by our method align closely with common errors in Korean English, it indicates the reliability of our method.

IV. RESULTS

A. Performance of an SSL-Based Phone Recognizer System via phone error rate(PER)

In Table 2, Wav2vec-xlsr-53 and Data2vec-large exhibit lower PER values for both the validation and test sets, indicating more accurate phone recognition compared to other models. Notably, Data2vec-large achieves the lowest validation PER (5.395), demonstrating good performance in self-supervised learning on the English corpus.

B. Comparison of our method and L2 Arctic human annotation on L2 English from various country, both with and without stress

In Table 4, the model demonstrates the highest accuracy (85.5%) for Korean, highlighting its effectiveness in recognizing phones for Korean English L2 learners. Arabic and Hindi also show high accuracy, indicating strong performance with the phonetics of these languages. Mandarin and Spanish follow with slightly lower accuracy, likely due to their tonal nature and unique phonological features. Vietnamese has the lowest accuracy among the languages tested, suggesting challenges in capturing its phonetic nuances.

C. Analyzing Error Patterns

As shown in Table 3, the results highlight the performance of the ROVER models. The ROVER method achieves an accuracy of 85.5% without stress markers and 85.3% with stress markers

In Table 5, comparing our phonetic transcription results with the manual annotations from L2Arctic for L2 Korean reveals variations in pronunciation, particularly with consonants. The findings include changes such as converting the open 'AH' sound to the fricative 'TH', altering the sibilant 'S' to its voiced counterpart 'Z', replacing 'D' with the voiced dental fricative 'DH', and interchanging the tense 'IY' with the lax 'IH'

We adapt our transcription method to the NIA144 dataset. We examined error patterns by analyzing 140 instances of Korean spoken English in NIA144 using our phone labeling method. In Table 6, we evaluate the efficacy of our system by comparing the error patterns identified through our phone annotation method with common consonant errors in non-native English speech by L2 Korean learners. Our findings align with the patterns documented in [5], including errors such as (/z/, /s/), (/dh/, /d/), (/v/, /h/), and (/f/, /p/)

	Substitutions	Deletion	Addition
1	'AH','TH'	'D'	'D'
2	'S', 'Z'	'DH'	'T'
3	'D', 'DH'	'IY'	'R'
4	'IY', 'IH'	'AH'	'HH'
5	'AO', 'OW'	'T'	'AO'
6	'AE', 'EH'	'AX'	'AH'
7	'AE', 'AH'	'S'	'K'
8	'IY', 'TH'	'Y'	'DH'
9	'Z', 'S'	'IH'	'TH'
10	'UW','OW'	'UW'	'W'

TABLE V
EXAMINING THE PATTERNS OF ERRORS PRESENT IN OUR PHONETIC TRANSCRIPTION WITH THE MANUAL ANNOTATIONS ON KOREAN SPOKEN ENGLISH

	Substitutions	reference [3]	Deletion	Addition
1	**'F', 'P'**	**/z/, /s/**	'T'	'T'
2	**'Z', 'S'**	**/dh/,/d/**	'N'	'N'
3	'L', 'R'	**/v/,/b/**	'D'	'D'
4	'D', 'T'	**/f/,/p/**	'S'	'S'
5	**'DH', 'D'**		'L'	'M'
6	**'S', 'Z'**		'R'	'L'
7	**'V', 'B'**		'W'	'R'
8	'B', 'P'		'Z'	'K'
9	'G', 'K'		'K'	'TH'
10	'T', 'D'		'Y'	'W'

TABLE VI
EXAMINING ERROR PATTERNS IN OUR PROPOSED PHONE ANNOTATION TO ENGLISH NIA144 DATASETS SPOKEN BY KOREAN: A COMPREHENSIVE INVESTIGATION OF 140 CASE; DATA2VEC-LARGE + HUBERT-LARGE + WAV2VEC-XLSR53

V. DISCUSSION

Our research focuses on developing automated phonetic transcription for non-native English speakers, particularly L2 Korean learners. While our system did not achieve high accuracy, the experiments have provided valuable insights into handling the issue of sparse phone annotations.

The accuracy of our phonetic transcription is reliable, especially when compared to previous research on non-native L2 Korean speakers. Earlier studies show an 88% agreement among human transcribers [5]. Despite a 3% margin of error, our transcription accuracy of 85.5% demonstrates a strong consensus and reliability in comparison with human transcriber agreement

Analyzing error patterns in phonetic transcription using real, unlabeled data from 140 cases in the NIA144 dataset reveals similarities with the error patterns found in previous research on Korean English learners [5]. These results suggest that our method is effective for transcribing the speech of non-native English speakers, especially Korean learners. Additionally, our method shows promise for analyzing pronunciation errors in other language pairs, particularly for minority languages where manual phonetic transcription is challenging.

VI. CONCLUSION

In summary, we have addressed the challenges by combining a self-supervised learning-based phone recognizer with ROVER. Our experiments using the L2arctic corpus for Korean learners achieved an 85.5% accuracy in phone recognition. This consistent accuracy and the analysis of error patterns on unlabeled corpora demonstrate our method's effectiveness for transcribing extensive non-native English speech. Future work will focus on adapting our approach for different non-native English corpora and conducting thorough error analyses to better understand phonetic transcription challenges.

REFERENCES

[1] G. Zhao, S. Sonsaat, A. Silpachai, *et al.*, "L2-arctic: A non-native english speech corpus.," in *Interspeech*, 2018, pp. 2783–2787.

[2] J. Zhang, Z. Zhang, Y. Wang, *et al.*, "Speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[3] L. Yang, J. Zhang, and T. Shinozaki, "Self-supervised learning with multi-target contrastive coding for non-native acoustic modeling of mispronunciation verification," *Proc. Interspeech 2022*, pp. 4312–4316, 2022.

[4] D. Zhang, A. Ganesan, S. Campbell, and D. Korzekwa, "L2-gen: A neural phoneme paraphrasing approach to l2 speech synthesis for mispronunciation diagnosis," 2022.

[5] H. Hong, S. Kim, and M. Chung, "A corpus-based analysis of english segments produced by korean learners," *Journal of Phonetics*, vol. 46, pp. 52–67, 2014.

[6] Y. Higuchi, N. Moritz, J. L. Roux, and T. Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," *arXiv preprint arXiv:2106.08922*, 2021.

[7] L. Lugosch, T. Likhomanenko, G. Synnaeve, and R. Collobert, "Pseudo-labeling for massively multilingual speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7687–7691.

[8] H. Zhu, D. Gao, G. Cheng, D. Povey, P. Zhang, and Y. Yan, "Alternative pseudo-labeling for semi-supervised automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[9] Y. Chen, W. Wang, and C. Wang, "Semi-supervised asr by end-to-end self-training," *arXiv preprint arXiv:2001.09128*, 2020.

[10] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*, PMLR, 2022, pp. 1298–1312.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[13] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, IEEE, 1997, pp. 347–354.

[14] R. Ardila, M. Branson, K. Davis, *et al.*, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

Unsupervised Discovery of Non-Categorical L2 Error Patterns Using Wav2Vec2.0 Code Vectors

Eunsoo Hong* Sunhee Kim† and Minwha Chung‡
* Seoul National University, Seoul, Korea
E-mail: stxlla13@snu.ac.kr Tel/Fax: +82-2-8806162
† Seoul National University, Seoul, Korea
E-mail: sunhkim@snu.ac.kr Tel/Fax: +82-2-8807693
‡ Seoul National University, Seoul, Korea
E-mail: mchung@snu.ac.kr Tel/Fax: +82-2-8809195

***Abstract*—L2 pronunciation is shaped by the interaction of two sound systems, which makes their identity more complex than a single phoneme category. The non-categorical nature demands assessment at a level finer than phonemes. As the granular requirement is highly labor-intensive, unsupervised methods emerged. Nevertheless, they either reverted to categorical diagnosis or used the supervised and phoneme-prescribed feature phonetic posterior-gram (PPG). Alternatively, this study adopts the unprescribed and unsupervised feature, the Wav2Vec2.0 code vector, to locate sub-phonemic variations. We first verify the features' L2 discernability by comparing their frequency across single-speaker data of L1 (CMU ARCTIC) and L2 (L2 ARCTIC). Clustering is performed on frequency vectors to test their separability on account of nativeness. Subsequently, sub-segmental patterns are analyzed among segmentally identical error samples in L2 Korean English NIA 037 data. After cataloging segmental errors detected by the model finetuned with L1 TIMIT, their corresponding code vector sequences are extracted by referencing the forced alignment result. We then derived dominant patterns of the sequences and compared them against L1 reference materials likewise constructed from TIMIT. Phoneme-code vector co-occurrence probability and code vector clustering were each used to check their attributes and uniqueness. The result confirmed the discernability, followed by linguistically interpretable common traits across patterns. (1) They formed a gradient error continuum along the changed articulatory value, reflecting the non-categorical nuanced understanding. (2) This trait is highlighted by intermediary typology, which assumed opposite values in two codebooks and was also rare in L1 for being L2 specific. Lastly, (3) distribution skewed towards the most approximate sound in the learner's L1, from which the patterns' complexity stems.**

I. Introduction

Articulation in L2 speech involves the mutual participation of the learner's native (L1) and target (L2) sound systems. Their complex interplay often spans the boundaries of two or more canonical phoneme categories. Cantonese English speakers, for instance, may utter a variation of [n] resembling both [l] and [n], mirroring their recent sound merger at the syllable initial position [1][2][3]. This between-categorical characteristic exists in a continuum, demanding gradient evaluation. Nevertheless, identifying subtleties of acoustic mismatch is an extremely laborious task. Unsupervised mispronunciation detection literature emerged under this setting, using the framework of acoustic pattern discovery[2][3][4][5][6][7][8][9][10]. If pattern discovery aims to extract recurring signals to substitute manual labeling [11], L2 errors could be set as one such target as they are recursive under systematic interaction. Nevertheless, the initial research focus reverted to categorical scope, as the obtained nuanced granular details were reduced to phonemic decoding [7][8] and binary decision-making [9][10]. The following [2][3][4][5][6] went beyond to describe the non-categorical attributes of error using the common analysis feature, phonetic posterior-gram (PPG). These posterior vectors were generated from MFCC-phoneme-label-trained neural networks, which contradicts two aspects. First, it is prescriptive of predefined phoneme thresholds, making the discovered pattern still tied to categorical circumscription. Requiring labels to map probability further notes its reliance on supervised learning. PPG is susceptible to the quality of labeled data used to train the instrumental model [2], which partially runs against the goal of sparing labor costs.

Concerning these drawbacks, this work adopts an alternative analysis feature, Wav2Vec2.0 code vectors. The choice is grounded on three factors. First, their generation process is fully unsupervised and unaffected by phonemic stipulation. Representation learning in self-supervised learning (SSL) does not accompany label training but rather self-discovers the operating units from the data's structural property. Second, this learning procedure conceptually resembles acoustic pattern discovery, which shares the goal of discovering lexically meaningful units to express speech content. As code vectors are already a rendition of discovered patterns, using them reinforces the research agenda. Third, linguistic probings in model documentation verify their phonetic relevancy. Ref. [12] proved that each feature specializes in representing different phonemes, while [13] showed that the usage pattern of multilingual pre-trained latent overlaps for close languages. The probing of [12] also implies that code vectors are more granular units than phonemes as the same phoneme is represented by multiple varieties.

With this understanding, we aim to survey the scope of variation capturable by code vectors to propose them as an

alternative means to discover non-categorical patterns. Accordingly, we first test the differentiability among L1 and L2 and subsequently use the feature to describe variations within each segmentally defined L2 error. This pattern discovery pipeline combines the aforementioned probing methods with frameworks of previous literature. To answer the question (1) Does the feature encode L2 acoustics differently from L1? codeword frequency probing in [13] is adopted. Featural usage of L1 and L2 speakers are compared through a frequency vector that maps occurrences of different code vector inventories. For question (2) How are non-categorical characteristics in L2 encoded in them? sequence analysis in [6] is applied to derive dominant patterns among segmentally identical error samples. The samples were pre-selected from the auxiliary segmental detection task. To interpret these patterns, two L1 reference materials are created. Co-occurrence-based phonetic probing in [12] is used to infer their phonetic attribute. KMeans clustering, a common similarity measure in [6][3][4][5], is performed on the union of available features to check relationships among discovered patterns.

II. Methods

A. Code Vector Inventory Probing

Code vectors' L2 discernability can be proved through the difference in the used inventory between L1 and L2 speech. Since product quantization in Wav2Vec2.0 concatenates two codebook entries, usage can be analyzed at the level of 1) individual entry and 2) entry pair. The former approach was taken in [13] to prove the feature's inter-language discernability. Among data of languages the model was pre-trained on, the degree of overlap in codeword distribution was in line with language similarity.

The same strategy is applied with the target of frequency calculation now set as individual L1 and L2 speakers. After segmenting speech by phoneme through forced alignment, the most representative codeword or codeword pair of the divided field was recorded as an index. At the entry-level illustrated in Fig. 1, codeword frequency vectors of size 1*320 are constructed per codebook whose values are normalized into probability. 320 is a size of codebook representation pre-defined by the model. These vectors are ultimately concatenated to a size of 1*640 for comparison. At the pair-level, the number of used pairs is not fixed, with a theoretical maximum reaching 102.4k i.e. 320*320 [12]. To plot the frequency in mutual space, we took the union of all available paired indices across the dataset and set them as counting bins to map occurrences. This amounted to 1*5712-sized vectors that were likewise normalized from the raw count.

KMeans clustering was eventually performed on each set of frequency vectors with the cluster number k set as 2. The expected result was to have L1 and L2 speaker data develop separate clusters, reflecting the usage pattern differences.

B. L2 Error Pattern Discovery

We next expand our survey to sub-segmental variations within L2. A list of error segments is first gathered through the

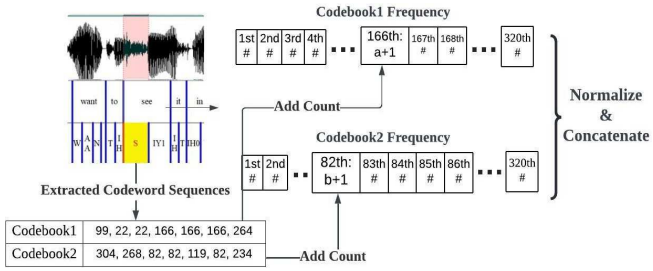


Fig. 1. Construction of Codeword Frequency Vectors. a and b are the hypothetical number of occurrences before adding count

standard practice of comparing recognized results with ground truth labels. The same pre-trained model used for code vector extraction was fine-tuned to create the recognition tool.

Then, code vector sequences of the error samples were retrieved as indices. The process first entails forced-aligning the L2 analysis data and spotting the time frames of falsely recognized phonemes. This was viable by employing the same phoneme set during forced alignment and grapheme-to-phoneme (G2P) conversion of fine-tuning labels. Subsequently, the sequence analysis method in [6] is applied to spot recurring patterns. The study [6] shared the goal of finding patterns within a canonically designated segment represented by summarizing indices. The analysis first involved filtering the sequence by removing minor presences and summing adjacent overlaps into one. Then, dominant patterns were selected based on frequency as their representation was once again refined by merging subsequences with their subsuming counterparts. In all analyzed cases, applying this procedure resulted in the 3 most dominant patterns, each represented by a single paired index.

To interpret these patterns, two reference materials were created from the same L1 data used for fine-tuning. We recorded 1) what phonemes each codeword (pair) statistically represents and 2) the raw numerical values of code vectors as opposed to indices. The former is a revisit to the phonetic probing in [12] that plotted the conditional probability of phoneme distribution into a graph. The same method of counting the feature's co-occurrence with human-annotated phoneme boundaries was executed, representing a cross-section of this graph along the vertical axis. Raw vectors were used to evaluate the distance among patterns. As our L1 data utilized 2202 pairs in total, the recorded 2202 concatenated vectors were clustered into 39 groups. The number 39 reflects the quantity of the mutual phoneme set as it serves as a minimum division criterion. Cluster IDs of the three dominant indices were compared to check if each discovered pattern formed a separate identity. If two or more indices belong to the same cluster, the patterns are merged into one. Fig. 2 illustrates the overall framework. While segmental detection is supervised, it is an optional step to limit the scope of sub-segmental analysis. In practice, the unsupervised discovery is applicable to any segment of interest.

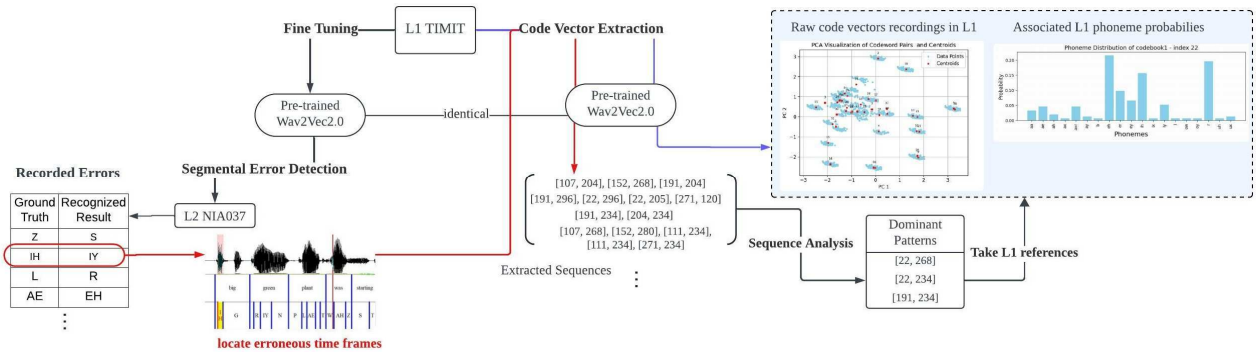


Fig. 2. Overview of L2 Error Pattern Discovery

III. Experiments

A. Data

Two different L1-L2 datasets were used for each task. For inventory comparison, an identical speech prompt across L1 and L2 was important to provide content-wise regulation. Speaker-level recordings were also needed to spot trends concerning a particular demographic. Accordingly, CMU ARCTIC (L1) and L2-ARCTIC (L2) v5.0 were chosen. For CMU ARTIC, only 5 speakers (bdl, rms, jmk, rms, clb) of North American accents were used as this research concerns pronunciations of US English. For pattern analysis, encompassing multiple speakers of diverse backgrounds was desirable to create robust speech recognition and references (L1), and make generalizable discoveries (L2). Manually confirmed phoneme duration was also preferred to extract code vector sequences accurately. Accordingly, TIMIT (L1) and NIA037, Korean Learners' L2 Speech corpus (English) developed by the National Information Society Agency (NIA), Korea, were used.

B. Implementation Detail

All SSL-related works and the model finetuning used fairseq **Framework**, a sequence modeling toolkit developed by the Facebook research group. The **Pretrained Model** used to extract code vectors was LARGE architecture trained on LibriVox (LV-60k) data. This is also the version phonetic probing in [12] was performed on. Using the TIMIT train split, **Finetuning** was conducted under the parameter of 40000 max update, 3e-4 learning rate, mask probability 0.65, mask channel probability 0.5. The final validation phoneme error rate was 1.63%. The mutual phoneme set used for finetuning label and forced alignment was the 39 ARPAbet symbols. For **G2P**, the set was derived by excluding the sentence stress numeric from the g2p tool kit [1]. For **Forced Alignment**, excluding the stress marker in English (US) ARPA dictionary v3.0.0 of Montreal Forced Aligner resulted in the same union of notation. The corresponding version of the acoustic model was also used.

[1]https://github.com/Kyubyong/g2pGitHub - Kyubyong/g2p: g2p: English Grapheme To Phoneme Conversion

Clustering of paired vectors and **Visualization** of individual codewords used FAISS library.

IV. Results

A. L1 to L2 Code Vector Usage Comparison

Fig. 3 displays the clustering results after performing the principal component analysis. Speakers of the frequency vectors are annotated next to the plotted location, and each element is color-coordinated according to its cluster membership. As can be noted from the color scheme, the cluster grouping coincides with the division between native and non-native speech. The concatenated format offers a clearer separation, allowing for a more nuanced examination of subtle pronunciation deviations. This motivates us to conduct detailed cross-speaker comparisons in a paired setting.

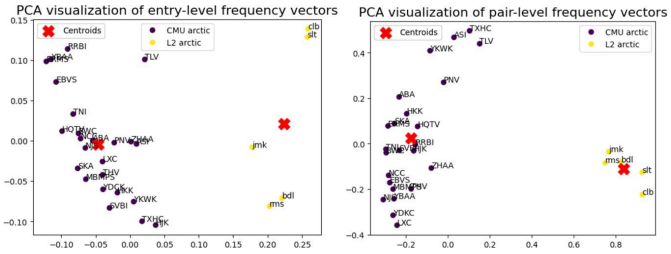
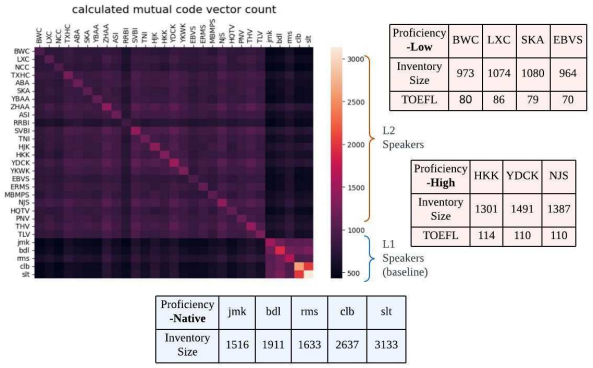


Fig. 3. Frequency Vector Clustering Results

The heatmap in Fig. 4 counts mutual codeword pair inventory between two speakers. Reflecting the former clustering result, the shared amount between native and non-native speakers is far below the rate within each speaker group. This sharing rate asymmetry is further marked by higher numbers of utilized pairs in L1. The left L2-ARCTIC grids have darker shades associated with fewer tokens compared to the rightmost five grids of CMU ARCTIC speakers. Interestingly, the inventory size increases as a function of the speaker's language proficiency. From the demographic information provided in [14], we have selected two speaker groups of opposing proficiency levels. The lower-level group had TOEFL iBT scores ranging under 90, whereas the higher-level group had scores over 110. The total usage-

-counting on the right shows that higher-level speakers overall use more paired tokens than lower-level counterparts. One possible explanation is that to fully utilize representations encoding English, one has to be phonetically aware of the sounds in the language. This awareness is marked by the utilization ratio of the code vector inventory acquired in the L1 standard. If the amount used by the native speakers is the full range of available acoustics, the less adept one is at articulating these sound units, the less amount of inventory in use there will be.



Proficiency -Low	BWC	LXC	SKA	EBVS
Inventory Size	973	1074	1080	964
TOEFL	80	86	79	70

Proficiency -High	HKK	YDCK	NJS
Inventory Size	1301	1491	1387
TOEFL	114	110	110

Proficiency -Native	jmk	bdl	rms	clb	slt
Inventory Size	1516	1911	1633	2637	3133

Fig. 4. Code Vector Inventory Counting

B. L2 Error Pattern Discovery Result

Subject of Gradience Our analyzed segmental errors concerned 5 different substitution routes. Patterns of each trajectory could be scaled by the assumed degree of changed articulatory value, forming a gradient error continuum. *Substitution of voicing identity*: in Z to S, the second highest probability in native phoneme distribution moved from sh[-voicing] to z[+voicing]. *Substitution of manner of articulation*: in DH to D, V to B, and F to P, native phoneme distribution moved from plosives and silence[-continuity] to fricatives, vowels, and approximants[+continuity]. This was a recurring dynamic across all three examples. *Substitution of laterality*: in L to R and R to L, native phoneme distribution moved from having more lateral[+laterality] to rhotic[-laterality] association (or vice versa). *Substitution of vowel height*: in AH to AA, the association ratio moved from allotting higher probability to AA[-high] to AO[+high]. *Substitution of tenseness*: in IH to IY, IY to IH, AE to EH, and EH to AE, tenseness calculation reflected Euclidean distance among raw code vectors. Fig. 6 charts these gradient movements. The intermediary patterns were decomposed into two codewords for the reason that will be explained below.

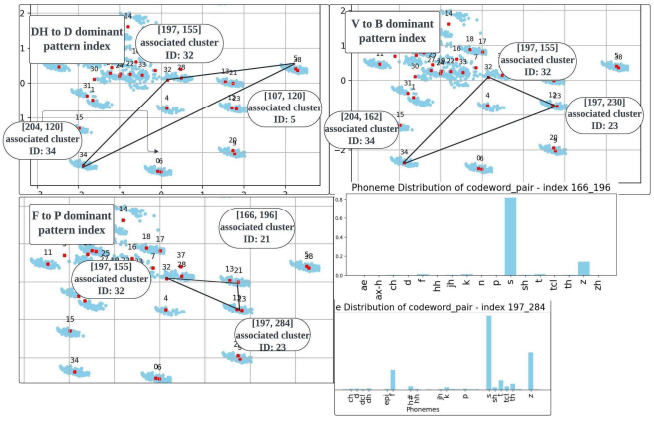


Fig. 5. Fricative to Plosive Dominant Pattern Dynamics

Intermediary Typology Patterns at the intermediate position were non-categorical by assuming opposite identities in two codebooks. If one displayed a positive value of the changed articulatory trait, the other displayed a negative value. Such contradictory pairing was rare in L1 data, attesting to the L2 particular nature of non-categoricity. In *Z to S*, intermediary [18, 51] had codebook1 bearing [-voicing], while codebook2 bearing [+voicing] identity. The paired index had no native speech presence. In *DH to D*, intermediary [204, 120] had codebook1 bearing [+continuity], while codebook2 bearing [-continuity] identity. It records a single co-occurrence with nasal [n], a sound that exemplifies duality with airflow obstructed in the oral cavity yet unhindered in the nasal cavity. In *R to L*, intermediary [191, 235] had codebook1 bearing [-laterality], while codebook2 bearing [+laterality] attribute. The pair had only three L1 occurrences. In *AH to AA*, codebook1 and 2 each assumed [-high] and [+high] identity. The middling pink figures in Fig. 6 showcase this ambivalence. Note that the ranking is comparative.

Distributional asymmetry Patterns were skewed toward the most approximate sound available in the learner's L1. This finding could be viewed at two levels: 1) dispersion rate across different substitution types and 2) within-substitution pattern distribution. The case of fricative to plosive substitution demonstrates the first view. Fig. 5 plots the distance among subsegmental patterns through their associated cluster IDs. Accordingly, *F to P* involves smaller between-pattern distances than the other two. The convergence gears towards the [+continuity] end as [197, 155] that once formed the [+continuity] end of the spectrum in *DH to D* in Fig. 6 is now the pattern bearing the least characteristic of continuity. The other two indices, [166, 196] and [197, 284], show an even higher association with the ground truth canonical value. The asymmetry is related to Korean fricative inventory consisting of voiceless but not voice fricatives, leading to fewer difficulties articulating a familiar phonation type. Namely, the more difficult target DH and V incurs greater dispersion. Meanwhile, the skewed distribution within substitution can be found in the case of liquid. Fig. 7 likewise visualizes the pattern dynamics with cluster IDs. The laterality spectrum is additionally superimposed by referencing Fig. 6. In *R to L*, the intermediary [191, 235] is closer to the [+laterality] pattern [204, 162] than [22, 101]. In *L to R*, two dominant patterns [191, 212] and [191, 162] merge into a single cluster that gears towards the same direction. The concentration on the

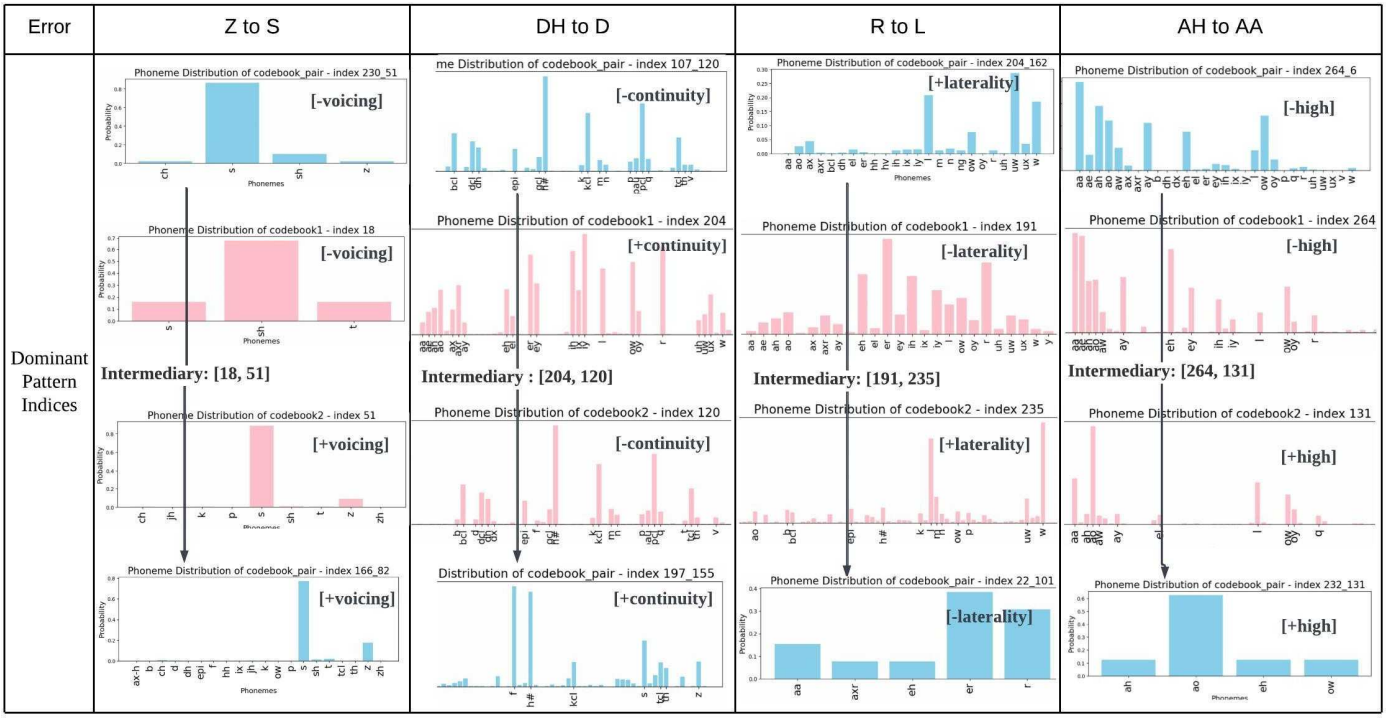


Fig. 6. Gradience of Error Continuum. Indices with numerous phoneme associations have been cropped for illustration purposes

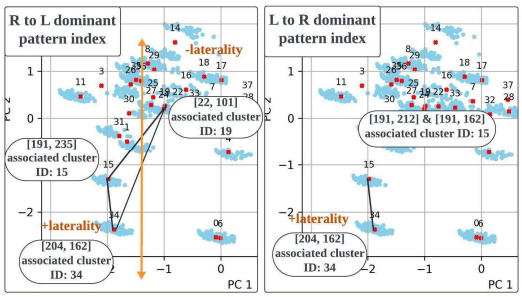


Fig. 7. Pattern Distribution in Liquid Substitution

[+laterality] end reflects greater difficulties in producing the rhotic variety, bunched r, unobserved in Korean compared to the existing lateral inventory.

V. Discussion

The three overarching findings are linguistically interpretable. First, it is expected for the variation spectrum to develop along the articulatory trait responsible for changes.

Second, since the non-categorical nature is L2-specific, the conflicting pairings were associated with non-mutual inventory between L1 and L2 speakers. As frequency probing confirmed the used inventory difference and its expanding range alongside articulation adeptness, one can expect that the more pronunciation deviation there is, the higher the chance for the speech to utilize these L1-unobserved non-mutuals. Accordingly, English front vowels reported to be particularly difficult for Korean

TABLE I
Dominant Index Pairs of Front Vowel Substitutions

IH to IY	[22,268] (1)	[191, 234]	[22, 268]
IY to IH	[191, 234] (4)	[22, 234]	[42, 234] (1)
AE to EH	[22, 268]	[42, 234]	[191, 234]
EY to EH	[42, 234]	[22, 234] (unobserved)	[22, 319]

learners in [15] commonly used code vector varieties rare in native speech. Their total count in TIMIT is written in parenthesis in Table I. Ref. [15] also notes how the difficulty is manifested as reduced frontal vowel space, when the articulatory confusion transfers to indistinguishability. In line with this understanding, patterns in Table. I experience overlap.

Third, relationships among sub-segmental patterns reflect the acoustic distance expected from L2 variations. The skewed distribution first implies that non-categorical traits are ultimately attributed to the learners' L1 influence. This between-categorical position of phonetic understanding further concurs with the numerical calculation between vectors. As our detected tenseness substitutions solely concerned front vowel pairs, they lacked enough representative samples for pair-level analysis. Hence, we have resorted to examining attributes of individual codewords, which revealed that the rarity stems from conflicting tenseness identity. That is, codebook1 displayed the identity of laxness while codebook2 displayed tenseness. This identity could be instantiated by calculating the tense-to-lax ratio of associated phoneme probabilities. Patterns can then be ranked on a tenseness spectrum with values multiplied from two codebooks. Inferred pattern-wise distance, hereafter, aligned with Euclidean distance. We have demonstrated this process with

a case of IH to IY substitution in Fig. 8.

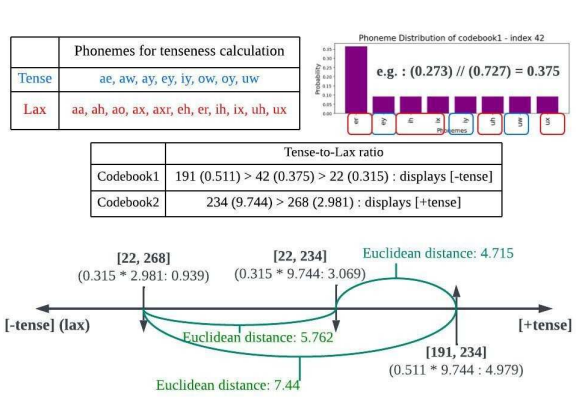


Fig. 8. Pattern Positioning in IH to IY. Although [22, 234] is unobserved, the distance between [22, 234] and [22, 268] is calculable through the second codeword vectors.

VI. Conclusions

[1] This work recreated the unsupervised L2 error pattern discovery experiments using an SSL representation, the Wav2Vec2.0 code vector. Leveraging its unprescribed yet phonetically relevant status, we have identified the range of variations within a single segmental error. While the inventory probing confirmed that encoding units of L2 differed from L1, the pattern discovery revealed that this difference results from an unlikely codebook combination in L2. Namely, the two codeword vectors assumed conflicting identities that formed opposite ends of the error pattern spectrum. As this conflict instantiates L2-specific non-categorical traits, increased usage of relevant index pairs coincided with a higher degree of phonetic divergence. Moreover, sub-segmental pattern distribution reflected the acoustic proximity of corresponding L1 phonemes to two segments participating in substitutions. Thus, the way Wav2Vec2.0 code vectors encode L2 variation is phonetically relevant, making it a valid tool to uncover sub-segmental gradience with the ability to quantify its details. Ultimately, the uncovered details can be employed for finer judgment and feedback that reflects the true nature of mispronunciations.

References

[1] C. L. C. Ng, "Merger of the syllable-initial [n-] and [l-] in hong kong cantonese," 2017.

[2] X. Li, X. Wu, X. Liu, and H. Meng, "Deep segmental phonetic posterior-grams based discovery of non-categories in l2 english speech," *arXiv preprint arXiv:2002.00205*, 2020.

[3] S. Mao, X. Li, K. Li, Z. Wu, X. Liu, and H. Meng, "Unsupervised discovery of an extended phoneme set in l2 english speech for mispronunciation detection and diagnosis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 6244–6248.

[4] Y.-B. Wang and L.-S. Lee, "Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 8232–8236.

[5] Y.-B. Wang and L.-s. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 564–579, 2015.

[6] X. Li, S. Mao, X. Wu, K. Li, X. Liu, and H. Meng, "Unsupervised discovery of non-native phonetic patterns in l2 english speech for mispronunciation detection and diagnosis.," in *INTERSPEECH*, 2018, pp. 2554–2558.

[7] A. Lee and J. Glass, "Mispronunciation detection without nonnative training data," in *Proc. Interspeech 2015*, 2015, pp. 643–647.

[8] A. Lee, N. F. Chen, and J. Glass, "Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6145–6149.

[9] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 382–387.

[10] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8227–8231.

[11] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2007.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[13] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[14] G. Zhao, E. Chukharev-Hudilainen, S. Sonsaat, *et al.*, "L2-arctic: A non-native english speech corpus," 2018.

[15] Y.-J. O. Hee-San Koo, "An Analysis of English Vowels of Korean Learners of English and English Native Speakers," *Korean Education Inquiry*, vol. 16, pp. 1–12, 2001.

[16] E. Hong, "Unsupervised Discovery of Non-Categorical L2 Error Patterns Using Wav2Vec2.0 Code Vector Features," M.A. thesis, College of Hum., Seoul National Univ., Seoul, Aug. 2024.

[1] This paper is based on a part of the first author's 2024 master's thesis [16]

Improved Architecture for High-resolution Piano Transcription to Efficiently Capture Acoustic Characteristics of Music Signals

Jinyi Mi, Sehun Kim and Tomoki Toda
Nagoya University, Japan
E-mail: {mi.jinyi, kim.sehun}@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

***Abstract*—Automatic music transcription (AMT), aiming to convert musical signals into musical notation, is one of the important tasks in music information retrieval. Recently, previous works have applied high-resolution labels, i.e., the continuous onset and offset times of piano notes, as training targets, achieving substantial improvements in transcription performance. However, there still remain some issues to be addressed, e.g., the harmonics of notes are sometimes recognized as false positive notes, and the size of AMT model tends to be larger to improve the transcription performance. To address these issues, we propose an improved high-resolution piano transcription model to well capture specific acoustic characteristics of music signals. First, we employ the Constant-Q Transform as the input representation to better adapt to musical signals. Moreover, we have designed two architectures: the first is based on a convolutional recurrent neural network (CRNN) with dilated convolution, and the second is an encoder-decoder architecture that combines CRNN with a non-autoregressive Transformer decoder. We conduct systematic experiments for our models. Compared to the high-resolution AMT system used as a baseline, our models effectively achieve 1) consistent improvement in note-level metrics, and 2) the significant smaller model size, which shed lights on future work.**

I. Introduction

Automatic music transcription (AMT) has gained considerable research interest in the fields of music signal processing and music information retrieval (MIR) for several decades [1]. The object of AMT is to convert acoustic musical signals into some form of musical notation [2], such as piano rolls, sheet music, and Musical Instrument Digital Interface (MIDI), to improve the time-consuming process of manual music transcription. AMT has been applied in automatic annotation of musical information [3], musical education through automatic instrument tutoring [4], [5], and musicological analysis [6]. In addition, a successful AMT system is also useful for the other tasks in MIR, such as beat tracking [7], chord recognition [8], and performance classification [9].

Piano transcription is a crucial task of AMT, which typically transcribes piano recordings into a series of note events with pitches, onset/offset timings, and velocities. This task is particularly challenging due to its inherent polyphonic nature, i.e., the multiple pitches are usually in the same frame, thereby causing a complex interaction and overlap of harmonics. To address this issue, previous methods, taking into account the piano's acoustic property that the note energy decays after an onset, have mainly focused on adapting models to notes with varying amplitude and harmonics. Inspired by image classification tasks [10], convolutional neural network (CNN)-based methods [11], [12] have been proposed to regress harmonic structures as acoustic representations for audio. On the other hand, recurrent neural network (RNN) methods, such as long short-term memory (LSTM) [13] and gated recurrent unit (GRU) [14], have been applied to capture the medium- and long-range dependencies between notes. Recently, drawing on advancements in speech recognition tasks, convolutional recurrent neural network (CRNN), which combines a CNN acoustic model with an RNN-based sequential model, has become popular for polyphonic piano transcription. Particularly, Kong et al. [15] proposed a high-resolution piano transcription system by regressing precise onset and offset times of notes at arbitrary time resolution, achieving effective performance in piano transcription.

However, there are several limitations of the high-resolution system [15]. First, the harmonics of notes are usually recognized as false positive notes. A potential reason is that the frequency components are calculated with the short-time Fourier transform (STFT), whereas the frequencies that have been chosen to make up the music scale are geometrically spaced, thus yielding components that do not map efficiently to musical frequencies. The second limitation is excessive time and resource consumption due to the large model size. This motivates us to build lightweight architectures for getting around resource constraints and achieve the higher accuracy of transcription.

With this in mind, this work applies an alternative front-end called the Constant-Q Transform (CQT) [16] instead of the STFT to achieve better simulation of the frequencies in music signals. Moreover, we have designed two architectures: the first is based on a CRNN with dilated convolution to well capture a harmonic structure of music signals on CQT feature, and the second is an encoder-decoder architecture that integrates CRNN with a non-autoregressive Transformer (NR-Transformer) decoder. Compared to the high-resolution system, we show that our systems effectively achieve consistent improvement in note-level metrics. Specifically, the proposed

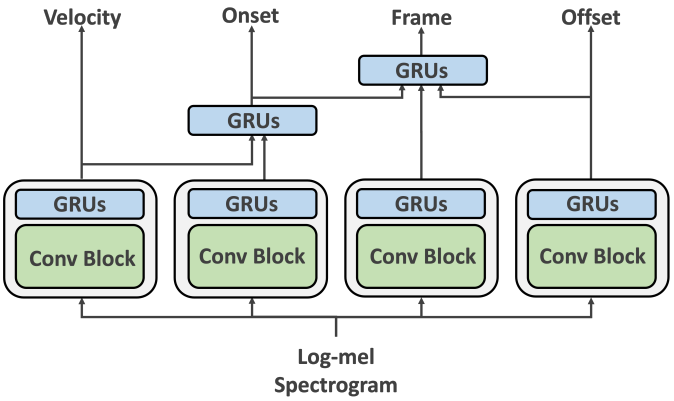


Fig. 1. The high-resolution model.

architectures use significantly fewer parameters of 2.7 million and 0.9 million, respectively, while the high-resolution system uses 20 million parameters. This demonstrates that our systems can achieve the ideal transcription performance without excessive resource consumption.

II. Related work

Since the high-resolution system can be seen as the step stone of our work, we here give a detailed overview of the high-resolution system in this section.

A. Onset and offset times detection

Onsets and offsets represent the beginning and ending of a piano note event, carrying rich information about the piano notes. [17] proposed a dual-objective system for onsets and offsets that conditions the detection of onsets to predict frame-wise outputs. For each piano note, [17] labeled only several consecutive frames of an onset or offset as 1, while other frames was labeled as 0, thereby limiting the transcription resolution. When the precise onset or offset time shifts within a frame, the information about these changes is lost after quantization. Additionally, this labeling method cannot handle cases where the precise onset or offset time is on the boundary between two frames. In light of this, [15] proposed a new labeling method to address the issue that transcription resolutions are limited by the hop size between adjacent frames, by predicting the continuous onset and offset times of piano notes. A new training target was applied to represent the time difference between the center of a frame and its nearest onset or offset times of a note. The process of encoding the time difference into the training targets can be formulated as

$$\begin{cases} g\left(\Delta_i\right) = 1 - \frac{|\Delta_i|}{J\Delta}, |i| \leq J \\ g\left(\Delta_i\right) = 0, |i| > J, \end{cases} \tag{1}$$

where Δ and Δ_i are the frame hop size time and the time difference, respectively. i is the index of a frame, where negative and positive i values indicate the previous and future frame indexes of an onset or offset. J is a hyperparameter that is used to control the sharpness of the targets, i.e., the smaller the J, the sharper the targets $g(\Delta_i)$.

B. High-resolution piano transcription system

Fig. 1 shows the architecture of the high-resolution model that we use as the baseline. The log-mel spectrogram with a shape of $T \times F$ calculated from STFT spectrum is the input feature, where T is the number of frames, and F is the number of mel frequency bins. The frame, onset, offset, and velocity tasks share the same acoustic module. This acoustic module, consisting of a convolution block with several convolutional layers and a bidirectional gated recurrent unit (biGRU) layer, is able to extract spectral and temporal information from the log-mel spectrogram. Then, a fully connected layer is applied to output the results of the acoustic model with a shape of $T \times K$, where K is the number of pitch classes. The prediction outputs of velocities are concatenated with outputs of the onsets from the acoustic model. This concatenated data is then fed into a biGRU layer to calculate the final onset predictions. Similarly, the predicted onsets and offsets are used as conditional information to predict frame-wise outputs in the same way.

III. Proposed method

At the core of our approach is the idea of effectively simulating the frequencies of music signals to improve the performance of the high-resolution system. We have designed two architectures: the first is based on a CRNN with dilated convolution (See Section III-A), while the second combines a CRNN with a NR-Transformer decoder in an encoder-decoder architecture (See Section III-B).

A. Improved CRNN Model for the High-Resolution System

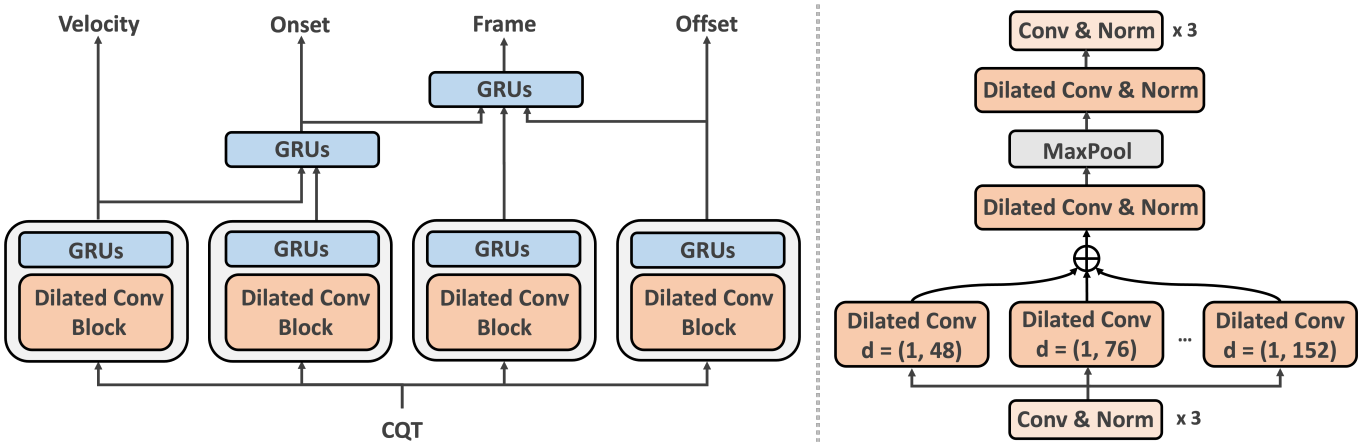
Fig. 2(a) shows the overall architecture of our proposed CRNN model for improved high-resolution system. The CQT spectrogram is applied as input instead of the log-mel spectrogram. Onset, offset, frame and velocity tasks share the same acoustic model that includes a dilated convolution block and a biGRU layer. The output of the velocity serves as conditional information for the onset, while the outputs of both onset and offset condition the frame. We denote this model as HRplus.

1) Inputs and Outputs: To better simulate the frequencies in music signals, we use CQT as the input representation. Unlike the traditional Fourier transform or the STFT which employ a linear frequency scale, the CQT utilizes a logarithmic frequency scale that closely approximates musical frequencies. As described in [18], the distance d_k between the fundamental frequency f_0 and the k-th overtone is given by

$$\begin{aligned} d_k &= \log_{2^{1/Q}}\left(k \cdot f_0\right) - \log_{2^{1/Q}}\left(f_0\right) \\ &= Q \cdot \log_2(k), \end{aligned} \tag{2}$$

where Q is the constant factor of the filter banks that indicates the number of frequency bins per octave.

The outputs of models consist of frame, onset, offset, and velocity events, where onsets and offsets are represented as continuous events, as described in Section II-A. Then, we post-process note-wise events into a set that contains onset time,



(a) Overall structure of HRplus.

(b) Dilated Conv Block.

Fig. 2. Illustration of HRplus model architecture. Norm denotes instance normalization with a ReLu activation, d denotes the dilation rate.

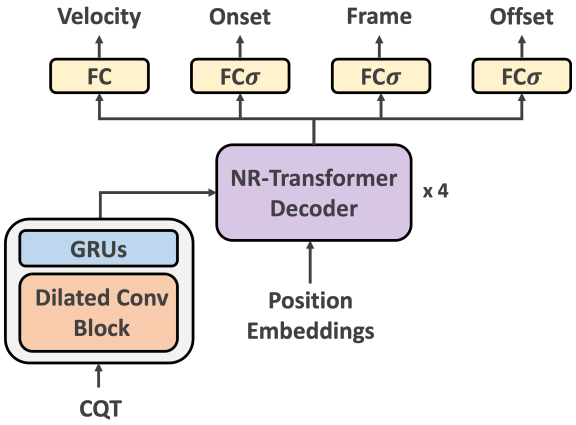


Fig. 3. Illustration of HRplus-hybrid model architecture. FC denotes a fully connected layer, σ denotes a sigmoid function.

offset time, and velocities, using the algorithm of the high-resolution system. This post-process proceeds in the following steps:

Step 1. Note onset detection. If the value of an onset event in a frame exceeds the onset threshold, and this value is a local maximum, this frame is detected to contain an onset. Then, the precise onset time is calculated.

Step 2. Velocity scaling. MIDI files use integers between 0 and 127 to represent the velocity of notes. Since we normalize the dynamic range of velocities from $[0, 127]$ to $[0, 1]$ during model training, the predicted velocities are scaled back to the $[0, 127]$ range if an onset is detected in Step 1.

Step 3. Note offset detection. For the onset detected in Step 1, an offset is detected if the offset prediction output exceeds the offset threshold, or if any frame prediction outputs fall below the frame threshold. Then, the precise offset time is calculated. Additionally, when consecutive onsets of the same pitch are detected, the previous onsets are truncated by adding offsets.

2) Acoustic Model for CQT Input Representation: Considering the challenge of capturing and analyzing multi-scale frequency information scattered in the CQT, we use a dilated convolution block and a biGRU layer as the acoustic model, inspired by HPPNet-sp [19]. The detail of dilated convolution block is shown in Fig. 2(b). The first three convolution layers, each with a kernel size of 7×7 and equipped with instance normalization and ReLU activation, extract local information from the CQT input. Subsequently, eight dilated convolution layers with kernel sizes of 1×3 apply different dilation rates of 48, 76, 96, 111, 124, 135, 144, and 152, respectively, to capture the harmonic series. These dilation rates, calculated by Eq. (2) with $Q = 48$, correspond to the intervals between adjacent harmonics in a harmonic series, i.e., d_2, d_3, ..., d_9 in Eq. (2). The outputs of the eight dilated convolution layers are combined and then fed into a dilated convolution layer with a dilation rate of 48 and a kernel size of 1×3, equipped with instance normalization and ReLU activation. Subsequently, a max-pooling layer with a pooling size of 4 along the frequency axis reduces the frequency bins to match the number of pitch classes. Finally, a dilated convolution layer with a kernel size of 1×3, a dilation rate of 12, instance normalization, and ReLU activation is applied, followed by three convolution layers with kernel sizes of 5×1.

B. Hybrid CRNN-Transformer Encoder-Decoder Model

We have designed an encoder-decoder architecture integrating the acoustic model of HRplus with the NR-Transformer decoder, as shown in Fig. 3. We denote this model as HRplus-hybrid. The motivations for designing this encoder-decoder architecture are derived from two main considerations: 1) the acoustic model of HRplus could serve exclusively as the encoder of HRplus-hybrid, so as to focus on extracting features from the input and transforming them into a intermediate high-dimensional representation; and 2) the strength of Transformer decoder in capturing long-term dependencies would be leveraged, thanks to which these intermediate representations could be effectively decoded to more accurate outputs even if reducing the model size. The decoder employs the NR-

Transformer decoder, which includes four decoder blocks stacked in series. Each NR-Transformer decoder block includes a self-attention module, a cross-attention module, and a feed-forward module. The outputs of the encoder and a trainable positional embedding are used to calculate the cross-attention. Finally, a fully connected layer outputs the predictions for onset, offset, frame, and velocity.

C. Loss functions

We use a similar loss calculation as the high-resolution system for the proposed models, with the total loss L_{total} being the sum of the losses for onset L_{on}, offset L_{off}, frame L_{fr}, and velocity L_{vel} as

$$L_{total} = L_{on} + L_{off} + L_{fr} + L_{vel}. \tag{3}$$

We denote the continuous target and prediction of onset and offset as g_{on}, $\hat{g}_{on}$, g_{off}, and $\hat{g}_{off}$, respectively, where they are regarded as the probability of a binary variable (i.e., from 0 to 1), and the binarized target and prediction of frame as b_{fr} and $\hat{b}_{fr}$ (i.e., 0 or 1 for the target and from 0 to 1 for the prediction). The onset loss L_{on}, offset loss L_{off}, and frame loss L_{fr} are represented as

$$L_{on} = \sum_{t=1}^{T}\sum_{k=1}^{K} L_{bce}(g_{on}(t,k), \hat{g}_{on}(t,k)), \tag{4}$$

$$L_{off} = \sum_{t=1}^{T}\sum_{k=1}^{K} L_{bce}(g_{off}(t,k), \hat{g}_{off}(t,k)), \tag{5}$$

$$L_{fr} = \sum_{t=1}^{T}\sum_{k=1}^{K} L_{bce}(b_{fr}(t,k), \hat{b}_{fr}(t,k)). \tag{6}$$

L_{bce} is a binary cross-entropy loss function defined as

$$L_{bce}(y, \hat{y}) = -y \cdot log(\hat{y}) - (1-y) \cdot log(1-\hat{y}), \tag{7}$$

where y denotes target and $\hat{y}$ denotes prediction. Since we predict the velocity only where the onset is detected, the velocity loss L_{vel} is represented as

$$L_{vel} = \sum_{t=1}^{T}\sum_{k=1}^{K} b_{on}(t,k) \cdot L_{bce}(b_{vel}(t,k), \hat{b}_{vel}(t,k)), \tag{8}$$

where b_{vel} and $\hat{b}_{vel}$ are the binarized target and prediction of velocity, $b_{on} \in \{0,1\}^{T\times K}$ indicates the presence or absence of note onsets.

IV. Experimental Evaluations

A. Datasets

To evaluate the performance of our systems on the piano transcription task, we used the MIDI and Audio Edited for Synchronous Tracks and Organization (MAESTRO) [20] dataset, a large-scale piano dataset containing about 200 hours of paired CD-quality audio recordings and MIDI files from ten years of the International Piano-e-Competition. These audio recordings and MIDI files are aligned with around 3 ms accuracy and sliced into individual musical pieces, which are annotated with composer, title, and year of performance. Virtuoso pianists performed on Yamaha Disklaviers, concert-quality acoustic grand pianos, integrated with a high-precision MIDI capture and playback system. To compare against the baselines, we trained and evaluated on MAESTRO v2 and MAESTRO v3 datasets, respectively. We used the train/validation/test split following by the official configuration of MAESTRO dataset without any extensions or augmentations. The total duration of each split in hours are 161.3/19.4/20.5 in MAESTRO v2 and 159.2/19.4/20.0 in MAESTRO v3, respectively.

B. Experimental Setup

We used PyTorch [21] to implement our systems. The audio recordings were split into 20-second pieces and resampled to 16 kHz so that all the frequencies of the piano could be covered. Then, we down-mixed the audio into a single channel and converted it to a CQT using the nnAudio toolkit [22] with a hop length of 320 points, 48 bins per octave, resulting in a total of 352 frequency bins. For high-resolution label for onset and offset, we set the hyperparameter $J = 5$, i.e., each onset or offset affects the regression values of $2 \times J = 10$ frames.

All the models we proposed were trained with the Adam optimizer [23] with a batch size of 2 and a learning rate of 6×10^{-4}. We used the ReduceLROnPlateau scheduler in PyTorch for learning rate scheduling, employing its default parameters. The best models were determined by the performance on the validation set. At inference, the velocity threshold was set to 0, while all other thresholds were set to 0.4. The outputs were converted to MIDI events as described in Sections III-A1.

C. Baselines

Since the proposed models are optimized based on the high-resolution system, and the acoustic model is inspired by HPPNet-sp, we use both High-resolution [15] and HPPNet-sp [19] as baselines. The results are shown in Table I and Table II. High-resolution applies the high-resolution labels for note onset and offset to construct the piano transcription system. HPPNet-sp uses dilated convolution and frequency-grouped LSTM to model the harmonic structure and pitch-invariance over time in piano transcription.

D. Evaluation Metrics

For evaluation metrics of the piano transcription systems, we used note-level metrics involving the standard precision, recall, and F1 score. These metrics match each predicted note with a ground truth note, considering onset times, pitches, and optional offset times and velocities. We used the mir_eval library [24] for all metric calculations. Following the default configuration of mir_eval, we applied a 50 ms tolerance for note onsets, a 20% offset ratio for note offsets, and a tolerance of 0.1 for velocities.

E. Experimental Results

Table I shows the evaluation results for the MAESTRO v2. For the MAESTRO v2 dataset, HRplus significantly outperforms High-resolution on all metrics. Besides, HRplus uses

TABLE I
TRANSCRIPTION RESULTS EVALUATED ON THE MAESTRO V2 DATASET
(P: PRECISION, R: RECALL, **BOLD**: BEST SCORE)

Model	Params	Onset			Onset & Offset			Onset, Offset & Velocity		
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
High-resolution [15]	20M	98.17	95.35	96.72	83.68	81.32	82.47	82.10	79.80	80.92
HRplus	**2.7M**	**98.96**	**95.81**	**97.34**	**86.05**	**83.36**	**84.67**	**84.21**	**81.59**	**82.86**

TABLE II
TRANSCRIPTION RESULTS EVALUATED ON THE MAESTRO V3 DATASET
(P: PRECISION, R: RECALL, **BOLD**: BEST SCORE, UNDERLINE: SECOND BEST SCORE)

Model	Params	Onset			Onset & Offset			Onset, Offset & Velocity		
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
High-resolution [reproduced]	20M	98.22	95.26	96.69	83.33	80.86	82.06	81.68	79.28	80.44
HPPNet-sp [19]	<u>1.2M</u>	98.45	**95.95**	<u>97.18</u>	84.88	82.76	83.80	83.29	<u>81.24</u>	82.24
HRplus	2.7M	**99.01**	<u>95.86</u>	**97.39**	**86.14**	**83.46**	**84.76**	**84.31**	**81.69**	**82.96**
HRplus-hybrid	**0.9M**	<u>98.68</u>	94.92	96.73	<u>85.98</u>	<u>82.77</u>	<u>84.32</u>	<u>83.91</u>	80.79	<u>82.30</u>

2.7 million parameters, markedly fewer than the 20 million parameters used by High-resolution. This demonstrates that our model architecture is more lightweight and cost-efficient.

To provide a more comprehensive comparison between our proposed models and the baselines, we train and evaluate High-resolution on the MAESTRO v3 dataset. The results are shown in Table II. From the baseline perspective, we observe that HPPNet-sp outperforms High-resolution in note-level metrics. Meanwhile, HPPNet-sp also has a significantly smaller model size compared to High-resolution. When comparing our proposed models with the baselines, it is observed that HRplus wins in eight out of nine test metrics, especially in the F1 score, which is significantly better than all other systems. Moreover, we find that even with the smallest model size of only 0.9 million parameters, HRplus-hybrid still achieves the second best results in six out of nine test metrics. It outperforms all the baselines in F1 score for Onset & Offset and Onset, Offset & Velocity. Specifically, HRplus-hybrid surpasses the High-resolution system in both F1 score and Precision across all metrics. This demonstrates our proposed models, combining the advantages of High-resolution and HPPNet, are very effective in improving AMT performance as well as reducing resource consumption.

We specifically analyze the model size among different systems. First, since the dilated convolution can significantly reduce the number of parameters, the model size of HRplus is greatly reduced compared to that of High-resolution. Second, in contrast to the approach taken by HRplus which utilizes different GRU+dilated conv block for individual outputs, HRplus-hybrid shares one GRU+dilated conv block across all outputs, thus achieving the smallest model size.

In addition, we note that the results of HR-hybrid are slightly inferior to those of HRplus. A potential reason is that HRplus-hybrid does not utilize conditional information in its prediction outputs. Conversely, the use of conditional information, especially onset information, has significantly enhanced the final transcription performance, as evidenced in the studies of [15], [17], [19].

V. CONCLUSIONS

In this paper, we utilize the CQT as the input representation instead of the STFT to better adapt to musical signals, thereby improving the transcription performance of the high-resolution system. We design two architectures: the first is based on a CRNN with dilated convolution, and the second is an encoder-decoder architecture that integrates CRNN with a NR-Transformer decoder. The experimental results demonstrate that our proposed methods can effectively achieve consistent improvements in note-level metrics without any extensions or augmentations. Specifically, the proposed architectures use only 2.7 million and 0.9 million parameters, respectively, compared with the 20 million used by the high-resolution system. Therefore, our systems can achieve the ideal transcription without excessive resource consumption. In the future, we will extend our models to other instruments and introduce transfer learning methods to enhance transcription performance.

ACKNOWLEDGMENT

This work was partly supported by JST CREST JPMJCR19A3, Japan. In addition, this work was also financially supported by JST SPRING, Grant Number JPMJSP2125. The author would like to take this opportunity to thank the "THERS Make New Standards Program for the Next Generation Researchers."

References

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2018.

[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, pp. 407–434, 2013.

[3] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1180–1188.

[4] E. Cano, C. Dittmar, J. Abeßer, C. Kehling, and S. Grollmisch, "Music technology and education," *Springer Handbook of Systematic Musicology*, pp. 855–871, 2018.

[5] P. Dunbar-Hall, "Music transcription as pedagogy: Discussion of a cross-disciplinary approach to teacher preparation," in *Celebrating Musical Communities: Proceedings of the 40th Anniversary National Conference, Perth 6th-8th July 2007*, Australian Society for Music Education Nedlands, WA, 2007, pp. 89–93.

[6] A. Klapuri and M. Davy, "Signal processing methods for music transcription," 2007.

[7] R. Vogl, M. Dorfer, G. Widmer, and P. Knees, "Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks.," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2017, pp. 150–157.

[8] Y. Wu and W. Li, "Automatic audio chord recognition with midi-trained deep feature and blstm-crf sequence decoding model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 355–366, 2018.

[9] S. Kim, H. Lee, S. Park, J. Lee, and K. Choi, "Deep composer classification using symbolic representation," *arXiv preprint arXiv:2010.00823*, 2020.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[11] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," *arXiv preprint arXiv:1612.05153*, 2016.

[12] R. Kelz, S. Böck, and C. Widnaer, "Multitask learning for polyphonic piano transcription, a case study," in *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*, IEEE, 2019, pp. 85–91.

[13] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2012, pp. 121–124.

[14] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, "An end-to-end framework for audio-to-score music transcription on monophonic excerpts," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2018, pp. 34–41.

[15] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[16] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[17] C. Hawthorne, E. Elsen, J. Song, *et al.*, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2018, pp. 50–57.

[18] W. Wei, P. Li, Y. Yu, and W. Li, "Harmof0: Logarithmic scale dilated convolution for pitch estimation," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2022, pp. 1–6.

[19] W. Wei, P. Li, Y. Yu, and W. Li, "Hppnet: Modeling the harmonic structure and pitch invariance in piano transcription," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2022.

[20] C. Hawthorne, A. Stasyuk, A. Roberts, *et al.*, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=r1lYRjC9F7.

[21] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[22] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans, "Nnaudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks," *IEEE Access*, vol. 8, pp. 161 981–162 003, 2020.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] C. Raffel, B. McFee, E. J. Humphrey, *et al.*, "Mir_eval: A transparent implementation of common mir metrics.," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, vol. 10, 2014, p. 2014.

Ev3DGS: Event Enhanced 3D Gaussian Splatting from Blurry Images

Junwu Huang, Zhexiong Wan, Zhicheng Lu, Juanjuan Zhu, Mingyi He, Yuchao Dai†
School of Electronics and Information & Shaanxi Key Laboratory of Information Acquisition and Processing
Northwestern Polytechnical University, Xi'an, Shaanxi 710129, China

***Abstract*—The novel view synthesis task involves inputting a source image, a source pose and a target pose, and rendering to generate a corresponding target image. However, obtaining a clear novel view synthesized image from only a set of blurred images and corresponding poses is a challenging problem, which often occurs in reality. To solve this problem, the good performance of 3D Gaussian Splatting (3DGS) in the field of 3D scene reconstruction is taken into account, as well as the remarkable effectiveness of event cameras in the deblurring problem. Inspired by the novel Event-Enhanced Neural Radiance Fields (E2NeRF) model, which is also based on event enhancement, a new 3D reconstruction framework, Event-Enhanced 3DGS (Ev3DGS), based on 3DGS is proposed by utilizing the combined data from event cameras and standard RGB cameras. We effectively introduce the event stream into the 3D Gaussian iterative process by constructing the blur rendering loss and event rendering loss, which guides the optimization of the network structure by predicting the blurred image and event generation processes. Compared with E2NeRF, the Ev3DGS proposed in this paper effectively improves the rendering performance with 4.8% and 2.5% improvement in PSNR and SSIM, and 15% reduction in LPIPS, while significantly reducing the training time consumption. Extensive experiments on both synthetic and real-world datasets show that Ev3DGS can effectively learn clear 3DGS from blurred image inputs, enabling high-quality novel view synthesis and making 3DGS more robust. Our code is publicly available at https://github.com/npucvr/Ev3DGS.**

I. Introduction

The novel view synthesis has a wide range of applications in 3D scene reconstruction, virtual/reality augmentation and other fields. In recent years, with the development of differentiable rendering technology, the novel view synthesis has received widespread attention again, and has shown superior performance and development prospects. Among them, the appearance of 3D Gaussian Splatting (3DGS) [1] has greatly accelerated the rendering speed of novel view synthesis. However, it is difficult for 3DGS to obtain a clear target image from a blurred source image input, as often occurs in reality, which poses a challenge to synthesize clear novel views using 3DGS.

Event camera, as a new type of bionic vision sensor, has a wide range of applications in the fields of feature detection and tracking, optical flow estimation [2] [3], video frame prediction [4], 3D reconstruction and pose estimation. It is promising to guide the learning process of 3DGS by introducing additional information carried by event data. Therefore, in order to solve the problem of 3DGS input source image blurring, this paper investigates event-enhanced 3DGS for blurred images.By utilizing event and image data acquired by an event camera and introducing blur rendering loss and event rendering loss, a new event-enhanced 3DGS-based model (Ev3DGS) is constructed, which enables Ev3DGS to efficiently learn from blurred images to a clear Gaussian scenes, thus realizing clear novel view synthesis.

Specifically, inspired by the Event-Enhanced Neural Radiance Fields (E2NeRF) [5] model, during the training process, we superimpose the clear images obtained by rendering multiple predicted poses at equal time intervals under one viewpoint as the predicted blurred images, and compare them with the input blurred images as our blur rendering loss. Secondly, the generation process of predicted event data is simulated based on the brightness change caused by the change of camera position and compared with the real event data to get the event rendering loss. By taking advantage of the high dynamic range, high temporal resolution, and low latency of the event camera, the event data can effectively enhance the 3DGS network so that we can learn a network with a clearer output image, which makes it possible to not only realize the deblurring of the input image, but also obtain high-quality synthetic images of the novel view.

Our proposed Event-Enhanced 3DGS (Ev3DGS) in this paper is tested and verified with E2NeRF synthetic dataset and real-world dataset. The comparisons are carried out with the E2NeRF model to test the effectiveness of the model. In addition this paper also carries out ablation experiments to demonstrate the effectiveness of event rendering loss. The contributions of this paper can be summarized as follows:

(1) An event-enhanced 3DGS model (Ev3DGS) is proposed, which is the first framework to reconstruct clear 3DGS from blurred images and corresponding event data using blur rendering loss and event rendering loss. The modeling framework effectively exploits the intrinsic relationship between events and images to significantly improve the performance and robustness of 3DGS;

(2) Both the Blur-rendering loss and event-rendering loss are introduced into 3DGS, which enhances the object clarity of reconstruction from 3DGS.

† Corresponding author (daiyuchao@nwpu.edu.cn)

This research was partly supported by the National Natural Science Foundation of China (62271410) and the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (CX2023013).

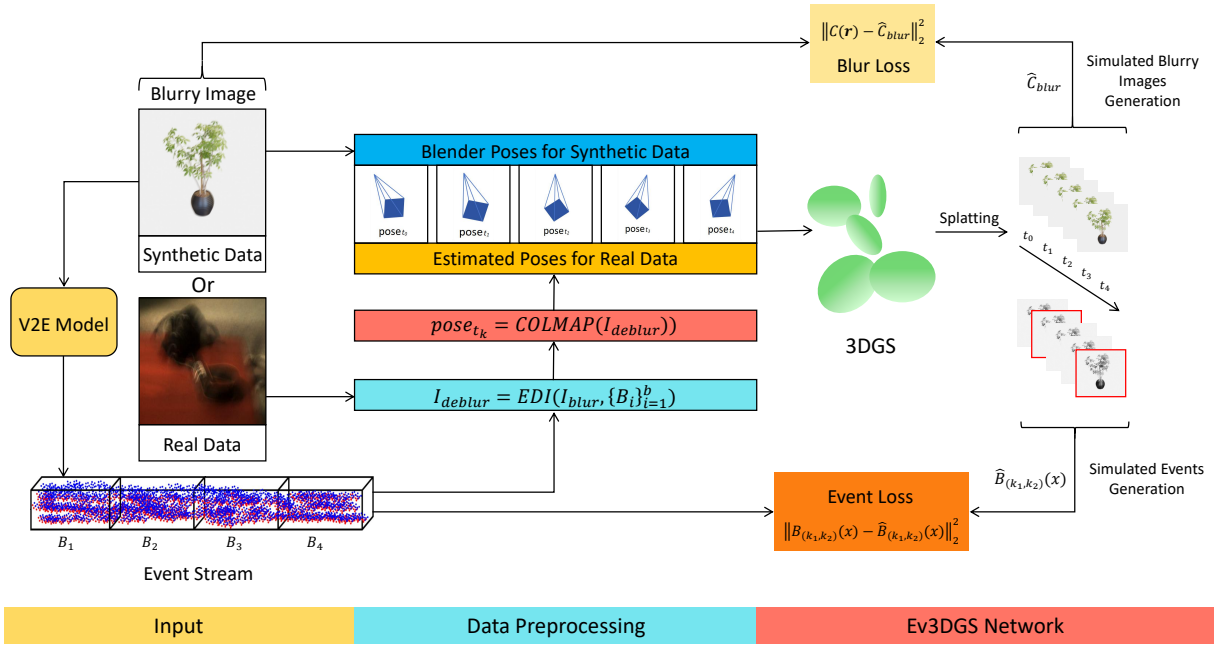


Fig. 1. Ev3DGS Model Framework. Its inputs are blurry images and corresponding event streams of every views.This Ev3DGS model introduces a blur rendering loss and an event rendering loss. The blur rendering loss simulates the generation process of blurred images to provide more information about the scene texture details to the network. The event rendering loss, on the other hand, introduces event data into the 3DGS training process, enabling the network to better learn the real 3D volume representation and obtain information about the motion characteristics of objects.

II. Related Work

A. 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [1] is a transformative technology in computer graphics in recent years. Recent innovative researches employing 3DGS in SLAM (Simultaneous Localization and Mapping) have demonstrated the great potential and versatility of 3DGS in SLAM. For example, GS-SLAM [6] employs an adaptive strategy to add or remove 3D Gaussian distributions to optimize the scene geometry reconstruction and improve the mapping of previously observed regions. To extend the concept of 3DGS to dynamic scenes, Lu *et al.* [7] proposes a 3D geometry-aware deformable Gaussian Splatting method for dynamic view synthesis, while SC-GS [8] utilizes sparse control points and deformed Multilayer Perceptrons (MLPs) to capture and represent the dynamics of a 3D scene. For the problem of blurring of the input image due to camera shake, Deblur-GS [9] relied on the interpolation method to estimate the motion trajectory information of the camera to realize the reconstruction of a clear 3D scene. BAD-Gaussian [10] performs batch clustering on top of this, which enhances the interrelationships of the images in the same exposure time and can better capture the complex motion trajectories during the exposure period.

B. Event Camera

Event Camera is a new type of biologically inspired vision sensor, sometimes called Dynamic Vision Sensor (DVS, dynamic vision sensor) or DAVIS (Dynamic and Active-Pixel Vision Sensor) [11]. Unlike standard cameras that capture images at a fixed frame rate, event cameras generate event streams by sensing changes in pixel brightness, which is characterized by low latency, high dynamic range, low power consumption, and high temporal resolution. It is also widely used in some traditional vision tasks such as feature detection and tracking, optical flow estimation, 3D reconstruction and pose estimation. In the field of optical flow estimation, Zhu *et al.* [12] utilized a self-supervised scheme to train the encoder and decoder of Convolutional Neural Network (CNN) for dense optical flow estimation, while Ye *et al.* [13] proposed an innovative monocular neural network structure that relies only on event data to estimate dense optical flow, depth and self-motion. In the field of 3D reconstruction, many active event-based 3D reconstruction methods have also been proposed. For example, Brandli *et al.* [14] successfully combined a DVS with a pulsed line laser for fast terrain reconstruction. Motion-contrast 3D scanning [15], on the other hand, uses structured light technology to synchronize high resolution, high speed and excellent performance in complex 3D scanning environments. For the event-enhanced new perspective synthesis task, Qi *et al.* [5] proposed the Event-Enhanced Neural Radiance Fields model E2NeRF, which constructs two losses and well utilizes the event properties to assist in solving the problem of blurring of the input image and realizes a clear three-dimensional reconstruction, but its training speed is slow due to the limitation of the Neural Radiance Fields (NeRF) model.

III. Prerequisite

A. 3D Gaussian Splatting

3D Gaussian Splatting [1] is an explicit rasterization technique for real-time radial field rendering described by 3D

Gaussian distributions, which allows for real-time rendering of photorealistic scenes learned from small image samples. The input to 3D Gaussian Splatting is a set of images of a static scene and a sparse point cloud obtained by the camera from the alignment. On the sparse points, a set of 3D Gaussian distributions are created, defined by the position $\boldsymbol{x}$, opacity α, covariance matrix $\boldsymbol{\Sigma}$, and Spherical Harmonic (SH) coefficients to model the view-dependent color, while parameter optimization is performed by an adaptive density control algorithm. The Gaussian distribution is defined as:

$$G(\boldsymbol{x}) = e^{-\frac{1}{2}(\boldsymbol{x})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x})}. \tag{1}$$

For regular optimization of the covariance matrix $\boldsymbol{\Sigma}$, given a scaling matrix $\boldsymbol{S}$ and a rotation matrix $\boldsymbol{R}$, the corresponding $\boldsymbol{\Sigma}$ can be found:

$$\boldsymbol{\Sigma} = \boldsymbol{R}\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{R}^{\mathrm{T}}. \tag{2}$$

To allow independent optimization of these two factors, they are further represented as a three-dimensional vector $\boldsymbol{s}$ for scaling and a quaternion $\boldsymbol{r}$ for rotation.

In addition, the key to 3DGS's improved model rendering rate is the construction of a fast micronizable rasterizer. The overall fast rendering and sorting is achieved by chunking the Gaussian sphere splash and allowing the Gaussian spheres with approximate a-values to be blended. And the fast rasterizer utilizes fast inverse passes to track through the accumulated a-values, allowing it to receive splat balls with gradients without restriction for scene representation. For each pixel, its color C is determined by all Gaussian distributions covering the pixel, which are represented as follows:

$$C = \sum_{i \in N_{cov}} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{3}$$

where N_{cov} represents the splats that cover this pixel, α_i represents the opacity of this Gaussian splat multiplied by the density of the projected 2D Gaussian distribution at the location of the pixel, and c_i represents the computed color.

B. Event Generation

The event camera is a novel bionic vision sensor that asynchronously and independently measures the brightness change of each pixel in the scene, and triggers the generation of an event signal when the brightness change of any pixel exceeds a threshold set by the event camera. Each event consists of the spatio-temporal coordinates of the triggering event (pixel position coordinates (x, y) and the triggering timestamp t with millisecond accuracy) and its polarity $\sigma = \pm 1$. The event camera is able to measure the brightness change of each pixel in the scene asynchronously and independently [16].

However, when the object moves, the event camera asynchronously generates event signals by sensing changes in image intensity, producing a series of event sequences, denoted as:

$$\sigma = \boldsymbol{\Gamma}\left(\log\left(\frac{L_{xy}(t)}{L_{xy}\left(t_{ref}\right)}\right), c\right), \tag{4}$$

where the potential image $L_{xy}(t)$ and $L_{xy}(t_{ref})$ represents the intensity at point (x, y) at moment t and t_{ref}, c is the response threshold for intensity change, and $\boldsymbol{\Gamma}(\cdot, \cdot)$ is a truncated function:

$$\boldsymbol{\Gamma}(d, c) = \begin{cases} +1, & d \geq c \\ 0, & d \in (-c, c) \\ -1, & d \leq -c \end{cases}, \tag{5}$$

where d is the intensity of the corresponding brightness change.

According to the general principles of event generation, the most primitive and complete representation of all the information in each event is to represent the event as an n × 4 dimensional matrix, where each column contains all the information of a single event: the 2D coordinates, the timestamp and the event polarity. Since only the brightness change of the object due to the motion process is of interest, in this paper, we consider equating the event data into b blocks and discretizing the time to superimpose the event data between any two potentially clear images to obtain the event change data containing only the 2D coordinates and event polarity.

C. Blurry Image Generation

The specific process of generating an image by means of a color camera can be represented as the process by which the camera sensor collects photons during the exposure and converts them into measurable charges. The degree of motion blur in the image depends on the motion of the camera during the exposure time. For example, a slow moving camera produces minimal relative motion, especially at shorter exposure times, whereas a fast moving camera produces a motion blurred image, especially in low light scenes with longer exposure times. The phenomenon may be expressed as an integral over a series of virtual potentially sharp images, denoted as follows:

$$I_{blur} = \phi \int_0^{\tau} I_t dt, \tag{6}$$

where I_{blur} denotes the real captured motion blurred image, ϕ is used as a normalization factor, τ is the camera exposure time, and I_t is the potentially clear image captured at timestamp $t \in [0, \tau]$ during the exposure time. The blurred image I_{blur} due to camera motion during the exposure time is calculated by averaging the potentially clear image I_t for each different timestamp t. The model is based on a discrete approximation. The discrete approximation of the model depends on the number k of discrete moments, denoted as follows:

$$I_{blur} = \frac{1}{b+1} \sum_{k=0}^{b} I_k. \tag{7}$$

Thus, for the blurred input of 3DGS, we can consider it as a superposition of a series of potentially clear images. So when acquiring the real world dataset and when making the synthetic dataset, we consider the blurred image at one viewpoint as an average of $b + 1$ potentially clear images. The preprocessing for the model input data will also be centered around this.

IV. METHOD

Fig. 1 shows the overall framework of Ev3DGS, which effectively improves the volumetric representation of 3DGS by introducing two new losses in the framework of 3DGS, and also designs an event-blurred image-based bit-pose estimation framework to efficiently deal with real-world data.Ev3DGS takes the blurred image's and the corresponding events as inputs for each viewpoint. Meanwhile Ev3DGS model introduces blur rendering loss and event rendering loss. The blur rendering loss simulates the process of blurred image generation and provides more information about the scene texture details to the network. The event rendering loss, on the other hand, introduces event data into the 3DGS training process, enabling the network to better learn the real 3D volumetric representation and obtain information about the motion characteristics of objects. After inputting the image bitmap into 3DGS, 3DGS generates an initial point cloud via for reconstruction rendering. After each rendering, the predicted blurred image is obtained by averaging the image sequences cumulatively, while the predicted event stream is calculated by converting the image sequences into gray-scale maps. The input blurred image and event stream are compared with the predicted values to obtain the blur rendering loss and event rendering loss for model supervision.

A. Blur Rendering Loss

With $b+1$ poses $\{\boldsymbol{P}_k\}_{k=0}^{b}$ of each view, we can get $b+1$ rays $\{\boldsymbol{r}_k\}_{k=0}^{b}$ emitted from each pixel. With the 3DGS network, each pixel can get $b+1$ color values $\left\{\widehat{C}_k = C\left(\boldsymbol{r}_k\right)\right\}_{k=0}^{b}$ regarded as the process of blurry pixel generation, and the average of the results is used as the predicted blurry color:

$$\hat{C}_{blur} = \frac{1}{b+1}\sum_{k=0}^{b} C\left(\boldsymbol{r}_k\right). \tag{8}$$

The loss function of 3DGS with blurred image as input is then expressed as:

$$\mathcal{L}_{blur} = \sum_{\boldsymbol{r}\in\mathcal{R}}\left[\left\|\hat{C}^{c}_{blur} - C(\boldsymbol{r})\right\|_2^2 + \left\|\hat{C}^{f}_{blur} - C(\boldsymbol{r})\right\|_2^2\right], \tag{9}$$

where 3DGS is designed using joint optimization of coarse and fine models.

3DGS uses stochastic gradient descent, utilizing standard GPU acceleration frameworks and adding custom CUDA cores for some operations. A sigmoid function is used for α to keep it constrained to $[0, 1)$, and an exponential activation function is used for the scaling factor of the covariance to ensure a smooth gradient.

Initialize the covariance to an isotropic Gaussian whose axis length is the same as the mean of the distances to the 3 nearest points. Standard exponential decay scheduling techniques were used for the location of the Gaussian. The loss function is the loss $\mathcal{L}_2$ and D-SSIM terms:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_2 + \lambda\mathcal{L}_{D-SSIM}, \tag{10}$$

where λ is a weighting factor.

The loss $\mathcal{L}_2$ and D-SSIM terms are standardized metrics.The loss function for 3DGS is slightly different from that of NeRF. Due to the high computational cost of NeRF by light sampling the sample points, NeRF is usually at the pixel level, whereas 3DGS performs the computation of the loss at the image level.

B. Event Rendering Loss

Image blur generation is a continuous process, however blur rendering loss only uses discrete $b+1$ frames corresponding to $b+1$ poses in the blurred image to model the blurring process, losing a large amount of information in the blur generation process. Therefore by introducing event rendering loss, the high temporal resolution property of event data is utilized to supervise the continuous blurring process between any two predicted frames. Given a pixel $\boldsymbol{x} = (x, y)$, two color values of C_{k_1} and C_{k_2} $(k_1 < k_2)$ are randomly selected from the $\left\{\widehat{C}_k\right\}_{k=0}^{b}$ of that pixel, and they are converted to grayscale values to obtain L_{k_1}, L_{k_2}. According to the principle of event generation, the difference between the logarithmic values of the two grayscale values of L_{k_1}, L_{k_2} is divided by a threshold θ to estimate the number of events between any two frames for a given pixel $\boldsymbol{x}$:

$$\widehat{B}_{(k_1,k_2)}(\boldsymbol{x}) = \begin{cases} \left[\frac{\log\left(L_{k_2}\right)-\log\left(L_{k_1}\right)}{\theta_{neg}}\right], L_{k_2} < L_{k_1} \\ \left[\frac{\log\left(L_{k_2}\right)-\log\left(L_{k_1}\right)}{\theta_{pos}}\right], L_{k_2} \geq L_{k_1} \end{cases}. \tag{11}$$

The mean square error between the estimated number of events $\widehat{B}_{(k_1,k_2)}(\boldsymbol{x})$ and the actual number of events $B_{(k_1,k_2)}(\boldsymbol{x})$ in $\{B_k\}_{k=1}^{b}$ is taken as our event rendering loss. For pixel $\boldsymbol{x}$ corresponding to $B_k(\boldsymbol{x})$ in the event block, set the number of negative events to its additive inverse so that positive and negative events can cancel each other out when stacking event blocks. Thus the expression for the event rendering loss is:

$$\begin{gathered} \mathcal{L}_{event} = \textstyle\sum_{\boldsymbol{x}\in\chi}\left\|\widehat{B}_{(k_1,k_2)}(\boldsymbol{x}) - B_{(k_1,k_2)}(\boldsymbol{x})\right\|_2^2, \\ B_{(k_1,k_2)}(\boldsymbol{x}) = \textstyle\sum_{k=k_1+1}^{k_2} B_k(\boldsymbol{x}), \end{gathered} \tag{12}$$

where χ is the set of pixels in each batch. The final loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{blur} + \omega\mathcal{L}_{event}, \tag{13}$$

where ω is the weight parameter.

V. EXPERIMENT

A. Dataset

Since this experiment improves on the method proposed by E2NeRF, in order to compare with the original model, the dataset constructed by E2NeRF is used in this paper. Based on the original NeRF dataset, camera shakiness was simulated using the camera shakify plugin in Blender. In each view, 17 clear images of the camera under different degrees of shaking

were rendered and their corresponding poses were recorded. These 17 images were then input into the simulated event generation model v2e [17] to simulate the event data generated during camera shaking. In addition, in order to obtain the simulated blurred images, the 17 images were processed to the original domain by inverse ISP operation and superimposed, and then ISP processing was utilized to obtain the final blurred images. Each scene contains 100 blurred images and corresponding event data.

The real-world dataset was captured using a DAVIS 346 color event camera. This camera is capable of capturing spatio-temporally aligned event data and RGB frames. Where the resolution of the camera is 346×260 and the exposure events of the RGB frames are set to 100 ms. 5 sets of challenging scenes containing rich color and texture details in low-light environments were acquired with this camera acquisition platform. Each scene contains 30 images with different levels of blurring on different views and the corresponding event data.

B. Experiment Details

The model in this thesis is constructed using 3DGS as pipeline and trained on a single NVIDIA RTX2080Ti GPU for 30,000 iters for each scene. For each hyper-parameter, set $\omega = 1/625$, $b = 4$ the rest of the parameters are kept the same as the default values of 3DGS, and set the positive threshold $\theta_{pos} = 0.2$ and negative threshold $\theta_{neg} = 0.2$ to simulate the generation of events.

For the Blender synthesized data, under each view, 5 poses are selected at equal intervals and sequentially input to the network for rendering. In order to reflect the generalization of the model, experiments were conducted under several synthetic scene data, including challenging scenes such as metallic materials.

For real-world data, since there are only blurred images and corresponding events, one pose estimation model is used to obtain the corresponding five poses. To characterize the high dynamic range of the event data, several scenes were captured under dark light conditions, each containing a total of 30 views.

C. Comparison Study

In order to measure the quality of the Ev3DGS model, we will compare the measurements from the synthetic dataset and under the real-world dataset. In addition this model will be compared with the E2NeRF model which is also based on blurring rendering loss and event rendering loss. The specific experiments are as follows:

Synthetic data. The experiments were performed under 6 synthetic scenes for blur rendering, and PSNR, SSIM and LPIPS were used to evaluate the experimental results, and the Ev3DGS rendering results are shown in Tab. I. The experimental results show that Ev3DGS achieves a good deblurring rendering level in all six scenes, with an average PSNR of 29.96. Especially in the hotdog scene, the PNSR reaches a level of 34.1, which is of very high rendering quality. In addition, in the challenging scenes such as materials and mic, the rendering quality is also good in the face of metallic materials and fine

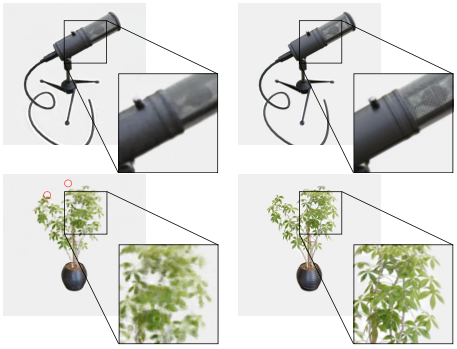


Fig. 2. Comparison of E2NeRF (left) and Our Ev3DGS (right) rendering results on synthetic data of mic (top) and ficus (bottom) scene.

meshes. However, by observing the rendering results of the synthesized scenes in the first line of Fig. 2, we can find that in some scenes, there are some "wrinkles" and "artifacts" on the edges of the objects, which may be due to the sensitivity of the event data to the edges of the moving objects.

TABLE I
SYNTHETIC DATA RENDERING RESULTS OF EV3DGS METHOD

Ev3DGS	Chair	Ficus	Hotdog	Lego	Materials	Mic	Average
PSNR↑	31.32	30.82	34.10	28.63	27.88	26.96	29.95
SSIM↑	0.9646	0.9654	0.9652	0.9238	0.9378	0.9488	0.9509
LPIPS↓	0.0503	0.0361	0.0610	0.1047	0.0708	0.0545	0.0629

The average training time is about 35min.

TABLE II
SYNTHETIC DATA RENDERING RESULTS OF E2NERF METHOD

E2NeRF	Chair	Ficus	Hotdog	Lego	Materials	Mic	Average
PSNR↑	30.36	25.81	33.43	27.76	27.42	26.61	28.57
SSIM↑	0.9525	0.9035	0.9541	0.8972	0.9255	0.9359	0.9281
LPIPS↓	0.0682	0.0890	0.0562	0.0889	0.0571	0.0831	0.0738

The average training time is about 13h.

The E2NeRF rendering results are shown in Tab. II. The experimental results show that in the same six scenes, the overall rendering level of the two is relatively close. However, the E2NeRF model rendering results show more obvious "patches" and white voids (circle in red), such as those shown in the second line of Fig. 2. In ficus and mic scenes, E2NeRF has a low level of differentiation for dense and small objects, while Ev3DGS has a clearer performance. Under the materials scene, Ev3DGS better reflects the reflection characteristics of metal materials, while E2NeRF's reflection characteristics are less obvious.

Real-world data. Since the real-world data lacks clear images of ground truth, the five real scenes were quantitatively analyzed using the reference-free image quality assessment metrics BRISQUE [18] and RankIQA [19], as shown in Tab. III, and the Ev3DGS results are also better. Under the blur rendering loss and event rendering loss, the event data effectively enhances the perception of 3DGS on dynamic blur. The experimental results show that Ev3DGS also exhibits a better rendering level in low dark real environments thanks to the high dynamic range characteristics of the event camera.

TABLE III
COMPARATIVE RESULTS ON REAL-WORLD DATA

Real-World Data	Ev3DGS	E2NeRF
BRISQUE↓	29.32	29.53
RankIQA↓	3.579	3.588

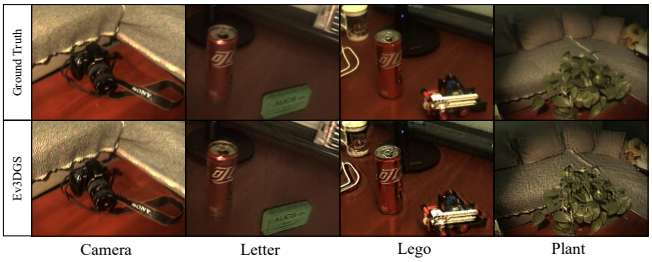


Fig. 3. Real-world data rendering results of Ev3DGS method.

D. Ablation Study

In this experiment, blurry rendering loss and event rendering loss are introduced to the model, and in order to verify the effect of introducing event rendering loss on the model, ablation experiments are carried out based on the synthetic dataset. Where the Ev3DGS model with event loss EventLoss removed is denoted by Ev3DGS*. The experimental results are shown in Tab. IV.

From the results, as shown in Tab. V, the rendered image of Ev3DGS model compared with Ev3DGS* model improves PSNR and SSIM but sacrifices a certain amount of LPIPS.Specifically for the rendering of each scene, it can be found that Event Loss has a large improvement in all the indicators for ficus and hotdog scenes; while for chair, mic scenes For the chair and mic scenes, there is no significant change; for the lego and materials scenes, there is a large improvement in PSNR and SSIM at the expense of certain LIPIS indicators.

TABLE IV
RESULTS OF THE SYNTHETIC DATA OF EV3DGS* METHOD

Ev3DGS*	Chair	Ficus	Hotdog	Lego	Materials	Mic	Average
PSNR↑	31.23	30.70	33.92	28.74	27.57	26.95	29.85
SSIM↑	0.9646	0.9651	0.9648	0.9266	0.9351	0.9487	0.9508
LPIPS↓	0.0496	0.0363	0.0614	0.0998	0.0702	0.0540	0.0618

TABLE V
COMPARATIVE RESULTS OF ABLATION STUDY

Ablation Study	Ev3DGS*	Ev3DGS	Δ
PSNR↑	29.85	29.95	0.34%↑
SSIM↑	0.9508	0.9509	0.01%↑
LPIPS↓	0.0618	0.0629	1.78%↑

VI. CONCLUSIONS

In this paper, a new 3DGS model based on event camera enhancement (Ev3DGS) is proposed, which explores the possibility of combining event camera with 3DGS and provides a solution for 3DGS with blurry input. In addition, the effectiveness of the model is demonstrated experimentally on both synthetic and real datasets. The results show that the Ev3DGS framework has a slight advantage in terms of rendering results over the existing deblurred 3D reconstruction model, the E2NeRF model. However, in terms of training and rendering speed, Ev3DGS is much faster than E2NeRF, which is also due to the ability of fast rendering of 3DGS.In conclusion, the relevant results and conclusions obtained in this paper are useful. It is hoped that the related work in this paper will contribute to the study of high-quality 3D representation learning using event data in complex and low-light scenes.

REFERENCES

[1] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *TOG*, 42(4):1–14, 2023.
[2] Zhexiong Wan, Yuchao Dai, and Yuxin Mao. Learning Dense and Continuous Optical Flow from an Event Camera. *TIP*, 31: 7237–7251, 2022.
[3] Zhexiong Wan, Yuxin Mao, Jing Zhang, and Yuchao Dai. Rpeflow: Multimodal Fusion of RGB-PointCloud-Event for Joint Optical Flow and Scene Flow Estimation. In *ICCV*, pages 10030–10040, 2023.
[4] Juanjuan Zhu, Zhexiong Wan, and Yuchao Dai. Video Frame Prediction from a Single Image and Events. In *AAAI*, volume 38, pages 7748–7756, 2024.
[5] Yunshan Qi, Lin Zhu, Yu Zhang, and Jia Li. E2NeRF: Event Enhanced Neural Radiance Fields from Blurry Images. In *ICCV*, pages 13254–13264, 2023.
[6] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. GS-SLAM: Dense Visual SLAM with 3D Gaussian Splatting. In *CVPR*, pages 19595–19604, 2024.
[7] Zhicheng Lu, Xiang Guo, Le Hui, Tianrui Chen, Min Yang, Xiao Tang, Feng Zhu, and Yuchao Dai. 3D Geometry-aware Deformable Gaussian Splatting for Dynamic View Synthesis. In *CVPR*, pages 8900–8910, 2024.
[8] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. SC-GS: Sparse-Controlled Gaussian Splatting for Editable Dynamic Scenes. In *CVPR*, pages 4220–4230, 2024.
[9] Jeongtaek Oh, Jaeyoung Chung, Dongwoo Lee, and Kyoung Mu Lee. DeblurGS: Gaussian Splatting for Camera Motion Blur. *arXiv preprint arXiv:2404.11358*, 2024.
[10] Lingzhe Zhao, Peng Wang, and Peidong Liu. BAD-Gaussians: Bundle Adjusted Deblur Gaussian Splatting. *arXiv preprint arXiv:2403.11831*, 2024.
[11] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-Based Vision: A Survey. *TPAMI*, 44(1):154–180, 2020.
[12] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras. In *RSS*, 2018.
[13] Chengxi Ye, Anton Mitrokhin, Cornelia Fermüller, James A. Yorke, and Yiannis Aloimonos. Unsupervised Learning of Dense Optical Flow, Depth and Egomotion with Event-Based Sensors. In *IROS*, 2020.
[14] Christian Brandli, Thomas A Mantel, Marco Hutter, Markus A Höpflinger, Raphael Berner, Roland Siegwart, and Tobi Delbruck. Adaptive Pulsed Laser Line Extraction for Terrain Reconstruction Using a Dynamic Vision Sensor. *Frontiers in neuroscience*, 7:275, 2014.
[15] Nathan Matsuda, Oliver Cossairt, and Mohit Gupta. MC3D: Motion Contrast 3D Scanning. In *ICCP*, pages 1–10, 2015.
[16] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High Frame Rate Video Reconstruction Based on an Event Camera. *TPAMI*, 44(5):2519–2533, 2020.
[17] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From Video Frames to Realistic DVS Events. In *CVPR*, pages 1312–1321, 2021.
[18] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-Reference Image Quality Assessment in the Spatial Domain. *TIP*, 21(12):4695–4708, 2012.
[19] Xialei Liu, Joost Van De Weijer, and Andrew D. Bagdanov. RankIQA: Learning from Rankings for No-Reference Image Quality Assessment. In *ICCV*, pages 1040–1049, 2017.

Automated prediction of loudness growth curve using EEG signals

Nitya Tiwari*, Arjun Reddy Vadyala *, and K. S. Nataraj†
* School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, Bhubaneswar, India
E-mail: nityatiwari@iitbbs.ac.in, 21ee01011@iitbbs.ac.in
† Department of Electronics and Communication Engineering, Indian Institute of Information Technology Dharwad, Dharwad, India
E-mail: nataraj@iiitdwd.ac.in

***Abstract*—This paper introduces an innovative and automated approach for the prediction of loudness growth curves based on auditory brainstem responses (ABRs), harnessing the power of deep learning and signal processing techniques. Hearing loss, affecting a significant portion of the global population, calls for accurate and efficient assessment methods to improve the quality of life for affected individuals. Our method entails preprocessing ABR signals, extracting informative features via empirical wavelet transform with Fourier Bessel series expansion, and subsequently mapping these features to loudness growth estimates using multi-target regression. Through evaluation employing mean squared error and Frechet distance, our approach demonstrates acceptable performance and consistency across subjects and stimulus levels. Importantly, it overcomes limitations inherent in existing methods that primarily rely on click ABRs and psychoacoustic measures.**

I. Introduction

According to findings in [1] based on the Global Burden of Disease Study 2019, every one person in five has hearing loss. It can be congenital or may occur due to aging or exposure to loud noise for prolonged periods. Hearing loss if left undiagnosed and unattended adversely affects the quality of life. It may hinder speech development in children and can cause loneliness, isolation, depression, dementia, and cognitive decline in the elderly [1]. Therefore, there is a need for timely hearing screening tests followed by a prescription of appropriate assistive hearing devices such as cochlear implants or hearing aids. As the hearing loss varies in nature and intensity, comprehensive hearing screening tests are needed to assess the loss type and severity. The commonly used procedures for assessment of auditory function include air conduction and bone conduction pure-tone thresholds, speech recognition thresholds, suprathreshold speech recognition, etc. [2]. However, speech-based audiometric procedures have been standardized for English-speaking adults and therefore not suitable for hearing assessment of non-native English speakers. Further, these tests require active participation and feedback from the listeners during assessment which may be difficult for children and infants.

Other methods of auditory function assessment include otoacoustic emissions (OAEs) and auditory evoked potentials (AEPs). They are useful for the assessment of infants and young children and may be performed in addition to routine audiometric measures for adults. AEP-based hearing threshold estimation makes the estimation language independent and is suitable for infants, children, and persons who lack the ability to provide a reliable response to the stimuli. Perceived loudness is a subjective quantity and is correlated to the physical sound intensity. The human perception of loudness is a function of stimulus intensity, stimulus frequency, and depends on individual listener. It varies in complex ways that cannot be accounted for by the listener's threshold level. For example, in listeners with sensorineural hearing loss the hearing thresholds are elevated, and the perceived loudness increases rapidly as compared to normal hearing listeners. Therefore, there is a need for estimating the loudness growth along with hearing thresholds and discomfort levels from physiological signals like AEP or auditory brainstem response (ABR) for normal as well as hearing-impaired listeners.

Numerous mathematical models have been developed to estimate loudness growth as a function of stimulus intensity, with early methods focusing on non-linear models for loudness prediction. An initial model proposed a power law relationship between loudness L (in Sones) and sound pressure level P (in Pascals) as $L = kP^{0.6}$, which is accurate for higher sound pressure levels relative to hearing threshold levels, was proposed in [3]. Hellman and Zwislocki [4] refined this model to $L = 0.01(P - P_0)^{0.6}$, where P_0 is a constant. Buus and Florentine [5] further advanced the model by incorporating the effect of Signal-to-Noise Ratio (SNR) using $L = k10^{s_{HL}\frac{HL}{10}}\left[\left(1+\left(snr_{th}.10^{\frac{SL}{10}}\right)^{s_{lo}}\right)^{\frac{s_{hi}}{s_{lo}}}-1\right]$, where snr_{th}, s_{lo}, and s_{hi} are free parameters and k is a scale factor that does not affect the predictions of the model. In this expression, s_{HL} determines the extent to which the loudness at threshold varies with the amount of hearing loss, snr_{th} may be considered the signal-to-noise ratio at threshold (reported as $\text{SNR}_{th} = 10\log(snr_{th})$ dB), s_{lo} is the asymptotic exponent of the loudness function at low levels, and s_{hi} is the asymptotic exponent at high levels. Both the functions proposed are suitable for use at higher stimulus. Further the exponent factor is very different for different loudness levels. Particularly, the loudness growth exponent was larger than 2 for stimulus levels closer to the threshold and was smaller than 0.6 for moderate and higher stimulus levels.

The procedures to measure loudness typically involve psychoacoustical methods such as Cross Modality Matching (CMM) and Magnitude Estimation (ME), which rely on patient responses and can yield inconsistent judgments [6]. Consequently, researchers have explored objective loudness estimation using auditory brainstem responses (ABRs) [6], [7]. Studies have correlated ABR features, especially Wave V latency and amplitude, with psychoacoustical measures of loudness growth. Pratt and Sohmer (1977) [8] utilized AME and assessed ABR features including latency, amplitude, and area of waves (I-V), finding correlations via power function fitting but noting variability across subjects and sessions. Davidson et al. ([9]) also explored Wave V correlations with AME, using Spearman's rank correlation coefficient across stimulus levels, revealing correlations mainly across sessions. In contrast, [10] identified Wave V latency correlations with loudness in normal listeners but not in those with sloping hearing loss. Objective evaluation of loudness discomfort levels has been attempted by correlating Wave V features with subjective assessments [11], [12].

Epstein Silva ([6], [7]) employed tone burst ABRs with AME and CMM measures for objective loudness growth estimation, employing MSE as an evaluation metric rather than Pearson correlation coefficient used in previous studies. Hoseingholizade [13] investigated chirp-evoked ABRs, averaging trials using Bayesian-informed weighted averaging and fitting linear or power functions to estimate loudness growth based on peak to trough amplitudes at different intensity levels.

Significant gaps persist in current research on loudness curve estimation using auditory brainstem responses (ABRs). Earlier methods have primarily relied on click ABRs, potentially limiting robust analysis of loudness curves. Moreover, existing methodologies heavily favor psychoacoustical measures like Cross Modality Matching (CMM) and Magnitude Estimation (ME), often overlooking other valuable ABR data aspects. This bias restricts the full potential of ABR signals for comprehensive feature extraction, typically focusing subjectively on parameters such as amplitude and latency. Furthermore, the field has predominantly used polynomial and exponential functions for loudness growth estimation, neglecting exploration into advanced methods like deep learning and signal processing for ABR data. Despite an extensive literature search, recent studies on automated prediction of loudness growth curves using EEG signals were notably absent, underscoring a critical research gap. Our study addresses these limitations by proposing a novel neural network-based approach, highlighting the need for further research to enhance ABR-based loudness curve estimation capabilities.

This paper introduces a novel approach employing deep learning for automated estimation of the loudness growth curves from the raw ABR signals. Moreover, we explore feature extraction technique, drawing from both signal processing and deep learning methodologies, to predict loudness growth curves. The proposed methodology encompasses several key steps: i) Pre-processing of ABR signals and loudness growth data, ii) Extracting features from ABR data through a signal processing approach called empirical wavelet transform with band boundaries determined by Fourier Bessel series expansion (EWT-FBSE), and iii) Mapping the extracted EWT-FBSE features to corresponding psychoacoustical estimates of loudness growth utilizing Multi-target Regression. We evaluating the efficacy of the proposed approach through objective measures such as mean squared error and Frechet distance and use the Physiobank database [14] for training and testing the proposed model.

II. Methodology

The steps involved in estimation of the relationship between the raw ABR signals and loudness growth curve are explained in the following subsections.

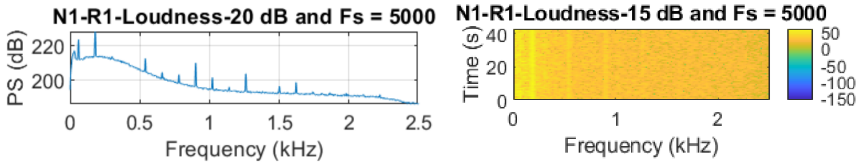
A. Pre-processing of ABR signals and loudness growth data

The processing of the raw files present in the Physiobank "Evoked Auditory Responses in Normals across Stimulus Level" database [14]. It involves extraction of the ABR signal and its metadata using functions from the WaveForm DataBase (WFDB) Toolbox from MATLAB. The database consists of two sets of signals - ABR and OAE and two sets of psychoacoustical estimates of loudness as a function of peak sound pressure level (peSPL)- Cross Modality Matching (CMM) and Magnitude Estimation (ME). There are 2002 samples per trial for 8 normal hearing listeners (four males, four females) with ages 19-31 with audiometric thresholds not exceeding 15 dB HL at octave frequencies 250 Hz - 2 kHz.

The raw ABR signal is averaged over the first 1000 trials for each loudness level to reduce the effect of noise in analysis. The stimuli are originally sampled at 48 kHz and they resampled to 2 kHz, 3 kHz, 5 kHz and 10 kHz. The power spectrum and spectrogram as shown in Figure 1 are used to examine the signal's frequency content. It can be seen that the frequency content is in the range of 0 – 2.5 kHz and in the power spectrum plots, peaks are primarily observed in the 0 - 1.5 kHz range. Thus, for further analysis, the sampling frequency is taken as 3 kHz.

B. Feature extraction

For feature extraction, we begin by decomposing each raw ABR signal using wavelet transform. Wavelet transform, known for its multiresolution analysis properties, is chosen as a suitable method for signal analysis. Empirical wavelet transform (EWT) is specifically employed to decompose the pre-processed signals into a set of intrinsic modes, effectively separating dominant modes from irrelevant ones. This approach has proven to be highly valuable, particularly when analyzing non-stationary and noisy signals. However, it's worth noting that EWT may not achieve an accurate representation of frequency compo nents in signals with shorter durations. Additionally, when frequency components are closely spaced, EWT may not be as effective. To address these limitations, we employ Fourier Bessel series expansion (FBSE) to compute band boundaries within the EWT framework. This combination



(a) a) Power spectrum (b) b) Spectrogram

Fig. 1: a) Power spectrum and b) spectrogram for averaged ABR signal sampled at 5 kHz.

of EWT and FBSE allows us to overcome these challenges, ensuring a more robust and accurate decomposition of the ABR signals for subsequent analysis. The computation of EWT-FBSE is done using the following procedure in [15]:

$$S(i) = \frac{2}{N^2 J_1(\tau_i)]^2}[\sum_{n=0}^{N-1} ns(n) J_0\left(\frac{\tau_i n}{N}\right) i = 1, 2, 3....N \quad (1)$$

where $J_0(.)$ and $J_1(.)$ are 0th and 1st order Bessel functions, respectively. The N positive roots of $J_0(\tau_k) = 0$ are τ_i for $i = 1, 2, \ldots, N$. The frequency of signal f_i (Hz) and τ_i are related mathematically as $\tau_i \approx \frac{2\pi f_i N}{F_s}$, where F_s represents a sampling rate of the signal. For designing the filter banks based on the FBSE-EWT method, the obtained FBSE spectrum was segmented based on i-th order range boundaries using the scale-space boundary detection method [16]. First (ω_1) and last (ω_N) boundary frequency values are 0 and π, respectively. The optimal set of boundary frequencies for adaptive segments are denoted as ω_i. The empirical scaling and wavelet functions are applied to design the band-pass filters on each adaptive segment of the FBSE spectrum. Wavelet-based band-pass filters were constructed based on [17].

C. Mapping the features to psychoacoustical estimate of loudness growth

The next step is mapping the extracted features from ABR signals to the corresponding psychoacoustical estimate of loudness growth. CMM is taken as the measure of the loudness growth (ranging from -1 to 2). This step involves mapping the features from ABR signals for all stimulus levels to the corresponding loudness growth curve for a given individual. This is achieved using multi-target regression (MTR) based convolutional neural network (CNN) that helps in learning the correlation between the features from ABR signals for different stimulus levels and the loudness growth curve. For prediction of the entire loudness growth curve in one go, MTR is beneficial as it facilitates simultaneous prediction of multiple continuous variables from a set of input variables. The output of MTR is given to a dense layer to obtain the final loudness growth curve for a listener. A deep hard sharing-based network is used for MTR which is connected to a fully dense layer with a linear activation in the third and final stage to predict the loudness growth curve with the number of outputs equal to the number of loudness levels of stimuli present in the input EWT data. The proposed architecture is shown in Figure 2.

The Multi-Target Regression Block in our methodology employs a CNN that captures both spatial and temporal characteristics within datasets. In our specific context, the extracted features are essentially two-dimensional, as they pertain to different stimulus levels and for each level it has 10-band EWT-FBSE coefficients. The CNN architecture consists of a convolutional layer, a max-pooling layer, and a dropout layer, which collectively capture the spatial dependencies within the data. The convolutional layer utilizes 32 filters of size 3x3, effectively discerning spatial patterns with padding set to "same" and a stride of 1. The dropout layer randomly omits a fraction of connections from the preceding layer, promoting model generalization. Batch Normalization is applied after each convolutional block to normalize the extracted features.

In a general MTR framework, inputs are represented as continuous variables. For an input vector X of dimension d, the output space has a dimension q. Each multi-target example, denoted as X_i and Y_i, defines the input-output relationship between vectors of dimensions d and q, respectively. In this supervised deep learning approach, the deep forward network incorporates shared layers where parameters are common among different input variables. The advantage of using a shared DNN architecture lies in its ability to reduce the number of parameters compared to separate networks for each band, thus enhancing computational efficiency and potentially improving generalization by leveraging shared knowledge across different inputs. The shared architecture is followed by non-shared layers, leading to the ultimate target output vector. In our implementation, we construct a shared deep neural network wherein the band outputs serve as features inputted into a shared CNN. The CNN's output is then connected to a fully connected dense layer, facilitating the prediction of the final set of loudness growth estimates using CMM. In essence, this process yields the loudness growth curve tailored to the individual under consideration.

III. Architecture Details of MTR CNN Model

In the implementation of the proposed architecture, a shared deep neural network is developed where the band outputs are first fed into the neural net as features in a shared CNN. The output of the shared CNN is then connected to a fully connected dense layer to predict the final set of loudness growth CMM estimates, i.e., the loudness growth curve for the given individual. The architecture details are given in Table I.

IV. Experiments and Evaluation

A. Model hyperparameters, loss function, and platform

For training the proposed model, we partitioned the dataset into training and testing subsets, following an 85-15 ratio. For the training process, we employed the Adam optimizer. The

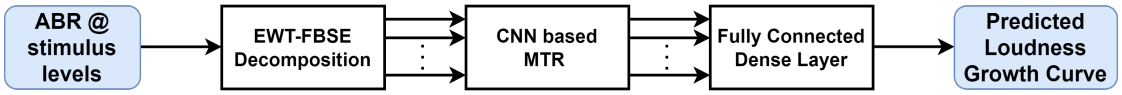


Fig. 2: Prediction of loudness growth using proposed approach.

TABLE I: Architecture Details of MTR CNN Model

Layer Type	Output Shape	Number of Parameters
Input	12, 10, 126	0
Conv 2D-1	12, 10, 32	36,320
Conv 2D-2	4, 3, 32	9,248
Conv 2D-3	2, 2, 32	9,248
Final Dense	12	1,548
Total Parameters		61,484

choice of Mean Squared Error (MSE) as the loss function is motivated by its superior performance in capturing the correlation between the output and non-linear features, as opposed to other loss functions such as Mean Absolute Error (MAE) and Pearson Correlation. To ensure that the validation loss remained under control during training, we implemented an early stopping mechanism with a patience parameter set to 40. Our model underwent training for a maximum of 500 epochs, utilizing two Nvidia RTX A4000 GPUs, each equipped with 16 GB of memory. The learning rate was set at 0.0001, and we used a batch size of 12 to optimize the training process efficiently.

B. Evaluation metrics

For evaluating the performance of our model, we employ two key metrics. The first metric is the Average Mean Squared Error, which quantifies the disparity between the CMM loudness growth estimate and the predicted loudness growth estimate. This metric serves as a crucial performance measure, gauging the accuracy of our predictions. Additionally, we utilize the Frechet distance [18] to assess the similarity between two curves. In our specific context, the Frechet distance is employed to measure the likeness between the predicted loudness growth curve and the ground truth loudness growth curve. This metric is defined as the smallest value among the maximum pairwise distances between the two curves. Notably, we utilize the discrete Frechet distance variant due to the discrete nature of our data points.

V. Results

In this study, we aimed to identify the optimal model for predicting a generalized loudness growth curve. To achieve this, we modified our approach by leveraging a dataset encompassing seven subjects. The dataset includes detailed measurements of auditory responses, allowing us to train and evaluate our models effectively.

To ensure a robust evaluation, we adopted a cross-validation strategy. In each iteration of this strategy, we set aside one subject's data as the test set while using the remaining subjects' data for training the model. This leave-one-out approach allowed us to systematically predict the loudness growth curve for each subject and subsequently compare the predictions to the ground truth curves. This method not only ensures that each subject is tested independently but also helps in understanding how well the model generalizes to unseen data.

To provide a comprehensive view of the model performance, we present a comparative analysis in Table II. The table includes the average Mean Squared Error (MSE) and standard deviation for models developed using K-Fold resampling to predict the loudness growth curve across all subjects. In this context, K-Fold resampling ensures that the training and testing processes are repeated multiple times, enhancing the reliability of the performance metrics. The columns in the table contain the average MSE (standard deviation) scores across stimulus intensity levels ranging from 45 dB to 100 dB for different subjects within a given fold. Each model, denoted as M_x, is the model that has not seen subject x during training. The rows correspond to the scores obtained when the model is evaluated using a particular subject, providing a detailed performance metric for each individual.

Key observations from Table II reveal insightful patterns in the performance of different models across subjects. Each model, denoted as M_x where x ranges from 1 to 7, was trained by leaving out data from one specific subject during each iteration of cross-validation. The table presents average Mean Squared Error (MSE) scores along with their standard deviations across stimulus intensity levels from 45 dB to 100 dB for each subject-model pair. The models exhibit relatively consistent performance with minor variations. Model M_6 consistently shows slightly better performance compared to others, indicated by lower average MSE scores across subjects. For instance, it achieves an average MSE of 0.01 (with a standard deviation of 0.09) for Subject 1 and 0.37 (with a standard deviation of 0.18) for Subject 7. This suggests that M_6 effectively generalizes well to unseen data, possibly due to the characteristics of the training data it was exposed to. Model M_1, while not consistently the best performer across all subjects, shows notable strength in predicting loudness growth curves for Subject 1, with an MSE of 0.06 (standard deviation 0.19). This performance makes it a promising candidate for further evaluation, especially considering its effectiveness in predicting unseen data for Subject 1.

Overall, the table underscores the robustness of the models in predicting loudness growth curves across different subjects, highlighting the importance of cross-validation in evaluating model performance. The standard deviations provide insights into the variability of predictions, further aiding in understanding the reliability of each model's performance metrics. These findings support the selection of Model M_1 for subsequent analyses and comparisons against existing methodologies, ensuring rigorous evaluation and validation of our approach.

Table III compares the Mean Squared Error (MSE) scores

TABLE II: MSE scores averaged across all loudness levels and test cases (Standard deviations) for 7 models and for 7 subjects

Subject	M_1	M_2	M_3	M_4	M_5	M_6	M_7
1	0.06 (0.19)	0.02 (0.13)	0.03 (0.15)	0.02 (0.15)	0.02 (0.14)	0.01 (0.09)	0.01 (0.11)
2	0.28 (0.05)	0.41 (0.07)	0.37 (0.05)	0.26 (0.05)	0.23 (0.05)	0.27 (0.11)	0.16 (0.09)
3	0.15 (0.14)	0.24 (0.10)	0.22 (0.14)	0.14 (0.14)	0.12 (0.13)	0.12 (0.08)	0.06 (0.09)
4	0.06 (0.24)	0.04 (0.20)	0.05 (0.30)	0.10 (0.30)	0.07 (0.24)	0.04 (0.18)	0.08 (0.21)
5	0.09 (0.24)	0.05 (0.19)	0.06 (0.22)	0.09 (0.25)	0.27 (0.18)	0.08 (0.23)	0.12 (0.27)
6	0.06 (0.24)	0.07 (0.27)	0.06 (0.24)	0.06 (0.23)	0.07 (0.24)	0.11 (0.32)	0.13 (0.29)
7	0.33 (0.12)	0.22 (0.14)	0.25 (0.12)	0.34 (0.11)	0.39 (0.12)	0.37 (0.18)	0.51 (0.17)

TABLE III: Comparison of MSE scores averaged across all loudness levels and test cases for Model M_1 and method in [6], [7]

Models	**1**	**2**	**3**	**4**	**5**	**6**	**7**
M_1	0.06	0.28	0.15	0.06	0.09	0.06	0.33
Silva and Epstein [6], [7]	0.19	0.03	0.08	0.08	0.15	0.29	0.11

TABLE IV: Frechet Distance Obtained for All Subjects for Model M_1

Models	**1**	**2**	**3**	**4**	**5**	**6**	**7**
M_1	0.37	0.75	0.62	0.61	0.83	0.53	0.98

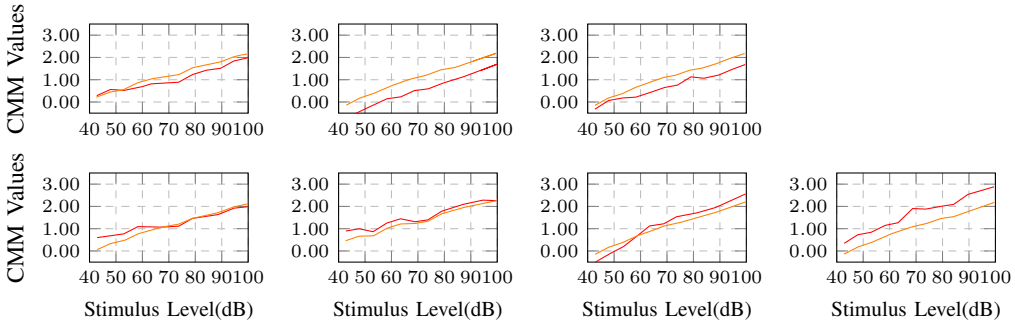


Fig. 3: Comparison of predicted (orange) and ground truth (red) loudness growth curve for seven subjects using model M_1.

averaged across all loudness levels and test cases for Model M_1 with those reported by Silva and Epstein [6], [7]. The table illustrates how Model M_1, selected based on its performance in predicting loudness growth curves for subjects at a stimulus frequency of 1 kHz, compares against established methodologies. Each entry in the table represents the MSE score for a specific subject, ranging from 1 to 7, indicating the predictive accuracy of Model M_1 relative to the referenced models.

Key observations reveal that Model M_1 consistently achieves competitive MSE scores across different subjects compared to Silva and Epstein's approach. Notably, Model M_1 exhibits strengths in predicting loudness growth curves for certain subjects, such as Subject 1 with an MSE of 0.06, demonstrating its efficacy in capturing variations in loudness perception across different stimuli intensities. These findings underscore the robustness of Model M_1 in predicting loudness growth curves and highlight its potential for advancing the understanding and application of predictive models in auditory perception research.

Table IV provides the Frechet distance scores calculated for Model M_1 across all seven subjects. The Frechet distance is a metric used to measure the similarity between two curves, typically comparing the predicted and ground truth loudness growth curves. A lower Frechet distance indicates a closer match between these curves. The scores in Table IV range from 0.37 to 0.98 across subjects 1 to 7, illustrating the varying degrees of accuracy in predicting loudness growth curves using Model M_1. Lower Frechet distances, such as 0.37 for Subject 1, indicate a closer match between the predicted and ground truth curves, suggesting that Model M_1 performs well in predicting loudness growth for some subjects but shows greater deviation for others.

To facilitate a thorough evaluation, we have generated visual representations of the predicted and ground truth loudness growth curves for all subjects using M1, presented in Figure 3. This figure depicts the relationship between CMM-value and intensity level (dB). In Subject 1, the predicted loudness growth curve closely mirrors the ground truth curve, indicative of model M_1's ability to accurately reproduce observed trends. This alignment is substantiated by a low Frechet distance of 0.37 (Table IV), underscoring the model's fidelity in predicting loudness responses in this specific case. Conversely, in Subject 7, discrepancies between the predicted and ground truth curves are more apparent, reflected in a higher Frechet distance of 0.98 (Table IV). Overall, the visual analyses presented in Figure 3 complement the quantitative metrics detailed in Tables II and IV, offering a comprehensive evaluation of model M_1's performance across a diverse cohort.

VI. Conclusion

Hearing loss represents a significant public health issue with profound implications for quality of life. In this study, we have introduced a novel approach employing deep learning to automate the estimation of loudness growth curves from raw Auditory Brainstem Response (ABR) signals, offering a valuable tool for hearing assessment. Our methodology encompasses several crucial steps: preprocessing of ABR signals and loudness growth data, feature extraction utilizing empirical wavelet transform with Fourier Bessel series expansion, and mapping these features to psychoacoustical estimates of loudness growth through multi-target regression.

The evaluation of our approach incorporates objective metrics such as Mean Squared Error (MSE) and Frechet distance, providing a rigorous quantitative assessment of our model's performance. Our experimental findings demonstrate the effectiveness of our method, showcasing competitive performance across diverse subjects and stimulus intensity levels. Notably, Model M_1 emerged as the top performer, particularly adept at predicting unseen loudness growth curves.

Comparative analysis with a prior study by Silva and Epstein underscores the favorable outcomes of our approach in terms of MSE and Frechet distance metrics. This highlights the robustness and reliability of our model in estimating loudness growth curves from ABR signals.

In summary, our study represents a significant advancement in hearing assessment techniques by introducing a robust and automated method for loudness growth curve estimation. This approach has the potential to deepen our understanding of auditory perception and facilitate the development of personalized solutions for individuals with varying degrees of hearing impairment. Future research directions should focus on validating and refining our model in populations with hearing impairments, thereby broadening its applicability and impact in clinical settings.

References

[1] GBD 2019 Hearing Loss Collaborators, "Hearing loss prevalence and years lived with disability, 1990–2019: Findings from the global burden of disease study 2019," *The Lancet*, vol. 397, no. 10278, pp. 996–1009, 2021.

[2] R. A. Dobie and H. S. Van, *Hearing Loss: Determining Eligibility for Social Security Benefits*. National Academies Press (US), 2004.

[3] L. Marks and M. Florentine, "Measurement of loudness, part i: Methods, problems, and pitfalls," in *Springer Handbook of Auditory Research*, vol. 37, 2010, pp. 17–56.

[4] R. P. Hellman and J. J. Zwislocki, "Loudness determination at low sound frequencies," *Journal of the Acoustical Society of America*, vol. 43, no. 1, pp. 60–64, 1968.

[5] S. Buus and M. Florentine, "Growth of loudness in listeners with cochlear hearing losses: Recruitment reconsidered," *Journal of the Association for Research in Otolaryngology*, vol. 3, no. 2, pp. 120–139, 2002.

[6] I. Silva and M. Epstein, "Estimating loudness growth from toneburst evoked responses," *Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3629–3642, 2010.

[7] I. Silva and M. Epstein, "Objective estimation of loudness growth in hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 353–362, 2012.

[8] H. Pratt and H. Sohmer, "Correlations between psychophysical magnitude estimates and simultaneously obtained auditory nerve, brain stem and cortical responses to click stimuli in man," *Electroencephalography and Clinical Neurophysiology*, vol. 43, no. 6, pp. 802–812, 1977.

[9] S. A. Davidson, L. G. Wall, and C. M. Goodman, "Preliminary studies on the use of an abr amplitude projection procedure for hearing aid selection," *Ear and Hearing*, vol. 11, no. 5, pp. 332–339, 1990.

[10] Y. C. Serpanos, H. O'Malley, and J. S. Gravel, "The relationship between loudness intensity functions and the click-abr wave v latency," *Ear Hear*, vol. 18, no. 5, pp. 409–419, 1997.

[11] J. Agnew and J. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[12] A. R. D. Thornton, L. Yardley, and G. Farrell, "The objective estimation of loudness discomfort level using auditory brainstem evoked responses," *Scandinavian Audiology*, vol. 16, no. 4, pp. 219–225, 2009.

[13] S. Hoseingholizade, "Investigation of the feasibility of using chirp evoked abr in estimation of loudness growth," M.S. thesis, McMaster University, 2015.

[14] *Evoked auditory responses in hearing impaired and normal*, MultiParameter Databases, Waveform Databases, PhysioBank Databases. [Online]. Available: https://archive.physionet.org/physiobank/database/.

[15] A. Bhattacharyya, L. Singh, and R. B. Pachori, "Fourier–bessel series expansion based empirical wavelet transform for analysis of nonstationary signals," *Digit. Signal Process. A Rev. J.*, vol. 78, pp. 185–196, Feb. 2018.

[16] J. Gilles and K. Heal, "A parameterless scale-space approach to find meaningful modes in histograms - application to image and spectrum segmentation," *Int. J. Wavelets, Multiresolut. Inf. Process.*, vol. 12, no. 6, p. 1 450 044, 2014.

[17] J. Gilles, "Empirical wavelet transform," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3999–4010, 2013.

[18] H. Alt and M. Godau, "Computing the fréchet distance between two polygonal curves," *International Journal of Computational Geometry & Applications*, vol. 5, no. 01n02, pp. 75–91, 1995.

Few-Shot Audio Classification Model for Detecting Classroom Interactions Using LaSO Features in Prototypical Networks

Rashed Iqbal, Christian Ritz, Jack Yang and Sarah Howard
University of Wollongong, NSW, Australia
E-mail: mri510@uowmail.edu.au, critz@uow.edu.au, jiey@uow.edu.au, sahoward@uow.edu.au

***Abstract*— This research introduces an innovative approach to few-shot sound classification applied to classroom sound recordings that integrates Label Set Operation (LaSO) features with Prototypical Networks. Traditional audio classification methods often require extensive labeled datasets, which can be impractical in real-world scenarios where obtaining large amounts of labeled audio data is challenging. This is particularly the case for the target application of automatically annotating long recordings of classroom audio to understand student learning in classrooms. This paper proposes an enhanced few-shot learning approach based on Prototypical Networks by incorporating LaSO features, to augment the feature space for the Prototypical Network. This methodology focuses on detecting and classifying teacher and student voices for future understanding and analysis of classroom interactions. Experimental results indicate the proposed approach incorporating LaSO features significantly improves classification accuracy of a prototypical network used for few-shot learning. This work paves the way for more advanced and automated solutions in educational environments, facilitating better monitoring and understanding of classroom dynamics.**

I. INTRODUCTION

A classroom environment is filled with various sounds that indicate the range of learning activities taking place. These sounds include teacher lectures, student questions, group discussions, and background noises. For education researchers, detecting patterns in these sounds can provide deeper insights into student learning dynamics. However, manually analyzing large volumes of recordings over extended periods can be impractical. Consequently, this study investigates an automated approach for classifying sounds in classroom audio recordings.

Previous techniques for classifying classroom audio involve training neural networks using features extracted from labeled classroom sound recordings, as seen in [1, 2]. The DART method uses straightforward features based on sound power levels, similar to the neural network approaches in [1]. In contrast, time-frequency features such as the mel-spectrogram [2] are frequently used. The neural networks applied for classroom sound classification in [1, 2] include Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and RNN variants like Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. This classification falls within the broader field of environmental sound or scene classification[3, 4].

In the recent past , the author explored utilizing sound power level features derived from the Decibel Analysis for Research in Teaching (DART) algorithm [5] to classify classroom audio captured by their developed system [6]. Other scholars have examined different techniques for analyzing classroom audio as well. For example, a Multi-Scale Audio Spectrogram Transformer (MAST) was developed to detect interactions between teachers and students during classroom activities [7]. Nonetheless, this research mainly focused on verbal exchanges between teachers and students, neglecting other important classes essential for understanding significant learning activities.

Environmental sound recognition aims to categorize different types of sound events within a recording. This fundamental task in machine listening has numerous practical applications, including smart cities [8, 9] and bioacoustics [10]. Although recent studies have made significant progress in recognizing sound events using extensive labeled datasets [11, 12], these approaches often fall short in real-world situations. This is due to the substantial effort required to collect enough annotated data for each category during the inference stage.
Recently, several studies have suggested employing few-shot learning for environmental sound classification [13-15]. These classifiers can quickly learn new acoustic patterns from a limited number of labeled examples, mainly because of their unique training objective. Nevertheless, they remain constrained to using ground truth as a binary attribute also they are not implemented on long audio streams.

This study involves first extracting LaSO [16] features from audio samples using a pre-trained base model. These features are then used as input to a Prototypical Network, which is trained to classify sound events with minimal labeled examples by calculating distances further details in section II. We apply this framework to the domain of classroom interaction

detection, focusing on identifying specific interactions and sounds within a classroom setting yield an even better understanding of classroom. The contribution of the study described in three parts:

i. We utilized scalogram audio features to train the few-shot learning model
ii. Extracted LaSO features from audio samples using a pre-trained based model
iii. LaSO features are then incorporated in prototypical network to generate protypes and classify long classroom audio.

Section II outlines the related work on few shot sound classification and prototypical network. Section III encompasses an in-depth methodology. Section IV describes the experimental setup and data collection method. Section V discussed the results and detailed analysis of the and the performance comparisons across different models.

II. RELATED WORK

The issue of few-shot learning has garnered significant interest in the computer vision field as well as sound classification recently. In the Meta-Learning, or learning-to-learn, approach [17-20], models are trained on instances of few-shot learning tasks instead of individual labeled samples.

A. Few shot Sound Categorization

There are several studies that apply few-shot learning to everyday sound recognition [13, 15, 19, 21, 22]. Heggan et al. experimented with various few-shot algorithms on everyday sound datasets for single-label classification [15]. Wang et al. addressed the multi-label few-shot problem by creating synthesized datasets, FSD-MIX-CLIPS and FSD-MIX-SED, and compared model performances by controlling generative factors in FSD-MIX-SED [13]. Cheng et al. adapted existing single-label few-shot algorithms to multi-label classification using a One-vs.-Rest strategy [22], and tested their methods on the AudioSet [23] dataset. Shi et al. applied meta-learning algorithms and linear regression on AudioSet, finding that meta-learning outperformed other few-shot methods [19]. It is important to note that while both Cheng and Shi used AudioSet for their experiments, their results cannot be directly compared as AudioSet is not publicly released, and neither study detailed how the database was adapted for few-shot learning.

B. Prototypical network

A training set $A = \{(M_i, O_i)\}_{i-1}^{A}$ where M_i represents the feature vector, $O_i \epsilon O$ signifies the discrete label of the i-th example, and O is the label set comprising O classes, $O = \{1, \dots, O\}$. Prototypical networks are trained using a series of "N-way K-shot" classification tasks created from the training set A.

In an "N-way K-shot" problem, the classification task consists of three components: (a) a subset O_s of N classes sampled from the set O, (b) G examples (support examples) drawn from A for each class in O_s, and (c) P examples (query examples) also drawn from A for each class in O_s. For any class $n \in O_s$, let Sn be the set of support examples for that class, with $|S_n| = G$. The prototype b_n (1) for class n is calculated as the mean of the embedding vectors of the support examples in S_n. Formally:

$$b_n = \frac{1}{G}\sum_{(M,O)\in S_n} g_\phi(M) \tag{1}$$

where g denotes the embedding mapping realized by the model whose parameters are denoted collectively as ϕ.

For a given query example M_p, the model classifies by generating a probability distribution over N classes in O_s using a softmax function applied to the distances between M_p and the N prototypes within the embedding space. Specifically, the likelihood that M_p belongs to class $n \in O_s$ is computed as follows (2):

$$r_\phi(\hat{z}_p = n | M_p) = \frac{\exp\left(-a\left(g_\phi(M_p), b_n\right)\right)}{\sum_{l\epsilon O_s} \exp\left(-a\left(g_\phi(M_p), b_l\right)\right)} \tag{2}$$

Where $\hat{z}_p$ represents the predicted label for M_p, and a is a distance metric, such as ℓ_2 or cosine distance. The network is optimized to reduce the negative log-probability of the correct class across the $N \times P$ query examples in (3):

$$\Gamma(\phi) = \sum_{(M,O)\epsilon P} -log r_\phi(\hat{z} = O|M) \tag{3}$$

where P is the set of query examples, $|P| = N \times P$.

While prototypical networks excel in various applications [24, 25], they are not directly applicable to multi-label few-shot classification. This is because formulating "N-way K-shot" problems becomes challenging when labels frequently co-occur in a multi-label context.

III. PROPOSED METHOD

This study presents an innovative method for few-shot sound classification, focusing on scalogram spectral utilization for detecting classroom interactions by combining Label Set Operation (LaSO) features with Prototypical Networks. Scalogram features extracted from the raw audio files and used as input to the label set operation features where a pretrained model utilized and backbone of the feature extraction.

A. Scalogram conversion

A scalogram, (4) similar to a spectrogram, is formed from the absolute values of the Continuous Wavelet Transform (CWT) coefficients across time and scale in a two-dimensional format [23]. This representation has proven to be more effective than other time-frequency features in neural network-based audio classification [17].

$$CWT_c(s,t) = \int_{-\infty}^{\infty} x(u)\frac{1}{\sqrt{s}}\psi^*\left(\frac{u-t}{s}\right)du \qquad (4)$$

In this context (1), $x(u)$ is the input signal, is s the scale parameter, which is related to frequency, ψ^* is the conjugate of the mother wavelet, t is the translation parameter that shifts the wavelet function along the time axis, and u denotes the segment of the signal.

B. Label Set Operation (LaSO)

The method is schematically depicted in Fig. 1 The input scalograms images U and V, each associated with a corresponding set of multiple labels $L(U)$ and $L(V)$, respectively, are mapped into the joint feature space $\mathcal{H}$ as H_U and H_V. This feature space $\mathcal{H}$ is implemented using a backbone feature extraction network $\mathcal{B}$; for our experiments, we utilized InceptionV3[26] and ResNet-50 [27] backbones. Three LaSO networks, named T_{int}, T_{uni}, and T_{sub}, process the concatenated H_U and H_V to generate synthesized feature vectors within the same feature space $\mathcal{H}$. As indicated by its name $int = intersection$, T_{int} aims to synthesize a feature vector in (5).

$$T_{int}(H_U, H_V) = W_{int} \epsilon\, \mathcal{H} \qquad (5)$$

This corresponds to a theoretical scalogram image E for which $\beta(E) = W_{int}$ and $L(E) = L(U) \cap L(V)$. Essentially, this means that if a human were to observe and label I, the label set would be $L(U) \cap L(V)$. Similarly, T_{uni} and T_{sub} generate outputs W_{uni} and W_{sub} in $\mathcal{H}$, which are anticipated to correspond to the union of the label sets $L(U) \cup L(V)$ and the subtraction of the label sets $L(U) \cup L(V)$, respectively [16]. In a nutshell the LaSO method can generate samples representing the intersection, union, or set difference of labels from two inputs, and generalizes to unseen labels, offering promising potential for our audio data.

C. LaSO based Prototypical network

InceptionV3 is used as the base model, pre-trained on ImageNet. Then we exclude the final classification layer. The global average pooling layer converts the feature maps from the base model to a fixed-size vector by taking the average of each feature map. The pooled feature vectors the fed into LaSO networks are designed to learn features based on the intersection (T_{int}), union (T_{uni}), and subtraction (T_{sub}) of label sets. Scalogram spectral images are concatenated then passed through multiple dense layers with ReLU activation and dropout layers.

For each pair of support and query features, we apply the LaSO T_{int}, T_{uni}, T_{sub} networks to compute intermediate, universal, and subset-specific feature in (6-8):

$$W_{int} = T_{int}(H_u, H_v) \qquad (6)$$
$$W_{uni} = T_{uni}(H_u, H_v) \qquad (7)$$
$$W_{sub} = T_{sub}(H_u, H_v) \qquad (8)$$

Where $W_{int}, W_{uni}, W_{sub} \in \mathbb{R}^k$.

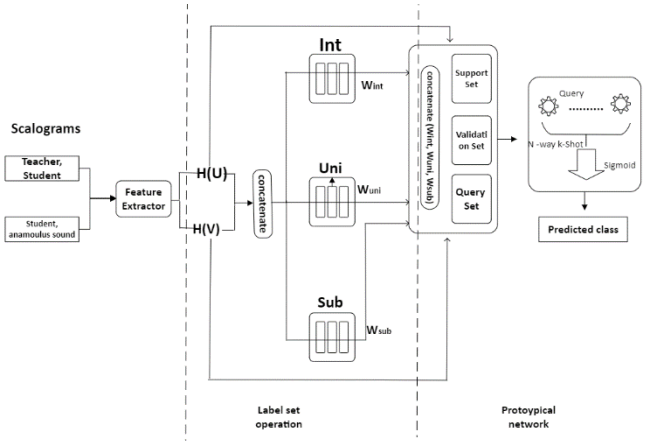


Fig. 1 LaSo based prototypical network process flow

Then Concatenated the LaSO features to form a combined feature vector (9):

$$W = concat(W_{int}, W_{uni}, W_{sub}) \in \mathbb{R}^{3k} \qquad (9)$$

In each episode of a prototypical network, we have a support set G and a query set P. For N-way K-shot learning, G contains N classes with K examples per class, and P contains query examples.

The prototype b_n (10) for each class n is found as the mean of its support examples

$$b_n = \frac{1}{K}\sum_{i=1}^{K} W_i^n \qquad (10)$$

Where W_i^n represents the combined LaSO feature for the i-th support examples of class n. Distance (11) between the query example W_P and each class prototype b_n:

$$a(W_p, b_n) = \|W_p - b_n\|_2^2 \qquad (11)$$

Then assign the query example W_p to the class with the nearest prototype $\hat{z}_p$ in (12):

$$\hat{z}_p = \arg min_m\, a(W_p, b_n) \qquad (12)$$

IV. Experimental Setup

We employed prototypical networks for our study. The proposed LaSO based ProtoNet models were trained using 3-way classification tasks to align with the evaluation scenario, where only 3 classes (Teacher, Student, Anomalous sound) are considered. We excluded irrelevant labels for classes and were not included in the "N-way K-shot" setup.

The audio in a ninth-grade science classroom at an urban Australian high school was captured. This study, which extends a previous project, received approval from the University of Sydney's ethics committee and the New South Wales (NSW) Department of Education. Consent was obtained from the students, their parents, and the teacher. The classroom followed a Bring Your Own Device (BYOD) policy, using Microsoft OneNote for learning and note-taking. It was equipped with four cameras and audio recorders. Audio was recorded using two side cameras and a digital recorder with an external

microphone focused on the teacher. The class consisted of 25 students and one teacher, meeting for 80 minutes four times a week.

In all experiments, audio recordings were sampled at 16 kHz, and the input was the Scalogram. The scalograms were generated from audio recordings using a 5 s window length. We applied inception V3 feature extractor for the audio then constructed LaSO based prototypical network and validated with validation set. Additionally, the network was tested on a long audio stream to evaluate the real-world implementation.

V. Results and Analysis

This section presents results from classifying classroom sounds using prototypical networks for an initial training and test database and for a separate unseen test database. Results are also presented when training these networks for classifying environmental sounds using two alternative well-known environmental sound databases (ESC [28] and Urban Sound [29]).

A. Training and initial Testing Performance

A database of 40 example audio segments across all 3 classes (13 teacher, 13 student and 14 anamolous classes) was prepared. From this database, a training set of 21 examples was selected and training is performed by randomly selecting 1 example of each class from this database, with the remaining 18 examples (6 per class) used for validation. This was repeated 3 times using a different example for each class each time (hence, three different models were trained). A separate test database was formed from the remaining 19 examples.

The validation accuracy in table 1 shows the validation accuracy for one of the trained models (similar results were achieved for the other two models). The LaSO+Prototypical Network consistently outperforms the Prototypical Network across all scenarios. In the 3-way 1-shot setting, it achieves a validation accuracy of 62.05%, which is significantly higher than the Prototypical Network's performance. This trend continues with 82.90% accuracy in the 3-way 3-shot scenario and 95.95% in the 3-way 5-shot scenario. The LaSO+Prototypical Network's superior performance can be attributed to the integration of LaSO features, which enhances the model's ability to capture and utilize important relational and contextual information within the data.

Table 1. Validation accuracy of the trained model

Model	3-way 1-shot	3-way 3-shot	3-way 5-shot
Prototypical net	54.45%	74.18%	87.75%
LaSO+Prototypical net	62.05%	82.90%	95.95%

Table 2 shows the average results for the test dataset across all 3 models. The Prototypical Network shows reasonable performance, with accuracy improving from 55.35% in the 1-shot scenario to 88.75% in the 5-shot scenario. However, the LaSO+Prototypical Network significantly outperforms the Prototypical Network alone in all scenarios, achieving 63.65%, 84.60%, and 99.55% accuracy in the 1-shot, 3-shot, and 5-shot settings.

Table 2. Test accuracy of the proposed model

Model	3-way 1-shot	3-way 3-shot	3-way 5-shot
Prototypical net	55.35%	75.18%	88.75%
LaSO+Prototypical net	63.65%	84.60%	99.55%

Table 3 shows average results using different performance measures across the 3 models when used to classify the test set. The LaSO Prototypical Network across different few-shot learning scenarios exhibit a notable performance. For the 3-way 1-shot scenario, the mean Average Precision (mAP) is 62.05%, with precision at 62.00%, recall at 61.89%, and an F1 score of 60.05%. As the number of shots increases, the model's performance improves significantly. In the 3-way 3-shot scenario, the mAP rises to 82.90%, precision to 82.10%, recall to 81.50%, and the F1 score to 82.90%. The 3-way 5-shot scenario shows the highest performance with a mAP of 95.76%, precision at 95.50%, recall at 95.23%, and an F1 score of 95.50%. These results highlight the LaSO Prototypical Network's strong capability to learn and generalize from limited data, with consistent improvements in precision, recall, and F1 score as more examples per class are provided.

Table 3. Various accuracy matrices of the proposed model

Model	3-way 1-shot	3-way 3-shot	3-way 5-shot
mAP	62.05%	82.00%	95.76%
precision	62.00%	82.10%	95.50%
Recall	61.89%	81.50%	95.23%
F1 Score	62.05	82.90%	95.50%

B. Results from evaluation on a second test database

A second test database was formed from the classroom audio recordings and consisting of 40 examples (13 teacher, 13 student, and 14 anomalous examples). Table 4 shows the accuracy results for this test database.

Table 4. Testing results on unseen data

Model	3-way 1-shot	3-way 3-shot	3-way 5-shot
Prototypical net	25.35%	61.25%	70.75%
LaSO+Prototypical net	60.65%	80.60%	95.00%

The Prototypical Network achieved an accuracy of 25.35%, while the combined model significantly outperformed it with 60.65%. In the 3-way 3-shot scenario, with three examples per class, the Prototypical Network's accuracy improved to 61.25%,

but the combined model still showed superior performance with 80.60%. Finally, in the 3-way 5-shot scenario, which involves five classes with five examples per class, the Prototypical Network reached 70.75% accuracy, whereas the LaSO+Prototypical Network achieved a notably higher accuracy of 95.00%.

As an example, Fig. 3 shows a 1-minute portion of the recorded classroom sounds with labeled sound class segments. The first image shows the predicted labels: '1 (Student)', '2 (Teacher)', and '3 (Anomalous Sound)', while the second image shows the ground truth graph with corresponding Letters: 'S (Student)', 'T (Teacher)', and 'A (Anomalous Sound)'.

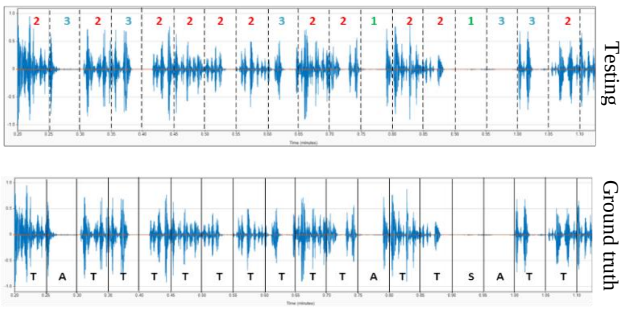


Fig 3. Testing and ground truth plot on long audio stream

There are a total of 36 instances which includes 27 Teachers, 5 anomalous sound and 4 students in the ground truth. While tested through our proposed model, the prediction shows that 24 instances are predicted as Teacher, the Student segment is predicted correctly, and anomalous sound predicted 4 times where actual ground truth is 5 and anomalous sound predicted 3 times over 4.

C. Testing over ESC and Urban Sound dataset

The proposed LaSO+Prototypical models trained on scalograms was further evaluated using two existing sound datasets, ESC [28] and Urban Sound [29], and three chosen classes from each database. The classes considered for ESC are Rain; Rooster; and person sneezing while for the Urban Sound dataset are: Gunshot; Dog Barking; and Siren. From each database, a set of 40 examples across all classes was created. Training proceeded using one randomly selected example for each class, leaving 37 examples that were used for testing (12 examples each for the first two classes and 13 examples for the third class). Accuracy results are shown in Table 5, including results for the classroom dataset using the LaSO+Prototypical model shown in table 5.

Table 5. Testing the Proposed model on ESC and US dataset

Model	3-way1-shot	3-way 3-shot	3-way 5-shot
ESC	52.45%	72.18%	79.75%
Urban sound	56.05%	82.90%	92.95%
Classroom	60.65%	80.60%	95.00%

In the 3-way 1-shot scenario, where the model classifies among three classes with only one example per class, it achieved 52.45% accuracy on the ESC dataset, 56.05% on the Urban Sound dataset, and 60.65% on the classroom dataset. In the 3-way 3-shot scenario, with three examples per class, the model's performance improved, reaching 72.18% accuracy on the ESC dataset, 82.90% on the Urban Sound dataset, and 80.60% on the classroom dataset. In the 3-way 5-shot scenario, with five examples per class, the model achieved 79.75% accuracy on the ESC dataset, 92.95% on the Urban Sound dataset, and the highest accuracy of 95.00% on the classroom dataset

VI. Conclusion

This study presents an innovative approach for few-shot sound classification by leveraging scalogram spectral features to detect classroom interactions, combined with Label Set Operation (LaSO) features and Prototypical Networks. Unlike traditional few-shot learning that deals with single-label classification, the synthesis of samples with multiple labels adds complexity. This requires advanced techniques for combining features in a meaningful way that preserves the semantics of multiple the audio labels. By converting raw audio files into scalograms and using these features as inputs for LaSO-based feature extraction, the study demonstrates the efficacy of scalograms over other time-frequency features in neural network-based audio classification. Experimental results reveal that the LaSO-enhanced Prototypical Network significantly outperforms the traditional Prototypical Network across various few-shot learning scenarios, achieving higher accuracy, mean Average Precision (mAP), precision, recall, and F1 scores. Specifically, the LaSO+Prototypical Network showed substantial improvements in 1-shot, 3-shot, and 5-shot settings. Validation and testing on a separate test dataset as well as two publicly available sound class datasets further confirmed the model's effectiveness. This approach demonstrates a robust capability to generalize from limited data, making it a valuable method for sound classification in educational settings and other domains requiring precise classification from minimal data.

VII. Acknowledgment

This research data was obtained from the ARC Discovery Project DP130100481.

References

[1] R. Cosbey, A. Wusterbarth, and B. Hutchinson, "Deep learning for classroom activity detection from audio," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: IEEE, pp. 3727-3731.

[2] A. Mou, M. Milanova, and M. Baillie, "Active Learning Monitoring in Classroom Using Deep Learning Frameworks," in

International Conference on Pattern Recognition, 2022: Springer, pp. 384-393.

[3] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206-219, 2019.

[4] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," Applied Sciences, vol. 10, no. 6, p. 2020, 2020.

[5] M. T. Owens et al., "Classroom sound can be used to classify teaching practices in college science courses," Proceedings of the National Academy of Sciences, vol. 114, no. 12, pp. 3085-3090, 2017.

[6] S. K. Howard, J. Yang, J. Ma, C. Ritz, J. Zhao, and K. Wynne, "Using data mining and machine learning approaches to observe technology-enhanced learning," in 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), 2018: IEEE, pp. 788-793.

[7] F. Liu and J. Fang, "Multi-Scale Audio Spectrogram Transformer for Classroom Teaching Interaction Recognition," Future Internet, vol. 15, no. 2, p. 65, 2023.

[8] G. Dove, C. Mydlarz, J. P. Bello, and O. Nov, "Sounds of New York city," Interactions, vol. 29, no. 3, pp. 32-35, 2022.

[9] A. Mitchell et al., "Deep learning techniques for noise annoyance detection: Results from an intensive workshop at the Alan Turing Institute," The Journal of the Acoustical Society of America, vol. 153, no. 3_supplement, pp. A262-A262, 2023.

[10] V. Morfi et al., "Few-Shot Bioacoustic Event Detection: A New Task at the DCASE 2021 Challenge," in DCASE, 2021, pp. 145-149.

[11] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880-2894, 2020.

[12] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," arXiv preprint arXiv:2104.01778, 2021.

[13] Y. Wang, N. J. Bryan, J. Salamon, M. Cartwright, and J. P. Bello, "Who calls the shots? Rethinking few-shot learning for audio," in 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2021: IEEE, pp. 36-40.

[14] J. Liang, Q. Phan, and E. Benetos, "Leveraging label hierarchies for few-shot everyday sound recognition," 2022.

[15] C. Heggan, S. Budgett, T. Hospedales, and M. Yaghoobi, "Metaaudio: A few-shot audio classification benchmark," in International Conference on Artificial Neural Networks, 2022: Springer, pp. 219-230.

[16] A. Alfassy et al., "Laso: Label-set operations networks for multi-label few-shot learning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6548-6557.

[17] C. So, "Exploring meta learning: parameterizing the learning-to-learn process for image classification," in 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021: IEEE, pp. 199-202.

[18] X. Zhong, C. Gu, W. Huang, L. Li, S. Chen, and C.-W. Lin, "Complementing representation deficiency in few-shot image classification: A meta-learning approach," in 2020 25th international conference on pattern recognition (ICPR), 2021: IEEE, pp. 2677-2684.

[19] B. Shi, M. Sun, K. C. Puvvada, C.-C. Kao, S. Matsoukas, and C. Wang, "Few-shot acoustic event detection via meta learning," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: IEEE, pp. 76-80.

[20] T. Zhang, L. Yang, X. Gut, and Y. Wang, "A Task-Specific Meta-Learning Framework for Few-Shot Sound Event Detection," in 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP), 2022: IEEE, pp. 1-6.

[21] J. Liang et al., "Adapting language-audio models as few-shot audio learners," arXiv preprint arXiv:2305.17719, 2023.

[22] K.-H. Cheng, S.-Y. Chou, and Y.-H. Yang, "Multi-label few-shot learning for sound event recognition," in 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), 2019: IEEE, pp. 1-5.

[23] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2017: IEEE, pp. 776-780.

[24] J. Huang, F. Chen, K. Wang, L. Lin, and D. Zhang, "Enhancing prototypical few-shot learning by leveraging the local-level strategy," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: IEEE, pp. 1660-1664.

[25] R. Li, J. Liang, and Q. Phan, "Few-shot bioacoustic event detection: Enhanced classifiers for prototypical networks," 2022.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818-2826.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[28] K. J. Piczak, "ESC: Dataset for environmental sound classification," in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015-1018.

[29] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 1041-1044.

Cuisine Image Synthesis with Improved Multiscale GANs Guided by CLIP

Weiyi Xia* and Satoru Fujita†
* Hosei University, Tokyo
E-mail: weiyi.xia.9a@stu.hosei.ac.jp Tel: +81-42-387-4545
† Hosei University, Tokyo
E-mail: fujita_s@hosei.ac.jp Tel: +81-42-387-4545

***Abstract*—In this study, a novel approach using a multiscale generative adversarial network (GAN) was proposed to generate high-quality, realistic cuisine images from ingredient and instruction texts. A model, FoodGAN++, that integrated text and image features within a hierarchical structure was also proposed for generating images at multiple resolutions. The proposed approach leveraged the capabilities of multiscale GANs to capture intricate details and variations required for realistic food imagery. FoodGAN was enhanced by incorporating the contrastive language–image pretraining (CLIP) model to guide image generation. Using CLIP as both a sentence encoder and a contrastive loss function, the generated images were realistic and accurately aligned with their textual descriptions. This study offers an efficient and effective solution for generating cuisine images, thereby significantly contributing to computational food analysis. FoodGAN offers three major advantages: (1) it produces detailed and realistic images that accurately represent complex cuisine compositions; (2) it efficiently integrates text and image features, ensuring that generated images closely align with their corresponding textual descriptions; and (3) it promotes advancements in the field of computational food analysis by addressing unique challenges associated with generating realistic cuisine images by integrating CLIP for enhanced guidance and precision.**

I. Introduction

Artificial intelligence has garnered considerable interest and is widely applied in various fields, fostering considerable advancements in image generation models. Thus, developing sophisticated models for creating high-quality, realistic images of cuisines from text descriptions is a promising research area. The unique challenges posed by the complexity of food or cuisine images must be addressed, which necessitate intricate representations of various ingredients and cooking procedures.

This study primarily identified and implemented effective image-generation techniques for capturing the multifaceted nature of cuisines. Unlike simpler, spatially compact image categories such as birds or flowers, food images require a sophisticated approach to accurately represent the diversity of ingredients and cooking processes. To this end, a specialized generative adversarial network (GAN) model was developed and implemented herein for generating cuisine images from textual data, including recipe titles, ingredient lists, and cooking instructions. Multiscale GANs, particularly those based on three-stage architectures, have shown significant potential in this domain. These models efficiently generate high-quality images, which is particularly crucial for cuisine image generation where the inherent complexity and diversity of ingredients necessitate advanced image-generation techniques. To build upon these foundations, we conducted a comprehensive review of existing work in text-to-image synthesis, with a particular emphasis on GANs and their efficacy in generating images from text descriptions.

One of the key contributions of this work is the integration of the contrastive language–image pretraining (CLIP)[1] model, which serves as both a sentence encoder and a guiding mechanism for image generation. CLIP's strong performance in aligning images with textual descriptions enhances the model's ability to generate realistic and accurately aligned images. This addresses a major limitation in previous models that struggled with consistency between text and image. Additionally, we employed word2vec to encode ingredient texts due to its efficiency in capturing semantic relationships in domain-specific contexts like recipes. While newer models such as BERT exist, word2vec was chosen for its lower computational overhead and effectiveness in handling short, structured texts typical of ingredient lists. These enhancements allow FoodGAN-v2 to integrate both text and image features for high-quality cuisine image generation, offering a specialized solution to challenges in computational food analysis.

II. Related Work

A. GANs

GANs[2] are universal generative models that generate images. These models combine a network that generates images from random input vectors and another network that attempts to distinguish between real and generated images. GANs primarily learn patterns from real images and generate high-quality images that resemble real ones. In traditional GANs, the interaction between the generator and discriminator is formulated as a minimax game. The generator produces images that are indistinguishable from real data, whereas the discriminator accurately classifies real and generated data.

B. Conditional GANs

Conditional GANs[3] (CGANs) were developed to further optimize the generator and discriminator by providing class labels. CGANs provide additional information (class labels or other modal data) to the generator and discriminator. These

labels are first self-encoded and concatenated with the encoded features, which improve the discriminant ability of the discriminator. Owing to this additional information, CGANs can process single-modal and multimodal image data.

C. Multiscale GANs

StackGAN++[4] extended on the concept of CGANs for sequentially generating images of different resolutions—from coarse to fine—by stacking three GANs. Recurrent neural networks (RNNs) were used to extract short text features related to images. However, cuisine images contain various ingredients mixed in intricate ways and not all ingredients necessarily contribute equally to the final image. An extension of StackGAN++, CookGAN[5], addresses the insufficiency of Stackgan++ by integrating an attention mechanism combined with a gated recurrent unit. This attention mechanism autonomously identifies the placement of crucial ingredients without supervision. Thus, more realistic cuisine images were generated using an attention-based model to associate ingredients with images.

These advancements in GANs, particularly the introduction of CGANs and multiscale GANs, provide a strong foundation for developing sophisticated models that can generate high-quality cuisine images from textual descriptions. This study builds upon these foundations to address the unique challenges in computational food analysis.

III. FoodGAN

A. Overview

A cuisine image generation model, FoodGAN++, was proposed herein, as shown in Fig.1. It identified a latent feature space from the ingredient and recipe texts to generate a fine cuisine image. Based on StackGAN++ [4], FoodGAN-v2 generated cuisine images at low-, medium,- and high-resolution image generation stages. The detailed architecture of this network is further elaborated in Sections III-B–III-F.

B. Recipe text embedding

As shown in Fig.1, features of a given recipe's ingredient and cooking procedure texts were computed using word2vec and CLIP text encoders [6], respectively. The rationale for using word2vec in this context is based on its proven ability to capture semantic relationships between words, particularly for domain-specific data such as ingredient lists, where certain words (e.g., 'flour,' 'sugar') frequently co-occur in related contexts. Although word2vec learns word feature vectors by predicting words appearing in the text, it was pretrained alongside images for use in the proposed method to capture the association between ingredient texts and cuisine images. In particular, food image features learned by a convolutional neural network (CNN) were compared with their text features embedded by a primary word2vec representation. Then, shared semantics between images and texts were embedded in the modified word2vec. The UPMC-Food-101 dataset [7] containing numerous food photographs generated effective word embeddings that captured the nuances of cuisine images.

In contrast, the cooking procedure text (figure) often contains lengthy sentences. The previous version of FoodGAN[8] used DistilBERT, a variant of the BERT model, as the instruction (sentence) encoder because it performed similar to BERT while considerably reducing computational costs. However, a pretrained CLIP model was used as the instruction encoder herein; this CLIP model also uses BERT as its text encoder. Additionally, the CLIP contrastive loss was incorporated to enhance the model performance in encoding complex textual data. The details are described in Section III-F. The CLIP model was used owing to its superior capability in understanding and representing sentences, which surpass the capabilities of long short-term memory (LSTM) and traditional BERT-based models.

C. Multiscale GAN structure

A hierarchical structure was used for the generation part of the proposed method. The embedded content derived from the recipe text was processed using two different types of encoder to gain the text feature c. It considered c concatenated with a noise vector z drawn from a normal distribution $N(0, I)$ as the input. Then, multiple generators produced images at different scales, particularly at resolutions of 64^2, 128^2, 256^2. Before generating images, the upsampling block F_0 and F_i received the condition vector c and noise vector z as inputs. This resulted in $h_0 = F_0(c, z)$ and $h_i = F_i(c, h_{i-1})$, and multistage samples were consequently generated via $s_i = G_i(h_i)$. Following each generator G_i, the discriminator D_i compared the real image x_i and generated image s_i and minimized the following cross-entropy loss:

$$\mathcal{L}_{D_i} = -\mathbb{E}_{x_i \sim p_{\text{data}i}}[\log D_i(x_i)] - \mathbb{E}_{s_i \sim p_{G_i}}[\log(1 - D_i(s_i))] \quad (1)$$

where x_i is derived from the real image distribution p_{data_i} at the ith scale and s_i is derived from the generated image distribution p_{G_i} at the same scale. Multiple discriminators were trained in parallel, with each focusing on different image scales.Based on discriminator training, generators minimized the following loss function:

$$\mathcal{L}_G = \sum_{i=1}^{2} \mathcal{L}_{G_i}, \quad \mathcal{L}_{G_i} = -\mathbb{E}_{s_i \sim p_{G_i}}[\log D_i(s_i)] \quad (2)$$

Here, $\mathcal{L}_{G_i}$ represents the loss function to approximate the ith image distribution (i.e., p_{data_i}). Throughout the learning process, the discriminator D_i and generator G_i were alternately optimized until they converged.

D. Conditional and unconditional losses

GANs can be categorized into two types based on the noise input. Unconditional GANs are provided only with random noise as the input, whereas conditional GANs are provided with controlled features that are added to random noise. Previous studies on conditional GANs [4], [5] generated high-quality images by internally combining losses of conditional and unconditional parts. A similar approach to such a combination was adopted herein. The objective function for

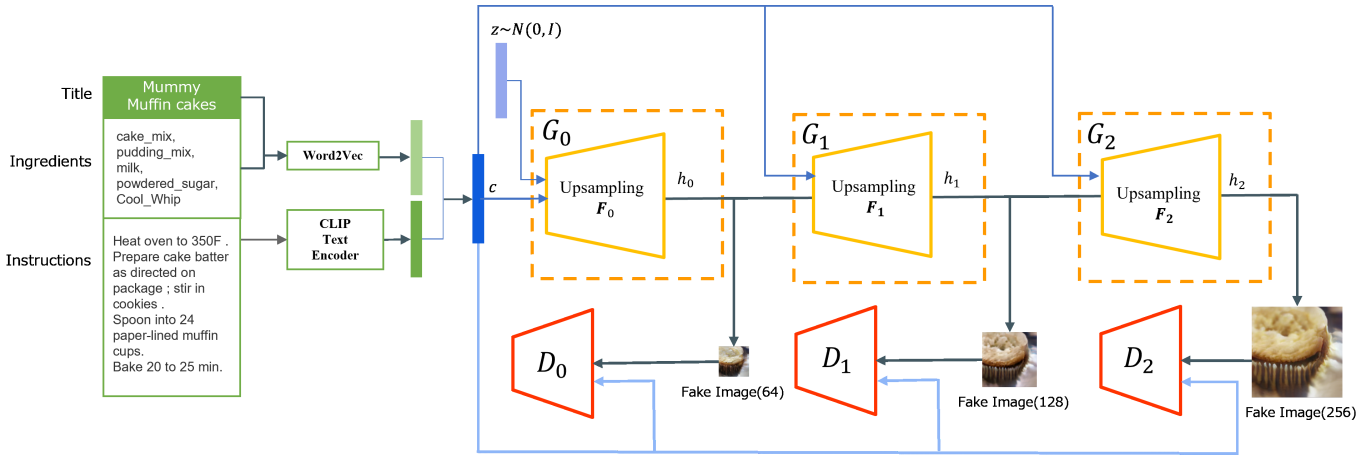


Fig. 1. Model structure

the conditional discriminator D_i comprises two terms, namely unconditional and conditional losses, as described by (3). In an implementation, wrong images were used, which were paired with a different recipe sampled from the training data as another input, except for the generated image, to maintain model robustness and enhance the discriminator performance.

$$\mathcal{L}_{D_i} = \underbrace{-\mathbb{E}_{x_i\sim p_{\text{data}_i}}[\log D_i(x_i)] - \mathbb{E}_{s_i\sim p_{G_i}}[\log(1-D_i(s_i))]}_{\text{Unconditional Loss}} + \underbrace{-\mathbb{E}_{x_i\sim p_{\text{data}_i}}[\log D_i(x_i,c)] - \mathbb{E}_{s_i\sim p_{G_i}}[\log(1-D_i(s_i,c))]}_{\text{Conditional Loss}}. \tag{3}$$

Similarly, the generator G_i of each scale approximated both unconditional and conditional image distributions. The loss for joint training was computed as shown in (4).

$$\mathcal{L}_{G_i} = \underbrace{-\mathbb{E}_{s_i\sim p_{G_i}}[\log(1-D_i(s_i))]}_{\text{Unconditional Loss}} + \underbrace{-\mathbb{E}_{s_i\sim p_{G_i}}[\log(1-D_i(s_i,c))]}_{\text{Conditional Loss}}. \tag{4}$$

E. Balanced consistency regularization

Balanced consistency regularization (bCR) [9]–[11] was introduced to address the issue of overfitting caused by standard data augmentation techniques. In traditional GAN setups, data augmentations (DA) such as color transformations and cutouts can degrade image quality by introducing artifacts. To mitigate this, we implemented bCR, which uses translation, scaling, and reflect padding to maintain the structural integrity of the images during augmentation. By ensuring that both the augmented and original images are treated as equivalent by the discriminator, bCR helps the model learn to focus on relevant features without overfitting to augmented distortions. In particular, a certain augmentation T was applied to both real images x and fake images $G(c)$. The augmented images were then represented by $T(x)$ and $T(G(c))$. At this point, L_{BCR} was included in the original discriminator loss:

$$\mathcal{L}_{\text{real}} = \|D(x) - D(T(x))\|^2 \tag{5}$$

$$\mathcal{L}_{\text{fake}} = \|D(G(c)) - D(T(G(c)))\|^2 \tag{6}$$

$$\mathcal{L}_{\text{bCR}} = \mathcal{L}_{\text{real}}\lambda_{bCR} + \mathcal{L}_{\text{fake}}\lambda_{bCR} \tag{7}$$

$$\mathcal{L}_{\text{D}} \leftarrow \mathcal{L}_{\text{D}} + \mathcal{L}_{\text{bCR}} \tag{8}$$

In our experiments, we explored varying the lambda value (λ_{bCR}) to determine its impact on model performance. This parameter controls the weight of the bCR loss in the overall objective function. Additionally, to prevent image quality degradation, color transformation and cutouts were eliminated and zoom-in, zoom-out, and rotation techniques were used instead. Moreover, reflect padding was implemented as part of the DA strategy to further enhance image generation quality and stability. Reflect padding involved extending the borders of an image by mirroring the pixels along the edges, which helped mitigate artifacts and improved the overall visual appeal of generated images.

F. CLIP contrastive loss

The pretrained CLIP model used as the sentence encoder was further leveraged to serve as a contrastive loss function[12]. Specifically, the generator was compelled to produce outputs that could be identified by the pretrained CLIP image and text encoders, designated as $\mathcal{E}$img and $\mathcal{E}$txt, using the contrastive cross-entropy loss; this loss was originally used to train these encoders.

During model implementation, the symmetric cross entropy loss was simultaneously computed in image and text alignment. This bidirectional computation ensured symmetry, allowing the model to consider matching accuracy in both directions during training. Such an approach enhanced the generalization ability of the model by accurately aligning generated images with the textual descriptions and vice versa. Considering the feature of the three-scale structure, this loss function was applied only in the third stage. By averaging the cross-entropy losses from both directions, potential training instability that

could arise from focusing on a single direction was mitigated. This averaging process contributed to a smoother and more effective training regime, ultimately improving the performance and robustness of the proposed model.

$$\mathcal{L}_{\text{CLIP}} = \mathbb{E}_{\{c_n\}} \left[-\log \frac{\exp(\mathcal{E}_{\text{img}}(G(c_0))^\top \mathcal{E}_{\text{txt}}(c_0))}{\sum_n \exp(\mathcal{E}_{\text{img}}(G(c_0))^\top \mathcal{E}_{\text{txt}}(c_n))} \right] \quad (9)$$

where $\{c_n\} = \{c_0, \ldots\}$ are sampled conditioning texts from the training data.

IV. Experiments

*a) **Dataset**:* Recipe1M[13] dataset containing 340,922 pairs of cooking recipes and images was used for evaluating the model performance. It is a rich resource for the proposed image generation model, providing diverse culinary data. Each recipe in this dataset is detailed and structured and contains three main components: title, ingredient list, and cooking instructions.

*b) **Evaluation metric**:* The quality of image generation was evaluated using the inception score (IS)[14] and Fréchet inception distance (FID)[15]. IS is one of the metrics used to evaluate the quality and diversity of images generated by the model. It is assessed using the Inception-v3 image recognition model [16], which employs CNNs. This metric evaluates the class classification score distribution of the generated images for determining the balance between quality and diversity. A high IS indicates that the generated images are diverse and realistic. As IS does not consider the distribution of real images, the FID was used herein for determining the distance between the distributions of real and generated images. Feature vectors are assumed to follow a multivariate normal distribution. Both IS and FID were calculated using 1000 generated images.

*c) **Implementation**:* FoodGAN-v2 was built based on the StackGAN++ code [4], with modifications to adapt it to the specific requirements of FoodGAN. The training was performed on a single GPU using the Adam optimizer with a batch size of 16. These parameters were adjusted to optimize the training process for the model and dataset.

A. Effectiveness of bCR

For the image generation experiments, the effectiveness of bCR in the proposed model was validated. A dataset comprising 5,000 cake images along with their corresponding recipes from the Recipe1M dataset was used in these experiments. The training process spanned 150 epochs. The baseline model for subsequent experiments was established without DA and bCR approaches. For a comprehensive evaluation, several experimental settings were considered:

- Baseline model (normal): The base model without any additional techniques.
- Model with DA: Augmented base model using DA techniques, including color transform[9].
- bCR variants:
  - bCR0: Applying translation.
  - bCR1: Incorporating parallel translation with reflect padding.
  - bCR2: Using random selection between parallel translation and scaling with reflect padding.
  - bCR3: Using scaling before parallel translation with reflect padding.
  - bCR4: Using scaling, parallel translation, and rotation with reflect padding.

TABLE I
Evaluation with different models

	λ_{bCR}	IS↑	FID↓
Baseline (normal)	/	4.19	56.57
Baseline + bCR0	5	4.17	60.41
	10	4.19	62.17
Baseline + bCR1	5	4.32	56.08
	10	4.09	53.45
Baseline + bCR2	5	4.27	52.26
	10	3.98	58.93
Baseline + bCR3	5	4.34	48.92
	10	4.28	48.21
Baseline + bCR4	5	4.33	45.67
	10	**4.42**	**41.08**

In this phase, bCR was introduced into the baseline model while the λ_{bCR} values were varied (Table I). Given that our initial tests with smaller λ_{bCR} values (1 to 4) showed minimal impact on results, we adopted values of 5 and 10, as used in the original bCR paper[11]. These experiments revealed the nuanced effects of bCR on image generation quality. Different variants exhibited varying degrees of improvement compared with the baseline model. When only the translation technique was applied to the images, the IS and FID decreased. However, after reflect padding was introduced in addition to DA techniques into the model (Section III-E), modest improvement in model performance was observed, particularly for bCR1. Subsequently, the impact of adding scaling (zoom out and zoom in) and rotation to the parallel translation process was explored, and significant increases in IS and FID were observed for bCR4 with $lambda$ of 10. This indicated its effectiveness in reducing image distortion. Encouraged by these findings, this specific setting for used for subsequent image generation experiments.

B. Effectiveness of CLIP and contrastive loss

Herein, the effectiveness of pretrained Word2Vec, DistilBERT, and CLIP models as word or sentence encoders in the proposed model was evaluated. The baseline model used Word2Vec and DistilBERT without the CLIP contrastive loss. To assess the improvement in model performance after incorporating the CLOP contrastive loss, either Word2Vec or DistilBERT was replaced with the CLIP encoder. Similar to the methodology described in Section IV-A, experiments were conducted using a dataset of 5,000 cake images. Training was conducted over 600 epochs. An optimal configuration for generating high-quality, realistic images from textual descriptions

TABLE II
EVALUATIONS WITH DIFFERENT TEXT ENCODERS

Word Encoder	Sentence Encoder	L_{CLIP}	IS↑	FID↓
Word2Vec	DistilBert	-	4.17	43.72
Word2Vec	CLIP	-	4.24	45.75
Word2Vec	CLIP	+	**4.30**	**37.98**
CLIP	DistilBert	-	4.23	52.24
CLIP	DistilBert	+	4.15	40.67
CLIP	CLIP	-	4.22	48.86
CLIP	CLIP	+	4.18	43.71

TABLE III
IMAGE GENERATION EVALUATION

	Salad		Cake	
	IS↑	FID↓	IS↑	FID↓
StackGAN++	3.26	50.14	2.95	96.70
CookGAN	3.31	71.36	3.72	82.79
FoodGAN(ours)	3.91	27.81	4.36	38.65
FoodGAN-v2(ours)	**3.93**	**22.23**	**4.62**	**17.69**
Real Images	5.09	/	5.42	/

was identified by systematically varying text encoders and including the CLIP contrastive loss. The corresponding results are shown in Table II.

The baseline model, which used Word2Vec for word encoding and DistilBERT for sentence encoding without the CLIP contrastive loss, provided a solid starting point for subsequent comparisons. When DistilBERT was replaced with CLIP as the sentence encoder, the model performance improved noticeably. This indicated the superior capability of CLIP to capture and encode the complexity of textual data. Then, the CLIP contrastive loss was introduced, primarily when Word2Vec and CLIP were used as the word and sentence encoders, respectively. Using these parameters, the model achieved best performance. This suggests that the CLIP contrastive loss considerably enhanced the model's ability to generate high-quality, realistic images that aligned well with their textual descriptions. The optimal configuration involved using Word2Vec for word encoding and CLIP for sentence encoding, coupled with the CLIP contrastive loss. This highlighted the significant impact of this configuration on improving image generation quality in text-to-image synthesis models.

C. Cuisine image synthesis

Two food categories, cakes and salads, were selected for the experiments due to their varying levels of visual complexity. Cakes often have more intricate textures, patterns, and decorations, making them challenging to synthesize accurately, while salads typically involve a diverse mix of ingredients with different colors and shapes. By testing the model on these two distinct categories, we aimed to assess its ability to handle both simple and complex food images. Additionally, these categories were chosen because they are well-represented in the Recipe1M dataset, ensuring sufficient data for training and evaluation. In the first experiment, the model was trained on cake images, for which a dataset of 19,035 cake images and its corresponding recipes was used. The second experiment used a dataset of 17,202 salad images. This efficient learning and generation process highlighted the adaptability and robustness of the proposed model across different food categories. The training lasted 300 epochs, signifying a short learning period in both experiments.

Table III compares the previously reported models[4], [5] and FoodGAN[8] in terms of IS[14] and FID[15]. As StackGAN++ does not support long text inputs, only the title and ingredient text were used as inputs for generating images. The IS of real Images is shown as a baseline. For both salad and cake categories, the proposed model achieved a higher IS and lower FID than those of previous research. The proposed model showed considerably improved performance, particularly in the cake generation experiment, where the FID was nearly halved. We assume that more diverse and higher-quality images can be generated by conducting experiments with a larger dataset. Furthermore, increasing the number of training iterations could potentially lead to considerably better outcomes.

Fig.2 shows the examples generated for "Green Salad," "Mummy Muffin Cakes," and "Easy Pancake." The images generated by StackGAN and CookGAN exhibited significant distortions. Although the previous method improved the quality of culinary image generation to some extent, some issues such as shape distortion, blurriness, and dull colors persisted. However, compared with StackGAN, CookGAN, and FoodGAN, the proposed model generated images with the sharpest edges, the most vibrant colors, and the highest overall quality. Simultaneously, there are other examples of generated images in various cuisine categories, as shown in Fig. 3.

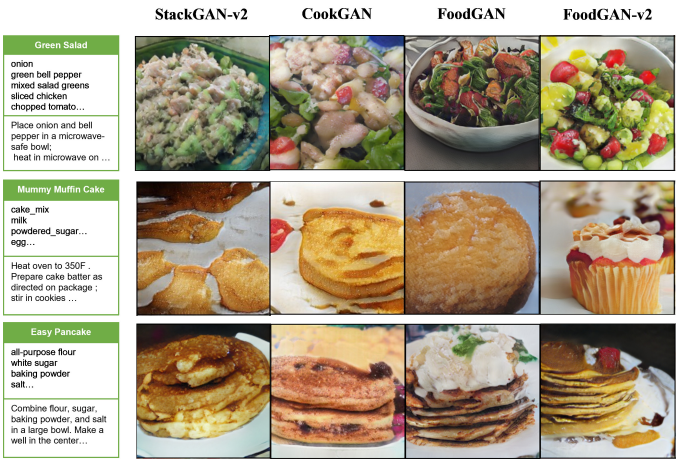


Fig. 2. Comparison of images generated by different models



Fig. 3. Examples of generated images

V. Conclusions

In summary, a GAN network was designed herein to generate cuisine images from recipe text, and the quality of generated images was evaluated. The proposed model exhibited excellent performance in generating images across various resolutions, from low to high, with realistic appearances and desirable diversity. Although the proposed model demonstrated strong performance in generating realistic cuisine images, certain limitations should be acknowledged. First, the model's reliance on text descriptions means that the quality of the generated images is highly dependent on the richness and accuracy of the input texts. Additionally, while the model performs well on the selected food categories, its generalization to other, more complex types of food remains to be tested. In future work, we aim to expand the model's capabilities by incorporating additional food categories and exploring diffusion-based models to further enhance the realism and variety of generated cuisine images.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP20K11777.

References

[1] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[3] M. Mirza and S. Osindero, *Conditional generative adversarial nets*, 2014. arXiv: 1411.1784 [cs.LG].

[4] H. Zhang, T. Xu, H. Li, *et al.*, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2019. DOI: 10.1109/TPAMI.2018.2856256.

[5] F. Han, R. Guerrero, and V. Pavlovic, "Cookgan: Meal image synthesis from ingredients," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1450–1458.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter*, 2020. arXiv: 1910.01108 [cs.CL].

[7] X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso, "Recipe recognition with large multimodal food dataset," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2015, pp. 1–6. DOI: 10.1109/ICMEW.2015.7169757.

[8] W. Xia and S. Fujita, "Foodgan: Realistic cuisine image synthesis with multi-scale generative adversarial networks," in *Proceedings of the IIEEJ International Conference on Image Electronics and Visual Computing*, 2024.

[9] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu, and S. Han, "Differentiable augmentation for data-efficient gan training," *Advances in neural information processing systems*, vol. 33, pp. 7559–7570, 2020.

[10] H. Zhang, Z. Zhang, A. Odena, and H. Lee, "Consistency regularization for generative adversarial networks," 2019.

[11] Z. Zhao, S. Singh, H. Lee, Z. Zhang, A. Odena, and H. Zhang, "Improved consistency regularization for gans," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 11 033–11 041.

[12] M. Kang, J.-Y. Zhu, R. Zhang, *et al.*, "Scaling up gans for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 124–10 134.

[13] J. Marín, A. Biswas, F. Ofli, *et al.*, "Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 187–203, 2021. DOI: 10.1109/TPAMI.2019.2927476.

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

A Novel LLM-based Two-stage Summarization Approach for Long Dialogues

Yuan-Jhe Yin*, Bo-Yu Chen†, and Berlin Chen*
* National Taiwan Normal University, Taiwan
† National Taiwan University of Science and Technology, Taiwan
E-mail: Yjyjames@gmail.com, patpatpat1015@gmail.com, berlin@ntnu.edu.tw

***Abstract*—Long document summarization poses a significant challenge in natural language processing due to input lengths that exceed the capacity of most state-of-the-art pre-trained language models. This study proposes a hierarchical framework that segments and condenses information from long documents, subsequently fine-tuning the processed text with an abstractive summarization model. Unsupervised topic segmentation methods identify semantically appropriate breakpoints. The condensation stage utilizes an unsupervised generation model to generate condensed data, and our current experiments employ ChatGPT (v3.5). The summarization stage fine-tunes the abstractive summarization model on the condensed data to generate the final results. This framework enables long documents to be processed on models even when the document length exceeds the model's maximum input size. The exclusion of the entire document from the summarization model reduces the time and computational resources required for training, making the framework suitable for contexts with constrained local computational resources.**

I. Introduction

Current summarization tasks are predominantly classified into extractive and abstractive summarization. Extractive summarization treats the task as a classification problem, determining which sentences in the article should be included in the summary. This method ensures that the summary accurately reflects the original article's meaning, preventing misinterpretations and potential hallucinations (factually incorrect or nonsensical content generated by language models). However, the direct extraction of summary content from the original text may hinder readability and present challenges in attaining high similarity scores compared to human-written labels during performance evaluation.

Abstractive summarization, on the other hand, employs language generation models to create summary content based on the input text. This approach often yields summaries that are more coherent and closely align with human-written labels. Nevertheless, the main challenges include achieving model convergence and preventing the generation of hallucinations or inaccuracies.

For longer text inputs, many studies utilize graph-based models to handle the summarization task [1]–[4]. By constructing graphs to represent the relationships between sentences, these models allow sentences that are relatively distant from each other in the text to influence one another. This approach mitigates the adverse effects of distance on summary content. During training, these models necessitate substantial VRAM, and as the input document length increases, the demand for computational resources becomes even more stringent.

A further challenge in summarizing lengthy documents occurs when the article length surpasses the model's input capacity, rendering effective model training impossible. Certain studies [5] tackle this issue through text segmentation techniques that partition articles into smaller, manageable segments. [6] employs a sliding window to encapsulate information throughout the document. Unlimiformer [7] modifies the Transformer architecture to accommodate longer input sequences.

We propose an architecture for preprocessing long articles, facilitating their training within commonly used language models. Prior research GoSum [8] indicates that segmentation is crucial for effective summarization. Extracting structural information from text enhances summary quality. Thus, our approach segments lengthy articles to prevent semantic discontinuity during text division, thereby ensuring quality of the summary. We utilizes the unsupervised segmentation method outlined in [9] for this architecture. This approach employs the TextTiling algorithm [10] and utilizes BERT embeddings to compute similarity, resulting in segmentation outcomes. Next, we employ zero-shot prompting with a large language model to condense the segmented text, allowing applicability to any dataset irrespective of golden labels. Subsequently, we fine-tune a Transformer-based abstractive summarization model to generate the final summary. The preceding segmentation and condensation steps facilitate processing and training with established summarization models.

The advantage of this architecture is that it enables efficient summarization of lengthy articles with reduced computational resources. When paired with prompts, this architecture yields superior outcomes on the ForeverDreaming dataset [11] . Furthermore, tailored prompts can be utilized based on the dataset and domain to enhance summary quality during the condensation stage.

II. Related Works

A. Unsupervised Topic Segmentation

The method outlined in [10] is widely employed for topic segmentation, identifying segment boundaries through shifts in word distribution. The process involves tokenization, determining lexical scores, and identifying boundaries to delineate topical segments. Building on this method, [9] uti-

lized BERT embeddings for sentence representation, achieving improved topic segmentation scores through incorporating BERT's strong semantic representation. Other studies focus on developing superior scoring models for unsupervised topic segmentation. For instance, [12] aims to enhance coherence scoring by training on diverse utterance pairs and defining relative scores. Additionally, [13] computes relevance scores by integrating topic similarity and coherence scores generated by trained topic and coherence encoders.

B. Dialogue Summarization

Dialogue summarization is an essential task that includes data types such as meetings, lectures, conversations, and scripts. The goal is to extract semantic information from dialogue content and succinctly present the key points.

DialogLM [14] proposed a pre-training framework featuring a window-based denoising approach that incorporates speaker masking, turn splitting and merging, text infilling, and turn permutation of the original dialogue. This framework enhances the model's comprehension of dialogue. Furthermore, they implemented Sinkhorn attention [15] in specific Transformer layers, enhancing local and global information interaction, thereby improving the model's suitability for long dialogue data. The framework is applied to UniLMv2 [16] and Longformer-Encoder-Decoder [6] in the study.

[17] proposes self-supervised methods that are similar yet distinct from DialogLM [14] for training a summarization model. For an input dialogue, interlocutors and utterances are alternated, with irrelevant utterances from other dialogues incorporated into the original dialogue. In the context of masking interlocutors, only the reference summary is utilized. A pre-trained BERT [18] is fine-tuned to detect modifications in the original dialogue. The weights are then shared to initialize a traditional encoder-decoder model, which is subsequently fine-tuned on an abstractive summarization task.

Some studies concentrate on augmenting the summarization task with supplementary information. In dialogues, the speaker's intentions are often implicit, posing challenges for models to fully comprehend the content. Thus, [19] suggests integrating commonsense knowledge to enhance the model's dialogue summarization. The presented framework utilizes commonsense knowledge model COMET [20] and PARA-COMET [21] to augment the input dialogue with additional knowledge. This augmented input is then fine-tuned using a pre-trained BART [22] model.

[23] proposes fusing static graphs representing dialogue discourse relations, keyword co-occurrences, speaker relations, and utterance positions. The static graphs are integrated with a dynamic graph created via a multi-head attention mechanism. This fused graph is used to generate summaries. This approach integrates both static information and the dynamic information potentially absent in static graphs.

C. Long Document Summarization

In long document summarization, two primary challenges are the document's input length exceeding the model's limit and the difficulty of capturing its structured information.

[24] addresses this issue by altering the cross-attention computation in Transformer architectures to enable unlimited input. Input data are encoded in chunks and stored in an index in GPU or CPU memory. In each cross-attention computation, the k-nearest neighbors of the query are selected from the index for inclusion in the computation. This approach guarantees the availability of each input sequence chunk for cross-attention computation.

In contrast to [25], which constructs a heterogeneous graph using word and sentence nodes, GoSum [8] builds the graph with sentence and section nodes, utilizing a reinforcement learning-based extractive summarization model. In this structure, sentence nodes connect to their respective section nodes, with both sentences and sections fully interconnected. This configuration conveys hierarchical semantic information and mitigates semantic drift across sections. Eliminating word nodes reduces the model's runtime and enhances the structural representation of lengthy documents.

[24] introduced a hierarchical propagation layer to facilitate information dissemination across multiple Transformer windows. The input data is segmented into multiple blocks and tokenized. The propagation layer initiates with a Transformer, subsequently processing the CLS token of each block through a BIGRU network to aggregate semantic information across blocks. This architecture enables BERT to handle input data exceeding its maximum size while retaining information from the entire document via updates to the CLS representation.

The study [26] introduces three methods for long document summarization: 'Direct', 'Chunk and Summarize', and 'Extract then Summarize'. The Direct approach omits input data preprocessing, with automatic truncation to the model's maximum token limit. The 'Chunk and Summarize' approach segments the input document at the maximum token limit, processes each segment independently, and merges the resulting summaries. Another variant entails inputting the generated summary with the subsequent segment into the model. The Extract then Summarize approach generates sentence embeddings for dialogues and summaries, subsequently calculating cosine similarity. A similarity threshold is established to exclude sentences with low similarity. Ultimately, the extracted dialogues are input into the models to generate summaries. The experiment utilized several state-of-the-art summarization models, including Longformer [6], T5 [27], Flan T5 [28], BART [22], and ChatGPT, and was evaluated on datasets such as Qmsum [25] and ForeverDreaming [11]. Since we are utilizing ForeverDreaming [11] in our evaluation experiments, its results provide baseline comparisons.

III. Proposed Framework

We propose a hierarchical framework for document summarization that involves multiple stages. First, the input document is divided into segments based on semantic information. These segments are then condensed using zero-shot prompting to generate a first-stage summary. Next, each first-stage summary is concatenated with k utterances extracted from the original

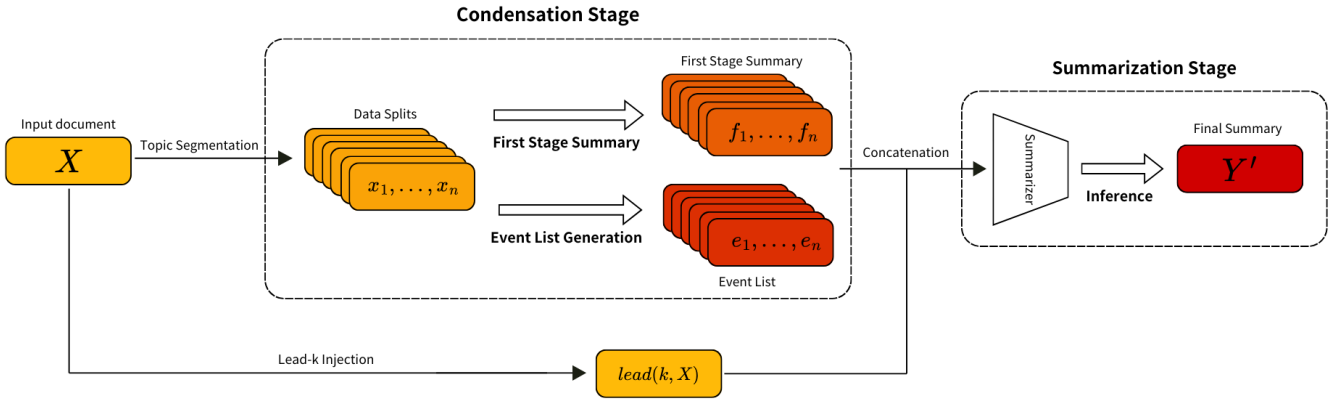


Fig. 1. Workflow of the proposed framework.

input. Finally, the processed data is fine-tuned using an abstractive summarization model to produce the final summary.

A. Topic Segmentation

A single input document consists of n utterances $X = \{u_1, u_2, ..., u_n\}$. We aim to determine the semantic breakpoints in the document for data splitting. Let semantic breakpoint sb_j be the index of last utterance in the jth segment, then we will have segment index set $SEG = \{sb_1, sb_2, ..., sb_l\}$, with l segments . We referenced the approach proposed in [9] and adjusted it according to our requirements and experimental results. Unlike [9], our approach does not apply the block-wise max pooling operation. Instead, a sliding window method is used for BERT embedding calculation. Specifically, the embedding of ith utterances eb_i is calculated by concatenating the previous and next utterances:

$$eb_i = tokenizer(u_{i-1} \oplus u_i \oplus u_{i+1})$$

$$sim_i = cos(eb_i \cdot eb_{i+1})$$

where $cos(\cdot)$ denotes cosine similarity function and sim_i denotes the cosine similarity between the BERT embeddings of the utterances u_i and u_{i+1}.

[9]'s work specifies the semantic breakpoint to be determined by every sim_i such that:

$$sim_i < \mu_S - \sigma_S$$

where μ_S, σ_S is the mean and variance of all cosine similarities $S = \{sim_1, sin_2, ...sim_n\}$. This method resulted in too many semantic breakpoints when the variance was too small, leading to excessive segmentation and loss of meaningful context. Therefore, we adopted a greedy approach for selecting semantic breakpoints that is more-suitable for our experiment. The steps are as follows: 1. Separate the utterance with the least cosine similarity as a semantic breakpoint. 2. Protect the nearby w utterances from segmentation. 3. Repeat steps 1 and 2 until the specified number of segments l is reached or no suitable breakpoint can be found.

w and l are customizable parameters. By protecting adjacent sentences from being split and allowing for an adjustable number of segments, the algorithm can better adapt to various scenarios.

B. Data Splitting

To accommodate the maximum input size of the summarization model, we perform a data split based on the results of topic segmentation, ensuring that semantic meaning is preserved. The pseudo segment size M is determined by the model's maximum input capacity. Starting from every M utterances in the input text, we locate the nearest semantic breakpoint. This process is conducted backwards to ensure that the length of each segment does not exceed M utterances. The ith data split denotes $x_i = \{u_{sb_j+1}, u_{sb_j+2}, ..., u_{sb_k}\}$, where j and k are some index in segment index set such that the length of utterances in segment $j+1$ to k is not greater than M. After data splitting, the input document $X = \{x_1 \cap x_2 \cap ... \cap x_m\}$ consists of m splits.

C. Condensation Stage

Performing summarization tasks using zero-shot prompting with ChatGPT may result in summaries presented from different perspectives or overly vague summaries (see the analysis section). Consequently, the generated summaries can deviate from the golden labels. By combining the first-stage summary with the event list method, our model can separately generate text with a summarization structure and neutral perspective descriptions. This approach allows the model in the summarization stage to obtain more comprehensive information. We use ChatGPT (v3.5) to generate the first-stage summary of the input document, ensuring that its length meets the input length requirements of the BART model. The first-stage summary generation process is as follows:

$$f_i = LLM(x_i, p_s)$$

$$F = f_1 \oplus f_2 \oplus ... \oplus f_m$$

where f_i is the first-stage summary of x_i , p_s denotes the prompt used to generate first-stage summary, and $LLM()$ denotes the zero-shot prompting model. The generated results are

concatenated to form the first-stage summary F corresponding to input document X.

The event list generation process is as follows:

$$e_i = LLM(x_i, p_e)$$

$$E = e_1 \oplus e_2 \oplus ... \oplus e_m$$

where e_i is the event list of x_i , p_e denotes the prompt used to generate event list, and E is the generated event list for X.

By generating the first-stage summary and event list, the information from the input document is condensed to a trainable length suitable for summarization models like BART.

D. Summarization Stage

To further enrich the training data, we inject the first k utterances of the original text into the training data. This ensures that the content processed by the LLM remains closely aligned with the original material, thereby enhancing summarization performance. We define enhanced input data as:

$$X' = E \oplus F \oplus lead(k, X)$$

where $lead(k, X)$ function extracts first k utterances from X. Finally, we use the enhanced input data from the training set to train a BART model. Let the golden summary be $Y = \{y_1, y_2, ..., y_k\}$, where y_i is the ith utterance of Y. The model minimizes the cross-entropy loss as follows:

$$L = -\sum_{t=1}^{T} logP(y_t|y_{<t}, X')$$

where y_t is the actual token at position t, $y_{<t}$ represents all tokens before position t, X' denotes the enhanced input data. Note that the enhanced input is also applied during the inference phase; the test set data will be processed in the same manner to generate the result summary.

IV. Experiment

A. Dataset

We conducted experiments on the ForeverDreaming dataset, which includes TV show transcripts paired with human-written summaries. The dataset contains 4,348 dialogues, comprising 3,673 training samples, 338 validation samples, and 337 test samples.

B. Baselines

We referenced the experimental data from the study [5], ensuring that the experimental details are consistent. The experiments were conducted on a randomly selected 15% subset of the ForeverDreaming dataset. Several approaches were presented, including 'Direct', 'Chunk and Summarize', and 'Extract then Summarize'. For detailed methods, please refer to the related works section. We present the result with the best ROUGE-1 score under their experimental settings for comparison. For the comprehensiveness of the baseline data, we also referenced DialogLM's results on the ForeverDreaming dataset. It is important to note that both our experimental results and those from DialogLM were obtained using the complete ForeverDreaming dataset.

TABLE I
Experiment Results and Baseline

Summarization	ROUGE-1	ROUGE-2	ROUGE-L
Our model	**29.15**	**6.72**	**24.54**
Baseline			
BART	19.35	2.02	17.28
LongFormer	15.39	1.42	12.52
T5	18.57	1.48	17.62
ChatGPT	25.04	4.34	18.31
DialogLM-large	**36.70**	**8.38**	**31.38**

BART: BART is a state-of-the-art denoising sequence-to-sequence pre-trained model, often used in natural language generation and translation. The BART-2048 model with the 'Chunk and Summarize' approach is used.

Longformer: Longformer replaces the standard self-attention mechanism for long document input, making it suitable for generative sequence-to-sequence tasks involving long documents. The Longformer model with the 'Extract then Summarize' approach is used.

T5: The Text-to-Text Transfer Transformer (T5) converts every language problem into a text-to-text format, commonly used for summarization tasks. The T5-base-2048 model with the 'Chunk and Summarize' approach is used.

ChatGPT: ChatGPT is a conversational generation model developed by OpenAI, based on the GPT-3 and GPT-4 architectures. It is designed for natural language interactions and provides high-quality exchanges in various application scenarios. ChatGPT (V3.5) with the 'Direct' approach is used.

DialogLM: DialogLM is a state-of-the-art pre-trained summarization model that focuses on long dialogue. Its window-based denoising method enables the model to gain a better understanding of long documents.

C. Experimental Results

Our framework outperforms the results obtained by [5]. The 'Chunk and Summarize' method is similar to ours in that it adopts the idea of segmenting input data. However, our approach extends further. By using an unsupervised topic segmentation method, the data splitting aligns more closely with the textual information. We then use the event list generation method to further integrate the segmented data, and finally, we fine-tune the summarization results using BART. This approach allows us to achieve a difference of more than 9 points in ROUGE-1 score compared to 'BART'.

The 'ChatGPT' result in the table does not use segmentation and performs better compared to 'Summary Generation' in Table 3. This indicates that using segmentation with ChatGPT may result in a lower summary score. ChatGPT achieves the best results when the article is not segmented. However, by employing event list generation and fine-tuning, we can achieve scores that surpass those of ChatGPT summarization.

DialogLM achieves better ROUGE scores in the Forever-Dreaming summarization task. The primary reason is the approach utilizes the full text of long documents and involves a

TABLE II
COMPARISON OF CHATGPT SUMMARIZATION

Methods	ROUGE-1	ROUGE-2	ROUGE-L
Event List Generation	22.47	3.39	19.69
First Stage Summarization	21.99	3.11	19.43
ChatGPT Second Stage Summarization	20.52	2.87	17.67
BART Second Stage Summarization	27.27	5.33	22.90

TABLE III
COMPIRSION OF LEAD-k INJECTION

Methods	ROUGE-1	ROUGE-2	ROUGE-L
Lead-0	28.96	6.58	24.40
Lead-1	29.02	6.59	24.45
Lead-3	28.38	6.32	23.80
Lead-5	**29.15**	**6.73**	24.54
Lead-10	28.97	6.72	**24.55**

relatively long pre-training process. The complete information is fed to the model instead of condensed text, requiring significant computational resources and training times. In contrast, our method requires fewer computational resources to achieve comparable results, making long document summarization feasible for computers with less VRAM.

V. ANALYSIS

A. ChatGPT Summarization

To assess the impact of different prompts on summary content, we conducted further analysis on results generated by the LLM during the zero-shot prompting data condensation stage. The result is presented in Table II. The 'Event List Generation' and 'First-Stage Summary' rows indicate the ROUGE scores of the concatenated event list E and the first-stage summary F compared to the golden label. The 'Second Stage Summarization' rows represent the further summarization of the concatenated first-stage summary.

The results for event list summarization marginally surpass those for direct text summarization. This discrepancy arises from the 'summarizing' prompt potentially producing vague or subjective summaries. Modifying the prompt to "event list" yields a text comprised of objective descriptions of the events, thereby more accurately reflecting the article's content.

The effects of ChatGPT summarization are evident in the 'First Stage Summarization' and 'ChatGPT Second Stage Summarization' rows. The same prompts were used to generate both results. The ROUGE score decreased with multi-stage zero-shot prompting summarization, indicating that while this method can be used to condense and shorten the original document, the information loss issue remains unsloved.

The results from the 'BART Second Stage Summarization' indicate that fine-tuning substantially enhances the ROUGE score, justifying our selection of BART as the final summarization model. Moreover, integrating 'Event List Generation' and 'First Stage Summarization' as inputs enables the model to capture the summarization structure and neutral descriptions of the original documents, thereby enhancing outcomes.

TABLE IV
COMPIRSION FOR DIFFERENT SUMMARIZATION MODEL

Model	ROUGE-1	ROUGE-2	ROUGE-L
Pegasus	25.57	5.20	21.59
BART	28.96	6.58	24.40

B. Lead-k Injection

The Lead-k Injection method also helps the model gain a better understanding of the original document, mitigating information loss during data preprocessing. To further understand the effects of Lead-k Injection, we tested how different values of k for injection affect the ROUGE score. Lead-0 indicates no injections, meaning the model is trained on the event list only. The table shows that Lead-k Injection enriches the input text, resulting in a better ROUGE score. However, excessively increasing k may lead to a loss of focus during the training phase, thereby decreasing model performance.

C. Summarization Model

In the Sammarization Stage, different summarization models can be applied to the framework. Experiments were conducted using both Pegasus and BART for comparison. Note that the presented results are generated with Lead-0 (no utterances injected), as the optimal k may vary for each model. The results show that BART outperforms Pegasus by approximately 6 points in ROUGE-1 score, indicating that BART is the superior model for our task.

VI. CONCLUSION

The proposed framework effectively summarizes long documents with low computational cost. Despite the data condensation process shortening the original document, some essential information may be lost. Currently, we use a naive solution that injects the first k utterances into the abstractive summarization model input. This method can be enhanced by employing ROUGE-based selection for the injected segment or by fine-tuning an extractive summarization model to identify and preserve key information from the original document. Summarizing long documents, such as those in the ForeverDreaming dataset, using common Transformer-based models, demands significant VRAM and extensive training time. Despite the additional costs of using the ChatGPT API, the framework is viable for scenarios lacking local computational resources.

REFERENCES

[1] D. Wang, P. Liu, Y. Zheng, X. Qiu, and X. Huang, "Heterogeneous graph neural networks for extractive document summarization," *arXiv preprint arXiv:2004.12393*, 2020.

[2] R. Jia, Y. Cao, H. Tang, F. Fang, C. Cao, and S. Wang, "Neural extractive summarization with hierarchical attentive heterogeneous graph network," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 3622–3631.

[3] Y.-N. Chen, Y. Huang, C.-F. Yeh, and L.-S. Lee, "Spoken lecture summarization by random walk over a graph constructed with automatically extracted key terms.," in *Interspeech*, 2011, pp. 933–936.

[4] N. Garg, B. Favre, K. Reidhammer, and D. Hakkani Tür, "Clusterrank: A graph based method for meeting summarization," 2009.

[5] A. Mullick, A. K. Bhowmick, R. Kokku, P. Dey, P. Goyal, N. Ganguly, *et al.*, "Long dialog summarization: An analysis," *arXiv preprint arXiv:2402.16986*, 2024.

[6] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[7] A. Bertsch, U. Alon, G. Neubig, and M. Gormley, "Unlimiformer: Long-range transformers with unlimited length input," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[8] J. Bian, X. Huang, H. Zhou, and S. Zhu, "Gosum: Extractive summarization of long documents by reinforcement learning and graph organized discourse state," *arXiv preprint arXiv:2211.10247*, 2022.

[9] A. Solbiati, K. Heffernan, G. Damaskinos, S. Poddar, S. Modi, and J. Cali, "Unsupervised topic segmentation of meetings with bert embeddings," *arXiv preprint arXiv:2106.12978*, 2021.

[10] M. A. Hearst, "Text tiling: Segmenting text into multi-paragraph subtopic passages," *Computational linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[11] M. Chen, Z. Chu, S. Wiseman, and K. Gimpel, "Summscreen: A dataset for abstractive screenplay summarization," *arXiv preprint arXiv:2104.07091*, 2021.

[12] L. Xing and G. Carenini, "Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring," *arXiv preprint arXiv:2106.06719*, 2021.

[13] H. Gao, R. Wang, T.-E. Lin, *et al.*, "Unsupervised dialogue topic segmentation with topic-aware utterance representation," *arXiv preprint arXiv:2305.02747*, 2023.

[14] M. Zhong, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Dialoglm: Pre-trained model for long dialogue understanding and summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 11 765–11 773.

[15] Y. Tay, D. Bahri, L. Yang, D. Metzler, and D.-C. Juan, "Sparse sinkhorn attention," *International Conference on Machine Learning*, pp. 9438–9447, 2020.

[16] H. Bao, L. Dong, F. Wei, *et al.*, "Unilmv2: Pseudo-masked language models for unified language model pre-training," in *International conference on machine learning*, PMLR, 2020, pp. 642–652.

[17] H. Lee, J. Yun, H. Choi, S. Joe, and Y. L. Gwon, "Enhancing semantic understanding with self-supervised methods for abstractive dialogue summarization," *arXiv preprint arXiv:2209.00278*, 2022.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] S. Kim, S. J. Joo, H. Chae, C. Kim, S.-w. Hwang, and J. Yeo, "Mind the gap! injecting commonsense knowledge for abstractive dialogue summarization," *arXiv preprint arXiv:2209.00930*, 2022.

[20] J. D. Hwang, C. Bhagavatula, R. Le Bras, *et al.*, "(comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 6384–6392.

[21] S. Gabriel, C. Bhagavatula, V. Shwartz, R. Le Bras, M. Forbes, and Y. Choi, "Paragraph-level commonsense transformers with recurrent memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 12 857–12 865.

[22] M. Lewis, Y. Liu, N. Goyal, *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[23] S. Gao, X. Cheng, M. Li, *et al.*, "Dialogue summarization with static-dynamic structure fusion graph," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 13 858–13 873.

[24] Q. Grail, J. Perez, and E. Gaussier, "Globalizing bert-based transformer architectures for long document summarization," in *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume*, 2021, pp. 1792–1810.

[25] M. Zhong, D. Yin, T. Yu, *et al.*, "Qmsum: A new benchmark for query-based multi-domain meeting summarization," *arXiv preprint arXiv:2104.05938*, 2021.

[26] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International conference on machine learning*, PMLR, 2020, pp. 11 328–11 339.

[27] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[28] H. W. Chung, L. Hou, S. Longpre, *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

Impulse response transforming method to control distance perception based on direct-to-reverberant energy ratio

Toru Takahashi* and Eita Morigaki† and Masato Nakayama‡
* Osaka Sangyo University, Japan
E-mail: takahashi@ise.osaka-sandai.ac.jp Tel/Fax: +81-735-3001
† Osaka Sangyo University, Japan
E-mail: s23mh02@ge.osaka-sandai.ac.jp Tel/Fax: +81-735-3001
‡ Osaka Sangyo University, Japan
E-mail: nakayama@ise.osaka-sandai.ac.jp Tel/Fax: +81-735-3001

***Abstract*—The perceptual distance can be controlled by convolving a dry source sound with a distance-dependent impulse response (IR). It is necessary to measure many IRs in advance to perform continuous distance control using this approach. However, this preliminary measurement is difficult in practice. In this article, we propose a method that uses a pre-existing IR for one distance as a seed and obtains IRs for other distances by transforming from the seed. The proposed method multiplies the late reverberation part of the IR by a constant value. In the proposed method, the constant value is determined to approximate the direct-to-reverberant energy ratio (DRR) of the IR at the target distance. We investigated the spatial resolution ability of five subjects to perceptual distance differences using measured IR. Similarly, when we investigated the spatial resolution ability using the transformed IR using the proposed method, we found a similar trend to that of the measured IR. It was suggested that in the 0.25 m step of the investigation range from 0.50 m to 4.00 m, the proposed method can generate IR that preserves distance perception.**

I. Introduction

Audio playback devices that reproduce sound environments have continued to be improved from monaural playback to stereo playback, then to 5.1 ch, 7.2 ch, and 22.2 ch playback, to approximate the actual sound environment and enhance the sense of presence. It is no exaggeration to say this is a history of reinforcing expressive abilities. These methods focus on precisely reproducing the audio waveform at the listening position using multiple speakers.

Recently, Egawa et al.[1] proposed a practical spatial representation method that combines Electro-Dynamic Loudspeaker(EDL) and Parametric Array Loudspeaker(PAL) and requires relatively few playback devices (Four EDLs and four PALs). The sound reproduced from PAL travels in a straight line, is less affected by the spatial transmission characteristics between the reproduction point and the listening point, and arrives as a sound close to the waveform of direct sound. The sound played from the EDL is affected by spatial transmission characteristics and arrives as a waveform that is a mixture of direct and indirect sounds.

The practical spatial representation method[1] uses four pairs of EDL and PAL to control the sense of distance and direction. The role of a pair of EDL and PAL is to change the sense of distance by controlling the direct-to-reverberant energy ratio(DRR)[2], [3]. The four pairs of EDL and PAL change the sense of direction by controlling the sound pressure level difference between the positions of the listener's ears. The method has the advantage that both distance and direction sense can be controlled by controlling only the interaural level difference (ILD) and DRR, without aiming for precise waveform reproduction. ILD and DRR control simply involves convolving the dry source played by EDL and PAL with an impulse response (IR) to reproduce distance and direction. The number of convolution control parts is limited to eight in the system, making it simpler to implement compared with methods that precisely reproduce waveforms. Control of the ILD and DRR is simply by convolving the IR with the waveforms sent from EDL and PAL. There have been many previous studies[4], [5] on distance perception to a sound source, and these studies have suggested that the direct-to-field ratio is an important clue. Therefore Egawa et al.'s approach is reasonable from a psychoacoustic perspective.

We hypothesize that if DRR is important in distance perception, distance perception should be preserved if DRR can approximate the DRR of the original IR without strictly reconstructing the IR waveform. If the hypothesis is true, the desired distance perception should be possible to construct from the sound obtained by transforming a certain IR to other distances by using the existing IR for a certain distance as a seed and transforming from the seed. Therefore, we propose a method that uses existing impulse responses as seeds and transforms impulse responses for other distances. Then, we conduct a subjective evaluation of the distance control speech using the IR transformed using the proposed method, and check whether the tendency of distance perception shows the same tendency as the distance perception using the measured IR.

II. Direct-to-reverberant energy ratio (DRR)

The direct-to-reverberant energy ratio (DRR) is one of the indicators that expresses the feature of the IR from a certain sound source to an observation point. If the IR is $h(n), n = 0, 1, \cdots, N$, the DRR of IR is defined by the following equation.

$$D_{rr} = \frac{\sum_{n=q}^{N} h^2(n)}{\sum_{n=0}^{q-1} h^2(n)} \tag{1}$$

N is sample length of IR, and q is the sample point that separates the direct sound component and the late reverberation component of the IR.

The shorter the distance from the sound source to the observation point, the greater the energy of the direct component received at the observation point than the energy of the late reverberant component. This is because the longer the distance to the observation point, the longer the propagation distance of the components of the late reverberation component, and the greater the distance attenuation effect. Therefore, the DRR is a clue to distance perception. It is known that humans use DRR as one of the cues to perceive the distance to a sound source. In addition to DRR, humans also use loudness as cues, but this study focuses on DRR. An example where DRR is an important judgment indicator is when determining distance between a loud sound in the distance and a soft sound nearby. When there are many objects that reflect sound near the sound source or observation point, it tends to show a high DRR. Although we would like to analyze DRR analytically as one of the important clues for distance perception, it is not easy to analyze it separately from other clues. Therefore, we begin by measuring IRs from 0.50 m to 4.00 m and investigating how people perceive distance when presented with these IRs. Then, when the distance perception obtained when IR is imitated by transform or synthesis is similar to the perceptual result using measuring IRs, we define that the transformation or synthesis method is superior.

In this article, we investigate whether subjects can discriminate sounds presented at any two distances among 0.50 m, 0.75 m, 1.00 m, 1.25 m, 1.50 m, 1.75 m, 2.00 m, 2.25 m, 2.50 m, 2.75 m, 3.00 m, 3.25 m, 3.50 m, 3.75 m, 4.00 m. Subjects compare stimulus sounds presented through headphones. This experiment consists of two experiments, Experiment-A and Experiment-B. The stimulus sound presented in the first experiment-A is a waveform generated by convolving a dry sound source with a measured IR corresponds distance. This experiment allows us to confirm the spatial resolution ability of humans in distance perception under certain conditions. The stimulus sound presented in the second experiment-B is a waveform generated by convolving a dry soud source with the IR whose DRR is controlled using the proposed method. Experiments-A and -B are both experiments in which distance differences are perceived, and if these results are similar, it can be shown that DRR-controlled IR is effective as a control method for distance perception. The similarity of these experimental results can demonstrate the effectiveness of our proposed method based on DRR control. The results of Experiment-A with five subjects confirmed that when there is a difference of 0.75 m to 1.0 m or more between two sound sources, people tend to be able to distinguish between near and far. The results of experiments-A and -B are similar, suggesting that the proposed method's DRR-controlled IR is effective in controlling the sense of distance.

III. Distance perception by Measured IR (Experiment-A)

In this section, we describe Experiment-A using actually measured IRs. First, we describe the measurement conditions for the real IR, and then we explain how to configure the stimuli presented in the TAFC test. Finally, based on the TAFC results, we show that when the two sound sources differ by more than 0.75 m to 1.0 m, humans tend to be able to discriminate between near and far when using actually measured IR.

A. Measuring IR

IRs are recorded in a university laboratory 15806 at Osaka Sangyo University (background noise level: $LA_{eq} = 44.6$ dB, reverberation time: $T_{60} = 0.4$ s). Fig.1 shows an overview of the laboratory 15806. In Fig.1, a recording point and sound source points are overlaied. We measure IR using multi-channel AD/DA RASP24 manufactured by System In Frontia[6]. The sampling rate and number of quantization bits are 48 kHz and 24 bit. IRs are measured using the Time Stretched Pulse (TSP) method[7]. The sample length of the TSP signal is 262144 samples (approximately 5.46 s). It is applied measured 16 TSPs to process time-synchronous overlapped-addition to improve signal-to-noise ratio. Fig.2 shows the IR measured under the condition that the distance between the sound source and the microphone is 1.0 m. The horizontal axis represents time, and the vertical axis represents amplitude. The time origin is the time at which the TSP signal starts playing.

B. Stimulus for confirming distance resolution

We present two stimuli (stimulus pairs) to subjects and confirm the resolution of distance perception by asking them about their sense of perspective. We force subjects to choose between two stimuli the one their auditory sense is closer to them (Two-alternative forced-choice (TAFC) test). In the following sections, the first stimulus sound of a stimulus pair is called the first sound, and the second stimulus sound is called the second sound. The first and second sounds are selected from 15 stimuli, corresponding to the distance between a listener and a sound source (Distance between a listener and sound souce: 0.50 m, 0.75 m, 1.00 m, 1.25 m, 1.50 m, 1.75 m, 2.00 m, 2.25 m, 2.50 m, 2.75 m, 3.00 m, 3.25 m, 3.50 m, 3.75 m, 4.00 m). The total number of combinations is 225 pairs. These 225 pairs are presented to the subject in a randomized order.

The presented sounds are from Japanese speech corpus for large vocabulary continuous speech recognition research (JNAS)[8]. One presented sound consists of single sentence

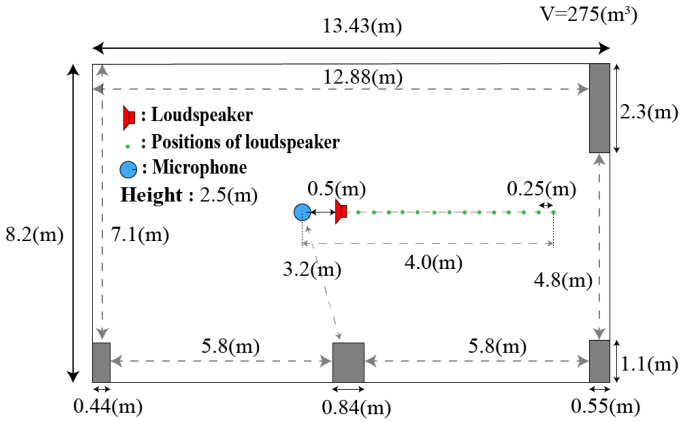


Fig. 1. Layout of the room where impulse responses were measured.

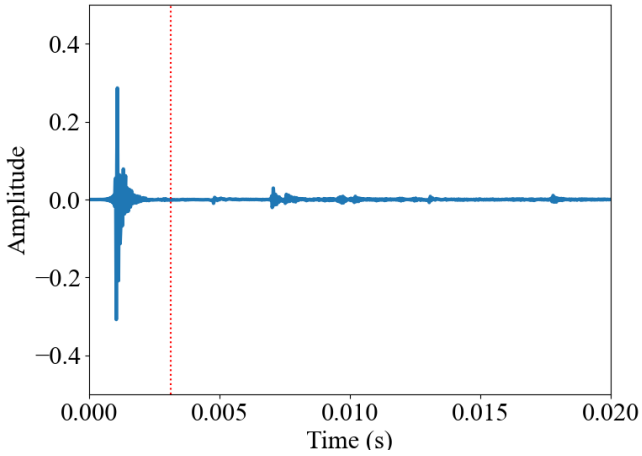


Fig. 2. Impulse response in the case of 1.0 m between sound source and observed points.

and it's length is about 5 s. The sampling rate is 16 kHz. The stimuli are generated from five sentences uttered by a male speaker and five sentences uttered by a female speaker.

For distance perception experiments, it is appropriate to present sounds with a high temporal resolution, that is, a high sampling rate, so we measure the IR at a sampling rate of 48 kHz. In our experiment, we upsample sounds in the corpus from 16 kHz to 48 kHz to match the audio and IR sampling rates. The upsampled signal $x_w(n)$ is regarded as a dry source sound, where $w, (w = 1, 2, 3, \cdots, 10)$ is sentence ID. The IDs $(w = 1, 2, 3, 4, 5)$ correspond to utterances by a male, and the IDs $(w = 6, 7, 8, 9, 10)$ correspond to utterances by a female. The sound obtained by convolving $y_(n)$ with the IR $h_d(n)$ at distance d presents as $y_{d,w}(n)$, where $d = 0.50$ m, 0.75 m, 1.00 m, 1.25 m, 1.50 m, 1.75 m, 2.00 m, 2.25 m, 2.50 m, 2.75 m, 3.00 m, 3.25 m, 3.50 m, 3.75 m, 4.00 m. According to the above notation, the stimulus sound $y_{d,w}(n)$ can be expressed as follows.

$$y_{d,w}(n) = x_w(n) * h_d(n) \tag{2}$$

A stimulus pair consists of the following sequence: At first,



Fig. 3. User interface of our application.

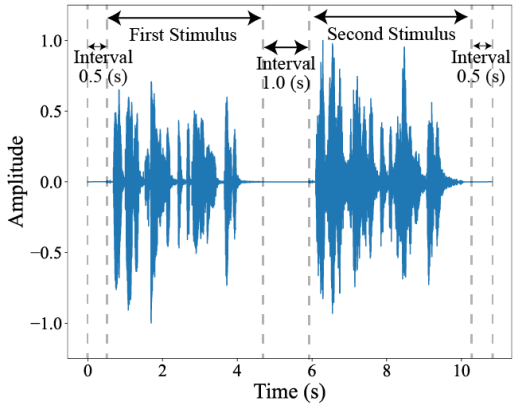


Fig. 4. Speech waveform presented in one trial.

there is a silence intarval of 0.5 s. Next, the first sound is presented. And then a silence intarval of 1.0 s is placed between the first and the second sounds. After that, the second sound is presented. Finally there is a silence intarval of 0.5 s. Fig.4 shows the configuration of stimulus pairs as audio waveforms.

C. About two-alternative forced-choice (TAFC) test

A two-alternative forced-choice test is a test in which the subject compares two presented stimuli and then selects one presented stimulus based on previously given criteria by a conductor of the experiment. The subject always has to choose one option, even if it is difficult to compare based on the selection criteria. It is called TAFC because the subjects are forced to choose one side. In this report, a subject has to choose the sound he senses closer after being presented with two sound stimuli.

We perform the TAFC test on 5 subjects (5 males). Two sessions are conducted for each subject, where one session consists of 225 stimulus pairs. Fig.3 shows the GUI of the experimental application. The Play button is a button to start playing the stimulus pair. Radio buttons A (choosing the first sound as closer) and B (choosing the second sound as closer) are buttons to select the closer stimulus. These stimulus selections become exclusively selectable when the stimulus pair is played at least once. The Enter button is used to confirm stimulus selection. The subject operates these buttons. Before confirming selection with the Enter button, a subject can replay stimulus pairs from the same trial by pressing the Play button again. If there is no obvious problem while

listening to the stimuli, the subjects are instructed not to listen to them again, and they are taught to make judgments based on their intuition based on the first listen in principle. The stimuli are AD converted and played back using a Roland Rubix-22 connected to a PC. They are presented through audio-technica ATH-AD500X headphones with the diotic method (a method in which the same stimulus sound is played to both ears).

D. Result of Two-alternative forced-choice (TAFC)

When five subjects take the TAFC test in two sessions each, we obtaine evaluation results for a total of 2250 stimulus pairs. Since there are 225 distance conditions, we obtain 10 subjective evaluation results for one distance condition. If the five subjects were able to completely discriminate the stimulus pairs, we would get a response that 10 out of 10 times they perceived the sound sources that were close to each other to be closer. In contrast, if they cannot discriminate at all, we will get answers that seem to be random variables since they respond uniformly at random. The expected result is that five subjects choose the closer sound source 5 out of 10 times. Through this experiment, we can determine the rate at which sound sources that are closer to each other are correctly perceived as being close.

When the first sound is at a distance of d_1 and the second sound is at a distance of d_2, the proportion of the subjects who correctly perceive d_1 as close under the condition $d_1 < d_2$, and the proportion of the subjects who correctly perceive d_2 as close under the condition of $d_1 > d_2$ are defined as *correctness*. The *correctness* is expressed as a heatmap on the $d_1 - d_2$ coordinate plane, where the correctness is uniformly random under the condition $d_1 = d_2$, so they are excluded from the heatmap. A heatmap obtained by averaging the responses of five subjects is shown in Fig.5. The horizontal axis represents the source distance d_1 of the first sound, and the vertical axis represents the source distance d_2 of the second sound. The *correctness* for each condition is expressed in gray scale and corresponds from 0 % to 100 % from white to black.

E. Discussion

The correctness tends to be unstable in the area where the distance difference between the first and second sounds is 0.25 m or 0.50 m. Judgments vary from 10 % to 100 %. Therefore, it can be confirmed that the perceived distance can be perceived as being reversed from the actual distance to the sound source.

Next, in the part where the distance difference between the first and second sounds is 0.75 m ($|d_1 - d_2| = 0.75$), the subject's response is correct to some extent, and the *correctness* is 30 % to 100 %. When the distance difference between the first and second sounds exceeds 1.0 m ($|d_1 - d_2| \geq 1.00$), the *correctness* is approximately 90 % or more when d_1 and d_2 are 2.0 m or less than 2.0 m. It can be seen that when the distance difference exceeds 1.0 m, it becomes easier to judge distance.

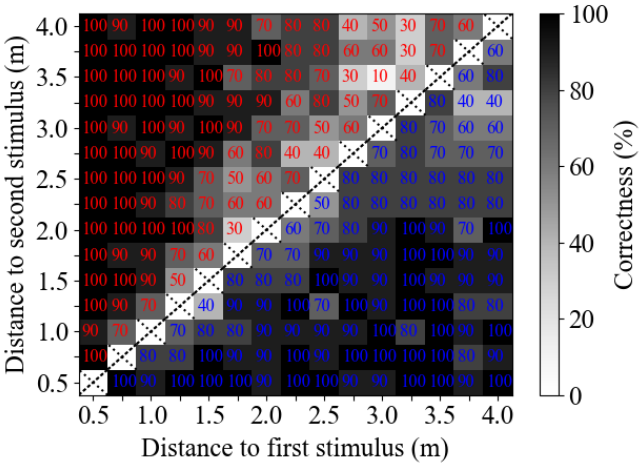


Fig. 5. The average *correctness* of distance perception for stimulus based on real IRs over the five subjects.

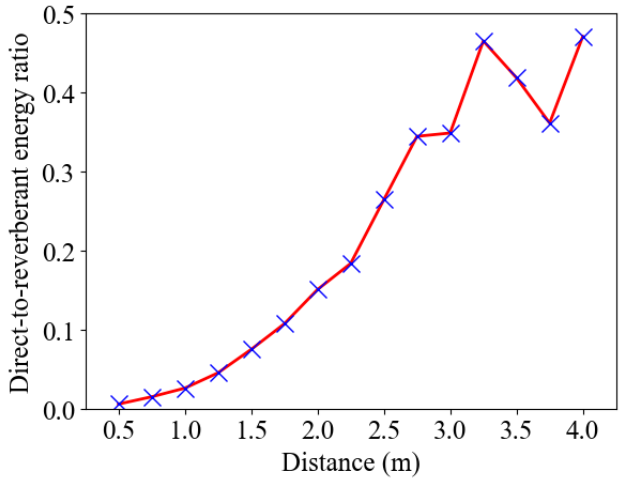


Fig. 6. DRR of some distances.

Finally, the overall trend is that when the distance d between the listening point and the sound source exceeds 2.0 m ($d \geq 2.0$), the resolution of distance perception tends to decrease. The reason that $d = 2.0$ is a kind of threshold is presumed to be due to the dimensions of the room. If the speculation is true, the distance over which high distance resolution can be maintained may increase depending on the room size when the room is large enough. Verification of this point is a future topic. Fig.6 shows the DRR of the measured IR. The horizontal axis is the IR distance d, and the vertical axis is the DRR. DRR increases monotonically up to $d = 2.75$ m. However, DRR does not increase monotonically in the range $d \geq 2.75$ m. It can be inferred that the DRR no longer increases monotonically due to the influence of multiple reflections from the walls due to the insufficient room size.

From the above, it can be predicted that the DRR of a room of sufficient size can be approximated by a monotonically

increasing function. Therefore, if there is a method to convert only the DRR from the IR of a certain distance to the desired distance, it is thought that distance perception can be controlled. In the next section, we propose such a method and discuss its details.

IV. Distance perception by DRR-controlled IR (Experiment-B)

It is not easy to record and collect IR from various distances in advance. In particular, if re-collection becomes necessary in response to changes in the environment, the burden is enormous. If it is possible to generate IRs at several distances from one IR, the burden of recording could be reduced. If it is possible to generate IRs of several distances from one IR, the burden of recording can be reduced. We conduct distance perception experiments by generating IRs from IR of 1.0 m. We compare the case of using generated IR and the case of using real IR.

A. Method of generating DRR-controlled IR (Proposed method)

We propose a method to convert IR measured at a certain distance d to IR at a different distance d'. The converted IR is converted so that it becomes the DRR of the real IR at the target distance. $h_d(n), n = 0, 1, \cdots, N-1$ is the measured IR waveform at distance d. The sample point that separates the first peak part corresponding to the direct wave component and the subsequent rear reverberant part in the IR waveform is expressed as q_d. Furthermore, the DRR of distance d can be expressed by the following equation.

$$D_{rr}(d) = \frac{\sum_{n=q_d}^{N-1} h_d^2(n)}{\sum_{n=0}^{q_d-1} h_d^2(n)} \tag{3}$$

where, N is sample length of the IR. In this way, DRR is determined by the energy ratio between the interval $n = 0, 1, \cdots, q_d - 1$ and the interval $n = q_d, \cdots, N-1$. In our experiment, $N = 32768$ samples (about 0.683 s at sampling rate 48 kHz).

Since we are focusing on the influence of DRR on distance judgment, we need to ensure that subjects do not use sound volume as a cue for sound distance judgment. Therefore, we perform an identity transformation on the direct wave of the IR ($h_d(n), n = 0, 1, \cdots, q_d - 1$), so that the loudness caused by the direct wave does not change significantly. Finally, the target of DRR transformation operation is limited to $n = q_d, \cdots, N-1$. We propose the simplest of these transformations. Specifically, the IR $h_d(n), n = q_d, \cdots, N-1$ is multiplied by a constant and transformed to the IR with the target IR's DRR. The formula for the proposed transformation from $h_d(n)$ to $\hat{h}_{d'}(n)$ is shown below.

$$\hat{h}_{d'}(n) = \begin{cases} h_d(n), & n = 0, 1, \cdots, q_d - 1 \\ g^* \cdot h_d(n), & n = q_d, \cdots, N-1 \end{cases} \tag{4}$$

The multiplyer g^* is positive value. It is formulated as follows:

$$g^* = \left(\frac{\sum_{n=q_{d'}}^{N-1} h_{d'}^2(n)}{\sum_{n=0}^{q_{d'}-1} h_{d'}^2(n)} \cdot \frac{\sum_{n=0}^{q_d-1} h_d^2(n)}{\sum_{n=q_d}^{N-1} h_d^2(n)} \right)^{\frac{1}{2}} \tag{5}$$

g^* is determined so that the DRR of the pre-transformed IR and the DRR of the post-transformed IR match the DRR of the reference IR, as shown below.

$$D_{rr}(\hat{h}_{d'}(n)) = D_{rr}(h_{d'}(n)) \tag{6}$$

$$\frac{\sum_{n=q_{d'}}^{N-1} \hat{h}_{d'}^2(n)}{\sum_{n=0}^{q_{d'}-1} \hat{h}_{d'}^2(n)} = D_{rr}(h_{d'}(n)) \tag{7}$$

Here we introduce the following approximation.

$$\frac{\sum_{n=q_{d'}}^{N-1} (g^* \cdot h_{d'}(n))^2}{\sum_{n=0}^{q_{d'}-1} h_{d'}^2(n)} \simeq \frac{\sum_{n=q_d}^{N-1} (g^* \cdot h_d(n))^2}{\sum_{n=0}^{q_d-1} h_d^2(n)} \tag{8}$$

By approximation, (7) can be transformed as follows.

$$\frac{\sum_{n=q_d}^{N-1} (g^* \cdot h_d(n))^2}{\sum_{n=0}^{q_d-1} h_d^2(n)} = D_{rr}(h_{d'}(n)) \tag{9}$$

$$\frac{\sum_{n=q_d}^{N-1} (g^* \cdot h_d(n))^2}{\sum_{n=0}^{q_d-1} h_d^2(n)} = \frac{\sum_{n=q_{d'}}^{N-1} h_{d'}^2(n)}{\sum_{n=0}^{q_{d'}-1} h_{d'}^2(n)} \tag{10}$$

$$(g^*)^2 \frac{\sum_{n=q_d}^{N-1} h_d^2(n)}{\sum_{n=0}^{q_d-1} h_d^2(n)} = \frac{\sum_{n=q_{d'}}^{N-1} h_{d'}^2(n)}{\sum_{n=0}^{q_{d'}-1} h_{d'}^2(n)} \tag{11}$$

$$(g^*)^2 = \frac{\sum_{n=q_{d'}}^{N-1} h_{d'}^2(n)}{\sum_{n=0}^{q_{d'}-1} h_{d'}^2(n)} \cdot \frac{\sum_{n=0}^{q_d-1} h_d^2(n)}{\sum_{n=q_d}^{N-1} h_d^2(n)} \tag{12}$$

Finally, we can obtain (5) by taking the square root of both sides of (12).

The real IRs and the IR transformed by the proposed method are shown in Fig7. In the top pannel of Fig.7, the targeted real IR $h_{2.00}(n)$ is shown with a gray line. This is reference IR. In the bottom, the seed IR $h_{1.00}(n)$ is shown with a red line. This is a seed IR. In the middle pannel, the transformed IR $\hat{h}_{2.00}$ from $h_{1.00}(n)$ to $h_{2.00}(n)$ is shown with a blue line. The horizontal axis is time and the vertical axis is amplitude.

As a result of the transformation, the amplitude of the late reverberation part of the seed's IR is amplified. It can be seen that the DRR of the transformed result in the middle row has increased and is approaching the reference DRR. It should be noted that the transformation increases or decreases the energy in the rear reverberation part. Normally, as the distance from the sound source to the listening point increases, the IR energy decreases, reflecting distance attenuation. In the proposed method, the IR energy slightly increases as the distance increases, but we believe that the change in energy is negligible. This is because the space we are dealing with here has little rear reverberation of IR.

B. Result of TAFC

In the previous section, we explained a control experiment for the experiment presented and the real IR was used as the stimulus sound. In this section, we describe an experiment similar to the previous section, using IR generated by the proposed method instead of real IR. The heatmap obtained from this experiment is shown in Fig.8.

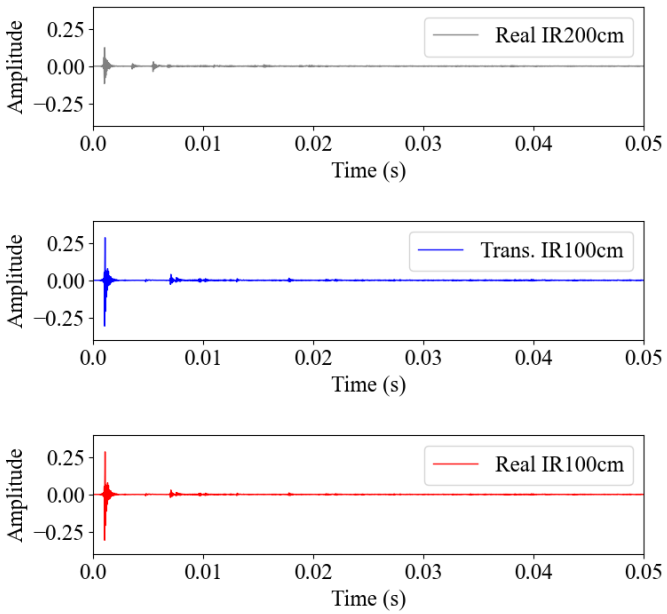


Fig. 7. Transforming a seed IR ($h_{1.00}(n)$) to a target IR ($\hat{h}_{2.00}(n)$).

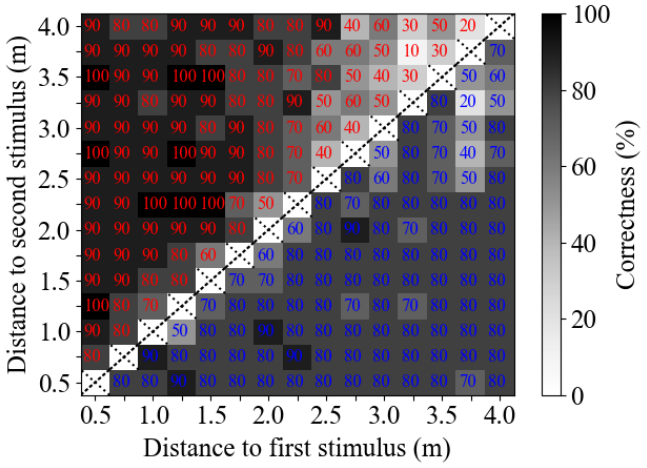


Fig. 8. The average *correctness* of distance perception for stimulus based on transformed IRs.

Comparing Fig.5 and Fig.8, we can confirm that the perceived distance follows a similar pattern. The value obtained by normalizing the cross-correlation between off-diagonal elements of the heatmap by the autocorrelation of the measured heatmap is 0.916, indicating high similarity. In particular, when limiting the heatmap under 2.00 m, the similarity is even higher at 0.928. The results suggest that there is no need to reproduce the IR in detail, and the sense of distance can be reproduced as long as the DRR is reproduced.

V. Conclusions

We proposed a method that uses a pre-existing IR for one distance as a seed and obtains IRs for other distances by transforming from the seed in this article. The key idea of the proposed method is approximating the DRR of the IR at the target distance. Subjective evaluation experiments suggested that the DRR is effective in IR transformation of perceptual distance.

The IR of 1.0 m is only used as the transformation seed in this experiment. Therefore, there remains a possibility that a similar trend to the measured IR strongly appears around 1.0 m. Additional evaluation is required regarding the behavior when the seed is set to an arbitrary distance, such as experiments where the seed is set to an IR of 2.0 m.

All experiments in this work is approved by ethics review committee for research involving human subjects of Osaka sangyo University (Number 2024-HR-16).

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP21H03488 and JP23K21691.

References

[1] T. Ekawa, M. Nakayama, and T. Takahashi, "Virtual sound source rendering based on distance control to penetrate listeners using surround parametric-array and electrodynamic loudspeakers," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1008–1015.

[2] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *Journal of the Acoustical Society of America*, no. 112, pp. 2110–2117, 2002. DOI: doi.org/10.1121/1.1506692.

[3] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using d/r spatial correlation matrix model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, 2011. DOI: 10.1109/TASL.2011.2134091.

[4] D. H. Mershon and J. N. Bowers, "Absolute and relative cues for the auditory perception of egocentric distance," *Perception*, vol. 8, no. 3, pp. 311–322, 1979.

[5] D. H. Mershon and L. E. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychopysics*, vol. 18, no. 6, pp. 409–415, 1975.

[6] *System in frontier inc.* https://https://www.sifi.co.jp/en/.

[7] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *Journal of Acoustic Society of America*, vol. 97, no. 2, pp. 1119–1123, 1995. DOI: doi.org/10.1121/1.412224.

[8] K. Itou, M. Yamamoto, K. Takeda, *et al.*, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999. DOI: 10.1250/ast.20.199.

New Abnormal Behavior Detection for Patient Surveillance System

Yujin Han and Taewan Kim
DataScience Major of Dongduk Women's University, Seoul, South Korea
E-mail: hanyujinius@gmail.com, kimtwan21@dongduk.ac.kr

***Abstract*—In contemporary healthcare settings, nursing staff encounter significant workload challenges due to high patient-to-nurse ratios, often ranging from 16 to 25 patients per nurse. This issue is particularly acute in psychiatric wards, where patients exhibit unpredictable behaviors. Continuous monitoring is essential for the care of these mental health patients, yet chronic staff shortages hinder the fulfillment of these needs. To address these challenges, this study proposes a new abnormal behavior detection for patient surveillance system, utilizing real-time video surveillance to identify abnormal behaviors in psychiatric wards. The system's accuracy has been enhanced through well-defined characteristic behaviors and a rule-based detection algorithm. Built on the *DeepStream* SDK, the system exhibits excellent performance, processing each frame in an average of 0.01 seconds and achieving a recognition accuracy exceeding 90%. This patient surveillance system significantly improves emergency response capabilities, mitigates potential risks, and enhances the work environment for medical personnel by providing accurate and timely patient status information. Ultimately, this approach is anticipated to contribute not only to improved health outcomes for patients but also to enhanced job satisfaction and safety for healthcare professionals.**

I. Introduction

The global prevalence of mental illness is rising due to various factors such as increasing social stress, environmental influences and drug-related issues. Patients with severe mental illnesses often require admission to closed psychiatric wards for continuous treatment and observation. Although the number of patients needing inpatient care steadily increases annually, there is a significant shortage of healthcare personnel to accommodate them [1]. Despite having more nursing and security staff in psychiatric wards compared to general wards, numerous dangerous incidents regularly occur. Monitoring patients through closed-circuit television (CCTV) or regular rounds is crucial for reducing these incidents [2]. However, the intensive focus on surveillance can delay other critical tasks, thereby increasing the risk of additional incidents. Therefore, an efficient and thorough monitoring system is essential. Implementing a real-time monitoring system utilizing room CCTV for early detection of potential risks will reduce the mental fatigue of nurses and enhance surveillance efficiency.

This study proposes an abnormal behavior detection system to identify early risks. Quantitative definitions for specific abnormal behavior patterns such as *Self-harm*, *Escape Attempt* and *Falldown* frequently observed in closed psychiatric wards have been established. Each abnormal behavior is defined based on various biomechanical characteristics, including joint angles, relative positions of body parts, and temporal repetitiveness of movements. According to these definitions, the system processes streaming video data to detect abnormal behaviors of psychiatric patients and immediately alerts the ward nurses. The implementation leverages *NVIDIA DeepStream* [3], resulting in a reduction of average processing time from 2.2 seconds to 0.95 seconds on a standard desktop equipped with a GPU 2080, thereby achieving a time savings of over one second. This system incorporates several significant enhancements:

- Definition of specific abnormal behaviors (*Self-harm*, *Escape Attempt* and *Falldown*) tailored to the unique hospital environment.
- Development of a human skeleton-based behavior detection algorithm designed to effectively identify the defined abnormal behaviors.
- Deployment of a high-performance real-time abnormal behavior detection system utilizing *DeepStream* technology.

II. Diagnostic Criteria for Abnormal Behaviors

A. Definition of Skeleton

The definition of the skeleton used in this system focuses on the positioning of the arms and legs, which are crucial for abnormal behavior detection. Specifically, this system utilizes several key points on the body: Shoulder (S), Elbow (E), Wrist (W), Hip (H), Knee (N) and Ankle (A). Additionally, it considers combinations such as Shoulder-Elbow-Wrist (SEW), Hip-Knee-Ankle (HNA) and Hip-Shoulder-Elbow (HSE). In this notation system, an angle is represented by the symbol θ followed by the abbreviations of the connected segments. For example, the angle formed by the right hip, right knee and right ankle is denoted as θ_{HNAR}. This method of skeleton mapping and angular notation forms the basis of the behavior detection methodology used in the patient surveillance system.

B. Self-harm

Considering the restrictions on bringing external objects into closed wards, this study focuses on two primary *Self-harm* behaviors using only the body. As summarized in Table I, 'Hit Head' involves one or both hands moving towards the head while the patient is not lying down and is identified based on arm angles and temporal changes in arm positions. 'Injure

TABLE I

CHARACTERISTICS OF SELF-HARM, ESCAPE ATTEMPT AND FALLDOWN. FOR SELF-HARM, $\square_c$ REFERS TO THE $\square$ IN THE CURRENT FRAME AND $\square_p$ REFERS TO THE $\square$ IN THE PREVIOUS FRAME, WITH W REPRESENTING THE Y POSITION OF THE WRIST IN IMAGE COORDINATES. FOR FALLDOWN, APY AND APX REPRESENT THE Y AND X POINTS OF ALL SKELETON POSITIONS IN IMAGE COORDINATES, RESPECTIVELY. CONSTANTS α AND β ARE SET TO 0.1*(HEIGHT OF IMAGE) AND 0.1*(WIDTH OF IMAGE).

	Self-harm		**Escape Attempt**	**Falldown**	
	Hit Head	**Injure Arms/Legs**	**Climb Window**	**Up→Down**	**Sideways**
Posture	Not Lie	Sit	Stand/Climb	Sit	Sit/Lie
Hands	Near Head	Near Arms/Legs	Reach/Grip Window Frame	Near floor	Near floor
Legs	Almost crouch	Almost crouch	One leg raised, one supporting	Slightly bent/straight	Slightly bent/straight
Position Algorithm	$W_c \leq W_p$	$W_c \geq W_p$	$\theta_{HSEL} \geq 70$ or $\theta_{HSER} \geq 70$	$APy_p + \alpha \leq APy_c$	$APx_c + \beta \leq APx_p$ or $APx_p + \beta \leq APx_c$
	$20 \leq \theta_{SEWL} \leq 90$ or $20 \leq \theta_{SEWR} \leq 90$	$\theta_{SEWL} \leq 180$ or $\theta_{SEWR} \leq 180$			
	$\theta_{SEWLc} \leq \theta_{SEWLp}$ or $\theta_{SEWRc} \leq \theta_{SEWRp}$		$\theta_{HNAL} \leq 30$ or $\theta_{HNAR} \leq 30$		

Arms/Legs' typically occurs in a sitting posture and is detected when hands are positioned near the arms or legs. This behavior is characterized by specific arm angles and changes in wrist positions, but it is more complex to detect due to the greater variability in angles and positions.

C. Escape Attempt

Given the high risk of injury or escape in psychiatric closed wards, this study focuses on 'Climb Window' as the most dangerous escape attempt behavior. As summarized in Table I, this behavior involves the patient in a standing or climbing posture with hands reaching towards or gripping the window frame. It is detected through characteristic postural changes, such as significant changes in shoulder angles and the elevation of one leg.

D. Falldown

Given the high risk of injury and the need for immediate response in psychiatric wards, this study categorizes fall incidents into two primary types. As summarized in Table I, 'Up→Down' involves a vertical fall from a sitting position, identified by an increase in all y-coordinates of the estimated skeleton in the image frame. 'Sideways' typically occurs from either a sitting or lying posture, detected by significant changes in the x-coordinates of the skeleton. This behavior is characterized by specific changes in overall body position but can be more complex to detect due to variations in fall trajectories.

III. ABNORMAL BEHAVIOR DETECTION

The system architecture for detecting abnormal behaviors in closed psychiatric wards is illustrated in Fig. 1. This framework is engineered to process real-time video streams, facilitating rapid and precise identification of potential risks. It employs the *DeepStream* SDK for efficient video processing and integrates a rule-based algorithm for behavior detection, thus enhancing its capacity to handle unforeseen scenarios.

A. Human Detection

The first step of the system is patient location detection. This process involves identifying individuals' positions in each frame from the streaming CCTV footage within hospital rooms and common areas. YOLOv4 [4], a widely adopted object detection model, is utilized for this task. Along with identification, tracking is implemented to enable temporal analysis of individual actions. Considering the unique environment of a closed psychiatric ward, the proposed approach prioritizes the accuracy of patient-specific recognition over general human detection. To enhance the overall capability, the object detection threshold was set to 0.4, lower than typical settings, allowing for the capture of more potential patient locations.

B. Bounding Box Expansion

When object detection fails, the system applies expanded coordinates of the bounding box(bbox) from the previous frame to the current frame. The expansion ratio is set to 20%, considering the characteristics of psychiatric ward patients who may exhibit the previously defined abnormal behaviors. This ratio is sufficient to capture sudden movements associated with these behaviors while minimizing unnecessary background information.

This expansion enables continuous tracking even when object detection fails, maintaining stable skeleton estimation. It compensates for temporary detection failures without complex algorithms, suitable for real-time processing. This enhances object detection robustness, enabling continuous skeleton estimation and improving rule-based behavior detection accuracy and efficiency.

C. Skeleton Estimation

The system uses bbox coordinates obtained from human detection or bbox expansion to estimate skeletons. For this task, HRNet [5], [6], which achieved state-of-the-art results in human pose estimation in 2019, is employed. The architecture of HRNet, featuring parallel subnets of various resolutions and continuous exchange of global and local information, enables accurate skeleton estimation across different bbox sizes. HRNet not only determines joint positions but also calculates their reliability. This reliability information serves as a valuable input for the subsequent rule-based behavior detection, enhancing the system's ability to accurately identify abnormal behaviors in psychiatric ward patients.

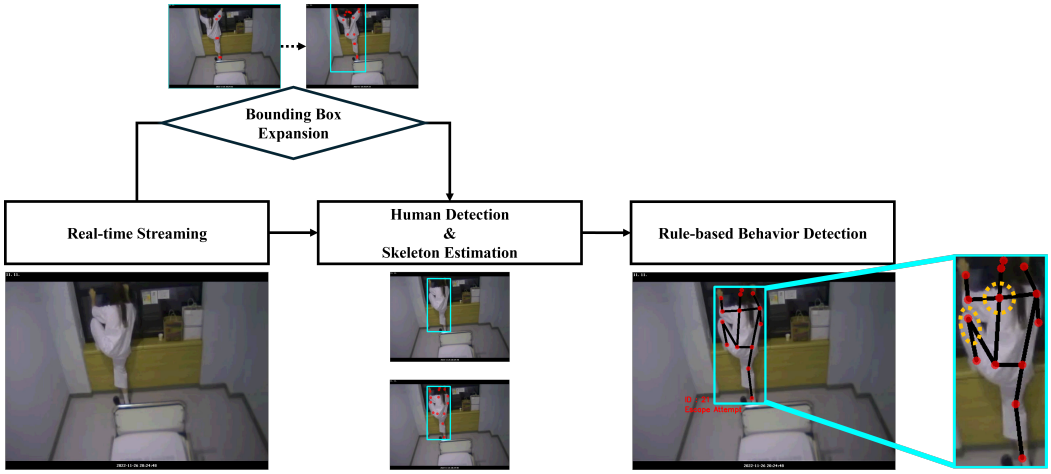


Fig. 1. The overall framework of the proposed abnormal behavior detection system.

D. Rule-based Behavior Detection

Separate recognition functions are established for each type of abnormal behavior (*Self-harm*, *Escape Attempt* and *Falldown*). Abnormal behaviors are detected based on various conditions such as posture and joint angles. Two common conditions applied before individual behavior recognition are:

- At least 10 out of 17 joints of the estimated skeleton must have a reliability of 0.8 or higher.
- For calculating joint angles such as θ_{SEW}, θ_{HSE} or θ_{HNA}, the reliability of all joints involved in the calculation must be 0.85 or higher.

The primary objective of the proposed system is to minimize false positives. Filtering out false detections based on the reliability of the estimated skeleton is crucial. Occlusions can lead to inaccuracies in skeleton predictions and reduce joint reliability. To address this, a threshold was set requiring at least 10 out of 17 joints (58.82%) to have high reliability. Joint reliability is particularly important for capturing detailed postures of arms or legs. Therefore, for recognizing abnormal behaviors, especially those involving joint angles, only joints with reliability exceeding 0.85 were used in angle calculations.

E. Abnormal Behavior Alarm

The rule-based behavior detection system is designed to trigger an alarm when dangerous behaviors are detected multiple times. While minimizing false negatives is crucial in closed wards, the possibility of false alarms must also be considered. The system adjusts the alarm triggering threshold according to the risk level associated with each behavior type. For *Falldown*, due to the high risk of immediate injury, the alarm is triggered with a relatively low number of detections. Self-harm, while potentially less dangerous in the short term, has an increasing risk with repetition, which is reflected in the threshold setting. *Escape Attempt* carries a high risk and thus has a low threshold applied for rapid response. For instance, behaviors such as window climbing are designed to trigger alarms quickly due to the need for immediate intervention.

This alarm threshold setting is tailored to the characteristics and potential risks of each behavior type. It aims to enhance the system's efficiency and accuracy in detecting abnormal behaviors. This approach balances the need for timely alerts while minimizing false alarms.

IV. Experiments

A. Dataset

To assess the performance of the proposed system, we utilized a dataset sourced from CCTV footage in actual closed wards at Inha University Hospital. This dataset includes a total of 11 videos: 5 videos each for the two primary behavior types (*Self-harm* and *Falldown*) and 1 video for the *Escape Attempt*. The video lengths vary from 1 to 3 minutes. Unlike standard behavior detection datasets, behavior labels were annotated for each frame. The labels were provided by professional nurses with specialized medical expertise from Inha University Hospital.

B. Performance of Abnormal Behavior Detection

The performance of the proposed abnormal behavior detection system was assessed by comparing the obtained abnormal behavior labels with the ground truth labels for each frame in a total of 11 videos. The confusion matrices for four behavior types: *No Action*, *Self-harm*, *Escape Attempt* and *Falldown* are presented in Fig. 2. In conventional behavior detection datasets, short videos have one behavior label each. However, for the abnormal behavior dataset constructed for this research, multiple behavior labels are present in a single long video. Hence, measuring detection accuracy for the *No Action* category is also essential.

The average abnormal behavior detection accuracy was measured at 90.8%. Particularly, the probability of correctly classifying *No Action* when it is indeed *No Action* was notably high at 97.3%. Detection accuracy for *No Action* is not

		Ground Truth Label			
		No Action	**Self-harm**	**Escape Attempt**	**Falldown**
Estimated Results	**No Action**	0.973	0.038	0.021	0.159
	Self-harm	0.007	0.905	0.017	0.002
	Escape Attempt	0.005	0.021	0.915	0.029
	Falldown	0.004	0.002	0.065	0.839

Fig. 2. Performance of the proposed method.

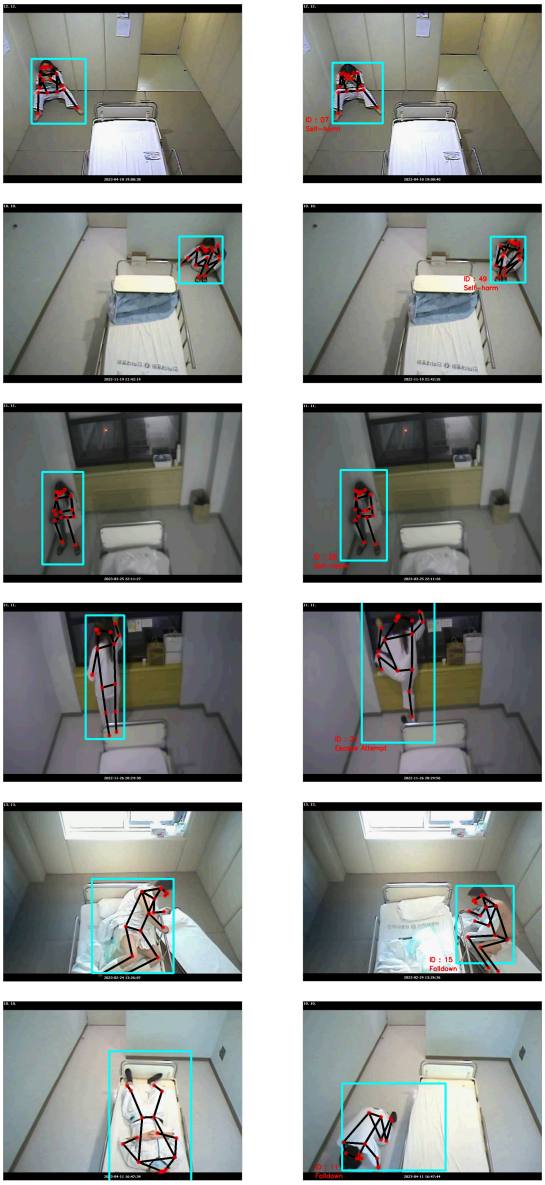
highly significant in conventional behavior detection algorithms. However, in the proposed system, it is crucial for the efficient handling of medical staff. Achieving high detection accuracy for *No Action* is integral to the system's performance, as it ensures accurate alarms for abnormal behaviors, allowing medical staff to respond appropriately to the situation. Self-harm also exhibited high accuracy at 90.5%. *Escape Attempt* showed a high accuracy at 91.5%. Analyzing the results, it was found that the system sometimes misinterpreted the descent after a patient climbs a window as a fall, due to the similar body movements involved. Despite this, the accuracy for *Escape Attempt* remained high due to the distinct and prolonged nature of the behavior, characterized by noticeable changes in leg and shoulder angles. Falldown showed slightly lower accuracy at 83.9% compared to other abnormal behavior types. This was often due to very short prelude symptoms of *Falldown*, which led to the system missing the action in the first 15-20 frames. These instances were mostly categorized as *No Action* and were not classified as other abnormal behaviors. Examples of the detected behaviors are illustrated in Fig. 3, showcasing the effectiveness of the system in identifying abnormal behaviors.

C. False Positive Rates of Abnormal Behavior Detection

In Section *B*, we analyzed the recognition accuracy for each type of abnormal behavior. The primary objective of the proposed system is to reduce the probability of false alarms while increasing the accuracy of abnormal behavior detection. To demonstrate the system's performance, we evaluate precision and recall values. Precision indicates the proportion of true positives among the instances classified as positive by the model, while recall represents the proportion of actual positives correctly identified by the model. The proposed system achieved a precision of 0.909 and a recall of 0.914, both indicating high performance. Additionally, the false positive rate was 1.7%. This suggests that the system effectively minimizes unnecessary alerts to medical staff while accurately signaling when a patient requires urgent attention.

V. Conclusions

In this study, new abnormal behavior detection for patient surveillance system was developed. This system defines frequent behaviors in closed psychiatric wards and uses rule-based algorithms to recognize abnormal behaviors. Implementation through *NVIDIA DeepStream* optimizes real-time processing capabilities, allowing for the quick detection of potential risks during patient surveillance in hospital settings. Despite the high variability in skeletal movements, the abnormal behavior detection system demonstrates an average recognition accuracy of 90.8%. Building on this, the aim is to further optimize the patient surveillance mechanism for identifying high-risk behaviors, with the goal of contributing to enhanced patient care and safety in medical facilities.



(a) (b)

Fig. 3. (a) Normal skeleton estimation. (b) Successful detection of defined abnormal behaviors (*Self-harm*, *Escape Attempt* and *Falldown*).

Acknowledgment

This work was supported by the Grant of the Korea Health Technology Research and Development Project through the

Korea Health Industry Development Institute (KHIDI), Ministry of Health and Welfare, Republic of Korea, under Grant HI22C1884.

References

[1] W. H. Organization *et al.*, "Mental health atlas 2020: Review of the eastern mediterranean region," 2022.

[2] H. Song, J. Kang, and T. Kim, "Continual learning based personalized abnormal behavior recognition alarm system," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 1, 2024.

[3] "Deepstream sdk — nvidia developer." (2023), [Online]. Available: https://developer.nvidia.com/deepstream-sdk.

[4] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[5] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[6] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *European Conference on Computer Vision (ECCV)*, 2018.

Utilizing Cross Layer Attentions for Semantic Segmentation of Small Objects

Chi-Hsuan Lu, Yu-Hsien Chung, Jung-Hui Cho and Chih-Chang Yu
Chung Yuan Christian University, Taoyuan, Taiwan
E-mail: ccyu@cycu.edu.tw Tel: +886-2659999

Abstract— **Segmenting small objects poses significant challenges in semantic segmentation tasks. This is because the downsampling process employed by many current deep learning networks tends to suppress the features of small objects, making them difficult to be retained in the encoded latent space. To mitigate this issue, we propose the Cross Layer Attention (CLA) encoder/decoder block, which aims to correlate informative features between deep feature layers. CLA calculates the correlation across different feature layers in the encoder and then subsequently integrating this information is integrated into the Feed-Forward Network (FFN) decoder, thus forming a dual decoding process. The designed CLA is able to preserve the features of objects at different positions, so that it can be used to enhance the feature of a small object during the upsampling process. The proposed CLA is experimented with the ADE20K public dataset using the Segformer – B0 and B1 as the backbone network, and the results are imrpoved by 1.37% and 2.38% in mean Intersection over Union (mIoU) compared to the original SegFormer – B1 and B0 model respectively. Furthermore, for objects which often appear smaller than 32×32 pixels in the validation set, the mIoU is also improved by 3.31% and 7.02%. The experiments demonstrates that the proposed CLA enhances the overall semantic segmentation results, particularly for small objects.**

I. INTRODUCTION

In an era of rapid technological advancement, computer vision has gained significant attention across various fields, demonstrating its applications in numerous areas. Generally, Convolutional Neural Networks (CNN) have been widely used in computer vision applications [1]. Most Convolutional Neural Networks (CNNs) utilize convolutional kernels with restricted sizes, thereby the networks have strong capabilities for analyzing local information within a scope. However, this characteristic also poses a limitation, obstructing CNNs from capturing long-range dependencies, such as the correlation between trucks in a distance and nearby cars, which leads to suboptimal performance, including imprecise segmentation outcomes. Although the receptive field can be expanded through the downsampling of feature maps, this process often results in the suppression or even loss of features pertinent to small objects. Consequently, while CNNs exhibit strong performance on larger objects, their efficacy diminishes significantly when dealing with smaller objects. To capture multi-scale features effectively, Lin et al. [2] introduced the Feature Pyramid Network (FPN). FPN uses multi-scale features to enhance the detection of small objects. However, the most models employing FPN remain CNN-based, thereby lacking the capability to capture long-range and cross layer dependencies.

Recently, the transformer architecture [3] has been adopted in computer vision. Unlike CNNs, which use limited-size convolutional kernels, transformers divide an image into several patches and compute self-attentions among patches. The self-attention mechanism allows each patch to attend to all other patches, eliminating information discontinuities and capturing long-range dependencies with global information, thus reducing information loss. However, while transformers capture long-range dependencies, the features of small objects tend to be sparse, and the computation of self-attention requires significant computational resources.

Addressing the difficulty of small object detection requires large models for training. These models often need expensive hardware, posing challenges for practical applications. Therefore, this study aims to improve small object segmentation within the constraints of affordable hardware (e.g., a consumer GPU). The proposed method utilizes the cross-layer attentions (CLA), which are computed in the encoder and store the correlation between different feature maps. In the decoder, CLA is then used to enhance segmentation performance.

The contributions of this study are as follows:

1. A novel CLA encoder/decoder block is introduced. The encoder blocks obtain correlations between different feature maps, while the decoder blocks use the encoder outputs to create upsampling branches, compensating for features lost during encoding process.
2. CLA requires only a small increase in parameters but enhances small object segmentation capabilities, particularly for lightweight models.
3. CLA can be applied to any model with a symmetric encoder-decoder structure.

II. Related Work

In this section, we will briefly describe the backbone network used in this study.

A. Small object segmentation

The segmentation of small objects is challenging due to their small size, which makes their features less prominent. In recent solutions for small object detection, multi-scale feature learning and Generative Adversarial Networks (GAN)-based detectors are widely adopted. For instance, Li et al. [4] proposed SNIP, which employs multiple feature maps at different scales. It filters out unsuitable targets based on the actual size of the objects in the current scale, thereby avoiding the search for small objects in large-scale feature maps. Additionally, Singh and Davis [5] introduced TridentNet, which divides the network into three parallel branches and uses dilated convolutions with different rates to generate feature maps. These feature maps are then merged to generate the final output. The use of parallel branches allows the model to have feature maps with varying receptive field sizes, thereby effectively identify objects of different scales. Although these methods effectively integrate multi-scale feature maps, they still face the challenge that the failure in the module responsible for detecting small objects results in the miss of small objects in the final results.

As GAN-based Detectors, Rabbi et al. [6] proposed a resolution-enhancing model. By enhancing the resolution, more features are generated to improve the recognition rate of small objects. Similar to the previous method, this approach also requires a significant amount of computational resources and large datasets to train the model.

B. Across Feature Map Attention

The Across Feature Map Attention (AFMA) method proposed by Sang et al. [7] is an attachable model that does not require additional training. Furthermore, this method can be applied to most models, such as U-Net [8], DeepLabV3 [9], FPN, and others. The main steps of AFMA are to find the correlation between the original image and a certain feature map in regard to the number of classes. AFMA processes the original image alongside a specific feature map to acquire the attention related to an object between the image and feature map. Subsequently, the decoder output is multiplied by these attention maps for enhancing the features of small objects.

C. SegFormer

The self-attention mechanism of the Transformer model allows the model to obtain image features more globally. SegFormer [10] is a transformer-based framework primarily used for semantic segmentation. Unlike the Vision Transformer (ViT) [11] [12], which uses position encoding or conditional position encoding to obtain positional information, SegFormer considers position encoding redundant for semantic segmentation. Therefore, SegFormer uses Mix-FFN, which directly extracts and utilizes positional information through 3×3 convolution layers and a multilayer MLP. In addition, to reduce the computation cost of transformer models, SegFormer proposed an efficient self-attention module. First it reshapes the input head tensor K from N×C to N/R×(R×C), where N is the size of the image and R is the reduction ratio. Then, a linear layer is used to change its dimension from R×C to C. This process can significantly reduce the computation cost of self-attention. With the use of efficient self-attention, Mix-FFN, and overlap patch merging, SegFormer distinguishes itself from other Transformer models and enhance both accuracy and efficiency.

Upon deploying SegFormer on several public datasets, it was observed that SegFormer still has some room for improving its performance in segmenting small objects. Although AFMA can address this problem, however, AFMA is restricted to be applied to the decoder output. This constraint requires the number of channels in AFMA to match the number of classes, which causes some information loss. Therefore, inspired by AFMA, this study proposed CLA as a method to address this limitation, which will be introduced in the next section.

III. Proposed Method

The proposed network with the CLA module is illustrated in Fig. 1. First, the feature maps generated by the SegFormer encoder are processed using the CLA encoder to capture the correlation across different layers. The decoder block accepts three inputs that contains a concatenated feature map, the CLA and another feature map. The final segmentation results are obtained through another multi-layer perceptron (MLP).

The key difference between the CLA encoder and the AFMA encoder is that AFMA only calculates the correlation of objects of the same class between the original image and the feature maps in a specific layer. In contrast, our model can compute the correlations between any two layers. Therefore, the model can have multiple CLA encoders to preserve the features of small objects to the greatest extent. Another benefit is that we can apply the results of multiple CLA encoders during the decoding process, ensuring that the model can compensate for the lost small object features during upscaling.

A. Encoder

The proposed CLA encoder is designed as a plug-in for the original network encoder. The process is shown in (1). First, a higher-resolution feature map, denoted as F_i, with size $H_i \times W_i \times C_i$ is passed through a 1×1 convolution to change the channel of the feature map to align with the channel size of a lower-resolution feature map, F_j, whose size is $H_j \times W_j \times C_j$. H_i is larger than H_j and W_i is larger than W_j. After this process, we can obtain another feature map whose size is $H_i \times W_i \times C_j$. Next,

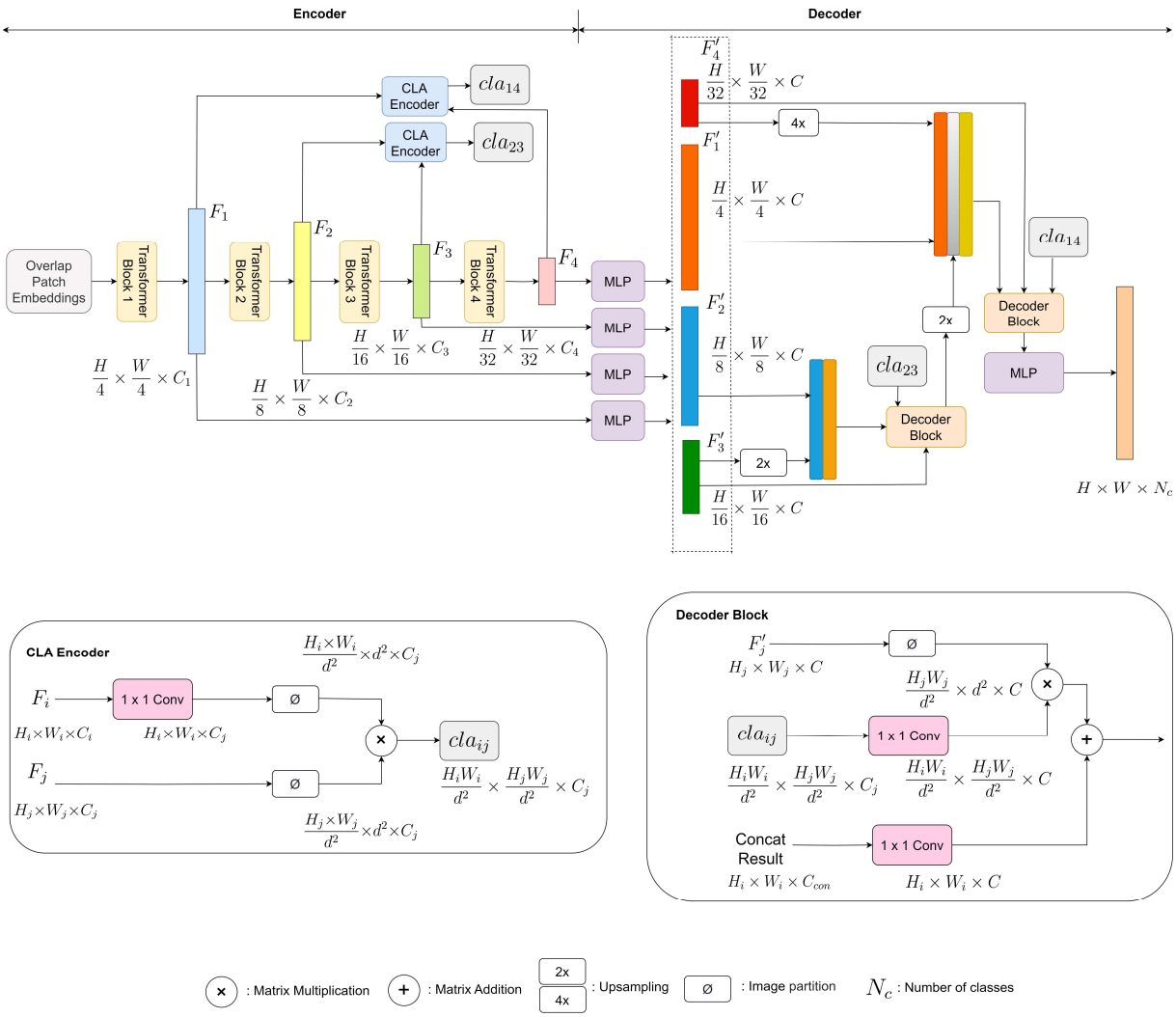


Fig. 1. The model structure proposed in this study. The cross attentions between two feature maps are computed through two CLA encoders. The output of corresponding encoders is then used in the decoder to compensate the feature of small objects.

both feature maps with channel size C_j undergo image patch partitioning (ϕ) with size d^2, changing their sizes to $(H_i \times W_i/d^2) \times d^2 \times C_j$ and $(H_j \times W_j/d^2) \times d^2 \times C_j$, respectively. Finally, we multiply and concatenate them to obtain the cla$_{ij}$ with size $(H_j \times W_j/d^2) \times (H_j \times W_j/d^2) \times C_j$.

$$cla_{ij} = \phi(\mathrm{Conv}_{1\times1}(F_i), d^2) \times \phi(F_j, d^2) \qquad (1)$$

In the original SegFormer, feature extraction is first performed through the transformer blocks, obtaining four different sizes of feature maps. We attach a CLA encoder between two transformer blocks to calculate the attentions. This study uses two different CLA encoders, cla_{14} and cla_{23}, which correspond to the use of F_1 with F_4 and F_2 with F_3 respectively.

B. Decoder Block

A decoder Block consists of a CLA decoder and a 1×1 convolution. To accommodate the cla_{ij} computation, cla_{ij} is first passed through a 1×1 convolution for matching the number of channels of the feature map, then the feature map with size $H_j \times W_j \times C$ also undergoes image patch partitioning (ϕ) with size d^2, changing its shape to $(H_j \times W_j/d^2) \times d^2 \times C$. Then it is multiplied by the cla_{ij} acquired from the encoder so that the shape of the CLA decoder output O_{ij} goes back to $(H_i \times W_i/d^2) \times d^2 \times C$, which contains the influence of large objects on small objects. The CLA Decoder operation is described in (2).

$$O_{ij} = \phi(F'_j, d^2) \times \mathrm{Conv}_{1\times1}(cla_{ij}) \qquad (2)$$

where O_{kl} is the CLA decoder output from other layers (if exists). Unlike the original SegFormer, we devise this decoder block which leverages the output of the CLA Decoder. First,

given two feature maps F_i and F_j, both through an MLP, the result is denoted as F_i' and F_j'. After that F_j' goes through two paths, one is being upscaled to the size of F_i' and concatenated with F_i'. The output O_{ij} of the previous decoder is also been concatenated with F_i' and F_j' if there exists one. If F_j' is the smallest feature map in the decoder, that block only takes the concatenated result of F_i' and F_j'. For example, in Fig. 1, the first concatenation is F_2' with F_3', then a decoder block is applied. After that, both the result and F_4' are upsampled to the same size as F_i' and concatenated with F_1' , then another decoder block is applied. The concatenated result goes through a 1×1 convolution in order to align the channel size with the CLA decoder output. Finally, an MLP is used to fuse all the features to obtain the final segmentation result.

IV. Experimental Results

The experiments in this study were conducted on a Intel core-i9 CPU and an NVIDIA GeForce RTX 4090 GPU using the ADE20K [13] dataset. The ADE20K dataset contains 20,210 images with 150 categories in various daily life scenes, including objects such as buildings, vehicles, furniture, and backgrounds such as sky, roads, and ground. We use SegFormer – B1 and B0 as the network backbone and trained the model for 320k iterations with a batch size of 8.

To demonstrate the performance of the proposed CLA, this study compares the proposed method with the SegFormer. We use two CLA encoders, denoted as cla_{14} and cla_{23}. Since the focus of this study is small objects, we also compute the mIoU of small objects, denoted as mIoU@small. We define objects that appear more than 60% of the time in the validation set with a size of less than 32×32 pixels as small objects. As a result, there are 17,134 images (77% of the entire dataset) that contain 12 categories of objects. Table 1 to 3 tabulates the performance with and without the proposed CLA module. From Table 1, we can see that the SegFormer – B1 with CLA module performs better than that without CLA module both in mIoU@small and mIoU. However, the performance is not as good as SegFormer – B2. This is because our model has almost half the number of parameters compared to SegFormer B2, the performance degradation is not surprised. Comparded to SegFormer – B1, with the CLA module, the number of parameters is only increased about 1%, indicating that our proposed technique imposes almost no burden on the model. For overall mIoU, our model improves 1.37% compared to SegFormer – B1. Regarding small objects, the mIoU@small is improved by 3.31%. We also applied our method on Segformer – B0, and the integration of the CLA module resulted in a 2.38% improvement in mIoU and a 7.02% imrpovement in mIoU@small, while only a 5.04% increase in the number of parameters.

Table 2 lists some of the categories improve the most from small objects in the experimental dataset. Compared to SegFromer – B1, the SegFormer – B1 with CLA model improves the mIoU for small objects such as clock, light, ashcan and glass with 16.42%, 4.86%, 4.36%, and 4.09% respectively. Table 3 lists four of the categories improve the most from non-small objects: ship, microwave, barrel and booth. The improvements in terms of mIoU are 41.80%, 24.93%, 18.06%, and 14.69%, respectively. Moreover, Table 2 and Table 3 indicate that the performance of SegFormer – B0 with CLA is still better than SegFormer – B0. These results demonstrate that our method not only improves the performance for non-small objects but also significantly enhances the performance for small objects.

We provide some visual comparisons of small objects (Fig. 2 (a) to (d)) and non-small objects (Fig. 3 (a) to (d)) between SegFormer – B1 and SegFormer – B1 with CLA. Considering small objects, as illustrated in Fig.2 (a), it is evident that the original SegFormer does not perform well in detecting distant skiers dressed in white, resulting in a smaller segmented region compared to the proposed model. This discrepancy is even more obvious in Fig.2 (b), where the number of distant small boats segmented by the original SegFormer is significantly

Table 1 Performance comparisons of models in different scales and the propsosed method

models	Parameter	mIoU@small	mIoU
SegFormer – B0	3.77M	22.64	36.79
SegFormer – B1	13.7M	31.63	41.75
SegFormer – B2	27.5M	35.69	46.33
SegFormer – B0 + CLA	3.96M	29.66	39.17
SegFormer – B1 + CLA	13.8M	34.94	43.12

Table 2 Performance comparison of models with CLA versus those without CLA for small objects.

models	clock	light	ashcan	glass
SegFormer – B0	14.42	35.02	18.98	3.39
SegFormer – B1	16.33	43.26	29.78	5.72
SegFormer – B0 + CLA	24.99	41.23	22.60	5.36
SegFormer – B1 + CLA	32.75	48.12	34.14	9.81

Table 3 Performance comparison of models with CLA versus those without CLA for non-small objects

models	ship	microwave	barrel	booth
SegFormer – B0	54.31	24.91	28.40	15.48
SegFormer – B1	24.51	32.76	27.39	36.96
SegFormer – B0 + CLA	56.22	28.15	44.98	54.25
SegFormer – B1 + CLA	66.31	57.69	45.45	51.65

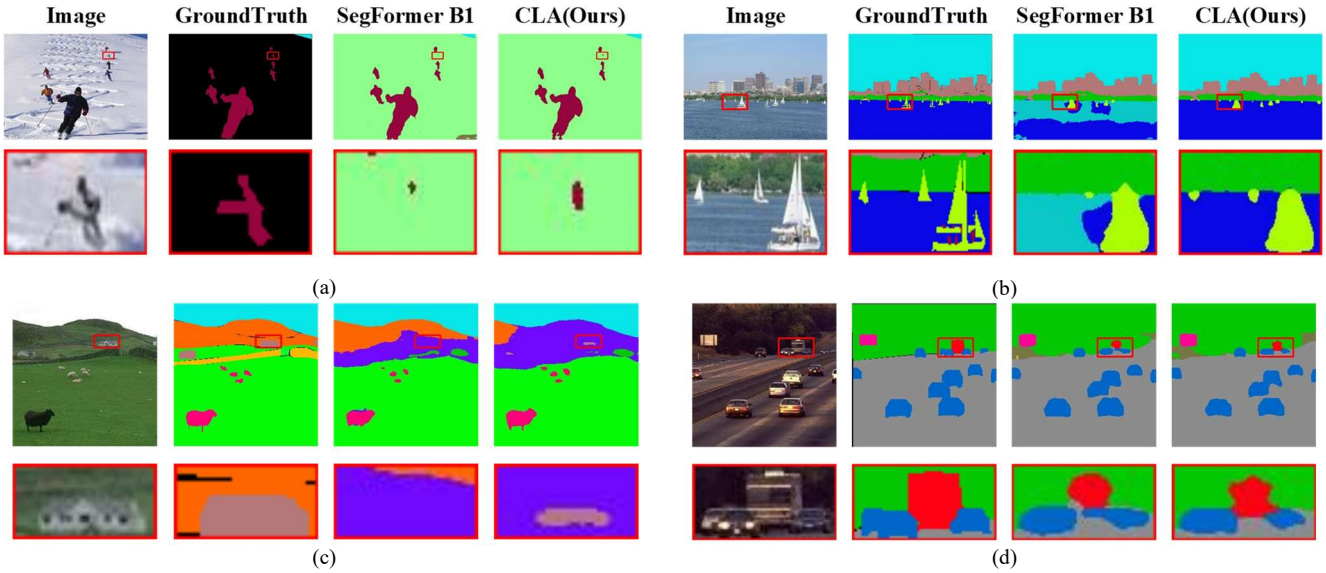


Fig. 2. Visual comparison of small objects. Each column from left to right in a subfigure represents the original image, ground truth, SegFormer B1 segmentation results, and our method. (a) person (b) boat (c) building (d) truck

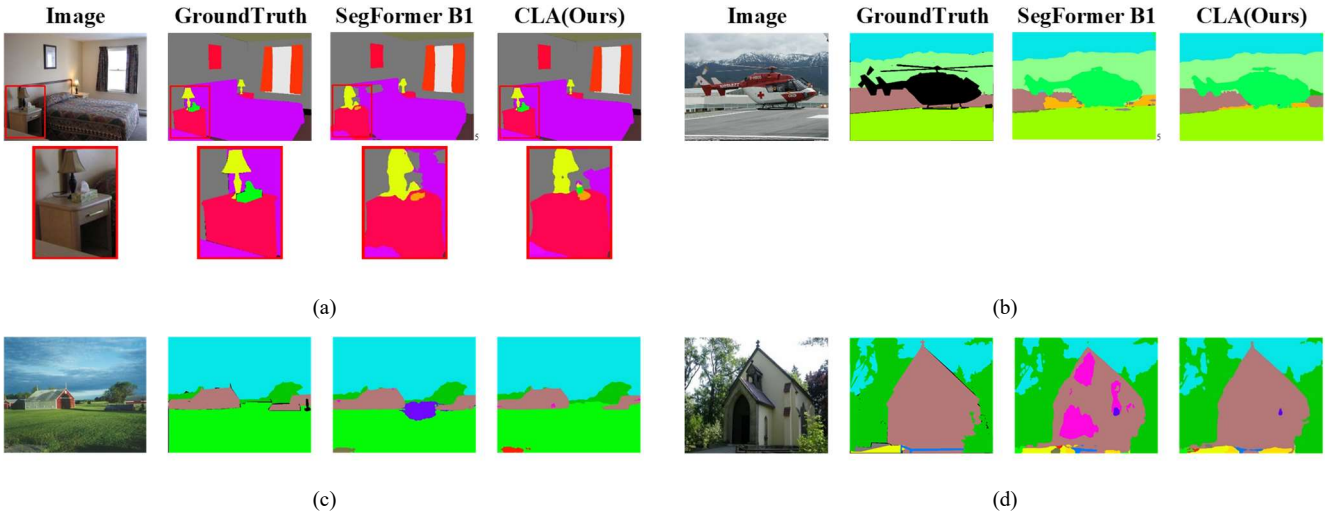


Fig. 3. Visual comparison of non-small objects. Each column from left to right in a subfigure represents the original image, ground truth, SegFormer B1 segmentation results, and our method. (a) lamp (b) helicopter (c) grass (d) church

fewer than those segmented by SegFormer with CLA. Fig. 2 (c) also exhibits a noticeable difference; the original SegFormer fails to segment distant small houses, whereas our model successfully segments both small houses. In Fig. 2 (d), part of a distant truck segmented by the original SegFormer is classified as a car, while our model clearly distinguishes different categories, and the segmented truck area is larger in our model compared to the original SegFormer. From the examples mentioned above, it can be observed that SegFormer with CLA performs better in handling small objects.

As shown in Fig. 3 (a), although both the original SegFormer and our model segment two lamps, it is evident that our results are more precise than the original SegFormer. In addition, for large objects such as helicopters and houses (see Fig. 3 (b) to (d)), the proposed method has less errors. To sum up, our method not only improves the ability of segmentation small object, but also non-small object.

V. Conclusions

This study aims to improve the semantic segmentation of small objects in natural scenes by using cross layer attentions, CLA, to obtain object information across different feature maps, thereby compensating for the features lost during downsampling. The extracted CLA is then utilized in the decoder, ultimately improves the semantic segmentation performance for both small and non-small objects.

It is noteworthy that in this study, only two CLA encoders were used: CLA_{14} and CLA_{23}. However, the CLA mechanism is not limited to these specific layers and can be extended to obtained from any two feature maps. Due to limitations of

hardware, we are unable to test the CLA mechanism with other combinations, such as extracting cla_{12} with cla_{34}. This is a topic we would like to explore in the future.

This study has successfully integrated SegFormer – B0 and B1 with CLA, which fuses multi-level feature information and compensates for the loss of features during the downsampling process. The proposed design improves the semantic segmentation performance in natural scenes on lightweight models, and the performance on small objects is particularly significant.

Due to the limitation of memory capacity, we were unable to implement SegFormer with AFMA as a baseline comparison to SegFormer with CLA. The implementation of AFMA requires computing cross-feature map attention using the original image (512×512) and the lowest level feature map (16×16). As a result, AFMA requires 128 GB of memory, which is 8 times more than CLA, which restrains us from trying more CLA connection methods. Therefore, in the future, we plan to continuously improve this method and apply CLA to different transformer or CNN architecture models to prove its generalizability on lightweight models.

VI. Acknowledgment

This study is funded by the National Science and Technology Council with grant no. 113-2221-E-033-041- and no. 113-2813-C-033-020-E

References

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998, doi: 10.1109/5.726791.

[2] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21-26 July 2017 2017, pp. 936-944, doi: 10.1109/CVPR.2017.106.

[3] A. Vaswani *et al.*, "Attention is all you need," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017.

[4] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-Aware Trident Networks for Object Detection," *2019 IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 6053-6062, 2019.

[5] B. Singh and L. S. Davis, "An Analysis of Scale Invariance in Object Detection - SNIP," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18-23 June 2018 2018, pp. 3578-3587, doi: 10.1109/CVPR.2018.00377.

[6] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-Object Detection in Remote Sensing Images with End-to-End Edge-Enhanced GAN and Object Detector Network," *Remote Sensing,* vol. 12, no. 9, p. 1432, 2020.

[7] S. Sang, Y. Zhou, M. T. Islam, and L. Xing, "Small-Object Sensitive Segmentation Using Across Feature Map Attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 5, pp. 6289-6306, 2023, doi: 10.1109/TPAMI.2022.3211171.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Cham, 2015: Springer International Publishing, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234-241.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," arXiv:1706.05587 doi: 10.48550/arXiv.1706.05587.

[10] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: simple and efficient design for semantic segmentation with transformers," presented at the Proceedings of the 35th International Conference on Neural Information Processing Systems, 2024.

[11] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ArXiv,* vol. abs/2010.11929, 2020.

[12] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, "Conditional Positional Encodings for Vision Transformers," p. arXiv:2102.10882doi: 10.48550/arXiv.2102.10882.

[13] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21-26 July 2017: IEEE, pp. 633-641, doi: CVPR.2017.544.

Data-Driven Sound Field Reproduction for Higher-Order Mode Matching Using a Circular Loudspeaker Array

Keiko Kawase*, Gen Sato*, Izumi Tsunokuni* and Yusuke Ikeda*
* Tokyo Denki University, Tokyo
E-mail: {24fmi07, 24udc03, 21udc02}@ms.dendai.ac.jp, yusuke.ikeda@mail.dendai.ac.jp

***Abstract*—The mode matching (MM) method is a well-known method of sound field reproduction. In MM, the driving functions of loudspeakers are derived by matching the reproduced sound field and the desired one through spherical harmonic function expansion. To fully represent a sound field, an infinite number of spherical harmonics is required. The accuracy of the MM method depends heavily on the number of microphones, and achieving higher-order harmonics necessitates a large number of microphones, leading to the truncation of spherical harmonic expansions. In this study, we propose an end-to-end deep learning approach based on the MM method to determine the driving functions, aiming to reduce the number of microphones required for recording the original sound field. Simulation results demonstrate that the proposed method can achieve sound field reproduction accuracy comparable to higher-order MM, while using approximately half the number of microphones.**

I. Introduction

Many sound field reproduction techniques have been proposed to create a physically accurate sound field, enabling a more immersive experience [1]–[3]. Sound field reproduction using multiple loudspeakers has the advantage of allowing the user to move freely within the reproduction area without wearing devices such as headphones [1]–[7]. This technology attracts attention for its potential applications in various fields, including virtual reality, concert halls, and automotive audio systems.

The mode matching (MM) method is a well-known numerical method for sound field reproduction [8]–[14]. In the MM method, the sound field is decomposed using spherical harmonic expansion to facilitate its reconstruction. The driving functions of the loudspeakers are determined to match the expansion coefficients of the spherical harmonic functions between the desired and reconstructed sound fields.

To fully represent a sound field, the expansion coefficients of infinite-order spherical harmonics are required [14]. However, when deriving the driving functions numerically in the MM method, the order of the spherical harmonics must be truncated. This truncation limits the size of the reproduction area. In particular, for a given reproduction area, a higher truncation order is required at higher frequencies. Moreover, the truncation order is dictated by the number of microphones used to record the desired sound field. As a result, a large number of microphones are required to record and reproduce a wide sound field [9].

In recent years, the development of deep learning techniques has led to the proposal of various methods for sound field reproduction using deep learning [15]–[18]. In particular, deep learning methods that focus on the structure of sound field data and extract more information from less spatial data have been proposed. For example, the pressure matching (PM) method with end-to-end learning to improve the reproduction accuracy at frequencies beyond the spatial Nyquist frequency has been proposed in [16]. In higher-order ambisonics (HOA), a data-driven learning method has been proposed by estimating higher-order spherical harmonics from lower-order ones [17]. Furthermore, we have previously proposed PM and MM based on sound field modeling using the sparse equivalent source method, and have achieved improvements in reproduction accuracy [19], [20].

In this study, we propose an end-to-end higher-order MM method based on deep learning that utilizes a small number of microphones. The proposed method estimates the higher-order harmonics from a reduced number of microphone signals. We specifically focus on estimating the original sound field during the recording process, assuming that the expansion coefficients of the sound field emitted by the loudspeakers are known. While in the previous study [17] estimated higher-order coefficients from lower-order ones, our proposed method directly estimates the higher-order coefficients from the microphone signals. Additionally the proposed method directly derives the driving functions of the loudspeakers.

The remainder of this paper is organized as follows. Section II presents the conventional MM method. In Sec. III, we describe the proposed method. In Sec. IV, simulation experiments are shown and the reproduced sound field in 2.5D is compared with the conventional MM method and the proposed method. Section V presents discussions. Finally, we concluded in Sec. VI.

II. Mode Matching

In the MM method, the driving functions are determined by matching the expansion coefficients of the spherical harmonics

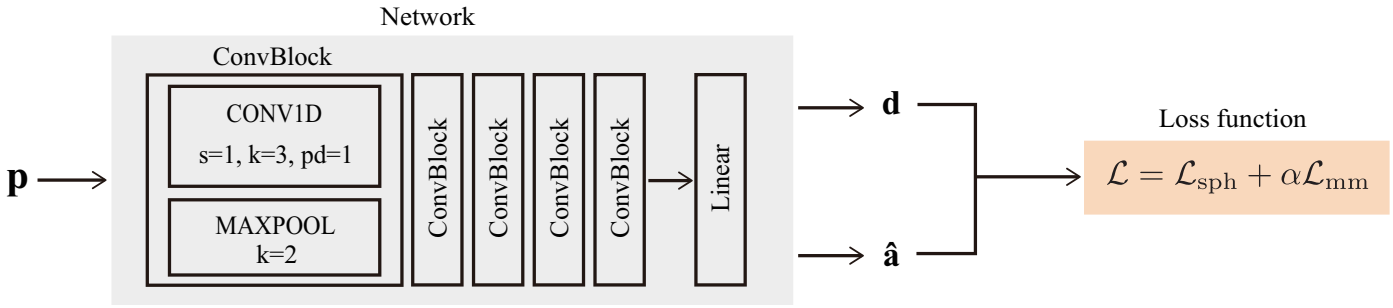


Fig. 1. Schematic diagram of the training and architecture of CNN in the proposed method. The microphone signal $\mathbf{p}$ is input to the network. The outputs are the driving functions $\mathbf{d}$ of L loudspeakers and the estimated expansion coefficients of the spherical harmonic function for the desired sound field $\hat{\mathbf{a}}$. The outputs $\mathbf{d}$ and $\hat{\mathbf{a}}$ are used in the loss functions $\mathcal{L}_{\mathrm{mm}}$ and $\mathcal{L}_{\mathrm{sph}}$. In the diagram, s is stride, k is kernel size and pd is padding.

between the desired and reproduced sound fields. For the desired sound field, the sound field of a point source at $(r_{\mathrm{s}}, \theta_{\mathrm{s}}, \phi_{\mathrm{s}})$ in the polar coordinate system can be represented by the coefficients $A_{n,m}^{(\mathrm{src})}$ for the spherical harmonics as follows [21]:

$$p(r,\theta,\phi) = \sum_{n=0}^{N}\sum_{m=-n}^{n} A_{n,m}^{(\mathrm{src})} j_n(kr) Y_n^m(\theta,\phi), \tag{1}$$

$$A_{n,m}^{(\mathrm{src})} = -ikh_n^{(2)}(kr_s) Y_n^{-m}(\theta_{\mathrm{s}},\phi_{\mathrm{s}}), \tag{2}$$

where i is the imaginary unit, k is the wave number, $j_n(\cdot)$ is n-th order spherical Bessel function, $h_n^{(2)}$ is the spherical Hankel function of the second kind and $Y_n^{-m}(\theta_s,\phi_s)$ is the spherical harmonic function. The maximum order N is determined by the wavelength λ and radius R of the reproduction area [21]

$$N = \lceil \frac{2\pi R}{\lambda} \rceil, \tag{3}$$

where $\lceil\cdot\rceil$ is the ceiling function. For a maximum order N, the number of required microphones M is obtained by [9]

$$M = (N+1)^2. \tag{4}$$

The reproduced sound field is represented as follows using L loudspeakers and its driving functions,

$$p(r,\theta,\phi) = \sum_{l=1}^{L} G(r_l,\theta_l,\phi_l) d_l, \tag{5}$$

where $G(\cdot)$ denotes the transfer function of l-th loudspeaker at (r_l,θ_l,ϕ_l) and d_l denotes the l-th driving function.

Thus, the following relationship holds from Eqs.(2) and (5),

$$\mathbf{a} = \mathbf{A}\mathbf{d}, \tag{6}$$

where $\mathbf{a}(\in \mathbb{C}^{(N+1)^2\times 1})$ is the expansion coefficient vector of the desired sound field, $\mathbf{A}(\in \mathbb{C}^{(N+1)^2\times L})$ is expansion coefficient matrix of sound field from all loudspeakers, and $\mathbf{d} = \{d_l\}(\in \mathbb{C}^{L\times 1})$ is driving function vector. In this study, the desired and loudspeaker's sound fields are emitted from single point source described by Eq. (2), $\mathbf{a} = \{A_{n,m}^{(\mathrm{src})}\}$, $\mathbf{A} = \{A_{n,m}^{(l)}\}$. $A_{n,m}^{(l)}$ denotes the expansion coefficients of l-th loudspeaker. In practice, spherical harmonic function expansions for arbitrary sound fields are required. In general, by solving Eq. (6) using the regularized least squares method or other methods [9].

III. Proposed Method

In the proposed method, the driving functions are determined using a deep neural network by estimating the higher-order expansion coefficients of the spherical harmonics for the desired sound field beyond the maximum order N limited by M microphones.

A. Networks

Fig. 1 shows the network architecture. A one-dimensional convolutional neural network (CNN) was mainly used for the architecture. This one-dimensional CNN had stride $\mathrm{s} = 1$, kernel size is $\mathrm{k} = 3$ and padding size $\mathrm{pd} = 1$. The CNN was followed by a pooling layer. The pooling layer used max pooling with kernel size $\mathrm{k} = 2$.

The microphone signal $\mathbf{p}(\in \mathbb{C}^{M\times 1})$ is the input to the network, and the output is the driving functions of L loudspeakers $\mathbf{d}(\in \mathbb{C}^{L\times 1})$. The estimated expansion coefficients of the spherical harmonics for the desired sound field $\hat{\mathbf{a}}(\in \mathbb{C}^{(N+1)^2})$ are output and used for learning in the loss function. $\hat{\mathbf{a}} = \{\hat{A}_{n,m}^{(\mathrm{src})}\}$. Note that in the proposed method, the estimated expansion coefficients are not directly used for deriving the driving functions, unlike in the conventional MM. The more details of losses are explained in Sec. III-B.

B. Losses

Using the estimated expansion coefficients $\hat{A}_{n,m}^{(\mathrm{src})}$ for the desired sound field and the driving function d_l of the l-th loudspeaker, the loss function is defined as follows,

$$\mathcal{L} = \mathcal{L}_{\mathrm{sph}} + \alpha\mathcal{L}_{\mathrm{mm}}, \tag{7}$$

$$\mathcal{L}_{\mathrm{sph}} = \frac{1}{N_{\mathrm{all}}}\sum_{n=0}^{N}\sum_{m=-n}^{n}\left|\hat{A}_{n,m}^{(\mathrm{src})} - A_{n,m}^{(\mathrm{src})}\right|^2, \tag{8}$$

$$\mathcal{L}_{\mathrm{mm}} = \frac{1}{N_{\mathrm{all}}}\sum_{n=0}^{N}\sum_{m=-n}^{n}\left|\left(\sum_{l=1}^{L} d_l A_{n,m}^{(l)}\right) - \hat{A}_{n,m}^{(\mathrm{src})}\right|^2 + \frac{\beta}{L}\sum_{l=1}^{L}|d_l|^2, \tag{9}$$

where α is the balance parameter, β is the regularization parameter, and $N_{\mathrm{all}} = (N+1)^2$ is the number of all elements

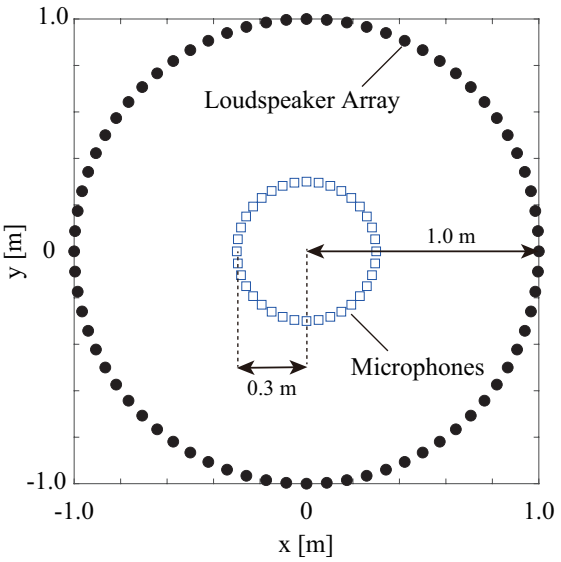


Fig. 2. Position of microphone array and loudspeaker array. Blue squares are microphones at radius 0.3 m. Black dots are loudspeakers at radius 1.0 m

in the expansion coefficient matrix $\mathbf{A}$. $\mathcal{L}_{\mathrm{sph}}$ is the loss on the expansion coefficient estimation for the desired sound field. In this study, the expansion coefficients for the loudspeakers $A_{n,m}^{(l)}$ are known. Thus, the loss $\mathcal{L}_{\mathrm{mm}}$ is calculated by the error between the estimated expansion coefficients $\hat{A}_{n,m}^{(\mathrm{src})}$ for the desired sound field and the expansion coefficients for the reproduced sound field, using the derived driving functions and the known expansion coefficients $A_{n,m}^{(l)}$. Furthermore, a regularization term $\frac{\beta}{L}\sum_{l=1}^{L}|d_l|^2$ is also introduced in the loss $\mathcal{L}_{\mathrm{mm}}$ to prevent over-fitting.

In this study, each loudspeaker and the desired sound source are regarded as single-point sources. Thus, the expansion coefficients $A_{n,m}^{(l)}$ for l-th loudspeaker can be obtained as in Eq. (2).

IV. Simulation Experiment

A. Conditions

Simulation experiments were conducted to evaluate the accuracy of the proposed method in reproducing the sound field using a circular loudspeaker array. The experimental conditions of the dataset are shown in Table I. In addition, the positions of the microphones and loudspeakers are shown in Fig. 2. The shape of the microphone array was also circular. The sound propagation was three-dimensional, and the evaluation region was the plane containing the circular loudspeaker array. In other words, the sound field condition is 2.5-dimensional. In this study, the network was trained using a single point source randomly placed sound field as the desired sound field.

The number of the desired sound fields in the dataset is 1800, and the ratio of training data to validation data is $8:2$. The batch size was 8. The number of epochs was 300. The learning rate was set to 0.0001. Adam was used as the optimization method. For comparison, in the conventional MM method, the driving functions were derived using the regularized least squares method.

TABLE I
Experimental conditions of dataset

Distances of point sources from the origin [m]	$1.5 \sim 3.5$
Frequency [Hz]	750
Number of loudspeakers L	72
Radius of loudspeaker array [m]	1
Number of microphones M	36
Radius of loudspeaker array [m]	0.3
Maximum order of learning data N	7

To evaluate the reproduction accuracies of sound fields with MM methods and DLMM (proposal) methods, the signal-to-noise ratio (SNR) was defined as

$$\mathrm{SNR} = 10\log_{10}\frac{|p_{\mathrm{des}}|^2}{|p_{\mathrm{des}}-p_{\mathrm{rep}}|^2}, \tag{10}$$

where p_{des} and p_{rep} are the desired and reproduced complex pressures. Similarly, the reconstructed accuracy of the sound field from the estimated expansion coefficients $\hat{A}_{n,m}^{(\mathrm{src})}$ of the spherical harmonics were evaluated by Eq. (10).

B. Results

The desired sound field for evaluation, which is not included in the training data, was a point source at position $(x, y = 2.0, 1.5)$ m. Based on Eq. (4), we compared the conventional MM with the condition at 36 microphones, i.e., maximum order $N = 5$.

First, we evaluated the estimation accuracy of the desired sound field. Fig. 3 shows the comparison of the desired and estimated sound fields ($N = 5, 7$) by the DLMM (proposed method). In conventional MM, the maximum order of spherical harmonic expansion is five, which is decided by the number of microphones as Eq. (4). From Fig. 3, the desired sound field ($N = 5$) was degraded compared to $N = 7$. As shown in Fig. 3(c), by the proposed method, the expansion coefficients estimated from the microphone signals can reconstruct the higher-order desired sound field ($N = 7$). Fig. 3(d) shows the SNR of DLMM compared to the desired sound field ($N = 7$). The mean of SNR was 22.0 dB. The figure shows that the estimation is sufficiently accurate within the evaluation area. Therefore, the proposed method can obtain spherical harmonic expansion coefficients with an accuracy of $N = 7$ using the number of microphones corresponding to $N = 5$.

Second, we evaluated the reproduction accuracies of the proposed method. Fig. 4(a)–(c) show the desired and reproduced sound fields. Fig. 4(d) and (e) show the SNR distribution. From Figs. 4(a)–(c), the proposed method (DLMM) can accurately reproduce the sound field over the entire evaluation region compared to the conventional MM method ($N = 5$), where the maximum order is constrained by the number of microphones.

From Figs. 4(d)–(e), there was a slight decrease in SNR near the origin for the proposed method compared to the conventional MM method. However, over the entire evaluation region, the proposed method can reproduce the sound field with a higher SNR. In addition, Table II compares the average SNR with MM ($N = 5, 7$) and proposed method for different radii of the evaluation region. For the 0.4 m radius evaluation area,

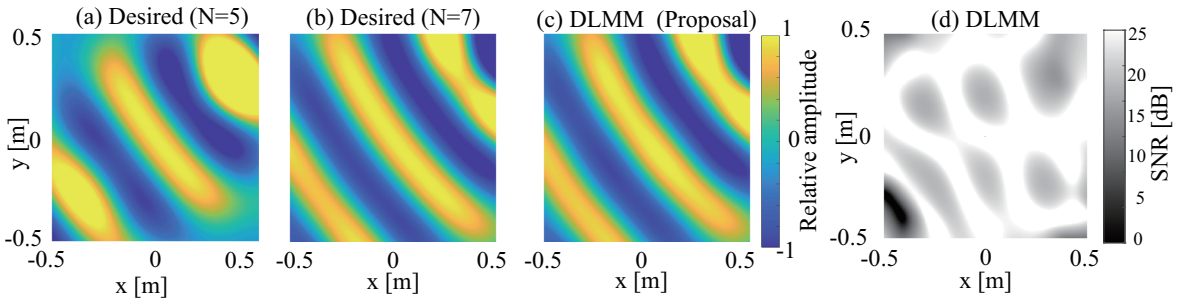


Fig. 3. Comparison of desired and DLMM-estimated sound fields. In the desired sound field, the point source is at (2.0 m, 1.5 m) (a) Desired sound field ($N = 5$), (b) Desired sound field ($N = 7$), (c) Estimated sound field with DLMM (proposal method), (d) SNR map with DLMM.

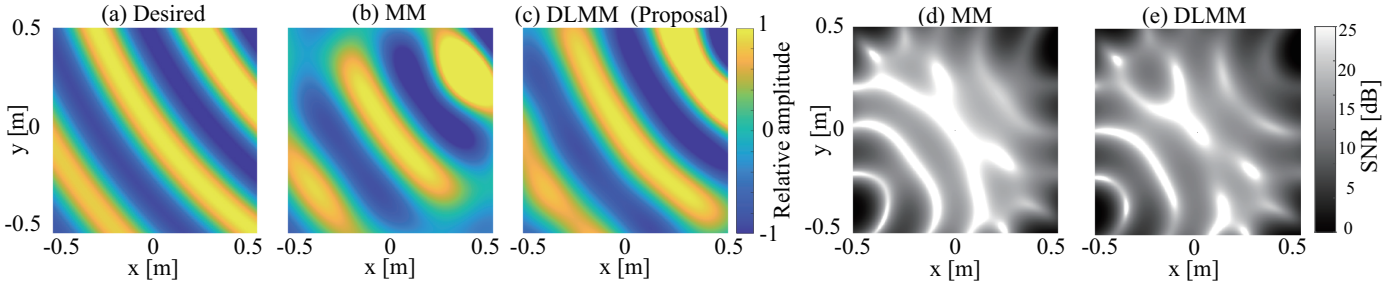


Fig. 4. Comparison of reproduced sound fields by the conventional MM and DLMM (proposed method) at 750 Hz:(a) Desired sound field, (b) Reproduced sound field with MM methods as a comparison at maximum order of MM $N = 5$, (c) Reproduced sound field with DLMM (proposal method), (d) SNR map with MM method, and (e) SNR map with DLMM.

TABLE II
COMPARISON OF AVERAGE SNR WITH RESPECT TO THE RADII OF THE EVALUATION REGIONS

Radius of eval. region	MM($N = 5$)	MM($N = 7$)	DLMM
0.4 m	15.5 dB	18.2 dB	15.7 dB
0.5 m	9.9 dB	14.9 dB	13.7 dB

the mean SNRs were similar for DLMM and MM ($N = 5$). On the other hand, in the 0.5 m radius evaluation region, the proposed method improved by approximately 4 dB compared to the conventional MM ($N = 5$). The proposed method also achieved SNR comparable to MM ($N = 7$).

V. DISCUSSIONS

The proposed method simultaneously estimated the higher-order spherical harmonic expansion and derived the driving functions based on MM. Simulation experiments confirmed that using 36 microphones can estimate the higher-order expansion coefficients up to $N = 7$, resulting in a wider range of sound field reproduction with higher accuracy.

In this study, the sound field was reproduced under 2.5D conditions. However, even when applied to 3D sound field reproduction, it is relatively easy to achieve by revising the network architecture because the expansion coefficients of spherical harmonics and driving functions are similarly represented as vectors.

One limitation of the proposed method is that it creates a separate model for each frequency, which requires many models to be applied to broadband signals. In addition, as the frequency increases, the wavelength may become significantly shorter relative to the spacing of the microphones, resulting in difficulty in estimating higher-order expansion coefficients. Therefore, achieving higher-order harmonics expansion at higher frequencies requires further investigation, such as examining appropriate microphone array configurations.

Other limitations include the fact that the model was trained only by a single point source as the desired sound field in this study. However, if a more general sound field is considered, the model should be trained with data that includes multiple sound sources and reflections. Furthermore, in this study, the spherical harmonic expansion coefficients of the sound field radiated from the loudspeakers were assumed to be known. However, it is expected that a similar model can be constructed for loudspeakers to estimate the expansion coefficients from the small number of microphone signals.

VI. CONCLUSIONS

In this study, we proposed a data-driven method for determining the driving functions of loudspeakers based on MM to reduce the number of microphones in recording the original sound field. Simulation experiments have demonstrated that the proposed method can achieve sound field reproduction with accuracy comparable to higher-order MM using a limited number of microphones. In future work, the proposed method will be evaluated with measured data, and other constraints, such as physical models, will be considered to achieve higher-order MM.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number 22K12099 and Research Institute for Science and Technology of Tokyo Denki University Grant Number Q24J-04 / Japan.

REFERENCES

[1] O. Kirkeby and P. A. Nelson, "Reproduction of plane wave sound fields," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2992–3000, 1993.

[2] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[3] D. Jerome, M. Sebastien, and N. Rozenn, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," in *in Proceedings of Audio Engineering Society Convention 114*, 2003.

[4] A. J. Berkhout, "A holographic approach to acoustic control," *Journal of the Audio Engineering Society*, vol. 36, pp. 977–995, 12 1988.

[5] S. Ise, "A principle of active control of sound based on the kirchhoff-helmholtz integral equation and the inverse system theory (in japanease)," *The Journal of the Acoustical Society of Japan*, vol. 53, no. 9, pp. 706–713, 1997.

[6] D. Jerome, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," in *in AES 23rd International Conference*, 2003.

[7] J. Ahrens and S. Spors, "Local sound field synthesis by virtual secondary sources," in *in AES 40th International Conference*, 2010.

[8] N. Ueno, S. Koyama, and H. Saruwatari, "Three-dimensional sound field reproduction based on weighted mode-matching method," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1852–1867, 2019.

[9] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the audio engineering society*, vol. 53, pp. 1004–1025, 11 Nov. 2005.

[10] P. N. Samarasinghe, M. A. Poletti, S. M. A. Salehin, T. D. Abhayapala, and F. M. Fazi, "3D soundfield reproduction using higher order loudspeakers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 306–310.

[11] N. Ueno, S. Koyama, and H. Saruwatari, "Sound field recording using distributed microphones based on harmonic analysis of infinite order," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 135–139, 2018.

[12] R. Matsuda and M. Otani, "Binaural-centered mode-matching method for enhanced reproduction accuracy at listener's both ears in sound field reproduction," *The Journal of the Acoustical Society of America*, vol. 150, no. 5, pp. 3838–3851, Nov. 2021.

[13] S. Koyama, K. Kimura, and N. Ueno, "Weighted pressure and mode matching for sound field reproduction: Theoretical and experimental comparisons," *Journal of the Audio Engineering Society*, vol. 71, pp. 173–185, 4 2023.

[14] E. G. Williams, *Fourier Acoustics*. London: Academic Press, 1999.

[15] S. Gao, J. Lin, X. Wu, and T. Qu, "Sparse dnn model for frequency expanding of higher order ambisonics encoding process," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1124–1135, 2022.

[16] X. Hong, B. Du, S. Yang, M. Lei, and X. Zeng, "End-to-end sound field reproduction based on deep learning," *The Journal of the Acoustical Society of America*, vol. 153, no. 5, pp. 3055–3064, 2023.

[17] L. Zhang, X. Wang, R. Hu, D. Li, and W. Tu, "Estimation of spherical harmonic coefficients in sound field recording using feed-forward neural networks," *Multimedia Tools and Applications*, vol. 80, pp. 6187–6202, 2021.

[18] L. Comanducci, F. Antonacci, and A. Sarti, "Synthesis of soundfields through irregular loudspeaker arrays based on convolutional neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 17, 2024.

[19] I. Tsunokuni, K. Kurokawa, and Y. Ikeda, "Pressure-matching-based 2d sound field synthesis with equivalent source array," in *23rd International Congress on Acoustics*, 2019, pp. 2701–2707.

[20] Y. Suzuki, I. Tsunokuni, and Y. Ikeda, "2.5 dimensional sound field reproduction based on mode matching and equivalent sources considering primary reflections," in *Proceedings of InterNoise 2023*, 2023.

[21] J. Ahrens, *Analytic Methods of Sound Field Synthesis*. Springer Berlin, Heidelberg, 2012.

Subset Random Sampling of Finite Time-vertex Graph Signals

Hang Sheng*, Qinji Shu*, Hui Feng*†and Bo Hu*†
* School of Information Science and Technology, Fudan University, Shanghai 200433, China
† State Key Laboratory of Integrated Chips and Systems, Fudan University, Shanghai 200433, China
E-mail: hsheng20@fudan.edu.cn, qjshu22@m.fudan.edu.cn, {hfeng, bohu}@fudan.edu.cn

***Abstract*—Time-varying data with irregular structures can be described by finite time-vertex graph signals (FTVGS), which represent potential temporal and spatial relationships among multiple sources. While sampling and corresponding reconstruction of FTVGS with known spectral support are well investigated, methods for the case of unknown spectral support remain underdeveloped. Existing random sampling schemes may acquire samples from any vertex at any time, which is uncommon in practical applications where sampling typically involves only a subset of vertices and time instants. In sight of this requirement, this paper proposes a subset random sampling scheme for FTVGS. We first randomly select some rows and columns of the FTVGS to form a submatrix, and then randomly sample within the submatrix. Theoretically, we prove sufficient conditions to ensure that the original FTVGS is reconstructed with high probability. Also, we validate the feasibility of reconstructing the original FTVGS by experiments.**

I. Introduction

Graphs can capture the underlying structure of data, making graph signals valuable for processing spatial information of data [1]. When data exhibit temporal variations, signal analysis also involves discussions in the time dimension. The concept of time-vertex graph signals (TVGS) was introduced in [2]. For example, Grassi *et al.* formed a product graph by stacking graph signals in finite time instances [2], which we refer to as finite time-vertex graph signals (FTVGS).

Sampling and reconstruction for TVGS is a crucial aspect, aiming to recover the entire signal from partial observations, thereby saving on the costs of acquisition, storage, and processing [3]. TVGS can be reconstructed from samples because real-world TVGS typically exhibit low-frequency characteristics in both the vertex and time domains. For example, when representing FTVGS with a matrix, this matrix would be low-rank and smooth.

There are some advancements in the research of sampling of FTVGS, provided that the joint time-vertex Fourier transform (JFT) spectrum support of FTVGS is known. The separate sampling scheme applies sampling to the graph and time domains of FTVGS respectively [4], while the joint sampling scheme further reduces the sampling rate based on the JFT spectrum of FTVGS [3]. The sampling framework based on subspace prior information applies the sampling framework of shift-invariant spaces to FTVGS [5]. Signal sampling and reconstruction algorithms based only on known graph spectral support were also proposed [6].

However, in reality, it is challenging to know the JFT spectrum of the FTVGS to be sampled beforehand. For instance, in FTVGS acquired by sensor networks, each row of the matrix represents a time series collected by a sensor, and each column represents a traditional graph signal. Before deploying the sensors, it is impossible to know the specific information of the FTVGS before sampling. Thus, the sampling methods for FTVGS based on known spectrum are not feasible.

To address the challenges posed by the unknown spectrum, we adopt a random sampling strategy. One approach is to sample the entire FTVGS through some unbiased stochastic process, which assumes that all elements in the matrix have an equal probability of being sampled [7], [8]. However, this assumption implies all the vertices could be sampled at any time (Fig. 1 (a)), which does not meet the requirement of reducing the number of sensors to be sampled. Cai *et al.* proposed a more flexible sampling scheme [9], but it still requires every row and column of the matrix to produce samples (Fig. 1 (b)). In practice, due to hardware, environmental, or cost limitations, we prefer to sample only a subset of rows and columns.

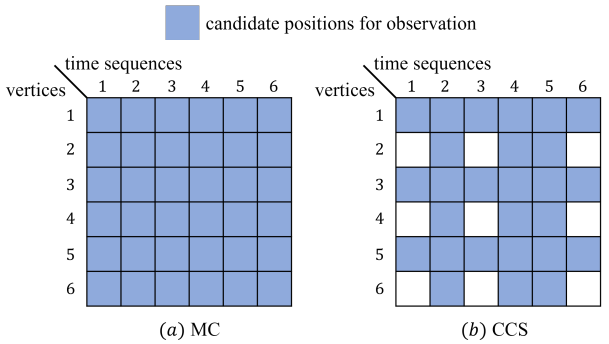
We propose a subset random sampling scheme for FTVGS to meet practical needs, significantly reducing the number of rows and columns from which samples are produced (Fig. 2 (c)). Specifically, we first randomly select subsets of rows and columns of the matrix to form a submatrix, and then randomly observe samples within the submatrix. Theoretically, under common assumptions, we prove sufficient conditions to ensure a high probability of reconstructing the original FTVGS. Additionally, the subset random sampling scheme impacts the complexity of the required samples and the accuracy of reconstructing the original signal, which we discuss in detail in Sections IV and V.

II. Related Works

The representation of FTVGS is a matrix, so the random sampling for FTVGS is also a random sampling problem on matrices. The pioneering work [7] studied the matrix completion (MC) problem based on the Bernoulli sampling model and proved a lower bound on the number of samples sufficient for exact recovery with high probability. Subsequent research on MC has extended beyond Bernoulli sampling to include uniform sampling with and without replacement [8], [10].

In subsequent studies, uniform sampling in MC was combined with CUR decomposition, resulting in a more flexible cross-concentrated sampling (CCS) [9]. But, in CCS, at least one element from each row or column of the FTVGS must still be sampled.

Fig. 1 visualizes the potential sampling regions for uniform sampling in MC and CCS. It is evident that neither method reduces the number of rows and columns being sampled in the matrix.



(*a*) MC (*b*) CCS

Fig. 1. The visualization of candidate positions for observation.

We aim to restrict the sample indices within subsets of the rows and columns of FTVGS, ensuring no samples are produced on unselected rows and columns. This approach is motivated by practical applications, where sampling is often desired on only partial vertices and time instants. For example, in sensor networks, we want to deploy as few sensors as possible, and it is not supported by hardware for each instant to have an equal probability of being sampled. In social networks, not everyone is willing to share their thoughts, and there may be times when participants cannot be recruited.

III. Model

An undirected graph can be represented as $\mathcal{G} = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$, where $\mathcal{V}_\mathcal{G}$ is the set of vertices with $|\mathcal{V}_\mathcal{G}| = N$, $\mathcal{E}_\mathcal{G}$ is the set of edges. We consider every vertex in $\mathcal{V}_\mathcal{G}$ relates to a time sequence of length T, which can be represented by a directed cyclic graph $\mathcal{T} = (\mathcal{V}_\mathcal{T}, \mathcal{E}_\mathcal{T})$. Such signals are FTVGS [2], whose graphical model is $\mathcal{G} \times \mathcal{T}$, where $\times$ is the Cartesian product. An FTVGS is $\mathbf{X} \in \mathbb{C}^{N\times T}$. Each row of $\mathbf{X}$ is a finite time sequence on a vertex, and each column is a static graph signal.

Suppose each edge of $\mathcal{G}$ is assigned an orientation, which is arbitrary but fixed. We then have the incidence matrix $\mathbf{Q} \in \mathbb{R}^{N\times M}$ and the graph Laplacian matrix $\mathbf{L}_\mathcal{G} = \mathbf{Q}\mathbf{Q}^T$. Then the gradient on $\mathcal{G}$ can be defined as $\mathbf{Q}^T\mathbf{X}$ and the graph total variation at time j is $\mathbf{X}(:,j)^T\mathbf{L}\mathbf{X}(:,j)$.

Then we denote first and second-order temporal differential operators by $\mathbf{D}_1 \in \mathbb{R}^{T\times(T-1)}$ and $\mathbf{D}_2 \in \mathbb{R}^{T\times(T-2)}$. For each $i \in \mathcal{V}_\mathcal{G}$, the first and second-order differences of the signal on $\mathcal{T}$ can be represented as $\mathbf{X}\mathbf{D}_1(i,j)$ and $\mathbf{X}\mathbf{D}_2(i,j)$. The gradient of $\mathbf{X}$ in the graph and time domains is

$$\triangledown\mathbf{X}(i,j) = \begin{bmatrix}\mathbf{Q}^T\mathbf{X}(i,j)\\ \mathbf{X}\mathbf{D}_1(i,j)\end{bmatrix}.$$

These are widely used assumptions.

Assumption 1. *An FTVGS $\mathbf{X} \in \mathbb{C}^{N\times T}$ is a rank-r matrix, where $r < \min\{N,T\}$. For all $i \in \mathcal{V}_\mathcal{G}, j \in \mathcal{V}_\mathcal{T}$,*

$$\max\left\{\left|\mathbf{X}(:,j)^T\mathbf{L}\mathbf{X}(:,j)\right|, \parallel \triangledown\mathbf{X}(i,j) \parallel, |\mathbf{X}\mathbf{D}_2(i,j)|\right\} \le C$$

for some constant C.

The low rank implies that $\mathbf{X}$ can be expressed as a matrix with only r non-zero rows (or columns) through a row (or column) linear transformation. This is consistent with the assumption of bandlimited JFT spectra of FTVGS, as JFT is a type of linear transformation. That is, $\mathbf{X} = \Psi_\mathcal{G}\mathbf{F}_\mathcal{G}, \mathbf{X}^T = \Psi_\mathcal{T}\mathbf{F}_\mathcal{T}$, where $\Psi_\mathcal{G}$ and $\Psi_\mathcal{T}$ are the graph Fourier transform (GFT) basis matrix and discrete Fourier transform (DFT) basis matrix, respectively [2]. $\mathbf{F}_\mathcal{G}$ and $\mathbf{F}_\mathcal{T}$ are the GFT and DFT spectra respectively, where each $\mathbf{F}_\mathcal{G}(:,j)$ is the bandlimited spectrum of the graph signal $\mathbf{X}(:,j)$ and each $\mathbf{F}_\mathcal{T}(:,i)$ is the bandlimited spectrum of the time sequence $\mathbf{X}(i,:)$ [3].

The graph total variation on $\mathcal{G}$, second-order difference on $\mathcal{T}$, and gradient are all bounded implies that $\mathbf{X}$ is smooth. That is, there are no abrupt changes in $\mathbf{X}$ in either spatial or temporal domains. Correspondingly, in the context of TVGS processing theory, the smoothness of $\mathbf{X}$ somehow indicates that its JFT spectrum exhibits low-frequency characteristics.

Since $\mathbf{X}$ is rank-r, its thin singular value decomposition (SVD) is

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

where $\mathbf{U} \in \mathbb{R}^{N\times r}$, $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ and $\mathbf{V} \in \mathbb{R}^{T\times r}$.

Assumption 2. *[7], [8] An FTVGS $\mathbf{X} \in \mathbb{C}^{N\times T}$ satisfying Assumption 1 is $\{\mu_1, \mu_2\}$-incoherence for some constants $1 \le \mu_1(\mathbf{X}) \le \frac{N}{r}, 1 \le \mu_2(\mathbf{X}) \le \frac{T}{r}$ provided*

$$||\mathbf{U}||_{2,\infty} \le \sqrt{\frac{\mu_1(\mathbf{X})r}{N}}, ||\mathbf{V}||_{2,\infty} \le \sqrt{\frac{\mu_2(\mathbf{X})r}{T}},$$

where $||\mathbf{U}||_{2,\infty} := \max_i \left(\sum_j (\mathbf{U}(i,j))^2\right)^{1/2}$.

Coherence indicates whether the energy of $\mathbf{X}$ is uniformly distributed across columns and rows. Low coherence means that the structure of $\mathbf{X}$ is relatively uniform, making it easier to obtain uniform samples during random sampling.

IV. Method

We propose *a subset random sampling scheme* for signals, as illustrated in Fig. 2. Formally, we first randomly select a subset of rows and columns from the FTVGS without replacement (Fig. 2 (a)), *i.e.*,

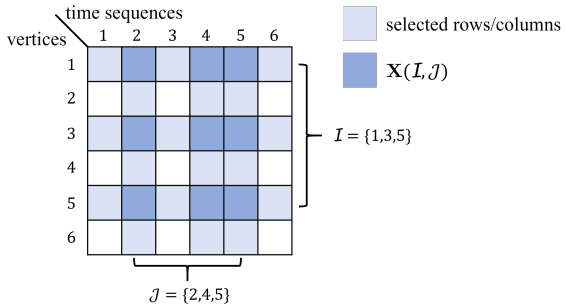
$$\mathcal{I} \subseteq \mathcal{V}_\mathcal{G}, \mathcal{J} \subseteq \mathcal{V}_\mathcal{T}.$$

Let $\mathbf{X}(\mathcal{I},\mathcal{J})$ be the submatrix of $\mathbf{X}$ composed of rows indexed by $\mathcal{I}$ and columns indexed by $\mathcal{J}$. Let the samples observed from FTVGS $\mathbf{X}$ be indexed by a set $\mathcal{S}$, and $|\mathcal{S}|$ denote the

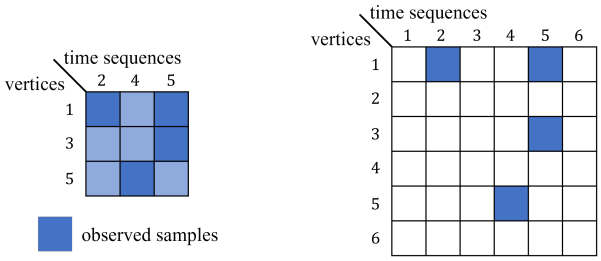
number of samples. Subsequently, we uniformly observe samples within the submatrix $\mathbf{X}(\mathcal{I}, \mathcal{J})$ with replacement (Fig. 2 (b), (c)), *i.e.*,

$$\mathcal{S} = \{(i, j) : i \in \mathcal{I}, j \in \mathcal{J}\}.$$

Let $\mathscr{P}_{\mathcal{S}}(\cdot)$ be the projection operator onto $\mathcal{S}$.



(a) Randomly select rows and columns



(b) Sampling on $\mathbf{X}(\mathcal{I}, \mathcal{J})$ (c) Position of the samples in $\mathbf{X}$

Fig. 2. Flowchart of the subset random sampling scheme.

Unlike existing methods, we would like to sample randomly on the submatrix $\mathbf{X}(\mathcal{I}, \mathcal{J})$ instead of $\mathbf{X}$, ensuring that no samples are observed on unselected rows and columns. We summarize the sampling process as Procedure 1.

Procedure 1 Subset random sampling

Input: $\mathbf{X}$

1: Uniformly sample row indices $\mathcal{I}$ without replacement, yielding $\mathbf{X}_R = \mathbf{X}(\mathcal{I}, :)$.
2: Uniformly sample column indices $\mathcal{J}$ without replacement, getting $\mathbf{X}_{RC} = \mathbf{X}(\mathcal{I}, \mathcal{J})$.
3: Uniformly sample entries in $\mathbf{X}_{RC}$ and get $\mathcal{S}$.

Output: $\mathscr{P}_{\mathcal{S}}(\mathbf{X})$, $\mathcal{I}$, $\mathcal{J}$, $\mathcal{S}$

The subset random sampling scheme is straightforward but violates the common assumption in MC problems that each row and column must produce at least one sample. To the best of our knowledge, theoretical guarantees for reconstructing the original FTVGS when the samples are structurally missing in entire rows and columns have not been explored. Therefore, we then study subset random sampling from a theoretical perspective and provide a sufficient condition that ensures reconstruction of the original signal with high probability from the samples generated by subset random sampling.

The following lemma ensures that the rank of $\mathbf{X}(\mathcal{I}, \mathcal{J})$ remains r with high probability, which is crucial for the subsequent theoretical derivation. Lemma 1 also provides the lower bounds on the number of selected rows and columns.

Lemma 1. *For an FTVGS $\mathbf{X} \in \mathbb{C}^{N \times T}$ satisfying Assumption 1 and Assumption 2, let $\mathbf{X}_R$ and $\mathbf{X}_{RC}$ be obtained by Procedure 1. Then if*

$$|\mathcal{I}| \geq 3r\mu_1(\mathbf{X})\frac{\ln(2r/\delta)}{\epsilon^2}, |\mathcal{J}| \geq 3r\mu_2(\mathbf{X})\frac{\ln(2r/\delta)}{\epsilon^2}$$

for some $0 < \delta < 1$ and $0 < \epsilon < 1$, $\operatorname{rank}(\mathbf{X}_R) = r$ holds with probability at least $1-\delta$ and $\operatorname{rank}(\mathbf{X}_{RC}) = r$ holds with probability at least $(1-\delta)^2$.

The $\mathbf{X}_R = \mathbf{U}_R\Sigma_R\mathbf{V}_R^T$ is shown in Eq. (3) and the SVD of $\mathbf{X}_{RC}$ is

$$\begin{aligned}\mathbf{X}_{RC} &= \mathbf{U}_R\Sigma_R(\mathbf{V}_R(\mathcal{J}, :))^T = \mathbf{U}_R\tilde{\mathbf{U}}_{RC}\Sigma_{RC}\mathbf{V}_{RC}^T \\ &= \mathbf{U}_{RC}\Sigma_{RC}\mathbf{V}_{RC}^T,\end{aligned} \tag{1}$$

where $\Sigma_R(\mathbf{V}_R(\mathcal{J}, :))^T$ has the compact SVD $\tilde{\mathbf{U}}_{RC}\Sigma_{RC}\mathbf{V}_{RC}^T$ with $\tilde{\mathbf{U}}_{RC} \in \mathbb{R}^{r \times r}$ being an orthonormal matrix, $\mathbf{V}_{RC} \in \mathbb{R}^{r \times |\mathcal{J}|}$, and $\mathbf{U}_{RC} = \mathbf{U}_R\tilde{\mathbf{U}}_{RC}$ is a matrix with orthogonal columns.

Next, Lemma 2 shows how the properties of the original signal $\mathbf{X}$ are converted to those of its submatrix $\mathbf{X}(\mathcal{I}, \mathcal{J})$. The proof of Lemma 2 is presented in APPENDIX B.

Lemma 2. *Under the assumption of Lemma 1, if $\mathcal{I}$ and $\mathcal{J}$ satisfy the inequality of Lemma 1, then the following conditions hold with probability at least $(1-p)^2, p = r\left[\frac{e^{-\eta}}{(1-\eta)^{1-\eta}}\right]^{\log r}$:*

$$||\mathbf{U}_{RC}||_{2,\infty} \leq \kappa(\mathbf{X})\sqrt{\frac{\mu_1(\mathbf{X})r}{(1-\eta)|\mathcal{I}|}},$$

$$||\mathbf{V}_{RC}||_{2,\infty} \leq \frac{\kappa(\mathbf{X})}{1-\eta}\sqrt{\frac{\mu_2(\mathbf{X})Nr}{|\mathcal{I}||\mathcal{J}|}},$$

where $\kappa(\mathbf{X})$ is the condition number of $\mathbf{X}$ and $\eta \in [0, 1)$.

By Lemma 1 and Lemma 2, we present our main theorem, and its proof is presented in APPENDIX C.

Theorem 1. *For an FTVGS $\mathbf{X} \in \mathbb{C}^{N \times T}$ satisfying Assumption 1 and Assumption 2, suppose that $\mathbf{X}_{RC}$ and $\mathcal{S}$ is derived from Procedure 1. Then if*

$$|\mathcal{S}| \geq \frac{32\beta(\kappa(\mathbf{X}))^4r^2N}{(1-\eta)^3}\mu_1(\mathbf{X})\mu_2(\mathbf{X})\frac{|\mathcal{I}|+|\mathcal{J}|}{|\mathcal{I}|}\log^2(2n),$$

where $\beta > 1$, $n = \max\{|\mathcal{I}|, |\mathcal{J}|\}$, the $\mathbf{X}$ can be recovered from $\mathcal{S}$ with probability at least

$$(1-\delta)^2(1-p)^2 - \frac{6\log n}{(|\mathcal{I}|+|\mathcal{J}|)^{2\beta-2}} - n^{2-2\beta^{1/2}}.$$

Lemma 1 and Theorem 1 provide the conditions for reconstructing the original FTVGS from subset random sampled samples. Among the probabilities in Theorem 1, $(1-\delta)^2$ is the probability that guarantees $\operatorname{rank}(\mathbf{X}_{RC}) = \operatorname{rank}(\mathbf{X}) = r$

and $(1-p)^2$ is the probability that guarantees the incoherence of $\mathbf{X}_{RC}$. We summarize the differences between random sampling of MC, CCS, and subset random sampling in Table I.

Our proposed subset random sampling scheme restricts the sample indices within $\mathbf{X}(\mathcal{I}, \mathcal{J})$, meaning that no samples will be drawn from unselected rows and columns. However, this comes at the cost of some reconstruction error, as detailed in Section V. Since the seminal work on the MC problem [7], the lower bounds on the number of randomly sampled entries have been relatively loose. Similarly, the lower bound provided in Theorem 1, based on the approach in [7], is also loose, particularly when N and T are small.

V. EXPERIMENTS

We validate Theorem 1 on a real-world dataset and propose a reconstruction algorithm tailored to the characteristics of subset random sampling. The details of the algorithm will be analyzed in our subsequent work.

The traffic data METR-LA is collected from loop detectors in the highway of Los Angeles County [12]. We chose data collected by 207 sensors for this experiment. Each sensor is regarded as a vertex in $\mathcal{V}_{\mathcal{G}}$ and we denote the data of every 512 consecutive timesteps as $\mathbf{X}$ for simulation. The METR-LA is divided into 100 FTVGS in total. That is, each $\mathbf{X}$ is a FTVGS with $N = 207$ vertices and $T = 512$ timesteps.

During the sampling stage, we follow the Algorithm 1. We reconstruct the original FTVGS by solving the optimization problem

$$\begin{aligned} \min_{\mathbf{F}_{\mathcal{G}}, \mathbf{F}_{\mathcal{T}}, \mathbf{E}} & \sum_i g(\lambda_i(\hat{\mathbf{X}}^T\hat{\mathbf{X}})) + \gamma_{\mathcal{G}}||\mathbf{W}_{\mathcal{G}} \odot \mathbf{F}_{\mathcal{G}}||_1 + \gamma_{\mathcal{T}}||\mathbf{W}_{\mathcal{T}} \odot \mathbf{F}_{\mathcal{T}}||_1 \\ \text{s.t.} \quad & \hat{\mathbf{X}} = \Psi_{\mathcal{G}}\mathbf{F}_{\mathcal{G}}, \\ & \hat{\mathbf{X}}^T = \Psi_{\mathcal{T}}\mathbf{F}_{\mathcal{T}}, \\ & \mathscr{P}_{\mathcal{S}}(\mathbf{X}) = \hat{\mathbf{X}} + \mathbf{E}, \mathscr{P}_{\mathcal{S}}(\mathbf{E}) = 0, \end{aligned} \tag{2}$$

where $\odot$ denotes the element-wise multiplication of two matrices, $\lambda_i(\hat{\mathbf{X}}^T\hat{\mathbf{X}})$ is the i-th eigenvalue of $\hat{\mathbf{X}}^T\hat{\mathbf{X}}$, $g(x) = \log(x^{1/2}+1)$, and $||\cdot||_1$ represents the ℓ_1-norm of the matrix. The two regularization terms cause each column of $\mathbf{F}_{\mathcal{G}}$ and $\mathbf{F}_{\mathcal{T}}$ to tend towards sparsity, which aligns with the assumption of bandlimited spectra in both the graph domain and the time domain. $\mathbf{W}_{\mathcal{G}}$ and $\mathbf{W}_{\mathcal{T}}$ signify the weighting matrices for weighted sparsity terms. $\gamma_{\mathcal{G}}$ and $\gamma_{\mathcal{T}}$ are regularization parameters. $\mathbf{E}$ is an error matrix.

The normalized root mean square error (NRMSE) is used to describe the error between signals $\mathbf{X}_a$ and $\mathbf{X}_b$

$$\mathrm{NRMSE}(\mathbf{X}_a, \mathbf{X}_b) = \frac{||\mathbf{X}_a - \mathbf{X}_b||_F}{||\mathbf{X}_a||_F}.$$

For each $\mathbf{X}$, we sample at different ratios, where ρ_{RC} represents the sampling ratio of randomly selected rows or columns, *i.e.*,

$$|\mathcal{I}| = \frac{\mathrm{Round}(\rho_{RC}N)}{N}, |\mathcal{J}| = \frac{\mathrm{Round}(\rho_{RC}T)}{T},$$

where $\mathrm{Round}(\cdot)$ represents rounding numbers to the nearest integer. $\rho_{\mathbf{X}_{RC}}$ represents the ratio of uniform sampling on $\mathbf{X}_{RC}$, *i.e.*, $|\mathcal{S}| = \mathrm{Round}(\rho_{\mathbf{X}_{RC}}|\mathcal{I}||\mathcal{J}|)/(|\mathcal{I}||\mathcal{J}|)$. And $\rho_{total} = |\mathcal{S}|/(NT)$ is the total sampling ratio. Under different sampling ratio settings, the NRMSE for the three reconstruction methods is calculated. The mean NRMSE of the 100 FTVGS are presented in Table II.

From the results in Table II, we can see that traditional MC methods are not very effective. Because these methods are based on the low-rank property and can accurately fill in the missing positions when the missing elements in the FTVGS follow a random distribution. However, the samples obtained through our subset random sampling violate their assumptions. The LIMC and our method (Eq. (2)) take into account the correlations between the rows and columns, demonstrating effectiveness and superiority in reconstructing the FTVGS from samples obtained through subset random sampling. This validates the conclusions of Theorem 1.

VI. CONCLUSIONS

In this work, based on the needs of practical applications, we propose a subset random sampling scheme for FTVGS, which involves uniform sampling from the submatrix $\mathbf{X}_{RC}$ instead of the entire $\mathbf{X}$. Theoretically, we proved sufficient conditions that guarantee the reconstruction of the original FTVGS with high probability under the subset random sampling scheme. Practically, we validated the effectiveness of reconstructing the original FTVGS from samples on a real-world dataset. Furthermore, for samples with randomly missing elements and some rows and columns are also randomly missing, we will consider targeted reconstruction algorithms to further reduce the NRMSE.

ACKNOWLEDGMENT

This work was supported in part by the Fundamental R&D Program of State Key Laboratory of Integrated Chips and Systems Under Grant KVH2327002.

APPENDIX A
PROOF OF THE LEMMA 1

From the SVD of $\mathbf{X}$ we know that $\mathbf{U} \in \mathbb{R}^{N\times r}$ is a matrix with orthogonal columns. Selecting the rows of $\mathbf{X}$ is essentially row sampling on $\mathbf{U}$.

Theorem 4.1 and Corollary 4.2 in [16] proved that $\mathrm{rank}(\mathbf{U}(\mathcal{I},:)) = r$ holds with probability at least $1-\delta$ if

$$|\mathcal{I}| \geq 3N\xi_1 \frac{\ln(2r/\delta)}{\epsilon^2}$$

for some $0 < \delta < 1$ and $0 < \epsilon < 1$, where $\xi_1 = ||\mathbf{U}||^2_{2,\infty}$.

According to Assumption 2, $\mu_1(\mathbf{X})r/N \geq ||\mathbf{U}||^2_{2,\infty} = \xi_1$, so we have

$$|\mathcal{I}| \geq 3r\mu_1(\mathbf{X}) \frac{\ln(2r/\delta)}{\epsilon^2}.$$

Further, we have $\mathbf{X}_R = \mathbf{U}(\mathcal{I},:)\Sigma\mathbf{V}^T$, where $\Sigma\mathbf{V}^T$ is a matrix with orthogonal rows. Then $\mathrm{rank}(\mathbf{X}_R) = r$ holds with probability at least $1-\delta$.

The SVD of $\mathbf{X}_R$ is

$$\mathbf{X}_R = \mathbf{U}(\mathcal{I},:)\Sigma\mathbf{V}^T = \mathbf{U}_R\Sigma_R\tilde{\mathbf{V}}_R^T\mathbf{V}^T = \mathbf{U}_R\Sigma_R\mathbf{V}_R^T, \tag{3}$$

TABLE I
DIFFERENCES BETWEEN THREE SCHEMES FOR RANDOM SAMPLING OF FTVGS.

	Random sampling of MC	CCS	Subset random sampling
Assumption	rank-r, $\{\mu_1, \mu_2\}$-incoherence		rank-r, smooth, $\{\mu_1, \mu_2\}$-incoherence
Complexity of required samples	$O(r^2 n_2 \log(n_2))$ [11] ($n_2 = \max\{N, T\}$)	$O(r^2 n_2 \log^2(n_2))$ [9]	$O(r^2 N \log^2(n))$

TABLE II
COMPARISON OF NRMSE FOR RECONSTRUCTION METHODS UNDER DIFFERENT SAMPLING RATIO SETTINGS.

ρ_{RC}	90%	80%	60%
$\rho_{\mathbf{X}_{RC}}$	90%	80%	60%
ρ_{total}	72.81%	51.37%	21.55%
SVT [13]	0.4542	0.6315	0.8421
TNNR [14]	0.4390	0.6006	0.8103
LIMC [15]	0.1199	0.1751	0.3531
Ours	**0.1171**	**0.1662**	**0.3442**

where $\mathbf{U}(\mathcal{I},:)\Sigma$ has the compact SVD $\mathbf{U}_R\Sigma_R\tilde{\mathbf{V}}_R^T$ with $\tilde{\mathbf{V}}_R \in \mathbb{R}^{r\times r}$ being an orthonormal matrix, and $\mathbf{V}_R = \mathbf{V}\tilde{\mathbf{V}}_R$.

The same argument yields $|\mathcal{J}| \geq 3T\xi_2 \ln(2r/\delta)/\epsilon^2$ for some $0 < \delta < 1$ and $0 < \epsilon < 1$, where $\xi_2 = ||\mathbf{V}_R||_{2,\infty}^2$ [16].

We have

$$\xi_2 = ||\mathbf{V}_R||_{2,\infty}^2 = ||\tilde{\mathbf{V}}_R^T\mathbf{V}||_{2,\infty}^2 = ||\mathbf{V}||_{2,\infty}^2 \leq \frac{\mu_2(\mathbf{X})r}{T}, \quad (4)$$

so we get

$$|\mathcal{J}| \geq 3r\mu_2(\mathbf{X})\frac{\ln(2r/\delta)}{\epsilon^2}.$$

Since $\mathbf{U}(\mathcal{I},:)\Sigma$ is a matrix with orthogonal columns with probability at least $1-\delta$. Then $\mathrm{rank}(\mathbf{X}_{RC}) = r$ holds with probability at least $(1-\delta)^2$.

APPENDIX B
PROOF OF THE LEMMA 2

According to the SVD of $\mathbf{X}_R$, Lemma 4.2 and Lemma 4.3 in [17] give

$$||\mathbf{U}_R||_{2,\infty} \leq \kappa(\mathbf{X})\frac{||\mathbf{X}_R^\dagger||_2}{||\mathbf{X}^\dagger||_2}\sqrt{\frac{\mu_1(\mathbf{X})r}{N}}, \frac{||\mathbf{X}_R^\dagger||_2}{||\mathbf{X}^\dagger||_2} \leq ||\mathbf{U}(\mathcal{I},:)^\dagger||_2.$$

Then we have

$$\begin{aligned}||\mathbf{U}_R||_{2,\infty} &\leq \kappa(\mathbf{X})||\mathbf{U}(\mathcal{I},:)^\dagger||_2\sqrt{\frac{\mu_1(\mathbf{X})r}{N}} \\ &= \kappa(\mathbf{X})\frac{1}{\sigma_{\min}(\mathbf{U}(\mathcal{I},:))}\sqrt{\frac{\mu_1(\mathbf{X})r}{N}}.\end{aligned}$$

According to the Lemma 3.4 in [18], $\sqrt{\frac{(1-\eta)|\mathcal{I}|}{N}} \leq \sigma_{\min}(\mathbf{U}(\mathcal{I},:))$ with failure probability at most $p = r\left[\frac{e^{-\eta}}{(1-\eta)^{1-\eta}}\right]^{\log r}, \eta \in [0,1)$. Then

$$\begin{aligned}||\mathbf{U}_R||_{2,\infty} &\leq \kappa(\mathbf{X})\sqrt{\frac{N}{(1-\eta)|\mathcal{I}|}}\sqrt{\frac{\mu_1(\mathbf{X})r}{N}} \\ &= \kappa(\mathbf{X})\sqrt{\frac{1}{1-\eta}}\sqrt{\frac{\mu_1(\mathbf{X})r}{|\mathcal{I}|}}\end{aligned}$$

holds with probability at least $1-p$. The Eq. (1) yields

$$\begin{aligned}||\mathbf{U}_{RC}||_{2,\infty} &= ||\mathbf{U}_R\tilde{\mathbf{U}}_{RC}||_{2,\infty} \\ &= ||\mathbf{U}_R||_{2,\infty} \leq \kappa(\mathbf{X})\sqrt{\frac{\mu_1(\mathbf{X})r}{(1-\eta)|\mathcal{I}|}}.\end{aligned}$$

Notice that $\mathbf{V}_{RC}^T = \Sigma_{RC}^{-1}\mathbf{U}_{RC}^T\mathbf{X}_{RC}$. Thus for any i,

$$||\mathbf{V}_{RC}^T\mathbf{e}_i||_2 \leq \frac{1}{\sigma_{\min}(\mathbf{X}_{RC})}||\mathbf{X}_{RC}\mathbf{e}_i||_2, \quad (5)$$

where $\mathbf{e}_i$ is a unit vector whose i-th element is 1. Since $\mathbf{X}_{RC} = \mathbf{X}(:,\mathcal{J}) = \mathbf{U}_R\Sigma_R(\mathbf{V}_R(\mathcal{J},:))^T$,

$$\begin{aligned}||\mathbf{X}_{RC}\mathbf{e}_i||_2 &\leq \sigma_1(\mathbf{U}_R)\sigma_1(\mathbf{X}_R)||(\mathbf{V}_R(\mathcal{J},:))^T\mathbf{e}_i||_2 \\ &= \sigma_1(\mathbf{X}_R)||(\mathbf{V}_R(\mathcal{J},:))^T\mathbf{e}_i||_2.\end{aligned} \quad (6)$$

Then combining (5) and (6), we have

$$\begin{aligned}||\mathbf{V}_{RC}||_{2,\infty} &\leq \frac{\sigma_1(\mathbf{X}_R)}{\sigma_{\min}(\mathbf{X}_{RC})}||(\mathbf{V}_R(\mathcal{J},:))||_{2,\infty} \\ &\overset{(a)}{\leq} \frac{\sigma_1(\mathbf{X}_R)}{\sigma_{\min}(\mathbf{X}_{RC})}||\mathbf{V}_R||_{2,\infty} \\ &= \kappa(\mathbf{X})\frac{\sigma_{\min}(\mathbf{X})}{\sigma_{\min}(\mathbf{X}_{RC})}||\mathbf{V}_R||_{2,\infty} \\ &= \kappa(\mathbf{X})\frac{||\mathbf{X}_{RC}^\dagger||_2}{||\mathbf{X}^\dagger||_2}||\mathbf{V}_R||_{2,\infty} \\ &\overset{(b)}{\leq} \kappa(\mathbf{X})\frac{1}{\sigma_{\min}(\mathbf{V}(\mathcal{J},:))\sigma_{\min}(\mathbf{U}(\mathcal{I},:))}||\mathbf{V}_R||_{2,\infty} \\ &\overset{(c)}{\leq} \frac{\kappa(\mathbf{X})}{1-\eta}\sqrt{\frac{NT}{|\mathcal{I}||\mathcal{J}|}}||\mathbf{V}_R||_{2,\infty} \\ &\leq \frac{\kappa(\mathbf{X})}{1-\eta}\sqrt{\frac{NT}{|\mathcal{I}||\mathcal{J}|}}\sqrt{\frac{\mu_2(\mathbf{X})r}{T}}\end{aligned}$$

with probability at least $(1-p)^2$, where Lemma 4.2 in [17] gives (a) and Lemma 3.4 in [18] gives (c). For (b), we have $\mathbf{X}_{RC} = \mathbf{U}(\mathcal{I},:)\Sigma(\mathbf{V}(\mathcal{J},:))^T$, so

$$||\mathbf{X}_{RC}^\dagger||_2 \leq ||((\mathbf{V}(\mathcal{J},:))^T)^\dagger||_2||\mathbf{X}^\dagger||_2||(\mathbf{U}(\mathcal{I},:))^\dagger||_2.$$

APPENDIX C
PROOF OF THE THEOREM 1

Since $\mathcal{I}$ and $\mathcal{J}$ are chosen uniformly from $\mathcal{V}_{\mathcal{G}}$ and $\mathcal{V}_{\mathcal{T}}$ based on Algorithm 1, according to Lemma 2 we have

$$||\mathbf{U}_{RC}\mathbf{V}_{RC}^T||_\infty \leq ||\mathbf{U}_{RC}||_{2,\infty}||\mathbf{V}_{RC}||_{2,\infty} \leq \frac{(\kappa(\mathbf{X}))^2}{(1-\eta)^{3/2}}\sqrt{\frac{\mu_1(\mathbf{X})\mu_2(\mathbf{X})Nr^2}{|\mathcal{I}|}}\sqrt{\frac{r}{|\mathcal{I}||\mathcal{J}|}}$$

for $\eta \in [0,1)$ with probability at least $(1-p)^2$.

By Theorem 2 in [8], we obtain that if

$$|\mathcal{S}| \geq \frac{32\beta(\kappa(\mathbf{X}))^4 r^2 N}{(1-\eta)^3}\mu_1(\mathbf{X})\mu_2(\mathbf{X})\frac{|\mathcal{I}|+|\mathcal{J}|}{|\mathcal{I}|}\log^2(2n) \quad (7)$$

for some $\beta > 1$, $\mathbf{X}_{RC}$ can be reconstructed with probability

$$1 - \frac{6\log n}{(|\mathcal{I}|+|\mathcal{J}|)^{2\beta-2}} - n^{2-2\beta^{1/2}}.$$

Based on Assumption 1, we know that $\| \nabla\mathbf{X}(i,j) \| < C$, so $|\mathbf{Q}^T\mathbf{X}(i,j)| < \infty$ and $\mathbf{X}\mathbf{D}_1(i,j) < \infty$. We define that

$$\| \nabla\mathbf{X}(i,j) \|_a = \sqrt{(\mathbf{Q}^T\mathbf{X}(i,j))^2 + (\mathbf{X}\mathbf{D}_1(i,j))^2 + a^2}$$

for $0 < a < \infty$, and $0 < \| \nabla\mathbf{X}(i,j) \|_a < \infty$. Similarly, we have

$$\left|\mathbf{Q}^T\mathbf{X}(i,j)\right|_a = \sqrt{(\mathbf{Q}^T\mathbf{X}(i,j))^2 + a^2} < \infty,$$

$$|\mathbf{X}\mathbf{D}_1(i,j)|_a = \sqrt{(\mathbf{X}\mathbf{D}_1(i,j))^2 + a^2} < \infty.$$

According to Section 5 in [19], we have

$$\nabla\left(\frac{\nabla\mathbf{X}(i,j)}{\| \nabla\mathbf{X}(i,j) \|_a}\right) = \frac{1}{\| \nabla\mathbf{X}(i,j) \|_a^3}(|\mathbf{X}\mathbf{D}_1(i,j)|_a^2\mathbf{X}(:,j)^T\mathbf{L}\mathbf{X}(:,j) + \left|\mathbf{Q}^T\mathbf{X}(i,j)\right|_a^2\mathbf{X}\mathbf{D}_2(i,j) + 2\mathbf{Q}^T\mathbf{X}(i,j)\mathbf{X}\mathbf{D}_1(i,j)\nabla\mathbf{X}(i,j))$$

and $\left|\nabla\left(\frac{\nabla\mathbf{X}(i,j)}{\|\nabla\mathbf{X}(i,j)\|_a}\right)\right| < \infty$. Thus, we can reconstruct $\mathbf{X}$ from $\mathbf{X}_{RC}$ by solving the TV inpainting model [19]. Then if Eq. (7) holds, we are able to reconstruct $\mathbf{X}$ from $\mathcal{S}$ with probability at least

$$(1-\delta)^2(1-p)^2 - \frac{6\log n}{(|\mathcal{I}|+|\mathcal{J}|)^{2\beta-2}} - n^{2-2\beta^{1/2}}.$$

REFERENCES

[1] G. Leus, A. G. Marques, J. M. Moura, A. Ortega, and D. I. Shuman, "Graph signal processing: History, development, impact, and outlook," *IEEE Signal Proc. Mag.*, vol. 40, no. 4, pp. 49–60, 2023.

[2] F. Grassi, A. Loukas, N. Perraudin, and B. Ricaud, "A Time-Vertex Signal Processing Framework: Scalable Processing and Meaningful Representations for Time-Series on Graphs," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 817–829, 2018.

[3] H. Sheng, H. Feng, J. Yu, F. Ji, and B. Hu, "Sampling theory of jointly bandlimited time-vertex graph signals," *Signal Process.*, vol. 222, p. 109 522, 2024.

[4] G. Ortiz-Jiménez, M. Coutino, S. P. Chepuri, and G. Leus, "Sampling and Reconstruction of Signals on Product Graphs," in *Proc. IEEE Global Conf. Signal, Inf. Process.*, Anaheim, CA, USA, Nov. 2018, pp. 713–717.

[5] D. Wei and Z. Yan, "Generalized sampling of multi-dimensional graph signals based on prior information," *Signal Process.*, p. 109 601, 2024.

[6] Z. Xiao, H. Fang, S. Tomasin, G. Mateos, and X. Wang, "Joint sampling and reconstruction of time-varying signals over directed graphs," *IEEE Trans. Signal Process.*, vol. 71, pp. 2204–2219, 2023.

[7] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[8] B. Recht, "A simpler approach to matrix completion.," *J. Mach. Learn. Res.*, vol. 12, no. 12, 2011.

[9] H. Cai, L. Huang, P. Li, and D. Needell, "Matrix completion with cross-concentrated sampling: Bridging uniform sampling and cur sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[10] O. Klopp, "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.

[11] T. Tong, C. Ma, and Y. Chi, "Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent," *J. Mach. Learn. Res.*, vol. 22, no. 150, pp. 1–63, 2021.

[12] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *Proc. ICLR*, 2018, pp. 1–26.

[13] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[14] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2117–2130, 2012.

[15] W. Wang and J. Wang, "Enhancing matrix completion using a modified second-order total variation," *Discrete Dyn. Nature Soc.*, vol. 2018, no. 1, p. 2 598 160, 2018.

[16] I. C. Ipsen and T. Wentworth, "The effect of coherence on sampling from matrices with orthonormal columns, and preconditioned least squares problems," *SIAM J. Matrix Anal. Appl.*, vol. 35, no. 4, pp. 1490–1520, 2014.

[17] H. Cai, K. Hamm, L. Huang, and D. Needell, "Robust cur decomposition: Theory and imaging applications," *SIAM J. Imag. Sci.*, vol. 14, no. 4, pp. 1472–1503, 2021.

[18] J. A. Tropp, "Improved analysis of the subsampled randomized hadamard transform," *Adv. Adapt. Data Anal.*, vol. 3, no. 01n02, pp. 115–126, 2011.

[19] J. Shen and T. F. Chan, "Mathematical models for local nontexture inpaintings," *SIAM J. Appl. Math.*, vol. 62, no. 3, pp. 1019–1043, 2002.

An Effective Contextualized Automatic Speech Recognition Approach Leveraging Self-Supervised Phoneme Features

Li-Ting Pai[1], Yi-Cheng Wang[1], Bi-Cheng Yan[1], Hsin-Wei Wang[1], Jia-Liang Lu[1],
Chi-Han Lin[2], Juan-Wei Xu[2], Berlin Chen[1]
Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
E.SUN Financial Holding Co., Ltd
E-mail: {61147095s, yichengwang, 80847001s, hsinweiwang, jialianglu, berlin}@ntnu.edu.tw;
E-mail: {finalspaceman-19590, weixu-22681}@esunbank.com

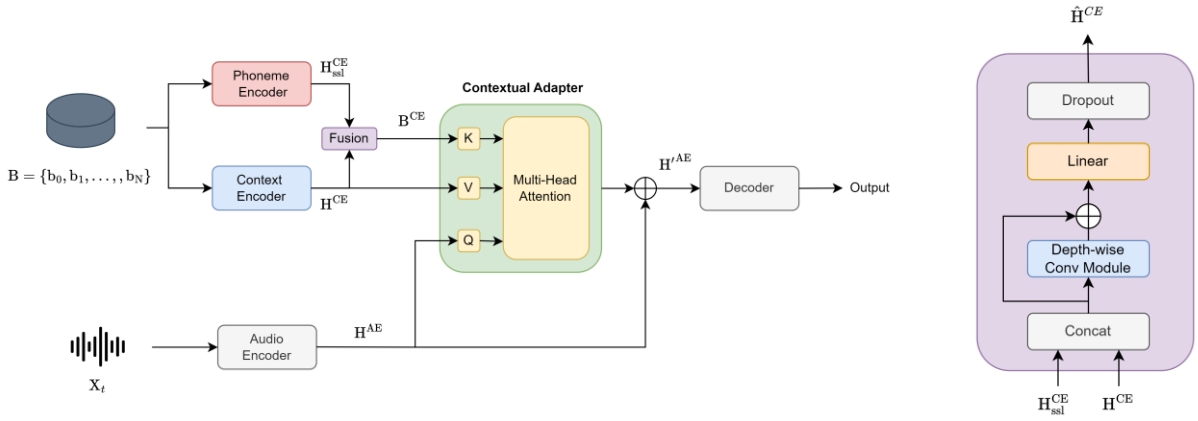
Abstract— **Years of scholarly efforts have led to extensive studies on end-to-end automatic speech recognition (E2E ASR), now demonstrating robust performance in everyday applications such as voice assistants, transcription services, and many others. However, E2E ASR struggles to recognize domain-specific phrases, such as keywords or name entities. To address this, contextualized ASR (CASR) has been developed to improve keyword recognition accuracy by incorporating specific contextual information, represented by a keyword list, into the ASR model. Despite their effectiveness, CASR systems still fall short in distinguishing between keywords with similar sounds, as well as generalizing to uncommon keyword pronunciations. Previous studies have focused primarily on enriching keyword representations by integrating keyword phoneme features derived from a simple sequence encoder with keyword grapheme features to overcome these obstacles. However, such phoneme representations are insufficient, as human pronunciation varies in different contexts, involving phenomena like linking and variations. In this paper, we argue that integration of more fine-grained phoneme features instrumental to accurate keyword recognition in CASR. To this end, we propose leveraging a self-supervised learning (SSL) phoneme encoder to provide more subtle phonemic details of keywords, effectively addressing these variations and alleviating the phonetic confusion between keywords. A series of experiments conducted on the SlideSpeech benchmark dataset demonstrates the effectiveness of our approach in alleviating keyword phonemic confusion and enhancing out-of-domain keyword recognition.**

INTRODUCTION

End-to-end (E2E) automatic speech recognition (ASR) [1] has garnered significant interest from both research communities and industrial applications in recent years due to its streamlined design and scalability compared to conventional hybrid DNN-HMM models. Although E2E ASR models demonstrate impressive recognition abilities for common words, they struggle with rare words, including keywords like personal names [2]. This difficulty arises from the imbalance of word distributions in the training set. Accurate keyword recognition is crucial for downstream tasks like natural language processing, making the improvement of keyword recognition accuracy essential for advancing E2E ASR models in real-world applications.

In contrast to machines, humans, with their rich sensory inputs, can easily recognize unfamiliar phrases by leveraging various clues from multiple modalities such as visual cues, body language, and environmental sounds. This ability enables them to adapt swiftly to different scenarios, including keynote speeches, financial discussions, and medical consultations. However, E2E ASR models rely solely on speech signals and lack these additional perceptual inputs, making it challenging for them to recognize domain-specific phrases. To address this limitation, researchers are developing Contextualized ASR (CASR) systems [3] that incorporate knowledge beyond speech signals. This approach enhances the understanding capabilities of ASR models and adapts them to diverse contexts, leading to more accurate and reliable speech recognition.

CASR research can be broadly categorized into three groups: shallow fusion, deep biasing, and prompting. 1) Shallow fusion [3, 4, 5] uses weighted finite state transducers (WFST) to combine the output of external *N*-gram language models with the output of the ASR model. While this can improve ASR performance, it requires careful tuning of the fusion weights for optimal results. 2) Deep biasing [6, 7, 8, 9] employs cross-attention [10] mechanisms to integrate keyword representations into the hidden layers of ASR models, effectively improving the recognition of domain-specific terms. 3) Prompting [11, 12, 13] involves providing contextual information as an input sequence to the ASR model. Among these CASR methods, the contextual adapter (CA) [14] stands out for its simplicity and effectiveness. It integrates a list of pre-defined keywords into the ASR model using a cross-attention mechanism. Notably, CA employs an adapter-style training method, necessitating the tuning of only about 3% of the model weights. This approach offers a significant advantage over other CASR methods. Despite their effectiveness, CA models still face challenges in distinguishing between keywords with similar sounds and



(a) Attention-based encoder-decoder (AED) model with contextual adapter (b) Fusion mechanism

Fig 1. (a) The architecture of phoneme-aware contextual adapter with attention-encoder-decoder architecture. (b) Fusion mechanism of the context feature and SSL feature.

generalizing to uncommon keyword pronunciations, especially in scenarios with limited resources.

Previous studies [15, 16, 17] have primarily focused on enriching keyword representations by integrating keyword phoneme features derived from a simple sequence encoder with keyword grapheme features to overcome these obstacles. However, such phoneme representations are insufficient due to limited keyword variation in the training corpus and the fact that human pronunciation varies in different contexts, involving linking and variations. For example, in English, the pronunciation of "black cab" may result in a linking sound, making it sound like "bla-cab," while "red apple" may sound like "re-dapple."

To address these challenges, recent advancements in self-supervised learning (SSL) offer promising solutions, particularly for scenarios with limited resources. SSL involves pre-training a model on large amounts of unlabeled data to obtain rich feature representations. This method does not rely on a large amount of manually labeled data but instead uses the intrinsic structure within the data to learn useful representations.

In the field of CASR, research on enhancing contextual adapters through phoneme-aware encoding using self-supervised learning is limited. In this paper, we argue that incorporating more fine-grained phoneme features is necessary to improve keyword recognition. We propose utilizing a self-supervised learning (SSL) phoneme encoder to provide detailed phoneme information, which can effectively address variations and reduce phonetic confusion between keywords. This approach aims to enhance the accuracy and reliability of CASR systems in recognizing domain-specific terms and uncommon keyword pronunciations.

In summary, our contributions are at least three-fold:

- **SSL Phoneme Encoder Integration:** We introduce a self-supervised learning (SSL) phoneme encoder to enhance keyword phoneme representation in CASR systems, capturing fine-grained phonemic details to improve recognition accuracy.
- **Phonetic Confusion Reduction:** Our approach significantly reduces phonetic confusion between keywords with similar sounds, leveraging detailed phonemic features for more precise keyword recognition.
- **Enhanced Out-of-Domain Generalization:** Experiments on the SlideSpeech benchmark dataset show the effectiveness of our methods in improving both in-domain and out-of-domain keyword recognition, demonstrating robustness and versatility.

RELATED WORK

Recently, contextual adapter (CA) [14] has demonstrated considerable success; however, its performance often declines when phoneme confusion increases in rare word lists. To address this issue, a phoneme-based encoding method was introduced to enhance the recognition of words with irregular pronunciations [15, 16, 17]. Despite its benefits, this method still lacks context consideration when encoding phonemes. To overcome this limitation, we propose combining detailed phoneme-aware features generated by the self-supervised learning (SSL) model XPhoneBERT [18].

XPhoneBERT was originally developed to enhance text-to-speech (TTS) tasks by learning robust phoneme representations. We propose utilizing XPhoneBERT to CASR models. XPhoneBERT is a pioneering multilingual model with a BERT-base architecture [19], trained using the masked prediction objective on a dataset of 330 million phoneme-level sentences from nearly 100 languages. This extensive training enables it to generate high-quality phoneme representations across various languages. Key features of XPhoneBERT include multilingual capability and a focus on phoneme-level data, capturing fine-grained phonetic nuances crucial for both TTS and ASR tasks. We adapt XPhoneBERT for contextualized ASR models by incorporating it as a phoneme encoder.

Table 1. In-domain testing: WER, K-WER and NK-WER results for History and Computer Science test datasets.

Domain	Model	WER	K-WER	NK-WER
Computer Science	Conformer	13.75	13.21	**13.90**
	+ CA	14.80	14.41	14.90
	+ SSL feature	**13.61**	**9.83**	14.64
History	Conformer	**13.90**	15.18	**13.66**
	+ CA	14.86	16.57	14.55
	+ SSL feature	14.80	**14.38**	14.86

PROPOSED METHOD

Before we delve into our proposed methods, we will first describe the architecture of our backbone E2E ASR model, which is an attention-based encoder-decoder (AED) network. Next, we introduce the contextual adapter (CA) methods. Finally, we elaborate on our proposed SSL phoneme features enriched CA approach.

A. Backbone E2E ASR Model

An AED model generally consists of an encoder network and a decoder network, as shown in Figure 1(a) in the gray part. Given an audio signal $\mathrm{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)$ of length T, represent the acoustic feature vectors, and $\mathbf{y} = (y_1, y_2, \dots, y_M)$ of length M be the corresponding subword sequence. The encoder network, $\mathrm{Enc}_{\mathrm{AE}}(\cdot)$, processes a sequence of acoustic feature vectors, X, and produces a sequence of high-level acoustic representations, $\mathrm{H}^{\mathrm{AE}} = \mathrm{Enc}_{\mathrm{AE}}(\mathbf{x}_t)$. The decoder network, $\mathrm{Dec}(\cdot)$, then fuses these acoustic embeddings H^{AE} with the previously decoded text tokens $\mathbf{y}_{1:m} = (y_1, y_2, \dots, y_m)$ using a cross-attention mechanism. The output probability of a possible upcoming token $\mathbf{y}_{m+1}$ can be derived by

$$\mathbf{y}_{m+1} = \mathrm{Dec}(\mathrm{H}^{\mathrm{AE}}, \mathbf{y}_{1:m}). \quad (1)$$

B. Contextual Adapter

Contextual adaptation (CA) is a mechanism designed to enhance the recognition accuracy of ASR models for rare or context-specific words by incorporating additional contextual information into the ASR model. CA introduces two additional components to the original ASR model: a context encoder and an attention-based adapter.

The Context Encoder, denoted as $\mathrm{Enc}_{\mathrm{CE}}(\cdot)$, processes a keyword list, consist of N keywords $\mathrm{B} = \{\mathbf{b}_0, \mathbf{b}_1, \dots, \mathbf{b}_N\}$ into the contextual representations $\mathrm{H}^{\mathrm{CE}} = \mathrm{Enc}_{\mathrm{CE}}(\mathrm{B})$.

The attention-based adapter then integrates these contextual embeddings H^{CE} into the ASR model using a multi-head cross-attention mechanism. The attention process can be derived as:

$$\mathrm{B}^{\mathrm{CE}} = \mathrm{softmax}\left(\frac{\mathrm{H}^{\mathrm{AE}}\mathrm{W}^{\mathrm{Q}}(\mathrm{H}^{\mathrm{CE}}\mathrm{W}^{\mathrm{K}})^{\mathrm{T}}}{\sqrt{d}}\right)\mathrm{H}^{\mathrm{CE}}\mathrm{W}^{\mathrm{V}}, \quad (2)$$

where $\mathrm{W}^{\mathrm{Q}}, \mathrm{W}^{\mathrm{K}}, \mathrm{W}^{\mathrm{V}}$ are trainable weight matrices, and d is the dimensionality of the embeddings. This biasing matrix B^{CE} is then used to contextualize the original acoustic embeddings, resulting in $\acute{\mathrm{H}}^{\mathrm{AE}} = \mathrm{H}^{\mathrm{AE}} + \mathrm{B}^{\mathrm{CE}}$. By incorporating the contextual information in this manner, the ASR model can more accurately recognize and transcribe rare words that are crucial for understanding the input audio sequence.

C. SSL Phoneme Features Enriched Contextual Adapter

We leverage phoneme features extracted from the SSL phoneme encoder, XPhoneBERT, alongside the context embedding H^{CE} generated by the context encoder to produce phoneme-aware context embeddings, as illustrated in Fig. 1(a).

The phoneme encoder, denoted as $\mathrm{Enc}_{\mathrm{pho}}(\cdot)$, extracts SSL features represented as $\mathrm{H}^{\mathrm{CE}}_{\mathrm{ssl}} = \mathrm{Enc}_{\mathrm{pho}}(\mathrm{B})$. We then concatenate these phoneme features $\mathrm{H}^{\mathrm{CE}}_{\mathrm{ssl}}$ with the context features H^{CE} to form a combined feature representation $\mathrm{H}^{\mathrm{CE}}_{\mathrm{cat}} = \mathrm{Concat}(\mathrm{H}^{\mathrm{CE}}_{\mathrm{ssl}}, \mathrm{H}^{\mathrm{CE}})$.

Next, we apply a depth-wise convolution to this combined representation to capture information from adjacent dimensions, as investigate in [20], resulting in $\mathrm{H}^{\mathrm{CE}}_{\mathrm{dw}} = \mathrm{DepthwiseConv}(\mathrm{H}^{\mathrm{CE}}_{\mathrm{cat}})$. The output of this convolution is summed with the original concatenated features, followed by a linear projection to produce the final phoneme-aware context embedding $\hat{\mathrm{H}}^{\mathrm{CE}} = \left(\mathrm{H}^{\mathrm{CE}}_{dw} + \mathrm{H}^{\mathrm{CE}}_{\mathrm{cat}}\right)\mathrm{W}$, as shown in Figure 1(b). These phoneme-aware context embeddings are subsequently used to generate the key embedding of the attention-based adapter, which is jointly optimized with the ASR model.

EXPERIMENTAL SETUP

A. Dataset and Evaluation Metrics

The SlideSpeech corpus [21] is a comprehensive English audio-visual dataset containing over 1,000 hours of slide presentations. It includes real-time synchronized slides, pre-processed Optical Character Recognition (OCR) results, and extracted keywords corresponding to the slides, making it an ideal dataset for evaluating the CASR task. SlideSpeech spans 22 diverse domain categories, with Computer Science and History being the largest. Therefore, our experiments focused on these two domains, as they are well-represented in the corpus. For the Computer Science dataset, we divided it into a training set with 109 hours, a development set with 20 hours, and a test set with 32 hours. Similarly, for the History dataset, the training set contained 173 hours, the development set 20 hours, and the test set 47 hours.

During training, we dynamically generated a keyword list that included keywords randomly selected from multiple utterances within

Table 2. Out of domain generalization: WER, K-WER and NK-WER results for History and Computer Science test datasets.

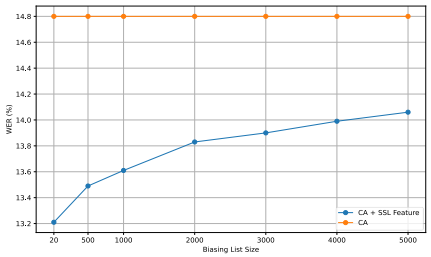
Training Domain	Testing Domain	Model	WER	K-WER	NK-WER
History	Computer Science	Conformer	16.59	16.15	**16.65**
		+ CA	17.48	16.95	17.57
		+ SSL feature	**16.45**	**13.79**	16.87
Computer Science	History	Conformer	**21.06**	**19.75**	**21.28**
		+ CA	22.42	21.32	22.60
		+ SSL feature	22.22	21.92	22.27

the batch as distractors, with the total number of distractors set to 200 for all experiments. "Distractors" refer to keywords that do not belong to the current utterance, used to simulate contextual interference in real-world scenarios. We adapted the keywords provided by the SlideSpeech corpus. Since training with the entire keyword list is memory-intensive, we dynamically generated a subset of the keyword list during training. This subset included keywords randomly selected from the full list, the reference keywords in the utterance, and a special token <OOV>, which serves as a fallback when no relevant context words are available. During testing, the keyword lists consist of keywords from the target utterance along with distractors, with the total number of distractors set to 1000. We evaluate contextual biasing performance using word error rate (WER), keyword word error rate (K-WER), and non-keyword word error rate (NK-WER). K-WER is calculated as the number of incorrect keywords in the keyword list divided by the total number of keywords in the test set. NK-WER is the number of incorrect non-keywords in the keyword list divided by the total number of non-keywords in the test set.

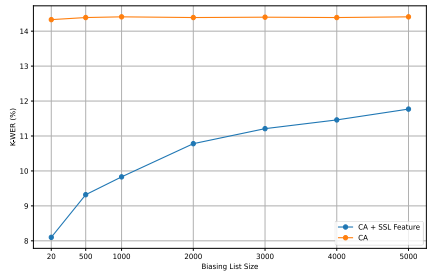
B. Baseline Models

We used a Conformer-Transformer model as the backbone for long-tailed speech recognition experiments. The network consists of a Conformer encoder [22] and a Transformer decoder [23] (denoted by Conformer for short). The Conformer encoder consists of 12 blocks, each with 2,048 hidden units and 8 attention heads. The Trans-former decoder network includes 6 blocks, each also with 2,048 hidden units. Our proposed method is compared against the iconic contextual adapter (CA) method, which serves as a strong baseline in this study. The contextual adapter is equipped with a cross-attention layer having 8 heads, each with a size of 64. The input features are audio features from the Slidepeech dataset, with each audio segment converted into 80-dimensional Mel spectrograms and normalized at the frame level. During training, we use the Adam optimizer with an initial learning rate set to 0.001, employing a learning rate decay strategy. The batch size is set to 64, and the model is trained for 35 epochs.

The components of the Contextual Adapter are as follows: The phoneme encoder uses XPhoneBERT [1], which converts phoneme sequences into 768-dimensional phonetic representations. The encoder



(a) WER

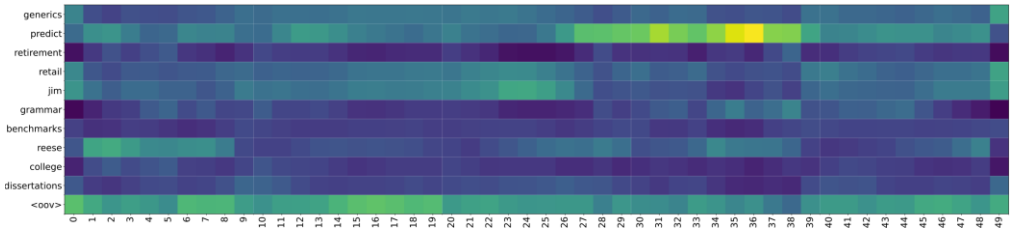


(b) K-WER

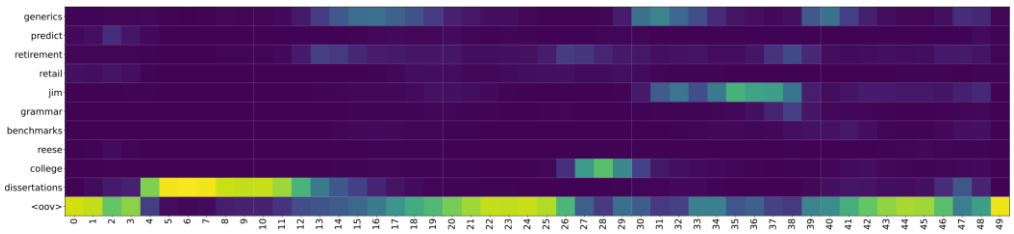
Fig 2. The impact of WER and K-WER under different keyword list sizes.

includes an embedding layer that maps input tokens to 512-dimensional embeddings, an OOV embedding layer that converts 512-dimensional embeddings to a single dimension to handle unknown tokens, and a padding embedding layer similar to the OOV embedding but used for padding tokens. The RNN encoder is composed of a bidirectional LSTM with two layers, each with 512 hidden units, where the final LSTM output is projected to 512 dimensions. A dropout layer with a rate of 0.1 is applied to prevent overfitting, and layer normalization is applied to intermediate features. Depth-wise convolution fusion is performed with a 1D convolutional layer with 1280 groups for channel feature fusion, followed by a projection layer that projects the fused features to 512 dimensions. The adapter includes a custom multi-headed attention mechanism, which consists of linear layers for query, key, and value transformations, each outputting 512 dimensions. A dropout rate of 0.1 is applied to the attention mechanism, and a projection layer projects the attention outputs back to 512 dimensions.

[1] https://huggingface.co/vinai/xphonebert-base



(a) Without SSL feature



(b) With SSL features

Fig 3. Visualization of the attention scores within the adapter, "dissertations" and "college" are the target keywords, in the exemplar utterance: "dissertations and college and alumni magazines." In (a), the contextual adapter is easily confused by phonetically similar keywords. In contrast, (b) demonstrates that our SSL feature-enriched contextual adapter effectively distinguishes keywords from other distractors.

To encourage further research, we will make our code publicly available for benchmarking and replication experiments.

Experimental Results

We begin our experiments by evaluating the performance on the in-domain dataset, comparing baseline Conformer models with and without contextual adapter (CA) and Self-Supervised Learning (SSL) features. Next, we assess the out-of-domain generalization adaptability of the CASR model by training and testing on different domain datasets. We then analyze how variations in the size of the keyword list affect the WER and K-WER. Finally, we visualize attention maps to understand the alignment between audio features and keywords, both with and without SSL features.

A. Main Results

Table 1 presents the in-domain testing results for Computer Science and History subset. The baseline Conformer model shows decent performance, but incorporating the CA results in a slight increase in error rates. In contrast, adding SSL features significantly improves performance, particularly in reducing the K-WER. These results indicate that while CA alone may not reduce errors, combining CA with SSL features enhances performance, especially in recognizing keywords, thus improving the overall accuracy of the ASR model in the Computer Science and History domain.

B. Cross-domain Generalizations

Table 2 highlights significant differences in performance across cross-domain datasets with various model configurations. When training on History domain dataset and testing on Computer Science domain dataset, the baseline model shows higher error rates, which increase further with CA. However, adding SSL features reduces these rates. Similarly, when training on Computer Science data and testing on History domain dataset, the baseline model shows high error rates, which increase with CA and decrease slightly with SSL features. These results suggest that while contextual adapters and SSL features can improve performance on cross-domain datasets, they may also lead to higher error rates in some cases.

C. Impact of Different Size of the Keyword list

Figure 2(a) shows that the WER decreases significantly as the size of the keyword list decreases. For instance, the WER is lower with a smaller keyword list, highlighting the effectiveness of our method in accurately handling rare words. Figure 2(b) illustrates that the K-WER increases as the size of the keyword list grows. This suggests that while larger keyword lists provide more context, they also introduce complexity, negatively impacting keyword recognition performance. Therefore, a balanced keyword list size is crucial for optimizing the trade-off between context and complexity, enhancing the accuracy of specific keywords.

D. Qualitative Analysis

Figure 3 visually depicts the corresponding results. On the X-axis stands for time stamps while the Y-axis displays keywords, with "dissertations," "collect" being the target keywords. Brighter pixels in the attention map signify higher attention weights assigned by the model to each rare word at a given time stamp. The figure illustrates the attention relationship between audio features and keywords, demonstrating the effectiveness of our method in integrating phoneme and contextual information. Without SSL features, the bias word does not effectively align with the frames. In contrast, with SSL features, the bias word aligns well with the frames, indicating better model convergence.

CONCLUSIONS

In this paper, we presented a novel method to enhance the performance of ASR systems by integrating fine-grained phoneme SSL features into a contextual adapter. Our approach leverages both phoneme and contextual information to improve the accuracy of recognizing context-specific words. By incorporating SSL models to extract phoneme representations from the keyword list, we provided dual representations that include both phonetic and grapheme information. This method significantly improves keyword recognition accuracy by comprehensively utilizing external knowledge and contextual information. Furthermore, by employing SSL models to extract phoneme features, we enhanced the accuracy of recognizing keywords. In future work, we will explore applying our approach to a wider range of datasets, investigate real-time ASR applications.

ACKNOWLEDGEMENT

This work was supported in part by E.SUN Bank under Grant Numbers 202308-NTU-02 and 202408-NTU-02. Any findings and implications in the paper do not necessarily reflect those of the sponsor.

REFERENCES

[1] J. K. Chorowski et al., "Attention-based models for speech recognition," in *Proc. Neural Information Processing Systems*, vol. 28, 2015.

[2] D. Bahdanau et al., "End-to-end attention-based large vocabulary speech recognition," in *Proc. ICASSP. IEEE*, 2016, pp. 4945-4949.

[3] I. Williams et al., "Contextual speech recognition in end-to-end neural network systems using beam search," in *Proc. Interspeech*, 2018, pp. 2227-2231.

[4] D. Zhao et al., "Shallow-fusion end-to-end contextual biasing," in *Proc. Interspeech*, 2019, pp. 1418-1422.

[5] S. Kim et al., "Improved neural language model fusion for streaming recurrent neural network transducer." in *Proc*. *ICASSP*, 2021, pp. 7333-7337.

[6] G. Pundak et al., "Deep context: end-to-end contextual speech recognition," in *Proc. IEEE SLT*, 2018, pp. 418-425.

[7] M. Han et al., "Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection," in *Proc. ICASSP*, 2022.

[8] D. Le et al., "Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion," *arXiv preprint arXiv:2104.02194*, 2021.

[9] F. J. Chang et al., "Context-aware transformer transducer for speech recognition," in *Proc. IEEE ASRU*, 2021, pp. 503-510.

[10] A. Vaswani et al., "Attention is all you need." in *Proc*. *NeurIPS*, 2017.

[11] X. Yang et al., "PromptASR for contextualized ASR with controllable style," in *Proc. ICASSP. IEEE*, 2024, pp. 10536-10540.

[12] G. Yang et al., "MaLa-ASR: Multimedia-Assisted LLM-Based ASR," *arXiv preprint arXiv:2406.05839*, 2024.

[13] Li, Yuang, et al. "Using Large Language Model for End-to-End Chinese ASR and NER." *arXiv preprint arxiv:2401.11382*, 2024.

[14] K. M. Sathyendra et al., "Contextual adapters for personalized speech recognition in neural transducers," in *Proc. ICASSP. IEEE*, 2022, pp. 8537-8541.

[15] R. Pandey et al., "PROCTER: pronunciation-aware contextual adapter for personalized speech recognition in neural transducers," in *Proc. ICASSP. IEEE*, 2023, pp. 1-5.

[16] Z. Chen et al., "Joint grapheme and phoneme embeddings for contextual end-to-end ASR," in *Proc. Interspeech*, 2019, pp. 3490-3494.

[17] A. Bruguier et al., "Phoebe: Pronunciation-aware contextualization for end-to-end speech recognition," in *Proc. ICASSP*, 2019, pp. 6171-6175.

[18] L. T. Nguyen et al., "XPhoneBERT: A pre-trained multilingual model for phoneme representations for text-to-speech," in *Proc. Interspeech*, 2023.

[19] J. Devlin et al., "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc*. *NAACL-HLT*, 2019, pp. 4171-4186.

[20] C. Si et al., "Inception Transformer," in *Proc. NeurIPS*, vol. 30, 2022, pp. 23495-23509.

[21] H. Wang et al., "SlideSpeech: A large scale slide-enriched audio-visual corpus," in *Proc. ICASSP. IEEE*, 2024, pp. 11076-11080.

[22] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[23] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

Real-Time Monophonic Dual-Pitch Extraction Model

Ngoc-Son Tran*, Pei-Chin Hsieh*, Yih-Liang Shen†, Yen-Hsun Chu‡ and Tai-Shi Chi*
* Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan
† Institute of Communications Engineering, National Yang Ming Chiao Tung University, Taiwan
‡ Realtek Semiconductor Corp., Hsinchu, Taiwan
E-mail: ngocsontran19398@gmail.com; wren.ee12@nycu.edu.tw; dennis831209@gmail.com;
yenhsun.chu@realtek.com; tschi@nycu.edu.tw

***Abstract*—Pitch estimation is a fundamental aspect of audio signal processing with applications in music information retrieval (MIR), speech analysis and more. This paper proposes a new model for real-time dual-pitch estimation to tackle the problem of estimating two pitches at the same time in duet songs. The model is designed for real-time processing, frame by frame and with streamlined components for efficiency. We also propose data augmentation methods for simulating duet singing scenarios. Our model is robust in single and dual-pitch scenarios through experiments with various datasets and metrics. Our results contribute to the pitch estimation techniques and provide a practical solution for real-time audio processing in MIR, speech analysis and beyond.**

I. Introduction

Pitch estimation has long been a pivotal topic in the domain of music information retrieval (MIR), finding applications in diverse areas such as melody extraction, and speech/music transcription. The existing landscape of pitch estimation methods can be broadly categorized into two types: those rooted in digital signal processing (DSP) and those driven by data. DSP-based methods encompass widely recognized algorithms like RAPT [1], YIN [2], and pYIN [3], which leverage time-domain periodicity features for pitch estimation. On the other hand, data-driven approaches, exemplified by neural network (NN)-based methods such as CREPE [4] and the hybrid model [5], rely on both time-domain and frequency-domain features, yielding superior performance compared with DSP-based methods.

Despite the effectiveness of these algorithms in pitch estimation tasks, their suitability for real-time applications remains limited — an imperative in scenarios like speaker identification [6], voice analysis [7], and music analysis [8]. Notably, the karaoke grading system should heavily rely not only on accurate single-pitch estimation for solo songs, but also on dual-pitch estimation for duet songs, to assess users' performance [9]. While the YIN algorithm stands out for its low computational overhead, due to its reliance on auto-correlation, it is faced with the challenge of handling two pitches simultaneously [2]. Recently, there has been a surge in the adoption of NN approaches for multi-pitch estimation. LATE/DEEP [10] uses a stack of 2D convolution layers to analyze the harmonic structure. RNN-BLSTM [11] uses the recurrent networks to capture the temporal information. However, a notable limitation of these methods lies in their inability to fulfill the real-time execution criterion. Motivated by this gap, our research endeavors to develop a real-time dual-pitch extraction algorithm for the karaoke grading system.

Our methodology builds upon the spectral branch in a previous study [12] while refining its efficiency. With the focus on efficiency, we adapt specific blocks to meet our objectives, prioritizing real-time performance on embedded systems with limited computational power. Concurrently, we leverage the log-scale spectrogram—an efficient representation to decipher the harmonic structure [5]—as the key feature input for our proposed NN model. Central to the architecture of the proposed model is the harmonic aware block, crucial for capturing underlying harmonic patterns essential for accurate pitch extraction [12]. Throughout the design process, careful attention is given to parameter optimization to minimize computational overhead while maintaining performance.

The rest of the paper is organized as follows. Section II presents the utilized features and details the proposed model. Section III outlines the datasets, and offers insights into the experimental configurations. Section IV delves into the experimental results and ensuing discussions. Lastly, Section V encapsulates the essence of our study through the concluding remarks.

II. Proposed Model

A. Input Features

Temporal and spectral features often serve as fundamental representations to NN models for audio applications [13]. To optimize computational efficiency, we adopted the log-scaled short-time Fourier transform (STFT)-based spectrogram, as the primary input representation for our model as in [5]. In addition to using the log-frequency representation, we applied the logarithmic scale to the magnitude of the log-spectrogram. As for the frequency resolution and the frequency coverage, we set 24 bins per octave, and took the first 176 frequency bins, covering 31 Hz to 4857 Hz, as the input representation to our model. The frequency coverage of 31 Hz to 4857 Hz was selected by following the setting in [14] to cover the pitch range of singing voices.

B. Proposed System Architecture

The original spectral branch proposed in [12] utilizes information from previous and subsequent frames to predict pitch of the current frame. However, for the real-time task, our system adopts a frame-based approach, necessitating adjustments to

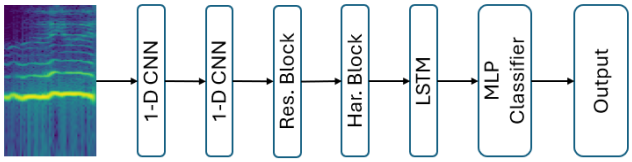


Fig. 1: The overall architecture of our proposed model.

the spectral branch architecture, as shown in Fig. 1. Specifically, we replaced all 2-D convolutional blocks in the original spectral branch with 1-D convolutional blocks and eliminated several blocks to streamline the model complexity. Details of the parameter settings in our model are outlined in Table I.

The model begins with two convolution blocks. Each block contains one layer of 1-D convolution with the kernel size of 8 and the stride of 1. The convolution layer is followed by batch normalization [15], ReLU activation, and the Maxpool layer with the pool size of 2 and the stride of 2. After the convolution blocks, there are the residual connection block (Res. Block) and the Harmonic-aware block (Har. Block). The residual connection block is shown in Fig. 2, and details of the Harmonic-aware block is illustrated in Fig. 3. At the end of the model, the layers of LSTM and MLP are deployed with the hidden size of 256 and 512, respectively. The zero-padding technique is employed at each convolution layer to maintain consistent input and output sizes.

The Harmonic-aware block is designed to capture specific harmonics based on Equation 1.

$$d_{i,j} = log_{2^{1/Q}}(jf_0) - log_{2^{1/Q}}(if_0) = Qlog_2(\frac{j}{i}) \quad (1)$$

where i, j represent the indexes of the harmonics, $d_{i,j}$ signifies the number of bins between i-th and j-th harmonics, Q denotes the number of bins per octave, and f_0 stands for the fundamental frequency. We set the parameter $Q = 24$. This setting yields interval computations between the 2nd, 3rd, and 4th to the 1st harmonic, resulting in approximately 24, 38, and 48 frequency bins, respectively. Two Max-pooling layers with the stride of 2 are used before the Harmonic-aware block, which implies the adoption of a $4\times$ downsampling. Therefore,

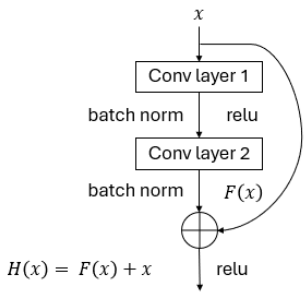


Fig. 2: The architecture of the residual block (Res. Block) in the proposed model.

TABLE I: Architecture configuration of the model. The Res. Block denotes for residual block, Har. block denotes for harmonic aware block.

Layer	Configuration	Input	Output
Conv1D	8 kernels: size 3, stride 1	1 x 176	8 x 176
MaxPool1D	pool 2, stride 2	8 x 176	8 x 88
Conv1D	8 kernels: size 3, stride 1	8 x 88	8 x 88
MaxPool1D	pool 2, stride 2	8 x 88	8 x 44
Res. Block	Conv1D 8 kernels: size 1, stride 1	8 x 44	8 x 44
	Conv1D 8 kernels: size 3, stride 1		
Har. Block	Conv1D 8 kernels: size 1, stride 1	8 x 44	1 x 44
	Conv1D 8 kernels: size 6/10/12, stride 1		
	Conv1D 1 kernel: size 1, stride 1		
MaxPool1D	pool 2, stride 2	1 x 44	1 x 22
LSTM	hidden size 256	1 x 22	1 x 256
MLP layer	hidden size 512	1 x 256	127

three parallel 1-D convolution kernels with sizes 6, 10, and 12 are employed to identify harmonic structures and enhance classification accuracy. The idea of using 1-D convolution kernels with fixed sizes to decipher harmonics on the log-spectrogram was proposed and evaluated in the previous study [5].

C. Dual-pitch Extraction

The estimated pitch is taken from the model's final output layer. The 127-dimensional output shows the most likely pitch as the highest value, called $peak_1$. For duet songs, the second singer's pitch ($peak_2$) is identified as the index with the second-highest value in the output layer. We distinguish between silence, solo, and duet scenarios based on the peak's magnitude as follows:

$$\text{Number of Singer} = \begin{cases} 2 & \text{if } peak_2/peak_1 \geq 0.5 \\ 1 & \text{if } peak_2/peak_1 < 0.5 \\ 0 & \text{if } peak_1, peak_2 < \alpha \end{cases} \quad (2)$$

In other words, if the ratio between the magnitudes of $peak_2$ and $peak_1$ exceeds or equals to 0.5, the scenario is classified as a duet, and two predicted pitches are reported. Otherwise, if the ratio is below 0.5, it is seen as a solo scenario, and only $peak_1$ is given. If both $peak_1$ and $peak_2$ are less than α, the scenario is considered a silence, and no pitch is given.

III. Experiment Setting

A. Model Setting

Initially, raw audio signals undergo resampling to achieve a 48 kHz sampling rate. Subsequently, the signals are transformed into spectrograms using STFT with a window size of 4096 and a hop size of 1024. Then, the linear frequency axis is transformed into the log frequency axis using the setting mentioned in Section II.A, and the magnitude is also transformed into the logarithmic scale.

In this work, the pitch estimation problem is treated as a classification task. We partition frequency spanning from 31 Hz (B0) to 1175 Hz (D6) into 127 pitch classes, excluding an unvoiced class. To establish the mapping between frequencies and pitch classes, we employ the following equation:

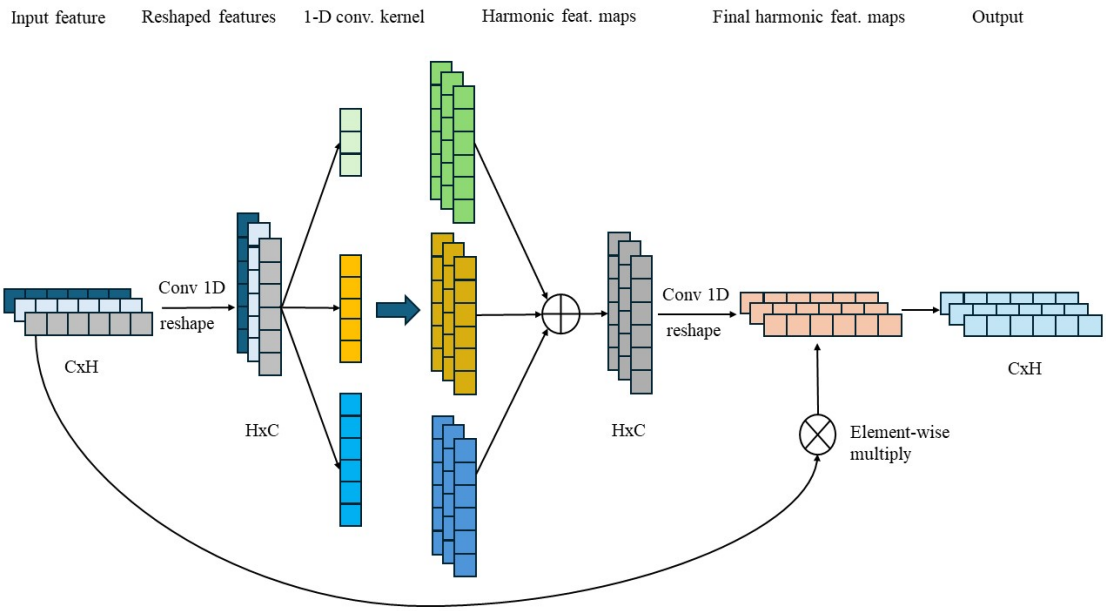


Fig. 3: Details of Harmonic-aware block (Har. Block) in our model. "C", and "H" respectively denote for the number of channels, and feature dimension.

$$f_n = 31 * 2^{(n-1)/2/12} \tag{3}$$

where n is the index of the pitch class, and f_n is the corresponding frequency.

For the labeling process, we employ a multi-label strategy [16], depicted in Fig. 4. Frames featuring two singers are allocated two pitches, labeled as 1 each, whereas frames with a single singer are labeled with a single pitch marked as 1. Frames without vocalization are identified with all pitch classes set to 0. Additionally, to ensure consistency between the audio signal and labels, we resample the labels, using mir_eval library [17], in the dataset to match the frame length of the log-spectrogram.

In this work, we employ the binary cross-entropy loss function to compute the error between the ground truth pitch vector y_i and the prediction pitch vector $\hat{y}_i$ for each frame:

$$L(y, \hat{y}) = \sum_{i=1}^{127} -y_i log(p(\hat{y}_i)) - (1 - y_i) log(1 - p(\hat{y}_i)) \tag{4}$$

We used the Adam optimizer [18] in this work, with the learning rate of 0.001. Adam facilitates faster convergence by dynamically adjusting the learning rate during training.

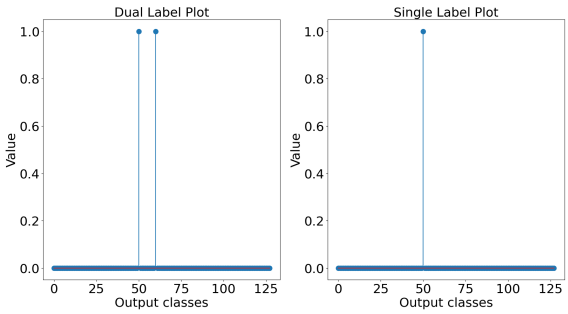


Fig. 4: Output samples of the labeling method (left panel: a duet frame; right panel: a solo frame).

B. Training Datasets and Data Augmentation

The MIR-1k dataset [19], comprising 1000 audio clips featuring 8 females and 11 males, serves as our primary training dataset. Each audio clip spans a duration ranging from 4 to 13 seconds, with a cumulative duration of 133 minutes. In particular, this dataset encompasses both instrumental and vocal tracks, separated into the left and right channels, respectively. Given our focus on monophonic scenarios, training exclusively utilizes data from the vocal channel. In addition, the vocal data are augmented to enhance the diversity and robustness of our model. Specifically, we employ shift-up (10 semitones) and shift-down (8 semitones) methods on vocal data to broaden the frequency range seen by the model.

Besides the single-pitch training dataset extracted from MIR-1k, we derive a new dataset from MIR-1k to simulate duet singing scenarios. This involves a mixing process where the original vocal track is first pitch-shifted upwards, followed by a mixing operation utilizing two conditions:

$$Duet = \begin{cases} Original + (0.8 \sim 0.95) * (Shift_up_{version}) \\ (0.8 \sim 0.95) * Original + (Shift_up_{version}) \end{cases} \tag{5}$$

The above two mixing conditions simulate situations where one singer's volume is slightly lower than the other's.

As for shifting pitch upward, we consider two augmentation methods. In the first augmentation method (AUG1), we upshift the pitch of the vocal track in intervals of 3, 4, 7, and 12 semitones, representing the minor third, major third, fifth, and octave, respectively. These intervals are commonly used in singing techniques [20]. In order to enhance our data-driven approach, we try the second augmentation method (AUG2), which considers all possible harmony singing scenarios by shifting up the vocal track with intervals from 1 to 12 semitones. For both augmentation methods, we use Eq. 5 to generate duet-singing datasets.

C. Evaluation Datasets and Metrics

For single pitch evaluation, we use the Vocadito dataset [21], which contains 40 short excerpts in 7 different languages of monophonic singing. Since we cannot find public duet singing datasets, we apply audio stems from the Dagstuhl ChoirSet dataset [22], which focuses on choral singing, to create our own test sets for duet singing. The ChoirSet dataset includes recordings from multiple singers performing three pieces in various harmonies: Soprano, Alto, Tenor, and Bass. Each singer was recorded using three types of microphones: a headset microphone (HSM), a larynx microphone (LRX), and a dynamic microphone (DYM). In practice, vocal pitch frequencies typically range from 100 Hz to 400 Hz [23]. Therefore, we extracted tenor and alto audios recorded by the LRX microphone from the dataset, pitch ranging from approximately 130 Hz to 698 Hz, and mixed them together to mimic duet singing for evaluation purposes.

To evaluate the models' efficacy on dual-pitch extraction, we employ metrics mentioned in [24]: Precision (P), Recall (R), F-score (F), FLOPs, and the number of parameters. Predicted pitches must match the true label exactly to be considered correct predictions without any tolerance window.

IV. Results and Discussion

In this study, we conduct comprehensive experiments with different model configurations denoted as Proposed 0, Proposed 1, and Proposed 2, as detailed in Table II. Within the Proposed 0 configuration, the number of channels in Harmonic blocks is 64. The Proposed 1 configuration maintains consistency with the Proposed 0 configuration regarding components, albeit with a reduction in the count of Harmonic blocks' channels to 8. The Proposed 2 configuration represents a different combination of modules from the Proposed 1 configuration, lacking one Harmonic-aware block.

A. Experiments on Different Thresholds

Firstly, we assess the performance of the proposed model using precision (P), recall (R), and F-score (F) metrics across varying thresholds, α, in Eq. 2. The threshold governs the trade-off between false positives and false negatives, crucial for refining the detection accuracy. Table III summarizes the results on dual-pitch extraction, using the Proposed 2 configuration, with thresholds ranging from 0.2 to 0.6.

Upon examination of the results, it is evident that the optimal F-score is achieved at a threshold of 0.2, with corresponding scores of P = 0.8669, R = 0.8592, and F = 0.8630. It indicates a good balanced precision and recall at this threshold. However, as the threshold is further increased, there is a significant decline in the performance of all metrics. Therefore, we set the threshold to 0.2 for all experiments.

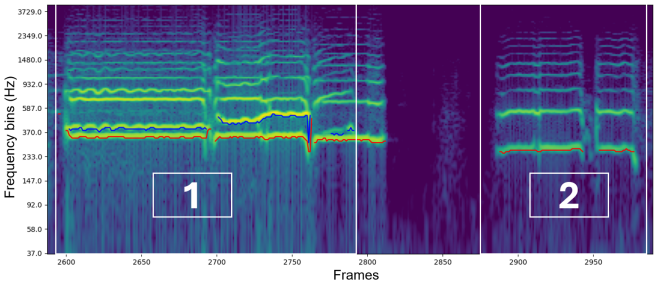


Fig. 5: The example output of the Proposed 2 configuration on a duet song. It shows the pitch tracks of two singers in red and blue colors respectively, overlaid on the log-spectrogram. The left region (1) indicates a duet singing situation, and the right region (2) indicates a solo singing situation.

TABLE II: Configurations of compared models. It shows the included (O) and excluded (X) components for different models. Used channel numbers (C) are shown in parentheses.

	Conv. 1	**Conv. 2**	**Res. + Har. Block 1**	**Res. + Har. Block 2**
Proposed 0	O	O	O (C: 64)	O (C:64)
Proposed 1	O	O	O (C:8)	O (C:8)
Proposed 2	O	O	X	O (C:8)

B. Performance Comparison of Data Augmentation Methods on Duet Songs

Results in Tables IV and V demonstrate a huge improvement in the model when using the second data augmentation method (AUG2) for training. All configurations have significant increases in precision, recall, and F-score. Notable, the Proposed 2 configuration in AUG2 outperforms all configurations in AUG1 with P=0.8669, R=0.8592, and F=0.8630. These results show that the AUG2 method is more effective in generalizing the model without increasing the computation. Although these scores are slightly lower than scores of LATE/DEEP [10], the

TABLE III: Performance scores of dual-pitch extraction thresholds α in Eq. 2.

α	**P**	**R**	**F**
0.2	0.8669	0.8592	0.8630
0.3	0.8494	0.8418	0.8455
0.4	0.8340	0.8266	0.8302
0.5	0.7974	0.7903	0.7938
0.6	0.7250	0.7186	0.7217

TABLE IV: Performance evaluation on duet songs using AUG1 duet dataset. FLOPS is computed per frame.

	P	R	F	FLOPs	Params
Proposed 0	0.8227	0.8151	0.8188	39.51M	1.28M
Proposed 1	0.8109	0.8033	0.8070	11.04M	1.26M
Proposed 2	0.7825	0.7749	0.7786	7.33M	0.80M

TABLE V: Performance evaluation on duet songs using AUG2 duet dataset. FLOPS is computed per frame.

	P	R	F	FLOPs	Params
Proposed 0	0.8829	0.8752	0.8790	39.51M	1.28M
Proposed 1	0.8834	0.8757	0.8795	11.04M	1.26M
Proposed 2	0.8669	0.8592	0.8630	7.33M	0.80M
LATE/DEEP [10]	0.8757	0.8684	0.8720	1.06G	0.59M

Proposed 2 configuration has a big advantage in computation efficiency with only 7.33M FLOPs compared to LATE/DEEP's 1.06G FLOPs. The huge reduction in computation makes Proposed 2 much more suitable for real-time applications. Fig. 5 shows an example output of Proposed 2 on a duet song. It showcases two estimated pitch tracks, represented by red and blue lines respectively, overlaid on the log-spectrogram.

C. Model Performance on Solo Songs

We also assess the performance of compared models on solo songs using a range of metrics, including overall accuracy (OA), raw pitch accuracy (RPA), raw chroma accuracy (RCA), voicing recall (VR), and voicing false alarm (VFA) in Mir_eval tool [17]. These metrics, which are often used in literature of single-pitch detection, are applied without applying any tolerance window. As illustrated in Table VI, Proposed 2 demonstrates commendable performance, achieving high accuracy while maintaining efficiency despite parameter reductions and reduced computational time. When compared against YIN [2], a trusted real-time single-pitch detection algorithm, Proposed 2 outperforms it across all metrics. In addition, results in the table also demonstrate Proposed 2 outperforms the spectral branch of the NN model [12] in almost all metrics. The spectral branch of [12] utilizes 11 frames of feature, which are from the first 352 frequency bin of the log-spectrogram, to predict the current pitch. Due to its high complexity, with FLOPs = 1.97G and parameters = 5.26M, it cannot be used for real-time inference.

V. Conclusions

In summary, we have presented a new neural network for real-time dual-pitch extraction to address the need for fast and accurate pitch estimation in audio processing. Our model works well for solo and duet songs. We get comparable results by using spectral features in one framework and reducing the computational cost. We also introduced a new data augmentation method for simulating duet singing. Our results show the potential of neural networks in audio processing and opens up new applications in MIR, speech processing and beyond. Next work will be to optimize and refine our model and explore more applications in real world scenarios.

TABLE VI: Performance evaluation on solo songs. SB is short for 'spectral branch'.

	OA	RPA	RCA	VR	VFA
Proposed 0	0.8386	0.8216	0.8230	0.9694	0.2507
Proposed 1	0.8407	0.8029	0.8066	0.9592	0.1957
Proposed 2	0.8289	0.7964	0.7978	0.9630	0.2215
YIN [2]	0.6106	0.6608	0.6710	0.9519	0.5011
SB of [12]	0.8449	0.7552	0.7604	0.9160	0.1179

VI. Acknowledgments

This research was partially supported by Realtek Semiconductor Corp., Taiwan.

References

[1] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[2] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[3] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE, 2014, pp. 659–663.

[4] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, *CREPE: A Convolutional Representation for Pitch Estimation*, 2018. arXiv: 1802.06182 [eess.AS]. [Online]. Available: https://arxiv.org/abs/1802.06182.

[5] H. Chou, M. T. Chen, and T. S. Chi, "A Hybrid Neural Network Based on the Duplex Model of Pitch Perception for Singing Melody Extraction, year=2018," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 381–385. DOI: 10.1109/ICASSP.2018.8461483.

[6] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 277–288, 2006. DOI: 10.1109/TSA.2005.853206.

[7] X. Sun, "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. I-333-I–336. DOI: 10.1109/ICASSP.2002.5743722.

[8] P. McLeod, "Fast, accurate pitch detection tools for music analysis," *Unpublished PhD thesis. Department of Computer Science, University of Otago*, 2008.

[9] O. Mayor, J. Bonada, and A. Loscos, "Performance analysis and scoring of the singing voice," in *Proc. 35th AES Intl. Conf., London, UK*, 2009, pp. 1–7.

[10] H. Cuesta, B. McFee, and E. Gómez, "Multiple f0 estimation in vocal ensembles using convolutional neural networks," *arXiv preprint arXiv:2009.04172*, 2020.

[11] J. Zhang, J. Tang, and L. R. Dai, "RNN-BLSTM based multi-pitch estimation.," in *Interspeech*, vol. 2016, 2016, pp. 1785–1789.

[12] S. Yu, Y. Yu, X. Sun, and W. Li, "A neural harmonic-aware network with gated attentive fusion for singing melody extraction," *Neurocomputing*, vol. 521, pp. 160–171, 2023.

[13] A. Natsiou and S. O'Leary, "Audio representations for deep learning in sound synthesis: A review," in *2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, 2021, pp. 1–8.

[14] W. Wei, P. Li, Y. Yu, and W. Li, "Harmof0: Logarithmic scale dilated convolution for pitch estimation," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2022, pp. 1–6.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pmlr, 2015, pp. 448–456.

[16] S. Stone and E. Spector, "Deep Neural Network for Multi-Pitch Estimation Using Weighted Cross Entropy Loss," in *2021 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, IEEE, 2021, pp. 1–3.

[17] C. Raffel, B. McFee, E. J. Humphrey, *et al.*, "MIR_EVAL: A Transparent Implementation of Common MIR Metrics.," in *ISMIR*, vol. 10, 2014, p. 2014.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] C. L. Hsu and J. S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 310–319, 2009.

[20] O. Zulfi, *How to Sing Harmony: The Ultimate Guide for Beginners*, Accessed: 2024-03-08, 2024. [Online]. Available: https://deviantnoise.com/singing/how-to-sing-harmony/.

[21] R. M. Bittner, K. Pasalo, J. J. Bosch, G. Meseguer-Brocal, and D. Rubinstein, "Vocadito: A dataset of solo vocals with f_0, note, and lyric annotations," *CoRR*, vol. abs/2110.05580, 2021. arXiv: 2110.05580. [Online]. Available: https://arxiv.org/abs/2110.05580.

[22] S. Rosenzweig, H. Cuesta, C. Weiß, F. Scherbaum, E. Gómez, and M. Müller, "Dagstuhl ChoirSet: A multitrack dataset for MIR research on choral singing," *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 3, no. 1, pp. 98–110, 2020. DOI: 10.5334/tismir.48. [Online]. Available: http://doi.org/10.5334/tismir.48.

[23] iZotope. "How to EQ Vocals." (2023), [Online]. Available: https://www.izotope.com/en/learn/how-to-eq-vocals.html (visited on 07/17/2023).

[24] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-f0 estimation and tracking systems.," in *ISMIR*, 2009, pp. 315–320.

Layer-Wise Feature Distillation with Unsupervised Multi-Aspect Optimization for Improved Automatic Speech Assessment

Chung-Wen Wu* and Berlin Chen†
* † Department of Computer Science and Information Engineering, National Taiwan Normal University
* E-mail: 40947040S@ntnu.edu.tw
† E-mail: berlin@ntnu.edu.tw

***Abstract*—Self-supervised features have shown promising progress across several domains. In Automatic Speech Assessment (ASA), SSL features have been widely utilized in recent research. However, few studies have dedicated efforts to explore the layer-wise features in pre-trained SSL models. Another key challenge in ASA is the high cost of labeling various aspects of speech proficiency, such as content relevance, delivery, and language use. In this paper, we propose three unsupervised subtasks to assist model training in ASA and examine the importance of embeddings from each layer of the acoustic model for various aspects. This provides preliminary research in this area. Extensive experiments demonstrate that model training with our tailored subtasks achieves superior performance in speech proficiency assessment tasks.**

I. Introduction

Automatic Speech Assessment (ASA) for second language (L2) proficiency offers significant benefits for both learners and educators. For L2 learners, ASA provides timely and feedback, which is crucial for effective language acquisition and skill improvement. This immediate response helps learners identify and correct errors, reinforcing proper pronunciation. For educators and testing institutions, ASA ensures more efficient and consistent scoring of language assessments. By automating the evaluation process, ASA reduces human error and bias, leading to fairer and more accurate results.

There are various methods to measure the speech proficiency of L2 learners, which can be broadly categorized into three components: content, delivery, and language use. Content assessment evaluates the relevance and coherence of the spoken material. Delivery encompasses prosody, pronunciation, stress, and other aspects related to how the speech is articulated. Language use focuses on grammar, morphology, and syntactic dependency, assessing the accuracy and complexity of the language structures employed.

However, labeling content, delivery, and language use separately is extremely time-consuming and requires professional training for the labelers. Moreover, recent research indicates that detailed scores assigned by experts may exhibit more bias compared to holistic scores.

Early studies in ASA [1]–[4] predominantly relied on handcrafted features related to different aspects of speech proficiency as input, ultimately making only holistic score predictions. While these features offer interpretability and align directly with human grading criteria, they are inherently limited by the holistic label. A holistic label alone is insufficient for providing comprehensive information necessary to effectively learn representations of various aspects of speech proficiency.

SSL representations have made significant strides across a range of speech processing tasks [5]–[10], including automatic speech recognition (ASR), speech enhancement, keyword spotting, speaker diarization, mispronunciation detection and diagnosis (MDD), and ASA. Models like wav2vec 2.0 [5], Whisper [11], and HuBERT [12], pre-trained on extensive datasets, leverage contextual information to extract features that are more robust and generalizable. In the realm of ASA, these models have proven highly effective in representing advanced speech-related features [13], thereby enhancing the accuracy of speech assessment.

In this paper, we propose a novel model architecture to explore SSL features across different layers of pre-trained acoustic encoders. We tailor various unsupervised subtasks for content, delivery, and language use to leverage the potential abilities of SSL features. By doing so, we provide more detailed information to enhance the training of models for ASA, avoiding the enormous consumption of human resources and ensuring a more comprehensive evaluation of L2 proficiency.

Extensive experiments demonstrate that model training with our tailored subtasks achieves superior performance in speech proficiency assessment tasks. Furthermore, our results show that information related to content, delivery, and language use is distributed across different layers of the acoustic encoder.

In summary, this paper presents three main contributions:

1) Tailoring unsupervised subtasks for content, delivery, and language use provides more concrete information than relying solely on holistic score labels, while also minimizing the need for additional human resources.
2) Investigating and analyzing the feature distribution of each aspect related to speech proficiency in the acoustic encoders.
3) Proposing a model trained with our tailored subtasks, we demonstrate that our approach significantly improves the accuracy of ASA tasks.

To the best of our knowledge, the relationship between speech proficiency features and latent representations at each layer of the acoustic encoder in ASA is an underexplored topic. This paper aims to provide preliminary research in this area.

II. Related Work

Recently, several studies in ASA have been conducted [14]–[16]. However, due to the scarcity of publicly available resources, these studies typically use only holistic labels as targets for model training. Although some research has access to detailed labels, such as topic development, delivery, and language use, this data is nonpublic. Even if we invest in and train human raters to label such detailed data, the collected data may not be sufficiently reliable due to the halo effect, where raters may rate all or some aspects of speech proficiency with similar scores.

SSL features from models like Whisper, wav2vec 2.0, and HuBERT have demonstrated great performance in several studies. However, to date, only a few studies have leveraged the features from the various layers of these pre-trained models to make assessments in ASA systems.

III. Methodology

A. Pseudo label generation

We generate pseudo labels for content, delivery, and language use from the data. These pseudo labels provide more detailed and stable information related to speech proficiency assessment, eliminating the need for human raters.

1) Content pseudo label: In content, we direct using content number associated with data as label.

2) Delivery pseudo label: In delivery, to measure a user's fluency and pronunciation, we begin by extracting some handcrafted features related to delivery and use the K-means algorithm to cluster the data into k clusters, which we then use as delivery pseudo labels. We employ two different types of acoustic models: monolingual wav2vec 2.0 and multilingual wav2vec 2.0. We start by using multilingual wav2vec 2.0 to recognize the transcriptions of user speech. Since L2 learners' pronunciation and accents are quite diverse, the multilingual training model can achieve better accuracy than the monolingual model in this scenario. On the other hand, we utilize monolingual wav2vec 2.0 to segment the audio signal into word-level segments using the CTC segmentation algorithm for force alignment with ASR transcriptions recognized by multilingual wav2vec 2.0. Using the monolingual model for alignment implies that if the L2 learner pronounces more accurately, the alignment result will be more correct. Within each word-level segment, we extract various delivery-related features, including pitch, duration (DUR), intensity, following silence, posterior probability from both monolingual and multilingual wav2vec 2.0, LM score from the n-gram LM during transcription, and confidence score from multilingual wav2vec 2.0. These features are utilized to construct an 8-dimensional continuous segment feature d_i, and subsequently, a sequence of segment features $[d_1, d_2, \ldots, d_M]$ is concatenated to represent the delivery feature, where M is a predefined length achieved by either truncating or padding with zeros. Finally, we use the K-means algorithm to obtain k clusters as pseudo labels.

3) Language use pseudo label: This part focuses on analyzing the grammar and syntax of sentences spoken by the user. We obtain the ASR transcription during the delivery pseudo label process. We extract language use features from each word to construct a sequence of language use features, which is then constrained to length M by padding or truncating. Next, we use spaCy, a natural language processing toolkit in Python, to extract Part-Of-Speech (POS) tags, dependency labels (DEP), and morphology (Morph.) labels. These features are encoded using one-hot encoding to obtain sparse language use features. Finally, similar to the delivery pseudo label generation, we leverage the K-means algorithm to cluster the data into k clusters as pseudo labels.

B. Proposed Model

To explore and leverage the representations of each layer in acoustic models, we propose a model architecture as illustrated in Fig. 1. This architecture trains three weighted vectors, w_c, w_d, and w_l, which separately combine acoustic embeddings into latent embeddings for content, delivery, and language use aspects. Subsequently, we use a linear layer to adapt the combined embeddings into latent representations with H dimensions separately. We employ CLS token embeddings to concatenate with combined latent representations, which are then passed through bidirectional Transformer blocks [17] of depth D separately. We extract the CLS token representation after the Transformer blocks to predict pseudo labels for content, delivery, and language use using multilayer perceptron (MLP) with H_o hidden dimensions to assist training. Finally, all CLS tokens are concatenated and using MLP to make the final holistic score assessment.

C. Training process

Based on our observations, although pseudo labels can provide rich information to assist holistic score prediction, they still contain some noise due to forced alignment errors or ASR transcript errors. To mitigate these side effects, we leverage a triplet loss approach across the three subtasks. This method helps learn well-represented continuous embeddings that cluster similar pseudo labels together and separate dissimilar ones. The benefit is a reduction in noise effects and the development of more robust representations. Finally, the holistic score is predicted using cross-entropy loss to train the model. The formulas are as follows:

$$L = L_c + \lambda(L_t(a_c, p_c, n_c) + L_t(a_d, p_d, n_d) + L_t(a_l, p_l, n_l)) \tag{1}$$

where L_c is cross entropy loss, L_t is triplet loss, a_c, p_c, n_c are anchor, positive and negative samples respectively.

IV. Dataset

Public corpora in ASA are quite scarce, with the most popular and commonly used datasets being ICNALE and Speechocean762. Both have their strengths and weaknesses.

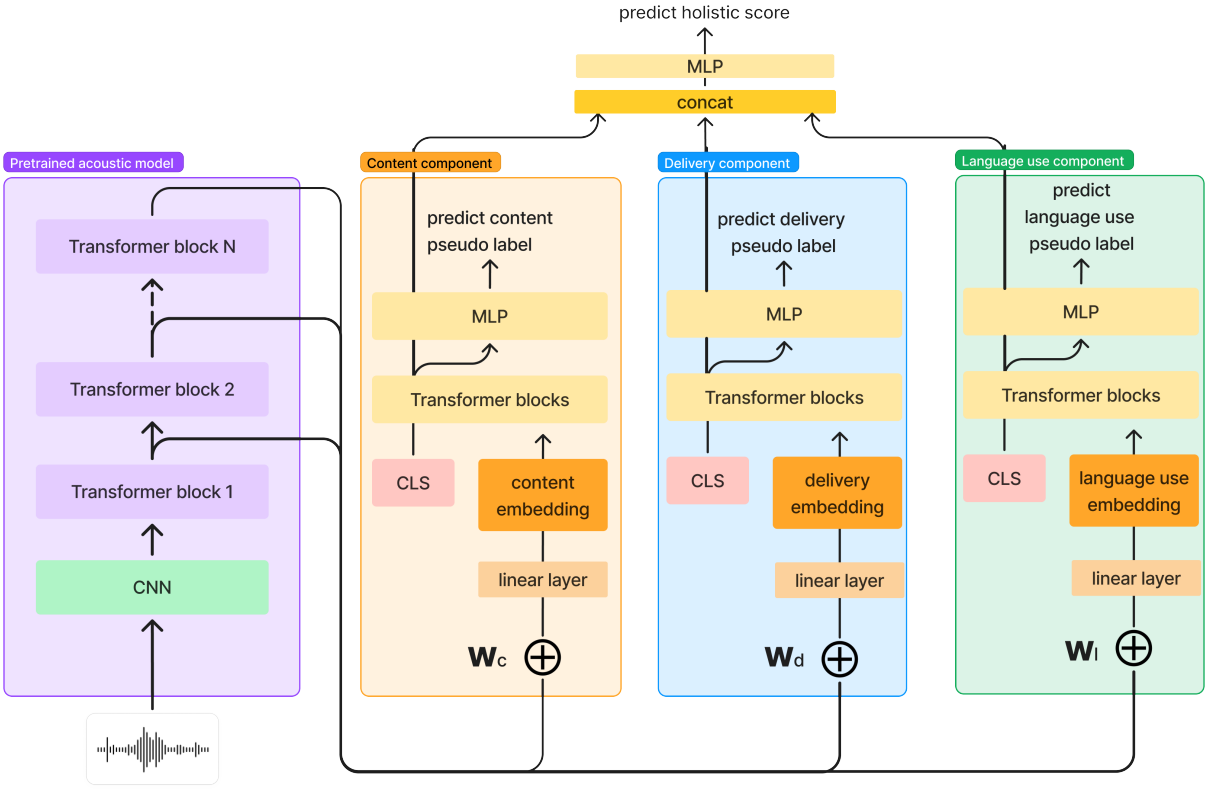


Fig. 1. The proposed model architecture combines latent representations from each layer of a pre-trained acoustic model to obtain embeddings for various aspects of speech proficiency. Leveraging pseudo labels to assist the training process.

The spoken part of the ICNALE dataset includes adequately lengthy monologues and dialogues, with each monologue being roughly 60 seconds and each dialogue lasting 30 to 40 minutes. However, ICNALE uses English test scores such as TOEFL, TOEIC, IELTS, and others as labels. These standard tests reflect the vocabulary and some reading skills of English learners but may not be sufficiently related to speaking skills.

In contrast, the Speechocean762 dataset has a large amount of data and detailed score labeling, such as accuracy, completeness, fluency, prosody, and even phoneme-level labeling. However, the speech in this dataset is quite short, typically around 2 to 5 seconds, and consists only of read-aloud tasks. This limitation prevents a comprehensive assessment in ASA systems.

Due to the above reasons, we used a corpus collected by the Language Training and Testing Center from the General English Proficiency Test (GEPT) intermediate level exam in this study. GEPT is an English certification test in Taiwan that covers listening, reading, writing, and speaking. For our study, we use data from the speaking section. This exam is a high-stake English proficiency test. It includes 1,199 responses divided equally among four sets of questions, each set answered by a different participant. The instructions given to participants were: "Below are a picture and four related questions. Please complete your answers in one and a half minutes. Do not read the number or the question when you answer. Please first look at the picture and think about the questions for thirty seconds."

Figure 3 illustrates the distribution of scores and response lengths. The scores, which range from 1 to 5, were calculated as averages from assessments provided by two professionals, with any floating-point values being discarded unconditionally. To create an unknown content test set, we randomly selected a specific set of questions. This set was designed to assess performance under cold start conditions, where the model encounters previously unseen content. The data not used for the unknown content test set was split into training, development, and known content test sets in an 8:1:1 ratio. This approach guarantees that the model receives substantial training data while also enabling thorough evaluation on both familiar and new content types.

V. Experiments and Results

A. Experiments Setup

The pretrained acoustic model used in our experiments is the Whisper-base model, which has 7 Transformer blocks in the encoder. The depth of the Transformer blocks in our model is set to 3. The dimensions for the latent representations in each aspect (H) and the hidden layers in the MLP (H_o) were configured to be 128 and 64, respectively. Throughout the training process, we employed the RAdam optimizer, incorporating a weight decay of 1×10^{-5}. The learning rate

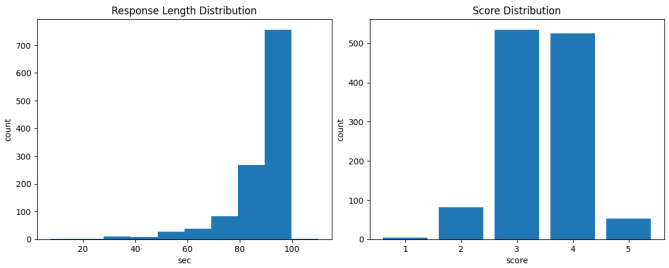


Fig. 2. Response length and score distribution in the GEPT intermediate level dataset.

was set to 2×10^{-4}, and all experiments were carried out over 16 epochs with a batch size of 8.

B. Ablation studies

TABLE I
ABLATION STUDIES ON THREE COMPONENTS

Model	Known Content Accuracy	Unknown Content Accuracy
Complete model	**0.722**	**0.727**
Without language use	0.667	0.697
Without delivery	0.644	0.697
Without content	0.667	0.683
Without any subtask[a]	0.633	0.647

[a] Without any subtask means that we only use a combined embedding from the acoustic model. We then sequentially pass this combined embedding through the same blocks of the model as illustrated in Fig. 1, extracting a CLS token embedding to predict the holistic score.holistic score predict.

As shown in Table I, the experiments demonstrate that training with three subtasks outperforms all other ablation models, indicating that each subtask improves the model's accuracy in holistic score assessment. The experiments also highlight that directly using the holistic score for training or combining embeddings from each layer for holistic score prediction is not optimal.

C. Visualize the weight distributions in combined embeddings

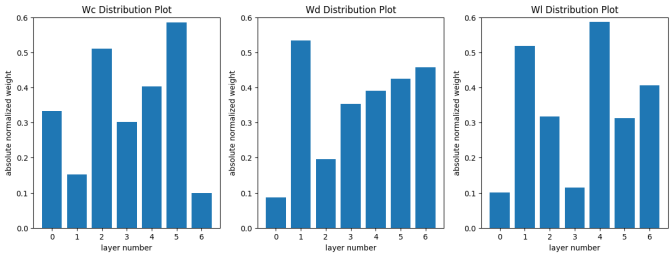


Fig. 3. Absolute weight distributions across each layer in the whisper-base encoder

We extract w_c, w_d, w_l from complete model, normaling it and take absolute value to shows each layers importance in whisper-base encoder, as shown in Fig. 3. Overall, the distributions for each aspect are quite distinct, proving that latent

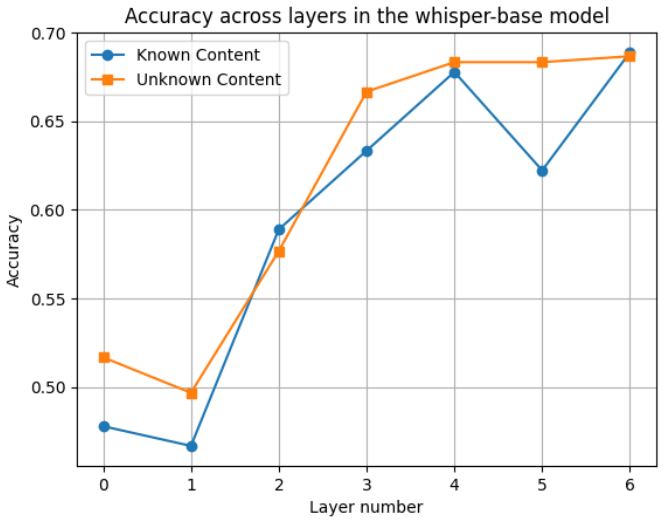


Fig. 4. We directly extract embeddings from each layer and sequentially pass them through the same blocks of the model, as illustrated in Fig. 1, to make holistic score predictions. The layer numbers range from 0 to 6, representing layers in the encoder from shallow to deep.

representations in different layers contain different meanings. Our method successfully directs attention to these various aspects, leveraging the unique information captured at each layer. This differentiation enhances the model's ability to make more accurate and nuanced assessments. For the content aspect, the features tend to focus on the middle layers. In the delivery aspect, layer 1 is the most significant, but other important features tend to focus on the deeper layers. For the language component, the features tend to focus on both shallow and deep layers.

D. Performance using features from each layer directly

In the Whisper-base model, when training using only the holistic score, the general tendency is that deeper layers provide better performance, as shown in Fig. 4. Experiments indicate that training with only the holistic score tends to yield better performance when using the last layer alone, achieving 0.689 and 0.687 in known content and unknown content, respectively. In contrast, combining all layers in the Whisper-base model, as shown in Table I, results in scores of 0.633 and 0.647 in known content and unknown content, respectively. This phenomenon points out that the holistic score alone may not provide sufficient information to effectively train a well-represented combined latent representation.

VI. CONCLUSIONS

We proposed a training method with three subtasks and demonstrated that this method can benefit model training in the ASA system using SSL features. By incorporating subtasks for content, delivery, and language use, our approach leverages rich pseudo labels to provide detailed and consistent information. This method improves the robustness and accuracy of the model, ensuring a more comprehensive assessment of L2 proficiency. The effectiveness of our approach is validated

through extensive experiments, demonstrating superior performance compared to relying solely on holistic labels.

Our experiments show that features of various aspects emerge in different layers of the model. This observation highlights the importance of leveraging multi-layer embeddings to capture the diverse and complex characteristics of speech proficiency. By analyzing and utilizing the representations from multiple layers, our proposed method can more effectively leverages content, delivery, and language use, leading to improved performance in ASA systems.

References

[1] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[2] Y. Qian, P. Lange, K. Evanini, *et al.*, "Neural approaches to automated speech scoring of monologue and dialogue responses," in *Proceedings of the 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2019, pp. 8112–8116.

[3] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, *et al.*, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 338–345. DOI: 10.1109/ASRU.2015.7404814.

[4] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6234–6238. DOI: 10.1109/ICASSP.2018.8462562.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[6] W.-N. Hsu, A. Sriram, A. Baevski, *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[7] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[8] H. S. Bovbjerg and Z.-H. Tan, "Improving label-deficient keyword spotting through self-supervised pre-training," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSP)*, 2023, pp. 1–5.

[9] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for mispronunciation detection.," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 4428–4432.

[10] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176 274–176 285, 2020. DOI: 10.1109/ACCESS.2020.3026823.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of International Conference on Machine Learning*, PMLR, 2023, pp. 28 492–28 518.

[12] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[13] S. Park and R. Ubale, "Multitask learning model with text and speech representation for fine-grained speech scoring," in *Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[14] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," *arXiv preprint arXiv:2204.03863*, 2022.

[15] J. Park and S. Choi, "Addressing cold start problem for end-to-end automatic speech scoring," *arXiv preprint arXiv:2306.14310*, 2023.

[16] S. Bannò and M. Matassoni, "Proficiency assessment of L2 spoken english using wav2vec 2.0," in *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1088–1095.

[17] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

Wind Noise Reduction with Orthogonal Polynomial Expansion

Li Du*, Chao Pan*, Lijun Zhang*
* Center of Intelligent Acoustics and Immersive Communications,
School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
E-mail: dulia@mail.nwpu.edu.cn, {cpan, zhanglj7385}@nwpu.edu.cn

***Abstract*—This paper is dedicated to mitigating unwanted ambient wind noise in outdoor multi-channel recordings. Different from conventional methodologies such as the multi-channel Wiener filter (MWF) and its extensions, we propose a wind noise reduction method from a beamforming perspective based on the spatial characteristics of wind noise captured by closely spaced microphones. Specifically, the noisy observation is decomposed into a parameterization space concerning frequency through orthogonal polynomial expansion. Subsequently, the estimated target source signals are reconstructed by selecting coefficients corresponding to desired components within the polynomial basis. Moreover, we introduce the Alternating Direction Method of Multipliers (ADMM) framework to attain a sparse solution for the coefficients during the decomposition phase. In comparison to existing multi-channel wind noise reduction methods, simulation results demonstrate the superior performance of the proposed method, particularly in low signal-to-noise ratio (SNR) scenarios.**

I. Introduction

Wind noise is a serious impediment to the intelligibility and quality of speech communication in outdoor scenarios, presenting a challenging task compared to other ambient noise [1], [2], [3], [4], [5], [6], [7], [8]. The inherent difficulties stem from the highly dynamic and nonstationary nature of wind-induced disturbances, compounded by a low signal-to-noise ratio (SNR). Traditional approaches for wind noise reduction can be divided into single-channel methods [2], [6], [9], [10] and multi-channel methods [7], [11], [12], [13]. In the multi-channel processings, common methods are based on the assumption that wind noise is uncorrelated. However, recent investigations showed that the spatial coherence of wind noise closely depends on the direction and the speed of free-field airflow, particularly apparent in the presence of strong winds [14], [15]. The non-zero spatial coherence of wind noise captured by closely spaced microphones is elucidated and can be modeled by a fluid-dynamics model in [16].

Utilizing the spatial coherence feature of wind noise, it is natural to design a multi-channel Wiener filter to effectively reduce wind noise. This approach relies on the accurate estimation of the power spectral densities (PSDs) for both wind noise and target signal, e.g., speech, as detailed in the works of [7], [17]. And extensions to this work include the estimation of a trade-off parameter for the wiener filter, denoted as the parametric multi-channel Wiener filter (PMWF) [17], [18] and improved PSD estimation methods [19].

A critical component of multi-channel noise reduction is beamforming, which exploits the spatio-temporal information to extract the target source signal from noisy observations [20]. The beamformers possess the ability to form a main beam directed towards the desired signal based on knowledge of the directions of the intended signal and interference signal, instead of the statistical properties. Among the diverse beamforming approaches, the differential beamformers provide the potential in speech communication systems due to their frequency-invariant beampatterns and small apertures [21], [22]. Inspired by the design of a differential beamformer with orthogonal polynomials [22], we present a novel method for wind noise reduction through orthogonal polynomial expansion.

To deal with the wind noise in microphone arrays, we first present the signal model and some important definitions. Then we investigate the spatial coherence between the speech signal and wind noise, clarifying their distinctions. An orthogonal polynomial expansion is performed to redefine the noisy observations in a frequency dependent parameterization space. Specifically, by analyzing the characteristics of speech source signals and wind noise on each order of orthogonal polynomials, we identify the orders in which speech source signals are dominant. Thus, the corresponding orders are selected to reconstruct an enhanced version of the desired source signal. An additional contribution is the introduction of the Alternating Direction Method of Multipliers (ADMM) scheme to optimize the decomposition coefficients. Simulations are conducted to illustrate the performance of the proposed method in terms of SNR, speech quality and speech intelligibility.

II. Signal Model

Consider an M-element uniform linear microphone array in the outdoor environment where the wind noise dominates the ambient noise. The distance between two successive sensors is δ_0 and the source signal is from the direction of θ. The reverberant effect can be omitted. We formulate the problem and devise the approach in the short-time Fourier transform (STFT) domain [23]. Hence, the observation of the mth sensor can be modeled as

$$Y_m(f,t) = X_m(f,t) + V_m(f,t), m \in \{0, \cdots, M-1\}, \quad (1)$$

where f is the frequency bin index, t is the time frame index, $X_m(f,t)$ is the clean speech signal captured by the

mth sensor, and $V_m(f,t)$ is the additive ambient noise. In this work, $V_m(f,t)$ mainly consists of wind noise.

For convenience, the signal model in (1) is concatenated into the vector form

$$\begin{aligned}\mathbf{y}(f,t) &\triangleq \begin{bmatrix} Y_0(f,t) & Y_1(f,t) & \cdots & Y_{M-1}(f,t) \end{bmatrix}^T \\ &= \mathbf{x}(f,t) + \mathbf{v}(f,t) \\ &= \mathbf{d}(f,\theta)X_0(f,t) + \mathbf{v}(f,t), \end{aligned} \tag{2}$$

where $\mathbf{x}(f,t)$ and $\mathbf{v}(f,t)$ are defined in the same way as $\mathbf{y}(f,t)$, and

$$\mathbf{d}(f,\theta) \triangleq \begin{bmatrix} 1 \\ e^{-j\frac{2\pi f}{c}\delta_0\cos(\theta)} \\ \vdots \\ e^{-j\frac{2\pi f}{c}(M-1)\delta_0\cos(\theta)} \end{bmatrix} \tag{3}$$

is the array manifold vector, where $j=\sqrt{-1}$ is the imaginary unit and c is the speed of sound in air. It should be pointed out that, for general array geometries, the array manifold vector is a function of source direction and sensor positions, which is not as simple as expressed in (3). Then, the correlation matrix of $\mathbf{y}(f,t)$ is

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(f,t) &= \mathbf{\Phi}_{\mathbf{x}}(f,t) + \mathbf{\Phi}_{\mathbf{v}}(f,t) \\ &= \phi_{X_0}(f,t)\mathbf{d}(f,\theta)\mathbf{d}^H(f,\theta) + \mathbf{\Phi}_{\mathbf{v}}(f,t), \end{aligned} \tag{4}$$

where $\mathbf{\Phi}_{\mathbf{x}}(f,t)$ and $\mathbf{\Phi}_{\mathbf{v}}(f,t)$ are the correlation matrices of $\mathbf{x}(f,t)$ and $\mathbf{v}(f,t)$, respectively, and $\phi_{X_0}(f,t) \triangleq \mathbb{E}\left[|X_0(f,t)|^2\right]$ is the variance of $X_0(f,t)$.

By applying a filter $\mathbf{h}(f,t)$ to the array observation $\mathbf{y}(f,t)$, the array output can be expressed as

$$\begin{aligned} Z(f,t) &= \mathbf{h}^H(f,t)\mathbf{y}(f,t) \\ &= X_{\mathrm{fd}}(f,t) + V_{\mathrm{rn}}(f,t), \end{aligned} \tag{5}$$

where

$$X_{\mathrm{fd}}(f,t) = \mathbf{h}^H(f,t)\mathbf{x}(f,t), \tag{6}$$

$$V_{\mathrm{rn}}(f,t) = \mathbf{h}^H(f,t)\mathbf{v}(f,t) \tag{7}$$

are the filtered desired source signal and the residual noise, respectively. The variance of the array output $Z(f,t)$ can be expressed as

$$\begin{aligned}\phi_Z(f,t) &= \mathbf{h}^H(f,t)\mathbf{\Phi}_{\mathbf{y}}(f,t)\mathbf{h}(f,t) \\ &= \phi_{X_{\mathrm{fd}}}(f,t) + \phi_{V_{\mathrm{rn}}}(f,t). \end{aligned} \tag{8}$$

Similarly, the variance/power of the desired source signal $\phi_{X_{\mathrm{fd}}}(f,t)$ and the variance/power of the residual noise $\phi_{V_{\mathrm{rn}}}(f,t)$ are

$$\phi_{X_{\mathrm{fd}}}(f,t) = \phi_{X_0}(f,t)|\mathbf{h}^H(f,t)\mathbf{d}(f,\theta)|^2, \tag{9}$$

$$\phi_{V_{\mathrm{rn}}}(f,t) = \mathbf{h}^H(f,t)\mathbf{\Phi}_{\mathbf{v}}(f,t)\mathbf{h}(f,t). \tag{10}$$

The output SNR can be expressed as

$$\mathrm{oSNR}(f,t) \triangleq \frac{\phi_{X_0}(f,t)|\mathbf{h}^H(f,t)\mathbf{d}(f,\theta)|^2}{\mathbf{h}^H(f,t)\mathbf{\Phi}_{\mathbf{v}}(f,t)\mathbf{h}(f,t)}. \tag{11}$$

Various optimal beamformers are built for different design targets and optimization criteria, where adaptive beamformers like the minimum variance distortionless response (MVDR) beamformer require the good estimations of the covariance matrices of both desired source signal and additive noise. In this work, we introduce considerations of the spatial coherence of wind noise in a samll aperture array and design a beamformer that fully exploits the difference between the correlation matrices $\mathbf{\Phi}_{\mathbf{x}}(f,t)$ and $\mathbf{\Phi}_{\mathbf{v}}(f,t)$.

III. Covariance Matrix and Orthogonal Expansion

A. Covariance Matrix of Wind Noise

The covariance matrix of wind noise is dependent on the far-field wind velocity U_c and the wind stream direction ϑ, as stated in [14] and [16], with ϑ being defined identically to θ. Covariance matrix of the noise can be expressed as

$$\mathbf{\Phi}_{\mathbf{v}}(f,t) = \phi_V(f,t)\mathbf{\Upsilon}(f,\vartheta) \odot \left[\mathbf{a}(f,\vartheta)\mathbf{a}^H(f,\vartheta)\right], \tag{12}$$

where $\odot$ stands for Hadamard product, $\phi_V(f,t)$ is the variance of the noise, the (i,j)th element of the spatial coherence matrix $\mathbf{\Upsilon}(f,\vartheta)$ modeled by the Corcos model can be expressed as

$$[\mathbf{\Upsilon}(f,\vartheta)]_{i,j} = e^{-\frac{2\pi f}{U_c}\delta_0|i-j|g(\vartheta)} \tag{13}$$

with

$$g(\vartheta) = \alpha_1\left|\cos(\vartheta)\right| + \alpha_2\left|\sin(\vartheta)\right| \tag{14}$$

denotes a coherence decay parameter, α_1, α_2 are the longitudinal and the lateral coherence decay rates respectively, experimentally provided in [15]. For a wind stream with constant direction and speed, (13) is assumed to be time-invariant. And the vector $\mathbf{a}(f,\vartheta)$ is defined as

$$\mathbf{a}(f,\vartheta) \triangleq \begin{bmatrix} 1 \\ e^{j\frac{2\pi f}{U_c}\delta_0\cos(\vartheta)} \\ \vdots \\ e^{j\frac{2\pi f}{U_c}(M-1)\delta_0\cos(\vartheta)} \end{bmatrix} \tag{15}$$

By comparing (3) and (15), the vector $\mathbf{a}(f,\vartheta)$ shares a similar form with the array manifold vector. Since U_c is often far smaller than the speed of sound in air, it is very likely that $\mathbf{d}(f,\theta)$ and $\mathbf{a}(f,\vartheta)$ belong to different subspaces, even if the intended source signal and wind noise are coming from the same direction, as do the $\mathbf{\Phi}_{\mathbf{x}}(f,t)$ and $\mathbf{\Phi}_{\mathbf{v}}(f,t)$.

B. The Orthogonal Expansion and the Decomposition

Following the work presented in [22], we express the array manifold vector as

$$\begin{aligned}\mathbf{d}(f,\theta) &= \begin{bmatrix} 1 & e^{-\jmath\varpi_1\cos\theta} & \cdots & e^{-\jmath\varpi_m\cos\theta} \end{bmatrix}^T \\ &= \sum_{n=0}^{N-1}\mathbf{c}_n(f)\mathcal{P}_n(\cos\theta), \end{aligned} \tag{16}$$

where $\varpi_m = 2\pi(m-1)f\delta_0/c$, and

$$\mathbf{c}_n(f) \triangleq \begin{bmatrix} C_{n,0}(f) & C_{n,1}(f) & \cdots & C_{n,M-1}(f) \end{bmatrix}^T \tag{17}$$

is a coefficient vector of length M, $n = 0, 1, 2, ..., N-1$ is the polynomial order, $\mathcal{P}_n(\cos\theta)$ is the orthogonal polynomial to approximate the exponential function of the array manifold vector. Several alternatives are available for expanding the array manifold, such as the MacLaurin's series, Jacobi polynomials [22], [24] and so on. Therefore, $\mathbf{c}_n(f)$ is determined by the geometric structure of the microphone array, the frequency band, the type of polynomial expansion and other factors.

We normalize the vector $\mathbf{c}_n(f)$'s according to

$$\mathbf{c}_n(f) \leftarrow \frac{1}{\|\mathbf{c}_n(f)\|}\mathbf{c}_n(f), \tag{18}$$

and then construct a matrix $\mathbf{C}(f)$ as

$$\mathbf{C}(f) \triangleq \left[\begin{array}{cccc} \mathbf{c}_0(f) & \mathbf{c}_1(f) & \cdots & \mathbf{c}_{N-1}(f) \end{array} \right], \tag{19}$$

which is clearly a matrix of size $M \times N$. Without loss of generality, we assume that $N \geq M$.

Given an array observation $\mathbf{y}(f,t)$, which is a function of the array manifold vector, we can get the optimal approximation according to (16) and divide it into

$$\begin{aligned} \mathbf{y}(f,t) &= \sum_{n=0}^{N-1} \beta_{y,n}(f,t)\mathbf{c}_n(f) \\ &= \mathbf{C}(f)\boldsymbol{\beta}_y(f,t), \end{aligned} \tag{20}$$

where

$$\boldsymbol{\beta}_y(f,t) \triangleq \left[\begin{array}{cccc} \beta_{y,0}(f,t) & \beta_{y,1}(f,t) & \cdots & \beta_{y,N-1}(f,t) \end{array} \right]^T. \tag{21}$$

Considering the speech signal in the STFT domain presents a large dynamic range, the normalization of the array observation $\mathbf{y}(f,t)$ is applied in practice

$$\mathbf{y}(f,t) \leftarrow \frac{1}{\|\mathbf{y}(f,t)\|_2}\mathbf{y}(f,t). \tag{22}$$

Similarly, we can define $\mathbf{x}(f,t) = \mathbf{C}(f)\boldsymbol{\beta}_x(f,t)$ and $\mathbf{v}(f,t) = \mathbf{C}(f)\boldsymbol{\beta}_v(f,t)$. It is evidently observed that a direct matrix inversion for solving the coefficient $\boldsymbol{\beta}_y(f,t)$ is prone to an ill-posed problem. Then we reformulate the matrix inversion as an optimization problem, introducing a regularization term into the cost function [25]. From a mathematical standpoint, the introduction of the ℓ_1 norm of $\boldsymbol{\beta}(f,t)$ is beneficial to the robustness of the solution. Furthermore, by the sparsity of the speech signal distribution exhibited in the frequency band, it is reasonable to assume that the $\boldsymbol{\beta}(f,t)$ is sparse in the STFT domain. The unconstrained optimization problem is shown in the following equation

$$\min \frac{1}{2}\|\mathbf{C}(f)\boldsymbol{\beta}_y(f,t) - \mathbf{y}(f,t)\|_2^2 + \lambda\|\boldsymbol{\beta}_y(f,t)\|_1. \tag{23}$$

Equation (23) is then solved by the ADMM framework [26]. Note that the variables in (23) are complex ones instead of real ones, so the optimization problem should be extended to the complex domain. A detailed discussion on the initialization and selection of parameters in the ADMM will be provided in the next section.

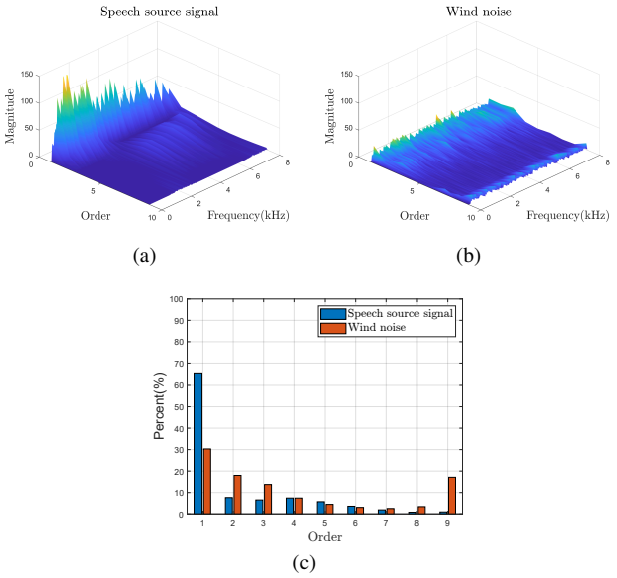


Fig. 1. $\boldsymbol{\beta}(f)$ coefficient distributions of (a) speech signal and (b)wind noise. (c) A percentage of distribution for the $\boldsymbol{\beta}(f)$ over each order.

Fig.1a and Fig.1b illustrate the values of $\boldsymbol{\beta}_x(f,t)$ and $\boldsymbol{\beta}_v(f,t)$, which correspond to the speech signal and wind noise respectively. It can be clearly seen that $\boldsymbol{\beta}_x(f,t)$ predominantly concentrates on the first and second orders at the lower frequencies, while $\boldsymbol{\beta}_v(f,t)$ is more evenly distributed across multiple orders. Fig.1c shows a bar chart depicting the magnitude of $\boldsymbol{\beta}_x(f,t)$, $\boldsymbol{\beta}_v(f,t)$ at each order as a percentage of the total. This above discussion implies that the low order components of $\boldsymbol{\beta}_y(f,t)$ contain the majority of speech signal components. As a consequence, the filter is constrained to these analytical findings in the reconstruction phase.

C. Reconstruction and Beamforming

According to (20), we have an estimation of the speech source signal $\mathbf{x}(f,t)$ as

$$\begin{aligned} \hat{\mathbf{x}}(f,t) &= \mathbf{C}(f)\hat{\boldsymbol{\beta}}_x(f,t) \\ &= \mathbf{C}(f)\mathbf{W}_L\boldsymbol{\beta}_y(f,t), \end{aligned} \tag{24}$$

where $\mathbf{W}_L$ is a constant matrix, which is defined as

$$\mathbf{W}_L \triangleq \left[\begin{array}{cc} \mathbf{I}_{L\times L} & \mathbf{0}_{L\times(N-L)} \\ \mathbf{0}_{(N-L)\times L} & \mathbf{0}_{(N-L)\times(N-L)} \end{array} \right], \tag{25}$$

and L is the order index at which the speech signal dominates, with $L \leq N$. Finally, the $\hat{\mathbf{x}}(f,t)$ is inverse normalized by

$$\mathbf{z}(f,t) \leftarrow \|\mathbf{y}(f,t)\|_2\hat{\mathbf{x}}(f,t). \tag{26}$$

In the case that we have an estimation of the array manifold vector toward the desired source, i.e., $\hat{\mathbf{d}}(f,t)$, we can calculate the array output according to

$$Z(f,t) = \frac{1}{\|\hat{\mathbf{d}}(f,t)\|^2}\hat{\mathbf{d}}^H(f,t)\mathbf{z}(f,t)$$
$$= \frac{1}{\|\hat{\mathbf{d}}(f,t)\|^2}\|\mathbf{y}(f,t)\|_2\hat{\mathbf{d}}^H(f,t)\mathbf{C}(f)\mathbf{W}_L\boldsymbol{\beta}_y(f,t) \quad (27)$$

In this paper, the direction of the source θ is known as a priori, and the $\mathbf{d}(f,\theta)$ is taken as the estimate of the array manifold vector, i.e., $\hat{\mathbf{d}}(f,t)$.

IV. Experiments and Analyses

A. Experimental Setup

Several experiments were conducted to validate and analyze the proposed method, with wind noise reduction methods developed in [17] serving as baselines. Sentences were randomly selected from the LibriSpeech dataset [22] and sampled at 16kHz. Wind noise was generated using an artificial wind noise generator [16] with temporal, spectral, and spatial characteristics matching measured multi-channel wind noise observations. The slow air stream velocity was set to $U_c = 1.8$ m/s and the fast velocity to 5 m/s in this work. Target speech signals were corrupted at three SNR values: -10dB, -5dB and 0dB. We utilized a closely spaced uniform linear array with eight microphones, each having an interelement spacing of 4 mm. The speech was kept in the endfire position, as well as wind noise in the direction $\vartheta = [0, 45°, 90°]$. We apply the STFT domain where the frame length is 32ms with half of the overlap between consecutive frames.

B. Results and Discussion

The parameter initialization and selection of the ADMM in the decomposition phase influence the performance of the proposed method. At each time-frequency frame, the dual variable λ in (23) is set to $\lambda = 0.1\mathbf{C}(f)^H\mathbf{y}(f,t)$. Fig.2 plots the spectrograms of the clean speech signal, wind noise, noisy signal at iSNR = 0dB, and the enhanced speech signal. Note that wind noise is predominantly focused in the low frequency band, which is the difference with other ambient noise characteristics. Fig.2d illustrates that the proposed method effectively addresses wind noise in the low frequency domain.

Considering the high fluctuation possessed by wind-induced disturbances, we mainly measure the fwSNR metrics of the proposed algorithm under various scenarios such as three wind directions and two wind speeds, with iSNR = -10dB. Averaged over all wind directions, the results in Table I show the effectiveness of the proposed method in complex wind turbulence situations.

In addition, a comparative analysis was conducted with the PMWF from [17] where the spatial complex coherence matrix of wind noise is also modeled as the classic Corcos model, and the baseline method MWF [7] and MVDR with the spatial coherence being an identity matrix. The results in Table II showcase the outstanding fwSNR, PESQ, and STOI improvements achieved by the proposed method at every iSNR level. *Highlighted values correspond to the best results*.

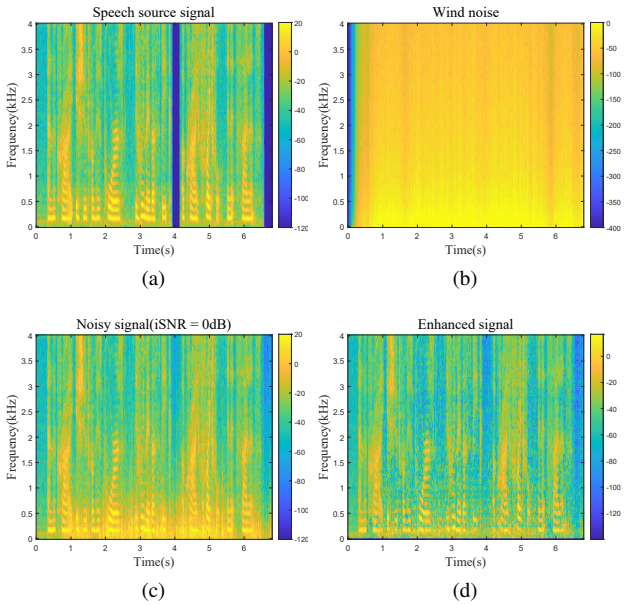


Fig. 2. The spectrograms of (a) clean signal, (b) wind noise, (c) noisy signal, (d) enhanced signal.

TABLE I
fwSNR improvements using the proposed method with the three wind directions and two wind speeds (iSNR = -10dB)

Speeds \ Directions	0	45°	90°	Average
1.8 m/s	14.61	18.14	17.02	16.59
5 m/s	15.10	17.93	15.84	16.29

TABLE II
fwSNR, PESQ and STOI Results With the proposed method and Baseline Methods

iSNR \ Algorithm / Metrics		ΔfwSNR	ΔPESQ	ΔSTOI
-10 dB	MVDR	4.16	0.45	0.0067
	MWF	12.12	0.93	0.0261
	PMWF	13.44	1.03	**0.1061**
	Proposed	**16.45**	**1.07**	**0.1061**
-5 dB	MVDR	4.03	0.44	0.0083
	MWF	7.44	1.04	0.0826
	PMWF	9.59	1.07	0.0827
	Proposed	**13.37**	**1.21**	**0.0926**
0 dB	MVDR	3.01	0.39	0.0156
	MWF	7.53	1.08	0.0383
	PMWF	7.71	1.27	0.0418
	Proposed	**10.21**	**1.48**	**0.0683**

Furthermore, the proposed method is more efficient and robust in the case of lower iSNR.

V. Conclusions

This paper proposes a multi-channel wind noise reduction method for outdoor recordings. Through analyzing the spatial characteristics of wind noise captured by closely spaced mi-

crophones and modeled by the Corcos model, a comprehensive analysis of the covariance matrix of target speech signals and ambient wind noise is conducted. Leveraging this analysis, we decompose the corrupted signals using the orthogonal polynomial expansion, highlighting the different coefficient distributions on the polynomial basis for speech source signals and wind noise. Then the estimated source signals are reconstructed by choosing low order coefficients corresponding to speech components. An additional contribution is the application of the ADMM framework, offering a more sparse and robust solution in the decomposition phase. The experiment results demonstrated that the proposed method outperforms baseline solutions in terms of fwSNR, PESQ, and STOI.

ACKNOWLEDGMENT

This work was supported in part by the NSFC (Nos. 62171373).

REFERENCES

[1] S. Kuroiwa, Y. Mori, S. Tsuge, M. Takashina, and F. Ren, "Wind noise reduction method for speech recording using multiple noise templates and observed spectrum fine structure," in *2006 International Conference on Communication Technology*. IEEE.

[2] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *2007 IEEE Workshop on Machine Learning for Signal Processing*. IEEE.

[3] B. King and L. Atlas, "Coherent modulation comb filtering for enhancing speech in wind noise," in *International Workshop on Acoustice Echo and Noise Control*, 2008, pp. 14–17.

[4] E. Nemer and W. Leblanc, "Single-microphone wind noise reduction by adaptive postfiltering," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE.

[5] J. A. Zakis, "Wind noise at microphones within and across hearing aids at wind speeds below and above microphone saturation," vol. 129, no. 6, pp. 3897–3907.

[6] C. Hofmann, T. Wolff, M. Buck, T. Haulick, and W. Kellermann, "A morphological approach to single-channel wind-noise suppression," in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*. VDE, 2012, pp. 1–4.

[7] P. Thuene and G. Enzner, "Maximum-likelihood approach to adaptive multichannel-wiener postfiltering for wind-noise reduction," in *Speech Communication; 12. ITG Symposium*. VDE, 2016, pp. 1–5.

[8] C. M. Nelke, *Wind noise reduction: signal processing concepts*. Wissenschaftsverlag Mainz, 2016.

[9] C. M. Nelke, N. Chatlani, C. Beaugeant, and P. Vary, "Single microphone wind noise psd estimation using signal centroids," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7063–7067.

[10] C. M. Nelke and P. Vary, "Wind noise short term power spectrum estimation using pitch adaptive inverse binary masks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5068–5072.

[11] ——, "Dual microphone wind noise reduction by exploiting the complex coherence," in *Speech Communication; 11. ITG Symposium*. VDE, 2014, pp. 1–4.

[12] J. Park, J. Park, S. Lee, J. Kim, and M. Hahn, "Coherence-based dual microphone wind noise reduction by wiener filtering," in *Proceedings of the 8th International Conference on Signal Processing Systems*, 2016, pp. 170–172.

[13] G. W. Elko, J. M. Meyer, and T. F. Gaensler, "Noise-reducing directional microphone array," Mar. 29 2016, uS Patent 9,301,049.

[14] G. M. Corcos, "The structure of the turbulent pressure field in boundary-layer flows," vol. 18, no. 03, p. 353.

[15] R. H. Mellen, "On modeling convective turbulence," vol. 88, no. 6, pp. 2891–2893.

[16] D. Mirabilii and E. A. Habets, "Simulating multi-channel wind noise based on the corcos model," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, sep 2018.

[17] ——, "Multi-channel wind noise reduction using the corcos model," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2019.

[18] A. Herzog, D. Mirabilii, and E. A. P. Habets, "Direction preserving wind noise reduction of b-format signals," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

[19] D. Mirabilii and E. A. P. Habets, "Spatial coherence-aware multi-channel wind noise reduction," vol. 28, pp. 1974–1987.

[20] J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds., *Springer Handbook of Speech Processing*. Springer Berlin Heidelberg.

[21] J. Benesty and C. Jingdong, *Study and design of differential microphone arrays*. Springer Science & Business Media, 2012, vol. 6.

[22] C. Pan, J. Benesty, and J. Chen, "Design of robust differential microphone arrays with orthogonal polynomials," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 1079–1089, aug 2015.

[23] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of signal enhancement and array signal processing*. John Wiley & Sons, 2017.

[24] L. Zhao, J. Benesty, and J. Chen, "Design of robust differential microphone arrays with the jacobi–anger expansion," *Applied Acoustics*, vol. 110, pp. 194–206, sep 2016.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York Inc., Aug. 2006.

[26] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

Sparse Blind Deconvolution and Demixing via Block Majorization-Minimization

Mengting Chen and Ziping Zhao
ShanghaiTech University, Shanghai, China
{chenmt, zipingzhao}@shanghaitech.edu.cn

***Abstract*—To support vastly growing intelligent devices, massive connectivity with low-latency communications has become a critical requirement. In this paper, we consider a multiple-input multiple-output network, where a large amount of devices are connected to an access point sporadically. We aim to simultaneously detect the active devices and recover the transmitted signals from the received mixed measurements, without a priori channel information. The problem is mathematically modeled based on the idea of blind deconvolution and demixing for sparse signals. We formulate the optimization problem via nonconvex matrix factorization, and propose an efficient block majorization-minimization algorithm, where the signals and filters are updated with analytical solutions in an alternating way. The proposed algorithm has much lower per-iteration computational complexity compared to state-of-the-art algorithms and hence is more scalable to large-size problems. Numerical results demonstrate that our method is able to recover the sparse signals and filters with higher precision as well as faster convergence in comparison with existing methods.**

I. Introduction

With the explosion of small and affordable computing devices, the capability of connecting a vast number of devices, or massive device connectivity, has become an essential requirement for future wireless communications. However, since the resources in a cellular network are limited, the devices are designed to keep inactive most of the time unless triggered by external event, leading to a bursty and sporadic signal transmission pattern. For example, in the mobile networks, applications running in the background of smartphones or tablets can periodically send or receive data. In the IoT networks, energy-constrained devices often remain idle to conserve power and only become active intermittently to send data. Furthermore, emerging applications in 5G, such as virtual reality, real-time video conferencing, online gaming, etc., require information to be delivered with ultra-low latency [1], [2]. It is thus urgent to support a communications network with both massive connectivity and low latency.

Massive connectivity in communication mainly has three key challenges. In scenarios with numerous devices, only a subset may need to transmit data intermittently. Therefore, it is necessary to identify active devices at a given time to help manage network resources to minimize idle communication overhead. Once active devices are identified, the system must estimate the channel conditions for each device, which is crucial for coding strategies and thus reduces signal interference. Finally, transmitted data from the active devices need to be accurately recovered. While the three challenges can be addressed separately, a more effective approach is to jointly detect active devices and estimate channel information [3], [4], followed by data recovery [5]. However, this two-stage procedure still introduces additional latency, since it requires channel information estimation for data recovery. To reduce channel signaling overhead, it could be more desirable to simultaneously detect active devices and recovery the transmitted data without knowing channel information [6]. This target can be mathematically modeled based on the idea of blind deconvolution and demixing for sparse signals (SBD^2).

Specifically, the SBD^2 for joint active device detection and transmitting data recovery relies on a bilinear model with a group-sparse assumption on the transmitting signal, which is introduced for active device detection. Since the bilinear structure essentially forms a low-rank matrix [7], [8], the SBD^2 problem can be formulated based on a sparse penalty [9] and a low-rank penalty [10]. A convex approach was proposed as a semidefinite programming (SDP) problem. However, SDP is computationally prohibitive especially for large-scale problems [11]. In [12], the authors used difference-of-convex-functions [13] to represent the rank-one property. An iteratively reweighted SDP (IR-SDP) method was then proposed for problem solving. In each iteration, an SDP is solved. Since solving an SDP needs to scale the problem to a higher-dimensional variable space, both SDP and IR-SDP are computationally expensive, especially for large-scale problems.

In this paper, we consider an SBD^2 problem that arises in a multiple-input multiple-output network. The problem is formulated based on nonconvex matrix factorization. We propose a block majorization-minimization (BMM) algorithm [14]–[16] to address the nonconvex problem. In each iteration, the signals and filters are updated with closed-form solutions. Consequently, BMM exhibits significantly lower per-iteration computational complexity compared to SDP and IR-SDP, making it more suitable for problems of large sizes. We also prove that the proposed BMM algorithm is equivalent to a proximal coordinate gradient descent algorithm. Numerical simulations verify that BMM recovers the unknown sparse signals and filters with higher precision and faster convergence than the existing methods.

II. Problem Formulation

We consider a multiple-input multiple-output (MIMO) communication system, where S devices with a single antenna are connected to a base station with R antennas; see Fig. 1. We

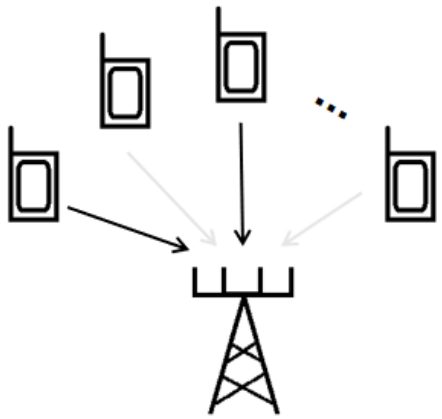


Fig. 1: In a MIMO system, multiple devices connect to a BS with several antennas. Black arrows show active uplink transmissions, while grey lines indicate inactive connections.

consider a block-fading model where channels are independent quasi-static flat-fading in each coherence time. In the massive access scenario, only a small fraction of S devices are active and access the base station in a coherence block. In the uplink transmission, let $\mathbf{x}_k \in \mathbb{C}^N$ be the data sequence of length N from the device k ($k = 1, \ldots, S$). When device k is active, $\mathbf{x}_k \neq \mathbf{0}$; otherwise, $\mathbf{x}_k = \mathbf{0}$. The signal sequence transmitted by the device k over M time slots is $\mathbf{A}_k\mathbf{x}_k$, where $\mathbf{A}_k \in \mathbb{C}^{M\times N}$ is a preassigned encoding matrix available to the base station. The signal $\mathbf{A}_k\mathbf{x}_k$ is passing through a channel $\mathbf{h}_{rk} \in \mathbb{C}^L$ to antenna r ($r = 1, \ldots, R$), where $L < M$ is the maximal delay spread of the finite impulse response. Denote $\mathring{\mathbf{h}}_{rk} = [\mathbf{h}_{rk}^{\mathsf{T}}, \mathbf{0}]^{\mathsf{T}} \in \mathbb{C}^M$ as the zero-padded filter. The transmission process is represented by a linear convolution of $\mathbf{A}_k\mathbf{x}_k$ and $\mathbf{h}_{rk}$. It can also be approximately represented by a circular convolution between $\mathbf{A}_k\mathbf{x}_k$ and $\mathring{\mathbf{h}}_{rk}$.[1] Then, the signal received by the antenna r at the base station is

$$\mathbf{y}_{\circledast,r} = \sum_{k=1}^{S} (\mathbf{A}_k\mathbf{x}_k) \circledast \mathring{\mathbf{h}}_{rk} + \mathbf{e}_{\circledast,r},$$

where $\circledast$ denotes the circular convolution and $\mathbf{e}_{\circledast,r}$ is an additive noise. Denote $\mathbf{F}$ as the M-point unitary discrete Fourier transform matrix. The received signal in frequency domain is

$$\mathbf{y}_r := \mathbf{F}\mathbf{y}_{\circledast,r} = \sum_{k=1}^{S} \sqrt{M}\,(\mathbf{F}\mathbf{A}_k\mathbf{x}_k) \odot (\mathbf{B}\mathbf{h}_{rk}) + \mathbf{e}_r, \qquad (1)$$

where $\odot$ denotes the Hadamard product, $\mathbf{B}$ contains the first L columns of $\mathbf{F}$, and $\mathbf{e}_r = \mathbf{F}\mathbf{e}_{\circledast,r}$.

Remark: In practice, the encoding matrix $\mathbf{A}_k$ is often designed to have a convenient structure which can lead to fast computations and minimal memory requirements. For example, it can be chosen as $\mathbf{A}_k = \mathsf{Diag}\,(\mathbf{q}_k)\,\mathbf{P}$, where $\mathbf{q}_k \in \mathbb{C}^M$ is a vector with entries being ± 1 and $\mathbf{P} \in \mathbb{C}^{M\times N}$ is a partial matrix containing the N columns of an $M \times M$ Hadamard/Fourier matrix [17].

Our goal is to recover nonzero signals $\mathbf{x}_k$'s of active devices from the observations $\{\mathbf{y}_r\}_{r=1}^R$. To reduce the channel signaling overhead, we assume that $\mathbf{h}_{rk}$'s are not available to both receivers and transmitters. Then, the problem is to simultaneously recover the message signals $\mathbf{x}_k$'s and the channels $\mathbf{h}_{rk}$'s from $\mathbf{y}_r$'s. Mathematically, the task of interest can be modeled as an SBD[2] problem.

Define

$$\mathbf{Y} = \left[\begin{array}{ccc} \mathbf{y}_1 & \cdots & \mathbf{y}_R \end{array}\right],$$

and

$$\mathbf{H}_k = \left[\begin{array}{ccc} \mathbf{h}_{1k} & \cdots & \mathbf{h}_{Rk} \end{array}\right],$$

for $k = 1, \ldots, S$. We propose a nonconvex optimization problem for SBD[2] in ($\star$),[2] where the objective is the combination of a fitting-error term and a group sparsity penalty $\sum_{k=1}^S \|\mathbf{x}_k\|$ with γ a tuning parameter. It can be easily verified that in model (1), each signal and filter pair $\{\mathbf{x}_k, \mathbf{h}_{rk}\}$ in model can be only identified up to a scaling factor; i.e., $\{\mathbf{x}_k^\star, \mathbf{h}_{rk}^\star\}$ and $\{\alpha\mathbf{x}_k^\star, \alpha^{-1}\mathbf{h}_{rk}^\star\}$ for any $\alpha \neq 0$ lead to the same output. To alleviate such ambiguity, we restrict the filters $\mathbf{h}_{rk}$ to have unit length in ($\star$) without affecting the optimality.

III. The Block Majorization-Minimization Method

Block majorization-minimization (BMM) [14], [16] can be viewed as a judicious combination of the block coordinate descent (BCD) method [18] and the majorization minimization (MM) method [15]. The BCD method aims to find a local optimal solution by optimizing the objective along one variable block each time and solving for different blocks successively. One potential limitation of BCD is the requirement that each subproblem needs to be solved exactly, which could be difficult for nonconvex subproblems. BMM removes the above restriction by optimizing in each variable block a subproblem where the original objective function in BCD is replaced by a surrogate function which can lead to cheap iterations.

Specifically, given an optimization problem as follows:

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \quad f(\mathbf{x}),$$

[1] In OFDM, with cyclic prefixing technique, the two types of convolution are equivalent.

[2] Throughout the paper, $\|\cdot\|$ is used to denote the ℓ_2-norm for vectors or the Frobenius norm for matrices.

$$\begin{array}{ll} \underset{\{\mathbf{x}_k,\mathbf{H}_k\}_{k=1}^S}{\text{minimize}} & \dfrac{1}{2}\left\|\displaystyle\sum_{k=1}^{S}\sqrt{M}\,((\mathbf{F}\mathbf{A}_k\mathbf{x}_k)\otimes\mathbf{1})\odot(\mathbf{B}\mathbf{H}_k) - \mathbf{Y}\right\|^2 + \gamma\displaystyle\sum_{k=1}^{S}\|\mathbf{x}_k\| \\ \text{subject to} & \|\mathbf{h}_{rk}\| = 1,\ k = 1,\ldots,S,\ r = 1,\ldots,R. \end{array} \qquad (\star)$$

where $\mathcal{X} \subseteq \mathbb{R}^N$. Suppose the optimization variable $\mathbf{x}$ can be partitioned into I blocks as $\mathbf{x} \triangleq (\mathbf{x}_1, \ldots, \mathbf{x}_I)$ where $\mathbf{x}_i \in \mathcal{X}_i$ and $\mathcal{X} = \prod_{i=1}^I \mathcal{X}_i$ with $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ and $\sum_{i=1}^I n_i = N$. At each iteration of BMM, one variable block, say, $\mathbf{x}_i$, is updated according to the following update rules:

$$\begin{cases} \mathbf{x}_i^+ \in \arg\min_{\mathbf{x}_i \in \mathcal{X}_i} \ \bar{f}_i(\mathbf{x}_i, \underline{\mathbf{x}}), \\ \mathbf{x}_{-i}^+ = \underline{\mathbf{x}}_{-i} \text{ with } \mathbf{x}_{-i} \triangleq (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_I), \end{cases}$$

where $\bar{f}_i$ is an upper bound function for f_i with respect to variable $\mathbf{x}_i$, and $\mathbf{x}_i^+$ and $\mathbf{x}_{-i}^+$ are the newly updated variables. The algorithm is monotonic and iteratively runs until some convergence criterion is met. Choosing the surrogate function $\bar{f}_i$ is the crucial part in BMM algorithm development. Generally speaking, they could be derived in multiple ways. However, a properly chosen one by taking the specific problem structure into account will make the iterative updating steps cheap while maintaining a fast algorithm convergence over iterations. In practice, surrogate functions will be much appreciated if they will lead to analytical solutions for the blockwise subproblems.

IV. Proposed Algorithm

In this section, we propose a BMM-based algorithm to effectively solve the problem ($\star$). The matrix factorization structure inherent in ($\star$) allows the algorithm to iteratively update the signals and filters in an alternating fashion. Furthermore, by finding suitable surrogate functions that exploit the blockwise structure of the problem, the two variables are updated with analytical solutions.

A. Updating the Signals

With fixed $\underline{\mathbf{H}}_k$'s,[3] we derive the update rule for $\mathbf{x}_k$'s. Let $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$ and $\mathbf{x} = \left[\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_S^\mathsf{T}\right]^\mathsf{T}$. Define a block matrix $\mathbf{C} \in \mathbb{C}^{MR \times NS}$ with $\mathbf{C}_{rk} = \sqrt{M}\mathrm{Diag}\left(\mathbf{B}\underline{\mathbf{h}}_{rk}\right)\mathbf{F}\mathbf{A}_k \in \mathbb{C}^{M \times N}$ as the block at the r-th row and the k-th column. Problem ($\star$) with respect to (w.r.t.) variable $\mathbf{x}$ can be written as

$$\underset{\mathbf{x}}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{C}\mathbf{x} - \mathbf{y}\|^2 + \gamma \sum_{k=1}^S \|\mathbf{x}_k\|. \tag{2}$$

To apply the BMM technique, we first introduce the following useful lemma.

Lemma 1. *[15] Let $\mathbf{L}$, $\mathbf{M}$ be $n \times n$ Hermitian matrices satisfying $\mathbf{M} - \mathbf{L} \succeq \mathbf{0}$. Then for any $\underline{\mathbf{x}} \in \mathbb{C}^n$, it follows that*

$$\mathbf{x}^\mathsf{H}\mathbf{L}\mathbf{x} \le \mathbf{x}^\mathsf{H}\mathbf{M}\mathbf{x} - 2\mathsf{Re}\left\{\mathbf{x}^\mathsf{H}(\mathbf{M} - \mathbf{L})\underline{\mathbf{x}}\right\} + \underline{\mathbf{x}}^\mathsf{H}(\mathbf{M} - \mathbf{L})\underline{\mathbf{x}},$$

where the equality is attained at $\mathbf{x} = \underline{\mathbf{x}}$.

The first term in the objective of (2) contains the quadratic term over $\mathbf{x}$, which is $\frac{1}{2}\mathbf{x}^\mathsf{H}\mathbf{C}^\mathsf{H}\mathbf{C}\mathbf{x}$. Based on Lemma 1, taking $\mathbf{C}^\mathsf{H}\mathbf{C}$ as $\mathbf{L}$, we can construct a surrogate (i.e., upper-bound) function for the quadratic term by choosing $\mathbf{M} = C\mathbf{I}$ where C is a scalar satisfying $C \ge \left\|\mathbf{C}^\mathsf{H}\mathbf{C}\right\|_2$. Here, $\left\|\mathbf{C}^\mathsf{H}\mathbf{C}\right\|_2$ is the spectral norm of $\mathbf{C}^\mathsf{H}\mathbf{C}$. A surrogate function for the fitting-error term can thus be designed as

$$\begin{aligned} \frac{1}{2}\|\mathbf{C}\mathbf{x} - \mathbf{y}\|^2 &\le \sum_{k=1}^S \left(\frac{C}{2}\mathbf{x}_k^\mathsf{H}\mathbf{x}_k - \mathsf{Re}\left\{\mathbf{m}_k^\mathsf{H}\mathbf{x}_k\right\}\right) + \text{const.} \\ &= \sum_{k=1}^S \frac{C}{2}\left\|\mathbf{x}_k - C^{-1}\mathbf{m}_k\right\|^2 + \text{const.}, \end{aligned} \tag{3}$$

where $\mathbf{m} = \left(C\mathbf{I} - \mathbf{C}^\mathsf{H}\mathbf{C}\right)\underline{\mathbf{x}} + \mathbf{C}^\mathsf{H}\mathbf{y}$ and $\mathbf{m}_k$ contains the $(N(k-1)+1, \ldots, Nk)$-th elements of $\mathbf{m}$ for $k = 1, \ldots, S$. Since the surrogate function in (3) has a decoupled structure over S devices, each $\mathbf{x}_k$ can be solved in a parallel way, leading to solve the following sub-problem

$$\underset{\mathbf{x}_k}{\text{minimize}} \quad \frac{1}{2}\left\|\mathbf{x}_k - C^{-1}\mathbf{m}_k\right\|^2 + C^{-1}\gamma\|\mathbf{x}_k\|,$$

which is a group soft-thresholding step [19] and the solution is given by

$$\mathbf{x}_k^+ = \left(1 - \gamma\left\|\mathbf{m}_k\right\|^{-1}\right)^+ C^{-1}\mathbf{m}_k, \tag{4}$$

where $(x)^+ \coloneqq \max\{x, 0\}$.

B. Updating the Filters

In the filter-updating step, we first rewrite the problem in a more compact way. Let $\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_S \end{bmatrix} \in \mathbb{C}^{LS \times R}$. The subproblem with respect to $\mathbf{H}$ is a constrained least-squares problem:

$$\begin{aligned} \underset{\mathbf{H}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{D}\mathbf{H} - \mathbf{Y}\|^2 \\ \text{subject to} \quad & \|\mathbf{h}_{rk}\| = 1, \ k = 1, \ldots, S, \ r = 1, \ldots, R, \end{aligned} \tag{5}$$

where $\mathbf{D} \in \mathbb{C}^{M \times LS}$ is a block matrix with its k-th block being $\mathbf{D}_k = \sqrt{M}\mathrm{Diag}\left(\mathbf{F}\mathbf{A}_k\underline{\mathbf{x}}_k\right)\mathbf{B} \in \mathbb{C}^{M \times L}$.

Similar to the last subsection, we aim to find a surrogate function for the objective in (5). Expanding the objective, we obtain a quadratic term $\frac{1}{2}\mathrm{Tr}\left(\mathbf{H}^\mathsf{H}\mathbf{D}^\mathsf{H}\mathbf{D}\mathbf{H}\right)$. Based on Lemma 1, by choosing D such that $D\mathbf{I} \succeq \left\|\mathbf{D}^\mathsf{H}\mathbf{D}\right\|_2$, we can find a surrogate function of the quadratic term as follows:

$$\begin{aligned} \frac{1}{2}\|\mathbf{D}\mathbf{H} - \mathbf{Y}\|^2 &\le \frac{D}{2}\mathbf{h}^\mathsf{H}\mathbf{h} - \mathsf{Re}\left\{\mathbf{n}^\mathsf{H}\mathbf{h}\right\} + \text{cst.} \\ &= \sum_{r=1}^R \sum_{k=1}^S \frac{D}{2}\left\|\mathbf{h}_{rk} - D^{-1}\mathbf{n}_{rk}\right\|^2 + \text{cst.}, \end{aligned} \tag{6}$$

where $\mathbf{N} = \left(D\mathbf{I} - \mathbf{D}^\mathsf{H}\mathbf{D}\right)\underline{\mathbf{H}} + \mathbf{D}^\mathsf{H}\mathbf{Y} \in \mathbb{C}^{LS \times R}$, $\mathbf{h} = \mathrm{vec}\left(\mathbf{H}\right) \in \mathbb{C}^{LSR}$, and $\mathbf{n}_{rk} \in \mathbb{C}^L$ is a subvector consisting of the $(L(rk-1)+1, \ldots, Lrk)$-th elements of $\mathbf{n} = \mathrm{vec}\left(\mathbf{N}\right)$. Then, $\mathbf{H}$ can be found by solving SR subproblems,

$$\underset{\|\mathbf{h}_{rk}\|=1}{\text{minimize}} \quad \left\|\mathbf{h}_{rk} - D^{-1}\mathbf{n}_{rk}\right\|^2, \tag{7}$$

[3] In this paper, underlined variables denote those whose values are given.

By solving the Karush–Kuhn–Tucker conditions of (7), we derive its closed-form solution

$$\mathbf{h}_{rk}^{+} = \frac{\mathbf{n}_{rk}}{\|\mathbf{n}_{rk}\|}. \tag{8}$$

The overall algorithm is summarized in Algorithm 1.

Algorithm 1 BMM algorithm for Problem ($\star$)

1: **Input:** $\{\mathbf{y}_r\}_{r=1}^{R}$, γ;
2: **Initialize:** $\left\{\mathbf{x}_k^{(0)}\right\}_{k=1}^{S}$, $\left\{\mathbf{h}_{rk}^{(0)}\right\}_{k,r=1}^{S,R}$;
3: **repeat**
4: update $\{\mathbf{x}_k\}_{k=1}^{S}$ in parallel by equation (4);
5: update $\{\mathbf{h}_{rk}\}_{k,r=1}^{S,R}$ in parallel by equation (8);
6: **until some convergence criterion is satisfied**
7: **Output:** $\{\mathbf{x}_k\}_{k=1}^{S}$, $\{\mathbf{h}_{rk}\}_{k,r=1}^{S,R}$.

C. Complexity Analysis

We analyze the per-iteration computational complexity of Algorithm 1. In each iteration, the computational cost of updating signals comes from matrix multiplication in $\mathbf{m}$, with a time complexity of $\mathcal{O}\left(MNRS^2 \log M\right)$; and the computational cost of updating filters comes from matrix multiplication in $\mathbf{N}$, with a time complexity of $\mathcal{O}\left(MNRS^2 \log M\right)$. The time complexity of $\mathcal{O}\left(MN \log M\right)$ is due to the use of fast Fourier transform in some matrix multiplications.

V. Interpreting Algorithm 1 as Proximal Coordinate Gradient Descent

In developing the BMM algorithm 1, the surrogates (3) and (6) are the isotropic quadratic approximations of the objectives in (2) and (5), respectively. This provides BMM with an intriguing parallel to the proximal coordinate gradient descent method. Denote the fitting-error term in the objective ($\star$) as f. The gradients of f with respect to $\mathbf{x}_k$ and $\mathbf{h}_{rk}$ are $\nabla_{\mathbf{x}_k} f$ and $\nabla_{\mathbf{h}_{rk}} f$, respectively.

In updating the signals, by plugging the expression of $\mathbf{m}_k$ into (4), we can verify that the update rule (4) of $\mathbf{x}_k$ is the proximal gradient descent with stepsize C^{-1}, i.e.

$$\begin{aligned}\mathbf{x}_k^{+} &= \operatorname{prox}_{C^{-1}\gamma\|\cdot\|_1}\left(C^{-1}\mathbf{m}_k\right) \\ &= \operatorname{prox}_{C^{-1}\gamma\|\cdot\|_1}\left(\underline{\mathbf{x}}_k - C^{-1}\nabla_{\mathbf{x}_k} f\right),\end{aligned}$$

where $\operatorname{prox}_{\lambda g(\cdot)}(\mathbf{z}) = \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|^2 + \lambda g(\mathbf{x})$. Likewise, the filter update rule (8) can be rewritten as

$$\begin{aligned}\mathbf{h}_{rk}^{+} &= \operatorname{prox}_{\mathbb{I}(\|\cdot\|=1)}\left(D^{-1}\mathbf{n}_{rk}\right) \\ &= \operatorname{prox}_{\mathbb{I}(\|\cdot\|=1)}\left(\underline{\mathbf{h}}_{rk} - D^{-1}\nabla_{\mathbf{h}_{rk}} f\right),\end{aligned}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Therefore, the two BMM updates can be viewed as proximal gradient descent steps for both $\mathbf{x}_k$'s and $\mathbf{h}_{rk}$'s.

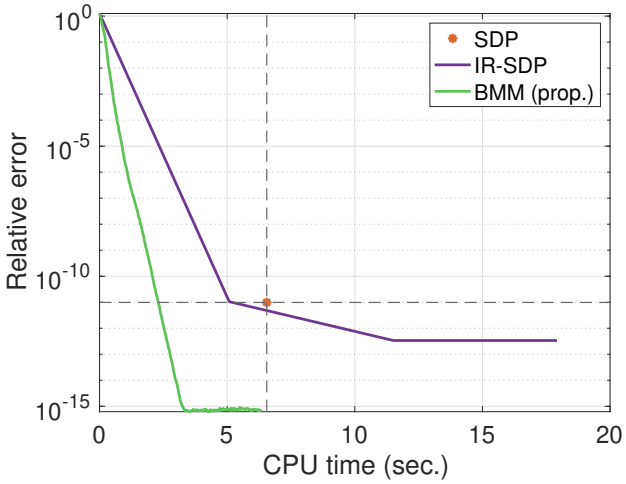


Fig. 2: Comparison of CPU time.

VI. Numerical Experiments

In this section, we illustrate the recovery performance of the proposed algorithm and compare it with existing methods, including SDP [9] and IR-SDP [12], where the first is the convex method and the second employs a nonconvex low-rank inducing function. Suppose we have $S = 10$ devices, of which only 2 send messages to a base station that has $R = 2$ antennas. We draw nonzero elements of the ground truth $\{\mathbf{x}_k^{\star}\}_{k=1}^{S}$ and $\{\mathbf{h}_{rk}^{\star}\}_{k,r=1}^{S,R}$ independently from the standard complex Gaussian distribution and set $\mathbf{h}_{rk}^{\star}$'s to have unit length. The encoding matrices $\mathbf{A}_k$'s are chosen based on the design in Section II. 20 Monte Carlo experiments are performed and the average results are presented. Each trial of the iterative methods is initialized from randomly selected starting points. The tuning parameter γ which controls the level of sparsity is found using grid search.

First we conduct experiments in the noiseless case to

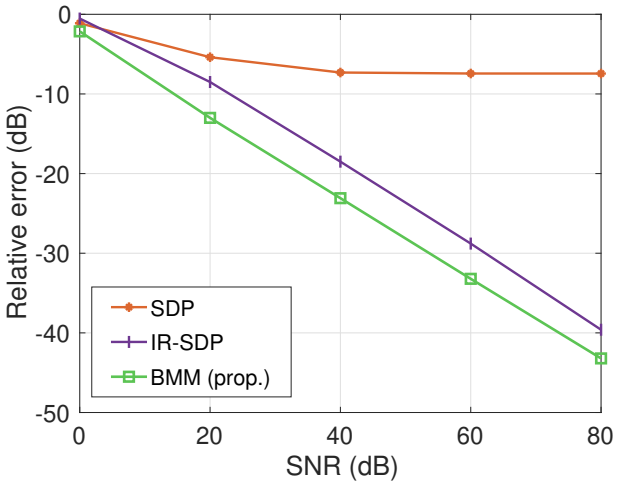


Fig. 3: Performance under different SNR.

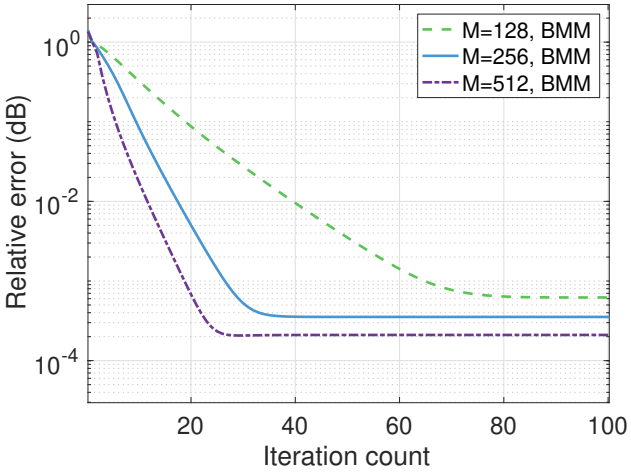


Fig. 4: Effectiveness of the proposed BMM.

compare the convergence speed of BMM with the existing methods. We set $M = 128$, $N = L = 8$, and $\gamma = 0.002$. The proposed BMM algorithm stops when the number of iterations exceeds 300, while IR-SDP stops when the difference of objective values between two sequential iterations, which is denoted as Δ, is no larger than 10^{-8}. The relative error (RE) between the estimate and the ground truth is defined as

$$\text{RE} = \frac{\sum_{r=1}^{R}\sum_{k=1}^{S}\left\|\hat{\mathbf{h}}_{rk}\hat{\mathbf{x}}_k^{\mathsf{H}} - \mathbf{h}_{rk}^{\star}\mathbf{x}_k^{\star\mathsf{H}}\right\|}{\sum_{r=1}^{R}\sum_{k=1}^{S}\left\|\mathbf{h}_{rk}^{\star}\mathbf{x}_k^{\star\mathsf{H}}\right\|}.$$

RE of each method against CPU time is presented in Fig. 2. It shows that both nonconvex algorithms achieve better estimation than SDP under this setting, and BMM converges the fastest among the three methods. Then we compare the recovery performance of the algorithms in a noisy scenario. We set $M = 128$, $N = L = 16$, and choose the noise to be complex white Gaussian. The stop criterion for IR-SDP is the same as above. We let BMM stop when $\Delta \leq 10^{-10}$ or the number of iterations exceeds 300. The signal-to-noise ratio (SNR) is defined as

$$\text{SNR} = 10\log_{10}\frac{\sum_{r=1}^{R}\sum_{k=1}^{S}\left\|\mathbf{h}_{rk}^{\star}\mathbf{x}_k^{\star\mathsf{H}}\right\|}{\sum_{r=1}^{R}\left\|\mathbf{e}_r\right\|^2}.$$

For different SNR, we set the tuning parameter to be $\gamma \in \{0.05, 0.02, 0.01, 0.005, 0.001\}$. RE in dB scale against SNR of the algorithms are illustrated in Fig. 3. It depicts that RE of BMM and IR-SDP decrease linearly with SNR, whereas RE of SDP changes flatly particularly at higher SNR levels. The underlying reason for the saturation problem of SDP may be that the convex relaxation is loose. The figure also shows that BMM achieves lower RE than IR-SDP, yielding better recovery performance. Moreover, we perform BMM with different sample sizes to verify its effectiveness. For $M \in \{128, 256, 512\}$, we set $N = L = 16$, $\gamma \in \{0.005, 0.003, 0.001\}$, SNR $= 60$, and run 100 iterations under each setting. It is interesting to observe that with arbitrary initial points, BMM converges linearly despite the problem size, as shown in Fig. 4. Finally, to see the scalability of these methods, we implement them under different settings and compare their CPU time costs when they converge. As displayed in Table I, BMM takes far less time than existing methods and hence is scalable to large-scale problems.

TABLE I: CPU time (sec.) comparison in the noiseless case.

M	$N = L$	IR − SDP	SDP	BMM (prop.)
256	8	28.05	8.93	**2.09**
	16	261.39	89.57	**4.52**
512	16	693.55	188.10	**3.49**
	32	11006.13	4370.12	**13.72**

VII. Conclusions

In this paper, we have proposed a BMM algorithm to jointly detect active devices and recover signals in MIMO communications, without prior knowledge of channel information. Based on the matrix factorization structure, the signals and the filters have been updated in an alternating way with low-cost closed-form solutions. Numerical results have indicated that BMM not only outperforms both SDP and IR-SDP methods in terms of recovery accuracy but also reduces the computational time significantly. Theoretical analyses of the convergence rate and performance guarantees in MIMO systems have not yet been provided, which suggests a potential future direction for our research.

References

[1] E. Bastug, M. Bennis, M. Medard, and M. Debbah, "Toward interconnected virtual reality: Opportunities, challenges, and enablers," *IEEE Comm. Mag.*, vol. 55, no. 6, pp. 110–117, 2017.

[2] K. S. Kim, D. K. Kim, C. Chae, *et al.*, "Ultrareliable and low-latency communication techniques for tactile internet services," *Proc. of the IEEE*, vol. 107, no. 2, pp. 376–393, 2019.

[3] L. Liu and W. Yu, "Massive connectivity with massive MIMO-Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, 2018.

[4] J. Ahn, B. Shim, and K. B. Lee, "EP-based joint active user detection and channel estimation for massive machine-type communications," *IEEE Trans. Comm.*, vol. 67, no. 7, pp. 5178–5189, 2019.

[5] L. Liu and W. Yu, "Massive connectivity with massive MIMO-Part II: Achievable rate characterization," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2947–2959, 2018.

[6] M. Chen and Z. Zhao, "Joint blind deconvolution and demixing of sparse signals via factorization and nonconvex optimization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 9176–9180.

[7] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1711–1732, 2014.

[8] A. Ahmed, "A convex approach to blind MIMO communications," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 812–815, 2018.

[9] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2886–2908, 2015.

[10] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proc. Amer. Control Conf.*, vol. 6, 2001, pp. 4734–4739.

[11] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM review*, vol. 38, no. 1, pp. 49–95, 1996.

[12] Y. Shi, J. Dong, and J. Zhang, *Low-overhead Communications in IoT Networks: Structured Signal Processing Approaches*, English. Springer, 2020.

[13] J. Y. Gotoh, A. Takeda, and K. Tono, "DC formulations and algorithms for sparse optimization problems," *Math. Program.*, vol. 169, no. 1, pp. 141–176, 2018.

[14] M. Razaviyayn, M. Hong, and Z. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[15] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, 2017.

[16] A. Nedić, J. Pang, G. Scutari, and Y. Sun, *Parallel and distributed successive convex approximation methods for big-data optimization*. Springer, 2018, vol. 2224, pp. 141–308.

[17] S. Ling and T. Strohmer, "Blind deconvolution meets blind demixing: Algorithms and performance bounds," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4497–4520, 2017.

[18] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[19] N. Parikh and S. P. Boyd, "Proximal algorithms," *Foundations Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.

An End-to-End Two-Stream Network Based on RGB Flow and Representation Flow for Human Action Recognition

Song-Jiang Lai[1,2], Tsun-Hin Cheung[1], Ka-Chun Fung [2], Tian-Shan Liu[2], Kin-Man Lam[1,2]

[1]*Center for Advances in Reliability and Safety, New Territories, Hong Kong*

[2]*Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. Kowloon, Hong Kong*

***Abstract*—With the rapid advancements in deep learning, computer vision tasks have seen significant improvements, making two-stream neural networks a popular focus for video-based action recognition. Traditional models using RGB and optical flow streams achieve strong performance but at a high computational cost. To address this, we introduce a representation flow algorithm to replace the optical flow branch in the egocentric action recognition model, enabling end-to-end training while reducing computational cost and prediction time. Our model, designed for egocentric action recognition, uses class activation maps (CAMs) to improve accuracy and ConvLSTM for spatio-temporal encoding with spatial attention. When evaluated on the GTEA61, EGTEA GAZE+, and HMDB datasets, our model matches the accuracy of the original model on GTEA61 and exceeds it by 0.65% and 0.84% on EGTEA GAZE+ and HMDB, respectively. Prediction runtimes are significantly reduced to 0.1881s, 0.1503s, and 0.1459s, compared to the original model's 101.6795s, 25.3799s, and 203.9958s. Ablation studies were also conducted to study the impact of different parameters on model performance.**

Keywords: two-stream, egocentric, action recognition, CAM, representation flow, CAM, ConvLSTM

I. INTRODUCTION

Early action recognition techniques primarily focused on detecting and representing interest points using methods such as spatio-temporal point detection [1] and optical flow histograms [2]. Although these methods are effective for feature extraction, they struggle with large datasets due to their lack of scalability. Heng et al. [3] introduced dense trajectory features, which improve motion representation but still face scalability issues. Recent advances in deep learning have surpassed traditional methods. Karpathy et al. [4] developed deep models that process temporal information and combine dense optical flow with deep features. Despite these advancements, limitations persisted, including challenges with long videos and pre-computed optical flow. Innovations, such as spatial-temporal CNN integration [5] and LSTM-based time fusion [6], further advanced the research field of action recognition.

Zhou et al. [7] improved action recognition accuracy through weighted fusion and sequential reasoning. Temporal attention models and ConvLSTM variants [8] have enhanced the focus on important frames and preserved spatial structures, proving valuable in applications like anomaly and violence detection. Concurrently, research on optical flow has evolved, with Fan et al. [9] enhancing accuracy with the TV-L1 method, and Sun et al. [10] introducing the optical flow-guided feature (OFF). Piergiovanni et al. [11] advanced this with the representation flow algorithm, offering a faster and more accurate alternative to traditional optical flow.

Our model combines the strengths of the egocentric activity recognition model proposed by Swathikiran et al. [12] with the representation flow algorithm, yielding an end-to-end trainable system. By fusing input RGB frames with representation flow, our method not only reduces prediction time but also enhances accuracy. This is achieved by leveraging the combined benefits of both the original EgoRCNN model [12] and the advanced representation flow model.

II. METHODOLOGY

The proposed activity recognition method comprises two components: deep feature encoding for the RGB stream employing spatial attention and class activation maps (CAMs), and a fully differentiable convolutional layer for the representation stream, which is developed from optical flow methods. The entire framework is designed to be trainable end-to-end, allowing all parameters to be optimized jointly to enhance model performance.

A. Class Activation Maps

A novel mechanism known as Class Activation Mapping (CAM) is introduced by Zhou et al. [13] to produce the class-specific saliency maps by leveraging the average pooling layer in contemporary deep CNN architectures, such as ResNet. Let w_l^c represent the weight for unit l corresponding to class c,

and $a_l(i)$ denote the activation values at spatial location i of unit l in the last convolutional layer. Furthermore, the CAM for class c, $F_l(i)$, can be calculated as follows:

$$F_c(i) = \sum_l w_l^c \, a_l(i). \quad (1)$$

Therefore, the CAM, F_c, can be used to identify the image regions that the CNN utilized to recognize the class was being considered (i.e., class c). Subsequently, we select the class that exhibits the highest probability, referred to as the winning class. This allows us to produce a saliency map using the CAM.

B. ConvLSTM Module

The structural block diagram of the RGB stream shows how RGB frames are encoded, is shown in **Figure 1**. ResNet-34 serves as the backbone, responsible for generating spatial attention maps, extracting relevant features, and predicting class categories. The training is conducted in two stages: Stage 1 focuses on training only the classifier and ConvLSTM layer, while Stage 2 involves training the additional convolutional and fully connection layers of the ResNet-34 network. The pre-trained weights based on the ImageNet dataset are utilized by the rest parts of the model. ConvLSTM is used for temporal aggregation, and Equation (1) is employed to compute the CAM for each RGB frame. The CAM is converted into an attention map via the softmax function along the spatial dimension, which is then multiplied with the output of the last convolutional layer. The attention map is computed as follows:

$$f_{SA}(i) = f_i \odot \frac{e^{M_c(i)}}{\sum_{i'} e^{M_c(i)}}, \quad (2)$$

Where f_i represents the output feature for the final layer of ResNet-34 at the spatial location i. $f_{SA}(i)$ denotes the Hadamard product of f_i and the softmax output, which indicates the spatial attention map applied to image features. $M_c(i)$ indicates the use of the winning category c to obtain the corresponding CAM. The symbol $\odot$ is used to denote the Hadamard product. After obtaining visual features with spatial attention approximated by Equation (2), temporal encoding for frame-level information is required.

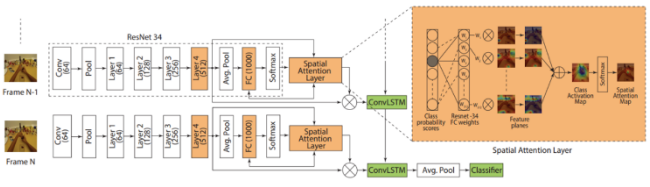


Figure 1: The RGB stream structure of the EgoRCNN model [12].

ConvLSTM works as shown by these following equations:

$$I_t = \sigma\left(w_x^i * f_{SA} + w_h^i * h_{t-1} + b^i\right), \quad (3)$$

$$f_t = \sigma\left(w_x^f * f_{SA} + w_h^f * h_{t-1} + b^f\right), \quad (4)$$

$$\tilde{c_t} = \tanh\left(w_x^{\tilde{c}} * f_{SA} + w_h^{\tilde{c}} * h_{t-1} + b^{\tilde{c}}\right), \quad (5)$$

$$C_t = \tilde{c_t} \odot f_{SA} + C_{t-1} \odot f_t, \quad (6)$$

$$O_t = \sigma(w_x^o * f_{SA} + w_h^o * h_{t-1} + b^o), \quad (7)$$

$$h_t = O_t \odot \tanh(C_t), \quad (8)$$

Where σ denotes the sigmoid function, I_t, f_t, and O_t are the input state, forget state, and hidden state, respectively, w and b represent the learnable weights and biases of the ConvLSTM.

C. The Representation Flow Algorithm

The representation flow algorithm shares many similarities but also has notable differences with the traditional optical flow algorithm. As detailed by Piergiovanni et al. [11], while both algorithms aim to capture motion information between frames, the representation flow layer is designed to extract features from deep feature maps, whereas optical flow is applied to video frames in the spatial domain. Additionally, the representation flow algorithm differs in its learning parameters: optical flow involves fixed hyper-parameters, such as τ, θ, and λ, which are manually set, whereas the representation flow algorithm utilizes trainable parameters that are updated during model training.

Denote F_1 and F_2 as the feature maps from two consecutive frames, with a single channel. These feature maps are then convolved with the Sobel operators, as shown in Equation (9), to compute the gradients of the feature maps.

$$\nabla F_{i,x} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} * F_i, \quad \nabla F_{i,y} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * F_i. \quad (9)$$

Firstly, we introduce the flow field: u, and the dual vector fields: p, which are used to minimize the energy, and set them to zero. Then, **Algorithm 1** is applied iteratively until the minimum total variational energy is achieved.

Algorithm 1: Method for the representation flow layer [11]

Initialize u and p, as follows:

$$u = 0, p = 0 \quad (10)$$

Compute the feature gradients using Equation (9), and then compute the difference as follows:

$$\rho_c = F_2 - F_1 \quad (11)$$

Perform the following operations, with n iterations:

$$\rho = \rho_c + \nabla_x F_2 \cdot u_x + \nabla_y F_2 \cdot u_y \quad (12)$$

$$v = \begin{cases} u + \lambda\theta\nabla F_2, & \rho < -\lambda\theta|\nabla F_2|^2 \\ u - \lambda\theta\nabla F_2, & \rho > \lambda\theta|\nabla F_2|^2 \\ u - \rho\dfrac{\nabla F_2}{|F_2|^2}, & \text{otherwise} \end{cases} \quad (13)$$

$$u = v + \theta \cdot divergence(p) \quad (14)$$

$$p = \frac{p + \frac{\tau}{\theta}\nabla u}{1 + \frac{\tau}{\theta}|u|} \quad (15)$$

Finally, after the loop, the flow u is obtained.

Following **Algorithm 1**, zero-pad p in the x and y directions and convolute them with the weights w_x and w_y to calculate the divergence:

$$divergence(p) = p_x * w_x + p_y * w_y, \quad (16)$$

∇u_x and ∇u_y are computed as follows:

$$\nabla u_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} * u_x, \nabla u_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * u_y. \quad (17)$$

D. Computing Flow-of-Flow

In principle, continuously stacking representation flow layers can enhance motion information extraction and representation, as more flow layers can more comprehensively capture the direction and magnitude of motion. However, a greater number of flow layers often results in more intricate motion features and textures. Applying the representation flow technique to sequential flow images may lead to inconsistent optical flow and non-rigid movements, hence impairing model performance. Inserting standard convolution layers between representation flow layers aids in preserving optical flow consistency and enhancing motion representation to resolve this issue. The comprehensive architecture of the video-CNN, integrated with the representation flow, which is demonstrated in **Figure 2**.

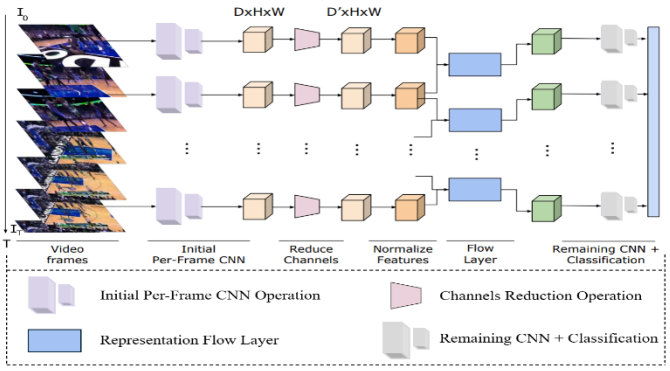


Figure 2: The comprehensive architecture of a video recognition CNN model with a representation flow layer [11].

E. Applying Representation Flow Layer for Action Recognition

In our proposed model, as shown in **Figure 3**, the representation flow layer replaces the original optical flow branch in a CNN model, integrating into the EgoRCNN framework. This refined model combines the RGB and representation flow branches, employing two classifier layers for joint training. Both branches take RGB frames as input. The CNN produces predictions for each time step, which are then averaged throughout the time dimension for generating the category probabilities. The hybrid network is trained by reducing the cross-entropy loss, as outlined below:

$$L(v, c) = -\sum_{i}^{k} (c == i) \log(p_i), \quad (18)$$

Where M is the classification model, v represents the video, c denotes a specific category among all categories, L refers to the value of the cross-entropy loss function, and $p = M(v, c)$.

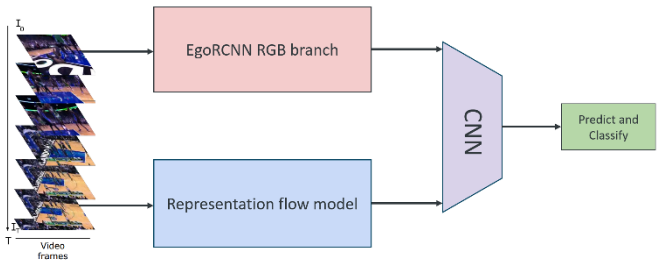


Figure 3: The overall structure of proposed model.

III. Experiment Results

A. Implementation Details

In this study, we employ RGB features and motion features, based on representation flow, for action recognition. The ConvLSTM with 512 hidden units is utilized by the RGB branch of our model for spatio-temporal coding and ResNet-34 as the backbone for extracting frame-level image features and spatial attention maps. Training is performed in two stages: Stage 1 spans 200 epochs with a learning rate of 10^{-3}, adjusted at the 25, 75, and 150 epochs, while Stage 2 involves 150 epochs with a learning rate of 10^{-4}, adjusted at the 25 and 75 epochs, utilizing the ADAM optimizer and a batch size of 32 with 25 frames sampled per video.

The representation flow branch utilizes ResNet-34 as the backbone and consists of two representation flow layers, separated by an intermediate convolutional layer to smooth the flow. The input video clip length is set to 16 frames for optimal performance. The number of channels is reduced to 32 by a 1×1 convolutional layer, balancing speed and performance, while a 3×3 convolutional layer restores the original channel count. The flow branch is trained for 750 epochs with

a batch size of 16, employing stochastic gradient descent with a learning rate of 10^{-3} and a decay factor of 0.5.

Then fuse the two branches at the decision-making level by adding a classifier to the top of the RGB branch and a new fully connected layer to integrate the outputs after separately training the RGB and representation flow branches. The final training phase consists of 250 epochs with a batch size of 32, utilizing a learning rate of 1 and a decay factor of 0.1 per epoch.

B. Experimental Results

GTEA61

Initially, we assessed the efficacy of our model using split2 of the GTEA61 dataset and contrasted it with the original EgoRCNN model that included the optical flow branch. The findings shown in **Table 1** indicate that the representation flow model's accuracy on split2 is 37.74, while the optical flow component of the original EgoRCNN model attains 39.42. Following the collaborative training of the two streams, both our proposed model and the original EgoRCNN model attained an accuracy of 69.84. Since the size of GTEA 61 dataset is too small, the superiority of our model cannot be reflected.

Model	Accuracy(%)	Time(s)
Optical flow	39.42	N/A
Representation flow	37.74	N/A
Two- stream (optical flow)	69.84	101.6795
Two-stream (representation flow)	69.84	0.1881

Table 1: Comparison of the original and proposed two-stream models in terms of accuracy and inference runtime on the GTEA61 dataset

The runtime required for classifying samples demonstrates significant differences between the models. Extracting optical flow is notably time-intensive, with the original model requiring 101.55 seconds pre video for this process. As shown in Table 1, the average runtime for classifying a sample using the original model is 0.1295 seconds, resulting in a total average runtime of 101.6795 seconds when optical flow extraction is performed. In contrast, our proposed model, which eliminates the need for optical flow extraction, achieves a significantly shorter average runtime of only 0.1881 seconds.

EGTEA GAZE+

We trained and assessed the proposed model on the larger EGTEA GAZE+ dataset. This dataset is significantly larger than the GTEA61 dataset, providing a more comprehensive basis for our analysis. The prediction accuracies of the two models are tabulated in **Table 2**. The assessment was conducted across three splits of the EGTEA GAZE+ dataset. The results illustrate the proposed model consistently outperforms the original model in terms of accuracy.

Method	Split1	Split2	Split3	Average
Two-stream (optical flow)	61.76	61.38	58.17	60.44
Two-stream (representation flow)	62.49	61.64	59.15	61.09

Table 2: Comparison of the original and proposed two-stream models in terms of accuracy on the EGTEA GAZE+ dataset

Table 3 shows that the original model requires an average runtime of 0.0999 seconds to predict the class category of a sample. However, it necessitates an additional 25.28 seconds for the extraction of optical flow, bringing the total runtime to 25.3799 seconds. In contrast, our proposed model exhibits remarkable efficiency, requiring only 0.1503 seconds for the classification process. This also represents a significant improvement in terms of speed and efficiency.

Method	Time(s)
Two-stream (joint train with optical flow)	25.3799
Two-stream (joint train with representation flow)	0.1503

Table 3: Comparison of the original and proposed two-stream models in terms of prediction runtime on the EGTEA GAZE+ dataset

HMDB

The superiority of our proposed model is previously assessed using the GTEA61 and EGTEA GAZE+ datasets, both of which are first-person perspective datasets. To further illustrate the generalizability and robustness of our proposed model, we expanded our experiments to the HMDB dataset, which provides a third-person perspective and has a modest sample size. The experimental findings on the HMDB dataset are shown in **Table 4**, illustrating the performance of our proposed model across various perspectives and sample sizes.

Our proposed model outperforms the original model can also be demonstrated by the prediction results on the HMDB dataset. Specifically, while the original model achieved an accuracy of 22.89, our proposed model reached 24.34. With joint training of both two streams, the accuracy for the original model improves to 49.87, compared to 50.71 for the proposed model. In terms of runtime, as shown in **Table 4**, the original model takes an average runtime of 203.9958 seconds, including 0.1058 seconds for classification and 203.89 seconds for optical flow extraction. In contrast, the proposed model significantly reduces the average runtime to only 0.1459 seconds.

Methods	Accuracy (%)	Time (s)
RGB flow only	50.00	N/A
Optical flow only	22.89	N/A
Representation flow only	24.34	N/A
Two-stream with optical flow	49.87	203.9958
Two-stream with representation flow	50.71	0.1459

Table 4: Comparison of the original and proposed two-stream models in terms of accuracy and runtime on the HMDB dataset

C. *Ablation study*

Optimal number of representation flow layers

The number of representation flow layers incorporated into a CNN model significantly impacts its performance. We investigated the effects of setting different number of representation flow layers to 1, 2, 3 and none on our proposed model. Evaluations were conducted on the HMDB dataset using consistent experimental settings, and the results are tabulated in **Table 5**. Our findings indicate the model with two representation flow layers achieved the highest accuracy compared to the configurations with 1, 3 or no flow layers.

Configuration	Accuracy(%)
Without representation flow layer	20.08
1 representation flow layer	23.15
2 representation flow layers	24.34
3 representation flow layers	23.11

Table 5: Recognition accuracy of the proposed model with different numbers of representation flow layers on the HMDB dataset

Optimal number of iterations for training

The performance of our model highly relies on the number of iterations used to compute the flow in the representation flow layer. We conducted experiments using 10, 20, 30, and 50 iterations, and the results are shown in **Table 6**. It is clear our model achieves the highest accuracy when 20 iterations are adopted for computing representation flow.

Number of iterations	Accuracy (%)
10	21.58
20	24.34
30	23.78
50	23.14

Table 6: Recognition accuracy of the proposed model with different number of iterations used for computing representation flow on the HMDB dataset

Effects of backbone size

The depth of a deep model can influence its tendency to overfit, thus affecting its accuracy and generalization. In order to investigate this, we implemented experiments that involved altering the size of backbone for the representation flow branch from ResNet-18 to ResNet-101. The performance is influenced by the model depth, as demonstrated in **Table 7**. It is evident the accuracy of the proposed model is at its highest when the backbone of our model is configured to ResNet-34.

Backbone size	Accuracy (%)
ResNet-18	21.03
ResNet-34	24.34
ResNet-50	24.27
ResNet-101	22.89

Table 7: Recognition accuracy of the proposed model with backbone of different sizes on the HMDB dataset

Comparison to other state-of-art methods

Table 8 and **Table 9** summarize the results of numerous experiments conducted to assess the efficacy of our proposed model and compare it to state-of-the-art methods. The methods in **Table 8** focus on egocentric activity recognition tasks, while **Table 9** addresses the third-person activity recognition tasks. The experiment results in Table 8 illustrate the accuracies on the fixed split2 of GTEA61, and for **Table 9**, it shows the prediction accuracies on each split of EGTEA GAZE+.

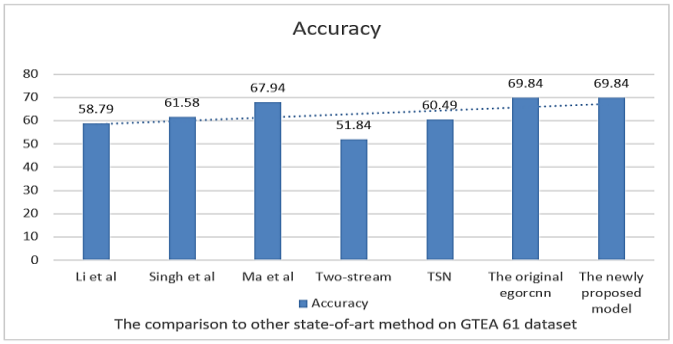


Table 8: Comparison of the proposed model with state-of-art methods in terms of prediction accuracy (%) on the GTEA 61 dataset

Li et al. [14] proposed a technique that uses hand segmentation and gaze information for detecting object locations in the first two datasets, alongside the Gaze+ dataset. Ma et al. [15] developed a network that integrates object localization and hand segmentation to enhance the accuracy of object and self-location detection. In contrast, Singh et al. [16] introduced the ego-flow input into a two-stream model for egocentric motion recognition tasks. Besides, Carreira et al. [17] introduced the I3D method, which extends 2D convolutions to 3D, enabling spatio-temporal feature learning for improved video action recognition. The TSN structure sparsely samples

video clips and aggregates their predictions for video-level human action recognition is proposed by Limin Wang et al. [18]. In the prediction accuracy results from **Table 8** and **Table 9**, we can observe a significant improvement when comparing the proposed model to other state-of-the-art methods.

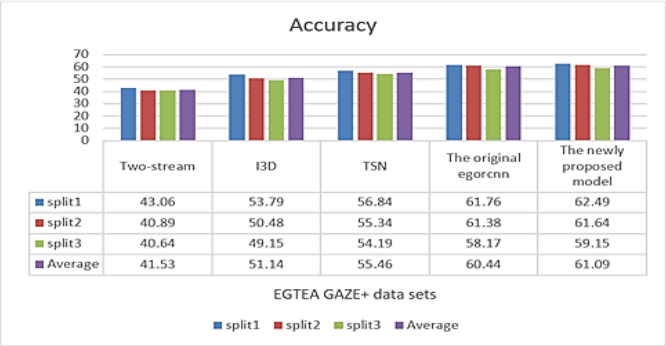


Table 9: Comparison of accuracy of the proposed model and state-of-the-art methods across different splits on the EGTEA GAZE+ data sets

IV. Conclusions

In this study, we have proposed a learnable representation flow layer as a replacement for the traditional optical flow branch in the original EgoRCNN model. Our model leverages the class attention map (CAM) of the RGB stream, thereby enhancing its ability to focus on activity-relevant regions. We further incorporate Conv-LSTM for spatio-temporal encoding with spatial attention. This has resulted in an improved performance over the original EgoRCNN model in terms of both accuracy and prediction runtime. This improvement is evident across three standard datasets: two of which are from the first-person perspective and the remaining one from the third-person perspective.

References

[1] Laptev, Ivan. "On space-time interest points." *International journal of computer vision* 64 (2005): 107-123.

[2] Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008, June). Learning realistic human actions from movies. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1-8). IEEE.

[3] Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." *Proceedings of the IEEE international conference on computer vision*. 2013.

[4] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).

[5] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[6] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4694-4702).

[7] Zhou, Bolei, et al. "Temporal relational reasoning in videos." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[8] Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2017, February). An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).

[9] Fan, L., Huang, W., Gan, C., Ermon, S., Gong, B., & Huang, J. (2018). End-to-end learning of motion representation for video understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6016-6025.

[10] Sun, Shuyang, et al. "Optical flow guided feature: A fast and robust motion representation for video action recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[11] Piergiovanni, A. J., and Michael S. Ryoo. "Representation flow for action recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[12] Sudhakaran, Swathikiran, and Oswald Lanz. "Attention is all we need: Nailing down object-centric attention for egocentric activity recognition." *arXiv preprint arXiv:1807.11794* (2018).

[13] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).

[14] Li, Yin, Zhefan Ye, and James M. Rehg. "Delving into egocentric actions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[15] Ma, Minghuang, Haoqi Fan, and Kris M. Kitani. "Going deeper into first-person activity recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[16] Singh, Suriya, Chetan Arora, and C. V. Jawahar. "First person action recognition using deep learned descriptors." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[17] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[18] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20-36). Springer, Cham.

Few-Shot Open-Set Keyword Spotting with Multi-Stage Training

Lo-Ya Li*, Tien-Hong Lo†, Jeih-Weih Hung‡, Shih-Chieh Huang§ and Berlin Chen¶
* Institute of AI Interdisciplinary Applied Technology, National Taiwan Normal University, Taiwan
E-mail: 612k0010c@ntnu.edu.tw
†¶ Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
E-mail: {teinhonglo, berlin}@ntnu.edu.tw
‡ Department of Electrical Engineering, National Chi Nan University, Taiwan
E-mail: jwhung@ncnu.edu.tw
§ Realtek Semiconductor Corp., Taiwan

***Abstract*—As the advance of human-computer interaction technologies continued, keyword spotting (KWS) systems have gained prominence in everyday devices. This study is dedicated to exploring innovative approaches for few-shot keyword recognition under open-set conditions, a challenging yet crucial area in speech processing. To this end, we design and develop a multi-stage training method that synergistically combines the advantages of acoustic and phonetic features, thereby substantially enhancing the ability of a KWS model. By learning multi-type features with joint training from only one dataset, our KWS model is equipped with a more robustness feature extractor to deal with few-shot KWS. Experimental results demonstrate that our model outperforms strong baselines by achieving a 15% improvement in recognition accuracy on open-set tests in a 10shot-10way setting. This research confirms the effectiveness of our multi-stage strategy and suggests promising directions for future development in keyword recognition technologies.**

Index Terms: Keyword spotting, few-shot learning

I. Introduction

Keyword spotting (KWS) is a pivotal technology for human-machine interaction, enabling seamless control of personal and household devices during multitasking activities such as driving and exercising. This technology can be categorized into two primary types: fixed vocabulary KWS [1], [2] and flexible customized KWS [3]–[6]. While traditional fixed vocabulary KWS methods have performed robustly in data-rich scenarios, they have faced significant challenges in environments with limited data. Moreover, the demand for flexible KWS systems has surged recently, driven by an increasing need for more personalized voice-activated control devices and applications. To address the issue of data scarcity in user-defined keywords, numerous studies have proposed various approaches employing few-shot learning (FSL) techniques in KWS, aiming to achieve effective recognition of keywords with minimal speech input from users.

The advent of few-shot learning (FSL) provided a new direction for KWS research, particularly for customized applications. Studies have explored the use of few-shot learning techniques to train KWS models that could adapt to new keywords with minimal user input. These studies primarily used large pre-trained datasets to extract generalizable features that could be fine-tuned with a small number of examples [7]. Recent advancements have focused on enhancing the robustness of feature extraction in FSL settings by integrating acoustic features with linguistic data. For instance, some models have been developed to combine traditional acoustic features with phoneme-based information, achieving a balance between generalization and specificity. This approach facilitated more accurate predictions of user-defined keywords by aligning audio features with their corresponding phonetic sequences [8].

In addition to leveraging phoneme-based information, the concept of prototype networks (ProtoNets) [9] has been crucial in the evolution of FSL for KWS. ProtoNets compute class prototypes as the mean of feature vectors, significantly simplifying the adaptation process to new and rare keywords. However, recent work extended this concept by incorporating standard deviation measures into the ProtoNet calculations, capturing a wider range of intra-class feature variability and enhancing the model's ability to handle ambiguously represented classes [10], [11].

To further improve the performance of few-shot keyword spotting (FS-KWS), our research has focused on using a multi-stage strategy. Previous studies on FS-KWS predominantly leveraged prior knowledge from extensively annotated pre-training datasets to develop a robust feature extractor, often depending exclusively on audio sample features or on a combination of phoneme features and labeled text information for alignment purposes [12].

In our approach, we proposed a multi-stage framework [13]–[15] to explore the synergy between acoustic and phonetic features, constructing a more sophisticated and robust feature extractor to enhance FS-KWS performance. Specifically, we leveraged the dynamic properties of the acoustic spectrum—such as spectral energy variations—and the articulatory characteristics inherent in phoneme sequences. To achieve this integration, we employed two distinct models: a feature extractor model focused on deriving robust representations from acoustic features, and an additional Transformer-based model that predicted the phoneme sequence of a keyword. This

phoneme sequence [16], [17] was then matched against a text-derived phoneme sequence to assist the main feature extractor model. Both models were trained jointly using a shared loss function, which encouraged them to learn complementary information and improved their generalization capabilities. By integrating these diverse feature sets, our method not only captured a broad spectrum of speech nuances but also adapted more effectively to new keywords and varied acoustic environments. This dual-model approach allowed for a richer and more adaptive learning process, resulting in improved performance across standard keyword spotting benchmarks.

Building on the aforementioned innovations, this paper presented a multi-stage framework for FSL architectures consisting of a feature encoder and a prototype-based open-set classifier. This classifier was initialized with few-shot open-set samples. We utilized the recent Multilingual Spoken Words Corpus (MSWC) dataset [18], from which we obtained both acoustic and textual information. Our methodology employed two distinct models for joint training to develop a robust feature extractor. Specifically, the feature extractor model used the ConvMixer architecture [19], [20], optimized with triplet loss, while the phoneme model utilized a Transformer architecture [21], [22] coupled with Connectionist Temporal Classification (CTC) loss. During each epoch, the losses from both models were aggregated and recursively applied in a joint training process. Upon completing the training of this powerful feature extractor model, we used the Google Speech Commands (GSC) v2 dataset to define prototypes for each category based on a few-shot, few-way setup. This approach not only simplified the adaptation to new or rare categories but also significantly enhanced the model's ability to generalize from a very limited number of examples. This made it particularly effective for applications such as keyword identification, where the ability to quickly adapt to new user-defined keywords was critical.

II. Method

In this section, we describe our approach to classifying speech commands using a multi-stage process. Our method consisted of three main stages: pretrain, prototype setting, and inference. Each stage played a crucial role in building and utilizing the model for accurate speech command classification. The complete architecture in Fig.1.

A. Stage 1: Pretrain

We utilized the MSWC dataset for pretraining. This dataset contained single-word audio files in multiple languages along with their corresponding text transcriptions.

First, we extracted features from the audio data. This step involved transforming raw audio signals into higher-dimensional feature vectors that better captured key information in the audio. We employed three different feature extractor models for this purpose:

a) Depthwise Separable Convolutional Neural Network (DSCNN): DSCNN [23] effectively extracted audio features using depthwise separable convolutions, reducing model parameters and computational complexity.

b) Broadband Convolutional Residual Network (BC_ResNet): This model [24], based on residual networks, was specifically designed to process audio features. It improved the training of deep networks through residual connections.

c) Convolutional Mixer (ConvMixer): ConvMixer [25] combined convolutional and mixing layers to extract multi-scale features, enhancing the model's ability to perceive different audio patterns.

The extracted audio features were used to compute triplet loss (1) [26]. A distance metric learning method that maximizes the distance between positive x_i^+ and negative x_i^- samples while minimizing the distance between positive samples and the anchor. This helped improve the discriminative ability of the feature extractor, ensuring that similar audio files were closer in the feature space.

$$\mathcal{L}_{TL} = -\frac{1}{B_t}\sum_{i=1}^{B_t}\max(\delta) \tag{1}$$

$$\delta = \big(0, d_{L2}(x_i, x_i^+) - d_{L2}(x_i, x_i^-) + \text{margin}\big) \tag{2}$$

Simultaneously, we utilized audio-to-phoneme and text-to-phoneme conversion models to convert the single-word audio from the MSWC dataset into corresponding phoneme sequences. For example, the word "Hello" was converted to ['HH', 'EH1', 'L', 'OW0']. To compare the phoneme sequences predicted by the model with the target phoneme sequences, we used CTC loss (3). It's suitable for aligning input and output sequences of variable lengths.

$$\mathcal{L}_{CTC} = -\log P(y|x) \tag{3}$$

where $P(y|x)$ is the probability of the target sequence y given the input sequence x. Then the losses from both models were aggregated and recursively applied in a joint training process. The total loss (4) as follows:

$$\mathcal{L}_{TOTAL} = \mathcal{L}_{TL} + \mathcal{L}_{CTC} \tag{4}$$

B. Stage 2: Set Prototype

In this stage, we used the GSC v2 dataset. This dataset contained a large number of speech samples, each belonging to a specific speech command category. The purpose of this stage was to extract feature prototypes for each category from the dataset for use in subsequent classification tasks.

We utilized the feature extractors pretrained in Stage 1 to process the GSC v2 dataset. The parameters of these extractors were frozen and not updated at this stage.

For each speech command category, we computed the mean of its feature vectors $x_{j,i}^S$ as the prototype for that category i. This mean feature vector c_i represented the centroid of the category (5). We also computed the standard deviation σ_i of

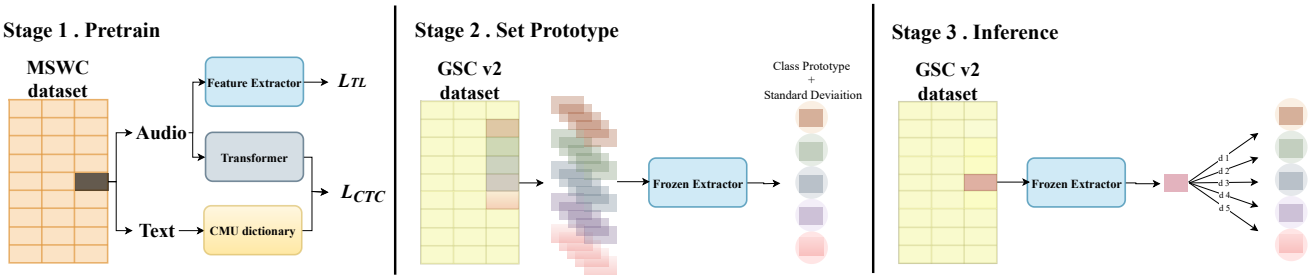


Fig. 1. Overall architecture of the multi-stage training process for few-shot keyword spotting. Used large dataset to pretrain a feature extractor, then GSC dataset to do few shot few way's keyword spotting.

the feature vectors for each category to capture the variability in the feature distribution of that category (6). These measures were used as baselines in subsequent classification tasks.

$$c_i = \frac{1}{S}\sum_{j=1}^{S} f(x_{j,i}^{S}) \tag{5}$$

$$\sigma_i = \sqrt{\frac{1}{S}\sum_{j=1}^{S}\left(f(x_{j,i}^{S}) - c_i\right)^2} \tag{6}$$

C. Stage 3: Inference

The purpose of this stage was to classify speech commands using the class prototypes set in Stage 2. We utilized the feature extractors pretrained in Stage 1 and frozen in Stage 2 to process the GSC v2 dataset. Each audio file was passed through the frozen feature extractors, generating corresponding high-dimensional feature vectors. These feature vectors captured key information in the audio file.

For each extracted feature vector, we calculated its distance to all class prototypes using two distance metrics: Euclidean distance, which refers to the length of the line segment in space connecting these two points. Mahalanobis distance (7) is an effective method to calculate the similarity of two unknown sample sets. Unlike Euclidean distance, it takes into account the connection between various characteristics. Then based on the calculated distances, we assigned the audio file to the class prototype with the smallest distance. This process ensured that each audio file was classified into the most matching category.

$$d_{MA}(\vec{X},\vec{Y}) = \sqrt{(\vec{x}-\vec{y})^T\Sigma^{-1}(\vec{x}-\vec{y})} \tag{7}$$

The objective of the inference stage was to achieve accurate speech command classification using pretrained models and the set class prototypes. This method leveraged the discriminative power of the pretrained feature extractors and the representational power of the class prototypes, making the classification process more reliable and precise.

The multi-stage approach described in this method section aimed to leverage the strengths of deep learning for feature extraction and distance-based classification for speech command recognition. By pretraining robust feature extractors, setting accurate class prototypes, and utilizing efficient distance metrics during inference, our method provided a reliable and precise classification framework.

III. EXPERIMENTS

A. Experimental Setup

a) Datasets: During the pretraining stage, we utilized the MSWC dataset. For each keyword, we selected a subset of samples based on the smallest sample size among all keywords. Specifically, we randomly selected 500 samples for each keyword, except for those from the GSC v2 dataset, resulting in an average of 5,470 samples per class. Additionally, we incorporated background noise from the DEMAND dataset, with Signal-to-Noise Ratio (SNR) values ranging from 0 to 5 dB. For fine-tuning, we employed the GSC v2 dataset, which comprises 35 keywords. From these, we selected 10 keywords (yes, no, up, down, left, right, on, off, stop, go) for our experiments. We conducted a 10-shot 10-way classification, utilizing 10 samples per keyword for training. Moreover, we implemented the openNCM method to incorporate an "unknown" class for words not included in the selected 10 keywords. The "unknown" class was comprised of samples from the words "backward," "forward," "visual," "follow," and "learn," averaging these five words to form the "unknown" class.

b) models: The DSCNN model had 407k parameters, the BC_ResNet model had 817k parameters, and the ConvMixer model had 119k parameters. The models utilized Mel Frequency Cepstral Coefficients (MFCC) for feature extraction. Audio signals were divided into short frames, each lasting 40 milliseconds, with a window stride of 50%. This configuration resulted in each window overlapping the previous one by 20 milliseconds. The resulting feature map had a size of 49x10, where each time step consisted of 49 frames, and each frame contained 10 MFCC features. The models were trained over 40 epochs, with each epoch consisting of 400 episodes. The Adam optimizer was employed with an initial learning rate of 0.001, and the models were evaluated every 20 episodes using a learning rate decay of 0.1. Furthermore, the models were also evaluated using an online learning method to adapt to unseen classes.

TABLE I
SUMMARY OF MODEL PERFORMANCE ON THE GSC TESTSET UNDER 5,10-SHOT 10-WAY OPEN-SET CLASSIFICATION SETTINGS.

Feature Extractor	Parameters	Transformer	5 shot		10 shot	
			ACC↑	AUROC↑	ACC↑	AUROC↑
TC_ResNet[7]	61k	X	0.52	0.63	0.58	0.67
Ours_TC_ResNet	61k	O	0.61	0.71	0.66	0.73
DSCNN[7]	407k	X	0.71	0.93	0.76	0.94
Ours_DSCNN	407k	O	0.80	0.87	0.91	0.93
Ours_BC_ResNet	817k	O	0.84	0.90	0.92	0.94
Ours_ConvMixer	119k	O	0.83	0.87	0.93	0.92

TABLE II
DIFFERENCES IN ACCURACY PERFORMANCE BETWEEN DISTANCES CALCULATED.

Feature Extractor	Distance Metrics		10 shot
	Euclidean	Mahalanobis	ACC↑ (%)
Ours_BC_ResNet	O		0.86
Ours_BC_ResNet		O	0.92
Ours_ConvMixer	O		0.84
Ours_ConvMixer		O	0.93

B. Experimental Results

Table 1 included the results for all the experiments. We initially utilized the TC_ResNet and DSCNN model's results on few-shot experiments with acoustic features of audio files as a reference baseline. Subsequently, we implemented our proposed method, employing the same dataset and simultaneously extracting both acoustic and phonetic features for joint training to develop a robust feature extractor. Additionally, we introduced the concept of standard deviation in the computation of prototypes and used Mahalanobis distance formulas. By incorporating a multi-stage training approach, the original accuracy increased from 76% to 86%. Further modifications involved changing the definition of few-shot, few-way prototypes and switching to the Mahalanobis distance, which elevated the experimental accuracy up to 91%, thereby achieving superior classification performance. Furthermore, considering the requirement for on-device operation, we opted for the BC_ResNet and ConvMixer feature encoder, which achieved a lower parameter count and also get a higher accuracy.

Table 2 illustrates the various methods of defining the open-set's few-shot prototype concept, as well as the different distances calculated between the test samples and the prototypes. We observed that the same pretrained model, as shown in rows 1 and 2, achieved better accuracy with the Mahalanobis distance when different distance metrics were applied. The table also indicates that averaging few-shot samples into a single vector prototype might result in some imbalanced outcomes when testing with 400 samples per class, which significantly contributes to the final evaluation phase.

IV. CONCLUSIONS

This study has systematically explored the implementation and optimization of keyword spotting systems under open-set conditions using few-shot learning frameworks. Our experimental results get an accuracy to 93%,and also affirm the critical influence of both multi-stage's feature extraction training and the choice of distance metrics on the performance of keyword spotting models. Specifically, the use of MFCC alongside sophisticated models has demonstrated considerable promise in enhancing the robustness and accuracy of keyword recognition.

A significant finding from our research is the superiority of the Mahalanobis distance over conventional Euclidean distance in computing the similarity between test samples and prototypes. This approach not only improved accuracy but also ensured more stable and reliable performance across various testing scenarios. Furthermore, our exploration into prototype averaging methods revealed that single vector representation might introduce imbalances, particularly when a large number of samples per class are used in testing phases. This insight underscores the necessity for adaptive and dynamic prototype calculation methods in few-shot learning environments.

In conclusion, our research contributes valuable insights into the development of adaptable, efficient, and accurate keyword spotting systems, paving the way for further innovations in the field of speech processing technology. Future studies may focus on refining the integration of phonetic features with acoustic signals and exploring the potential of neural network architectures in enhancing the generalizability of keyword spotting systems under diverse operational conditions.

REFERENCES

[1] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural network," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4087–4091.

[2] R. Tang and J. J. Lin, "Deep residual learning for small footprint keyword spotting," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5484–5488.

[3] J. Jung, Y. Kim, J. Park, Y. Lim, B. -Y. Kim, Y. Jang and J. -S. Chung, "Metric learning for user-defined keyword spotting," in 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5.

[4] L. Lei, G. Yuan, H. Yu, D. Kong and Y. He, "Multilingual customized keyword spotting using similar-pair contrastive learning," in 2023 IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2437-244.

[5] H.-K. Shin, H. Han, D. Kim, S.-W. Chung, and H.-G. Kang, "Learning audio-text agreement for open-vocabulary keyword spotting," in 2022 Proc. Interspeech, pp. 1871–1875.

[6] Z. Yang, S. Sun, J. Li, X. Zhang, X. Wang, L. Ma and L. Xie, "CaTT-KWS: A multi-stage customized keyword spotting framework based on cascaded transducer-transformer," in 2022 Interspeech, arXiv: 2207.01267.

[7] M. Rusci and T. Tuytelaars, "Few-shot open-set learning for on-device customization of keyWord spotting systems," in 2023 Interspeech, pp. 1-5.

[8] Y. Li, C. Yu, G. Sun, H. Jiang, F. Sun, W. Zu, Y. Wen, Y. Yang and J. Wang, "Cross-utterance conditioned VAE for non-autoregressive text-to-speech," in 2022 Association for Computational Linguistics (ACL), pp. 1-10.

[9] J. Snell, K. Swersky and R. S. Zemel, "Prototypical networks for few-shot learning," in 2017 Advances in neural information processing systems 30, arXiv: 1703.05175.

[10] S. Yang, B. Kim, K. Shim and S. Chang, "Improving small footprint few-shot keyword spotting with supervision on auxiliary data," 2023, arXiv: 2309.00647.

[11] W. Lin, X. Tang, X. Han, J. Ma, X. Zhang and L. Jiao, "NQ-Protonet: Noisy query prototypical network for few-shot remote sensing scene classification," in 2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, pp. 3620-3623.

[12] Y.-H. Lee and N. Cho, "iPhonMatchNet: Zero-shot user-defined keyword spotting using implicit acoustic echo cancellation," in 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5.

[13] Z. Liu, T. Li, and P. Zhang, "Rnn-t based open-vocabulary keyword spotting in mandarin with multi-level detection," in 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5649–5653.

[14] R. Yang, G. Cheng, H. Miao, T. Li, P. Zhang, and Y. Yan, "Keyword search using attention-based end-to-end asr and frame-synchronous phoneme alignments," in 2021 IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3202–3215.

[15] S. Sigtia, J. Bridle, H. Richards, P. Clark, E. Marchi, and V. Garg, "Progressive voice trigger detection: Accuracy vs latency," in 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6843–6847.

[16] K. Nishu, M. Cho and D. Naik, "Matching latent encoding for audio-text based keyword spotting," 2023, arXiv: 2306.05245.

[17] K. Nishu, M. Cho, P. Dixon and D. Naik, "Flexible keyword spotting based on homogeneous audio-text embedding," in 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5050-5054.

[18] M. Mazumder, S. Chitlangia, C. Banbury, Y. Kang, J. M. Ciro, K. Achorn, D. Galvez, M. Sabini, P. Mattson, D. Kanter, et al, "Multilingual spoken words corpus," in Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[19] D. Ng, R. Zhang, J.Q. Yip, C. Zhang, Y. Ma, T.H. Nguyen, C. Ni, E.S. Chng and B. Ma, "Contrastive speech mixup for low-resource keyword spotting," in 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5.

[20] ND. Ng, Y. Chen, B. Tian, Q. Fu and E. S. Chng, "Convmixer: Feature interactive convolution with curriculum learning for small footprint and noisy far-field keyword spotting," in 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3603-3607.

[21] B. Axel, O'C. Mark and C. M. Tairum, "Keyword Transformer: A self-attention model for keyword spotting," in 2021 Proc. Interspeech, pp. 1-5.

[22] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu et al, "Conformer: Convolution-augmented transformer for speech recognition," in 2020 Proc. Interspeech, pp. 1-5.

[23] M. Li, C. Ma, W. Dang, R. Wang, Y. Liu and Z. Gao, "DSCNN: Dilated shuffle CNN model for SSVEP signal classification," in 2022 IEEE Sensors Journal, vol. 22, no. 12, pp. 12036-12043.

[24] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, pp. 5484-5488.

[25] D. Ng, Y. Chen, B. Tian, Q. Fu and E. S. Chng, "Convmixer: Feature interactive convolution with curriculum learning for small footprint and noisy far-field keyword spotting," in 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3603-3607.

[26] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in 2017 Interspeech, pp. 1487–1491.

Self-Supervised Augmented Diffusion Model for Anomalous Sound Detection

Jiawei Yin[1, 2, †], Wenbin Zhang[2], Mingjun Zhang[2] and Yu Gao[2, *]
[1] Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education,
East China University of Science and Technology, Shanghai, 200237, China
[2] AI Research Center, Midea Group (Shanghai) Co.,Ltd., Shanghai 201702, China
E-mail: y30221023@mail.ecust.edu.cn, {zhangwb87, zhangmj139, gaoyu11}@midea.com

***Abstract*—Generative models have significantly enhanced the capability of unsupervised anomalous sound detection (ASD) with their strong data modeling capabilities. However, many existing ASD methods based on generative models focus solely on accurately reconstructing sound data itself, neglecting the use of metadata. This results in limited features learned by the models. Additionally, these methods suffer from issues such as low generation quality and mode collapse. To address these challenges, we propose a self-supervised augmented diffusion model (SSDM) to improve ASD performance. SSDM learns expressive embeddings through a self-supervised learning module with a dual-path time-frequency self-attention ASD framework and then uses a denoising diffusion module to learn the distribution of these embeddings as the basis for anomaly detection. The reconstruction loss, which is derived from the reconstruction of test data embeddings measured by the self-supervised learning module in the denoising diffusion module, is used as the anomaly score. Experiments on the DCASE Challenge 2023 Task 2 development dataset demonstrate the effectiveness of the proposed method.**

I. Introduction

Anomalous sound detection (ASD) is a task that involves distinguishing between normal and abnormal states of a machine by analyzing the sounds it emits [1]. However, due to the infrequency and diversity of anomalous sounds, it is challenging to collect data that covers all possible anomalies. Consequently, this task primarily involves learning the distribution of normal sounds to detect sounds that deviate from this distribution and classify them as anomalies. The main approach in existing methods is to use generative models to minimize the reconstruction error of normal sounds during the training phase, and then use the reconstruction error as an anomaly score during the inference phase. By effectively capturing the complex distribution patterns in the data, generative models have become an intuitive and widely applied method for ASD [2]–[4].

Despite the success of generative models, they still face several challenges. Firstly, generative models are designed to learn the distribution of normal data, so the optimization objectives of these methods are primarily focused on the audio data itself [5]. This often leads to insufficient utilization of metadata [6], such as machine type or operating conditions. This approach is suboptimal because operating conditions often contain features that reflect the differences in sounds of the same type of machine [7]. As a result, the audio features learned by generative models for the same machine type can become more complex, leading to unstable ASD performance. Secondly, Variational Autoencoders (VAEs) [8]–[10] and Generative Adversarial Networks (GANs) [3], [4], [11] are the most commonly used generative models for ASD. However, VAEs tend to generate samples of lower quality, while GANs are prone to mode collapse, which hinders the accurate detection of anomalous sounds.

Recently, several self-supervised classification methods have been proposed [7], [12]–[14], leveraging metadata such as machine type and operating conditions as classification conditions. These methods utilize advanced classification networks to learn discriminative high-level features from normal data. K-nearest neighbors (KNN) are then employed as anomaly detectors to measure the distance between test features and normal features to assess anomalies. This approach has achieved satisfactory performance in ASD tasks. However, it is not always stable [15]. This instability arises due to the difficulty of the classification task, as the features learned through auxiliary tasks may not be sufficiently fine-grained [16], leading to the anomaly detector's inability to effectively distinguish between normal and anomalous sounds.

To this end, we propose a self-supervised augmented diffusion model (SSDM) for ASD. First, we design a self-supervised learning module that leverages auxiliary classification tasks to learn expressive and discriminative latent features. Self-attention mechanisms are incorporated into the module to enhance the learning of features along the time and frequency axes. Then, we design a denoising diffusion module based on the diffusion model [17]. The denoising diffusion module learns the data distribution directly from these discriminative latent features rather than from the original data, enabling more effective anomaly detection. In the denoising diffusion module, during the training phase, the latent features are progressively corrupted through a forward diffusion process, and a reverse denoising process is employed to reconstruct and capture the distribution of normal data. During the testing phase, anomalous sounds are identified based on the reconstruction error of the input. We conducted experiments on the DCASE 2023 Challenge Task 2 development dataset, and the results demonstrate the effectiveness of our proposed method.

*Corresponding author.
†Work was done when Jiawei Yin interned at Midea.

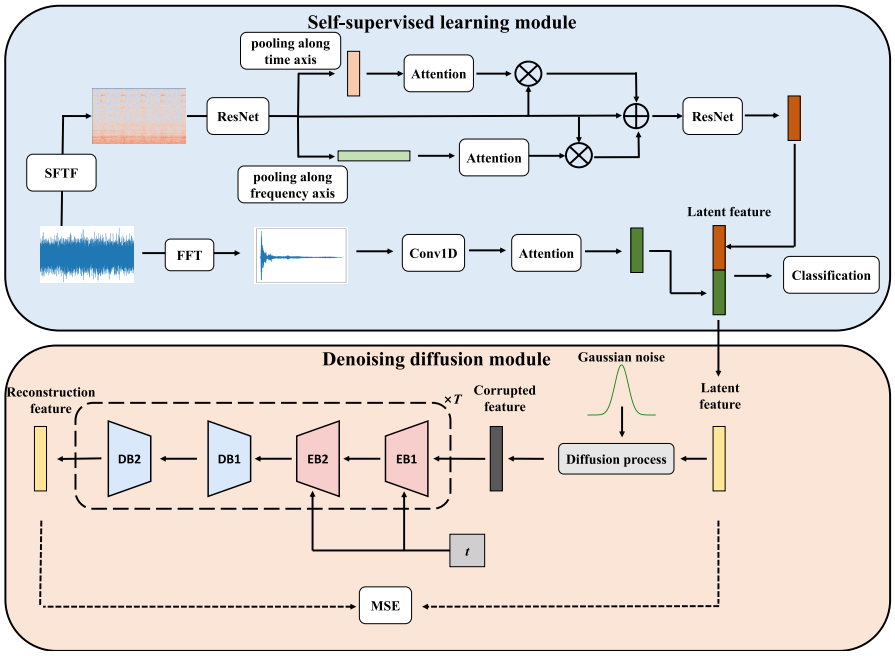


Fig. 1: The overview of our proposed SSDM method

The main contributions of this paper can be summarized as follows:

1) To effectively utilize the meta-information of sound data, we design a self-supervised learning module with a dual-path time-frequency framework to extract discriminative feature embeddings. We incorporate time-frequency self-attention mechanisms to enhance the focus on features along the time and frequency axes.
2) To learn the fine-grained distribution of the embeddings, we design a denoising diffusion module that reconstructs the data through a diffusion and denoising process, serving as the basis for anomaly sound detection. The reconstruction error of the embeddings from the test data is used as the anomaly score to identify anomalous sounds.
3) Experiments on the DCASE 2023 development set show that our proposed SSDM method outperforms baseline and state-of-the-art methods.

II. METHOD

Figure 1 provides an overview of our proposed SSDM method. First, in the self-supervised learning module, we use a dual-path ASD framework to accomplish an auxiliary classification task, introducing a self-attention module to enhance focus on the temporal and frequency aspects of the input. The self-supervised learning module outputs latent feature embeddings. These embeddings are then used in a denoising diffusion module to learn the distribution. The latent features are first progressively corrupted by a forward diffusion process and then reconstructed through a reverse denoising process to capture the distribution of normal data. Finally, anomalies are identified by using the reconstruction error as the anomaly score to detect anomalous sounds.

A. *Self-supervised learning module*

The primary purpose of generative models is to learn the distribution of input data, but they cannot fully utilize the metadata within the data to learn high-level features with discriminative semantics. To address this, we introduce a self-supervised learning module designed to extract data representations suitable for the ASD task. We employ a dual-path ASD framework [6] as the backbone of our self-supervised learning module to separately extract spectrogram and full magnitude frequency spectrum features. The spectrum path is processed by 1D convolutional layers, while the spectrogram path is handled by ResNet [18] layers. The dual-path framework has been proven to effectively extract expressive features from sounds compared to using a single path alone. We integrate self-attention modules along specific dimensions into both branches. An improved version of the Squeeze-and-Excitation Networks (SEnet) block [19] is used in this work, enhancing information along specific dimensions, as defined below:

$$y = x + \sigma\left(avgpool_i(x) \cdot W^{\mathrm{T}} + b\right) \cdot x \tag{1}$$

where $avgpool_i$ is the average pooling operation along a specific i dimension and σ is the sigmoid function operation. W and b are learning parameters, and x represents the input features. In the spectrogram path, we employ self-attention modules along the time and frequency dimensions. In the spectrum path, we use self-attention modules along the frequency dimensions. The encoded representations from both paths are combined to form the latent feature z, which is

used for training an auxiliary classification task through the classification head.

B. Denoising diffusion module

To effectively capture the distribution of the latent features z of normal data output from the self-supervised learning module, we designed a denoising diffusion module based on the denoising diffusion probabilistic model (DDPM) [17]. DDPM has demonstrated its excellent performance in terms of sample data generation quality and training stability [23]–[25]. The denoising diffusion module consists of a forward diffusion process $q(\,z_t|\,z_{t-1})$ and the reverse denoising process $p_\theta\,(z_{t-1}\,|z_t\,)$. In the diffusion process with T steps, random Gaussian noise progressively corrupts the original data distribution, which follows a Markov process:

$$q(\,z_t|\,z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}\right) \tag{2}$$

where $t = 1, \ldots, T$. The amount of noise added at each step is defined by the variance schedule β_t. This allows the distribution of z_t at any given time to be computed solely based on the original input z_0 and the noise addition step t. It can be defined as:

$$z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}e \tag{3}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$. In the denoising process, the model learns to optimize the parameters θ. The reverse process is defined as follows:

$$p_\theta\,(z_{t-1}\,|z_t\,) = \mathcal{N}\left(z_{t-1}; \mu_\theta\,(z_t, t)\,, \beta_t\mathbf{I}\right) \tag{4}$$

where $\mu_\theta\,(z_t, t)$ is mean of noise predicted by the network.

The training of DDPM can be regarded as an autoencoder. In our work, we use two encoder blocks (EBs) and two decoder blocks (DBs), each composed of dense layers, to train the DDPM. The input and output sizes are consistent and the output is predicted noise ε_t. The goal is to make the predicted noise consistent with the real noise e, so MSE loss can be used.

$$Loss = \arg\min\left[\|\varepsilon_t - e\|^2\right] \tag{5}$$

Based on the predicted noise ε_t , the denoising process can be defined as follows:

$$z'_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_t\right) + \sqrt{\beta_t}e \tag{6}$$

In ASD, the anomaly score is derived from the reconstruction error, with data points showing high deviations being more likely to be classified as anomalies. The anomaly score, denoted as $error$, is defined as follows:

$$error = \|z_0 - z'_0\|_2^2 \tag{7}$$

III. EXPERIMENTS AND RESULTS

A. Dataset

We use the development set from Task 2 of the DCASE 2023 challenge to validate the effectiveness of our proposed method. This dataset includes seven types of machines: Bearing, Fan, Gearbox, Slider, ToyCar, ToyTrain, and Valve. The dataset is divided into training and test sets. The training set contains 1000 normal sound samples for each machine type, while the test set includes 100 normal and 100 abnormal sound samples. Additionally, the training set provides attributes for each sample, which can be used during training. We utilize these attributes as classification labels in the self-supervised learning module.

We evaluate performance using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) and the partial AUC (pAUC). In our paper, we report the harmonic mean of all AUCs and pAUCs for each machine type, as well as the harmonic mean for the entire dataset.

B. Implementation details

In the self-supervised learning module, we use the mixup strategy to augment the original waveforms. The sub-cluster AdaCos [6] is employed as the classification loss function, and we use all possible combinations of machine types and machine attributes as classification labels. The training consists of 50 epochs with a batch size of 128. The optimizer is Adam with an initial learning rate of 0.01, which decreases to 90% of its value every 5 epochs. In the denoising diffusion module, the network hidden layer size is set to 512. During training, the number of diffusion steps is set to 500, and during testing, it is set to 200. The training consists of 500 epochs with a batch size of 128. The optimizer is Adam with a fixed learning rate of 0.001.

C. Performance comparison and ablation studies

We compare our proposed method with the baseline system [20], [21] and other SOTA methods [12], [22], as shown in Table 1. Our method achieves the highest overall harmonic mean, with improvements of 10.41% and 6.09% compared to the baseline model, and a 1.21% increase compared to the SOTA self-supervised method FeatEx. The self-implement baseline in the table is a combination of a dual-path ASD framework and KNN as the anomaly detector. The results demonstrate that the proposed SSDM enhances overall ASD performance.

To validate the roles of the self-supervised learning module and the denoising diffusion module in SSDM, we conducted ablation experiments. Table 2 shows the harmonic mean of the performance across all categories for different modules in SSDM. The results indicate that SSDM is highly effective when both modules are included. The self-supervised learning module significantly improves overall ASD performance, while the denoising diffusion module complements it to achieve even better ASD performance.

TABLE I: Comparison of different methods with harmonic mean of AUCs and pAUCs.

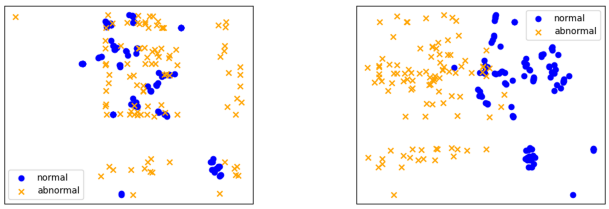
Method	ToyCar	ToyTrain	Bearing	Fan	Gearbox	Slider	Valve	All hmean
AE (Mahala). [20]	52.83	48.23	60.02	59.90	64.60	71.75	52.83	57.68
MobileNetV2. [21]	55.18	59.03	68.00	53.61	63.63	82.24	60.06	62.00
FeatEx. [12]	52.35	55.45	64.30	63.80	75.39	85.56	88.13	66.88
Han et al. [22]	63.47	57.35	57.10	62.76	67.52	79.11	67.79	64.31
Self-implement Baseline	53.04	53.13	66.00	63.42	72.36	84.23	63.05	63.55
SSDM (Ours)	56.26	57.75	65.84	67.77	72.09	89.55	78.55	68.09

TABLE II: Ablation studies on the modules of SSDM with harmonic mean of AUCs and pAUCs.

Module	SS+KNN	DM	SSDM
Self-supervised learning module	✓	×	✓
Denoising diffusion module	×	✓	✓
All hmean	65.87	56.11	68.09

D. Visualization analysis

To demonstrate the effectiveness of incorporating time-frequency self-attention in our self-supervised learning module, we use t-SNE to visualize the embeddings of the slider. The slider has been shown in previous studies to have highly distinctive time-frequency characteristics [11], which can help distinguish normal sounds from abnormal ones. As shown in Figure 2, after adding time-frequency self-attention, the overlap between abnormal and normal embeddings is significantly reduced, which will aid in subsequent anomaly detection module.



(a) without time-frequency self-attention (b) with time-frequency self-attention

Fig. 2: Illustration of t-SNE of time-frequency self-attention on slider.

To better illustrate the performance of each module in our proposed SSDM, we visualized the data using kernel density distribution plots. After training with normal data, we input the test data into the trained model and detect anomalies by evaluating the reconstruction error. As shown in Figure 3, different kernel density plots correspond to the probability density distributions of the test data within various modules of SSDM. Figure 3(a) shows the original data distribution, where the distribution of anomalous data largely overlaps with that of normal data. After processing by the self-supervised learning module, as shown in Figure 3(b), the anomalous data becomes roughly separated from the normal data, with the difference in distribution becoming more pronounced. Subsequently, the embedded features are reconstructed using the denoising diffusion module, where the reconstructed anomalous data follows the distribution pattern of normal data, resulting in significantly higher reconstruction errors for anomalies. As shown in Figure 3(c), accurate anomaly detection can be achieved based on the reconstruction error using SSDM.

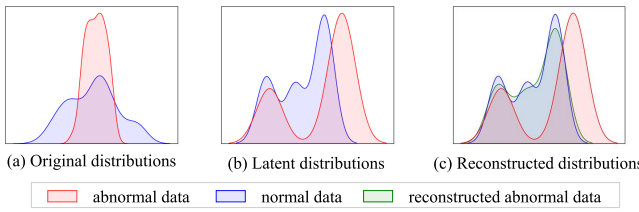


Fig. 3: Kernel density distribution plots of test data in different modules. (a). Original distribution of the test data. (b). Distribution of the latent features after processing by the self-supervised learning module. (c). Distribution of the latent features after reconstruction by the denoising diffusion model.

IV. Conclusions

In this paper, we introduce a self-supervised augmented diffusion model (SSDM) for anomaly sound detection. SSDM utilizes a self-supervised learning module with dual-path time-frequency self-attention framework to leverage metadata from sounds and learn expressive and discriminative features. The denoising diffusion module learns the distribution of these discriminative latent features and uses the reconstruction error to detect anomalous sounds. Experimental results on the development dataset of DCASE 2023 Task 2 demonstrate the effectiveness of our proposed method.

References

[1] T. Nishida, N. Harada, D. Niizumi, *et al.*, "Description and discussion on dcase 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2406.07250*, 2024.

[2] H. Yun, H. Kim, Y. H. Jeong, and M. B. Jun, "Autoencoder-based anomaly detection of industrial robot arm using stethoscope based internal sound sensor," *Journal of Intelligent Manufacturing*, vol. 34, no. 3, pp. 1427–1444, 2023.

[3] A. Jiang, W.-Q. Zhang, Y. Deng, P. Fan, and J. Liu, "Unsupervised anomaly detection and localization of machine audio: A gan-based approach," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[4] S. Liu, J. Li, W. Ke, and H. Yin, "Multi-attention enhanced discriminator for gan-based anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 6715–6719.

[5] S. Li, J. Yu, Y. Lu, G. Yang, X. Du, and S. Liu, "Self-supervised enhanced denoising diffusion for anomaly detection," *Information Sciences*, vol. 669, p. 120 612, 2024.

[6] K. Wilkinghoff, "Sub-cluster adacos: Learning representations for anomalous sound detection," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8.

[7] H. Hojjati and N. Armanfard, "Self-supervised acoustic anomaly detection via contrastive learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 3253–3257.

[8] M.-H. Nguyen, D.-Q. Nguyen, D.-Q. Nguyen, C.-N. Pham, D. Bui, and H.-D. Han, "Deep convolutional variational autoencoder for anomalous sound detection," in *2020 IEEE Eighth International Conference on Communications and Electronics (ICCE)*, IEEE, 2021, pp. 313–318.

[9] H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Hierarchical conditional variational autoencoder based acoustic anomaly detection," in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 274–278.

[10] V. Zavrtanik, M. Marolt, M. Kristan, and D. Skočaj, "Anomalous sound detection by feature-level anomaly simulation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 1466–1470.

[11] X. Huang, F. Guo, and L. Chen, "A res-ganomaly method for machine sound anomaly detection," *IEEE Access*, 2024.

[12] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 276–280.

[13] S. Choi and J.-W. Choi, "Noisy-arcmix: Additive noisy angular margin loss combined with mixup for anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 516–520.

[14] H. Chen, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu, "Self-supervised representation learning for unsupervised anomalous sound detection under domain shift," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 471–475.

[15] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 816–820.

[16] J. Guan, F. Xiao, Y. Liu, Q. Zhu, and W. Wang, "Anomalous sound detection using audio representation with machine id based contrastive learning pretraining," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[20] K. Dohi, K. Imoto, N. Harada, *et al.*, "Description and discussion on dcase 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2305.07828*, 2023.

[21] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *2023 31st European Signal Processing Conference (EUSIPCO)*, IEEE, 2023, pp. 191–195.

[22] B. Han, Z. Lv, A. Jiang, *et al.*, "Exploring large scale pre-trained models for robust machine anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 1326–1330.

[23] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2022, pp. 35–45.

[24] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 650–656.

[25] X. Zhang, N. Li, J. Li, T. Dai, Y. Jiang, and S.-T. Xia, "Unsupervised surface anomaly detection with diffusion probabilistic model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6782–6791.

Music2Fail: Transfer Music to Failed Recorder Style

Chon In Leong, I-Ling Chung, Kin-Fong Chao,
Jun-You Wang, Yi-Hsuan Yang, Jyh-Shing Roger Jang
National Taiwan University, Taiwan
E-mail: r11922174@csie.ntu.edu.tw, r11942103@ntu.edu.tw, r12922179@ntu.edu.tw,
junyou.wang@mirlab.org, yhyangtw@ntu.edu.tw, jang@csie.ntu.edu.tw

***Abstract*—The goal of music style transfer is to convert a music performance by one instrument into another while keeping the musical contents unchanged. In this paper, we investigate another style transfer scenario called "failed-music style transfer". Unlike the usual music style transfer where the content remains the same and only the instrumental characteristics are changed, this scenario seeks to transfer the music from the source instrument to the target instrument which is deliberately performed off-pitch. Our work attempts to transfer normally played music into off-pitch recorder music, which we call "failed-style recorder", and study the results of the conversion. To carry out this work, we have also proposed a dataset of failed-style recorders for this task, called "FR109 Dataset". Such an experiment explores the music style transfer task in a more expressive setting, as the generated audio should sound like an "off-pitch recorder" while maintaining a certain degree of naturalness.** [1]

I. Introduction

Generally, the goal of *music style transfer* is to change the style of the input audio while preserving the content of the input audio[2]. In particular, the *content* of the input music data refers to music features such as rhythm and melody, while the *style* refers to the unique perceived characteristics that an instrument expresses in music performance. According to the wide range of conversion goals and targets, the tasks of music style transfer can be broadly classified into three categories: *one-to-one* [1], [2], *many-to-one* [3], [4], *many-to-many* [5], [6]. Several music style transfer methods have been inspired by research from different fields, such as voice conversion (VC) [5]–[8] and image-based style transfer [9], [10].

Most works mainly use well-played audio as training data for both the source and target domains. This implies that well-played input audio will be converted into equally well-played output audio. Although this ensures faithful style transfer, assuming all audio to be well-played in the real world restricts the expressiveness of instruments. The Singing Voice Beautifying (SVB) task [11] is a special case. They used paired data of amateur and professional singing voices for training, aiming to correct the pitch and improve the vocal tone, indicating that the source audio was not well-played. In this paper, we focus on another special case, where the target domain is not well-played. For example, the target domain may be a soprano recorder that is deliberately performed poorly[3]. We refer to this case as *failed-music style transfer*. The motivation is that such failed music contains a wider range of characteristics, but humans can still distinguish between a "failed recorder" and another instrument. This poses a more difficult scenario to music style transfer: Can a music style transfer model not only tackle well-played instruments but also instruments that are *not* well-played?

Take a soprano recorder as an example, a fail-style recorder may contain many types of errors, such as:

- **Cracked voice.** Producing a harsh sound.
- **Weird dynamics.** Unnatural volume while playing.
- **Failed tonguing.** Mistakes in the articulation.
- **Overblowing.** Blowing too hard, causing the voice to sound raspy.
- **Underblowing.** Blowing not hard enough, causing the voice to sound hissing.

Generally, these are considered errors that should not occur in live performances. However, by definition, such errors should not make a style transfer model malfunction. Instead, a style transfer model should generate audio that sounds like a failed recorder (sometimes has an unpleasant style, but still sounds like a recorder). Such failed music style transfer might be useful in the fields of entertainment, it could serve as the score for some comedies or some humorous scenes.

In this paper, we investigate the music style transfer scenario of failed recorders, treating it as a type of instrument. We apply various general style transfer methods and analyze the conversion results. However, there are no existing datasets for such scenarios which were deliberately recorded as failed-style. To facilitate our research, we propose the "FR109" dataset, a collection of failed-style recorder music recorded by a professional, with deliberately included failures.

To sum up, the main contributions of this paper are two-fold:

- We discuss the special scenario of *failed-music style transfer* that serves as a more challenging task for style transfer. Regarding the experimental results, we analyzed them from the perspectives of the Mel spectrogram and Wiener entropy, providing corresponding analyses and interpretations.
- To carry out the work for *failed-music style transfer*, we propose the FR109 dataset, a dataset of failed recorder

[1]Our demo website is now available on: https://navi0105.github.io/demo/music2fail/

[2]In practice, due to the differences in pitch ranges of different instruments, we may perform additional pitch shifting in the experiments.

[3]A famous example can be found at https://www.youtube.com/watch?v=X2WH8mHJnhM

performance, created intentionally by an experienced individual playing a recorder.

II. Datasets

In this work, we adopted three datasets in the experiments, including two publicly available datasets (URMP [12] and Bach10 [13]), and our FR109 dataset. The comparison of different datasets is shown in Table I.

A. The URMP dataset

The URMP dataset [12] contains 44 music pieces ranging from duets to quintets, with separated tracks for individual instrument recordings. There are 14 distinct instruments in this dataset. In our experiments, we only used the violin, clarinet, and saxophone tracks as the training data.

B. The Bach10 dataset

The Bach10 dataset [13] consists of audio recordings of 10 J.S. Bach chorales performed separately with violin, clarinet, saxophone, and bassoon. In our experiments, we only used the violin, clarinet, and saxophone tracks as the testing data. Since the training data (URMP) and testing data (Bach10) belong to different datasets, such an evaluation scenario is more challenging.

C. The proposed FR109 dataset

As for the failed-music style transfer, we proposed the FR109 dataset, which consists of 109 songs recorded with a soprano recorder played by a professional, with a total duration of 5.05 hours. Errors are introduced to each performance intentionally.

As discussed in Section I, the types of errors include cracked voice, weird dynamics, failed tonguing, overblowing, and underblowing. To compute the statistics of the dataset, we extract the pitch of recorder music using CREPE [14]. The pitch mean of the FR109 dataset is around 905Hz (between A5 and A#5), and the maximum pitch value is 1990Hz (around B6). These statistics match the actual pitch range of the soprano recorder, which spans from C5 to D7.

Since there is no other dataset for failed recorders, we use the FR109 dataset as both the training dataset and the testing dataset in the failed-music style transfer experiments. A 90%/10% split is employed to divide the dataset into a training dataset and a testing dataset. In our experiments, we trained style transfer models to perform style transfer between all 4 instruments (violin, clarinet, saxophone, and failed recorder).

We are making the FR109 dataset publicly available for reproducibility, you can check out more information on our official GitHub page: https://github.com/navi0105/Music2Fail

TABLE I
THE LIST OF DATASETS UTILIZED IN THIS WORK, INCLUDING THE PROPOSED FR109 DATASET. "PIECES" STANDS FOR THE NUMBER OF MUSIC PIECES.

Dataset	Instrument	Pieces	Total duration
URMP	Violin	34	1.02 hours
	Clarinet	10	0.30 hours
	Saxophone	11	0.26 hours
Bach10	Violin	10	0.09 hours
	Clarinet	10	0.09 hours
	Saxophone	10	0.09 hours
FR109	(Failed) recorder	109	5.05 hours

III. Method

In this work, we experiment with three different well-known style transfer methods for failed-music style transfer, they are StarGAN [15], VAE-GAN [7], and DDSP [3].

StarGAN [15] introduced domain labels to the generator and discriminator, the generator uses the domain label to specify the target domain, while the discriminator needs to predict the input's domain. During training, the generator and the discriminator contest with each other, the generator's objective is to fool the discriminator, making it mispredict the domain label, while the discriminator's objective is to avoid being fooled by the generator. StarGAN only used a single generator and discriminator for learning a multi-mapping between different styles, instead of one generator and one discriminator for each pair of styles.

VAE-GAN [7] used one generator and one discriminator for each domain, the generator used a variational autoencoder composed of two parts: the universal encoder and a decoder, the universal encoder shared across each generator to encode the input to latent code, and a decoder to transfer the latent code to the target domain. Since it uses the same encoder for every input domain and target domain, the performance was increased due to the variation of the input data. The decoder is domain-specific so it can be specialized to that domain. The discriminator only needed to predict whether the data was generated for that domain.

DDSP [3] integrates classic signal processing with deep learning. This method employs an autoencoder architecture for style transfer within a single domain. The encoder extracts key features from the source audio, including loudness, fundamental frequency, and residual information, while the decoder maps these features to control parameters for synthesizers to generate the output audio. DDSP has an assumption that the pitch component extracted from the source audio should closely match the fundamental frequency of the output audio, which may not be suitable for failed-music style transfer.

StarGAN and VAE-GAN are two-stage style transfer pipelines, where we first convert the source audio to Mel spectrogram. Then, the Mel spectrogram was transferred to the failed recorder style using the generator. Finally, the vocoder generates waveform from the transferred Mel spectrogram. Here, we use BigVSAN [16] as our vocoder, its pretrained weight are available in their official repository[4], which was pretrained on the LibriTTS dataset's training dataset [17] for 10 million steps.

[4]https://github.com/sony/bigvsan

TABLE II
THE FAD METRICS OF STYLE TRANSFER MODELS TO FAILED RECORDER IN THE BACH10 DATASET.

Models	FAD (↓)
StarGAN	13.87
VAE-GAN	**7.27**
DDSP	38.91

IV. EXPERIMENTS

As discussed in Section II, we combined the URMP dataset [12] and the FR109 dataset's training dataset for training and the Bach10 dataset [13] for evaluation. Three different methods are compared in the experiments, they are StarGAN [15], VAE-GAN [7], and DDSP [3].

A. Training

1) Data preprocessing: We refer to the arguments for calculating the Mel spectrogram from BigVSAN [16] to compute the Mel spectrograms of music data in our dataset, 24,000 for sampling rate, 100-bands of Mel filter bank, 1024 for FFT / Hann window, hop size is 256 and the frequency range is from 0 to 12,000 Hz.

B. Evaluation

In this section, we compare the performance between StarGAN, VAE-GAN, and DDSP. Both objective and subjective experiments are conducted.

1) Objective evaluation: Fréchet Audio Distance (FAD) [18] is a reference-free metric to compute the Fréchet Inception Distance (FID) between audio embedding sets extracted from the reference set and evaluation set. In the experiments, the reference set is music from the testing dataset and the evaluation set is music generated by the model. FAD represents the degree of dissimilarity between the two sets. The audio embeddings are extracted by a pretrained VGGish audio classification model [19]. We use FAD as an objective evaluation metric to assess the distance between the audio files converted by StarGAN / VAE-GAN / DDSP and the real audio performance of a target instrument.

Table II shows the FAD score of Bach10 dataset's music converted to failed-style recorder music using these three different models. The results indicate that StarGAN performs slightly worse than VAE-GAN on both datasets. Considering that StarGAN only utilizes one unified decoder while VAE-GAN uses one decoder for each instrument, such a performance gap is acceptable.

As for the DDSP model, results show that DDSP has a significantly larger FAD compared to StarGAN and VAR-GAN. By inspecting the audio converted by DDSP, we found that they contain a large amount of noise, which is the reason for the high FAD. The results of DDSP demonstrate that its assumptions about pitch invariance can lead to better performance on well-played instrument transitions, but do not apply well to our task.

TABLE III
THE MOS OF THE LISTENING TEST ON THE BACH10 DATASET. THE NUMBERS INSIDE THE CELLS REPRESENT THE MOS AND THEIR STANDARD DEVIATIONS. SS, MS, AND SQ INDICATE THE STYLE SIMILARITY, MELODY SIMILARITY, AND SOUND QUALITY, RESPECTIVELY.

Models	SS (↑)	MS (↑)	SQ (↑)
StarGAN	2.54 ± 1.26	3.15 ± 1.19	2.46 ± 1.27
VAE-GAN	**2.98** ± 1.23	**3.56** ± 0.93	**3.00** ± 0.98
DDSP	1.33 ± 0.77	2.19 ± 1.11	1.38 ± 0.70

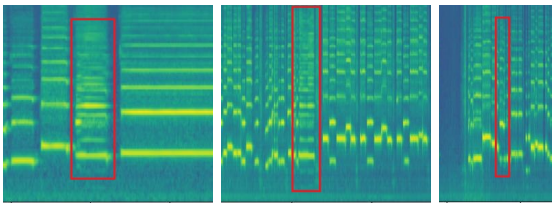


Fig. 1. Mel spectrograms of failed recorder music, the red rectangular parts show the inharmonic partials.

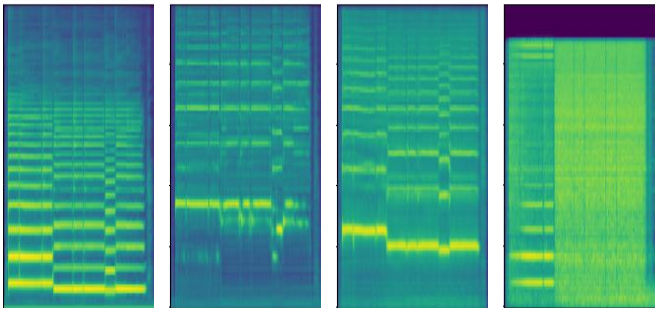
2) Subjective evaluation: We performed a listening test that evaluates the performance of converting these three (well-played) instruments into failed recorder in the Bach10 dataset (3 source-target pairs).

For each source-target pair, we randomly choose one audio clip for the listening test. We employed a rating scheme based on the Mean Opinion Score (MOS) [20]. For each audio clip (converted by one of the models), we asked the participants to evaluate its quality in three aspects: (1) *Style similarity* (SS) to the target instrument, (2) *Melody similarity* (MS) to the source audio, (3) *Sound quality* (SQ) of the converted audio. The scoring ranges from 1 to 5, where 1 is the worst and 5 is the best. In total, we received 16 valid responses from the listening test. Table III shows the MOS of the Bach10 dataset.

The results indicate that StarGAN's overall performance falls behind VAE-GAN's on all metrics in the conversion to failed recorder. The p-values between StarGAN and VAE-GAN are 0.09 (SS), 0.06 (MS), and 0.02 (SQ). Although only the sound quality (SQ) is considered statistically significant, overall, we can still conclude that StarGAN is slightly inferior to VAE-GAN in converting to failed recorder music. As for DDSP, similar to the objective results, the MOS results of DDSP are significantly worse than those of StarGAN. This is likely because DDSP generates noise more frequently. All the t-test yielded p-values well below 0.05. This reflects the specific challenges involved in music style transfer to failed instrument music for DSP-based synthesizers.

V. ANALYSIS

In this section, we analyze the results of failed recorder style transfer, and further compare the tasks between the conversion to well-played instruments (violin, clarinet, saxophone) and failed recorder.



(a) Source (b) StarGAN (c) VAE-GAN (d) DDSP

Fig. 2. The Mel spectrograms of a failed-music style transfer example. (a) The Mel spectrogram of the source audio, which is performed by a saxophone; (b) The Mel spectrogram of the converted audio (to failed recorder) by StarGAN; (c) The Mel spectrogram of the converted audio (to failed recorder) by VAE-GAN. (d) The Mel spectrogram of the converted audio (to failed recorder) by DDSP.

TABLE IV
WIENER ENTROPY OF THE URMP DATASET AND FR109 DATASET. VN., CL., SAX., AND REC. REPRESENT VIOLIN, CLARINET, SAXOPHONE, AND RECORDER, RESPECTIVELY.

Dataset	Instrument	Wiener entropy
URMP	Vn. / Cl. / Sax.	0.0005
FR109	Rec.	0.0345

A. Mel spectrogram analysis

To understand the challenge of failed-music style transfer, we first visualize the spectrograms of the failed recorder music in the FR109 dataset, as shown in Figure 1. The red rectangular parts of the Mel spectrograms implies that the sound has *inharmonic partials*, meaning that the frequencies of the overtones do not align with integer multiples of the fundamental frequency. This creates a more complex and less predictable timbre. These inharmonic partials are considered features of the failed recorder because they are present in the Mel spectrograms of every failed recorder sample. Such inharmonic partials are rarely found in well-played instrument performances.

Next, we visualised an example of failed-music style transfer, Figure 2(a) shows the Mel spectrogram of source audio performed by a saxophone, in which there is no clear inharmonic partial. Figure 2(b) and Figure 2(c) show the audio converted to a failed recorder by StarGAN and VAE-GAN, respectively. We can clearly see that inharmonic partials occur throughout the whole Mel spectrogram of StarGAN, showing that it does capture the characteristic of failed recorders and performs style transfer accordingly. For VAE-GAN, inharmonic partials can still be seen, but not as clearly as that of StarGAN. This shows that in this particular case, while both StarGAN and VAE-GAN do perform style transfer to some extent, StarGAN achieves a better style similarity to a failed recorder. Our informal listening test also confirms this observation. We attached these audios in the supplementary material.

TABLE V
WIENER ENTROPY OF THE STYLE TRANSFER RESULTS OF STARGAN, VAE-GAN AND DDSP ON DIFFERENT TARGET INSTRUMENTS. VN., CL., SAX., AND REC. REPRESENT VIOLIN, CLARINET, SAXOPHONE, AND RECORDER, RESPECTIVELY.

Model/Target	Vn.	Cl.	Sax.	Rec.
StarGAN	0.0007	0.0004	0.0002	0.0153
VAE-GAN	0.0003	0.0003	0.0006	0.0159
DDSP	0.0008	0.0008	0.0005	0.2041

Based on Figure 1 and Figure 2, it can be seen that failed-music style transfer does show a clearly different characteristic to the style transfer of other well-played instruments. To achieve style transfer to failed music, a model has to generate audio with unique properties that do not usually occur in well-played music. Discussing such a task would help understand the performance and the limitation of a style transfer model in another aspect.

B. Wiener entropy

Furthermore, we utilized the STFT-based *Wiener entropy* [21] to quantify how much the noise-like sound is in the results produced by StarGAN and VAE-GAN, along with the Wiener entropy of the URMP dataset and FR109 dataset, which serve as the benchmark for real performance of well-played music and failed music. Table IV shows the Wiener entropy of the URMP dataset and the FR109 dataset, i.e. well-played instrument music and failed recorder music, we can see that failed recorder music exhibits a higher proportion of noise-like characteristics compared to well-played instrument music. Table V shows the Wiener entropy of each of the models on different target instruments. We can see that when the target instruments are well-played instruments, the noise in the results from StarGAN and VAE-GAN are similar to the URMP dataset since their Wiener entropy is very similar. For failed recorder music, we can see that there is a gap between the Wiener entropy of FR109 and the Wiener entropy of StarGAN or VAE-GAN for converting to failed recorder, this shows that there is still room for improvement in converting music to the failed recorder style.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have conducted a series of experiments on the failed-music style transfer, and analysed the characteristics of this relatively special transfer in different aspects through various evaluations.

Furthermore, we have released the FR109 dataset, consisting of failed recorder performances, which is useful for investigating the expressiveness of different style transfer model. Through this study, we hope to propose a music style transfer task that is different from the usual music style transfer task that pursues sound quality and accuracy, but rather a music style transfer task that is more versatile.

References

[1] M. Pasini, "Melgan-vc: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms," *arXiv preprint arXiv:1910.03713*, Oct. 2019.

[2] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, and R. B. Grosse, "Timbretron: A wavenet(cyclegan(cqt(audio))) pipeline for musical timbre transfer," *CoRR*, vol. abs/1811.09620, Nov. 2018.

[3] J. H. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: differentiable digital signal processing," *CoRR*, vol. abs/2001.04643, Jan. 2020.

[4] S. Nercessian, "Differentiable WORLD synthesizer-based neural vocoder with application to end-to-end audio style transfer," *CoRR*, vol. abs/2208.07282, 2022. arXiv: 2208.07282.

[5] A. Bitton, P. Esling, and A. Chemla-Romeu-Santos, "Modulated variational auto-encoders for many-to-many musical timbre transfer," *CoRR*, vol. abs/1810.00222, Oct. 2018.

[6] Y. Wu, Y. He, X. Liu, Y. Wang, and R. B. Dannenberg, "Transplayer: Timbre style transfer with flexible timbre control," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, 2023, pp. 1–5.

[7] R. S. Bonnici, M. Benning, and C. Saitis, "Timbre transfer with variational auto encoding and cycle-consistent adversarial networks," in *2022 International Joint Conference on Neural Networks*, Padua, Italy, 2022, pp. 1–8.

[8] L. Comanducci, F. Antonacci, and A. Sarti, "Timbre transfer using image-to-image denoising diffusion models," *arXiv preprint arXiv:2307.04586*, Jul. 2023.

[9] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference*, 2018, pp. 2100–2104.

[10] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[11] J. Liu, C. Li, Y. Ren, Z. Zhu, and Z. Zhao, "Learning the beauty in songs: Neural singing voice beautifier," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, 2022, pp. 7970–7983.

[12] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, Feb. 2018.

[13] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Speech Audio Process.*, vol. 18, no. 8, pp. 2121–2133, Nov. 2010.

[14] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 161–165.

[15] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, Utah, USA, 2018, pp. 8789–8797.

[16] T. Shibuya, Y. Takida, and Y. Mitsufuji, "Bigvsan: Enhancing gan-based neural vocoders with slicing adversarial network," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, 2024, pp. 10 121–10 125.

[17] H. Zen, V. Dang, R. Clark, *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," in *The 20th Annual Conference of the International Speech Communication Association*, Graz, Austria: ISCA, 2019, pp. 1526–1530.

[18] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, Dec. 2018.

[19] S. Hershey, S. Chaudhuri, D. P. W. Ellis, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, Louisiana, USA, 2017, pp. 131–135.

[20] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: Methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[21] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.

Murmur Separation and Classification from Heart Sound Using Constrained Singular Spectrum Analysis and Wavelet Transform

Yuanyang Qi* and Saeid Sanei†
* Imperial College London, London, UK
E-mail: yuanyang.qi23@imperial.ac.uk Tel/Fax: +44-7422951891
† Imperial College London, London, UK
E-mail: s.sanei@imperial.ac.uk Tel/Fax: +44-7919098002

***Abstract*—Detection and classification of heart murmurs are crucial for the early and timely diagnosis and treatment of cardiovascular diseases. However, heart murmurs often overlap with normal heart sounds, making their detection difficult. In this paper, we propose a two-stage method for separating murmurs before their classification. The first stage employs Constrained Singular Spectrum Analysis to exploit the distinct statistical characteristics of heart sounds and murmurs, facilitating their initial separation. The second stage leverages the capability of Wavelet Transform to effectively localize sound components in both time and frequency domains, which is essential for separating subtle murmurs that are not completely removed during the first stage. After these two stages, normal heart sounds are precisely extracted, and the remaining audio components, predominantly murmurs, are effectively classified. Our results demonstrate that pre-separating murmurs enhances the classification performance, achieving an accuracy of over 99% after using transfer learning. The proposed method can improve the clarity and accuracy of heart sound analysis, thereby facilitating better diagnostic evaluations and interventions in clinical settings.**

I. Introduction

Cardiovascular disease is one of the leading threats to human health. Murmurs are very important indicators of such diseases. Therefore, detecting heart murmurs is crucial for the early and timely diagnosis and treatment of cardiovascular conditions [1]. A heart murmur is an abnormal sound heard during the cardiac cycle, typically audible with the aid of a stethoscope, and is often described as a whooshing or swishing noise [2]. This sound results from turbulent blood flow through the heart. Different types of murmurs are associated with various underlying heart diseases [3]. For instance, systolic murmurs, such as those associated with aortic stenosis or mitral regurgitation, occur during ventricular contraction and often indicate abnormalities in valve structure or function. Diastolic murmurs, like those linked to aortic regurgitation or mitral stenosis, manifest during ventricular relaxation and suggest issues such as valve incompetence or narrowing [4]. Fig. 1 illustrates the normal heart sound and different types of murmurs. Often, murmurs are obscured by normal heart sounds and other noises [5]. Extensive experience and training for physicians are needed to identify heart murmurs.

In recent years, various computer-aided technologies have been applied to classify murmurs. In [6] the authors presented a method using wavelet transform and Hilbert phase envelope to classify abnormal heart signal sounds into ten classes depending on the occurrence moment. Many machine learning approaches have also been proposed to detect and classify normal heart sounds and heart murmurs, such as the K-Nearest Neighbour (KNN) algorithm, Support Vector Machines (SVM) algorithm [7], Convolution Neural Network (CNN) [8], Residual Networks (ResNet) [9], and Attention Mechanism [10]. However, they don't attempt the murmur separation method before classification. In [11], [12] the authors have used deep learning methods for source separation. Although the separation results are competitive, their methods require a large data size for training [13].

In this paper, we introduce and employ signal processing and machine learning methods to separate and classify different types of murmurs. Normal heart sounds generally have slow variation and hence a lower Zero Crossing Rate (ZCR) [14] and more peaky distribution, denoted by higher kurtosis [15] compared with murmurs. We use Constrained Singular Spectrum Analysis (CSSA) with ZCR and kurtosis to initially separate the normal heart sound from the murmur. Subsequently, we apply Wavelet Transform (WT) for further extraction of the murmur to achieve high separation precision. Finally, we use machine learning techniques to classify the separated murmurs into their respective types.

II. Methodology

A. Separation of Heart Sound and Murmur

Fig. 2 demonstrates the overall separation process. The method is a two-stage process for the separation of murmurs from original sounds. In the first stage, CSSA is used to initially segregate the normal heart sound from the original sound. This stage can filter out the regular heartbeats, although some murmurs may remain mixed within the extracted sound. The second stage focuses on refining the normal heart sound obtained from the first stage using WT to achieve a high-precision separation of the normal heart sounds.

Singular Spectrum Analysis (SSA) is a non-parametric method particularly advantageous for analyzing non-linear and

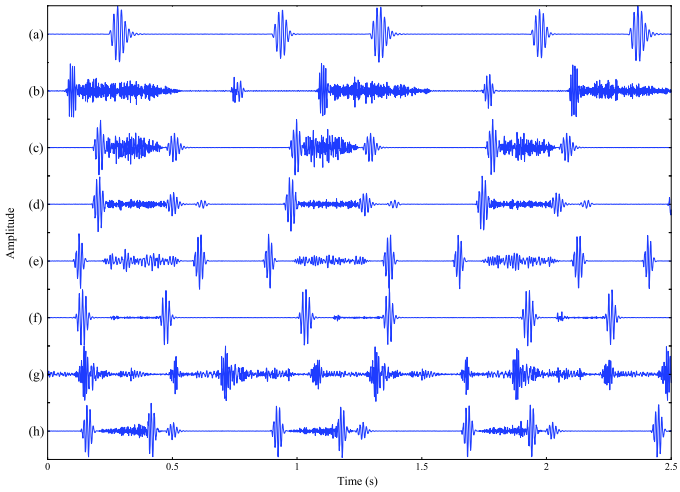


Fig. 1. Normal Heart sound (a) and heart sounds with (b) aortic regurgitation murmur, (c) aortic stenosis murmur, (d) mitral regurgitation murmur, (e) mitral stenosis murmur, (f) mitral valve prolapse murmur, (g) pericardial friction murmur, and (h) pulmonary stenosis murmur.

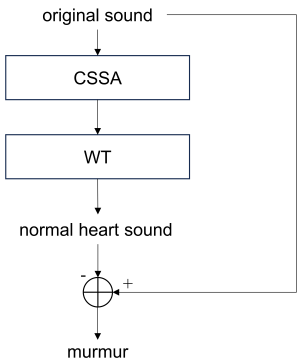


Fig. 2. The overall flowchart for separating heart sound and murmur.

non-stationary time series data. It has many applications such as noise reduction, signal reconstruction, forecasting future trends and biomedical signals [16]. Using SSA, the signal is decomposed into multiple components first, and then the desired signals are reconstructed according to some constraints, referred to as CSSA. The normal heart sounds tend to have lower ZCR and higher kurtosis compared with the murmurs. Therefore, setting ZCR or kurtosis as constraints in the CSSA process is an approach to differentiating these sounds. The above process can be summarized as follows.

After selecting a fixed window length L, the original signal x is transformed from a one-dimensional signal of length N into an $L \times K$ trajectory matrix X, where $K = N - L + 1$.

$$X = \begin{bmatrix} x_0 & x_1 & x_2 & \cdots & x_{K-1} \\ x_1 & x_2 & x_4 & \cdots & x_K \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{L-1} & x_L & x_{L+1} & \cdots & x_{N-1} \end{bmatrix} \tag{1}$$

In the next step, the covariance of the Hankel matrix X, $S = XX^T$, is decomposed into $U\Lambda U^T$ where Λ is a diagonal matrix consisting of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_L$ $(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L)$ and $U = [U_1, U_2, \cdots, U_L]$ is the orthonormal eigenvectors of S. Here, $X = \sum_{i=1}^{d} X_i$ where $d = \text{Rank}(X)$. After applying diagonal averaging, the original signal x is converted into a $N \times d$ matrix R, with each component R_i representing a one-dimensional signal of length N.

$$R = \begin{bmatrix} R_0 & R1 & R2 & \cdots & R_{d-1} \end{bmatrix} \tag{2}$$

Next, we apply the constraint along with W, a column vector filled with binary values, to selectively extract the desired component from R.

$$W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{d-1} \end{bmatrix}, \; w_i = \{0, 1\} \tag{3}$$

This selection process allows us to reconstruct the final signal x_r, which is obtained by multiplying R by W, denoted as $R \times W$.

$$x_r = R \times W \tag{4}$$

CSSA with ZCR: ZCR measures how frequently the signal changes between positive and negative values. For each component R_i of the signal, we calculate the ZCR. If the computed rate is below or equal to a predefined threshold, then we assign w_i a value of 1; otherwise, w_i is set to 0. CSSA with ZCR enables selective signal reconstruction based on the dynamic characteristics of each component.

$$w_i = \begin{cases} 1 & \text{ZCR}(R_i) \leq \text{threshold}, \\ 0 & \text{ZCR}(R_i) > \text{threshold}. \end{cases} \tag{5}$$

CSSA with Kurtosis: Kurtosis is a statistical measure that describes the distribution's peakedness relative to a normal distribution. In this method, we optimize a vector of parameters W to maximise the kurtosis of $R \times W$. This task constitutes a nonlinear integer programming problem, which can effectively be solved using Genetic Algorithm (GA). Such an algorithm is effective for solving complex optimization challenges involving the nonlinear cost function and constraint. Fig. 3 illustrates the process by which the CSSA with kurtosis separates the heart sounds from murmurs. CSSA with kurtosis enables selective signal reconstruction based on the statistical characteristics of each component represented by its distribution.

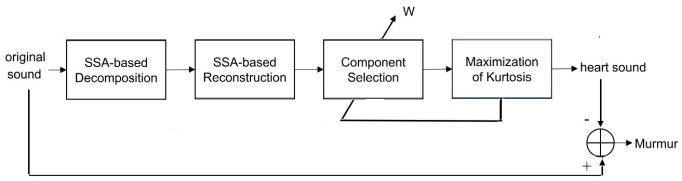


Fig. 3. The process for CSSA with kurtosis.

$$W = \max_{W} \text{ kurtosis}(R \times W) \tag{6}$$

In the first stage of separating heart sound and murmur, we employ CSSA both with ZCR and kurtosis. Each method is utilized to separate the normal heart sound, leaving the murmur as the residual component. Following this, we calculate the correlation between the separated normal heart sound and the murmur. The method—either CSSA with ZCR or CSSA with kurtosis—that yields the lower correlation between these components is then selected as the input to the second stage of the analysis. Fig. 4 shows the flowchart of CSSA in the first stage. This approach ensures that the method leading to the least similarity between the normal sound and the murmur is advanced, enhancing the effectiveness of the separation process.

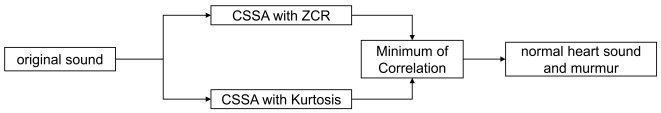


Fig. 4. The flowchart of CSSA in the first stage of separation.

In the second stage of separating heart sound and murmur, we employ WT to meticulously filter out the residual murmur from the heart sound obtained in the first stage. WT is known for its time-frequency analysis capabilities, making it particularly effective for analyzing non-stationary signals, where the signal characteristics change over time [17]. This makes WT an important tool in various applications, such as signal compression, noise reduction, and feature extraction. WT works by decomposing a signal into a set of wavelets, which are functions that can capture both frequency and temporal location information simultaneously. This decomposition allows for a multi-resolution analysis of the signal, enabling the identification of features that may not be apparent in the original time-domain representation. Given that our signal is discrete, we employ the Discrete Wavelet Transform (DWT) to distinguish the residual murmurs from the normal heart sounds identified initially. During the application of DWT, the heart sound signal is decomposed into various frequency components at different scales. The filtering process involves thresholding and reconstructing the signal to ensure that the residual murmur is effectively separated from the normal heart sounds.

After completion of these two stages, the process results in precise extraction of normal heart sounds. The use of CSSA and DWT ensures that both stages complement each other, leading to a more refined separation process. The separated murmurs, are then separated and can be used for classification. This separation is crucial as it allows for a focused analysis of murmur characteristics without the interference of normal heart sounds. This two-stage method enhances the clarity and accuracy of heart sound analysis, facilitating better diagnostic evaluations and interventions in clinical settings.

B. Classification of Murmur

Once the murmur has been separated, we proceed to extract its features and employ machine learning for classification. For this purpose, we utilize four distinct classification methods: SVM [18], Random Forest (RF) [19], CNN [20] and transfer learning [21].

SVM is a supervised machine learning algorithm and its goal is to find an optimal hyperplane that can best separate different classes by maximizing the margin between the closest data points of each class, known as support vectors. RF is a bagging method within ensemble learning and can enhance the accuracy and stability of predictions by constructing multiple decision trees and aggregating their outputs by majority voting for classification. By training each tree on different subsets of data and features, Random Forest enhances model robustness and accuracy and reduces overfitting. For SVM and RF, we use Mel Frequency Cepstral Coefficients (MFCCs) as the feature input.

CNN is a deep learning algorithm and our proposed architecture is shown in Fig. 5. This CNN comprises two convolutional layers and two max pooling layers which help in extracting and downsampling spatial hierarchies of features, two fully connected layers that synthesize these features into predictions, and three dropout layers strategically placed to prevent overfitting by randomly omitting a subset of features during training. For CNN, we use the Mel Spectrogram as the feature input. The Mel Spectrogram is considered as a coloured image which has been resized to 224×224 for three colour channels of red, blue and green.

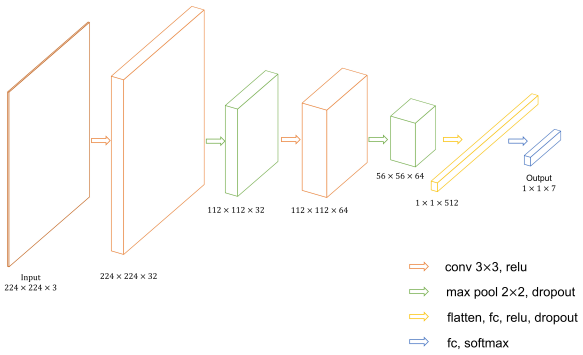


Fig. 5. The architecture of the designed CNN.

Transfer learning aims to learn from diverse and comprehensive data and then apply the learned knowledge to a new task. We use Wav2cec 2.0 [22] as the pre-trained model which is trained on a large amount of audio data. Wav2Vec 2.0 is an end-to-end self-supervised learning model that directly takes raw audio waveforms as input [23]. After pretraining, the model can be fine-tuned to perform classification directly. Fig. 6 shows the architecture of transfer learning using Wav2vec 2.0 as the pre-trained model.

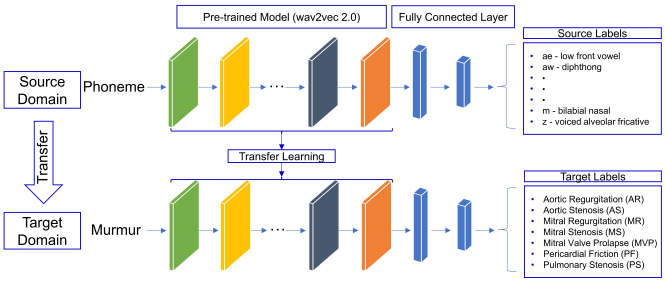


Fig. 6. The architecture of transfer learning using Wav2vec 2.0 as the pre-trained model.

III. Experiment and Results

Data: We utilized the publicly available dataset from [24] that includes four classes of abnormalities: Aortic Stenosis, Mitral Regurgitation, Mitral Stenosis, and Mitral Valve Prolapse. Additionally, we expanded the data by collecting data for three more abnormal classes: Aortic Regurgitation, Pericardial Friction, and Pulmonary Stenosis from online open sources, such as Google.

Separation: Fig. 7 shows the separation results of seven types of murmurs using CSSA and WT.

Classification: Accuracy, F1 Score, Sensitivity, and Specificity are widely used to evaluate the performance of classification models, providing a comprehensive assessment of a model's predictive accuracy and precision in handling both positive and negative classifications.

Accuracy is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

F1 Score is the harmonic mean of Precision and Recall, calculated as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

where Precision is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

and Recall (also known as Sensitivity) is calculated as:

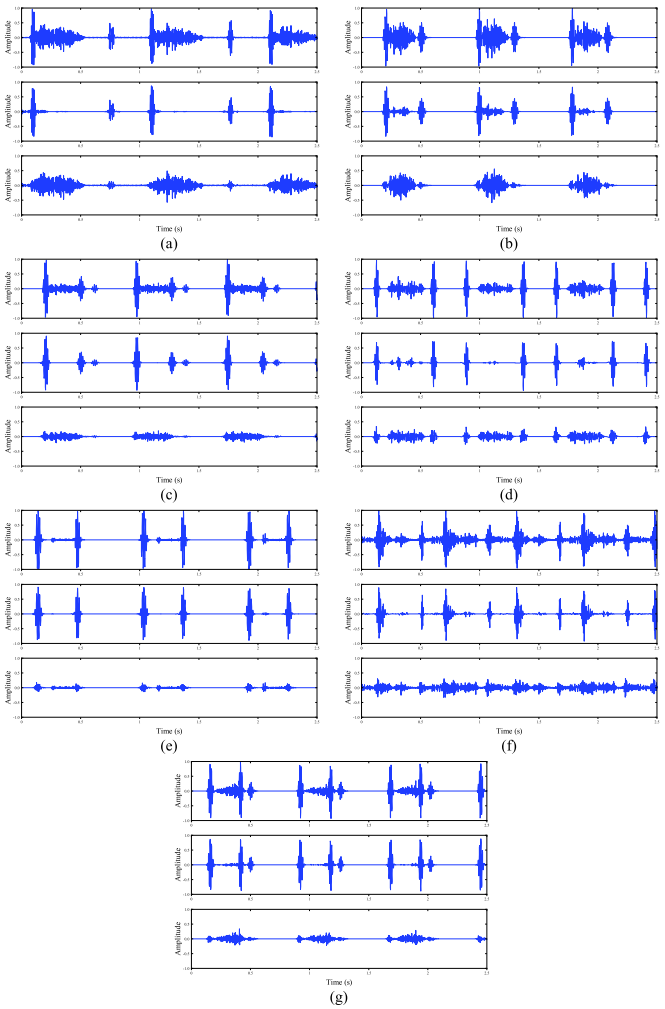


Fig. 7. Separation of seven types of murmurs from the original sounds: the top signal is the original sound, the middle signal is the separated normal heart sound, and the bottom signal is the separated murmur; (a) aortic regurgitation murmur, (b) aortic stenosis murmur, (c) mitral regurgitation murmur, (d) mitral stenosis murmur, (e) mitral valve prolapse murmur, (f) pericardial friction murmur, and (g) pulmonary stenosis murmur.

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN} \tag{10}$$

Specificity is calculated as:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{11}$$

Together, these metrics provide a detailed understanding of a model's performance, allowing evaluation and fine-tuning of models for optimal predictive accuracy and precision in classification tasks.

Tables I and II respectively present the results for two different cases: (i) classification using the original, unseparated sound, and (ii) classification of murmur signals following their separation using CSSA and WT. The results from these tables

clearly demonstrate that the pre-separation of murmurs significantly enhances the performance metrics, achieving higher Accuracy, F1 Score, Sensitivity, and Specificity.

TABLE I
ACCURACY, F1 SCORE, SENSITIVITY, AND SPECIFICITY METRICS FOR THE CLASSIFICATION OF SEVEN MURMUR TYPES (AS, AR, MR, MS, MVP, PF, PS) USING SVM, RF, CNN, AND TRANSFER LEARNING; THE ORIGINAL SOUND HAS BEEN USED HERE.

	SVM	RF	CNN	Transfer Learning
Accuracy	91.27%	94.76%	97.38%	99.13%
F1 Score	92.03%	94.68%	97.33%	98.97%
Sensitivity	91.56%	94.20%	97.47%	98.86%
Specificity	98.50%	99.10%	99.56%	99.85%

TABLE II
ACCURACY, F1 SCORE, SENSITIVITY, AND SPECIFICITY METRICS FOR THE CLASSIFICATION OF SEVEN MURMUR TYPES (AS, AR, MR, MS, MVP, PF, PS) USING SVM, RF, AND TRANSFER LEARNING; THE SEPARATED MURMURS USING CSSA AND WT HAS BEEN USED HERE.

	SVM	RF	CNN	Transfer Learning
Accuracy	94.76%	96.94%	98.69%	99.56%
F1 Score	94.88%	97.42%	98.73%	99.49%
Sensitivity	94.53%	97.29%	98.93%	99.47%
Specificity	99.10%	99.49%	99.78%	99.92%

IV. CONCLUSIONS

In this study, we introduced a novel two-stage method to separate murmurs from heart sounds before applying murmur classification. The initial stage involves utilizing CSSA to decompose the heart sound signal, followed by reconstructing the signal using ZCR and kurtosis as constraints. Subsequently, WT is applied in the second stage to further extract the murmur. This two-stage approach is designed to maximize the precision of murmur separation, thereby improving the performance of subsequent classification tasks. Our experimental results have been promising, showing that pre-separating murmurs can enhance classification performance, with accuracy levels exceeding 99% after using transfer learning. This methodological advancement has the potential to greatly improve the quality and effectiveness of medical care, enabling more accurate and reliable diagnosis of heart conditions.

REFERENCES

[1] C. Xu, X. Li, X. Zhang, *et al.*, "Cardiac murmur grading and risk analysis of cardiac diseases based on adaptable heterogeneous-modality multi-task learning," *Health Information Science and Systems*, vol. 12, no. 1, p. 2, 2023.

[2] M. M. Milani, P. E. Abas, and L. C. De Silva, "A critical review of heart sound signal segmentation algorithms," *Smart Health*, vol. 24, p. 100 283, 2022.

[3] J. Prince, J. Maidens, S. Kieu, *et al.*, "Deep learning algorithms to detect murmurs associated with structural heart disease," *Journal of the American Heart Association*, vol. 12, no. 20, e030377, 2023.

[4] C. J. Riley, "Assessment of the heart," *History and Physical for the Pediatric Dental Patient: Establishing a Systematic Approach for Procedural Sedation*, pp. 43–60, 2024.

[5] A. F. Gündüz and A. Karci, "Heart sound classification for murmur abnormality detection using an ensemble approach based on traditional classifiers and feature sets," *Computer Science*, vol. 5, no. 1, pp. 1–13, 2022.

[6] V. Nivitha Varghees and K. I. Ramachandran, "Heart murmur detection and classification using wavelet transform and hilbert phase envelope," in *2015 Twenty First National Conference on Communications (NCC)*, 2015, pp. 1–6.

[7] A. D. Levin, A. Ragazzi, S. L. Szot, and T. Ning, "A machine learning approach to heart murmur detection and classification," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2020, pp. 521–525.

[8] S. Sudarsanan and J. Aravinth, "Classification of heart murmur using cnn," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2020, pp. 818–822.

[9] B. Walker, F. Krones, I. Kiskin, G. Parsons, T. Lyons, and A. Mahdi, "Dual bayesian resnet: A deep learning approach to heart murmur detection," in *2022 Computing in Cardiology (CinC)*, vol. 498, 2022, pp. 1–4.

[10] V. Vakamullu, M. Mishra, and A. Mukherjee, "Real-time heart murmur classification using attention based deep learning approach," in *2022 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2022, pp. 1–6.

[11] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings 13*, Springer, 2017, pp. 258–266.

[12] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[13] K.-H. Tsai, W.-C. Wang, C.-H. Cheng, *et al.*, "Blind monaural source separation on heart and lung sounds based on periodic-coded deep autoencoder," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3203–3214, 2020.

[14] H. Tang, M. Wang, Y. Hu, B. Guo, and T. Li, "Automated signal quality assessment for heart sound signal by novel features and evaluation in open public datasets," *BioMed Research International*, vol. 2021, no. 1, p. 7 565 398, 2021.

[15] S. Sanei, M. Ghodsi, and H. Hassani, "An adaptive singular spectrum analysis approach to murmur detection from heart sounds," *Medical engineering & physics*, vol. 33, no. 3, pp. 362–367, 2011.

[16] S. Sanei and H. Hassani, *Singular spectrum analysis of biomedical signals*. CRC press, 2015.

[17] C. S. Burrus, R. A. Gopinath, and H. Guo, "Wavelets and wavelet transforms," *rice university, houston edition*, vol. 98, 1998.

[18] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[20] Y. LeCun, B. Boser, J. S. Denker, *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[21] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010, pp. 242–264.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] D. Krishna, P. Wang, and B. Bozza, "Using large self-supervised models for low-resource speech recognition.," in *Interspeech*, 2021, pp. 2436–2440.

[24] Yaseen, G.-Y. Son, and S. Kwon, "Classification of Heart Sound Signal Using Multiple Features," *Applied Sciences*, vol. 8, no. 12, p. 2344, 2018.

VietSing: A High-quality Vietnamese Singing Voice Corpus

Minh Vu*, Zhou Wei†, Binit Bhattarai‡, Kah Kuan Teh§ and Tran Huy Dat¶
Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
* E-mail: minhducvu1311@gmail.com
† E-mail: zhouwei_2001@hotmail.com
‡ E-mail: binit.bhattarai10@gmail.com
§ E-mail: teh_kah_kuan@i2r.a-star.edu.sg
¶ E-mail: hdtran@i2r.a-star.edu.sg

***Abstract*—This paper introduces a comprehensive Vietnamese dataset designed for singing voice synthesis (SVS). While there are extensive datasets available for widely spoken languages such as English and Chinese, resources for less common languages like Vietnamese are still scarce. The dataset, VietSing, comprises high-quality audio recordings and corresponding phonetic annotations, meticulously curated to support the development and evaluation of SVS systems. Detailed phonetic transcriptions and alignment with musical scores are provided to facilitate precise modeling of Vietnamese phonetics and prosody in song. We outline the data collection process, annotation methodology, and the challenges faced in ensuring linguistic, musical accuracy and how we tackled these problems using deep learning methods. Initial experiments using popular SVS model demonstrate the potential of VietSing to enhance the naturalness and intelligibility of synthesized Vietnamese singing voices. The dataset is licensable at ai@artiligent.net. Some audio samples and the code for our baseline evaluation methods can be found at this link[1].**

I. Introduction

Recently, singing voice synthesis (SVS) using deep learning model [1]–[6] has drawn a lot of attention from both industry and academic communities. The progress in deep learning, especially in neural network based text-to-speech (TTS) systems [7]–[9], has made the production of emotional speech [10] and singing voices more attractive.

In addition to improving deep learning architectures, one of the biggest challenges in the SVS task is obtaining an appropriate singing database. Unlike TTS task, which only requires audio and corresponding textual transcriptions, SVS models must also include the musical information of the audio. The process of collecting the singing voice data tends to be costly and typically necessitates self-recording to ensure the desired quality. Although numerous SVS datasets have been released for widely spoken languages such as English [11], [12], Chinese [1], [13], [14], Japanese [15], [16], there are no available Vietnamese singing corpora. As shown in Table I, despite several singing datasets in different languages having more hours and singers, not all of them include alignment and musical scores, which are crucial for SVS models. Different from many other tasks, Singing Voice Synthesis often requires fairly high-quality audio with clear voice, high sampling rate, etc. We find it impossible to obtain music data with the desired quality from open internet sources, and it is also very challenging to completely eliminate background music noise. Therefore, we decide to build the dataset from self-recorded songs.

TABLE I
Comparison of different singing datasets

Corpus	Language	#Hours	#Singers	Alignment	Score
NUS-48E [11]	English	1.91	12	✓	✗
NHSS [12]	English	7	10	✗	✗
JVS-MuSiC [15]	Japanese	2.28	100	✗	✗
Tohoku Kiritan [16]	Japanese	1	1	✗	✗
PopCS [1]	Chinese	5.89	1	✗	✗
M4Singer [14]	Chinese	29.77	20	✓	✓
Opencpop [13]	Chinese	5.25	1	✓	✗
VietSing (Our)	**Vietnamese**	**2.05**	**1**	✓	✓

This paper introduces VietSing, a high-quality Vietnamese singing dataset designed for the SVS task. In this corpus, we select 60 songs to be recorded by a professional female singer. The audio quality is guaranteed since all songs are recorded in a studio using modern devices. Further statistics about the corpus will be discussed in the following sections. Additionally, we present experimental results from testing the dataset with DiffSinger [1] model.

II. VietSing Dataset

In this section, we introduce the construction pipeline and the statistical aspects of VietSing. This dataset is specifically designed for the SVS task and created using an optimized strategy that minimizes the need for manual labeling. Fig. 1 shows the steps of creating the corpus. The details of each step are described in the following subsections.

A. Songs and singer

All the songs chosen for the corpus are Vietnamese songs from the late 1990s and early 2000s, as songs from this period typically contain fewer special sound effects, making it easier for our singer to sing them. These songs must meet the following criteria:

[1]https://github.com/rosotron/VietSing

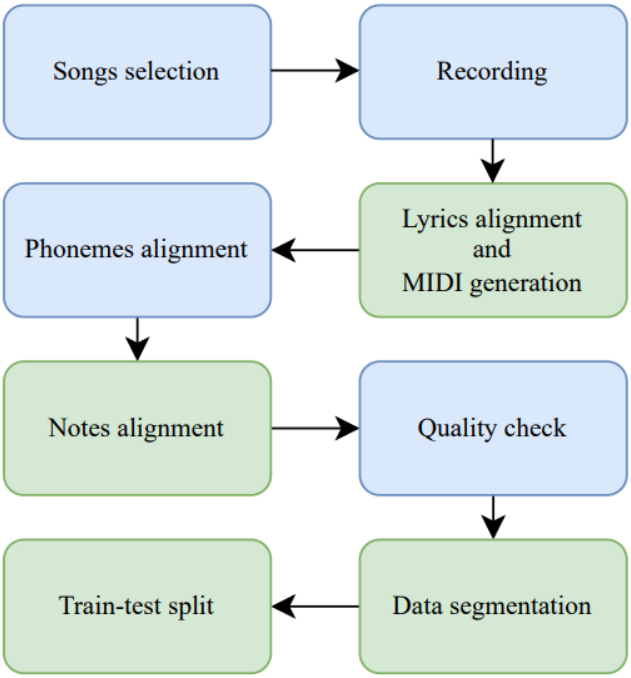


Fig. 1. Pipeline for creating VietSing. Blue boxes show manual steps and green boxes show code-based steps.

- Lyrics are clearly sung
- Songs with code-switching lyrics are not chosen

B. Recording

All the songs are recorded in a professional studio environment with minimal reverb. During the recording, the singer wears headphones to control the singing voice. The recordings are made using a Neumann U87 condenser microphone, capturing only the singing voice without any background music. The entire corpus is recorded at a sampling rate of 88.2 kHz with a bit depth of 24 bits.

C. Data labeling

To minimize manual labeling effort for VietSing, we employ various code-based techniques (detailed in Fig.2). Subsequently, we implement a random quality assurance check on the automatically annotated data, which we will describe in this section.

1) MIDI annotation: In the context of Singing Voice Synthesis, detailed musical scores, including elements like note pitch, duration, and tempo, are crucial for training. Annotating such scores typically necessitates expertise in music theory. However, the initial absence of a qualified annotator presents a significant challenge. To address this, we construct a semi-automated annotation process that leverages publicly available, code-driven solutions to handle the musical score annotation steps.

After recording, we create MIDI file from singing voice audio file using Omnizart [17], a Python library and a streamlined solution for automatic music transcription. Omnizart [17] provides options to extract note pitch and note duration from singing vocal. The quality of the generated MIDI files is subsequently assessed through subjective listening tests. The pitch values of the notes are translated into their respective note names and octaves, with A4 being equivalent to 69 in MIDI notation and corresponding to 440 Hz. While note duration were initially set, they are adjusted to align with manually annotated phoneme duration. The process of phoneme annotation will be explained in the following subsections.

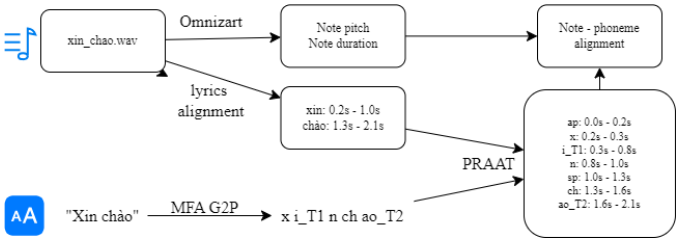


Fig. 2. Pipeline for extraction and alignment of note information and phoneme.

2) Phonemes annotation: In Vietnamese, phonemic tone is crucial for both spoken and sung communication. While Mandarin, another tonal language, possesses a large number of Singing Voice Synthesis databases, its tonal simplicity allows listeners to comprehend lyrics independent of tone information. In Vietnamese songs, however, tone realization is important due to the interplay between semantic meaning and musical pitch. Vietnamese has a complex sound system with distinct features depending on the dialect. The standard Vietnamese phonology includes 6 tones, 14 vowel phonemes, and 19 consonant phonemes, but the Southern and Northern dialects have additional consonants unique to each region.

Due to the prevalence of the Northern dialect in Vietnamese music, particularly songs sung in the Hanoi accent, we use the Vietnamese (Hanoi) Montreal Forced Alignment (MFA) [18] G2P model v1.0.0. This model encompasses a comprehensive phonological set of 223 elements, incorporating both tones and phonemes. After splitting the lyrics into corresponding phonological elements, we perform manual phoneme alignment. We label the timestamp of each phoneme in the lyrics using PRAAT [19], this process also includes labeling special phonemes: 'sp' - spoken noise, 'sil' - silence, 'ap' - aspirate. Before manual annotation, we create a preliminary alignment using the Vietnamese song lyric alignment framework from Zalo AI Challenge 2022[2], this has significantly improved the annotating pace and accuracy. Fig. 2 illustrates the steps of phoneme alignment.

D. Note - Phoneme alignment

As described in previous sections, we extract the note pitch and duration using a deep-learning-based tool and we label the phoneme information manually. Consequently, the initial timestamps for phonemes and notes do not match. We assume that each note corresponds to a single phoneme, while a

[2]https://github.com/nguyenvulebinh/lyric-alignment

TABLE II
THE NUMBER OF OCCURRENCE OF SEVERAL PHONEMES BEFORE AND AFTER DATA FILLING AND AUGMENTATION.

Phoneme	Before	After
s	1548	942
x	108	714
ch	2202	1575
tr	75	702
ã	0	196

phoneme may extend across multiple slurred notes. To address this, we duplicate each phoneme according to the number of notes occurring within the same timestamp. These notes are designated as slurred, with the first slur note marked as 0 and the subsequent slur notes marked as 1. Conversely, for word boundaries, the initial phoneme of a word is marked with 1, while the remaining phonemes in that word are marked with 0. Special phonemes such as 'ap', 'sp', and 'sil' are automatically aligned with a note pitch of 0, and their duration are calculated based on the labels.

E. Phonemes filling and augmentation

Following the above steps, we analyze the database and observe differences in the frequency of phoneme occurrences, with some rare phonemes being absent from the recorded songs. To resolve this issue, we augment the dataset by adding and balancing the missing phonemes. Table II illustrates changes in occurrence number of several phonemes before and after data filling and augmentation.

1) Phoneme augmentation: We augment the data by leveraging the unique pronunciation characteristics of the Vietnamese Hanoi accent. Specifically, in speaking and singing with the Hanoi accent, certain phonemes are pronounced identically, such as 'ch' and 'tr', 'd', 'r', and 'gi', and 's' and 'x'. To balance the corpus, we replace the less frequent phonemes with their pronounced equivalents until we achieve a more balanced ratio between these phonemes.

2) Missing phoneme filling: After the phoneme augmentation step, 26 phonemes are still missing from the corpus due to their rarity and distinct pronunciation from other phonemes. To ensure the dataset fully covered the MFA [18] phone set, we record these missing phonemes. However, due to many reasons, the original singer of the corpus is unavailable for additional recordings. Therefore, we record these new phonemes with a different singer and convert the new singer's voice to the original singer's voice using the Singing Voice Conversion (SVC) technique. The singer is asked to sing each of the missing phonemes individually in random melodies, and we save each phoneme in a separate audio file. We utilize So-VITS [20] as our SVC system to clone the original singer's voice, training the model with the full set of audio data from the original singer over 200 epochs. The cloned voice audio files are then evaluated through listening tests to ensure they closely resemble the original singer's voice. After that, we run the process of annotating notes and phonemes on the newly recorded data.

A representative data sample is presented in Table III, comprising raw audio, lyrics, phonemes, notes, note and phoneme durations, slur notes and the word boundary. This example illustrates the annotation process and the resulting transcription of an audio clip.

F. Post processing

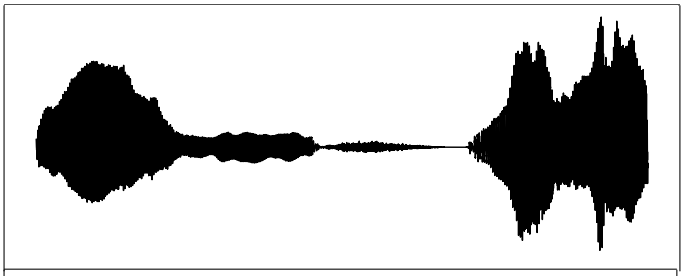
1) Audio segmentation: After the labeling steps, we segment each song into small parts, with each part not exceeding 11 seconds. The segmentation points are determined based on the end times of words in the lyrics. This segmentation is necessary for the SVS system to train effectively. The final corpus consists of 729 utterances and 2.05 hours of singing voice audio.

2) Train test split: We select 3 songs out of the 60-song corpus for the testing process. The chosen songs represent a range of average pitch values: the second lowest, the second highest, and a medium average pitch value. This selection provides a comprehensive evaluation of the SVS model. The chosen songs are song 2021 (57.35 average pitch value), song 2026 (60.20 average pitch value), and song 2034 (64.63 average pitch value). Song 2026 presents an additional challenge due to its numerous vocal ornaments and pronunciation adjustments by the singer to fit the notes, which happens sometimes in Vietnamese singing.

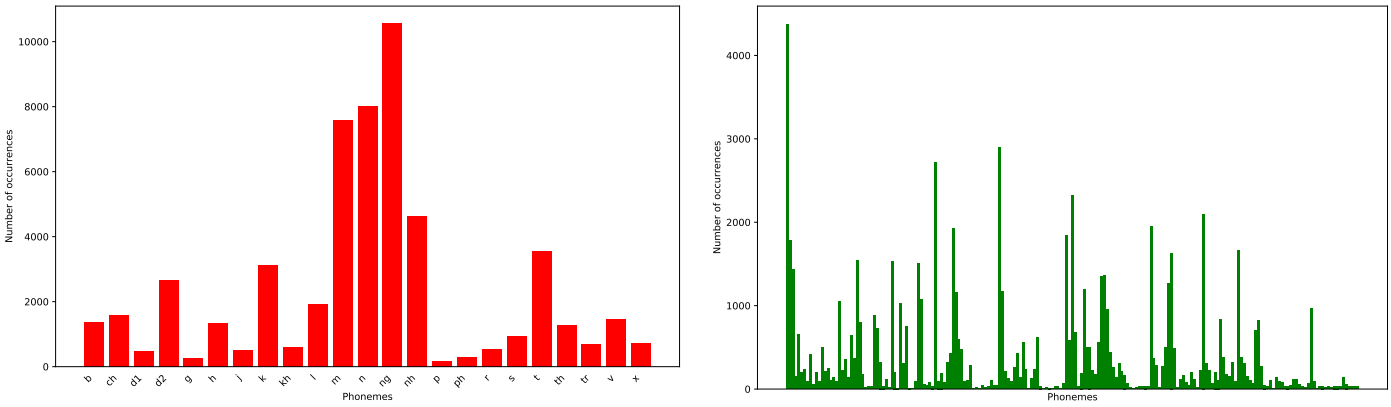
G. Statistics

As mentioned in Section II-C2, our phoneme set is composed of tonal phonemes, resulting in a relatively large number of distinct elements. Despite the natural rarity of certain phonemes in Vietnamese, our VietSing corpus covers all possible tonal phonemes. However, the frequency of some phonemes is higher than others. Fig. 3a and Fig. 3b display the distribution of consonant and vowel phonemes in our dataset, respectively.

TABLE III
TRANSCRIPTION OF A SAMPLE AUDIO CLIP.
FROM TOP TO BOTTOM: WAVEFORM, LYRICS, SYLLABLE, PHONEME, NOTES, DURATION, SLUR NOTES, WORD BOUNDARY.



anh yêu anh												
anh				SP	yêu					anh		
a1_T1	nh			SP	i_T1		e2_T1		u1_T1	a1_T1	nh	
C#4	C#4	E4	D#4	Rest	D#4	A#3	A#3	A#3	D4	D4	D4	D#4
0.224	0.078	0.139	0.137	0.29	0.083	0.037	0.073	0.029	0.067	0.103	0.016	0.111
0	0	1	1	0	0	1	0	0	1	0	0	1
1	0	0	0	1	1	0	0	0	0	1	0	0



(a) Consonant phonemes

(b) Vowel phonemes

Fig. 3. The statistical distribution of phonemes

VietSing comprises 60 songs with varying BPMs, ranging from 80 to 220. The distribution for BPM is illustrated in Fig. 4. This diversity in tempo is beneficial for the SVS system, enabling it to recreate singing voices across a wide range of BPMs. Fig. 5 represents the statistical distribution of note pitch in VietSing. The note pitch values range from 47 (B2, 123.5 Hz) to 76 (E5, 659.3 Hz) with 60 (C4, 261.6 Hz) occurs the most.

III. Benchmarks

We perform several benchmark tests to verify the quality of VietSing for singing voice synthesis task.

A. Method

Singing Voice Synthesis has experienced notable progress due to the emergence of deep learning methodologies. These advancements encompass innovations in non-autoregressive models [2], [21], [22], diffusion models [1], and end-to-end models [5]. Among the diverse systems designed for SVS, DiffSinger [1] and ViSinger2 [5] stands out as a prominent framework for producing high-quality singing voices. Despite its achievements, ongoing efforts persist in enhancing the richness and expressiveness of synthesized voices, including those for Vietnamese songs.

In our experimental setup for our baseline, we employ two approaches to SVS. We use DiffSinger [1] as the acoustic model and HifiGAN [2] as the vocoder for the two stage approach and ViSinger2 [5] for the end to end approach for the singing voice synthesis task. DiffSinger [1] is grounded in the diffusion probabilistic model, while Hifi-GAN [2] serves as the vocoder. The generation process in this model determined by the intersection of diffusion trajectories between the ground-truth mel-spectrogram and the prediction from a straightforward mel-spectrogram decoder. Unlike conventional acoustic models that rely solely on basic L1 or L2 loss functions, which may lead to excessively smooth outputs. DiffSinger [1] adopts a hybrid approach, it computes the average of the L1 loss and the structural similarity (SSIM) index, resulting in mel-spectrograms that closely approximate the ground truth. To generate audio waveforms from the synthesized mel-spectrograms, we leverage a pretrained Hifi-GAN [2] vocoder.

On the other hand, ViSinger2 [5] is an end-to-end SVS model based on conditional variational autoencoder models with two separate encoders and a decoder. The posterior

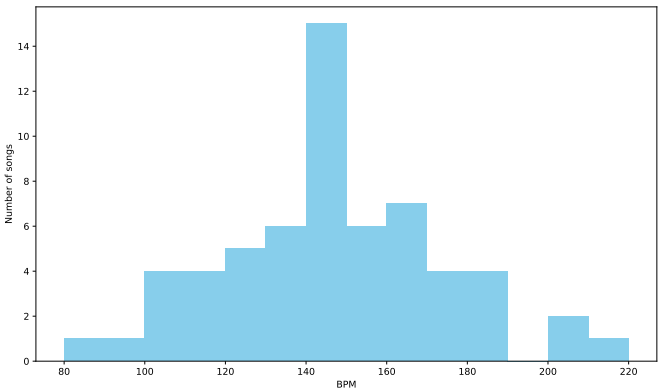


Fig. 4. The statistical distribution of BPM.

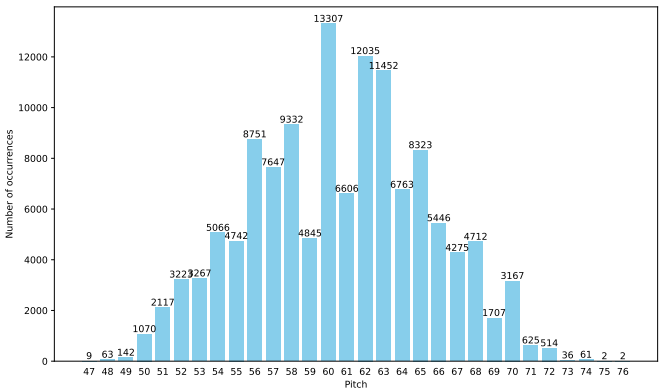


Fig. 5. The statistical distribution of note pitch with note A4 = 69 = 440 Hz.

encoder uses 1-D convolution layers to extract latent representations from mel-spectrograms, while the decoder generates waveforms from these representations, incorporating DSP synthesizers to produce periodic and aperiodic components. These components are used to enhance HiFi-GAN [2]'s performance, mitigating phase and glitch issues. The prior encoder, based on FastSpeech [4], provides constraints by predicting fundamental frequency and mel-spectrograms. The model's loss functions include GAN loss, KL divergence between the two encoders, DSP synthesizer loss, duration loss, and auxiliary feature loss, aiming to improve synthesis quality.

B. Experimental setup

To perform the baseline test using DiffSinger, the audio is down-sampled to 24,000 Hz and the mel-spectrogram is extracted using 80 Mel bins and linearly scaled from -1 to 1. A pretrained vocoder, trained on 70 hours of singing data, is used as a universal vocoder. The model is trained for 1000 epochs with a learning rate of 0.001. The training process takes approximately 2 days using an NVIDIA Quadro RTX5000 16G. For training VISinger2[5], the model is trained up to 720K steps with a batch size of 4 using 3 NVIDIA Quadro RTX5000 16G.

C. Results

To evaluate the SVS performance, we use duration accuracy (duracc) and mean opinion score (MOS). The score of MOS test ranges from 1 to 5, in which 1 means very bad and 5 means excellent. Each audio is rated by 15 qualified listeners. Within this evaluation, MOS-Q indicates the quality of the audio, while MOS-P measures the coherence and naturalness of the prosody. Table IV presents the results. The model excels in accurately reproducing duration length. The MOS scores range between 3 (average) and 4 (good), demonstrating the reliability of our corpus. Overall, the dataset produces stable audio quality (MOS-Q). For voice coherence and naturalness (MOS-P), the corpus receives slightly lower scores, but they remain at an acceptable level.

TABLE IV
OBJECTIVE AND SUBJECTIVE EVALUATION FOR VIETSING DATASET. FOR ALL THE METRICS, HIGHER VALUE MEANS BETTER PERFORMANCE, WITH "DURACC" MEANS DURATION ACCURACY OF GENERATED AUDIO COMPARED TO THE ORIGINAL AUDIO, "MOS-Q" INDICATES AUDIO QUALITY AND "MOS-P" REPRESENTS COHERENCE AND NATURALNESS OF PROSODY.

Test Set	Method	duracc	MOS-Q	MOS-P
low-pitch	Ground Truth	-	4.79±0.24	4.41±0.09
	DiffSinger[1]	0.996	3.80±0.20	3.79±0.21
	VISinger2[5]	0.994	3.65±0.05	3.70±0.10
mid-pitch	Ground Truth	-	4.70±0.20	4.57±0.07
	DiffSinger[1]	0.992	3.94±0.06	3.85±0.05
	VISinger2[5]	0.996	3.75±0.05	3.75±0.15
high-pitch	Ground Truth	-	4.65±0.05	4.62±0.02
	DiffSinger[1]	0.989	3.71±0.19	3.69±0.21
	VISinger2[5]	0.992	3.60±0.10	3.69±0.11

TABLE V
THE AUDIO PERFORMANCE BEFORE AND AFTER DATA AUGMENTATION BY DIALECT CHARACTERISTICS.

Method	duracc	MOS-Q	MOS-P
Ground Truth	-	4.32±0.18	4.41±0.09
DiffSinger[1] - before augment	0.966	3.70±0.10	3.44±0.06
DiffSinger[1] - after augment	0.987	3.73±0.17	3.59±0.11

TABLE VI
THE AUDIO PERFORMANCE OF VOICE-CLONED PHONEMES BY DIFFSINGER[1].

Metric	Score
MOS-Q	3.15±0.10
MOS-P	3.26±0.04

Moreover, to assess the effectiveness of phoneme augmentation and filling as discussed in Section II-C2, we prepare a separate test song for the model trained with data both before and after augmentation. To evaluate the missing phonemes filled by the SVC [20] technique, we modify the lyrics and phonemes of the test song to include these voice-cloned phonemes. The performance of these tests are measured using MOS-Q and MOS-P and is presented in Tables V and VI. As shown, data augmentation generally improves the synthesis quality. For the phonemes filled by So-VITS [20], the scores remain at acceptable levels, though it is true that the tonality of these phonemes is not very well covered.

IV. CONCLUSION

In this paper, we present the VietSing dataset, a Vietnamese singing voice corpus specifically designed for SVS systems. All audio songs in our corpus are self-recorded using professional equipment. The annotation process combines manual and automatic methods, and we augment the data by leveraging unique features of the Vietnamese dialect.

We acknowledge several limitations of VietSing. First, despite random checks for annotation quality, there are inevitable errors in musical labeling when using deep learning methods. Second, to achieve the best synthesized results with our corpus, we recommend users to add simulated slur notes, as we have marked many notes as slurred to align with the notes detected by Omnizart [17]. Finally, we acknowledge that several tonal phonemes appear less frequently than others due to their natural rarity in music lyrics, which may result in occasional lack of clarity for these phonemes.

REFERENCES

[1] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, pp. 11 020–11 028.

[2] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[3] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Applied Sciences*, vol. 7, no. 12, p. 1313, 2017.

[4] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[5] Y. Zhang, H. Xue, H. Li, *et al.*, "VISinger2: High-Fidelity End-to-End Singing Voice Synthesis Enhanced by Digital Signal Processing Synthesizer," in *Proc. INTERSPEECH 2023*, 2023, pp. 4444–4448. DOI: 10.21437/Interspeech.2023-391.

[6] K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on convolutional neural networks," *arXiv preprint arXiv:1904.06868*, 2019.

[7] X. Tan, J. Chen, H. Liu, *et al.*, "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[8] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[9] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6588–6592.

[10] "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022, ISSN: 0167-6393.

[11] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, IEEE, 2013, pp. 1–9.

[12] B. Sharma, X. Gao, K. Vijayan, X. Tian, and H. Li, "Nhss: A speech and singing parallel database," *Speech Communication*, vol. 133, pp. 9–22, 2021.

[13] Y. Wang, X. Wang, P. Zhu, *et al.*, "Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis," in *Proc. Interspeech 2022*, 2022, pp. 4242–4246. DOI: 10.21437/Interspeech.2022-48.

[14] L. Zhang, R. Li, S. Wang, *et al.*, "M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6914–6926, 2022.

[15] H. Tamaru, S. Takamichi, N. Tanji, and H. Saruwatari, "Jvs-music: Japanese multispeaker singing-voice corpus. arxiv 2020," *arXiv preprint arXiv:2001.07044*,

[16] I. Ogawa and M. Morise, "Tohoku kiritan singing database: A singing database for statistical parametric singing synthesis using japanese pop songs," *Acoustical Science and Technology*, vol. 42, no. 3, pp. 140–145, 2021.

[17] Y.-T. Wu, Y.-J. Luo, T.-P. Chen, *et al.*, "Omnizart: A general toolbox for automatic music transcription," *Journal of Open Source Software*, vol. 6, no. 68, p. 3391, 2021. DOI: 10.21105/joss.03391. [Online]. Available: https://doi.org/10.21105/joss.03391.

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[19] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[20] Z. Ning, Y. Jiang, Z. Wang, B. Zhang, and L. Xie, "Vits-based singing voice conversion leveraging whisper and multi-scale f0 modeling," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–8.

[21] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, "Xiaoicesing: A high-quality and integrated singing voice synthesis system," *arXiv preprint arXiv:2006.06261*, 2020.

[22] C. Wang, C. Zeng, and X. He, "Xiaoicesing 2: A high-fidelity singing voice synthesizer based on generative adversarial network," *arXiv preprint arXiv:2210.14666*, 2022.

Learning a Sequence of Cursive-Style Japanese Characters in Classical Literary Works

Satoru Fujita* and Keizo Oyama†
* Hosei University, Japan
E-mail: fujita_s@hosei.ac.jp Tel: +81-42-387-4545
† National Institute of Japanese Literature, Japan and
National Institute of Informatics, Japan
E-mail: oyama.keizo@nijl.ac.jp Tel: +81-50-5533-2900

***Abstract*—This paper presents a method for analyzing and recognizing Kuzushiji characters in images using Transformer models. Kuzushiji are cursive-style Japanese characters historically employed in the writing and printing of Japanese classical literature from the eighth to the nineteenth century. Japanese characters are categorized into Hiragana, Katakana, and Kanji characters. Hiragana and Katakana are phonetic symbols, each comprising approximately 50 classes of basic symbols in contemporary Japanese. Classical Kuzushiji writings, however, include more than just the basic forms of Hiragana and Katakana, as some of them originated from modifications of different Chinese characters with similar phonetic sounds. Additionally, Kuzushiji writings include numerous Kanji characters, which number in the thousands and are individually simplified and altered. In addition to such numerous classes of characters in Kuzushiji, these characters are written in a cursive style, where strokes are simplified within a character and are interconnected between characters. Our approach for recognizing cursive-style Kuzushiji characters employs Transformer architecture, as it is well-suited for analyzing characters that are vertically connected without explicit borders or spaces. We anticipate that the attention mechanism within the Transformer will effectively learn the relationships among adjacent characters, thereby aiding in the accurate separation and identification of character regions and boundaries. Our project has just started, but we have already achieved 90% accuracy for transcription of pre-modern literary works in our experiments.**

I. Introduction

Cursive-style Japanese characters, known as Kuzushiji, were used in writing and printing of Japanese classical literature from the eighth to the nineteenth century[1]. Kanji characters originated from Chinese characters, extensively used in Japanese historical records around the sixth century, following Chinese grammatical structures. Subsequently, Japanese native phonetic sounds were mapped to Chinese characters with similar pronunciations, leading to the invention of Hiragana and Katakana, which were derived from cursive and simplified versions of their Chinese counterparts. Since they were written in a connecting cursive style from top to bottom direction, it presents significant challenges for machine and contemporary Japanese to correctly read them. Additionally, Japanese language includes Kanji characters, which are semantic ideograms closely resembling Chinese characters. Japanese literary works were thus written in a connected format that integrates cursive styles of Hiragana, Katakana and Kanji characters.

Many valuable documents that provide insights into Japanese classical culture were written in Kuzushiji. However, only a few Japanese people today are able to read Kuzushiji due to the current educational system's lack of support for Kuzushiji literacy. Consequently, there is a need to transcribe Kuzushiji documents into contemporary Japanese characters to make them accessible to public.

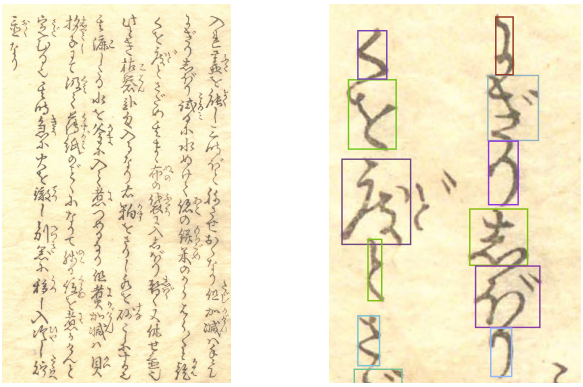
Research on Kuzushiji recognition has been active, since the Kuzushiji dataset became available[2]. Most of existing Kuzushiji research employed CNN models to capture the features of character shapes[1][3][4]. However, these approaches often ignore the valuable sequential relationships between characters. Kuzushiji characters are actually connected by cursive lines created through brush strokes, which are essential for accurately interpreting their sequences and structures. Attention mechanisms were proposed to solve such sequential issues using BLSTM[5][6]. According to following these directions, this paper proposes utilizing the Transformer model[7] for the analysis of sequential connection of characters. Given that a sequence of characters forms a word, and a sequence of words forms a sentence, we initially designed the Transformer architecture for character image recognition, but can extend it to sentence and context recognition. These results of extension could provide another valuable feedback to enhance the basic performance of character recognition. Our ultimate objective is not only to accurately identify Kuzushiji characters but also to recognize their contextual meanings within classical literary works. We would expect that the Transformer model could serve as the optimal bridge from the sequential character recognition to the semantic analysis of entire documents.

Section II briefly introduces Kuzushiji characters in history and in the related existing research work. Section III explains our proposed architecture that consists of CNN frontend and the Transformer. Section IV conducts experiments of the proposed architecture. Section V discusses the issues found in this study, and Section VI describes conclusions in this paper.

II. Kuzushiji Characters and the Existing Approaches for Character Recognition

A. Kuzushiji

Kuzushiji are cursive-style characters found in Japanese classical literature. Fig. 1 shows a sample document written



(a) Japanese classical literary work written in Kuzushiji.

(b) Zoom-in image with bounding boxes.

Fig. 1. Sample of Kuzushiji document in Kaggle[8].

in Kuzushiji from the dataset of Kaggle competition[8]. Fig. 1(a) depicts a full-page image in which a sequence of Kuzushiji characters are written vertically from top to bottom, with multiple sequences arranged horizontally from right to left. This top-to-down and right-to-left arrangement is a common order for reading vertically written documents in Japanese. Fig. 1(b) is a zoomed-in image focusing on a portion of Fig. 1(a). The rectangles represent bounding boxes for individual characters, although there are no explicit spaces between the connecting characters, and the bounding boxes are sometimes overlapped. Since Kuzushiji is written with a Japanese writing brush, the strokes within a character often continue into subsequent strokes, and characters themselves often continue into the next characters. These characteristics of Kuzushiji make it difficult to read them as distinct separate characters.

Here is a list of some notable difficulties involved in Kuzushiji recognition.

- Within a character, the strokes of Hiragana, Katakana, and Kanji flow continuously into subsequent strokes without interruption.
- Adjacent characters are sometimes connected by their strokes.
- Some Hiragana characters of completely different shapes, each originating from different Chinese characters, are used interchangeably as the same phonetic symbols in a text.
- The bounding boxes for these characters are not clearly separated. Their sizes and aspect ratios vary and are inconsistent.
- In addition to Hiragana and Katakana of around fifty classes respectively, Kanji of several thousand classes are contained in texts. Furthermore, many of these Kanji characters appear only once or twice in a text. Consequently, it is very difficult to learn Kanji characters that are rarely encountered.

This paper employs 'Kuzushiji Recognition' dataset from the Kaggle competition[8] for training and evaluation. The dataset comprises 683,464 characters, with 4,212 distinct classes of characters. Among these, 461 classes are Hiragana, Katakana, and simple symbols, which collectively account for 66% of all characters. Additionally, the top 10% of character classes account for 88% of occurrences, while nearly half of the character classes appear fewer than ten times each. It is observed that achieving 88% accuracy is possible by correctly recognizing only the top 10% of character classes without errors. However, raising the recognition rate beyond 90% is very challenging, as there are very few sample instances of less common characters in the training data.

B. Prior Kuzushiji Recognition

In the Kaggle competition, most of the top teams adopted the two-stage method, such as a Faster R-CNN or CenterNet to detect character regions, and ResNet to recognize individual characters[2]. All of these were deviations of CNN families, which were suitable for pattern recognition but had limitation to recognize a sequential relationship among characters.

Kuronet is an end-to-end method for recognizing Kuzushiji using neural networks[1][3]. It processes a 976x976 pixel image using a U-Net architecture[9] to generate a feature representation. This representation indicates the probabilities of the candidate characters for each pixel. The character with the highest probability at each pixel is selected, and cluster of pixels with the same character are grouped to form bounding boxes. This architecture is notable because it does not require any predefined bounding boxes. Instead, it analyzes the entire image to simultaneously determine the correct class of character and its boundary in an end-to-end process through convolutional calculations around each pixel.

Hu et al. applied an attention mechanism for recognizing consecutive Kuzushiji characters within phrases[6]. They utilized an encoder-decoder architecture, and the decoder consisted of the attention and LSTM. The attention could determine a character sequence from a image feature map in an order. Similarly, Le et al. achieved human-like focus movement from a character to the next, utilizing an attention mechanism with LSTM[5]. DKNet adopted a pre-trained MobileNet to extract the features of Kuzushiji characters[10]. It utilized a trilateral filter to preserve edges of Kuzushiji, and reduced over-fitting and gradient vanishing effects in Kuzushiji recognition by ReLU with dropout layers.

C. Object Detection

Object detection is a technique used to find and identify objects within an image. Traditionally, this process involved two separate steps: identifying the bounding boxes of the targets, and determining their categories. In the two-step approach, however, the performance of each step impacts the final detection quality, often resulting in lower accuracy than expected. Recently, end-to-end approaches using the Transformer model have been proposed for object detection. DETR (Detection Transformer) is a pioneering example of such an approach[11]. In DETR, the image is first processed by a CNN frontend and then passed to the Transformer encoder, which incorporates two dimensional positional encodings. Query bounding boxes are input to the Transformer decoder, and the outputs from

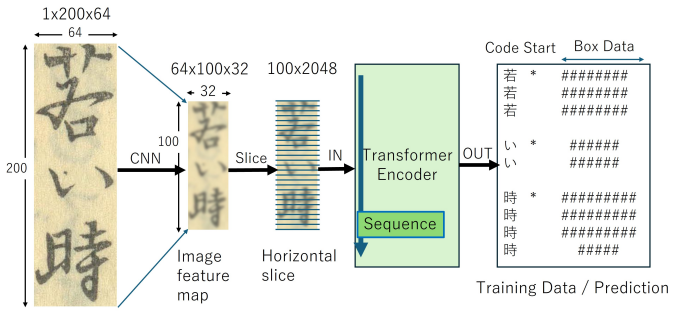


Fig. 2. Schematic flow of Kuzushiji recognition.

the Transformer decoder are bounding boxes containing the detected objects. These outputs are compared with a set of ground truth objects using bipartite matching, as the order of the detected objects is undetermined. In this framework, the Transformer effectively identifies vertical and horizontal relationships within an image. Character recognition can be considered as a type of object detection, but the sequential relationship along one direction is much more crucial than another direction.

III. Kuzushiji Recognition

A. Overview

The ultimate objective of this project is to analyze and interpret the contents of Japanese classical literary works preserved in ancient cultural heritage texts. We adopted the Transformer architecture to recognize sequences of characters, words, and sentences, similar to the process used in large language models. Our goal is to develop a comprehensive architecture that incorporates semantic and contextual analysis of literary works, leveraging the capabilities of the Transformer, and Kuzushiji character recognition in images represents the first stage of this project.

The object detection mechanism in DETR suggests a promising research direction for utilizing the Transformer to achieve our goal. The point is, that is to say, DETR can simultaneously detect bounding boxes and categorize their contents using the Transformer's two-dimensional attention mechanism. In contrast to DETR, however, our task involves analyzing a one-dimensional sequence of characters in a vertical direction, following the reading order of Japanese texts. Furthermore, we do not require bipartite matching between predictive results and ground truth for loss function calculation, as the correct answers are sequentially aligned with the predictions. This sequential alignment simplifies our architecture compared to DETR.

The schematic flow of the proposed Kuzushiji recognition process is abstractly illustrated in Fig. 2. The input image is vertically long rectangle with a width of 64 pixels. This rectangle region, which surrounds a sequence of characters, is extracted from the original image and resized to a width of 64 pixels. The height of the rectangle depends on the length of the character sequence. A sample image labeled '1x200x64' in Fig. 2 represents one channel with 200 rows and 64 columns. The rectangular image is input to the CNN frontend, detailed in Section III-B, which generates an image feature map with 64 channels. In Fig. 2, the width of the feature map is downscaled to 32-pixels as a sample value, although this can vary depending on the adopted CNN architecture. Subsequently the feature map is sliced by rows. The data length of each row is the product of the image width and the number of channels generated by the CNN. In the case illustrated in Fig. 2, there are 100 rows each containing 2048 nodes. These sliced data, each with 2048 nodes per row, are input to the 6-layer Transformer encoder with d_model=2048 and nhead=8 as a sequence of data. The output of the Transformer retains the same sequence length. Each output in the sequence is directed to three fan-out layers, which calculate the corresponding character code, the indicator for the top position of the character, and the positional occupations of the character's bounding box. The character code is represented as a one-hot vector, with the final value determined by the position of the maximum value via the softmax function. The top position, marked as '*' in Fig. 2, serves as a border signal between the connected bounding boxes of characters. The positional occupations are represented in an array with the same size as the width of the image feature map, indicating which positions are part of character's bounding box. To simplify the illustration, Fig. 2 illustrates an output with a smaller number of occupied boxes and rows. However, in this example, they are actually 32 occupied positions across 100 rows, represented as a rectangle table of '#' marks and spaces. The output data are compared with the ground truth data during training, and three cross entropy loss functions are calculated for the predicted character code $L_{\text{map-code}}$, the predicted top position L_{top} and the predicted occupied positions in a bounding box L_{box}.

$$Loss_{\text{map}} = L_{\text{map-code}} + L_{\text{top}} + L_{\text{box}} \tag{1}$$

In this flow, each row only contains partial and sliced features of a character, but the adjacent features are mutually referred in the attention matrix of the 6-layer Transformer. These mutual references help in identifying a correct character even at the top of the rows where the character starts. The top position indicator functions to separate two connecting characters, as the positional occupations for one bounding box may sometimes connect to the next bounding box.

B. CNN Frontend

The CNN frontend plays a crucial role in extracting local features from the image and can be implemented in various forms, each affecting system performance. We examined several variations and chose a simplified U-Net like architecture with skip connections[9], as shown in Fig. 3. In this architecture, R-Block represents a CNN block with a residual connection[12]. The input image with a width of 64 pixels is downsampled to the feature map with a width of 4 pixels at the bottom, and then upsampled to a width of 32 pixels. At the bottom of this U-Net architecture, local patterns around

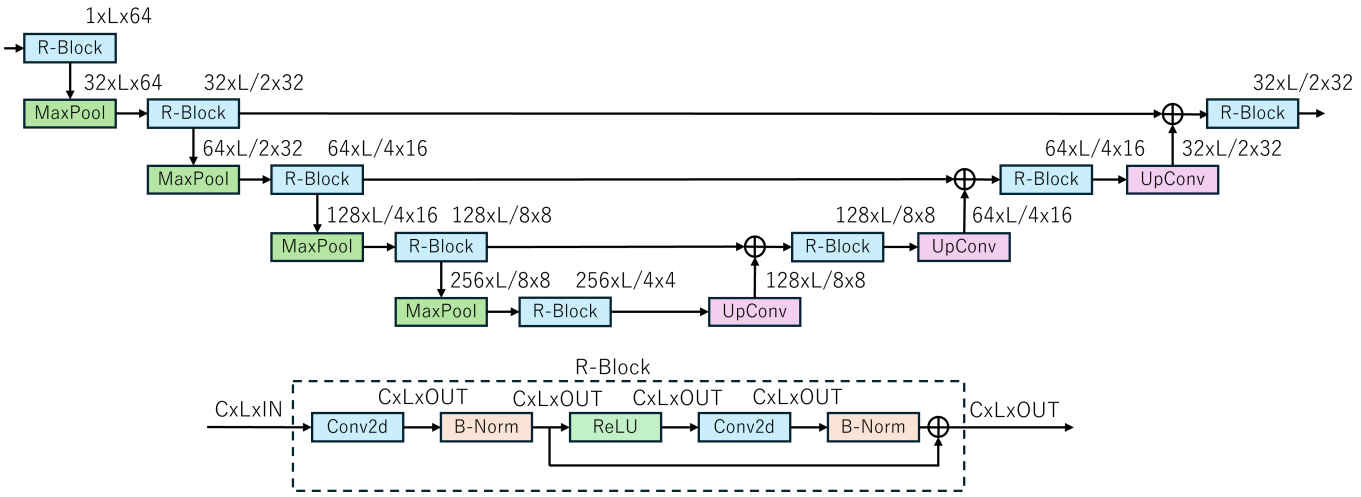
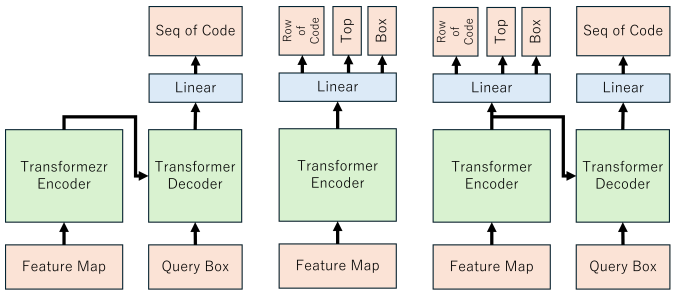


Fig. 3. CNN frontend.



(a) First Design (b) Second Design (c) Third Design

Fig. 4. Three Transformer models.

a character could be aggregated to form the entire pattern of the character. The upsampling step is halted before reaching the original input image resolution to provide rich contextual information to the single row data, allowing the Transformer to more easily capture the content. In Fig. 3, the CNN frontend stops upsampling one level below the original resolution. However, we also tested other deviations, such as stopping at the original or at two levels below the original, described in Section IV.

C. Transformer Models

The input sequence data to the Transformer has a length corresponding to the height of a vertically long region that surrounds a sequence of characters, and the final output has a length corresponding to the identified character sequence, which is different from the input length. The initial system design, as shown in Fig. 4(a), therefore adopted the Transformer encoder-decoder architecture[7] similar to the DETR. The inputs to the decoder are query boxes whose contents are zero but added with positional encodings. However, this design failed to converge to a stable solution. This failure could be attributed to the significant representational distance between the input and the output, making it difficult to find a correspondence. Through experiments of the first design, we redrew it into the second design, as shown in Fig. 4(b). This new design utilized only the Transformer encoder, where the output directly corresponded to the input of the sliced feature map, as illustrated in Fig. 2. This design change significantly improved convergence to a solution. However, the output consists of narrowly sliced pieces of a character sequence image, appearing as an intermediary result of character recognition. Therefore, the second design requires another program to generate the final character sequence from this intermediary result. The third design, as shown in Fig. 4(c), reintroduces the Transformer encoder-decoder architecture to generate the final output of the character sequence. The inputs to the decoder are the same as those in the first design, where the query-box contents are zero with positional encodings. The key difference from the first design is that both of the encoder and decoder produce outputs and the loss function $Loss_{\text{sequence}}$ accounts for all output errors, including the error in the predicted sequence of character codes $L_{\text{seq-code}}$.

$$Loss_{\text{sequence}} = L_{\text{map-code}} + L_{\text{top}} + L_{\text{box}} + L_{\text{seq-code}} \quad (2)$$

While learning, image features of Kuzushiji sequence are input to the encoder, the grand truth of the Kuzushiji code maps corresponding to the input data shape are compared with the output from the encoder, and that of the final character sequences are compared with the output from the decoder. The third design successfully found a stable solution after loss convergence and produced a sequence of characters. This approach can be extended to sentence recognition, making it the most suitable for our ultimate goal.

IV. Experiments

A. Dataset

We evaluated the proposed methods using the Kuzushiji dataset from the Kaggle competition[8]. The dataset comprises 3,605 page images from classical literary books, and includes 4,212 classes of characters. We used 3,000 images for training,

TABLE I
FEATURES OF KAGGLE DATASET.

	Total size	For training	For testing
images	3,605	3,000	500
lines	37,680	32,502	4,097
characters	683,464	579,904	83,528

TABLE II
RESULTS OF KUZUSHIJI RECOGNITION IN BOUNDING-BOX.

Image size	Dropout	Accuracy
128x128	-	0.925
64x64	-	0.920
32x32	-	0.912
128x128	0.5	0.915
64x64	0.5	0.914
32x32	0.5	0.909

and 500 images for testing. Each page typically contains around ten lines of characters, although some pages feature drawings with short sentences. The features of the Kaggle dataset are detailed in Table I.

B. Kuzushiji Recognition in Bounding Box

The first experiment measured the feasibility of recognizing Kuzushiji characters within the given bounding boxes. We used a CNN depicted in a left half of Fig. 3, combined with a fully connected layer and softmax function to identify the codes of Kuzushiji characters. The optimizer was Adam with a learning rate of 0.0001, and the entire training data were learned over 10 epochs. It is important to note that more than a thousand characters appeared only once or twice in the dataset, giving them only 10 or 20 opportunities to be learned. In every epoch, data augmentation was performed using translation and scaling. The accuracy for 64x64 pixel images was 0.920, and the best accuracy reaches 0.925 for 128x128 pixel images, as shown in Table II. Although the 128x128 pixel images achieved the highest score, the training time was significantly longer. Therefore, we decided the width of the vertically long rectangle for learning a sequence of characters should be 64 pixels, balancing performance and computation cost.

C. Consective Kuzushiji Recognition in Rectangular Region

The second experiment measured the performance of recognizing a sequence of Kuzushiji characters in a vertically long rectangular region. We prepared two Transformer models with three CNN implementations. The CNN architectures are illustrated in Fig. 3, where the image with a width of 64 pixels is downsampled to 64, 32 and 16 pixels. In the 32 pixel case, one output node corresponds to 2x2 nodes in the input, and in the 16 pixel case, one output node corresponds to 4x4 nodes in the input. This experiment aimed to confirm the adequate levels of downsampling, and evaluate the performance of the recognition. The optimizer was Adam with a learning rate of 0.00001, and the entire training data were learned over 400 epochs. Training a single model required more than three days using an H100 GPU. Data augmentation, including translation and scaling was performed in every epoch.

TABLE III
RESULTS OF KUZUSHIJI RECOGNITION.

System	Width	Dropout	Precision	Recall	F1 score
Second	16	-	0.862	0.854	0.858
Second	32	-	0.896	0.900	0.898
Second	64	-	0.836	0.861	0.843
Second	32	0.5	0.867	0.867	0.867
Second	64	0.5	0.835	0.848	0.841
Third	-	0.5	0.865	0.864	0.865
Kuronet*	-	-	0.889	0.902	0.896

* Kuronet was evaluated under the different conditions[1].

Table III presents the results, where 'Second' refers to the second design of the Transformer model, and 'Third' refers to the third design. 'Kuronet' is included as a reference, but the experimental conditions differ from those of the other methods. Notably, the second design with a width of 32 pixels without dropout achieved the best F1 score of 0.898, although the score was lower than 0.925 reported in Table II. The decrease in the score might be attributed to the difficulty in determining the bounding boxes. The third design achieved an F1 score of 0.865, which is lower than that of the second design. The cause of this decrease has not yet been determined, but it is likely related to the loss function, as the loss values for $L_{\text{map-code}}$ were significantly larger in the third design compared to the second design during training.

The scores of our proposed methods are comparable to those of the Kuronet, despite our systems being preliminary implementation of our idea. Since Kuronet achieved its performance using regularization algorithms for the input images, we should also explore similar tuning methods to enhance our performance.

V. DISCUSSION

We designed a Kuzushiji recognition architecture that leverages the Transformer to calculate attention among a sequence of sliced images along a vertical direction. This architecture draws inspiration from DETR, which utilizes the Transformer for object detection by recognizing two-dimensional relationships. In contrast, our architecture uses the Transformer to recognize one-dimensional relationships, as a sentence or sequences of characters are arranged in a linear order. The recognition accuracy of our architecture has reached the same level as the leading technology, Kuronet. A notable advantage of our approach is its extendability to understand the contextual meanings of sentences, which in turn could provide feedback to enhance character recognition. This research is in its early stage, and we will continue to explore the potential of our architecture, aiming to contribute well to the recognition of Japanese classical literature.

Through our experiments, we encountered several difficulties in dealing with Kuzushiji characters. The most critical one is the long tail of the character-class distribution. While object detection or classification typically deals with ten to a hundred classes of objects, the number of Kuzushiji character classes exceeds several thousand. This number will only increase as we apply our system to more documents and books. We



Fig. 5. Variations of 'FUJI' character.

may require an alternative architecture to handle the vast number of character classes. Additionally, the long tail of the character occurrence distribution includes many character classes that are rarely found in documents, making it difficult to improve the accuracy rate. Rare characters need more learning opportunities, but their infrequent occurrence in documents prevent to improve accuracy. Overcoming this contradiction is essential for enhancing the system's performance.

Another difficulty we encountered was the wide variation of cursive styles for Kanji characters. Upon examining the list of false answers in the Kuzushiji recognition results, we found that the Kanji character 'FUJI' was often misinterpreted as different characters. Fig. 5 shows three images of the same 'FUJI' character. This character consists of 19 strokes and includes many horizontal lines, making it difficult to write strokes in detail accurately. To make matters worse, this character 'FUJI' is a part of famous family names, leading to frequent and inevitable occurrences of this character in texts. As a result, simplified versions of Kuzushiji characters have been invented and are frequently used in documents. This variation makes it difficult to learn such a single character with multiple different representations.

VI. Conclusions

This paper proposes a Kuzushiji recognition architecture using the Transformer. The architecture processes an image of a vertically long rectangle that surrounds a sequence of characters as input. The input data are first analyzed by a CNN frontend and then sliced by rows. These sliced data are then sent to the Transformer as a sequence of data, and the Transformer calculates relationships among the sliced data to identify the correct characters and their bounding boxes in the output. The experimental results demonstrated that the proposed architecture effectively identifies characters within a rectangular region. The project's ultimate goal is for the architecture not only to identify characters but also to understand the context of documents in classical literary works. This contextual understanding will provide feedback to the character recognition process, thereby enhancing the accuracy of character identification. This paper serves as a preliminary report on our project. We aim to address the identified weaknesses in this paper. We continue to improve them to contribute to the research on Japanese classical literature.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP20K11777. This work was conducted as a "collaborative study under the NIJL Project".

References

[1] A. Lamb, T. Clanuwat, and A. Kitamoto, "KuroNet: Regularized residual U-Nets for end-to-end kuzushiji character recognition," *SN COMPUT.SCI*, vol. 1, no. 177, 2020.

[2] K. Ueki and T. Kojima, "Survey on deep learning-based kuzushiji recognition," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII*, Springer, 2021, pp. 97–111.

[3] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical Japanese literature," *Neural information processing systems (Neurips) creativity workshop*, 2018.

[4] H. T. Nguyen, N. T. Ly, K. C. Nguyen, C. T. Nguyen, and M. Nakagawa, "Attempts to recognize anomalously deformed kana in Japanese historical documents," in *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, 2017, pp. 31–36.

[5] A. D. Le, T. Clanuwat, and A. Kitamoto, "A human-inspired recognition system for pre-modern japanese historical documents," *IEEE Access*, vol. 7, pp. 84 163–84 169, 2019. DOI: 10.1109/ACCESS.2019.2924449.

[6] X. Hu, M. Inamoto, and A. Konagaya, "Recognition of kuzushi-ji with deep learning method a case study of kiritsubo chapter in the tale of genji," in *The 33rd Annual Conference of the Japanese Society for Artificial Intelligence*, 2019.

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] *Kuzushiji Recognition in Kaggle Competition*, https://www.kaggle.com/c/kuzushiji-recognition.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[10] D. Singh, C. V. Aravinda, M. Kaur, *et al.*, "Dknet: Deep kuzushiji characters recognition network," *IEEE Access*, vol. 10, pp. 75 872–75 883, 2022. DOI: 10.1109/ACCESS.2022.3191429.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with Transformers," in *Computer Vision – ECCV 2020*, Springer International Publishing, 2020, pp. 213–229, ISBN: 978-3-030-58452-8.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Inertial Strengthened CLIP model for Zero-shot Multimodal Egocentric Activity Recognition

Mingzhou He, Haojie Wang, Shuchang Zhou, Qingbo Wu*, King Ngi Ngan, Fanman Meng, Hongliang Li
School of Information and Communication Engineering,
University of Electronic Science and Technology of China, Chengdu 611731, China
* {Correspondence: qbwu@uestc.edu.cn}

***Abstract*—The popular CLIP model has empowered various zero-shot learning tasks by unifying them into the vision-language alignment framework. However, due to the dynamic subject-object interactions and complex motion variations, the egocentric videos become more diverse and widen the gap between the vision and language data, which limits the applicability of the CLIP model. Thanks to the widely used Inertial Measurement Unit (IMU) in wearable devices, this paper proposed an Inertial Strengthened CLIP (IS-CLIP) model to refine the visual representation of egocentric videos, which achieves highly effective zero-shot multimodal egocentric activity recognition. Our IS-CLIP is composed of two modules, i.e., the Subject-object Interaction Refinement Module (SIRM) and the Subject-motion Guided Aggregation Module (SGAM). On the one hand, SIRM embeds the motion patterns of IMU into the visual representation of each frame to enhance the action-related features. On the other hand, the SGAM maps the IMU to the frame-wise weights to aggregate the visual representations of all frames, which reduces the differences between the videos of the same action with different speeds, directions, and temporal locations. Experiments on the UESTC-MMEA-CL dataset show that the proposed IS-CLIP outperforms many state-of-the-art methods in the zero-shot multimodal egocentric activity recognition task.**

***Index Terms*—egocentric vision, multimodal activity recognition, IMU, zero-shot learning, CLIP**

I. Introduction

Egocentric activity recognition (EAR) has become a significant research direction, with the widespread application of wearable devices in fields such as health monitoring, human-robot interaction, etc [1], [2]. In recent years, numerous outstanding fully supervised deep learning methods for egocentric activity recognition have emerged [3]–[5], leveraging wearable device sensors such as cameras, inertial sensors, and optical flow. For example, Huang et al. [6] introduce the knowledge graph and present a knowledge-driven egocentric activity recognition framework. Hao et al. [7] extend a single inertial sensor to two-hand inertial sensors and propose a two-branch late-fusion framework. With the development of technology, practical application scenarios have proposed zero-shot demand, which requires that the model be trained on base categories while being capable of recognizing novel categories [8], to reduce the labor-intensive data acquisition and annotation. However, none of the previously mentioned methods possess zero-shot recognition capabilities.

Contrastive Language Image Pre-training (CLIP) [9], which is trained on hundreds of millions of image-caption pairs collected from the Internet, has good feature representation

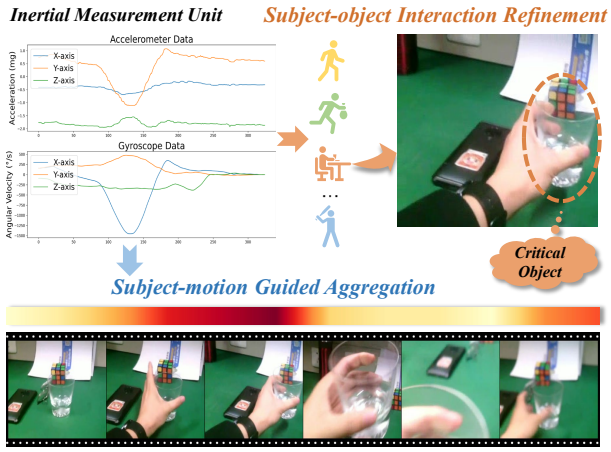


Fig. 1. The IMU provides supplementary information on subject motion, which is utilized to refine CLIP's visual representation. (a) Subject-object Interaction Refinement: In hand movements mode, we should focus on the cup in hand rather than the irrelevant phone. (b) Subject-motion Guided Aggregation: Utilize the IMU's temporal features to emphasize video frames with significant motion.

for images and texts. There are many CLIP-based image recognition works [10]–[12] that show amazing zero-shot performance. Recently, some research has applied CLIP to video input and achieved zero-shot third-person video activity recognition. ActionCLIP [13] proposed the "pre-train, adapt, and fine-tune" paradigm for activity recognition and expanded text labels to enhance the matching between video and text. OpenVCLIP [14] models spatial-temporal relationships in videos and proposes the interpolated weight optimization method to maintain the generalization of CLIP after training. FROSTER [15] adopts a residual feature distillation approach to ensure that CLIP retains its generalization while effectively adapting to activity recognition.

However, zero-shot methods based on videos typically encounter two significant challenges when applied to egocentric activity recognition. First, egocentric videos capture complex external environments rather than the action subject. Numerous irrelevant objects appear in the video, severely impacting CLIP's visual representation. Second, changes in first-person video may reflect environmental object movement rather than the subject's motion, causing significant interference with activity recognition. Inertial sensors on wearable devices provide additional motion data of the action's subject, offering crucial

compensation to mitigate these interferences and achieve more accurate visual-language alignment. To our knowledge, there is currently a lack of research on integrating IMU data into zero-shot egocentric activity recognition based on CLIP.

Therefore, we introduce the IMU modality and utilize IMU's subject motion information to strengthen the CLIP's visual representation. To this end, we design the Subject-object Interaction Refinement Module (SIRM) and Subject-motion Guided Aggregation Module (SGAM). Specifically, egocentric videos usually contain multiple background objects, most of which are irrelevant to the activity. IMU time-frequency features correspond to typical motion mode, such as hand and head movements. The SIRM utilizes IMU time-frequency features as query vectors to guide the focus on objects related to behavior patterns. On the other hand, the SGAM module leverages the temporal features of IMU to emphasize video frames with significant motion, reducing the interference from frames without subject motion and the differences between instances of the same action caused by the environment. In addition, our method employs feature distillation [16] to more effectively embed the knowledge of CLIP into the multimodal model.

Overall, we are the first to propose a zero-shot multimodal (video and IMU) egocentric activity recognition framework based on CLIP, called IS-CLIP. We introduce IMU data to refine the CLIP visual representation of video to enhance the zero-shot performance. Our method achieves state-of-the-art performance in two zero-shot settings on the UESTC-MMEA-CL [17] dataset, establishing an exemplary benchmark for zero-shot multimodal egocentric activity recognition.

II. Method

A. Preliminaries

First, we introduce the procedure of zero-shot recognition based on CLIP and apply it to a single video modality. Zero-shot recognition aims to classify the unseen category set $\mathcal{C}_{\mathcal{N}}$ through the classification network trained on the seen category image set $\mathcal{C}_{\mathcal{B}}$, where $\mathcal{C}_{\mathcal{B}} \cap \mathcal{C}_{\mathcal{N}} = \emptyset$. In this context, the traditional classification head relying on fully connected layers loses its effectiveness. Thanks to CLIP's rich image-text knowledge and image-text matching paradigm [9], CLIP is often used as a classifier for various tasks to easily achieve zero-shot capabilities [18]–[20]. Specifically, an image $\mathcal{I}_i$ is passed through the visual encoder $\mathcal{V}(\cdot)$ to obtain a visual embedding v_i. On the other side, constructed text inputs like "a photo of a $[Class]$" are fed into the language encoder $\mathcal{L}(\cdot)$ to obtain text embeddings $\{t_1, t_2, ...t_n\}$, where n is the number of classes. Then, the cosine similarity between the visual embedding v_i and each text embedding $t_j, j \in [1, n]$ is calculated by:

$$cos(v_i, t_j) = \frac{v_i \cdot t_j^T}{\|v_i\| \cdot \|t_j\|}, j \in [1, n] \tag{1}$$

and the category c_i of the image $\mathcal{I}_i$ is defined as:

$$c_i = \underset{j \in [1,n]}{\arg\max} \cos(v_i, t_j) \tag{2}$$

In this way, zero-shot recognition can be achieved by just appending semantic inputs of novel categories. For video input, the method of equally spaced frame extraction is usually used, and the input is actually a set of time-ordered images $\{\mathcal{I}_{t_1}, \mathcal{I}_{t_2}, ...\mathcal{I}_{t_N}\}$. All of them passed through the visual encoder to obtain visual embeddings $\{v_{t_1}, v_{t_2}, ...v_{t_N}\}$. These embeddings are then averaged into one visual embedding and matched with the semantic embeddings for classification according to Eq. (1)(2). This CLIP-based single video modality activity recognition method should be considered as our baseline.

B. Framework Pipeline

The overall framework of zero-shot multimodal egocentric activity recognition is illustrated in Fig. 2. The input of the model is N frames extracted from the video and IMU data of the same time period. IMU data includes accelerometer and gyroscope, which are one-dimensional time series data. In order to better extract the time domain and frequency domain features of IMU data, we performed the short-time Fourier transform (STFT) [21] on the IMU data. The STFT divides a long-time signal into several short periods and then applies the Fourier transform to each period. In this way, the spectrum of the signal in each time period can be obtained, thereby describing the time-frequency characteristics. The formula for STFT is defined as:

$$\mathrm{STFT}(x(t))(\tau, \omega) = \int_{-\infty}^{\infty} x(t)\, w(t-\tau)\, e^{-j\omega t}\, dt \tag{3}$$

where $x(t)$ is the time signal, $w(t-\tau)$ is the window function, τ is the time offset, and ω is the frequency.

In the visual branch, we built two identical ViT-B/16 visual encoders and initialized them with CLIP pre-trained weights. In order to maintain the generalization of CLIP, we adopt knowledge distillation. One of the encoders is regarded as the teacher model, and its parameters are completely frozen. The other one is the student model whose last two blocks in the transformer and the last fully connected layer are fine-tuned during training. N frames are respectively passed through the teacher encoder and the student encoder to obtain visual features f_v^t and $f_v \in \mathbb{R}^{N \times 512}$, and feature distillation is performed between them. The distillation loss $\mathcal{L}_{kd}$ is defined as:

$$\mathcal{L}_{kd} = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{f_{vi}^t \cdot f_{vi}}{\|f_{vi}^t\| \|f_{vi}\|} \tag{4}$$

In the IMU branch, the accelerometer and gyroscope spectrograms obtained by STFT are sent to the IMU encoder to extract their respective time-frequency features. Note that the two sensors share one encoder to reduce the number of parameters. The two sensor features are then concatenated and reduced dimension through a linear layer to obtain the IMU feature $f_{IMU} \in \mathbb{R}^{1 \times 512}$.

Next, we need to consider how to integrate visual and IMU features. The presence of motion-irrelevant objects in first-person videos and pseudo-motion caused by these objects can

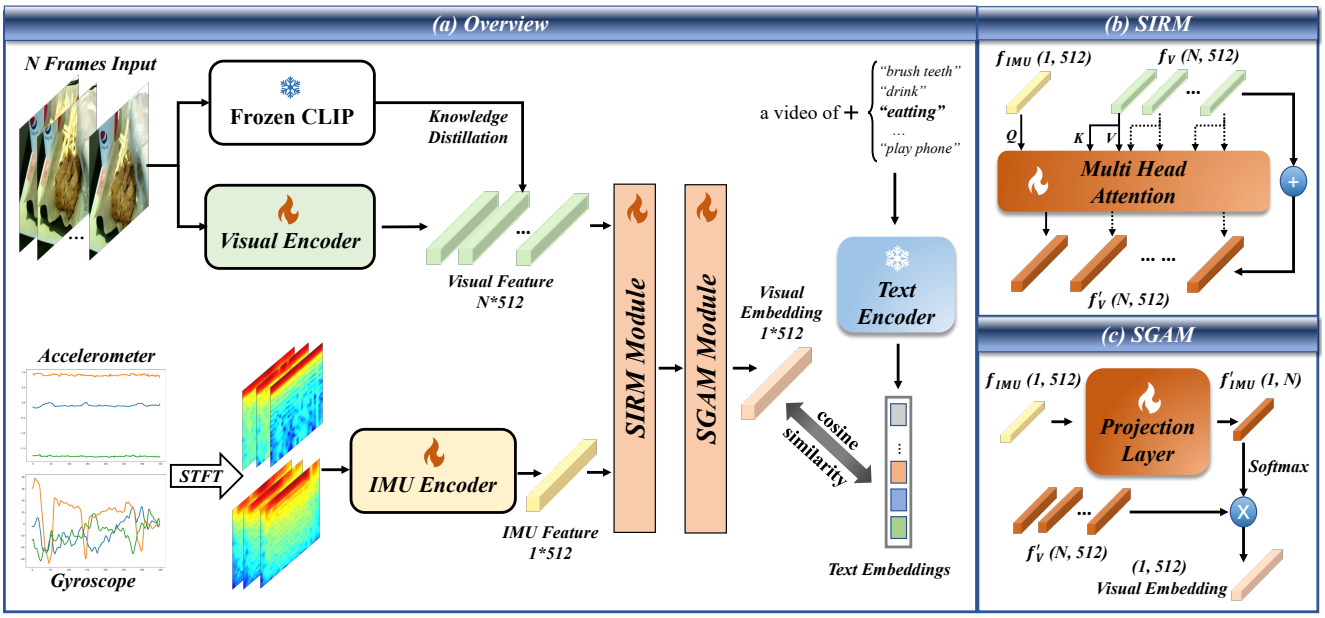


Fig. 2. (a) An overview of our proposed zero-shot multimodal egocentric activity recognition framework. The input is temporally ordered frames and IMU data (accelerometer and gyroscope). Among them, STFT stands for short-time Fourier transform, (b) SIRM: Subject-object Interaction Refinement Module. (c) SGAM: Subject-motion Guided Aggregation Module. Note that N in the figure represents N frames of an instance rather than the batch size.

interfere with CLIP's visual representation. The supplementary information on subject motion provided by IMU data can mitigate these interferences. Meanwhile, CLIP has never encountered IMU data, directly fusing IMU and visual features would harm generalization. For these reasons, we design the SIRM and SGAM modules, which use IMU features f_{IMU} to refine visual representation capabilities, resulting in enhanced visual embeddings $v \in \mathbb{R}^{1\times512}$.

Finally, the visual embedding is matched with the text embedding for category prediction, as Eq. (1)(2). On the text processing side, it is exactly the same as vanilla CLIP, with all parameters frozen. The overall loss $\mathcal{L}$ of the framework is defined as:

$$\mathcal{L} = \mathcal{L}_{CrossEntropy} + \mathcal{L}_{kd} \tag{5}$$

C. Subject-object Interaction Refinement Module (SIRM)

In egocentric videos, there are usually many objects that are irrelevant to the behavior, which are mixed into the visual features to affect the image-text matching of CLIP. The main advantage of IMU data is that it can provide subject's motion information, which can reflect the basic motion mode, and we should focus on the objects related to motion in vision. Inspired by this, we present the Subject-object Interaction Refinement Module (SIRM). We use the IMU feature $f_{IMU} \in \mathbb{R}^{1\times512}$ as query vectors and visual features $f_v \in \mathbb{R}^{N\times512}$ as both 'key' and 'Value' for multi-head attention calculation. Note that f_{IMU} performs attention calculations with each of the N dimensions in f_v. The calculation process is as:

$$Q_i = f_{IMU}W_{Q_i},\ K_i = f_vW_{Ki},\ V_i = f_vW_{Vi} \tag{6}$$

$$head_i = \text{softmax}\left(\frac{Q_iK_i^T}{\sqrt{d_k}}\right)V_i \tag{7}$$

$$f_v^{'} = \text{Concat}(\text{head}_i)W_O,\ i \in [1,8] \tag{8}$$

Among them, W_Q, W_K, W_V, W_O are weight matrices, and i is the number of heads. In this way, the interaction between IMU and visual features generates the weighted mask, which is applied to enhance key object attention to refine the visual features. Furthermore, we also add residual connections to the visual features before and after refinement, thereby preserving the generalization capabilities of CLIP.

D. Subject-motion Guided Aggregation Module (SGAM)

Among the N frames extracted at equal intervals from the video, not all frames are in the period of subject movement. Furthermore, changes between frames may be caused by the motion of unrelated objects in the video rather than the movement of the subject. The video-based single-modal method cannot effectively identify critical frames, and directly averaging visual features will dilute essential information. The IMU's temporal features can accurately reflect the strength of the subject's movement. Therefore, we propose the Subject-motion Guided Aggregation Module (SGAM), as shown in Fig. 2(c). We map the IMU features $f_{IMU} \in \mathbb{R}^{1\times512}$ to N dimensions in the time series through a projection layer (MLP) and apply softmax to obtain the temporal weighted mask. Then, weighted fusion is performed on the N temporally ordered visual features $f_v^{'} \in \mathbb{R}^{N\times512}$ so that the final visual embedding v contains more information from periods of the subject's intense activity.

III. Experiments

We mainly report the results of our method on zero-shot egocentric activity recognition and compare them with other

TABLE I
ZERO-SHOT EGOCENTRIC ACTIVITY RECOGNITION RESULTS UNDER TWO EXPERIMENT SETTINGS. THE FINE-TUNED CLIP IS OUR BASELINE, WHICH USES KNOWLEDGE DISTILLATION. HM IS THE HARMONIC MEAN.

Method	Publication	Prompt Augment	Base(16) Novel(16)			Base(24) Novel(8)		
			Acc^{novel}	Acc^{base}	HM	Acc^{novel}	Acc^{base}	HM
ActionCLIP [13]	Arxiv2021	✓	35.0	99.1	51.7	44.7	98.8	61.5
XCLIP [22]	ECCV2022	×	41.6	99.7	58.7	54.6	98.8	70.4
VPT [23]	ECCV2022	×	31.1	94.6	46.8	39.1	94.5	55.3
Text4Vis [24]	AAAI2023	×	43.9	98.8	60.8	55.2	98.2	70.7
ViFi-CLIP [25]	CVPR2023	×	55.3	98.2	70.7	63.7	96.9	76.9
Open-VCLIP [14]	ICML2023	✓	63.9	96.6	76.9	75.9	95.3	84.5
FROSTER [15]	ICLR2024	✓	64.0	97.0	77.1	71.4	97.1	82.3
CLIP [9]	ICML2021	×	58.5	55.2	56.8	68.0	44.5	53.8
Fine-tuned CLIP	-	×	60.5	97.6	74.7	74.2	96.6	83.9
IS-CLIP(ours)	-	×	**66.3**	97.7	**79.0**	**77.8**	97.2	**86.4**

state-of-the-art CLIP-based activity recognition methods. Further, we conduct an in-depth analysis of the proposed modules' effectiveness. Finally, we conduct ablation experiments on both modalities and method components.

Dataset and metric. We conduct experiments on the UESTC-MMEA-CL [17] dataset, which is a multimodal dataset of first-person perspective activity recognition acquired by smart glasses. Alongside first-person perspective video from the camera, it also includes data streams from acceleration and gyroscope sensors. The dataset consists of 32 categories of human daily activities, with 4,553 samples in the training set and 1,316 samples in the test set. We adopt two zero-shot settings on this dataset, i.e., the first 16 classes as base classes and the first 24 classes as base classes, separately. Since we focus on zero-shot performance, the critical evaluation metric is the accuracy of novel classes.

Implementation details. Our model adpot ViT-B/16 as the visual encoder and initialize with CLIP pretrained weights. We train the model on base catrgories for 50 epochs on an RTX4090 GPU, and the batch size is 16. We utilize different optimizers and learning rates for various model parameters. For the visual encoder of CLIP, we employ the Adam optimizer with a learning rate of 1e-5. SGD is the optimizer for other parameters with a learning rate of 1e-3.

A. Zero-shot Egocentric Activity Recognition

To validate the superiority of our multimodal framework, comparative experiments are conducted under two different zero-shot settings. The experimental results are presented in Table I. All comparison methods use only the video modality as input, as they inherently lack the capability to adapt to inertial measurement unit (IMU) data input. These methods are all based on CLIP, thus possessing certain zero-shot recognition capabilities. Therefore, comparing our method with these advanced methods is reasonable. We adopt the same fine-tuned method as described in this paper, including knowledge distillation, to obtain a robust baseline, which is referred to as Fine-tuned CLIP in Table I.

The experimental results show that our method outperforms the baseline by 5.8% and 3.6% in novel categories' accuracy (Acc) under the two settings, respectively. This demonstrates that our proposed modules effectively adapt to IMU input and leverage the advantages of IMU data to enhance zero-shot detection performance. Compared to the second-best method, our novel categories' Acc is higher by 2.3% and 1.9%, achieving state-of-the-art performance. Notably, we don't use prompt augmentation as some other methods do. This further showcases the superiority of the multimodal approach. Additionally, we also achieved the best performance in the harmonic mean metric, indicating that our multimodal method improves the overall network performance rather than boosting zero-shot recognition at the expense of base class performance.

B. Analysis of Visual Feature

Our goal is to embed CLIP knowledge into a multimodal framework based on video and IMU data, to achieve zero-shot egocentric activity recognition that is superior to single video modality. To this end, we design the SIRM and SGAM modules to adapt to the input of IMU data and enhance zero-shot capabilities. To verify the effectiveness of our method, two aspects should be considered: (1) whether the two modules enhance the representation of visual features. (2) whether the addition of IMU modality enhances or impairs CLIP's image-text matching ability. In this subsection, we examine the effect of our method at the feature level.

Firstly, we project the visual embeddings of our method and the baseline into 2D space using the t-SNE algorithm to observe their distribution. As shown in Fig. 3, the visual embeddings of our method are significantly more separable after the two modules guided by IMU data. Among them, the distance between classes is increased dramatically, and the distribution within the class is more compact. More importantly, the problematic categories that were initially mixed are also successfully distinguished. These results prove that our

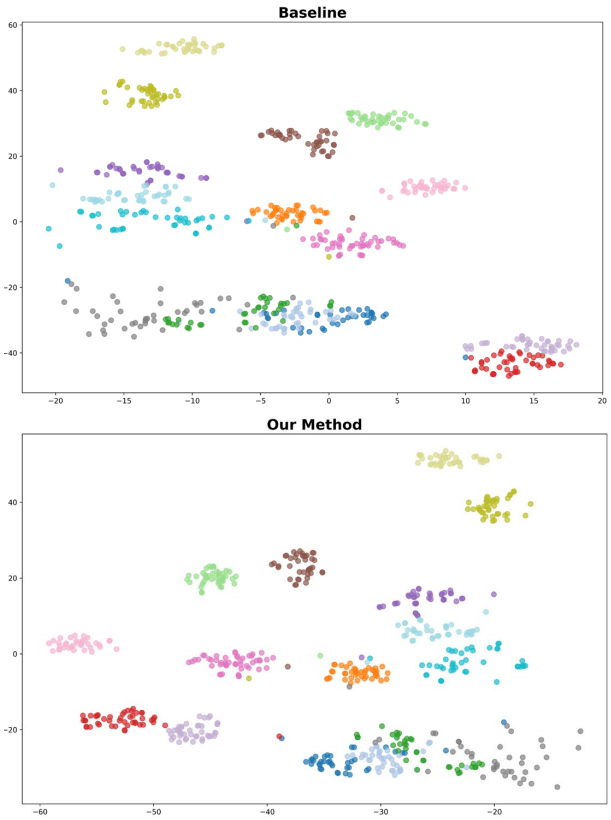


Fig. 3. t-SNE plots of our method and baselines' visual embeddings on novel categories. Our method shows better separability.

method exactly enhances the visual representation. In addition, we count the cosine similarity of each novel category instance's visual embeddings and their corresponding text embeddings, and show the mean of category dimension. As shown in Fig. 4, the results show that the similarity between visual and semantic embedding has increased in almost all categories. This verifies that the integration of IMU data does not destroy the knowledge of CLIP, and the ability of zero-shot recognition is improved as expected after being refined by our method.

C. Ablation Study

We conduct ablation experiments in (Base:16, Novel:16) zero-shot setting of UESTC-MMEA-CL dataset. Table II shows that SIRM and SGAM modules can effectively improve the accuracy of novel categories. Among them, the improvement brought by adding SGAM is better, and the improvement in the base categories is also more apparent. This is because the recognition method based only on video frame extraction naturally lacks temporal information regardless of whether it is zero-shot. However, temporal features can be easily obtained through IMU data. The employment of SIRM will slightly reduce the performance of base categories, which may be because the feature representation learned based on full supervision is more accurate than the refinement guided by IMU.

Furthermore, we also conduct ablation experiments on accelerometer and gyroscope modalities, as shown in Table III.

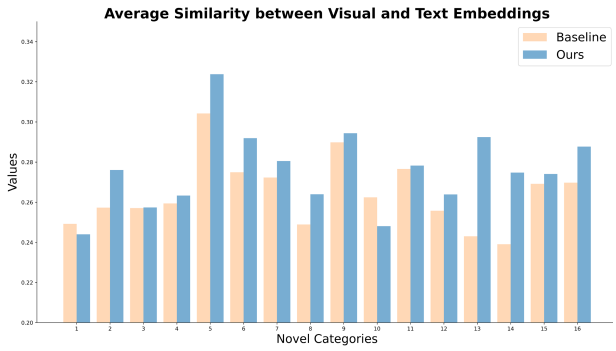


Fig. 4. Comparison of the average cosine similarities of visual and text embeddings on novel categories.

The results show that the integration of both modalities can effectively improve the performance of the novel categories, which further proves the superiority of our proposed multimodal framework.

TABLE II
ABLATION STUDY OF PROPOSED MODULES.

SIRM	SGAM	*Base*	*Novel*	*HM*
		97.6	60.5	74.7
✓		97.1	63.9	77.0
	✓	**98.6**	64.6	78.1
✓	✓	97.7	**66.3**	**79.0**

TABLE III
ABLATION STUDY OF ACCELEROMETER AND GYROSCOPE MODALITIES.

Accelerometer	Gyroscope	*Base*	*Novel*	*HM*
		97.6	60.5	74.7
✓		97.7	64.1	77.4
	✓	**98.3**	65.3	78.5
✓	✓	97.7	**66.3**	**79.0**

IV. Conclusions

In this paper, we pioneeringly proposed a multimodal zero-shot egocentric activity recognition framework based on video and Inertial Measurement Unit (IMU). We embed the knowledge of CLIP into this framework to achieve zero-shot recognition capabilities and design two modules to adapt IMU input to enhance zero-shot performance. Our method achieves state-of-the-art zero-shot egocentric activity recognition performance, to establish a benchmark for subsequent research based on video and IMU in this field.

Acknowledgment

This work was supported in part by the National Science and Technology Major Project (No. 2021ZD0112001), the National Natural Science Foundation of China (No. U23A20286), the Independent Research Project of Civil Aviation Flight Technology and Flight Safety Key Laboratory (FZ2022ZZ06), and the Natural Science Foundation of Sichuan Province (2023NSFSC1972).

References

[1] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106 970, 2021.

[2] S. Singh, C. Arora, and C. Jawahar, "First person action recognition using deep learned descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2620–2628.

[3] X. Wang, Y. Wu, L. Zhu, and Y. Yang, "Symbiotic attention with privileged information for egocentric action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 12 249–12 256.

[4] M. Lu, Z.-N. Li, Y. Wang, and G. Pan, "Deep attention network for egocentric action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3703–3713, 2019.

[5] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei, "Jointly learning energy expenditures and activities using egocentric multimodal signals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1868–1877.

[6] Y. Huang, X. Yang, J. Gao, J. Sang, and C. Xu, "Knowledge-driven egocentric multimodal activity recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 4, pp. 1–133, 2020.

[7] Y. Hao, A. Kanezaki, I. Sato, R. Kawakami, and K. Shinoda, "Egocentric human activities recognition with multi-modal interaction sensing," *IEEE Sensors Journal*, 2024.

[8] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 384–400.

[9] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[10] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[11] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 113–19 122.

[12] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[13] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.

[14] Z. Weng, X. Yang, A. Li, Z. Wu, and Y.-G. Jiang, "Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization," in *International Conference on Machine Learning*, PMLR, 2023, pp. 36 978–36 989.

[15] X. Huang, H. Zhou, K. Yao, and K. Han, "Froster: Frozen clip is a strong teacher for open-vocabulary action recognition," *arXiv preprint arXiv:2402.03241*, 2024.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[17] L. Xu, Q. Wu, L. Pan, *et al.*, "Towards continual egocentric activity recognition: A multi-modal egocentric activity dataset for continual learning," *IEEE Transactions on Multimedia*, vol. 26, pp. 2430–2443, 2024. DOI: 10.1109/TMM.2023.3295899.

[18] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[19] Y. Zhong, J. Yang, P. Zhang, *et al.*, "Regionclip: Region-based language-image pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 793–16 803.

[20] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 084–14 093.

[21] S. Kumawat, M. Verma, Y. Nakashima, and S. Raman, "Depthwise spatio-temporal stft convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4839–4851, 2021.

[22] B. Ni, H. Peng, M. Chen, *et al.*, "Expanding language-image pretrained models for general video recognition," in *European Conference on Computer Vision*, Springer, 2022, pp. 1–18.

[23] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *European Conference on Computer Vision*, Springer, 2022, pp. 105–124.

[24] W. Wu, Z. Sun, and W. Ouyang, "Revisiting classifier: Transferring vision-language models for video recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 2847–2855.

[25] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan, "Fine-tuned clip models are efficient video learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6545–6554.

COIN-AT-PVAD: A Conditional Intermediate Attention PVAD

En-Lun Yu[†], Ruei-Xian Chang[‡], Jeih-weih Hung[§], Shih-Chieh Huang[¶] and Berlin Chen[‖]
[†‖] Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
E-mail: {enlunyu, berlin}@ntnu.edu.tw
[‡§] Department of Electrical Engineering, National Chi Nan University, Taiwan
E-mail: {s110352016, jwhung}@ncnu.edu.tw
[¶] Realtek Semiconductor Corp., Taiwan

***Abstract*—Personalized voice activity detection (PVAD), compared to conventional VAD, shows more developmental potential in scenarios with multiple speaker interference. Among the various methods for integrating speaker and acoustic features, performance may be limited due to the weaker representational capability of speaker embeddings derived from external speaker verification models. This study proposes a new architecture called Conditional Intermediate Attention PVAD (COIN-AT-PVAD) to address this issue. This architecture builds upon the Attentive Score (AS) module and incorporates the Feature-wise Linear Modulation (FiLM) scheme to better integrate multimodal information. Through comparing various fusion strategies, we show that COIN-AT-PVAD significantly surpasses the baseline model, especially when external embedding features have limited representational capacity. Experimental findings also indicate that, when compared to some state-of-the-art models, COIN-AT-PVAD achieves superior average precision and accuracy while retaining a compact model size, showcasing its efficacy in real-world applications on resource-limited devices.**

I. Introduction

Voice Activity Detection (VAD) [1]–[3] is specifically developed to detect and identify speech within received audio signals. It serves as the front-end component for a variety of speech processing applications, including automatic speech recognition (ASR) [4], keyword spotting (KWS) [5], and speech enhancement (SE) [6]. VAD reduces computational load by filtering away non-speech content before further processing, which is particularly important for high-performance systems with heavy computational requirements.

However, personalized speech applications in real-world situations often encounter voices from non-target speakers, which causes conventional VAD systems to generate disruptive false alarms, thereby impairing the performance of downstream tasks. To address this issue, recent research has developed a new variant of VAD known as Personal VAD (PVAD). In contrast to conventional VAD, PVAD is designed to recognize speech specific to a known speaker. This helps reduce interference from other speakers and mitigate false positives commonly seen in conventional VAD systems.

Ding *et al.* pioneered a straightforward PVAD approach in their seminal work [7]. This method involves generating speaker embeddings from enrollment speech via a speaker verification model, concatenating speaker embeddings with handcrafted acoustic features, and employing the concatenation as input features into the VAD framework. The authors further enhanced this system in PVAD 2.0 [8] by mapping speaker embeddings to the same domain as acoustic features and using a linear transformation to fuse the speaker information with acoustic features. The resulting compact input features simplify the subsequent processing stages and enhance PVAD performance.

Building upon the PVAD paradigm, which intertwines speaker information and acoustic data, subsequent studies have explored various methods to improve this integration [9]–[16]. A recent study [9] conducted a comprehensive comparison and analysis of a series of integration methods, including comparing the effects of static fusion at different points in the model pipeline and incorporating dynamic speaker estimation on both performance and complexity. Expanding on this foundation, [10] focused on integrating both pieces of information using Cross-Attention in the acoustic feature space. Furthermore, the work in [11] replaced the external speaker verification model with an internal embedding extractor and introduces an innovative Attentive Score (AS) loss function. This enables the attention score module to derive the attention weight from the concatenated features to draw attention to specific acoustic features.

Although [11] offers an intriguing solution for PVAD, there remains room for further validation and improvement. One of the concerns is that the AS module integrates acoustic and speaker information using a concatenation mechanism, which may be too simple and ineffective.

This study introduces a novel architecture dubbed Conditional Intermediate Attention PVAD (COIN-AT-PVAD) to address the above concern with the AS module. This architecture refines the AS module by conditionally fusing its multimodal inputs in order to improve its performance. In addition, we provide multiple options for locating the the presented novel AS module within the PVAD process. The experimental results indicate that the new COIN-AT-PVAD surpasses the baseline model and some state-of-the-art PVAD frameworks in performance while requiring fewer model parameters.

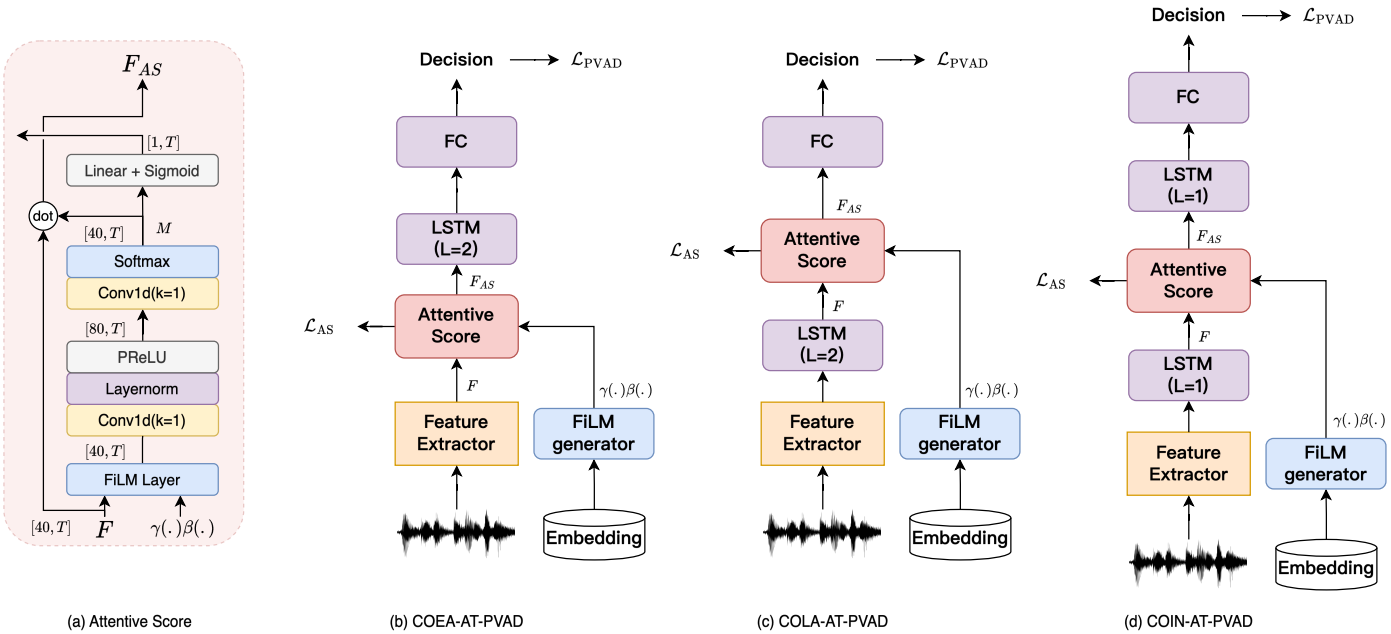


Fig. 1. Diagrams of (a) the advanced Attentive Score module; (b) Conditional Early Attention PVAD; (c) Conditional Latent Attention PVAD; (d) Conditional Intermediate PVAD.

II. Methods

A. Attentive PVAD with Conditioned Fusion

To begin with, we briefly review the AS module of [11]. Following that, we introduce a variant of the AS module that incorporates conditional fusion of input features.

Liu *et al.* proposed an AS module for PVAD that combines features from the target (enrollment) speaker and input utterances. This module extracts a similarity score between both features through two convolutional layers:

$$\hat{\mathbf{F}}_t = [\mathbf{F}_t; \mathbf{e}^{\text{target}}], \tag{1}$$

$$\mathbf{M}_t = \text{AS}(\hat{\mathbf{F}}_t), \tag{2}$$

where $\mathbf{F}_t$ and $\mathbf{e}^{\text{target}}$ represent the input acoustic feature at time frame t and target speaker embedding, respectively, and $\text{AS}(\cdot)$ denotes the overall function responsible for computing the similarity score $\mathbf{M}_t$ in the AS module.

Subsequently, by performing an element-wise multiplication of the similarity score $\mathbf{M}_t$ with $\mathbf{F}_t$, a weighted feature $\mathbf{F}_{\text{AS},t}$ is obtained:

$$\mathbf{F}_{\text{AS},t} = \hat{\mathbf{F}}_t \odot \mathbf{M}_t. \tag{3}$$

However, the use of concatenation as in (1) to combine speaker and acoustic information may have drawbacks: 1) It increases the feature dimensions, leading to a parameter-intensive model; 2) These two types of information are of different modalities, and thus concatenating them limits the model's learning capability.

Among fusion strategies, feature-wise Linear Modulation (FiLM) [17] has been demonstrated in multiple studies to effectively combine multimodal information and reduce the number of feature parameters [8]–[10]. Therefore, we propose to introduce FiLM into the AS module, as shown in Fig. 1(a), to obtain the conditioned feature $\tilde{\mathbf{F}}_t$, which replaces the concatenated feature $\hat{\mathbf{F}}$ in (1). The FiLM-wise conditioned feature $\tilde{\mathbf{F}}_t$ is obtained by:

$$\tilde{\mathbf{F}}_t = \text{FiLM}(\mathbf{F}_t) = \gamma(\mathbf{e}^{\text{target}}) \cdot \mathbf{F}_t + \beta(\mathbf{e}^{\text{target}}). \tag{4}$$

Here, γ and β represent the scaling vector and biasing vector, respectively, and they are produced by the FiLM generator using the target speaker embedding $\mathbf{e}^{\text{target}}$ as conditional information.

B. Fusion Strategies for Attention PVAD

According to subsection II-A, the advanced AS module learns the speaker-conditioned mask $\mathbf{M}_t$ and applies it to the input acoustic features $\mathbf{F}_t$. The location arrangement of this AS module in the PVAD pipeline inevitably influence the performance. Partially Inspired by the different fusion manners in [9], here we propose three variants of attentive PVAD (AT-PVAD), which mainly differ in the positioning of the advanced AS module inside the PVAD process, and their flowcharts are depicted in Figs. 1(b), 1(c), and 1(d). The underlying attention strategies are explained as follows:

- **Conditional Early Attention (COEA):** In this method, speaker and acoustic information are combined through conditional fusion at the early stage of the PVAD pipeline. The output of AS module is then used to train the subsequent classification module, which consists of a two-layer LSTM and two fully connected layers (FC). This method directly modulates the raw acoustic features and

learn a deeper classifier. Therefore, if the information from the acoustics and the speaker differs significantly, it may limit the fusion and lead to suboptimal results.

- **Conditional Latent Attention (COLA):** Similar to many other architectures that employ conditional fusion, COLA incorporates conditional fusion at a later stage in the model pipeline, specifically before the fully connected layers which serve as a simple classifier. This enables better training of the two-layer LSTM and the FiLM generator in front of the AC module to effectively integrate multimodal information across speaker and acoustics compared to COEA.
- **Conditional Intermediate Attention (COIN):** Here, the AS module is incorporated between two one-layer LSTMs. This configuration allows the front LSTM to capture improved acoustic features, while the back LSTM aids in the final classification. Clearly, COIN is intended to leverage the strengths of COEA and COLA in order to achieve superior results without increasing computational overhead.

C. Loss Function

As with other PVAD works, we employ the averaged binary cross entropy (BCE) [18] as one source of the objective function for the PVAD binary classification task, which is calculated as follows:

$$\mathcal{L}_{\text{PVAD}} = \frac{1}{T}\sum_{t=0}^{T-1} \text{BCE}(y_t, p_t). \tag{5}$$

Here, $y_t \in \{0, 1\}$ and $p_t \in [0, 1]$ respectively represent the ground-truth label and the PVAD model prediction outcome at time frame t, and T is the total number of time frames.

Additionally, based on [11], we introduce the AS loss function $\mathcal{L}_{\text{AS}}$ to specifically learn the AS module. The AS loss encourages the AS module to focus on learning the similarities across multimodal information by calculating the Mean Squared Error (MSE) loss between the ground-truth label y_t and the estimated weighted score $\tilde{m}_t$:

$$\tilde{m}_t = \text{Sigmoid}(\text{Linear}(\mathbf{M}_t)), \tag{6}$$

$$\mathcal{L}_{\text{AS}} = \frac{1}{T}\sum_{t=0}^{T-1} (y_t - \tilde{m}_t)^2, \tag{7}$$

where $\mathbf{M}_t$ is the similarity score shown in Eq. (2), and Sigmoid and Linear denote the sigmoid-layer and linear-layer operations, respectively. Finally, the above two losses are added together to be the total loss used to train the entire PVAD network:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PVAD}} + \mathcal{L}_{\text{AS}}. \tag{8}$$

III. Experimental Setup

A. Dataset

The LibriSpeech corpus [19] is employed as the source data to evaluate the presented PVAD systems. The training set consists of three subsets, totaling 960 hours of speech data from 2,338 different speakers: the training subsets train-clean-100 and train-clean-360 provide a total of 460 hours of clean speech, while train-other-500 provides 500 hours of noisy speech. Similarly, the LibriSpeech test set includes both clean and noisy speech, totaling 10 hours of speech from 73 speakers. To proceed with the experiments for PVAD, each individual utterance in the LibriSpeech corpus cannot be directly used since it just corresponds to a single speaker. Therefore, we prepare a dataset that contains concatenated utterances from multiple speakers using this corpus. We choose utterances from one to three speakers at random following a uniform distribution and concatenate them. For each concatenated utterance, one of the speakers is randomly chosen as the target speaker. In addition, we randomly select the concatenated sentences with a probability of 0.2 and replace their corresponding target speaker embeddings with the embeddings from non-target speakers that are absent from those sentences. This arrangement aims to prevent the model from detecting any target speaker activity for these sentences in training. It also improves the model's generalization capabilities and prevents biased outputs, which could result in high false-positive rates.

Regarding speaker embedding, utterances from each speaker are randomly selected and fed into a pre-trained speaker verification model to generate window-level 256-dim d-vectors. These d-vectors are then L2-normalized and averaged to produce the utterance-level d-vector, which serves as the target speaker embedding $\mathbf{e}^{\text{target}}$. Additionally, to prevent overfitting and improve the model's robustness, we employ the MTR data augmentation technique [20]. This technique introduces random noise sources with different room impulse responses, effectively improving the model's simulation of various noise and reverberation conditions. Consequently, the model is expected to perform well in a broader range of real-world environments.

B. Implementation details

We extract 40-dimensional log Mel-filterbank energies as raw acoustic features from utterances with a frame size of 25 ms and a step of 10 ms. For three AT-PVAD variants presented, the model details are as follows:

1) COEA: The 2-layer (front) LSTM contains 64 cells for each layer.
2) COLA: The 2-layer (back) LSTM contains 40 cells for each layer.
3) COIN: The one-layer front LSTM contains 40 cells, and the one-layer back LSTM contains 64 cells.

Additionally, we prepare an **AT-PVAD baseline model**, which is nearly identical to the COEA-AT-PVAD structure, with the exception that it employs **the concatenation of raw acoustic features (40-dim Mel-filterbank energies) and target speaker embedding as the AS module input**. This baseline model allows us to examine whether conditional fusion outperforms direct concatenation for the two sources of information. Finally, all AT-PVAD structures end with two fully connected layers that serve as the classifier.

TABLE I
AVERAGE PRECISION (AP) FOR THE TARGET-SPEAKER SPEECH (TSS), ACCURACY AND MODEL SIZE OF THE AT-PVAD VARIANTS.

Model	AP (**tss**)	Accuracy (%)	Parameters (k)
AT-PVAD baseline	0.868	83.79	84.949
COIN-AT-PVAD	**0.912**	**86.74**	71.869
COLA-AT-PVAD	0.901	86.27	**55.285**
COEA-AT-PVAD	0.903	86.08	92.029

TABLE II
AVERAGE PRECISION (AP) FOR THE TARGET-SPEAKER SPEECH (TSS), ACCURACY AND MODEL SIZE OF THE ORIGINAL PVAD, PVAD 2.0, THE AT-PVAD BASELINE AND COIN-AT-PVAD

Model	AP (**tss**)	Accuracy (%)	Parameters (k)
PVAD	0.884	84.34	130.242
PVAD 2.0	0.908	86.56	97.602
AT-PVAD baseline	0.868	83.79	84.949
COIN-AT-PVAD	**0.912**	**86.74**	**71.869**

We implemented all models using PyTorch [21]. For model training, we initially employed the Adam optimizer [22] with a learning rate of 1×10^{-3} for the first epoch, and subsequently reduced the learning rate to 1×10^{-5} for the following epochs.

IV. RESULTS AND DISCUSSIONS

We evaluate PVAD models with a variety of metrics. Accuracy (%) is calculated as the ratio of the number of correctly detected frames to the total number of detected frames. It serves as an evaluation metric to determine how well the PVAD system identifies target speaker speech versus non-target speaker speech frames. Average Precision (AP) is the area under the Precision-Recall Curve with respect to the target-speaker speech (tss). The metric AP is important because the PVAD dataset contains fewer positive samples than negative ones. In addition, we assess the suitability of models for resource-limited devices by taking into account the number of model parameters.

A. Comparison of the AT-PVAD methods

Table I displays the results of the AT-PVAD baseline and three proposed AT-PVAD variants for comparison and analysis. From this table, we have the following observations:

1) The three AT-PVAD variants behave better than the AT-PVAD baseline model, revealing that the conditioned fusion is a better way than the concatenation to merge speaker embeddings and acoustic features as the input to the AS module.
2) COIN-AT-PVAD exhibits the optimal performance among the three AT-PVAD variants in the two metrics, AP and Accuracy, and it has fewer model parameters compared to the AT-PVAD baseline. The central placement of the AS module in the PVAD pipeline provides COIN-AT-PVAD with potential advantages:
   - The front LSTM in the model acts like an acoustic encoder, enhancing the model's non-linear capability to achieve better feature representation. This allows the speaker information to be more effectively integrated with the enhanced acoustic features through the collaborative operation of the FiLM generator.
   - As the AS module highlights the target speaker's characteristics in the deep acoustic features, the back LSTM captures complex patterns and temporal relationships of features before final classification, further enhancing feature representation.
3) COLA-AT-PVAD has the fewest model parameters while it behaves closely to COEA-AT-PVAD. The smaller size of the LSTM used in COLA-AT-PVAD makes it more suitable for deployment on resource-limited devices.

B. Comparison of AT-PVAD and the SOTA PVAD

Table II includes the evaluation results of the original PVAD [7], PVAD 2.0 [8], the AT-PVAD baseline and the newly presented COIN-AT-PVAD for a more comprehensive comparison. As a note, PVAD 2.0 mainly revises the original PVAD by introducing an FiLM layer to make the acoustic feature conditioned with speaker embedding. We have some findings from this table:

1) Equipped with the FiLM layer, PVAD 2.0 and the novel COIN-AT-PVAD behaves better than the original PVAD and AT-PVAD baseline. In particular, COIN-AT-PVAD outperforms PVAD 2.0 in all of the three metrics by providing moderate better AP and Accuracy scores but requiring much fewer model parameters (71.869k versus 97.602k). Thus COIN-AT-PVAD offers substantial benefits for practical applications.
2) The AT-PVAD baseline, which grabs the idea of the AS module in AS-pVAD [11], does not perform as well as expected. It can be attributed to the fact that, compared to the AT-PVAD baseline, AS-pVAD utilizes a more advanced speaker embedding extractor (ECAPA-TDNN) and acoustic feature encoder (TCNN) to offer superior input representation for the AS module.

V. CONCLUSIONS

In this study, we propose a novel PVAD framework which adopts an advanced attentive score module. This module uses acoustic features conditioned on the target speaker embedding to generate the attention weight. We have conducted comparative analyses on the PVAD variants that incorporate this conditional attention mechanism at the early, mid, and late stages of the PVAD pipeline. The experimental results show that all the PVAD variants equipped with conditional attention perform well. Specifically, the PVAD with mid-stage conditional attention (COIN-AT-PVAD) shows excellent performance and has a compact model size, making it suitable for deployment on resource-constrained devices.

ACKNOWLEDGMENT

This work was supported by Realtek Semiconductor Corporation. Any findings and implications in the paper do not necessarily reflect those of the sponsor.

REFERENCES

[1] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[2] S.-Y. Chang, B. Li, T. N. Sainath, G. Simko, and C. Parada, "Endpoint Detection Using Grid Long Short-Term Memory Networks for Streaming Speech Recognition," *Proc. Interspeech*, 2017.

[3] M. Shannon, G. Simko, S.-Y. Chang, and C. Parada, "Improved End-of-Query Detection for Streaming Speech Recognition," *Proc. Interspeech*, 2017.

[4] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-End Automatic Speech Recognition Integrated With CTC-Based Voice Activity Detection," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[5] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[6] Y. G. Thimmaraja, B. G. Nagaraja, and H. S. Jayanna, "Speech enhancement and encoding by combining SS-VAD and LPC," *Int J Speech Technol*, 2021.

[7] S. Ding, Q. Wang, S.-y. Chang, L. Wan, and I. L. Moreno, "Personal VAD: Speaker-Conditioned Voice Activity Detection," *Speaker Odyssey*, 2020.

[8] S. Ding, R. Rikhye, Q. Liang, *et al.*, "Personal VAD 2.0: Optimizing Personal Voice Activity Detection for On-Device Speech Recognition," *Proc. Inrespeech*, 2022.

[9] S. Kumar, S. S. Buddi, U. O. Sarawgi, *et al.*, "Comparative Analysis of Personalized Voice Activity Detection Systems: Assessing Real-World Effectiveness," *Proc. Interspeech*, 2024.

[10] B. Zeng, M. Cheng, Y. Tian, H. Liu, and M. Li, "Efficient Personal Voice Activity Detection with Wake Word Reference Speech," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[11] F. Liu, F. Xiong, Y. Hao, K. Zhou, C. Zhang, and J. Feng, "AS-pVAD: A Frame-Wise Personalized Voice Activity Detection Network with Attentive Score Loss," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[12] I. Medennikov, M. Korenevsky, T. Prisyach, *et al.*, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," *Proc. Interspeech*, pp. 274–278, 2020.

[13] A. Jayasimha and P. Paramasivam, "Personalizing Speech Start Point and End Point Detection in ASR Systems from Speaker Embeddings," *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pp. 771–777, 2021.

[14] M. He, D. Raj, Z. Huang, J. Du, Z. Chen, and S. Watanabe, "Target-speaker Voice Activity Detection with Improved I-Vector Estimation for Unknown Number of Speaker," *arXiv preprint arXiv:2108.03342*, 2021.

[15] N. Makishima, M. Ihori, T. Tanaka, A. Takashima, S. Orihashi, and R. Masumura, "Enrollment-less training for personalized voice activity detection," *Proc. Interspeech*, 2021.

[16] E.-L. Yu, K.-H. Ho, J.-W. Hung, S.-C. Huang, and B. Chen, "Speaker Conditional Sinc-Extractor for personal VAD," *Proc. Inrespeech*, 2024.

[17] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual Reasoning with a General Conditioning Layer," *AAAI*, 2018.

[18] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2018.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[20] C. Kim, A. Misra, K. Chin, *et al.*, "Generation of Large-Scale Simulated Utterances in Virtual Rooms to Train Deep-Neural Networks for Far-Field Speech Recognition in Google Home," *Proc. Interspeech*, pp. 379–383, 2017.

[21] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2019.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Separation of Cardiopulmonary Sound Signals for Classification of Respiratory Diseases

Ruxin Zheng* and Saeid Sanei†
* Imperial College London, London, United Kingdom
E-mail: 18272020090@163.com Tel: +86-15729321697
† Imperial College London, London, United Kingdom
E-mail: s.sanei@imperial.ac.uk Tel: +44-7919098002

***Abstract*—Due to having the mixture of heart and lung sounds, the traditional automatic cardiopulmonary sound signal analyses have some challenges in achieving sufficient classification accuracy and robustness. The aim of this study is to improve the performance of respiratory disease classification model by cardiopulmonary sound separation. We used 920 auscultated audio signals involving 127 subjects who were healthy or had different respiratory conditions. For separation of lung sounds from heart sounds and other interferences we have developed an adaptive subspace analysis which iteratively maximizes the kurtosis. The separated lung sound is then classified using random forest classifier, and the classification performance is evaluated by cross-validation. The results show that the classification performance of cardiopulmonary sound for different age groups (infants and adults) significantly improves.**

I. Introduction

Lung diseases mainly including chronic obstructive pulmonary disease (COPD), lung cancer and asthma are the major global public health problem. These diseases not only affect the quality of life of patients, but also place an enormous burden on global healthcare systems. Pulmonary sound classification plays an important role in clinical diagnosis, personalized treatment, public health monitoring, medical research and teaching, automation and telemedicine. By improving the accuracy and efficiency of pulmonary sound classification, the management and treatment of respiratory diseases can be significantly improved, and the quality of patients' lives can be improved.

Lung sounds are usually heard or recorded using stethoscope, which is also used for heart sound auscultation. The sounds heard by the stethoscope from the chest wall consist mainly of heart sounds, lung sounds, and other underlying sounds, such as thorax and muscle frictions. Heart sounds are mainly caused by the mechanical activity of the heart, including closing of the heart valves and the vibration of the heart.

A parallel source separation system is proposed to extract heart and lung sounds from single-channel mixed signals based on non-negative matrix decomposition and clustering strategies [1]. In addition, a hard threshold is used in the wavelet transform domain to separate the non-stationary part of the input signal (heart sound) from the stationary part (lung sound). By placing two microphones on the left and right chest walls to collect signals, the fast ICA algorithm is applied to separate the heart and lung sound signals [2]. Blind source separation is performed using non-negative matrix factorization (NMF). By constructing spectral mask and using NMF to decompose spectral features, heart and lung sound signals are separated [3]. A new period-encoding depth-using autoencoder (PC-DAE) method is proposed to isolate mixed cardiopulmonary sounds in an unsupervised manner by assuming different periodicity between the heart and respiratory rates [4].

Singular spectrum analysis (SSA) reveals the intrinsic structure of the data by breaking down the time series into interpretable components [5][6][7][8][9]. Therefore, in this paper, the collected cardiopulmonary sound signals are separated by SSA algorithm advance to extract purer lung sound signals, so as to improve the accuracy of subsequent classification of lung diseases.

VGG16 deep learning model has been used to detect and classify lung diseases [10]. The study has a large, publicly available dataset containing X-ray images of COVID-19, pneumonia, and pneumothorax. Classification of the lung diseases using deep learning and interpretable artificial intelligence (XAI) techniques has been discussed [11]. It mainly uses a variety of deep learning models, including convolutional neural networks (CNNs), hybrid model, integrated model and transformer model, and combines XAI technology to improve the interpretability of the model. A multi-class classification method based on CNN to learn from pulmonary disease images has been proposed in [12]. The pre-processing has been performed using center clipping, the fine-tuning learning performed using the EfficientNet B7 model, and the Multi GAP structure used to maximize the features of each layer.

Random forest is an ensemble learning method based on decision trees that is widely used in classification and regression tasks. It constructs multiple decision trees by randomly selecting samples and features during the training process, significantly improving the stability and generalization ability of the model, while being able to assess the importance of features and provide an interpretation of the classification results [13].

In this paper, we propose a method combining SSA and random forest classifier to separate cardiopulmonary sound signals from infants and adults, and classify the extracted lung sounds into various pulmonary diseases. The method is implemented through three main steps: signal separation, feature extraction, and classification including model training

and validation.

II. Description of Database

In this study we use the ICBHI 2017 Challenge dataset as the baseline dataset for pulmonary auscultation sounds for the detection of respiratory diseases. It includes 920 recordings ranging in length from 10s-90s from 127 subjects. In total, there are 5.5 hours of recordings, containing 6898 breathing cycles. The data includes both clean breathing sounds and noisy recordings in real life. The data are from either healthy subjects or those having one of the seven categories of respiratory diseases.

III. Methodology

A. Singular Spectrum Analysis

SSA is a powerful time series analysis and processing tool, which is often used for single-channel signal separation, noise reduction, and pattern recognition [9]. The basic principle is to decompose and reconstruct the one-dimensional time series signal by embedding it into the high-dimensional trajectory space and extracting meaningful subspace structure from it.

1) Embedding: First, the one-dimensional time series is transformed into a high-dimensional trajectory matrix by stacking overlapping signal segments. Given a time series $X = \{x_1, x_2, \ldots, x_N\}$ and window length L. In general, L satisfies $2 \leq L < N/2$, where N is the length of the time series. In this research, L takes the value of 250. Then, we construct a trajectory matrix X by the Hankel transform of the signal, where each column is an overlapping segment of the time series, defined as:

$$X = \begin{bmatrix} x(1) & x(2) & \cdots & x(K) \\ x(2) & x(3) & \cdots & x(K+1) \\ \vdots & \vdots & \ddots & \vdots \\ x(L) & x(L+1) & \cdots & x(N) \end{bmatrix} \tag{1}$$

where the trajectory matrix X is a matrix of L rows and K columns, where each column of X is part of the original signal, L is the window length or embedding dimension, and K is the number of columns of the trajectory matrix, equal to $K = N - L + 1$.

2) The Covariance Matrix and Eigenvalue Decomposition: Next, we compute the covariance matrix S of the trajectory matrix X. The covariance matrix S is defined as

$$S = XX^T \tag{2}$$

Then, the EVD is performed, and the eigenvalues are sorted, and the relevant principal components are extracted for the separation of subsequent signals.

3) Adaptively Grouping the Desired Signal Components: From the morphology of the signals it is evident that the heart sound has a higher kurtosis than the lung sound. This property is exploited as a criterion for our adaptive single-channel source separation system. To select the desired eigentriples, we multiply a diagonal matrix W by EVD factors and optimize it in each iteration. The W matrix is initialized as an identity matrix, and it is iteratively optimized to maximize the kurtosis of the signal. Kurtosis measures the sharpness of the signal distribution. In each iteration, the W matrix is updated by evaluating the gradient of kurtosis with respect to W and using the gradient descent method. Let the kurtosis function be estimated as follows:

$$Kurt(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^4 \Bigg/ \left(\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2\right)^2 \tag{3}$$

where X is the separated signal, $X_i = W_{ii}\lambda_i^{1/2}U_i$, N denotes the total number of data points, $Kurt$ stands for kurtosis, and μ is the average value of the signal X. In the optimization process, W is then updated as:

$$W(i+1) = W(i) - \alpha \cdot \nabla_W(Kurt) \tag{4}$$

where α is the learning rate and $\nabla_W(Kurt)$ is the gradient concerning W. λ_i and U_i are respectively, the ith eigenvalue and eigenvector of S in eq.(2).

4) Reconstruction: According to the optimized W matrix, the covariance matrix S is decomposed into the lung sound interferences subspaces including the heart sound subspace from which each signal can be reconstructed.

Then, the time domain signal is reconstructed by diagonal averaging [9], transforming the matrix back into a one-dimensional signal while retaining the main characteristics of the desired lung sound signal. Through these steps, SSA can effectively separate lung sound and heart sound signals, providing accurate data for subsequent signal analysis and diagnosis.

5) Kurtosis: Kurtosis is a statistic that measures the peakedness of the distribution of a signal's probability distribution. By maximizing the kurtosis, the accuracy of signal separation can be improved, especially when separating non-Gaussian signal components [14]. In this study, we propose to regularize the SSA-based signal reconstruction by maximizing the kurtosis of the reconstructed signal component to optimize the separation effect of heart and lung sounds. The overall is shown in Fig. 1.

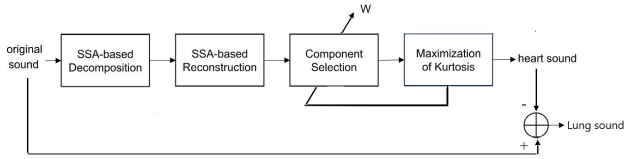


Fig. 1: The adaptive SSA algorithm regularized by kurtosis maximization criterion.

B. Random Forest Classifier

A random forest is an ensemble learning method that builds multiple decision trees for classification or regression. Each tree is trained on a different sample set and feature subset,

and the final prediction is voted on by the results of all trees [12].The process of building and evaluating a random forest classifier is shown in Fig. 2.

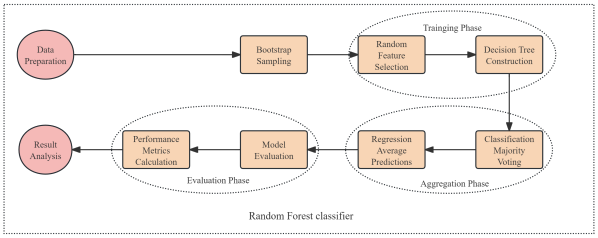


Fig. 2: The process of building and evaluating a random forest classifier.

For the original dataset, random forest generates multiple training sets through Bootstrap sampling [15]. Each training set is obtained by randomly drawing samples from the original data set, and the size of each training set is the same as the original dataset. Assume T is a training set generated through Bootstrap sampling: $\{x_i\}_{i=1}^{n}$ of which each x_i is randomly drawn from the original data set, there are put back, so each sample may be repeated in the training set.

For each of the Bootstrap sampling to generate the training set, build a decision tree. Split at each node, randomly select m characteristics (namely feature subset), and then select the best split point from these features [11]. If the total number of features is p, it is usually $m \ll p$ (classification problem) or $m \approx p$ (regression problem). For each node, select the split point that maximizes information gain or minimizes Gini impurity in the feature subset. If the feature subset is $\{x_1, x_2, \ldots, x_m\}$, the optimal split point is selected as s, where s represents the point of split, the information gain brought about by s, or the reduction of Gini impurity [15].

After all the trees are built, the random forest makes a final prediction by integrating the results of all the tree predictions. For the classification tasks, the random forest determines the final category through a majority voting mechanism. Let N be the number of trees in the forest and y_j the prediction of the j-th tree for sample x, then the final prediction category $\hat{y}$ is:

$$\hat{y} = \text{mode}(y_1, y_2, \ldots, y_N)$$

where mode represents the mode function, that is, the class that returns the most occurrences. For the regression task, the random forest determines the final predicted value by averaging the predicted values of all trees. Let $\hat{y}_j$ be the predicted value of the j-th tree for a new sample x, then the final predicted value $\hat{y}$ is:

$$\hat{y} = \frac{1}{N} \sum_{j=1}^{N} \hat{y}_j$$

IV. Results and Discussion

According to the characteristics of cardiopulmonary sound data selected in this study, the embedding dimension (window length) of SSA was set to 250 for constructing the trajectory matrix. For the W matrix optimization part, the learning rate is set to 0.01 for the gradient descent optimization of diagonal W matrix. At the same time, the convergence threshold is set to 1e-6 and the maximum number of iterations is set to 100, which is used to determine the convergence condition of the gradient descent algorithm and control the number of iterations of the gradient descent algorithm. The following takes one of the speech signals as an example to illustrate the effectiveness of using SSA and Kurtosis algorithm to separate the cardiopulmonary sound signals.

A. Results of separation

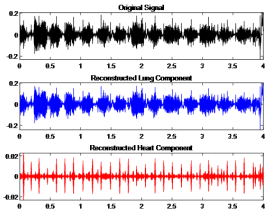
From the signal waveform and the histogram of Fig. 3, we can see that the amplitude of the original signal varies greatly at different time points, and the amplitude value presents a symmetric distribution between -0.2 and 0.2, and a wide frequency band. The time domain waveform and histogram of separated lung sound signal are similar to the original signal, including a wide frequency range and complex time domain changes. It is consistent with the characteristics that lung sounds usually contain high and low frequency components and vary with the respiratory cycle. The time domain waveform of separated heart sound signal is regular, the amplitude is relatively small and the periodicity is consistent. The amplitude values of the histogram are concentrated between -0.02 and 0.02, and the distribution is narrow. This is consistent with the physiological characteristics of heart sounds, which are mainly concentrated in the lower frequency range and produce a distinct pulse with each heartbeat.

In Fig. 3, the spectrum diagram shows the frequency distribution of signals at different time points, and power spectral density diagram shows the power distribution of signals at different frequencies. The reconstructed pulmonary sound signal has a significant energy distribution in the frequency range of 0 to 10 kHz, especially with a strong component above 1 kHz, and has a high power over a wide frequency range.

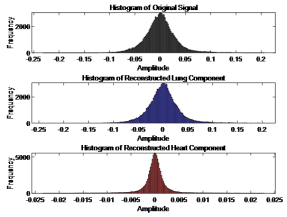
The reconstructed heart sound signals are mainly concentrated in the low frequency range (0 to 1 kHz), have low power, and show periodic changes on the timeline. This is consistent with the low-frequency nature of the heart sound signal and the pulses produced by each heartbeat. The spectral diagram and the power spectrum density analysis further quantified the distribution of the separated signal in the frequency domain, and proved the efficiency of the separation process.

In addition, in terms of numerical value, the dot product of reconstructed lung sound signal and heart sound signal is 1.480311, and Pearson correlation coefficient is 0.172788, indicating that the linear correlation between the two signals is weak, and there is a certain independence in the time domain. This further illustrates that the separation effect is better and the correlation between the signals is low.

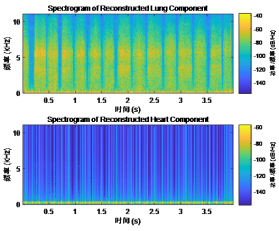
In addition, in terms of numerical value, the dot product of reconstructed lung sound signal and heart sound signal is 1.480311, and Pearson correlation coefficient is 0.172788, indicating that the linear correlation between the two signals



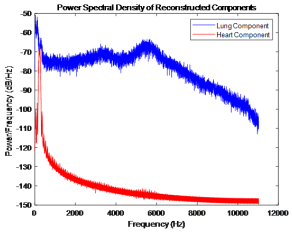
(a) Time domain waveform of original signal (left top), reconstructed lung sound signal (left middle) and heart sound signal (left bottom).



(b) Histogram of the original signal (right top), reconstructed lung sound signal (right middle), and heart sound signal (right bottom).



(c) Spectrograms of reconstructed lung sound signals (left top) and heart sound signals (left bottom).



(d) Power spectral density (PSD) of reconstructed lung sound signals and heart sound signals (right).

Fig. 3: Analysis of Time Domain, Histogram, Spectrogram, and Power Spectral Density of Cardiopulmonary Sound Signals.

are weak, and there is a certain independence in the time domain.

The overall classification accuracy with balanced dataset of infant samples before and after separation was 0.3333 and 0.5. The classification performance of the cardiopulmonary sound after separation has different performance in different categories.

B. Results of lung disease classification

Here is a summary of the respiratory diseases in the database, including the breakdown for children aged 7 years or younger in Table I. Because the number of samples of LRTI and Asthma patients is too small to be accurately classified, these two types of samples were not considered in the actual experiment.

TABLE I: Summary of the respiratory diseases in the database

Disease Category	Number of All Patients	Number of All Samples	Number of Infants ($\leq$ 7 years)	Number of Samples ($\leq$ 7 years)
Healthy	26	35	16	21
COPD	64	793	0	0
URTI	14	23	12	21
Bronchiectasis	7	16	0	0
Pneumonia	6	37	1	3
Bronchiolitis	6	13	6	12
Total	123	917	35	57

Table II and Table III illustrate respectively the performance of the classifier performance before and after applying SSA with balanced dataset. Precision represents the percentage of all samples correctly classified. Recall represents the proportion of all samples. F1 Score is the harmonic average of accuracy rate and recall rate,and is a comprehensive evaluation of the two.

TABLE II: The classification performance before and after lung sound separation (infants).

Disease Category	Precision (After)	Precision (Before)	Recall (After)	Recall (Before)	F1 Score (After)	F1 Score (Before)
Bronchiolitis	0.66667	NAN	0.66667	0	0.66667	NAN
Healthy	0	0	0	0	NAN	NAN
Pneumonia	0.33333	0.5	0.66667	1	0.44444	0.66667
URTI	1	0.25	0.66667	0.3333	0.8	0.28571

TABLE III: The classification performance before and after lung sound separation (infants and adults).

Disease Category	Precision (After)	Precision (Before)	Recall (After)	Recall (Before)	F1 Score (After)	F1 Score (Before)
Bronchiectasis	0.75	0.73333	0.92308	0.84615	0.82759	0.78571
Bronchiolitis	0.41667	0.4375	0.38462	0.53846	0.4	0.48276
COPD	0.92308	0.92308	0.92308	0.92308	0.92308	0.92308
Healthy	0.4	0.36364	0.30769	0.30769	0.34783	0.33333
Pneumonia	0.75	0.66667	0.92308	0.76923	0.82759	0.71429
URTI	0.36364	0.125	0.30769	0.076923	0.33333	0.095238

In general, the classification model of cardiopulmonary sound after separation has different performance in different categories. For infant data, the classification performance of the bronchitis and upper respiratory infection categories improved significantly. However, for the healthy and pneumonia categories, the model performance did not significantly improved or deteriorate. When the adult data were added, the performances of the model in the identification of bronchitis, healthy and pneumonia were different, and the overall performance of the identification of pneumonia and upper respiratory tract infection was significantly improved.

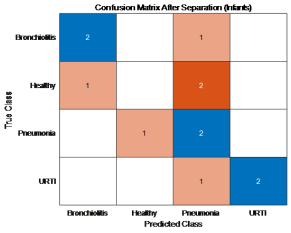
Classification of the separated single infant samples and the addition of the adult samples were generally better than classification of the unseparated samples. After separation, the model has higher accuracy and lower loss in training and validation data. The unseparated models showed obvious signs of overfitting, the verification accuracy and loss fluctuated greatly, and the classification performance was poor.

The overall classification accuracy with balanced dataset of infant samples before and after separation was 0.3333 and 0.5, and that of all samples before and after separation was 0.57692 and 0.62821, and the accuracy with unbalanced dataset of all samples before and after separation was 0.91385 and 0.91603, respectively. In conclusion, the separation of cardiopulmonary sounds is helpful to improve the accuracy and stability of respiratory disease classification.

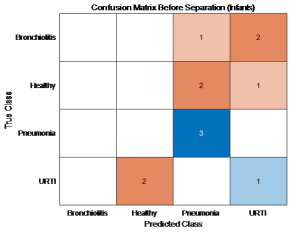
V. Conclusions

In this study, SSA algorithm has been regularized by maximization of kurtosis to enhance auscultated sound signal separation. The classification performance before and after separation was evaluated for a random forest classifier and used cross-validation and provided a comprehensive evaluation process and detailed analysis of the results.

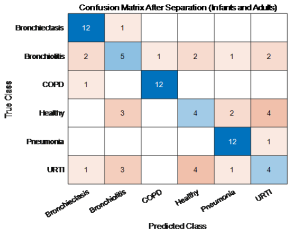
The separation of cardiopulmonary sound signals from infants and adults significantly improved the performance of the respiratory disease classification model. The results suggest that by incorporating kurtosis for SSA optimization, the model is better equipped to handle the complexities of cardiopulmonary sounds, leading to more accurate classifications. This



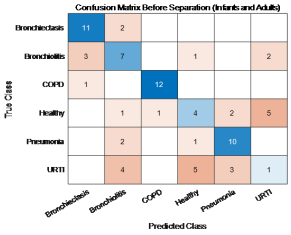
(a) Confusion Matrix After Separation (Infants).



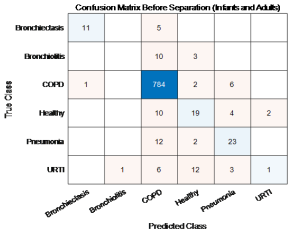
(b) Confusion Matrix Before Separation (Infants).



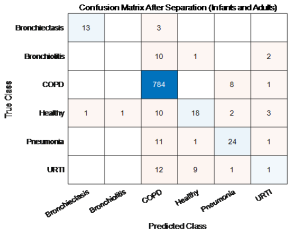
(c) Confusion Matrix After Separation (Infants and Adults).



(d) Confusion Matrix Before Separation (Infants and Adults).

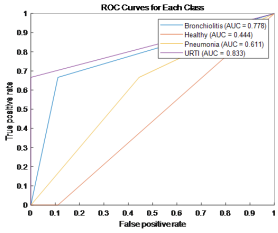


(e) Confusion Matrix After Separation (Unbalanced Dataset).

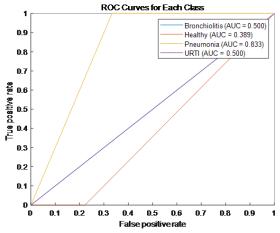


(f) Confusion Matrix Before Separation (Unbalanced Dataset).

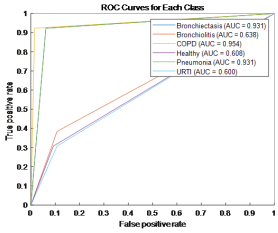
Fig. 4: Confusion Matrices for Different Conditions.



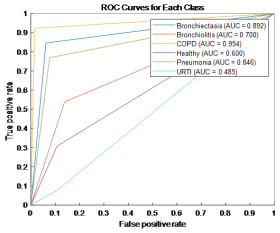
(a) ROC Curve After Separation (Infants).



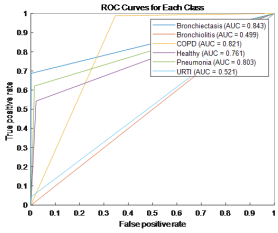
(b) ROC Curve Before Separation (Infants).



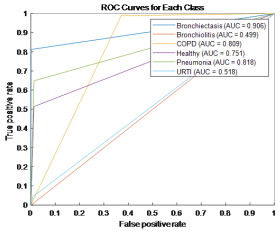
(c) ROC Curve After Separation (Infants and Adults).



(d) ROC Curve Before Separation (Infants and Adults).



(e) ROC Curve After Separation (Unbalanced Dataset).



(f) ROC Curve Before Separation (Unbalanced Dataset).

Fig. 5: ROC Curves After (left of top) and Before (middle of top) Separation of Infants and After (right of top) and Before (left of bottom) Separation of Infants and Adults with Balanced Dataset and After (middle of bottom) and Before (right of bottom) Separation of Infants and Adults with Unbalanced Dataset.

study provides empirical support for the practical application of cardiopulmonary sound separation technology in clinical settings, demonstrating its potential to enhance diagnostic accuracy and contribute to better patient outcomes.

REFERENCES

[1] A. J. Muñoz-Montoro, D. Suarez-Dou, and R. Cortina, "Parallel source separation system for heart and lung sounds," *J Supercomput*, vol. 77, pp. 8135–8150, 2021.

[2] J. C. Chien, M. C. Huang, Y. D. Lin, and F. C. Chong, "A study of heart sound and lung sound separation by independent component analysis technique," in *Conf Proc IEEE Eng Med Biol Soc*, 2006, pp. 5708–5711.

[3] C. Lin and E. Hasting, "Blind source separation of heart and lung sounds based on nonnegative matrix factorization," in *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, Naha, Japan, 2013, pp. 731–736.

[4] K.-H. Tsai and others, "Blind Monaural Source Separation on Heart and Lung Sounds Based on Periodic-Coded Deep Autoencoder," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 3203–3214, 2020.

[5] S. Sanei and H. Hassani, *Singular Spectrum Analysis of Biomedical Signals*, CRC, 2015.

[6] S. Sanei, T. K. M. Lee, and V. Abolghasemi, "A new adaptive line enhancer based on singular spectrum analysis," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 2, pp. 428-434, 2012.

[7] F. Ghaderi, H. Mohseni, and S. Sanei, "Localizing heart sounds in respiratory signals using singular spectrum analysis," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 12, pp. 3360-3367, 2011.

[8] S. Sanei, M. Ghodsi, and H. Hassani, "A constrained singular spectrum analysis approach to murmur detection from heart sounds," *Elsevier Journal of Medical Engineering and Physics*, vol. 33, no. 3, pp. 362-367, 2011.

[9] N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for Time Series*, Springer, New York, NY, 2013, pp. 11.

[10] A. Chatchaiwatkul and others, "Lung Disease Detection and Classification with Deep Learning Approach," in *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 2021, pp. 1–4.

[11] T. Taher, S. Ahmed, S. Shahriar, and T. Towhid, "Explainable Lung Disease Classification from Chest X-Ray Images Utilizing Deep Learning and XAI," 2024, pp. 1–5.

[12] M. Hong, B. Rim, H. Lee, H. Jang, J. Oh, and S. Choi, "Multi-Class Classification of Lung Diseases Using CNN Models," *Applied Sciences*, vol. 11, no. 19, p. 9289, 2021.

[13] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[14] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
[15] A. Liaw and M. Wiener, "Classification and Regression by RandomForest," *Forest*, vol. 23, 2001.

Optimization of the Intensity Aware Loss for Dynamic Facial Expression Recognition

Davy Tec-Hinh Lau [1], Jian-Jiun Ding [2], and Guillaume Muller [3]
National Taiwan University, Taipei, Taiwan
r12942162@ntu.edu.tw +33-783258324
National Taiwan University, Taipei, Taiwan
jjding@ntu.edu.tw +886-2-33669652
Ecole des Mines de Saint-Etienne, Saint-Etienne, France
guillaume.muller@emse.fr +33-0477420271

***Abstract*— Dynamic facial expression recognition (DFER) is a challenging topic in computer vision. One of the critical problems in DFER is the variation in intensity in the different scenes. Different intensities lead to different traits and in the low intensity scenarios most emotions will be misclassified as neutral expression. The intensity aware loss has been introduced as a method to put more focus on the low intensity samples and better capture the traits of low intensity cases to acquire better discriminative abilities. In this work, an alternative way to define the intensity aware loss is proposed. With the use of the logarithmic Euclidian distance, the cross-entropy loss, and the exponential decrease model, and model combination, the discriminative ability in the low-intensity case can be further improved. Moreover, several ways to further optimize the model was proposed. Experiments show that, with these proposed techniques, even better performance of DFER can be achieved.**

I. INTRODUCTION

Facial Expression Recognition is one of the most popular fields in computer vision now, because of the need of better human-computer interactions which requires the computer to understand how humans feel. After the failure of Static Facial Expression Recognition (SFER) – which provides results with high accuracy for single images within lab conditions – the focus is now on Dynamic Facial Expression Recognition (DFER), which aims to classify videos into expressions. There are a lot of difficulties in DFER because of the nature of the samples. Indeed, in the datasets (like DFEW) that has been created to develop DFER, most of the samples are in 'in-the-wild' conditions, which means that a lot of parameters (head motion, head angle, lighting, obstruction, intensity of the expression) impact the classification.

New models have been developed to tackle these problematics: Former DFER [4] introduces the use of transformer for DFER, which allows the model to have a better understanding of which channels are the more relevant for the prediction thanks to the attention mechanism. M3DFEL [10] is a multi-instance learning model, the principle is to remove the noisy frames that will have a bad influence on the classification, the frames are grouped into instances and an instance is considered noisy if all the frames within this instance are negative, in which case this instance will not be considered in the training. NR-DFERNet [12] is also a denoising method, the model captures the dynamic-static features of each frame, then each frame's features are compared to the average and the influence of one frame is determined by the gap between its features and the mean. The bigger the gap is, the lower is its influence on the training process. AEN [6] uses emotion grouping, the prediction is done in 2 steps: the first prediction is the group of emotion (positive, neutral, negative) and the second prediction is done within the group that has been predicted in the previous step. The positive group contains (happy, surprise), the neutral group contains only (neutral) and the negative one contains (sad, angry, disgust, fear). MIDAS [5] is a data augmentation method; data augmentation is useful in DFER because the number of samples are very limited and for example data in DFEW are imbalanced which cause a bias in the training.

In [1], Li *et al* introduced the intensity aware loss to tackle the problem of intensity of the expression. Indeed, they notice that all the emotions tend towards the neutral emotion when the intensity tends towards 0. Therefore, when the intensity of the emotion is low the emotions are more likely to be misclassified because the traits are more subtle and have more resemblance with the other emotions, in other words depending on the intensity the intra-class difference is bigger than the inter-class difference which will cause a bias in the classification. The intensity aware loss puts more focus on those low intensity samples to try to prevent this phenomenon. Nevertheless, the loss function that is proposed by [1] also influences the high intensity samples which are supposed to be clear predictions, therefore this can influence negatively the classification training. Also, the function put more focus on the low intensity samples but we believe that the value that it takes is not optimized and therefore the function can be improved to better assess the purpose.

To tackle these 2 issues, we design a new loss function that takes inspiration from the principle proposed by [7]. This loss function will have a influence closer to 0 when the intensity is very high and therefore the prediction is very clear, and will put more focus than the original Intensity Aware Loss on the low intensity samples.

We also design 3 ways to further optimize this loss function. Firstly, we decrease its influence on the training over the epochs because once the model has learnt to discriminate the emotions with different intensities, its influence will negatively impact the training. Then, we implement a pretrained model to tackle the data imbalance issue and lastly, we take inspiration from the MAN model (Mining Ambiguity and Noise) to combine 2 models in order to take the strength of both models to reduce the uncertainty.

II. Intensity Aware Loss and MAN Models

The Intensity Aware Loss is used to put more focus on the low intensity samples during the training and is defined as:

$$L_{IA} = -\log(P_{IA}), \tag{1}$$

$$P_{IA} = \frac{e^{x_t}}{e^{x_t} + e^{x_{max}}} \tag{2}$$

where x_t is the target logit and x_{max} is the largest logit excluding the target. The intensity is calculated by the gap between x_t and x_{max}. After analyzing the function, we can notice that the values of L_{IA} are included between [0.14;0.3], which represents the max and min values for the highest intensity and the lowest. We notice that even when the intensity is at its highest, the loss function has an influence that is equal to almost half of its influence when the intensity is at the lowest. Therefore, the loss function impacts the model even when it should not because when the intensity is high the prediction is clear. Also, the values for the low intensity samples are pretty low and it could be more emphasized.

The mining ambiguity and noised (MAN) model has been developed to tackle the issue of the noise in the labeling. Indeed, it is sometimes hard even for humans to tell what expression the subject is trying to make. Therefore, it is possible that the labelling is wrong, which will make the model learn bad material and the classification will be worse. They divide the annotations in 3 categories: clean, ambiguous and noisy. In order to separate the annotations, they make 2 different models predict the emotions. If both of them predict the target emotion, the annotation is considered clean. If only one of them manage to predict the target emotion, the annotation is ambiguous and finally if none of them predict the target then it is a noisy annotation. Then to reduce the impact of the noise in the labelling, different strategies are applied according to the category of the annotation. We will take inspiration from this model to combine 2 models to take the best out of their strength and get a model that is more robust and adaptative.

III. Proposed Method

A. Overview

To further improve the performance of DFER, we first propose the improved ways to define the loss function, and then the 3 methods for optimization the improved Intensity Aware Loss are introduced.

B. New Intensity Aware Loss

To tackle the issues that have been highlighted previously, we use the Euclidian distance to get the difference between x_t and x_{max}"

$$L_{IA}^{*} = -\log(\sqrt{({x_t}^2 - {x_{max}}^2)}) = -\log(|x_t - x_{max}|). \tag{3}$$

We then define:

$$P_{IA}^{*} = |x_t - x_{max}|. \tag{4}$$

If we consider a simpler model in which the prediction is binary between 2 emotions, we have:

$$x_t + x_{max} = 1, \tag{5}$$

$$L_{IA}^{*} = -\log(|2x_t - 1|). \tag{6}$$

After analyzing this new function, we get that the values of L_{IA}^{*} are contained in [0.009; 1.699]. We can see that now the focus put on the low intensity samples is much bigger and the influence on the high intensity samples is almost null. Nevertheless, the influence on the low intensity samples might be too big and might have to be adjusted. The final loss is obtained by using the cross-entropy loss and L_{IA}^{*}.

$$L_{total} = L_{CE} + \alpha L_{IA}^{*}$$

where α is a hyper-parameter controlling the loss coefficients, which we will determine in the next subsection.

C. Exponential Decrease

The newly introduced L_{IA}^{*} has almost no influence on the training when the prediction is clear and therefore x_t and x_{max} have a big gap. Nevertheless, in practice, it is unlikely to happen that the gap between x_t and x_{max} will be that big because in practice, the prediction is made between 7 emotions. Therefore, we well reduce the impact of L_{IA}^{*} over the training because the model mostly learns the discriminative behavior at the beginning of the training, and L_{IA}^{*} will only cause a bias if it keeps the same influence during all the training. Therefore L_{IA}^{*} becomes L_{final}^{*}:

$$L_{final}^{*} = A * e^{-\lambda * n} * L_{IA}^{*} \tag{7}$$

TABLE I. DISTRIBUTION OF THE EMOTIONS IN THE SUBSET

Emotion	Total of sample	%
Happy	98	12.47
Sad	119	15.14
Neutral	100	12.72
Anger	94	11.96
Surprise	107	13.61
Disgust	116	14.76
Fear	152	19.34

TABLE II. LIST OF COMBINATIONS

Combination	Without L^*_{final}	With L^*_{final}	Pretrained without L^*_{final}	Pretrained with L^*_{final}
a	x			x
b	x		x	
c		x		x
d		x	x	
e	x	x		
f			x	x

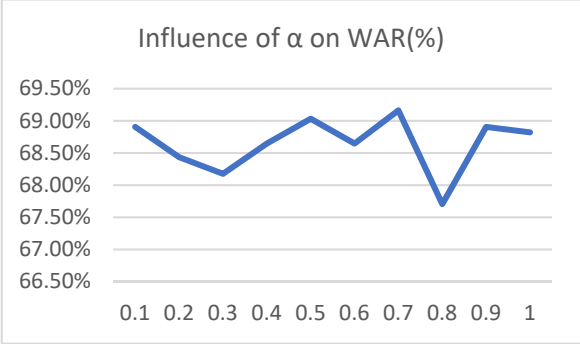


Fig. 1 Evolution of WAR with α

where A is the amplitude of the exponential, λ is the parameter that controls the speed of the decrease, and n the current epoch of the training. These parameters will be determined in the next subsection.

Moreover, we adopt 2 strategies for this method: the first one will be to have a big amplitude but also a quick decrease, whereas the other one will be to keep $A = 1$ but to decrease slowly the influence of the loss function.

D. Pretraining

One of the major issues of DFER is the lack of data for some emotions. This causes data imbalance in the dataset and therefore some emotions are very hard to predict. This is the case for "Disgust" that only has 146 samples in DFEW, so 1.22% of the dataset. To tackle this issue, we create a subset of the training data containing balanced data and we pretrain a model for $n_{pretrain}$ epochs.

TABLE III. PERFORMANCES OF THE SECOND STRATEGY

Model	**A**	**λ**	**WAR(%)**
2.1	1	0.05	69.119
2.2	**1**	**0.03**	**69.204**
2.3	1	0.02	69.162
2.4	1	0.035	69.076
2.5	3	0.05	68.135

TABLE IV. PERFORMANCES OF THE PRETRAINED MODELS

$n_{pretrain}$	Use of L^*_{final}	WAR (%)	UAR (%)
5	Yes	68.862	54.324
5	No	68.862	54.160
10	Yes	69.247	54.399
10	**No**	**69.290**	**54.693**
15	Yes	69.204	53.927
15	No	69.162	54.184

TABLE V. BEST PERFORMANCES OF EACH COMBINATION

Combination	**Best WAR**	β
a	69.418	0.2
b	69.504	0.6
c	69.461	0.6
d	69.461	0.7
e	**69.717**	**0.4**
f	69.632	0.9

As we can see in Table I. We use almost all the samples of "Disgust" in the subset. This could cause an overfitting. That is why we will pretrain the model with $n_{pretrain} \in \{5,10,15\}$, and for each of them we will conduct the pretraining with and without L^*_{IA}.

E. Model Combination

As mentioned previously, we take inspiration on the MAN model. We trained different models and combine them with a weighted average to have a more robust prediction.

$$\hat{Y} = \beta \hat{Y}_1 + (1-\beta)\hat{Y}_2 \quad (8)$$

where $\hat{Y}_1$ and $\hat{Y}_2$ are the predictions of the 2 models and β the parameter that controls the influence the importance of each model in the final prediction. We will conduct the combinations displayed on Table II.

IV. EXPERIMENTS

A. Experiment Details

TABLE VI. COMPARISON OF THE PERFORMANCES OF THE PROPOSED MODEL.

Methods	NE	AN	SU	UAR	WAR
Former DFER [4]	67.52	70.03	56.43	53.69	65.70
M3DFEL [10]	67.88	74.24	59.69	56.10	69.25
AEN [6]	70.67	72.08	59.07	56.66	69.37
NR-DFERNET [12]	70.03	75.09	61.60	54.21	68.19
Resnet18 + MIDAS [5]	58.64	68.06	59.65	**57.45**	69.16
IAL+GCA [1]	70.10	**76.06**	62.22	55.71	69.24
Our model	**72.28**	75.12	**67.69**	55.60	**69.72**

TABLE VII. PREDICTION ON THE NEW SAMPLES

	Final model	Benchmark	Pretrained
Happy low	0.466	0.437	0.563
Happy mid	0.686	0.786	0.987
Happy high	0.583	0.618	0.857
Surprise low	0.136	0.101	0.114

We use the dataset DFEW to conduct the experiments and the same implementation details as in [1]. Therefore, the images are resized in 112x112 pixel, the optimizer is SGD, with a batch size of 40, the learning rate equal to 0.001 which decrease exponentially over 80 epochs. Also, we use the Weighted Average Recall (WAR) and Unweighted Average Recall (UAR) as metrics.

B. Evaluation of the Parameters

We conduct studies on DFEW to determine the parameters of our final model.

We first begin by determining α: as we can see on Fig. 1. we achieve the best performances for $\alpha = 0.7$. We only consider the WAR here because what we want is the best model overall. On Table II. we can see the performances of the second strategy for the exponential decrease. The second strategy is better than the first one with an average WAR of 68.94% instead of 68.54%. Therefore, we take the model 2.2, which reduces the impact of the loss function to 9% after 80 epochs. Next, we achieved the best results with the pretrained models which has $n_{pretrain} = 10$ and without the use of L^*_{final} for the UAR and the WAR. Nevertheless, the aim of this method is not met. Indeed, the performances for "Disgust" and "Fear" are worse than before the pretraining even though it has better performances overall. The accuracy for "Fear" is still 0.00% and decreased by 4% for "Disgust".

Finally, we achieve the best result for combination *f* which combines the model without L^*_{final} and the model with L^*_{final}. This final model has a WAR of 69.72% and a UAR of 55.60% which makes the best model among those presented before. Also, we can note that this time the performances of 'Fear' have been improved by 5.6%.

C. Comparison with the Benchmark

We compare the benchmark with our final model and the best pretrained model. We use the best pretrained model to assess the performances of the new loss function because the final model is obtained by combining a model that uses L^*_{final} and one that does not use an IAL. Therefore, the prediction might be more accurate but it will be less clear. We filmed clips of ourself doing emotions with different intensities and make the 3 models predict the emotion. We have done happiness with low, mid and high intensity and also surprise with low intensity.

We can see on Table VII. that the pretrained model has indeed clearer predictions on the "Happy" samples.

On all the samples, the pretrained model have clearer predictions than the benchmark model. This is especially true for "Happy mid" with 0.987% instead of 0.786% for the benchmark.

V. CONCLUSION

In this work, we develop a new loss function inspired by the Intensity Aware Loss. This new loss function combined with 3 optimization methods that we implemented provides a model that achieves great performances with an increase of 0.35% compared to the second-best model. Also, as intended this new model puts more focus on the low intensity samples than the original IAL and the influence on the high intensity samples is negligible. To further optimize the model, it could be useful to use a data augmentation method to generate samples of 'Fear' and 'Disgust' instead of using the pretraining method.

ACKNOWLEDGE

The authors thank for the support of National Science and Technology Council under the contract of NSTC 113-2221-E-002-146.

REFERENCES

[1] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Intensity-aware loss for dynamic facial expression recognition in the wild," in *AAA*I Conf. Artificial Intelligence, vol. 37, pp.67-75, 2023.

[2] A. Psaroudakis and D. Kollias, "MixAugment & mixup: Augmentation methods for facial expression recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2367-2375, 2022.

[3] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "FERV39k: A large-scale multi-scene dataset for facial expression recognition in videos," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp.20922-20931, 2022.

[4] Z. Zhao and Q. Liu, "Former-DFER: Dynamic facial expression recognition transformer," in *ACM Int. Conf. Multimedia*, pp.1553-1561, 2021.

[5] R. Kawamura, H. Hayashi, N. Takemura and H. Nagahara, "MIDAS: Mixing ambiguous data with soft labels for dynamic facial expression recognition," in *IEEE Winter Conf. Applications of Computer Vision*, pp.6552-6562, 2024.

[6] B. Lee, H. Shin, B. Ku, and H. Ko, "Frame level emotion guided dynamic facial expression recognition with emotion grouping," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp.5681-5691, 2023.

[7] Z. Zhang, X. Sun, J. Li, and M. Wang, "MAN: Mining ambiguity and noise for facial expression recognition in the wild," *Pattern Recognition Letters*, vol. 164, pp.23-29, 2022.

[8] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, issue 2, article 199, 2023.

[9] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *Int. J. Cognitive Computing in Engineering*, vol. 2, pp.57-64, 2021.

[10] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou, "Rethinking the learning paradigm for dynamic facial expression recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp.17958-17968, 2023.

[11] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu and J. Liu, "DFEW: A large-scale database for recognizing dynamic facial expressions in the wild," in *ACM Int. Conf. Multimedia*, pp.2881-2889, 2020.

[12] H. Li, M. Sui, and Z. Zhu, "NR-DFERNet: Noise-robust network for dynamic facial expression recognition," *arXiv:2206.04975*, 2022.

Dynamic Sensor Placement on Graphs Based on Graph Signal Sampling Theory

Saki Nomura*, Junya Hara†, Hiroshi Higashi†, and Yuichi Tanaka†
* Tokyo University of Agriculture and Technology, Tokyo, 184–8588 Japan
† Osaka University, Osaka, 565–0871 Japan
E-mail: s_nomura@msp-lab.org, {j.hara, higashi, ytanaka}@comm.eng.osaka-u.ac.jp

***Abstract*—In this paper, we consider a sensor placement problem where sensors can move within a network over time. Most existing methods assume that sensor positions are static, i.e., they do not move, however, many mobile sensors like drones, robots, and vehicles can change their positions over time. Moreover, underlying measurement conditions could also be changed, which are difficult to cover with statically placed sensors. We tackle the problem by allowing the sensors to change their positions in their neighbors on the network. We dynamically determine the sensor positions based on graph signal sampling theory such that the non-observed signals on the network can be best recovered from the observations. For signal recovery, the dictionary is learned from a pool of observed signals. It is also used for the sensor position selection. In experiments, we validate the effectiveness of the proposed method via the mean squared error of the reconstructed signals. The proposed dynamic sensor placement outperforms the existing static sensor placement in synthetic data.**

I. Introduction

Sensor placement problem is one of the main research topics in sensor networks [1]–[3]. Its purpose is to find K sensor positions among N candidates $(K < N)$ where sensors can be placed optimally so that the reconstruction error is minimized in some sense. It has been extensively studied in machine learning and signal processing [2], [4]–[6].

Data on a network can be measured by sensors located at its nodes like wireless sensor networks, the Internet, and power grids [7]. Therefore, sensor placement on a network, i.e., graph, has been widely studied [8]. In the graph-based sensor placement problem, nodes of the graph represent possible sensor positions, edges correspond to relationships among the sensors, and the signal value on a node is a measurement.

Many existing works of sensor placement problem assume the static sensor placement, i.e., sensors do not or cannot move [9]–[11]. However, the measurement conditions could be changed over time. In this case, the selected sensors in the static positions do not reflect the current signal statistics. This may lead to the limitation of reconstruction qualities.

To overcome the limitation of the above-mentioned problem, in this paper, we propose a dynamic sensor placement where sensors can flexibly change their positions over time. To realize the dynamic sensor placement, we sequentially learn the dictionary for signal reconstruction from a pool of observed signals on a network (i.e., graph signals) by utilizing sparse coding. We then dynamically determine the sensor positions at every time instance based on graph signal sampling theory [12], such that the non-observed graph signal can be best recovered from the sampled graph signal.

In experiments, we demonstrate that the proposed method outperforms existing sensor placement methods in synthetic datasets.

Notation: We consider a weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ represent sets of nodes and edges, respectively. The number of nodes is $N = |\mathcal{V}|$ unless otherwise specified. The adjacency matrix of $\mathcal{G}$ is denoted by $\mathbf{W}$ where its (m, n)-element $\mathbf{W}_{mn} \geq 0$ is the edge weight between the mth and nth nodes; $\mathbf{W}_{mn} = 0$ for unconnected nodes. The degree matrix $\mathbf{D}$ is defined as $\mathbf{D} = \mathrm{diag}\,(d_0, d_1, \ldots, d_{N-1})$, where $d_m = \sum_n \mathbf{W}_{mn}$ is the mth diagonal element. We use graph Laplacian $\mathbf{L} \coloneqq \mathbf{D} - \mathbf{W}$ as a graph variation operator. A graph signal $\mathbf{x} \in \mathbb{R}^N$ is defined as a function $x : \mathcal{V} \to \mathbb{R}$. Simply speaking, $x[n]$ is located on the node $n \in \mathcal{V}$ of $\mathcal{G}$.

The graph Fourier transform (GFT) of $\mathbf{x}$ is defined as $\hat{x}(\lambda_i) = \langle \mathbf{u}_i, \mathbf{x} \rangle = \sum_{n=0}^{N-1} u_i[n] x[n]$, where $\mathbf{u}_i$ is the ith column of an orthonormal matrix $\mathbf{U}$. It is obtained by the eigendecomposition of the graph Laplacian $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with the eigenvalue matrix $\mathbf{\Lambda} = \mathrm{diag}\,(\lambda_0, \lambda_1, \ldots, \lambda_{N-1})$. We refer to λ_i as a *graph frequency*.

II. Graph Signal Sampling

In this section, we introduce graph signal sampling. We first review sampling theory on graphs and then describe a sampling set selection on graphs in [13].

A. Sampling Theory on Graphs

Here, we assume graph signals (i.e., sensor measurements in this paper) inherit the underlying graph information like smooth signals which dominate at low graph frequencies on a network [14]. Suppose that a graph signal is characterized by the following model [12]:

$$\mathbf{x} \coloneqq \mathbf{A}\mathbf{d}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N \times M}$ $(M \leq N)$ is a known generator matrix and $\mathbf{d} \in \mathbb{R}^M$ is a vector containing expansion coefficients. We can also assume that the small number of atoms (i.e., columns in $\mathbf{A}$) essentially contributes to the signal $\mathbf{x}$. The generator matrix $\mathbf{A}$ specifies the signal subspace $\mathcal{A}$, i.e., $\mathcal{A} = \mathcal{R}(\mathbf{A})$, where $\mathcal{R}(\cdot)$ represents the range space. A piecewise constant (PC) model is well-studied generators [12], [15], [16].

If we do not know the exact $\mathbf{A}$ but have its prior knowledge, it may be learned from the observations. Let $\mathbf{X} \in \mathbb{R}^{N\times M}$ be a collection of M graph signals. Typically, $\mathbf{A}$ has been determined by $\mathbf{A} = \boldsymbol{\mathcal{U}}$, where $\boldsymbol{\mathcal{U}}$ is obtained by the singular value decomposition (SVD) of the collection of observed signals $\mathbf{X} \coloneqq \boldsymbol{\mathcal{U}}\boldsymbol{\Sigma}\boldsymbol{\mathcal{V}}^\top$ [17], and $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ are unitary matrices.

Let $\mathbf{S}^\top \in \mathbb{R}^{K\times N}(K \leq N)$ be an arbitrary linear sampling operator where columns of $\mathbf{S}$ are linearly independent. Hereafter, we assume $K = M$ for simplicity, while we can easily extend the case of $K < M$. It specifies the sampling subspace $\mathcal{S}$, i.e., $\mathcal{S} = \mathcal{R}(\mathbf{S})$. A sampled graph signal is represented by $\mathbf{c} = \mathbf{S}^\top\mathbf{x}$. It is best recovered by the following transform [18]–[20]:

$$\tilde{\mathbf{x}} = \mathbf{A}(\mathbf{S}^\top\mathbf{A})^\dagger\mathbf{c}, \tag{2}$$

where $\cdot^\dagger$ is the Moore-Penrose pseudoinverse.

If $\mathcal{A}$ and $\mathcal{S}$ together span $\mathbb{R}^N$ and only intersect at the origin, perfect recovery, i.e., $\tilde{\mathbf{x}} = \mathbf{x}$, is guaranteed. This is referred to as the *direct sum* (DS) condition [12]. Note that the DS condition implies whether $\mathbf{S}^\top\mathbf{A}\mathbf{A}^\top\mathbf{S}$ is invertible, since $(\mathbf{S}^\top\mathbf{A})^\dagger = \mathbf{A}^\top\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{A}^\top\mathbf{S})^{-1}$ holds if and only if the DS condition is satisfied.

From the perspective of the sensor placement, $\mathbf{S}^\top$ can be considered as a node-wise sampling operator. In the following, we introduce a design method of $\mathbf{S}^\top$ on $\mathcal{G}$ for an arbitrary generator.

B. D-optimal Sampling Strategy

We introduce a graph signal sampling method utilized for our dynamic sensor placement [20]. Here, we define a node-wise sampling operator as follows:

Definition 1 (Node-wise graph signal sampling)**.** *Let $\mathbf{I}_{\mathcal{M}\mathcal{V}} \in \{0,1\}^{K\times N}$ be the submatrix of the identity matrix indexed by $\mathcal{M} \subset \mathcal{V}$ ($|\mathcal{M}| = K$) and $\mathcal{V}$. The sampling operator is defined as*

$$\mathbf{S}^\top \coloneqq \mathbf{I}_{\mathcal{M}\mathcal{V}}\mathbf{G}, \tag{3}$$

where $\mathbf{G} \in \mathbb{R}^{N\times N}$ is an arbitrary graph filter.

While $\mathbf{G}$ is arbitrary, a typical selection of $\mathbf{G}$ is the identity matrix or a graph lowpass filter as that in classical sampling theory. In a sensor placement perspective, $\mathbf{I}_{\mathcal{M}\mathcal{V}}$ specifies the sensor positions on $\mathcal{G}$ and $\mathbf{G}$ can be viewed as a coverage area of sensors. For example, we consider the following polynomial filter

$$[\mathbf{G}]_{nm} = \sum_{i=0}^{N-1}\sum_{p=0}^{P}\alpha_p\lambda^p u_i[n]u_i[m] = \left[\mathbf{U}\left(\sum_{p=0}^{P}\alpha_p\boldsymbol{\Lambda}^p\right)\mathbf{U}^\top\right]_{nm}, \tag{4}$$

where $\{\alpha_p\}$ are arbitrary coefficients. In graph signal processing, (4) is referred to as a P-hop localized filter [21], whose nonzero response is limited within P-hop neighbors of the target node. This implies that the coverage area of all sensors is limited to P-hop.

Sensor placement problem is generally formulated as

$$\mathcal{M}^* = \arg\max_{\mathcal{M}\subset\mathcal{V}} e(\mathcal{M}), \tag{5}$$

where e is a properly designed cost function. Typically, e is designed based on the optimal experimental designs [22], such as A-, E-, and D-optimality. D-optimality is one of the popular experimental designs and is widely used in graph signal sampling. The D-optimality is more appropriate than other optimality designs if we aim to recover graph signals from non-ideally sampled signals [20].

In the D-optimal design, the following problem is considered as selecting an optimal sampling set for graph signals [20]:

$$\mathcal{M}^* = \arg\max_{\mathcal{M}\subset\mathcal{V}} \det(\mathbf{Z}_{\mathcal{M}}), \tag{6}$$

where $\mathbf{Z} = \mathbf{G}\mathbf{A}\mathbf{A}^\top\mathbf{G}$ and thus $\mathbf{Z}_{\mathcal{M}} = \mathbf{I}_{\mathcal{M}\mathcal{V}}\mathbf{G}\mathbf{A}\mathbf{A}^\top\mathbf{G}\mathbf{I}_{\mathcal{M}\mathcal{V}}^\top$. Recall that $\mathbf{S}^\top\mathbf{A}\mathbf{A}^\top\mathbf{S}$ is invertible when the DS condition is satisfied. Therefore, the maximization of $\det(\mathbf{Z}_{\mathcal{M}}) = \det(\mathbf{S}^\top\mathbf{A}\mathbf{A}^\top\mathbf{S}) = |\det(\mathbf{S}^\top\mathbf{A})|^2$ leads to the best recovery of graph signals.

The direct maximization of (6) is combinatorial and is practically intractable. Therefore, a greedy method is applied to (6). Suppose that $\mathrm{rank}(\mathbf{Z}) \geq K$. In [23], (6) can be converted to a greedy selection as follows:

$$\begin{aligned} y^* &= \arg\max_{y\in\mathcal{M}^c} \det(\mathbf{Z}_{\mathcal{M}\cup\{y\}}) \\ &= \arg\max_{y\in\mathcal{M}^c} \mathbf{Z}_{yy} - \mathbf{Z}_{y\mathcal{M}}(\mathbf{Z}_{\mathcal{M}})^{-1}\mathbf{Z}_{\mathcal{M}y}. \end{aligned} \tag{7}$$

The selection is performed so that the best node y^* is appended to the existing sampling set $\mathcal{M}$ one by one.

III. Dynamic Sensor Placement on Graphs

In this section, we propose a dynamic sensor placement problem based on sampling theory on graphs. First, we derive online dictionary learning based on sparse coding. Second, we consider the control law of sensors. Fig. 1 shows the overview of the proposed method. The flow of the proposed method is outlined as follows:

1) The dictionary is sequentially learned from previous observations (Section III-A).
2) Time-varying graph signals are sampled and reconstructed based on graph sampling theory (Section II).
3) Sensor positions are dynamically determined based on the dictionary (Section III-B).

A. Online Dictionary Learning

In sampling and reconstruction introduced in Section II-A, $\mathbf{A}$ in (1) is assumed to be fixed. Note that (7) is regarded as a static sensor placement problem since the sensor positions are determined only once. In contrast, we consider the time-varying $\mathbf{A}_t$ in this paper where the subscript t denotes the time instance. In this setting, the optimal sensor positions can be changed according to $\mathbf{A}_t$.

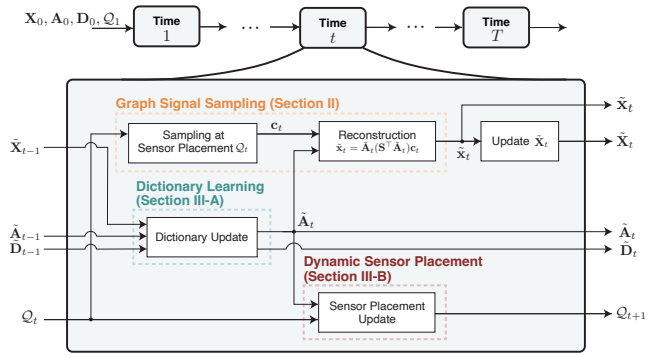


Fig. 1: Overview of the proposed method.

We infer a time-varying generator $\mathbf{A}_t$ from previous D observations

$$\mathbf{X}_{t-1} = [\mathbf{x}_{t-D}, \dots, \mathbf{x}_{t-1}], \tag{8}$$

in every time instance for dynamic sensor placement.

We assume the measurement model as follows:

$$\mathbf{X}_t = f(\mathbf{X}_{t-1}), \tag{9}$$

where f is a mapping from $\mathbf{X}_{t-1}$ to $\mathbf{X}_t$. That is, every observation is generated from signals in the previous time instance. Although the mapping of the time-evolution, f, is generally unknown, we suppose that f is Lipschitzian, i.e., there exists some constant L that satisfies the following inequality.

$$\|f(\mathbf{X}_{t-1}) - f(\mathbf{X}_t)\|_F \leq L\|\mathbf{X}_{t-1} - \mathbf{X}_t\|_F. \tag{10}$$

By definition of f, (10) can be transformed into

$$\|\mathbf{X}_t - \mathbf{X}_{t+1}\|_F \leq L\|\mathbf{X}_{t-1} - \mathbf{X}_t\|_F. \tag{11}$$

This assumption implies that signals vary smoothly in time[1].

Next, we consider estimating the generator $\mathbf{A}_t$ as a dictionary from (the estimated) $\tilde{\mathbf{X}}_{t-1}$ in (8) based on the assumption in (11), where $\tilde{\cdot}$ corresponds to the estimated vector/matrix.

Since the exact value of the number of atoms, M, may not be known in general, herein, we suppose $M \ll D$, and denote a dictionary and expansion coefficient by $\mathbf{A}_t \in \mathbb{R}^{N\times D}$ and $\mathbf{d}_t \in \mathbb{R}^D$, respectively. Therefore, we seek the optimal $\mathbf{A}_t$ such that the number in nonzero values of $\mathbf{d}_t$ is close to M, i.e., $\mathbf{d}_t$ is sparse. This setting is similar to well-studied dictionary learning problems [24], [25].

Let $\mathbf{D}_t = [\mathbf{d}_{t-D+1}, \dots, \mathbf{d}_t] \in \mathbb{R}^{D\times D}$ be the collection of expansion coefficients from time $t-D+1$ to t. Then, we can express $\mathbf{X}_t = \mathbf{A}_t\mathbf{D}_t$ (see (1)).

Here, we assume that the graph $\mathcal{G}$, initial generator matrix $\mathbf{A}_0$, and expansion coefficients matrix $\mathbf{D}_0$ are given for training, where $\mathbf{A}_0$ and $\mathbf{D}_0$ are arbitrary matrices. Following from (11), we suppose that $\mathbf{X}_t$ is sufficiently close to $\mathbf{X}_{t-1}$. As a result, we formulate the following problem:

$$\arg\min_{\tilde{\mathbf{A}}_t, \tilde{\mathbf{D}}_t} \|\tilde{\mathbf{X}}_{t-1} - \tilde{\mathbf{A}}_t\tilde{\mathbf{D}}_t\|_F^2 + \mu\|\tilde{\mathbf{D}}_t\|_1 + \eta\|\tilde{\mathbf{A}}_t - \tilde{\mathbf{A}}_{t-1}\|_F^2, \tag{12}$$

[1]Note that (11) is necessary to stably learn the dictionary, while we do not explicitly estimate f in the following part.

where $\tilde{\mathbf{X}}_{t-1} = \tilde{\mathbf{A}}_{t-1}\tilde{\mathbf{D}}_{t-1}$, and μ and η are the parameters. The first term in (12) is the data fidelity term for $\tilde{\mathbf{A}}_t$. The second and third terms are for the regularization controlling the sparsity of $\tilde{\mathbf{D}}_t$ and the temporal variation of $\tilde{\mathbf{A}}_t$, respectively.

Note that we need to solve (12) with respect to two variables $\tilde{\mathbf{A}}_t$ and $\tilde{\mathbf{D}}_t$, which jointly form a nonconvex optimization. In this paper, we divide (12) into two independent subproblems with respect to $\tilde{\mathbf{A}}_t$ and $\tilde{\mathbf{D}}_t$, and solve them alternately, similar to the method in [25].

First, we solve (12) with respect to $\tilde{\mathbf{A}}_t$ by fixing $\tilde{\mathbf{D}}_t$ as $\tilde{\mathbf{D}}_{t-1}$. In this case, we easily obtain the closed-form solution for the dictionary as follows:

$$\begin{aligned}\tilde{\mathbf{A}}_t &= \arg\min_{\tilde{\mathbf{A}}_t} \|\tilde{\mathbf{X}}_{t-1} - \tilde{\mathbf{A}}_t\tilde{\mathbf{D}}_{t-1}\|_F^2 + \eta\|\tilde{\mathbf{A}}_t - \tilde{\mathbf{A}}_{t-1}\|_F^2 \\ &= (\eta\tilde{\mathbf{A}}_{t-1} + \tilde{\mathbf{X}}_{t-1}\tilde{\mathbf{D}}_{t-1}^\top)(\eta\mathbf{I}_D + \tilde{\mathbf{D}}_{t-1}\tilde{\mathbf{D}}_{t-1}^\top)^{-1}. \end{aligned}\tag{13}$$

Second, we update $\tilde{\mathbf{D}}_t$ in (12) by using $\tilde{\mathbf{A}}_t$ in (13), i.e.,

$$\tilde{\mathbf{D}}_t = \arg\min_{\tilde{\mathbf{D}}_t} \|\tilde{\mathbf{X}}_{t-1} - \tilde{\mathbf{A}}_t\tilde{\mathbf{D}}_t\|_F^2 + \mu\|\tilde{\mathbf{D}}_t\|_1. \tag{14}$$

To solve (14), we utilize the proximal gradient method [26]–[28], whose problem is given by the following form.

$$\tilde{\mathbf{D}}_t = \arg\min_{\tilde{\mathbf{D}}_t} g(\tilde{\mathbf{D}}_t) + h(\tilde{\mathbf{D}}_t), \tag{15}$$

where g is a differentiable function and h is a proximable function [22]. Note that (15) is set to the entire problem being convex. As a result, the optimal solution is obtained by the following update rule.

$$\tilde{\mathbf{D}}_{t,n+1} = \mathrm{prox}_{\gamma h}(\tilde{\mathbf{D}}_{t,n} - \gamma\nabla g(\tilde{\mathbf{D}}_{t,n})), \tag{16}$$

where $n+1$ is the number of iterations, γ is the step size, ∇g is the gradient of g, and $\mathrm{prox}_{\gamma h}$ is defined as

$$\mathrm{prox}_{\gamma h}(\mathbf{D}) = \arg\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{C} - \mathbf{D}\|_F^2 + \gamma\mu\|\mathbf{C}\|_1 = S_{\gamma\mu}(\mathbf{D}), \tag{17}$$

where $S_{\gamma\mu}$ is the soft thresholding operator defined as follows:

$$[S_{\gamma\mu}(\mathbf{D})]_{ij} = \begin{cases} \mathbf{D}_{ij} - \gamma\mu, & \mathbf{D}_{ij} \geq \gamma\mu \\ 0, & |\mathbf{D}_{ij}| < \gamma\mu \\ \mathbf{D}_{ij} + \gamma\mu, & \mathbf{D}_{ij} \leq -\gamma\mu. \end{cases} \tag{18}$$

By applying (17) to (16), we have the following iteration with ISTA [29].

$$\begin{aligned}\mathbf{D}_{t,n+1} &= S_{\gamma\mu}(\mathbf{D}_{t,n} - \gamma\nabla g(\mathbf{D}_{t,n})) \\ &= S_{\gamma\mu}\{\mathbf{D}_{t,n} - 2\gamma\mathbf{A}_t^\top(\mathbf{A}_t\mathbf{D}_{t,n} - \tilde{\mathbf{X}}_{t-1})\}, \end{aligned}\tag{19}$$

We set the step size as $\gamma \leq 1/\lambda_{\max}(\mathbf{A}_t^\top\mathbf{A}_t)$ according to the convergence condition of ISTA [29]. We iterate (19) until $\|\mathbf{D}_{t,n+1} - \mathbf{D}_{t,n}\|_F^2 < \epsilon$ where ϵ is a small constant. The optimization of (12) is summarized in Algorithm 1.

In the following, we formulate the dynamic sensor placement based on the learned dictionary. The control law of sensors is also derived based on the sampling strategy introduced in (7).

Algorithm 1: Online Dictionary Learning

Input: $\tilde{\mathbf{X}}_{t-1}, \tilde{\mathbf{A}}_{t-1}, \tilde{\mathbf{D}}_{t-1}, t$
if $t = 1$ **then**
  $\tilde{\mathbf{A}}_1 = \mathbf{A}_0$
else
  $\tilde{\mathbf{A}}_t = (\eta\tilde{\mathbf{A}}_{t-1} + \tilde{\mathbf{X}}_{t-1}\tilde{\mathbf{D}}_{t-1}^\top)(\eta\mathbf{I}_D + \tilde{\mathbf{D}}_{t-1}\tilde{\mathbf{D}}_{t-1}^\top)^{-1}$
$\tilde{\mathbf{D}}_t = \tilde{\mathbf{D}}_{t-1}$
while $\|\tilde{\mathbf{D}}_{t,n} - \tilde{\mathbf{D}}_{t,n-1}\|_F^2 \geq \epsilon$ **do**
  $\tilde{\mathbf{D}}_{t,n+1} = S_{\gamma\mu}\{\tilde{\mathbf{D}}_{t,n} - 2\gamma\tilde{\mathbf{A}}_t^\top(\tilde{\mathbf{A}}_t\tilde{\mathbf{D}}_{t,n} - \tilde{\mathbf{X}}_{t-1})\}$
  $n \leftarrow n + 1$
Output: $\tilde{\mathbf{A}}_t$

B. Dynamic Sensor Placement

For brevity, we define the following matrix:

$$\mathbf{N}_t \coloneqq \tilde{\mathbf{A}}_t^\top \mathbf{G}. \tag{20}$$

Note that we now assume that $\tilde{\mathbf{A}}_t$ is given. With (20), (7) can be rewritten by a dynamic form as[2]

$$\begin{aligned} y_t^* &= \arg\max_{y\in\mathcal{M}^c} \mathbf{N}_{:y}^\top\mathbf{N}_{:y} - \mathbf{N}_{:y}^\top\mathbf{N}_{:\mathcal{M}}(\mathbf{N}_{:\mathcal{M}}^\top\mathbf{N}_{:\mathcal{M}})^{-1}\mathbf{N}_{:\mathcal{M}}^\top\mathbf{N}_{:y} \\ &= \arg\max_{y\in\mathcal{M}^c} \|\boldsymbol{\nu}_y\|^2 - \|\mathbf{P}_{\mathcal{R}(\mathbf{N}_{:\mathcal{M}})}\boldsymbol{\nu}_y\|^2, \end{aligned} \tag{21}$$

where $\boldsymbol{\nu}_y = \mathbf{N}_{:y}$ and $\mathbf{P}_{\mathcal{R}(\mathbf{Q})}$ is the orthogonal projection onto $\mathcal{R}(\mathbf{Q})$. Since $\mathbf{N}$ can temporally vary, the optimal solution in (21) can also change in every time instance.

Note that (21) is based on a greedy selection introduced in Section II-B. However, in practice, this optimization may not be efficient for large graphs: The control law with (21) implies that we need a large computational burden if all sensor positions are sequentially determined at every time instance. This results in a delay for sensor relocations which should be alleviated.

In this paper, instead, we consider selecting the sensor positions independently, i.e., (21) is solved for a sensor independently of the other ones. In the following, we rewrite (21) as a distributed optimization.

Here, we denote the position of the ith sensor at the tth time instance by $q_{i,t}$. We define the ith Voronoi region in the graph $\mathcal{G}$ as follows:

$$\mathcal{W}_{i,t} = \{v \in \mathcal{V} \mid d(v, q_{i,t}) < d(v, q_{j,t}), \forall q_j \in \mathcal{Q}_t\}, \tag{22}$$

where $d(v, q_{i,t})$ is the shortest distance between v and $q_{i,t}$, and $\mathcal{Q}_t = \{q_{j,t}\}_{j=1,\dots,K}$ denotes the set of all sensor positions at t. Note that, in each $\mathcal{W}_{i,t}$, only one sensor is contained: We illustrate its example in Fig. 2.

By utilizing (22), we can rewrite (21) for seeking the best position of a sensor within its Voronoi region $\mathcal{W}_{i,t}$ as follows:

$$q_{i,t+1}^* = \arg\max_{y\in\mathcal{W}_{i,t}} \|\boldsymbol{\nu}_y\|^2 - \|\mathbf{P}_{\mathcal{R}(\mathbf{N}_{:\mathcal{M}_{i,t}})}\boldsymbol{\nu}_y\|^2. \tag{23}$$

[2]For notation simplicity, we omit $\cdot_t$ in $\mathbf{N}_t$ hereafter.

Algorithm 2: Dynamic Sensor placement

Input: Sensing matrix $\mathbf{G}$, initial positions of sensors $\mathcal{Q}_1$, time duration T
for $t = 1, \dots, T$ **do**
  Update the dictionary $\tilde{\mathbf{A}}_t$ with Algorithm 1
  Update the data matrix $\tilde{\mathbf{X}}_t$
  Calculate the Voronoi regions $\{\mathcal{W}_{i,t}\}_{i=1,\dots,K}$
  $\mathbf{N}_t \leftarrow \tilde{\mathbf{A}}_t^\top\mathbf{G}$
  par for $i = 1, \dots, K$
    Calculate $\mathbf{P}_{\mathcal{R}(\mathbf{N}_{:\mathcal{M}_{i,t}})}$
    Calculate the P hop neighbors $\{\mathcal{N}_{i,t}^P\}_{i=1,\dots,K}$
    $q_{i,t+1}^* = \arg\max_{y\in\mathcal{W}_{i,t}\cap\mathcal{N}_{i,t}^P} \|\boldsymbol{\nu}_y\|^2 - \|\mathbf{P}_{\mathcal{R}(\mathbf{N}_{:\mathcal{M}_{i,t}})}\boldsymbol{\nu}_y\|^2$
  $\mathcal{Q}_{t+1} \leftarrow \{q_{i,t+1}^*\}_{i=1,\dots,K}$
Output: Sensor positions $\{\mathcal{Q}_t\}_{t=1,\dots,T+1}$

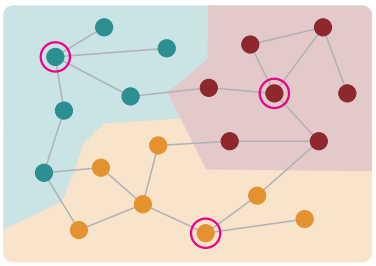


Fig. 2: Voronoi diagram on a graph. Nodes with circles represent the current sensor positions. Each colored area represents the Voronoi region corresponding to each sensor.

Since (23) depends only on the Voronoi region of a sensor, all sensor selections in (23) are independent of each other. Therefore, we can select sensors by (23) in a distributed fashion.

When we assume mobile sensors, their movable areas could be restricted to their neighbors. To reflect this, we can further impose a constraint on the Voronoi region where the movable nodes are restricted to the P-hop neighbors of the target node. The P-hop neighbor of the ith sensor (node) is defined as

$$\mathcal{N}_{i,t}^P = \{v \in \mathcal{V} \mid d(v, q_{i,t}) \leq P\}. \tag{24}$$

As a result, the proposed control law of sensors in (23) is further modified as follows:

$$q_{i,t+1}^* = \arg\max_{y\in\mathcal{W}_{i,t}\cap\mathcal{N}_{i,t}^P} \|\boldsymbol{\nu}_y\|^2 - \|\mathbf{P}_{\mathcal{R}(\mathbf{N}_{:\mathcal{M}_{i,t}})}\boldsymbol{\nu}_y\|^2. \tag{25}$$

Algorithm 2 summarizes the proposed dynamic sensor placement method[3]. Note that our main contribution is to formulate dynamic sensor placement as graph signal sampling. We may be able to use other methods, such as those described in [30] and [31], as internal algorithms for dictionary learning.

[3]We utilize the `MATLAB` notation where we denote the parallel processing in Algorithm 2 as **par for**.

IV. EXPERIMENTS

In this section, we compare reconstruction errors of time-varying graph signals measured by selected positions of sensors. We perform recovery experiments for synthetic datasets.

We suppose that the graph $\mathcal{G}$ is given a priori. The nodes represent the candidates of sensor positions, and the edges indicate the paths where sensors can move. Sampling is performed at every time instance. After sampling, the reconstructed signal at t, i.e., $\tilde{\mathbf{x}}_t$, is stored as the latest signal collection $\tilde{\mathbf{X}}_t$, which is used for the dictionary/sensor positions update. Simultaneously, $\tilde{\mathbf{x}}_{t-D}$ is discarded from $\tilde{\mathbf{X}}_t$.

A. Setup

We compare the signal reconstruction accuracy of the proposed method with those of the following existing methods.

1) *Static1*: Static sensor placement with the fixed dictionary learned by the SVD only once at $t = 0$ [17]. This method does not change the dictionary over time.
2) *Static2*: Static sensor placement (similar to Static1) with the non-fixed dictionary learned by the SVD at every time instance.

Since the SVD-based dictionary learning is still widely-used in many sensor placement applications [32], Static1 and Static2 are possible baseline methods. We also compare the proposed method to the extreme case of the proposed method, i.e., $P = \infty$. In this case, sensors can move anywhere in Voronoi region regardless of the current sensor positions.

The recovery experiment is performed for signals on a random sensor graph with $N = 256$. The number of sensors is set to $K = 8$.

We consider the following the piecewise-constant (PC) model:

$$\mathbf{x} = \sum_{i=0}^{M-1} d[i]\mathbf{1}_{\mathcal{T}_i} = [\mathbf{1}_{\mathcal{T}_1}, \ldots, \mathbf{1}_{\mathcal{T}_M}]\mathbf{d}, \tag{26}$$

where $\mathcal{T}_i (i = 1, \ldots, M)$ is the nonoverlapping subset of nodes for the ith cluster: $[\mathbf{1}_{\mathcal{T}_i}]_n = 1$ if the node n is in $\mathcal{T}_i$ and 0 otherwise [33]. In this case, the generator matrix is $\mathbf{A} = [\mathbf{1}_{\mathcal{T}_1}, \ldots, \mathbf{1}_{\mathcal{T}_M}]$ where M is the number of clusters. The number of clusters is set to $M = 3$. We randomly separate nodes into M clusters so that nodes in a cluster are connected. We generate $\mathbf{d}_t$ as follows:

$$\mathbf{d}_t = \begin{bmatrix} 3\{\exp(-\frac{t}{25})\sin(t + \frac{\pi}{3}) + 1\} \\ 3\{\exp(-\frac{t}{25})\sin(2t) + 1\} \\ 3\{\exp(-\frac{t}{25})\sin(3t - \frac{\pi}{3}) + 1\} \end{bmatrix}. \tag{27}$$

In this model, signal values in a cluster change simultaneously (but the temporal frequencies are different for different clusters). It can be a possible model for clustered sensor networks.

For the PC model, we generate the data matrix $\mathbf{X}_t$ with $T = 40$. The number of time slots is set to $D = 20$. The initial expansion coefficients and dictionary can be arbitrary as mentioned in Section III-A. In the experiment, however, we used an SVD-based dictionary for a fair comparison to alternative methods. Therefore, $\mathbf{A}_0 = \boldsymbol{\mathcal{U}}$ at $t = 0$, where $\boldsymbol{\mathcal{U}}$ is obtained by the SVD of the known measurement data, i.e., $\mathbf{X}_0 = \boldsymbol{\mathcal{U}}\boldsymbol{\Sigma}\boldsymbol{\mathcal{V}}^\top$. The initial value of $\mathbf{D}_t$ is also set to $\mathbf{D}_0 = \mathbf{1}\mathbf{1}^\top$. We set the temporal sampling period $T_s = \frac{\pi}{30}$ to avoid aliasing.

The parameters of the dictionary learning are experimentally set to $(\gamma, \eta, \mu) = (10^{-4}, 3, 1)$, and $P = 1$ is used for the sensor movable area where we assume a sensor can only be moved to its one-hop neighbor based on reasonable sensor mobility.

Recovery accuracy is measured in noisy scenarios where white Gaussian noise $w[i] \sim \mathcal{N}(0, 0.1)$ is added to the signals. For all methods, the initial position of sensors is determined by a state-of-the-art sampling set selection method on graphs [2]. We calculate the averaged mean squared errors (MSEs) for 50 independent runs.

B. Results

Fig. 3 shows the experimental results for the noisy case. We also show examples of the original and reconstructed graph signals of PC models in Fig. 4. Fig. 3 indicates that the proposed method shows consistently (and significantly) smaller MSEs than the other methods in all time instances, and the MSE of the proposed method gradually decreases over time. This validates that the proposed method can adapt to the change in measurement conditions, while the static methods fail to do so. In this experiment, $P = 1$ and ∞ do not results in a large difference. However, as shown in Fig. 4 at $t = 13$, we observe that $P = \infty$ produces a larger error than $P = 1$ because several sensor locations are very close to each other. A possible reason for this is that the sensor positions were biased toward the right side.

V. CONCLUSION

In this paper, we propose a dynamic sensor placement method based on graph sampling theory. We sequentially learn the dictionary from a time series of observed graph signals by utilizing sparse coding. Using the dictionary, we dynamically determine the sensor placement at every time instance such that the non-observed graph signal values can be best recovered from those of the observed (selected) nodes. In experiments, we demonstrate that the proposed method outperforms existing static sensor placement methods in synthetic datasets.

More details on theoretical aspects and comprehensive experimental results with respect to this paper can be found in [34].

ACKNOWLEDGMENT

This work is supported in part by JSPS KAKENHI under Grant 23K26110 and 23K17461, and JST AdCORP under Grant JPMJKB2307.

REFERENCES

[1] F. Y. Lin and P.-L. Chiu, "A near-optimal sensor placement algorithm to achieve complete coverage-discrimination in sensor networks," *IEEE Commun. Lett.*, vol. 9, no. 1, pp. 43–45, 2005.

[2] A. Sakiyama, Y. Tanaka, T. Tanaka, and A. Ortega, "Eigendecomposition-free sampling set selection for graph signals," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2679–2692, 2019.

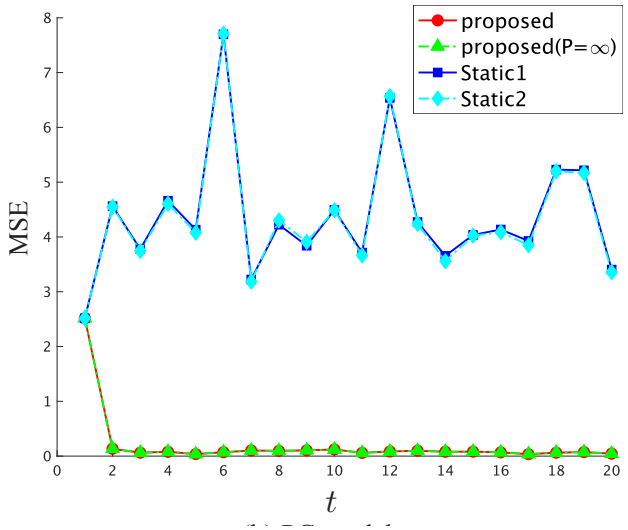


Fig. 3: MSE comparison of the reconstructed synthetic data.

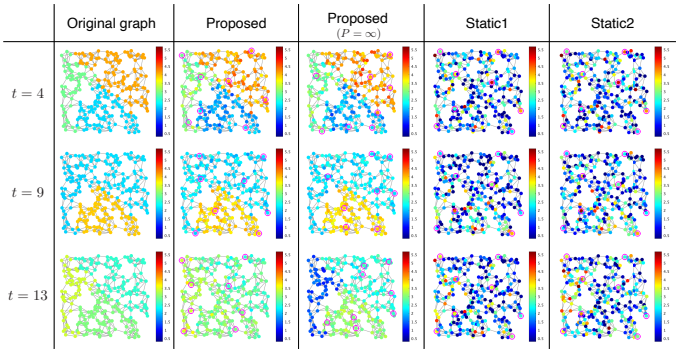


Fig. 4: Reconstructed graph signals for PC signals at time $t = 4, 9, 13$. Circled nodes represent selected sensor positions.

[3] Y. Jiang, J. Bigot, and S. Maabout, *Sensor Selection on Graphs via Data-driven Node Sub-sampling in Network Time Series*, 2020. arXiv: 2004.11815.

[4] A. Downey, C. Hu, and S. Laflamme, "Optimal sensor placement within a hybrid dense sensor network using an adaptive genetic algorithm with learning gene pool," *Struct. Health Monit.*, vol. 17, no. 3, pp. 450–460, 2018.

[5] J. Li, "Exploring the potential of utilizing unsupervised machine learning for urban drainage sensor placement under future rainfall uncertainty," *J. Environ. Manage.*, vol. 296, p. 113 191, 2021.

[6] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, 2008.

[7] L. M. Oliveira and J. J. Rodrigues, "Wireless Sensor Networks: A Survey on Environmental Monitoring.," *J. Commun.*, vol. 6, no. 2, pp. 143–151, 2011.

[8] A. Sakiyama, Y. Tanaka, T. Tanaka, and A. Ortega, "Efficient sensor position selection using graph signal sampling theory," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 6225–6229.

[9] B. Li, H. Liu, and R. Wang, "Efficient sensor placement for signal reconstruction based on recursive methods," *IEEE Trans. Signal Process.*, vol. 69, pp. 1885–1898, 2021.

[10] H. Zhou, X. Li, C.-Y. Cher, E. Kursun, H. Qian, and S.-C. Yao, "An information-theoretic framework for optimal temperature sensor allocation and full-chip thermal monitoring," in *DAC Design Automation Conference 2012*, IEEE, 2012, pp. 642–647.

[11] K. Visalini, B. Subathra, S. Srinivasan, G. Palmieri, K. Bekiroglu, and S. Thiyaku, "Sensor placement algorithm with range constraints for precision agriculture," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 34, no. 6, pp. 4–15, 2019.

[12] Y. Tanaka, Y. C. Eldar, A. Ortega, and G. Cheung, "Sampling signals on graphs: From theory to applications," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 14–30, 2020.

[13] J. Hara and Y. Tanaka, "Sampling set selection for graph signals under arbitrary signal priors," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 5732–5736.

[14] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, 2013.

[15] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6510–6523, 2015.

[16] M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo, "Signals on graphs: Uncertainty principle and sampling," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4845–4860, 2016.

[17] B. Jayaraman and S. M. Mamun, "On data-driven sparse sensing and linear estimation of fluid flows," *Sensors*, vol. 20, no. 13, p. 3752, 2020.

[18] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge University Press, 2015.

[19] Y. Tanaka and Y. C. Eldar, "Generalized sampling on graphs with subspace and smoothness priors," *IEEE Trans. Signal Process.*, vol. 68, pp. 2272–2286, 2020.

[20] J. Hara, Y. Tanaka, and Y. C. Eldar, "Graph signal sampling under stochastic priors," *IEEE Trans. Signal Process.*, 2023.

[21] D. I. Shuman, B. Ricaud, and P. Vandergheynst, "Vertex-frequency analysis on graphs," *Appl. Comput. Harmon. Anal.*, vol. 40, no. 2, pp. 260–291, 2016.

[22] L. Condat, "A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 460–479, 2013.

[23] F. Zhang, *The Schur Complement and Its Applications*. Springer Science & Business Media, 2006, vol. 4.

[24] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[25] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *1999 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1999, pp. 2443–2446.

[26] G. B. Passty, "Ergodic convergence to a zero of the sum of monotone operators in Hilbert space," *J. Math. Anal. Appl.*, vol. 72, no. 2, pp. 383–390, 1979.

[27] P. Tseng, "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities," *SIAM J. Control Optim.*, vol. 29, no. 1, pp. 119–138, 1991.

[28] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[29] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," en, *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[30] K. Zhang, M. Coutino, and E. Isufi, "Sampling graph signals with sparse dictionary representation," in *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, pp. 1815–1819.

[31] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 689–696.

[32] K. Manohar, B. W. Brunton, J. N. Kutz, and S. L. Brunton, "Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns," *IEEE Control Syst. Mag.*, vol. 38, no. 3, pp. 63–86, 2018.

[33] S. Chen, R. Varma, A. Singh, and J. Kovačević, "Representations of piecewise smooth signals on graphs," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 6370–6374.

[34] S. Nomura, J. Hara, H. Higashi, and Y. Tanaka, "Dynamic sensor placement based on sampling theory for graph signals," *IEEE Open Journal of Signal Processing*, 2024, in press.

Automatic Prompt Generation and Grounding Object Detection for Zero-Shot Image Anomaly Detection

Tsun-Hin Cheung*, Ka-Chun Fung*†, Songjiang Lai*†, Kwan-Ho Lin*, Vincent Ng*, and Kin-Man Lam*†
* Centre for Advances in Reliability and Safety, New Territories, Hong Kong
† Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong

***Abstract*—Identifying defects and anomalies in industrial products is a critical quality control task. Traditional manual inspection methods are slow, subjective, and error-prone. In this work, we propose a novel zero-shot training-free approach for automated industrial image anomaly detection using a multimodal machine learning pipeline, consisting of three foundation models. Our method first uses a large language model, i.e., GPT-3. generate text prompts describing the expected appearances of normal and abnormal products. We then use a grounding object detection model, called Grounding DINO, to locate the product in the image. Finally, we compare the cropped product image patches to the generated prompts using a zero-shot image-text matching model, called CLIP, to identify any anomalies. Our experiments on two datasets of industrial product images, namely MVTec-AD and VisA, demonstrate the effectiveness of this method, achieving high accuracy in detecting various types of defects and anomalies without the need for model training. Our proposed model enables efficient, scalable, and objective quality control in industrial manufacturing settings.**

***Index Terms*—Image Anomaly Detection, Prompt Generation, Object Localization, Multimodal Model**

I. Introduction

Ensuring product quality is a major challenge in industrial manufacturing. Identifying defects, flaws, and anomalies in produced goods is crucial to maintaining high standards and avoiding costly recalls or customer dissatisfaction. Traditionally, this quality control process has relied on manual visual inspection by trained human experts. However, this approach is inherently slow, subjective, and prone to human error [1].

Advancements in computer vision and machine learning have enabled the development of automated industrial inspection systems [2]. These systems can rapidly analyze images of products, detect defects, and classify anomalies in an objective and scalable manner. A key challenge in this domain is capturing the wide variability in the appearance of normal, defect-free products. Conventional approaches using supervised learning often require large, labeled datasets containing both normal and anomalous examples, which can be expensive and time-consuming to collect [3].

Recent approaches, such as CLIP [4] and WinCLIP [5], have achieved promising performance in zero-shot image anomaly detection. However, these CLIP-based methods utilize a set of fixed templates to generate the text prompts of normal and abnormal products, which require human experts and the generated text prompts are not class-specific. This results in limited performance in industrial image anomaly detection because defects are mostly object-specific. Moreover, WinCLIP [5] applies the sliding windows method to obtain multi-scale features. However, these methods often ignore the different sizes of industrial products.

Methods	Text Prompt Generation	Image Region Partition
CLIP [4]	Template-generated	Entire image
WinCLIP [5]	Template-generated	Entire image + sliding windows
Ours	**LLM-generated**	**Entire image + object detector**

TABLE I
COMPARISON OF TEXTUAL PROMPT GENERATION AND IMAGE FEATURE EXTRACTION OF OUR PROPOSED METHOD TO OTHER CLIP-BASED IMAGE ANOMALY DETECTION METHODS.

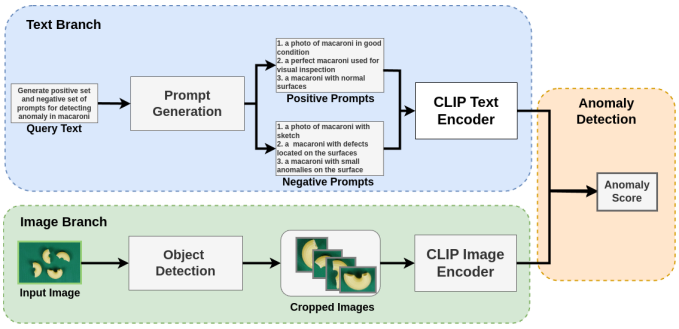


Fig. 1. Illustration of the prompt generation and object localisation-enhanced CLIP for zero-shot anomaly detection.

To address these two challenges, we propose a novel multimodal approach for industrial image anomaly detection that can address these limitations, as shown in Table I. Our method leverages large language models to automatically generate text descriptions of the expected appearance of the product. These prompts are then used in conjunction with an object detection model and a zero-shot image-text matching model to identify any deviations from the normal state. This approach enables efficient, accurate, and salable quality control without the need for extensive labeled training data. The illustration of our multimodal pipeline is shown in Fig. 1.

Our contributions are summarized as follows:

- We propose to use a language model, such as GPT-3 [6], to generate object-specific prompts for describing normal and anomaly industrial images, instead of relying on a set of prompts generated by fixed templates.
- We propose to use grounding DINO [7] to locate the objects in images, effectively suppressing the background noise and addressing multi-resolution challenges in zero-

shot image anomaly detection.
- Finally, we incorporate the two modules with pretrained CLIP [4] to perform image anomaly detection, outperforming the vanilla-CLIP and the current state-of-the-art (SOTA) baseline, i.e., WinCLIP [5], in zero-shot settings.

II. Related Work

A. Image Anomaly Detection

Automated image anomaly detection has been a long-standing challenge in computer vision, with applications in various domains such as industrial quality control, medical imaging, and security surveillance. Traditional approaches have often relied on hand-crafted feature extraction and unsupervised anomaly detection techniques, such as one-class support vector machines (OC-SVM) and isolation forests [8]. Recently, deep learning-based methods have shown promise in addressing the limitations of conventional anomaly detection algorithms. These approaches typically utilize convolutional neural networks (CNNs) or Transformer models to learn representations of normal data and identify deviations from this learned distribution [9]. However, a key challenge with these supervised methods is the requirement of labeled anomaly data, which can be costly and time-consuming to obtain in real-world scenarios.

CLIP (Contrastive Language-Image Pretraining) [4] is a deep learning model that is pretrained on a large-scale dataset of image-text pairs, enabling it to learn a joint representation of visual and textual data. This learned representation can then be leveraged for zero-shot image anomaly detection, where the model is not explicitly trained on anomalous samples but rather learns to identify deviations from the normal data distribution based on its general understanding of the visual world. WinCLIP [5] extends the CLIP approach by using a sliding window mechanism instead of analyzing the whole image. In this method, the model is trained to learn a joint representation between image patches and associated textual descriptions. During inference, the model evaluates the aggregated patches and identifies anomalies based on the discrepancy between the visual representation and the normal data distribution.

In this paper, we also consider the use of CLIP to generate anomaly scores for product images. However, our focus is on improving the quality of text prompts using large language models and enhancing the accurate localisation of products in a CLIP-based pipeline.

B. Prompt Generation with Large Language Models

The use of large language models, such as GPT-3 [6] and LLaMA [10], has revolutionized the field of natural language processing, enabling numerous applications, including text generation, question answering, and sentiment analysis [5, 6].

More recently, researchers have explored the potential of leveraging these powerful language models for tasks beyond pure text processing, such as multimodal learning [11] and AI agent reasoning [12]. Our application is slightly different from the language models used for automatic prompt generation [13] in multimodal modeling, where the language model is tasked with producing and enhancing the human written prompts. These language-guided prompts have proven effective in a variety of computer vision tasks, including image classification and object detection [14]. However, the application of prompt generation techniques to industrial image anomaly detection has not been extensively explored and presents a promising area for further research.

C. Grounding Object Detection

Object detection is a fundamental computer vision task that involves localizing and classifying objects within an image. Traditional object detection methods, such as Fast R-CNN [15] and YOLO [16], have relied on supervised learning approaches that require a large amount of annotated training data. Recent advancements in unsupervised and weakly-supervised object detection have aimed to reduce the need for extensive manual labeling. These include techniques that leverage saliency maps, image-level annotations, or self-supervised learning approaches to learn object representations [17].

Grounding object detection, also known as text-guided object detection, aims to address the limitations of traditional object detection by leveraging natural language descriptions to guide the detection process [18]. One pioneering work in this area is Referring Expression Comprehension (REC) [19], which focuses on localizing an object in an image based on a natural language description. REC models learn to align visual and textual representations to identify the referred object. Building upon REC, Grounded CLIP [20] and GroundingDINO [7] extend the grounding concept to generic object detection tasks. These methods utilize the joint representation learning of CLIP [4] to associate textual and visual information, enabling zero-shot or few-shot object detection without requiring extensive object-level annotations. In our work, we utilize the zero-shot capability of GroundingDINO to accurately locate the objects, which are then used for anomaly detection.

III. Methodology

Our proposed method leverages recent advancements in large language models, object detection, and zero-shot image-text matching to develop an efficient and scalable quality control system. The proposed method consists of three foundation models. The core idea is to first generate textual prompts describing the expected appearance of the product using language models, i.e., GPT-3 [6]. Then we use a grounding object detection model, namely Grounding DINO [7] to locate the objects in the input image, and finally compare the located product images to the generated prompts to compute anomaly scores using pretrained CLIP [4].

This multimodal approach allows us to combine the strengths of different deep learning techniques to address the challenges of industrial image anomaly detection. The language model-based prompt generation captures the rich domain knowledge required to define the expected product characteristics, while the object detection and image-text matching components enable robust and accurate anomaly identification.

A. Prompt Generation

In the text branch, the first step in our method is to generate two sets of textual prompts that describe the expected appearance of the "normal" and "anomaly" product. We utilize the pre-trained GPT-3 large language to generate these prompts.

For the text prompts, we provide the language model with relevant information about the product, including its category. The inputs to the language models to generate the "normal" prompt and "anomaly" prompts are denoted as x_{normal} and x_{anomaly}, respectively. The model then generates two sets of prompts that describe the desired appearance of the product, one for the normal class, and another for the anomaly class, as shown in Equations (1) and (2).

$$\mathbf{P}_{\text{normal}} = \text{GPT3}(x_{\text{normal}}), \tag{1}$$

$$\mathbf{P}_{\text{anomaly}} = \text{GPT3}(x_{\text{anomaly}}), \tag{2}$$

where, $\mathbf{P}_{\text{normal}}$ and $\mathbf{P}_{\text{anomaly}}$ are two sets of text prompts that describe normal and anomaly products, respectively.

B. Object Localization

The next step in our pipeline is to locate the product in the input image. We utilize an object detection model, Grounding DINO to identify the bounding boxes of the products within the image. This step ensures that subsequent anomaly detection is performed on the relevant regions of the image, rather than the entire frame solely.

To locate the product of interest within the input image, we leverage the Grounding DINO object detection model [7]. GroundingDINO is a powerful Transformer-based system that can precisely identify the bounding box coordinates of an object given only its class name as input. We feed the input image I and the class label c of the product into the GroundingDINO model, which then outputs the bounding box coordinates $b = [x, y, w, h]$ that specify the location of the product within the image. This allows us to extract the relevant image patches that contain the product for further analysis. Specifically, we crop the original image I based on the bounding box coordinates to obtain the cropped product image $\mathbf{I}_{\text{product}}$:

$$\mathbf{I}_{\text{object}} = \mathbf{I}[y : y + h, x : x + w] \tag{3}$$

As multiple bounding boxes can be located in the image, we repeat the above steps to obtain multiple image patches that contains the objects.

C. Zero-Shot Anomaly Detection

The final step of our method is to detect any anomalies or defects in the cropped product image. We employ a zero-shot image-text matching model, i.e., CLIP [4], to compare the cropped image to the textual prompts generated in the first step.

The image-text matching model is trained to learn a joint embedding space between images and their corresponding textual descriptions. This allows the model to assess the similarity between the input image and the generated prompts, detecting any significant discrepancies that may indicate the presence of an anomaly.

To detect anomalies in product images, we leverage the powerful CLIP,which encodes both text and image data into a shared embedding space, allowing us to compute meaningful similarities between text prompts and image content.

First, we compute the two text embeddings using CLIP - a "normal" prompt that describes the expected appearance of the product, and an "anomaly" prompt that describes potential defects or anomalies we want to detect. The two embedding vectors are denoted as $\mathbf{t}_{\text{normal}}$ and $\mathbf{t}_{\text{anomaly}}$, for "normal" and "anomaly" prompts respectively.

To compute the image embedding, for each cropped product image $\mathbf{I}_{\text{object}}$, we use the pretrained CLIP to obtain the average embedding vector of the object patches, denoted as $\mathbf{e}_{\text{object}}$. We also obtain the image embedding of the whole image as the global feature, denoted as $\mathbf{e}_{\text{image}}$. Then, the fused feature $\mathbf{e}_{\text{fused}}$ is obtained by averaging the global feature $\mathbf{e}_{\text{image}}$ and the local features $\mathbf{e}_{\text{object}}$.

Following the extraction of the CLIP embeddings for both normal and anomaly prompts, the next step involves determining the anomaly score for each product image. The anomaly score quantifies the deviation of an image from the expected norm, facilitating the identification of potential anomalies. The anomaly score (s) is defined as the following:

$$\text{s} = \frac{\mathbf{e}_{\text{fused}} \cdot \mathbf{t}_{\text{anomaly}}}{\mathbf{e}_{\text{fused}} \cdot \mathbf{t}_{\text{anomaly}} + \mathbf{e}_{\text{fused}} \cdot \mathbf{t}_{\text{normal}}} \tag{4}$$

In summary, our anomaly detection method exploits the semantic understanding capabilities of the CLIP model by associating textual prompts with image embeddings, thereby being enable to detect anomalies in product images based on predefined normal and anomaly criteria.

IV. Experiments

A. Experimental Details

We evaluate our proposed method on two industrial product image datasets:

MVTec-AD dataset [21]: The MVTec-AD dataset is a widely used benchmark for industrial image anomaly detection. It contains more than 5,000 high-resolution images of 15 different industrial product categories, anootated with various types of anomalies and defects.

VisA dataset [22]: The VisA dataset is another industrial product image dataset, containing more than 10,000 images of 20 different product categories. This dataset includes a diverse set of anomalies and defects commonly found in industrial manufacturing environments.

To quantify the performance of our proposed method, we employed two widely used evaluation metrics: Area Under the Receiver Operating Characteristic (AUROC) and Area Under the Precision-Recall Curve (AUPR). AUROC provides a comprehensive measure of the trade-off between the true

Methods	AUROC	AUPR
SPADE [23] (1-shot)	0.810	0.906
PaDiM [24] (1-shot)	0.766	0.881
PathCore [25] (1-shot)	0.834	0.922
WinCLIP [5] (0-shot)	0.918	0.965
Ours	**0.932**	**0.966**

TABLE II
COMPARISON OF ZERO-SHOT IMAGE ANOMALY DETECTION ON THE MVTEC-AD DATASET.

Methods	AUROC	AUPR
SPADE [23] (1-shot)	0.795	0.820
PaDiM [24] (1-shot)	0.628	0.683
PathCore [25] (1-shot)	0.799	0.828
WinCLIP [5] (0-shot)	0.781	0.812
Ours	**0.829**	**0.857**

TABLE III
COMPARISON OF ZERO-SHOT IMAGE ANOMALY DETECTION ON THE VISA DATASET.

positive rate and the false positive rate, capturing the overall classification performance. AUPR is a useful metric for evaluating the performance of anomaly detection models, as it takes into account the balance between precision and recall.

B. Comparison to State-of-the-Art Methods

Table II shows the comparison of our method with several state-of-the-art zero-shot and few-shot image anomaly detection methods on the MVTEC-AD dataset. Our method achieves an AUROC score of 93.2% and an AUPR score of 96.6%, outperforming the previous best zero-shot method, WinCLIP [5], by a significant margin of over 1 percentage point in both metrics.

The strong performance of our approach on MVTEC-AD demonstrates the effectiveness of our multimodal pipeline in capturing the rich domain knowledge required to accurately define the expected product characteristics and localize anomalies in a zero-shot setting. The language model-based prompt generation, coupled with robust object detection and zero-shot image-text matching components, allows our system to handle a wide variety of defect types, including surface defects, structural anomalies, and missing parts.

Table III presents the results on the VISA dataset, which is another challenging industrial anomaly detection benchmark. Our method achieves an AUROC score of 82.9% and an AUPR score of 85.7%, outperforming the current state-of-the-art zero-shot and few-shot methods by a notable margin.

The VISA dataset consists of a diverse set of industrial products, including electronics, automotive parts, and consumer goods, which require a more generalized method for anomaly detection. Our multimodal method is able to adapt to a wider range of product types and defect categories demonstrates its flexibility and robustness. The language model-based prompt generation, combined with the object detection and zero-shot image-text matching, allows our system to effectively capture the unique characteristics of each product and accurately identify anomalies across the diverse VISA dataset.

Methods	AUROC	AUPR
Ours w/o prompt generation	0.905	0.951
Ours w/o object detection	0.913	0.953
Ours	**0.932**	**0.966**

TABLE IV
ABLATION STUDY ON THE MVTEC-AD DATASET.

Methods	AUROC	AUPR
Ours w/o prompt generation	0.815	0.848
Ours w/o object detection	0.792	0.821
Ours	**0.829**	**0.857**

TABLE V
ABLATION STUDY ON THE VISA DATASET.

C. Ablation Studies

Table IV shows the results of the ablation studies on the MVTEC-AD dataset. We consider two variants of our method: one without the prompt generation component (Ours w/o prompt generation) and another without the object detection component (Ours w/o object detection).

The results demonstrate the importance of both the prompt generation and object detection components in our multimodal pipeline. When removing the object detection, the AUROC and AUPR scores drop by over 1 percentage point, indicating that the language model-based prompt generation is crucial for capturing the rich domain knowledge required for accurate anomaly detection.

Similarly, excluding the prompt generation component leads to a significant performance degradation, with the AUROC and AUPR scores decreasing by more than 2 percentage points. This highlights the importance of object-level features in localizing anomalies in industrial products.

The full multimodal method combines the prompt generation, object detection, and zero-shot image-text matching, achieving the best performance on the MVTEC-AD dataset, with an AUROC of 93.2% and an AUPR of 96.6%.

The ablation study results on the VISA dataset are presented in Table V. Similar to the MVTEC-AD analysis, we observe a substantial drop in performance when removing either the prompt generation or the object detection components.

Without the prompt generation, the AUROC and AUPR scores decrease by more than 1 percentage point, demonstrating the value of the language model-based domain knowledge in adapting to the diverse range of products and defect types in the VISA dataset.

The impact of the object detection component is even more pronounced, with the AUROC and AUPR scores dropping by over 3 percentage points when this module is removed. This underscores the critical role of the object-level features in accurately identifying anomalies in the complex VISA dataset. The full multimodal method, combining all the key components, achieves the best performance on the VISA dataset, with an AUROC of 82.9% and an AUPR of 85.7%.

The results of the ablation studies on both datasets highlight the complementary nature of the prompt generation, object

detection, and zero-shot image-text matching components in our multimodal method. Each module contributes significantly to the overall performance, and the synergistic integration of these elements is crucial for achieving state-of-the-art zero-shot industrial image anomaly detection capabilities.

V. Conclusion

In this work, we have proposed a comprehensive approach for industrial anomaly detection that leverages advancements in image anomaly detection, automatic prompt generation, and grounded object localization. By integrating these state-of-the-art techniques, we have developed a robust and versatile method capable of identifying and localizing defects or irregularities in industrial settings.

Our method first employs a deep learning-based anomaly detection model to identify deviations from the learned distribution of normal product samples. This is followed by the use of a large language model to automatically generate descriptive prompts that capture the salient features of the detected anomalies. Finally, we incorporate object detection to localize the regions of interest, providing valuable information for targeted analysis and root cause investigation.

The experimental results on a diverse dataset of industrial products have demonstrated the effectiveness of our method, achieving high accuracy in anomaly detection and precise localization of the detected issues. Furthermore, the language-guided prompts have shown potential for improved interpretability and explainability, which are crucial for industrial applications where the ability to understand and communicate the nature of detected anomalies is of paramount importance.

Moving forward, we aim to further explore the synergies between these technologies, investigating novel architectures and optimization techniques that can enhance the overall performance and robustness of the method. Additionally, we plan to extend the capabilities of our method to handle a wider range of industrial scenarios and explore integration with real-time monitoring and decision-making systems.

Overall, this work demonstrates the promising future of industrial anomaly detection, where the combination of advanced computer vision, natural language processing, and object localization can revolutionize the way we identify and address quality issues in industrial manufacturing and production processes.

References

[1] H. Sun, Y. Cao, and O. Fink, "Cut: A controllable, universal, and training-free visual anomaly generation framework," *arXiv preprint arXiv:2406.01078*, 2024.

[2] M. W. Hridoy, M. M. Rahman, and S. Sakib, "A framework for industrial inspection system using deep learning," *Annals of Data Science*, vol. 11, no. 2, pp. 445–478, 2024.

[3] A. Saberironaghi, J. Ren, and M. El-Gindy, "Defect detection methods for industrial products using deep learning techniques: A review," *Algorithms*, vol. 16, no. 2, p. 95, 2023.

[4] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[5] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 606–19 616.

[6] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[7] S. Liu, Z. Zeng, T. Ren, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[8] S. Goyal, A. Raghunathan, M. Jain, H. V. Simhadri, and P. Jain, "Drocc: Deep robust one-class classification," in *International conference on machine learning*, PMLR, 2020, pp. 3711–3721.

[9] A. S. Khan, Z. Ahmad, J. Abdullah, and F. Ahmad, "A spectrogram image-based network anomaly detection system using deep convolutional neural network," *IEEE access*, vol. 9, pp. 87 079–87 093, 2021.

[10] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[12] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] A. Lu, H. Zhang, Y. Zhang, X. Wang, and D. Yang, "Bounding the capabilities of large language models in open text generation with prompt constraints," *arXiv preprint arXiv:2302.09185*, 2023.

[14] W. He, W. Chen, B. Chen, *et al.*, "Unsupervised prompt tuning for text-driven object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2651–2661.

[15] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[17] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5866–5885, 2021.

[18] A. Sadhu, K. Chen, and R. Nevatia, "Zero-shot grounding of objects from natural language queries," in *Pro-*

ceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4694–4703.

[19] S. KAZEMZADE, V. Ordonez, M. MATTENV, *et al.*, "Referring to objects in photographs of natural scenes," in *Empirical Methods in Natural Language Processing*, vol. 28, 2014, pp. 787–789.

[20] L. Xiao, X. Yang, F. Peng, M. Yan, Y. Wang, and C. Xu, "Clip-vg: Self-paced curriculum adapting of clip for visual grounding," *IEEE Transactions on Multimedia*, 2023.

[21] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9592–9600.

[22] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *European Conference on Computer Vision*, Springer, 2022, pp. 392–408.

[23] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," *arXiv preprint arXiv:2005.02357*, 2020.

[24] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: A patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition*, Springer, 2021, pp. 475–489.

[25] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 318–14 328.

A Non-Intrusive Speech Quality Assessment Model using Whisper and Multi-Head Attention

Guojian Lin*, Yu Tsao† and Fei Chen*
* Southern University of Science and Technology, Shenzhen, China
E-mail: 12432635@mail.sustech.edu.cn, fchen@sustech.edu.cn Tel/Fax: +86-755-88018554
† Academia Sinica, Taiwan
E-mail: yu.tsao@citi.sinica.edu.tw Tel/Fax: +886-2-2787-2390

***Abstract*—Speech quality assessment serves as an important tool for speech related applications. In this study, we propose a non-intrusive model QUAL-Net, which is able to estimate subjective quality scores of the target speech. QUAL-Net combines acoustic features extracted by a large-scale model Whisper with spectral features and time-domain waveform features. Furthermore, QUAL-Net employs a CNN-BiLSTM-Attention architecture and introduces multi-head attention mechanism into attention layer to enhance model's performance. Experimental results demonstrate that Whisper embedding features have more powerful speech quality characterization ability than other self-supervised learning (SSL) embedding features. Additionally, the feature combination utilizing all three types of acoustic features obtains optimal improvement in model performance. Moreover, the results prove that multi-head attention has potential to capture more key information from acoustic features than multiplicative self-attention. We tested QUAL-Net's performance on the noisy and enhanced track of VoiceMOS Challenge 2023. Compared with MOSA-Net and other speech quality assessment models, QUAL-Net achieves significant improvement when it is trained to estimate subjective quality scores. QUAL-Net outperforms the top-ranked MOSA-Net+ in all evaluation metrics. QUAL-Net uses a simpler CNN architecture compared to the MOSA-Net+, contributing to reduction of the model complexity.**

I. Introduction

In real life, accurate speech quality assessment is of great significance to the development of speech-related applications, such as speech enhancement, speech synthesis and hearing aids. Listening test based on listeners is acknowledged as the most direct and accurate method to evaluate speech quality. The mean opinion score (MOS) is the most widely used evaluation metric of speech quality in subjective listening test, ranging from one to five. However, listening test is time-consuming and costly, and requires specific listening environments. Due to the limitation of subjective listening test, objective metrics for speech quality assessment have been proposed, such as perceptual evaluation of speech quality (PESQ) [1], perceptual objective listening quality analysis (POLQA) [2], signal-to-distortion ratio (SDR) [3] and hearing aid speech quality index (HASQI) [4]. However, these metrics based on signal processing algorithm need clean speech with the same frequency and time length as reference.

In recent years, researchers have developed non-intrusive speech quality assessment models based on deep learning. By learning features from a large amount of speech data, these models are able to accurately predict speech quality scores without clean reference. Deep learning-based methods can be divided into two types on the target evaluation metrics. The first type is to predict objective evaluation metrics. Quality-Net [5] uses the network BiLSTM to predict scores of PESQ. STOI-Net [6] employs a CNN-BiLSTM architecture with attention to predict scores of objective speech intelligibility. AMSA [7] utilizes reference-less multi-task learning (MTL) framework to predict multiple objective speech quality and intelligibility scores. The second type focus on predicting scores from subjective listening test. MOSNet [8], a CNN-BiLSTM based model, is proposed to estimate quality of the converted speech. MBNet [9] uses two networks to seprately predict the mean quality score of an utterance and the difference between the mean score and listener score, respectively. LDNet [10] employs raw speech and listener-dependency information of listeners as input to the model for MOS prediction. In VoiceMOS Challenge 2022 [11], which aims to encourage the research on development of MOS predictor for synthesized speech, numerous innovative systems [12], [13], [14], [15] based on self-supervised learning (SSL) models have been proposed, yielding great improvement in the performance of MOS prediction.

Whisper [16], a large-scale pre-trained model, has demonstrated its advanced performance and powerful generalization ability in various speech processing tasks. Zezario et al. [17] utilizes Whisper to extract phonetic embedding representations to assess HASPI scores for hearing aids, contributing to an improvement of approximately 30% in the ranking correlation between predicted scores and actual scores compared to the baseline model using WavLM [18] embedding features. It is more difficult to obtain MOS from human listening test than quality scores from objective evaluation methods, which results in the limited amount of labeled speech data for training. Therefore, it is necessary to consider the importance of each speech frame's information to ensure accurate evaluation. For example, there are quiet segments in speech frames that contain redundant information, which should receive less attention. More attention should be paid to the frames that contain more useful information. However, self-attention mechanism overly focuses on its own location information when encoding features, ignoring the importance of other location information. In contrast, the multi-head attention mechanism maps feature information to multiple subspaces to compute location weights

in parallel, enabling the model to focus on different subspace information at different locations, thus capturing more effective information. Liang et al. [19] utilizes multi-head attention mechanism to design a non-intrusive speech quality evaluation model for hearing aids. The value of the Pearson Correlation Coefficient (PCC), which describes the correlation between the predicted quality scores and the actual quality scores, is improved from 0.943 to 0.985.

Based on CNN-BiLSTM with multiplicative self-attention, MOSA-Net [20] combines cross-domain features with embedding representations from SSL model to evaluate quality and intelligibility of the noisy speech. In VoiceMOS Challenge 2022, MOSA-Net has achieved performance close to that of the baseline system [11]. We tested the performance of MOSA-Net using noisy and enhanced speech dataset from track 3 of VoiceMOS Challenge 2023 [21] and found its prediction accuracy can be further improved. Based on the architecture of MOSA-Net, this work proposes a novel model QUAL-Net for speech quality evaluation. SSL pre-trained models are replaced with Whisper to extract acoustic embedding features. QUAL-Net employs multi-head attention mechanism instead of multiplicative self-attention. QUAL-Net also utilizes a simpler CNN architecture with less convolution layers than MOSA-Net. In the experiments, we first compare the performance of Whisper-based features with other four SSL-based features. Then, we investigate the effect of different speech feature combinations on the accuracy of quality prediction. Subsequently, we analyze the impact of multi-attention and compare it with other attention mechanism. Finally, we compare QUAL-Net's performance on noisy and enhanced speech dataset with other systems.

II. QUAL-NET

The overall architecture of the proposed QUAL-Net is presented in Fig. 1. QUAL-Net is composed of the feature extraction module and the quality prediction module. The specific parameters of the model are described in Table I. In the feature extraction module, given a speech waveform $X = [x_1, x_2, \ldots, x_n, \ldots, x_N]$, the model takes three input branches. In the first branch, X is converted by 512-point STFT (Short Time Fourier Transformation) with a Hamming window of 32 ms and a hop of 16 ms to obtain a 257-dimensional spectrogram. In the second branch, X is fed into SincNet [22], a convolution network based on "Sinc" function, with filter dimension of 257. The output of the SincNet is a 257-dimensional filtered time-domain waveform, namely learnable filter banks (LFB) features. Subsequently, spectral features and LFB features are fed into CNN. CNN has five convolution networks, with a two-dimensional convolution layer, a batch normalization layer, a ReLU activation function and a power average pooling layer in each network. The pooling layers calculate the p-th root of the p-th power sum of X in the moving window. The calculation process is described as follows, where the value of p is 4:

$$f(x) = \left(\sum_{x \in X} x^p \right)^{\frac{1}{p}} \tag{1}$$

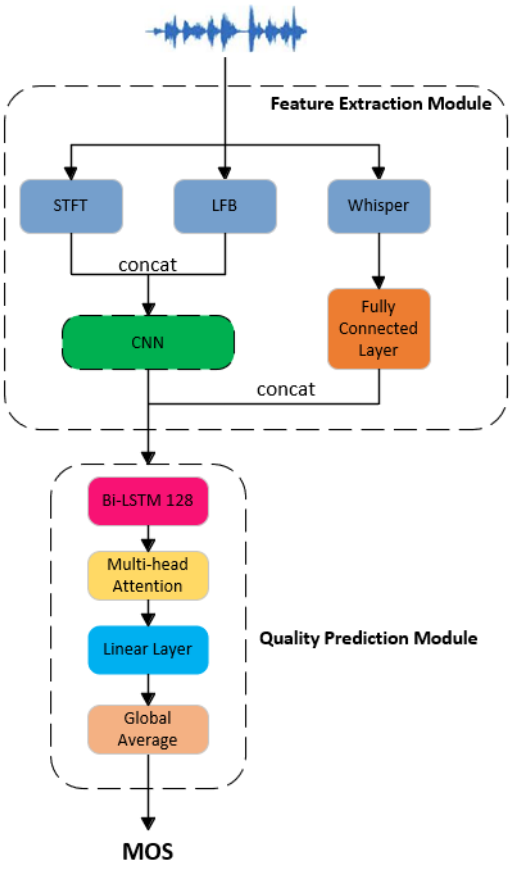


Fig. 1. Architecture of the proposed QUAL-Net

In the third branch, X is first transformed into a spectral signal by STFT, the frequency axis is divided into a series of Mel frequency bands, and then the energy within each Mel frequency band is summed and logarithmized to obtain the log mel-spectrogram M. Then, M is fed into Whisper pre-trained model to obtain 1024-dimensional Whisper embedding features WE. A fully connected layer is utilized to reduce the dimension of WE to 512 to ensure WE have the same feature dimension as deep frame-level features extracted by CNN. Subsequently, WE are concatenated with deep frame-level features. The combined features are mapped into quality prediction module to predict quality scores.

Quality prediction module consists of a BiLSTM, a multi-head attention layer, a linear layer and an adaptive average pooling function. The input features are first processed by BiLSTM with 128 nodes. BiLSTM comprises a forward LSTM and a backward LSTM. It models temporal information of feature sequences using backward and forward propagation, contributing to effective leveraging context information of long time sequences. BiLSTM is utilized to process combined features frame-by-frame and capture contextual dependencies in acoustic features. Multi-head attention layer with 8 attention heads is applied to learn different attention weights based on correlations within combined features, enabling the model to focus on more effective feature information to ensure accurate

TABLE I
PARAMETERS OF THE QUAL-NET

Name	Layer	Parameter	Size of output
CNN	Conv1	32,3×3	32×N×257
	Batch Normaliztion	32	32×N×257
	ReLU	/	32×N×257
	LPPool2d	4,1×4	32×N×64
	Conv2	32,3×3	32×N×64
	Conv3	64,3×3	64×N×64
	Batch Normaliztion	64	64×N×64
	ReLU	/	64×N×64
	LPPool2d	4,1×4	64×N×16
	Conv4	64,3×3	64×N×16
	Conv5	128,3×3	128×N×16
	Batch Normaliztion	128	128×N×16
	ReLU	/	128×N×16
	LPPool2d	4,1×4	128×N×4
	Reshape	/	N×512
Score Prediction Module	BiLSTM	128	N×256
	Dense	128	N×128
	Multi-Head Attention	128,heads=8	N×128
	Dense	1	1×128
	Global Average	1	1×1

quality prediction. Then, a linear layer with one node is utilized to generate frame-level scores. The number of frames of a speech utterance is the sum of frame number of the deep frame-level features and WE. The output of linear layer is processed through global average pooling to calculate the final quality score.

Moreover, the model integrates both frame-level quality scores and utterance-level quality scores into the loss function for training. The loss function is described as follows:

$$L_{quality} = \frac{1}{N}\sum_{n=1}^{N}\left[\left(Q_n - \hat{Q}_n\right)^2 + \frac{\alpha_Q}{F_u}\sum_{l=1}^{F_u}\left(Q_n - \hat{q}_{nl}\right)^2\right] \quad (2)$$

where Q_n represents actual Quality scores of the n-th training utterance. $\hat{Q}_n$ represents the predicted Quality scores of the n-th training utterance. The total number of training utterances is denoted by N. F_u represents the total number of frames in the n-th training utterance, which is the frame number of combined feature. $\hat{q}_{nl}$ is the predicted frame-level scores of the l-th frame of the n-th training utterance. For the n-th training utterance, there are F_u predicted frame-level scores. The weights between utterance-level and frame-level losses are determined by α_Q.

III. EXPERIMENTS

A. Dataset

The dataset in this experiment is from the noisy and enhanced track of VoiceMOS Challenge 2023 [21]. Training data is based on TMHINT-QI [23] dataset, a Mandarin corpus containing 24,408 ten-word utterances. It contains totally 8201 samples, including 360 clean speech samples, 1874 noisy speech samples with four types of noises (babble, street, pink, and white) at four signal-to-noise ratio (SNR) levels (-2, 0, 2, and 5) and 5967 enhanced speech samples derived from five speech enhancement systems: KLT, MMSE, FCN, DDAE, and Transformer. 226 listeners were recruited to take the listening test and predicted speech quality scores on a range from 1 to 5. The mean of the subjective quality scores for each utterance was used as the ground-truth score of the speech.

A seperate dataset TMHINT-QI (S) was created as test set, containing 360 noisy samples with the same noise type as those in training data and 1600 enhanced samples processed by five speech enhancement models: MMSE, FCN, Trans, DEMUCS, and CMGAN, including two unseen enhancement systems. A total of 110 listeners were recruited for listening test of the test set.

B. Model Training

In the training phase, 90% of training data were used for training and 10% for validation. We used Adam optimizer with initial learning rate of 0.001 and adopted a dynamic learning rate adjustment strategy. If the validation loss did not decrease after 10 training iteration rounds, the learning rate of Adam optimizer decreased by a factor of 10 with the minimum learning rate of 0.000001 to help the model find the optimal parameters better as well as to avoid overfitting. The training process involved totally 50 epochs and the sampling rates of all the speech are 16 kHz.

C. Evaluation Metrics

To evaluate the performance of the model, system-level and utterance-level Mean Squared Error (MSE), Linear Correlation Coefficient (LCC) and Spearman Rank Correlation Coefficient (SRCC) are used. MSE indicates the difference between predicted scores and actual quality scores. LCC describes the linear correlation between predicted scores and actual scores. SRCC represents the rank correlation between predicted scores and actual scores. Since it is more useful for quality assessment models to predict the ranks of systems accurately than to predict actual quality scores, we use system-level SRCC as the primary evaluation metric for model performance.

D. Comparison of Different Speech Embedding Features

In the first experiment, we aim to compare different speech pre-trained models and select the optimal pre-trained model. Five pre-trained models, BEATs [24], Hubert [25], XLSR [26], WavLM, and Whisper are employed to extract speech embedding features, where the feature dimensions extracted by the Hubert, XLSR, WavLM, and Whisper models are 1024, while the feature dimension of the BEATs model is 768.

As presented in Table II, Hubert, WavLM, and Whisper embedding features have respectively high correlation with speech quality. The models using BEATs and XLSR embedding features achieve low prediction accuracy, particularly

TABLE II
PERFORMANCE OF QUAL-NET WITH DIFFERENT SPEECH EMBEDDING FEATURES

Embedding Feature	System-level			Utterance-level		
	LCC	SRCC	MSE	LCC	SRCC	MSE
BEATs [24]	0.584	0.582	0.792	0.537	0.505	1.090
Hubert [25]	0.961	0.955	0.067	0.793	0.753	0.352
XLSR [26]	0.888	0.883	0.257	0.721	0.693	0.586
WavLM [18]	**0.964**	0.962	0.063	0.794	0.757	0.348
Whisper [16]	0.961	**0.966**	**0.053**	**0.807**	**0.780**	**0.323**

TABLE III
PERFORMANCE OF QUAL-NET WITH DIFFERENT FEATURES COMBINATIONS

Combination	Feature	System-level			Utterance-level		
		LCC	SRCC	MSE	LCC	SRCC	MSE
1	STFT+LFB	0.808	0.796	0.227	0.638	0.572	0.599
2	WE	0.926	0.921	0.106	0.786	0.752	0.374
3	STFT+WE	0.954	0.954	0.058	**0.808**	0.777	**0.323**
4	LFB+WE	0.950	0.949	0.066	0.803	0.773	0.327
5	STFT+LFB+WE	**0.961**	**0.966**	**0.053**	0.807	**0.780**	**0.323**

BEATs embedding features show significantly weaker correlation with speech quality than other features. The model with Whisper embedding features achieves the best performance in all evaluation metrics except for the system-level LCC, which is slightly lower than that of the WavLM. Based on training data amount of labeled speech of 680,000 hours and training parameters of large scale, Whisper reveals its robustness in audio feature extraction. Experimental results demonstrate the advantage of utilizing Whisper to deploy quality prediction model.

E. Comparison of Different Feature Combinations

In the second experiment, we investigate the effect of different acoustic features on the prediction performance through ablation experiment. As shown in Table III, five different feature combinations are employed, where STFT represents spectral features, LFB represents waveform features after SincNet filter processing and WE represents the embedding features extracted by Whisper. Comparing combination 1, combination 2 and combination 5, WE significantly outperforms STFT+LFB in both system-level and utterance-level evaluations, indicating that the Whisper embedding features play a major role in quality prediction. Comparing combination 3 and combination 4, with the same embedding features, model using STFT and WE has slightly better performance than that of using LFB and WE. It indicates that although LFB features retain the raw waveform more completely, STFT features which retain the speech phase and amplitude can capture more useful phonetics information. The performances of combination 3 and combination 4 have been very close to that of combination 5. It further demonstrates that compared with WE, both STFT and LFB features do not have a large magnitude of enhancement on prediction performance of the model.

Combination 5 achieves the most accurate quality prediction, demonstrating the effectiveness of combining different acoustic features for speech quality estimation.

F. Impact of Multi-Head Attention

In the third experiment, to investigate the impact of multi-head attention in the model, we compare QUAL-Net (*No-ATT*), which does not have attention layer with QUAL-Net (*Mul-ATT*) and QUAL-Net (*Multi-Head-ATT*). It needs to be clarified that Mul-ATT denotes that the model uses the self-attention layer which employs multiplicative attention and Multi-Head-ATT denotes that the model uses multi-head attention layer.

As present in Table IV, the models with an attention layer achieve more accurate quality prediction than the model without that. Compared with QUAL-Net (*No-ATT*), QUAL-Net (*Multi-Head-ATT*) improves on system-level SRCC by 0.051. Additionally, QUAL-Net (*Multi-Head-ATT*) outperforms QUAL-Net (*Mul-ATT*) in all evaluation metrics, improving on system-level SRCC by 0.022. It confirms that multi-head attention enables the model to focus on more useful information in time frames and effectively capture contextual association of speech, thus enhancing model's performance. Multi-head attention also exhibits superior ability to process frame-level feature of speech over multiplicative self-attention.

G. Comparison with Other Systems

In this experiment, we compare QUAL-Net with other speech quality evaluation models on the noisy and enhanced speech dataset of VoiceMOS Challenge 2023. Table V exhibits the performance of different systems. MOS-SSL [12] uses finetuned SSL models to predict MOS and LE-SSL-MOS [27] constructs a MOS predictor based on SSL models and the scores of individual listener augmentation branches, introducing new unsupervised metrics to improve prediction accuracy. UTMOS [13] builds strong and weak learners based on SSL models and classical machine learning algorithms to predict MOS. It is noted that MOSA-Net serves as the baseline model of our system, and its enhanced version MOSA-Net+ [28] achieves top-ranked performance on the noisy and enhanced track of VoiceMOS Challenge 2023.

TABLE IV
PERFORMANCE OF QUAL-NET WITH DIFFERENT ATTENTION LAYER

Model	System-level			Utterance-level		
	LCC	SRCC	MSE	LCC	SRCC	MSE
QUAL-Net (*No-ATT*)	0.920	0.915	0.243	0.771	0.742	0.521
QUAL-Net (*Mul-ATT*)	0.947	0.944	0.179	0.786	0.758	0.454
QUAL-Net (*Multi-Head-ATT*)	**0.961**	**0.966**	**0.053**	**0.807**	**0.780**	**0.323**

TABLE V
PERFORMANCE OF ALL SYSTEMS ON ENHANCED AND NOISY SPEECH TRACK OF VOICEMOS CHALLENGE 2023

System	System-level			Utterance-level		
	LCC	SRCC	MSE	LCC	SRCC	MSE
MOS-SSL [12]	0.637	0.487	2.986	0.518	0.403	3.356
UTMOS [13]	0.769	0.621	1.763	0.611	0.477	2.216
LE-SSL-MOS [27]	0.769	0.749	0.635	0.684	0.636	0.688
MOSA-Net [20]	0.954	0.941	0.067	0.781	0.749	0.358
MOSA-Net+ [28]	0.952	0.956	0.082	0.803	**0.780**	0.343
QUAL-Net	**0.961**	**0.966**	**0.053**	**0.807**	**0.780**	**0.323**

Our system QUAL-Net yields notable improvement in the performance of MOS prediction, outperforming all the compared systems including the top-ranked model MOSA-Net+ [28]. Compared with MOSA-Net, our system improves system-level SRCC from 0.941 to 0.966. In comparison to MOSA-Net+ [28], our system improves system-level SRCC from 0.956 to 0.966. It demonstrates the advantage of using Whisper to create multi-domain acoustic features and adopting multi-head attention mechanism for enhancing model's speech quality evaluation performance. It is worth noting that although UTMOS achieves top-ranked performance in VoiceMOS Challenge 2022, its performance on the noisy and enhanced dataset is much lower than that of QUAL-Net and MOSA-Net, which to some extent reflects the difficulty in realizing the quality assessment of multi-domain speech dataset.

IV. CONCLUSIONS

This study proposes a non-intrusive speech quality assessment model QUAL-Net. QUAL-Net incorporates pre-trained model Whisper into feature extraction and introduces multi-head attention mechanism to quality prediction. Experimental results demonstrate that Whisper embedding features can capture more effective acoustic information compared to other SSL-based features. Additionally, ablation experiment confirms that Whisper embedding features play a major role in speech quality evaluation. The model using spectral features, time-domain waveform features and Whisper features achieves the optimal performance. Furthermore, multi-head attention has superior performance over multiplicative self-attention and further improves prediction accuracy of the model. QUAL-Net achieves a notable improvement over MOSA-Net and outperforms MOSA-Net+ [28] and other speech quality assessment models. Since QUAL-Net uses CNN architecture with less number of convolution layers than MOSA-Net+ [28], it achieves better performance with simpler model structure. In future work, we plan to test the performance of QUAL-Net on different types of datasets and explore QUAL-Net's potential for semi-supervised speech quality assessment.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62371217). Part of this study was the basis for the Bachelor's thesis of the first author (G.J.L.).

REFERENCES

[1] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, IEEE, vol. 2, 2001, pp. 749–752.

[2] J. G. Beerends, C. Schmidmer, J. Berger, *et al.*, "Perceptual objective listening quality assessment (POLQA), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[3] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[4] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[5] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," *arXiv preprint arXiv:1808.05344*, 2018.

[6] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "STOI-Net: A deep learning based non-intrusive speech intelligibility assessment model," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2020, pp. 482–486.

[7] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 911–915.

[8] C.-C. Lo, S.-W. Fu, W.-C. Huang, *et al.*, "MOSNet: Deep learning based objective assessment for voice conversion," *arXiv preprint arXiv:1904.08352*, 2019.

[9] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNet: MOS prediction for synthesized speech with mean-bias network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 391–395.

[10] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 896–900.

[11] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4536–4540.

[12] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 8442–8446.

[13] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4521–4525.

[14] W.-C. Tseng, W.-T. Kao, and H.-y. Lee, "DDOS: A MOS prediction framework utilizing domain adaptive pre-training and distribution of opinion scores," in *Proc. Interspeech 2022*, 2022, pp. 4541–4545.

[15] Z. Yang, W. Zhou, C. Chu, *et al.*, "Fusion of self-supervised learned models for MOS prediction," in *Proc. Interspeech 2022*, 2022, pp. 5443–5447.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, PMLR, 2023, pp. 28 492–28 518.

[17] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Utilizing Whisper to enhance multi-branched speech intelligibility prediction model for hearing aids," *arXiv preprint arXiv:2309.09548*, 2023.

[18] S. Chen, C. Wang, Z. Chen, *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[19] R. Liang, Y. Xie, J. Cheng, C. Pang, and B. Schuller, "A non-invasive speech quality evaluation algorithm for hearing aids with multi-head self-attention and audiogram-based features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2166–2176, 2024.

[20] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2022.

[21] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–7.

[22] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 1021–1028.

[23] Y.-W. Chen and Y. Tsao, "InQSS: A speech intelligibility and quality assessment model using a multi-task learning network," in *Proc. Interspeech 2022*, 2022, pp. 3088–3092.

[24] S. Chen, Y. Wu, C. Wang, *et al.*, "BEATs: Audio pre-Training with acoustic tokenizers," in *International Conference on Machine Learning*, PMLR, 2023, pp. 5178–5193.

[25] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[26] A. Babu, C. Wang, A. Tjandra, *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[27] Z. Qi, X. Hu, W. Zhou, *et al.*, "LE-SSL-MOS: Self-supervised learning MOS prediction with listener enhancement," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–6.

[28] R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang, and C.-S. Fuh, "A study on incorporating Whisper for robust speech assessment," *arXiv preprint arXiv:2309.12766*, 2023.

Development of Simple Algorithm to Detect and Filter Motion Artifact Noise in Non Invasive Blood Pressure (NIBP) Measurement

Trio Adiono*, Rd Elviana La'salina Muhlis*, Clarence Amadeus*, Sindy Novaria Cicilya Sinaga†
*School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia
†PT. Xirka Dama Persada, Indonesia

***Abstract*—Motion artifact is a noise induced by patient during Non Invasive Blood Pressure (NIBP) measurement due to the patient's intentional or unintentional movements. It can reduce the accuracy of the measurement. In this paper, we will conduct a study to identify, detect, and reduce the motion artifact effect to the NIBP measurement. MATLAB will be used to analyze the measurement signal, detecting and reducing the motion artifact present on the data. Our experiment shows that measurement data corrupted by motion artifacts has larger peaks in their frequency response. To reduce this noise, the noisy measurement signal is fit into a straight line, then the noisy measurement signal is subtracted with this straight line. The method has decreased Mean Arterial Pressure (MAP) error by 67.48%, Systolic Blood Pressure (SBP) error by 39.28%, and Diastolic Blood Pressure (DBP) error by 72.22%.**

***Keywords*—** motion artifact; NIBP; oscillometric method

1. Introduction

Blood pressure is an important parameter for human health. This is because it can lead to various diseases. Low blood pressure can cause hypotension, while high blood pressure can cause hypertension. Low blood pressure can cause extreme fatigue, shortness of breath, and lightheadedness, while high blood pressure can increase the probability cardiovascular disease such as stroke, dissecting aneurysm, and heart attack, renal failure, and retinopathy [1]. Based on information provided by the Ministry of Health of the Republic of Indonesia, the higher someone's blood pressure, the higher the risk of heart disease, kidney failure, and stroke [2].

The movements performed by patients during measurements can affect the accuracy of Non-Invasive Blood Pressure (NIBP) readings [3]. Generally, these movements can cause higher Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) values than they should be [4]. The movements that occur during the measurements can be caused by environmental factors, such as when patients are being transferred, or by movements made by the patients themselves, such as tremors, arms flexing, palm rotation, and body movements [5]. Therefore, an algorithm is needed in automatic blood pressure measuring devices that can detect and eliminate motion artifacts, ensuring the obtained values are accurate.

Based on the above background, the authors will conduct a study on the influence of motion artifacts on oscillometric signals and methods to identify motion from the existing signal processing. Furthermore, the authors will develop a filter to mitigate the impact of motion artifacts on the measurement results, with the limitation that the equipment used is an automated non-invasive blood pressure measurement device utilizing the oscillometric method with a valve leak. The focus of the identification and detection process will be on the filter algorithm used and data processing. The authors will not make any modifications or additions to the device components, including the pump operation or adding motion detection sensors.

2. Previous Work

The basic filtering technique is generally effective in eliminating high-frequency noise. However, it is less effective in suppressing motion artifacts and arrhythmias. This is because both of these phenomena occur at the same frequency range as the data of Arterial Branchial Pressure(ABP) [6]. In the research [7], the Kalman filter is used to predict the pressure when the cuff is deflating. Furthermore, the Empirical decomposition mode is employed to separate the frequency of data and artifacts in searching for PPG data (Photoplethysmogram) [8]. Additionally, an accelerometer sensor is utilized to eliminate motion artifacts during ambulance transportation. The use of this sensor can estimate the magnitude of pressure caused by movement so that the uncorrupted pressure value is the initial pressure minus the pressure due to movement [9].

3. Proposed Method

The Oscillometric method is similar to the auscultation method, with the only difference being that the determination of SBP and DBP is not based on Korotkoff sounds [10] but rather on pressure data [11]. This algorithm makes blood pressure simpler and remove the requirement of expert [12]. First, a cuff wrapped around the arm is inflated with air until the pressure exceeds the SBP pressure, stopping the blood flow. Then, the air in the cuff is slowly released, reducing the pressure. When the pressure in the cuff drops below the peak arterial pulse pressure, the machine starts detecting small pressure waves that reflect the difference between the pressure in the cuff and the pressure in the artery. With further deflation of the cuff, this pressure difference becomes more prominent until the cuff starts detaching from the body part, and the detected pulse volume decreases. Therefore, the machine records a series of pulse waves within it, which start flat, then become very small, rise to a peak, and then decrease until they are almost undetectable [6]. An illustration of this algorithm can be seen in Figure 1. As seen in Figure 1, this algorithm takes the deflation raw cuff pressure data read by the pressure sensor as input. From this input, the following steps are performed.

Converting raw pressure data into oscillometric signals requires signal processing. An example of such raw pressure data can be seen in Figure 2. Common signal processing techniques used in this method involve the use of filters and detrending. One

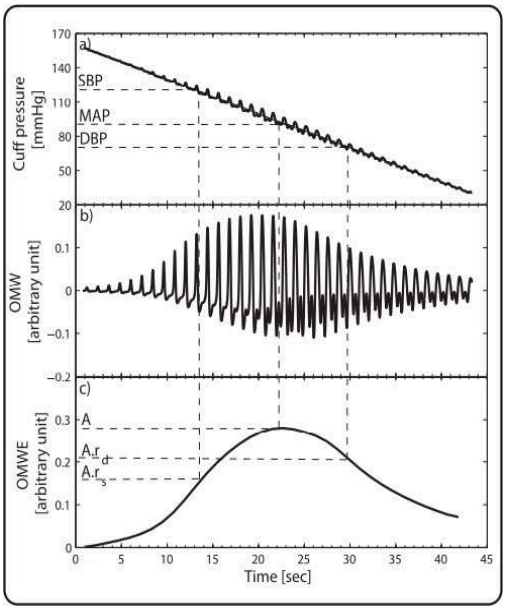


Figure 1. An Illustration of The Oscillometric Method Algorithm for SBP and DBP Measurement

of the commonly used filters is the bandpass filter, which is a combination of lowpass and highpass filter. The lower cut-off frequency value is usually between 0.1 to 0.5 Hz, while the upper cut-off frequency value is 20 Hz [6]. Detrending is performed to remove the downward trend in the raw data curve, making the curve more horizontally flat. An example of filtered signal can be seen in Figure 3. In Figure 3, lowpass filter will extract the decreasing cuff pressure signal called Cuff Pressure Waveform(CPW), while the highpass filter will extract the oscillometric waveform(OMW).

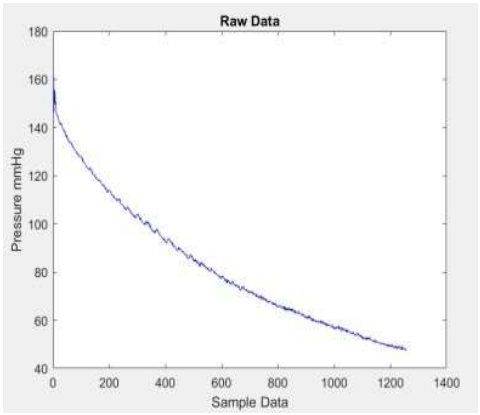


Figure 2. An Example of Raw Pressure Data

After obtaining the OMW, the process continues with estimating blood pressure values. The most popular estimation process is using the Maximum Amplitude Algorithm (MAA) to estimate Mean Arterial Pressure (MAP). The obtained oscillation signal is used to create an envelope, which can be found using indices from that oscillation signal. The highest amplitude of this envelope is taken as the MAP value. An example of this process can be seen in Figure 4.

Next, the SBP and DBP values are obtained by multiplying ratios relative to the MAP value in the envelope, with ratios used ranging from 0.5 to 0.8. More specific details are 0.45 - 0.73

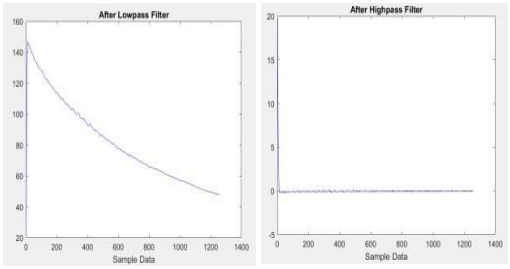


Figure 3. The Output of Lowpass and Highpass Filter

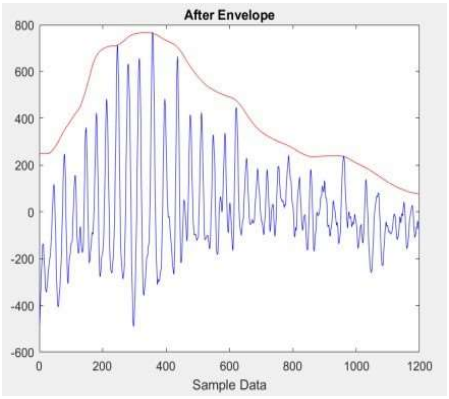


Figure 4. An Example of Enveloping the OMW

for systolic and 0.69 - 0.83 for diastolic. The indices of the SBP and DBP values on the envelope graph are then matched with the corresponding values in the raw data[6]. An example of this step can be seen in Figure5.

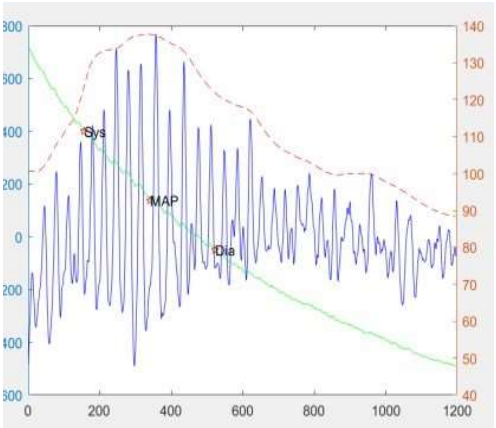


Figure 5. Estimation of MAP, SBP, and DBP

The frequency response of the oscillometric signal is also shown in Fig6.

3.1. Parameter Selection

In this paper, oscillometric based algorithm is used to measure blood pressure. The parameters in the algorithm is shown in Table 1.

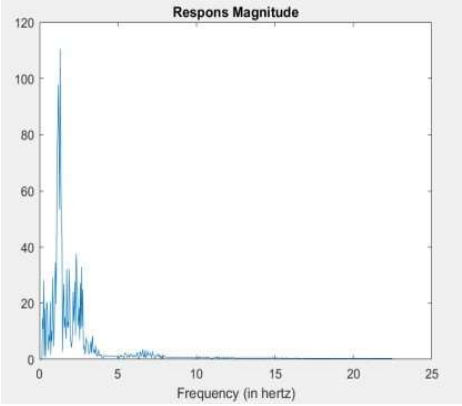


Figure 6. An Example of Raw Pressure Data

TABLE 1. PARAMETER OF THE ALGORITHM

parameters	value
low pass filter cutoff	20Hz
high pass filter cutoff	0.5Hz
moving average filter frame	10 , 5
envelope frame length	65
moving average filter frame	10
Systole Ratio	0.742
Diastole Ratio	0.712

3.2. Dataset Creation

The dataset used to evaluate the accuracy of the signal was derived from the blood pressure data of 18 individuals owned by PT Xirka Dama Persada under normal conditions. Furthermore, for the motion data, the authors collected data from 18 volunteers on different days under normal conditions and move conditions. As a comparison, the authors also used normal blood pressure data from the Omron HEM7156 Blood Pressure Monitor.

3.3. Program Implementation

The authors implemented this algorithm using MATLAB R2023a due to its suitability in signal processing applications and to facilitate signal plotting and analysis at each algorithmic step. Although MATLAB has libraries and toolboxes for various DSP algorithms, the authors attempted to minimize the use of such functions in this paper. This was done to ease the porting of the algorithm to other programming languages, such as C.

4. Result and Discussion

In this section, the testing result will be displayed and discussed

4.1. Algorithm Performance Without Motion Artifact Filter

The performance of the algorithm with and without motion artifact can be seen in Table 2. As seen in Table 2, the algorithm works quite well if no motion artifact is present. The SBP error (SBPE) is slightly above 5 mmHg, which is quite tolerable, while the DBP error (DBPE) and MAP error (MAPE) is below 5 mmHg. However, the algorithm's performance dropped significantly when motion artifact is present. As seen in the Table 2, the errors in MAP, SBP, and DBP increased drastically to 36.111, 18.667, and 34.611 mmHg, respectively.

TABLE 2. ERROR MEASUREMENTS BEFORE AND AFTER MOTION ARTIFACT NOISE IS PRESENT

#Sub	MAPE (mmHg)		SBPE (mmHg)		DBPE (mmHg)	
	Before	After	Before	After	Before	After
1	3	27	5	17	1	22
2	13	33	1	13	9	40
3	7	15	8	24	7	63
4	3	24	6	20	1	22
5	3	21	3	17	3	18
6	10	40	7	17	6	37
7	6	55	19	17	0	32
8	3	37	4	32	3	39
9	2	56	6	13	8	47
10	1	36	0	13	1	37
11	0	43	1	23	1	34
12	1	48	3	20	1	25
13	5	51	9	47	2	38
14	8	10	9	3	5	18
15	0	52	2	17	1	48
16	2	40	15	11	2	45
17	1	15	0	12	3	20
18	5	47	6	20	4	38
Max	**13**	**56**	**19**	**47**	**9**	**63**
Min	**0**	**10**	**0**	**3**	**0**	**18**
Average	**4.056**	**36.111**	**5.778**	**18.667**	**3.222**	**34.611**

4.2. Motion Artifact Detection

In the previous subsection, it is shown that the presence of a motion artifact heavily affects the algorithm's performance. Therefore, it is important to identify, and hopefully minimize, the effect of motion artifact to the NIBP data.

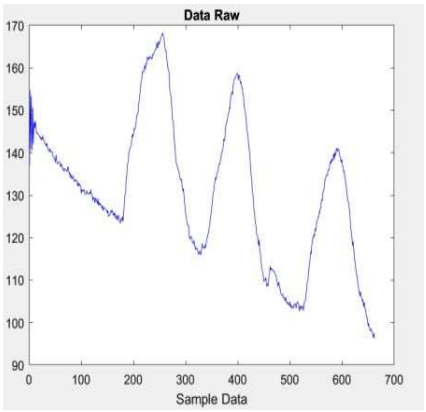


Figure 7. Raw Pressure Data with Motion Artifact

It can be observed in Figure 7, that the data values increase in correlation with movements, and these increases overlap with the genuine data. Therefore, even when smoothed with a moving average, it does not have much impact. Since the motion artifact values persist despite applying a moving average filter, the artifact data remains present in the subsequent processing. Furthermore, even when a bandpass filter is applied, the artifact effects persist,

resulting in inaccurate measurements and significantly different ranges of systolic and diastolic values, as shown in Figure8 and Figure9. Sometimes, the DBP value is higher than the Mean Arterial Pressure (MAP), which indicates an incorrect result. On other occasions, the MAP value is higher than the SBP value, even though the SBP value should be the highest and the MAP value should represent the average arterial pressure. The frequency response of the noisy signal also exhibit very high values, as shown in Figure 10.

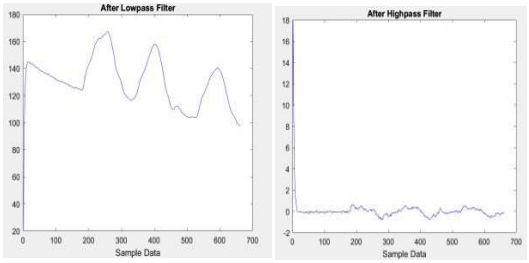


Figure 8. Signal After Bandpass Filtering Still Has Motion Artifact Effects

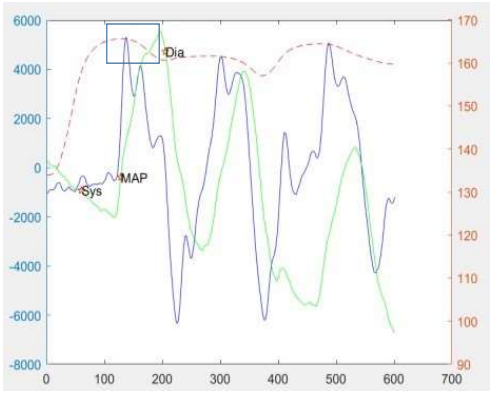


Figure 9. Estimation of MAP, SBP, and DBP of Measurement Data with Motion Artifact

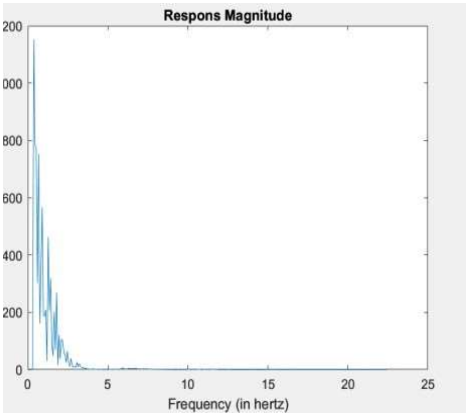


Figure 10. Frequency Response of The Motion Artifact Corrupted Raw Pressure Data

To detect the presence of motion artifact in a NIBP data, we propose 3 category, which are max oscillometric amplitude(Am), frequency of peak magnitude in the frequency response(Fm), and peak magnitude of frequency response(Afm). Based on this category, we measure the F score of the performance of each category to detect the presence of motion artifact in the raw measurement data. The result can be seen in Table 3.

TABLE 3. PARAMETER OF THE ALGORITHM

parameters	threshold	TP	TN	FP	FN	f score
Am	1500	46	50	0	4	0.9583
Fm	0.5Hz	41	50	0	9	0.9011
Afm	400dB	7	50	0	43	0.2456

Table 3 shows the performance comparison of various motion artifact detection parameter. TP stands for True Positive, a case when a raw measurement data with motion artifact is detected as noisy and TN stands for True Negative, a case when a raw measurement data without motion artifact is detected as noise free, while FP stands for False Positive, a case when a raw measurement data without motion artifact is detected as noisy, and FN stands for False Negative, a case when raw measurement data with motion artifact is detected as noise free. From this TP, FP, TN, and FN data, F score is calculated. The parameter with the highest F score is Am, which is the maximum amplitude of the oscillometric waveform. This indicates that raw measurement data corrupted with motion artifact noise is best characterized with large peaks on their oscillometric waveform.

4.3. Algorithm Performance with Motion Artifact Filter

Based on the raw data results in Figure 7, a filter is needed to remove artifact values. Observing the pattern of the raw data resembling a decreasing exponential function, the author created a decreasing exponential function to serve as a subtractor for the raw data, as shown in Figure 11 . The result of subtracting the raw data with this function will then be smoothed using moving average filter with an 80-wide window. This smooth result will provide an estimation of the motion artifact, as seen in Figure 12. Subsequently, the raw data containing the motion artifact will be subtracted by the motion artifact value to obtain an approximation of the original data. After filtering, the signal needs to undergo preprocessing such as moving averages or smoothing filters to optimize the raw data results. However, this carries a risk of potential loss of blood pressure data as well.

Figure 13 shows the raw pressure data after pre-processing. As seen in Figure 13, the filter has succesfully reduces the motion artifact, and after the preprocessing, the data is almost completely free from motion artifact. This means that the data is ready to be processed by the usual algorithm.

Figures 14 - 15 show the data after cutting and bandpass filtering until the MAP, SBP, and DBP calculation. Figure 16 shows the frequency response of the OMW. As seen in these figures,the OMW is more similar to normal data, and MAP, SBP, and DBP can be calculated normally.

The performance of the algorithm after this simple filtering is done can be seen in the Table 4

As seen in Table 4, the simple filter has succesfully reduces the error in NIBP measurement caused by motion artifacts. The mean MAP error (MAPE) is decreased by 67.48%, the mean Systolic error (SBPE) is decreased by 39.28%, and the mean Diastolic Error (DBPE) is decreased by 72.22%. Although there are significant improvement, the final mean errors are still above

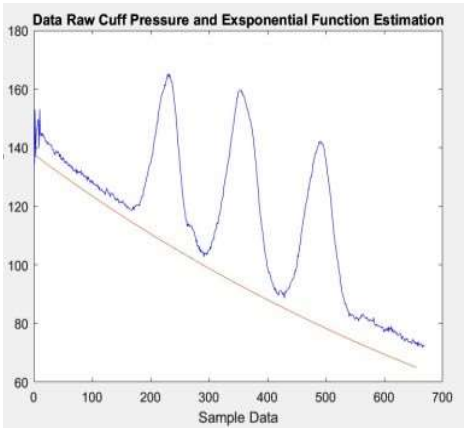


Figure 11. Exponential Function Created to Estimate the Deflation Rate of the Cuff

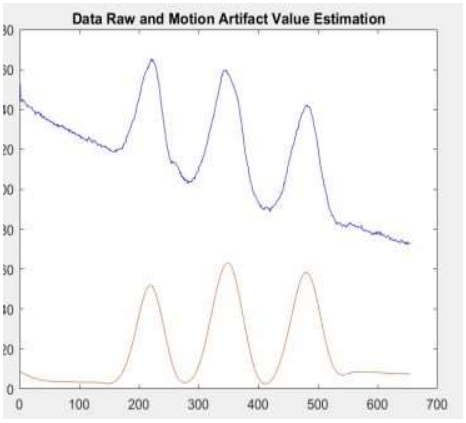


Figure 12. Exponential Function Created to Estimate the Deflation Rate of the Cuff

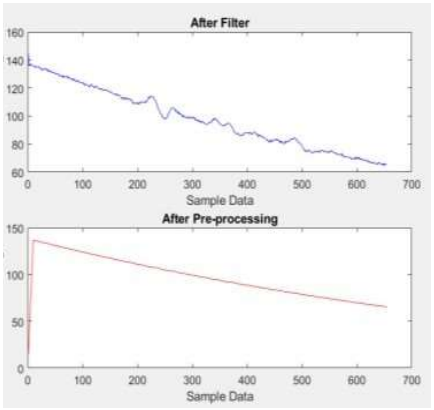


Figure 13. Raw Pressure Data After Pre-processing

5mmHg and haven't satisfy the Indonesian Ministry of Health's standard. Therefore, more complex algorithm is needed to analyze the noisy signal to derive a better filtering method.

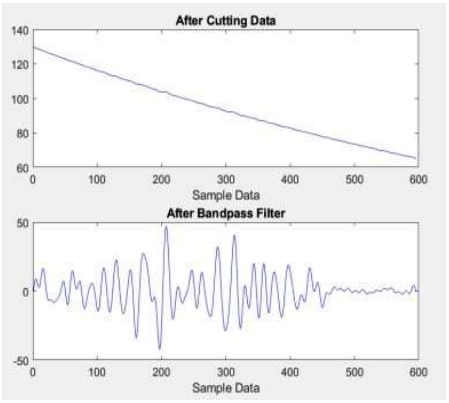


Figure 14. Raw Pressure Data After Cutting and Bandpass Filtering

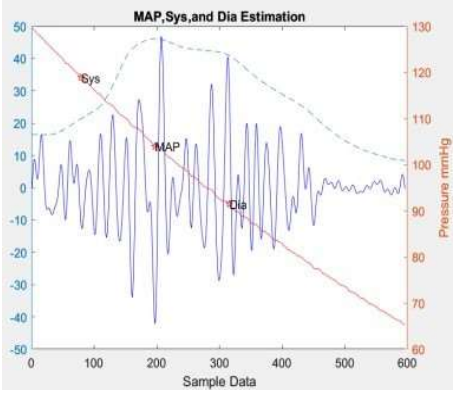


Figure 15. Estimation of MAP, SBP, and DBP

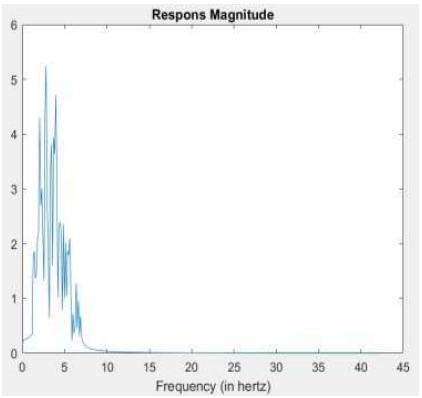


Figure 16. Frequency Response of the Motion Artifact-corrupted OMW Signal After Filtering

5. Conclusion

Motion Artifact is one of the main reasons behind inaccuracy of NIBP measurements. It is characterized by large peaks on the raw measurement data. To minimize the effects of this noise, a simple filtering has been proposed. The filtering involves finding a exponential line to fit the noisy raw measurement data, then substracting it from the noisy raw measurement data. The proposed algorithm has reduced MAP error by 39.28%, SBP error by 39.28%, and DBP error by 72.22%. Although this is a significant improvement, the actual errors are still bigger than 5 mmHg, which

TABLE 4. PERFORMANCE OF THE ALGORITHM BEFORE AND AFTER SIMPLE FILTERING

#Sub	MAPE (mmHg)		SBPE (mmHg)		DBPE (mmHg)	
	Before	After	Before	After	Before	After
1	27	15	17	12	22	9
2	33	15	13	3	40	16
3	15	15	24	35	63	6
4	24	7	20	14	22	7
5	21	9	17	4	18	12
6	40	17	17	6	37	16
7	55	13	17	34	32	8
8	37	9	32	11	39	14
9	56	11	13	22	47	3
10	36	14	13	8	37	10
11	43	9	23	9	34	11
12	48	8	20	9	25	12
13	51	11	47	13	38	7
14	10	13	3	3	18	4
15	52	14	17	7	48	9
16	40	8	11	5	45	9
17	15	8	12	5	20	9
18	47	13	20	4	38	3
Average	**36.111**	**11.611**	**18.667**	**11.333**	**34.611**	**9.611**

means a more complex algorithm is needed to analyze the signal and filter the motion artifact noise.

Acknowledgment

This work was supported in part by Indonesia Endowment Fund for Education Agency (LPDP) under Grant PRJ-50/LPDP/2020, S-248/LPDP.4/2021, and S-446/LPDP.4/2023.

References

[1] *Clinical methods: The history, physical, and laboratory examinations*, 3rd ed. LexisNexis UK, 1990.

[2] D. P2PTM, "Hipertensi, penyakit jantung dan pembuluh darah," n.d. [Online]. Available: https://p2ptm.kemkes.go.id/informasi-p2ptm/hipertensi-penyakit-jantung-dan-pembuluh-darah

[3] R. V. Horn, R. Kahlke, L. Taylor, and T. Dorsett, "Noninvasive blood pressure performance: a reproducible method for quantifying motion artifact tolerance in oscillometry," *Biomed Instrum Technol*, vol. 35, pp. 395–41, 2001.

[4] J. N. Amoore, W. B. Geake, and D. H. T. Scott, "The effects of pulse rate, artifact, and pulse strength on oscillometric non-invasive blood pressure measurements," in *Proceedings of 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1996, n.d.

[5] B. Alpert, D. Quinn, B. Friedman, P. Matsumura, R. Dart, and R. Donehoo, "Evaluating the impact of motion artifact on noninvasive blood pressure devices," *J Clin Hypertens (Greenwich)*, vol. 22, no. 4, pp. 585–589, Apr 2020.

[6] M. Forouzanfar, H. R. Dajani, V. Z. Groza, M. Bolic, S. Rajan, and I. Batkin, "Oscillometric blood pressure estimation: Past, present, and future," *IEEE Reviews in Biomedical Engineering*, vol. 8, pp. 44–63, 2015.

[7] T. J. Dorsett and C. L. Davis, "Method of artifact rejection for noninvasive blood pressure measurement by prediction and adjustment of blood pressure data," Patent U.S. Patent 4 949 710 A, Aug 21, 1990.

[8] M. E. Miller, "A near-optimum filter for suppressing motion artifact in oscillometric signal," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Nov 1989, pp. 640–641.

[9] Y. Koo, J. Kang, H. Shin, M. Y. Jung, G. J. Suh, and H. C. Kim, "Preliminary study of motion artifact rejection for nibp measurement in an ambulance," in *Proc. IEEE 29th Annu. Int. Conf. Eng. Med. Biol. Soc.*, Aug 2007, pp. 705–708.

[10] M. Ward and J. A. Langton, "Blood pressure measurement," *Contin. Educ. Anaesth. Crit. Care Pain*, vol. 7, pp. 122–126, 1997.

[11] M. Mafi, S. Rajan, M. Bolic, V. Z. Groza, and H. R. Dajani, "Blood pressure estimation using the maximum slope of oscillometric pulses," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012.

[12] P. S. Lewis, "Oscillometric measurement of blood pressure: A simplified explanation. a technical note on behalf of the british and irish hypertension society," *Journal of Human Hypertension*, vol. 33, no. 5, pp. 349–351, 2019.

AI-generated image detectors are surprisingly easy to mislead... for now

Zihang Lyu, Jun Xiao, Cong Zhang and Kin-Man Lam
The Hong Kong Polytechnic University, Kowloon, Hong Kong
E-mail: {zihang.lyu, jun.xiao, cong-clarence.zhang}@connect.polyu.hk, enkmlam@polyu.edu.hk

***Abstract*—AI-generated image detectors, also known as fake image detectors, have demonstrated remarkable performance across different datasets and generators, achieving superior detection accuracy and generalizability. Considering the structures of existing AI-generated detectors, which perform binary classification, we propose a simple yet effective adversarial attack method, namely Binary Fast Gradient Sign Method (BFGSM), in this paper. We demonstrate that existing AI-generated image detectors are sensitive to subtle and imperceptible distortions, which raises serious safety risk for these models and makes them inadequate for real-world applications. Experimental results show that our proposed attack successfully misleads current AI-generated image detectors, reducing the attack distortion level by 7.72% with negligible impact on the misleading success rate.**

I. Introduction

As large language models and deep generative models have made great progress in recent years, artificial intelligence-generated content (AIGC) has become increasingly prevalent in people's daily lives. Consequently, the associated security and privacy issues have garnered significant attention from researchers, with generated image detection being one of the most critical aspects.

The detection of generated images initially focused on generative adversarial networks (GANs). Some studies first applied co-occurrence matrices [1] and frequency domain analysis [2]. Wang *et al.* [3] found that, with proper data augmentation and pre-processing, CNN-generated images can be easily spotted, and a detector trained on a single type of generator can generalize to different types of generated images. Recent studies [4], [5] further extend this idea into diffusion models with enhanced feature encoding and artifact representation. However, despite the fact that current generative image detectors have demonstrated progressively higher detection accuracy and precision, these methods typically simplify the detection problem into a binary classification task and employ straightforward classifiers. This approach introduces a significant security vulnerability: these detectors lack the ability to resist adversarial attacks.

Adversarial attacks [6], [7] play an important role in evaluating and enhancing the robustness of classifiers. These attacks involve manipulating the input data to deceive machine learning models. For image classification, a general strategy is to generate a small and unnoticeable perturbation to input images, leading classifiers to produce incorrect labels. Fig. 1 shows some adversarial examples generated by the Binary Fast Gradient Sign Method (BFGSM) proposed in this paper.

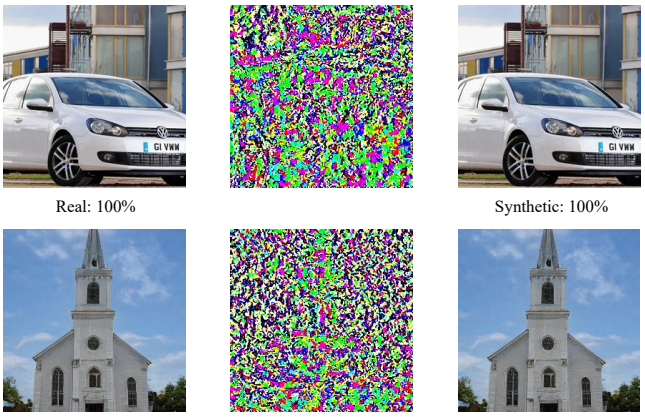


Fig. 1. Visualization of two adversarial examples generated by our proposed Binary Fast Gradient Sign Method (BFGSM): the **left column** represents the input image from real and fake datasets, respectively; the **middle column** shows the adversarial noise to be applied to the original image; the **right column** demonstrates the generated adversarial example. Common AI-generated image detectors are successfully misled with unnoticeable perturbations.

It clearly demonstrates that by adding a small noise that is unnoticeable to human eyes, the classifier can be easily misled. It is worth noting that for conventional classifiers, the security threat posed by adversarial attacks is relatively minor, as misleading these classifiers does not yield significant benefits. However, in the domain of generative image detection, evading the discriminator is a primary objective for malicious users, thereby imparting exceptional importance to adversarial attacks in this field.

In this paper, to address the research gap in adversarial attacks within the domain of generated image detection and to provide a paradigm for future studies, we first analyze the weaknesses of existing detectors when facing adversarial attacks and the underlying reasons for these vulnerabilities. Subsequently, we propose a simple yet effective adversarial attack method based on the traditional Fast Gradient Sign Method (FGSM), named BFGSM, to specifically mislead generated image detectors. This approach retains the simplicity of the traditional FGSM [9] structure while requiring only minimal modifications to the images. We aim for this method to provide a new baseline for misleading binary classification and to contribute to the development of more robust detectors in future research.

The main contributions of this paper are summarized as

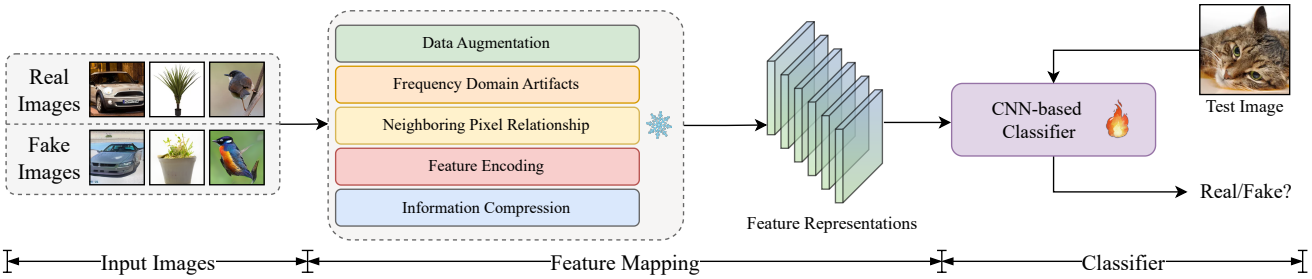


Fig. 2. Illustration of the overall framework of existing AI-generated image detectors: A variety of research studies have been conducted to generate feature representations that aim to distinguish generated images from real images, including data augmentation [3], frequency domain artifacts [2], neighboring pixel relationship [5], feature encoding [4], and information compression [8]. These feature mapping techniques are usually fixed in the training process, with only the CNN-based classifier being trainable. Such framework design makes it vulnerable to adversarial attack.

follows:

1) We introduce adversarial attacks to the field of AI-generated image detection and show that existing detectors can be easily misled with a simple adversarial attack method, which is an essential step for AI-generated image detectors in industrial-level practice.
2) We propose a novel Binary Fast Gradient Sigh Method that can mislead current AI-generated image detectors effectively.
3) Experimental results show that our proposed method can mislead existing detectors with smaller distortion compared to current attacking methods, demonstrating superior effectiveness.

II. Related Work

A. AI-generated Image Detection

While generated images can refer to a large number of images, the detection of AI-generated image mainly focuses on deep generative models, e.g. GANs and diffusion models, and some low-level vision tasks like image enhancement [10]–[12] and super-resolution [13]–[18]. This task initially focused on forged faces, known as deepfakes [19], and gradually tends to generalize fake image detection. Some methods are tailored for domain-specific generation, for example, frequency-level artifacts [2] for GAN images and DIRE [20] for Diffusion images. Other research aims for a universal fake image detector that trains the detectors with images from a single generator and is capable of detecting images generated from various sources. Wang *et al.* [3] initially proposed a simple yet effective detector that using data from ProGAN [21] and has the ability to generalize to detect images from other GAN models and some low-level vision tasks. Ojha *et al.* [4] further analyzed the detection of GAN and Diffusion images and proposed a feature representation for generated images, which extends the idea of universal fake image detectors to diffusion models.

B. Adversarial Attack

With the widespread application of deep neural networks [22]–[28], adversarial attacks have become a highly focused research area. The study of adversarial attacks began with Szegedy *et al.* [29], who discovered several intriguing properties of neural networks. They demonstrated, for the first time, that neural networks could be misled with slight distortions. The initial objective was to develop the equation of minimum distortion to mislead general classifier. However, due to high complexity, this objective was turned to find the minimum loss function addition that leads to misclassification. As the first approach, the proposed method constructs reliable adversarial examples, but with low efficiency and a complex algorithm structure. This idea was further enhanced through the Fast Gradient Sign Method (FGSM) [9], a gradient-based attack method that generates adversarial examples by computing the gradient of the loss function with respect to the input image and adding small perturbations in the direction of the gradient. This method is simple and fast, and some research also extends FGSM into iterative versions [30] and integrates a momentum term [31] to achieve a stronger attack effect. These attacking methods require full knowledge of the model structure and are known as white-box attack. Other research [32], [33] also focuses on black-box attacks where the attacker knows nothing about the model's internal information.

III. Methodology

A. Analysis of AI-generated image detectors

Revisiting the model structure of existing AI-generated image detectors, we find that they commonly focus on feature representation and data pre-processing. Fig. 2 illustrates the overall pipeline of these detectors as explored in various research studies. Researchers have experimented with various sophisticated methods to identify representations that differentiate between real and fake images. However, they have often overlooked the design of the classifier. Typically, a CNN-based network is chosen, sometimes incorporating a ResNet [34] block. This simplistic design leaves the detectors vulnerable to adversarial attacks. Additionally, the binary classification design represents another reason that existing detectors fail to resist adversarial attacks, which will be elaborated on in the following section.

B. Misleading AI-generated image detectors

We propose a simplified version of FGSM [9] targeted for binary classification, namely Binary FGSM (BFGSM), to

retain the original simplicity and success rate while reducing the distortion level applied to images. Suppose x and y are the original input image and label, θ represents the model parameters, the adversarial example $\bar{x}$ is generated by the traditional FGSM [9] as follows:

$$\bar{x} = x + \epsilon \cdot sign(\nabla_x(\mathcal{L}(\theta, x, y)), \tag{1}$$

where $\mathcal{L}$ represents the model loss function and $\nabla_x(\mathcal{L}(\theta, x, y))$ calculates the model gradient with respect to x. Here the perturbation level is controlled by the hyper-parameter ϵ. However, unlike multi-class classification problems, where multiple categories compete, misleading a binary classifier only requires the target class output to fall below a fixed threshold, typically set to 0.5 with the assumption that the input dataset is balanced. This means that when attacking binary classifiers, we can apply relatively smaller noise to samples where the model is less certain, without worrying about a decrease in success rate. In contrast to cases where the model is highly confident, for inputs near the threshold, even a very small perturbation can cause the model to make incorrect predictions. Specifically, it can be formulated as follows:

$$\min_{\alpha_x} \|x - \underbrace{(x + \alpha_x \cdot sign(\nabla_x(\mathcal{L}(\theta, x, y))))}_{\bar{x}} \|_2, \tag{2}$$

subject to $P(\bar{x}) = P(x)$ and $D(\bar{x}) \neq D(x)$, where $P(\cdot)$ denotes the data distribution, and $D(\cdot)$ denotes the AI-generated image detector. Intuitively, after adding the small perturbation, the data distribution is preserved but the AI-generated image detectors should produce inaccurate prediction labels. Therefore, in this paper, we consider the perturbation controlling term α_x to follow the distribution of classification confidence: for images that hold a high classification confidence, we apply a relatively larger distortion and vice versa, denoted as:

$$\alpha_x = \epsilon \cdot \frac{|\hat{y} - \gamma|}{z}, \tag{3}$$

where γ represents the classification threshold, $\hat{y}$ is the model output of x through a sigmoid activation function, and $z = max(|\hat{y} - \gamma|)$ is the normalizing term. The proposed BFGSM then generates the adversarial example as follows:

$$\bar{x} = x + \alpha_x \cdot sign(\nabla_x(\mathcal{L}(\theta, x, y)). \tag{4}$$

Let N represent the number of samples. We adopt the binary cross entropy as the loss function, represented as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}(y \log(\hat{y}) + ((1-y)\log(1-\hat{y}))). \tag{5}$$

Fig. 3 shows the framework of the proposed BFGSM. We can observe that the adversarial examples are produced dynamically with the model output.

IV. Experiments

A. Dataset and Implementation Details

Dataset Information. We evaluate our method on the UnivFD(UniversalFakeDetection) [4] dataset, which consists of various of generative models from GAN and diffusion models to some low-level vision tasks, including ProGAN [21], CycleGAN [35], BigGAN [36], StyleGAN [37], GauGAN [38], StarGAN [39], Deepfakes [19], SITD [10], SAN [13], CRN [40], IMLE [41], Guided Diffusion Model [42], LDM [43], Glide [44] and DALL-E [45]. Different variants of LDM on the number of steps and classifier-free diffusion guidance (CFG) are adopted. Glide has different settings for downsampling and upsampling.

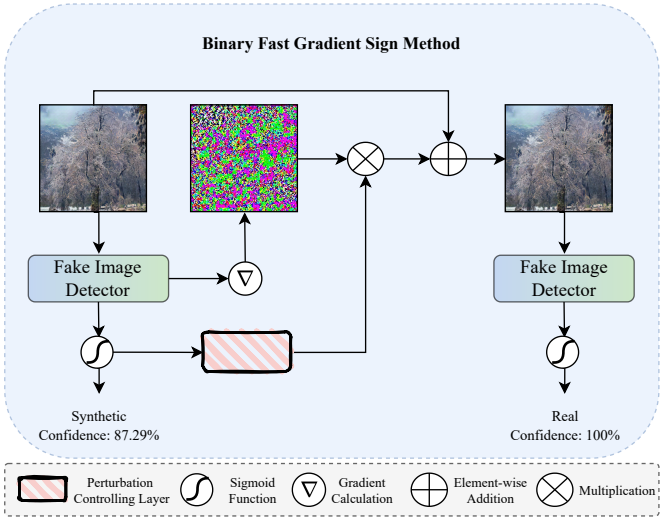


Fig. 3. Illustration of our proposed BFGSM. The perturbation controlling layer controls the magnitude of the adversarial noise, making it adaptive to the original image and corresponding target detector, thus reducing the perturbation level.

Evaluation Metrics. Following the target classifiers [3], [4], we select the classification accuracy and average precision as the evaluation metrics. To demonstrate the effectiveness of the proposed attacking method, we also adopt the L2 norm as a standard metric to evaluate the magnitude of the perturbations.

Implementation Details. We selected CNN Detection [3] and UnivFD [4] as targets for our attacks. The former method pioneered the field of AI-generated image detection through appropriate data augmentation, while the latter advanced the field with its feature extraction techniques. These two classifiers aptly represent the current state of the field. Both classifiers are trained on real and fake images from ProGAN [21] and have the ability to generalize to other generative models. All experiments were conducted on PyTorch on a single NVIDIA RTX 4090 GPU.

B. Effect of Binary Fast Gradient Sign Method

The accuracy and average precision (AP) of the target detectors before and after the attack are illustrated in Table I and Table II. It can be observed that with our proposed BFGSM, the performance of both target detectors has significantly declined, with an average decrease of 45.23% in accuracy and 49.28% in AP. The experimental results demonstrate the capability of our proposed attacking method to mislead classifiers, confirming that existing fake image detectors are indeed vulnerable to

TABLE I

THE ACCURACY OF THE TARGET DETECTORS FOR AI-GENERATED IMAGES ON THE UNIVFD DATASET BEFORE AND AFTER THE ATTACK. THE AVERAGE ACCURACY AFTER THE ATTACK IS HIGHLIGHTED IN BOLD.

Detection method	Adversarial attack	Generative Adversarial Networks						Deep fakes	Low level vision		Perceptual loss		Guided	LDM			Glide			DALL-E	Total
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/ CFG	100 steps	100 27	50 27	100 10		Avg. Acc.
CNN classifier [3]	✗	99.99	85.20	70.20	85.70	78.95	91.70	53.47	66.67	48.69	86.31	86.26	60.07	54.03	54.96	54.14	60.78	63.80	65.66	55.58	69.58
	✓	68.01	15.52	15.60	14.43	22.03	3.68	0.02	10.83	29.45	7.46	13.37	12.65	18.10	18.20	18.10	18.60	18.30	18.90	18.05	**17.96**
UnivFD [4]	✗	100.0	98.50	94.50	82.00	99.50	97.00	66.60	63.00	57.50	59.50	72.00	70.03	94.19	73.76	94.36	79.07	79.85	78.14	86.78	81.38
	✓	77.35	53.2	53.75	45.85	63.65	29.05	6.40	6.00	38.00	45.75	46.00	34.15	47.30	41.10	48.65	42.15	41.50	41.50	46.05	**42.49**

TABLE II

ILLUSTRATION OF THE ACCURACY OF DIFFERENT DETECTORS FOR AI-GENERATED IMAGES ON THE UNIVFD DATASET. THE BEST AND THE SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

Detection method	Adversarial attack	Generative Adversarial Networks						Deep fakes	Low level vision		Perceptual loss		Guided	LDM			Glide			DALL-E	Total
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/ CFG	100 steps	100 27	50 27	100 10		mAP
CNN classifier [3]	✗	100.0	93.47	84.50	99.54	89.49	98.15	89.02	73.75	59.47	98.24	98.40	73.72	70.62	71.00	70.54	80.65	84.91	82.07	70.59	83.58
	✓	92.83	31.38	30.77	30.87	31.93	31.03	30.63	30.83	30.80	32.16	35.61	30.71	30.75	30.77	30.75	30.83	30.80	30.97	30.75	**34.48**
UnivFD [4]	✗	100.0	99.46	99.59	97.24	99.98	99.60	82.45	61.32	79.02	96.72	99.00	87.77	99.14	92.15	99.17	94.74	95.34	94.57	97.15	93.38
	✓	87.04	55.21	54.03	44.37	68.19	34.93	30.79	31.00	32.59	31.21	33.05	32.84	49.44	37.45	50.27	37.43	37.28	37.43	46.24	**43.73**

adversarial attacks. Moreover, unlike multi-class classifiers, adversarial attacks on binary classifiers only need to decrease their performance to chance levels rather than zero in order to significantly mislead them, as such reduced performance is practically meaningless. An interesting phenomenon is that generative models evaluated on the same dataset of real images always exhibit similar performance, as shown in Table I with examples, such as LDM, Glide, and DALL-E. This can be explained by the fact that images more resistant to adversarial attacks are concentrated in the real dataset. However, malicious users can easily mislead detectors by presenting generated images as real ones. This further illustrates the vulnerability of current detectors to adversarial attacks.

C. Effect of Dynamic Perturbation Controlling

Table III compares the misleading success rate of our proposed BFGSM with the traditional FGSM on the two target detectors, as well as the L2 distance between the generated adversarial examples and the original images. To mislead a binary classifier, it only needs to reduce the classification accuracy to chance performance. Therefore, we maintain a balance between the level of perturbation and the success rate to compare the effectiveness of different attack methods under the premise of successfully misleading the classifiers. It is clear that our proposed method significantly reduces the perturbations to the original images with almost no impact on the success rate. The noise level is reduced by 7.72% with only a negligible success rate decrease of 0.04%, demonstrating the superiority of our method. Such performance indicates that the proposed BFGSM is a more suitable adversarial attack method in the field of AI-generated image detection. It is sufficiently simple and efficient while significantly reducing the distortion level of the attack, making it more difficult to detect under the same level of perturbation and serving as a proper baseline for further adversarial attack and detector robustness evaluation.

TABLE III

ILLUSTRATION OF THE MISLEADING SUCCESS RATE (%) AND AVERAGE L2 DISTANCE BETWEEN ORIGINAL IMAGES AND THE ADVERSARIAL EXAMPLES GENERATED BY FGSM AND OUR PROPOSED BFGSM W.R.T THE TARGET DETECTORS ON THE UNIVFD DATASET.

Detection methods	CNN classifier [3]		UnivFD [4]	
	Success rate	L2 Distance	Success rate	L2 Distance
FGSM	74.20	2.29	47.86	1.94
BFGSM (Ours)	74.19	**2.22**	47.79	**1.70**

V. CONCLUSIONS

As AI-generated content continues to integrate into human society, the importance of AI-generated image detection in fields, such as security, privacy, and copyright, continues to increase. Despite the remarkable performance achieved by existing AI-generated image detectors, their vulnerability to adversarial attacks allows malicious users to easily bypass them, hindering their true industrial-level application. In this paper, we first analyze the structure of general detectors and point out that the lack of classifier design and the nature of binary classification are the reasons for existing shortcomings. Subsequently, we propose a simple yet effective adversarial attack method tailored for AI-generated image detection, namely the Binary Fast Gradient Sign Method (BFGSM). This method utilizes an adaptive controlling term that changes dynamically with the classification confidence of the input image with respect to the target detector, applying only the necessary noise level to mislead each specific image. Experimental results demonstrate that the proposed BFGSM can easily mislead existing classifiers with minimal perturbations, showcasing the effectiveness and superiority of our method. This study establishes a new research paradigm for adversarial attacks in the field of AI-generated image detection and provides a new adversarial attack baseline for misleading binary classification. We anticipate it will contribute to the development of more robust detectors in future studies.

REFERENCES

[1] L. Nataraj, T. M. Mohammed, S. Chandrasekaran, *et al.*, "Detecting GAN generated fake images using co-occurrence matrices," *Electronic imaging*, 2019.

[2] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," in *WIFS*, 2019.

[3] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *CVPR*, 2020.

[4] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *CVPR*, 2023.

[5] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection," in *CVPR*, 2024.

[6] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *ICPR*, 2021.

[7] T. Zheng, C. Chen, and K. Ren, "Distributionally adversarial attack," in *AAAI*, 2019.

[8] Z. Lyu, J. Xiao, C. Zhang, and K.-M. Lam, "AI-generated image detection with Wasseerstein distance compression and dynamic aggregation," in *ICIP*, 2024.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[10] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018.

[11] J. Xiao, Q. Ye, T. Liu, C. Zhang, and K.-M. Lam, "Deep progressive feature aggregation network for multi-frame high dynamic range imaging," *Neurocomputing*, vol. 594, p. 127 804, 2024.

[12] J. Xiao, Z. Lyu, C. Zhang, Y. Ju, C. Shui, and K.-M. Lam, "Towards progressive multi-frequency representation for image warping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2995–3004.

[13] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *CVPR*, 2019.

[14] J. Xiao, Q. Ye, R. Zhao, K.-M. Lam, and K. Wan, "Deep multi-scale feature mixture model for image super-resolution with multiple-focal-length degradation," *Signal Processing: Image Communication*, vol. 127, p. 117 139, 2024.

[15] J. Xiao, Q. Ye, R. Zhao, K.-M. Lam, and K. Wan, "Self-feature learning: An efficient deep lightweight network for image super-resolution," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4408–4416.

[16] C. Yang, J. Xiao, Y. Ju, G. Qiu, and K.-M. Lam, "Improving robustness of single image super-resolution models with monte carlo method," in *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 2135–2139.

[17] C. Yang, R. Dong, J. Xiao, *et al.*, "Geometric distortion guided transformer for omnidirectional image super-resolution," *arXiv preprint arXiv:2406.10869*, 2024.

[18] J. Xiao, W. Jia, and K.-M. Lam, "Feature redundancy mining: Deep light-weight image super-resolution model," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 1620–1624.

[19] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV*, 2019.

[20] Z. Wang, J. Bao, W. Zhou, *et al.*, "Dire for diffusion-generated image detection," in *CVPR*, 2023.

[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *ICLR*, 2018.

[22] C. Zhang, K.-M. Lam, and Q. Wang, "Cof-net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023.

[23] C. Zhang, J. Su, Y. Ju, K.-M. Lam, and Q. Wang, "Efficient inductive vision transformer for oriented object detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023.

[24] C. Zhang, K.-M. Lam, T. Liu, Y.-L. Chan, and Q. Wang, "Structured adversarial self-supervised learning for robust object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, 2024.

[25] R. Dong and K.-M. Lam, "Bi-center loss for compound facial expression recognition," *IEEE Signal Processing Letters*, vol. 31, pp. 641–645, 2024.

[26] Y. Ju, M. Jian, J. Dong, and K.-M. Lam, "Learning photometric stereo via manifold-based mapping," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, 2020, pp. 411–414.

[27] Y. Ju, B. Shi, Y. Chen, H. Zhou, J. Dong, and K.-M. Lam, "Gr-psn: Learning to estimate surface normal and reconstruct photometric stereo images," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[28] Y. Ju, K.-M. Lam, W. Xie, H. Zhou, J. Dong, and B. Shi, "Deep learning methods for calibrated photometric stereo and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[29] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," in *ICLR*, 2014.

[30] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, 2018.

[31] Y. Dong, F. Liao, T. Pang, *et al.*, "Boosting adversarial attacks with momentum," in *CVPR*, 2018.

[32] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017.
[33] Y. Dong, H. Su, B. Wu, *et al.*, "Efficient decision-based black-box adversarial attacks on face recognition," in *CVPR*, 2019.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[35] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.
[36] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *ICLR*, 2018.
[37] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.
[38] J. F. O'brien and H. Farid, "Exposing photo manipulation with inconsistent reflections," *ACM Transactions on Graphics*, 2012.
[39] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018.
[40] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *ICCV*, 2017.
[41] K. Li, T. Zhang, and J. Malik, "Diverse image synthesis from semantic layouts via conditional imle," in *ICCV*, 2019.
[42] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *NeurIPS*, 2021.
[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.
[44] A. Nichol, P. Dhariwal, A. Ramesh, *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *ICML*, 2022.
[45] A. Ramesh, M. Pavlov, G. Goh, *et al.*, "Zero-shot text-to-image generation," in *ICML*, 2021.

Dictionary Learning for Directed Graph Signals via Augmented GFT

Tsubasa NAITO*, Ryuto ITO*, Yuichi TANAKA†, Shogo MURAMATSU*
* Niigata University, Niigata, Japan
† Osaka University, Osaka, Japan

***Abstract*—This paper proposes a method for designing directed graph (digraph) filters through a dictionary learning approach. Practical digraph filtering methods have not yet been established because of the difficulties posed by the asymmetry of the adjacency matrix of a digraph. *Augmented graph Fourier transform* (AuGFT), proposed by Kitamura *et al.,* defines a new graph Laplacian and extends the conventional graph Fourier transform (GFT) for undirected graph signals to directed ones. However, challenges remain in filter design through AuGFT, particularly in determining the skew intensity parameters. Therefore, this study aims to establish a design method for digraph filters with AuGFT. Filters are parameterized with AuGFT, and parameter optimization is performed using a dictionary learning technique. To verify the effectiveness of the proposed method, experimental results of digraph filtering are shown for temperature data of contiguous US and the GSP-Traffic Dataset. Compared with undirected graph filtering, the proposed method is shown to have high steerability in the design of digraph filters.**

I. Introduction

Graph signal processing (GSP) is a theoretical framework for signal processing based on analysis of the spectral properties of the graph, where the signal domain is defined on the vertices [1], [2]. When the graph structure is known, graph signal processing provides an efficient means of storing, transmitting, and analyzing a variety of graph signals. Graphs can reflect known or estimated connectivity as a prior knowledge in target models. The theory and algorithms of graph signal processing have primarily been designed for undirected graphs, where the connectivity information between vertices is represented by an adjacency matrix. The graph Laplacian, being a graph operator, is a symmetric matrix and can thus be diagonalizable by an orthonormal matrix. Such orthonormal matrices are used as a basis for the graph Fourier transform (GFT). The eigenvalues of graph operators, known as graph frequencies, are non-negative real numbers, allowing the spectrum corresponding to the frequencies to be sorted according to their magnitudes. Consequently, filtering in the spectral domain is realized and utilized as the primary operation in undirected graph signal processing. However, real-world networks, which often include directional information, are difficult to represent using only connectivity information and are instead represented using digraphs. However, digraph operators are generally not symmetric, offering no guarantee that the graph operator can be diagonalized or that its eigenvalues are real [3], [4]. Therefore, unlike undirected graphs, defining the GFT and graph frequency for digraphs poses significant challenges.

To address these challenges, various approaches have been proposed. For example, studies using symmetric components of graph operators [4], [5], studies on minimizing spectral variance [6], studies on constructing a new GFT by applying Jordan decomposition [7], [8] or singular value decomposition [9] to the adjacency matrix, and studies on symmetrization of the adjacency matrix [10]. However, many of the existing methods are not easy to apply to GSP, which is based on filtering in the spectral domain.

In [11], Furutani *et al.* redefined graph operators for digraphs as Hermitian symmetric matrices by embedding the directional information into the imaginary part of the graph operator. This approach enables the sorting of the graph frequency as a real number and performs filtering in the graph frequency domain. However, a challenge arises with the Hermite-Laplacian filtering technique: it practically produces complex-valued outputs for real-valued inputs.

In order to solve the conventional digraph filtering problem, in the article [12], Kitamura *et al.* proposed the augmented GFT (AuGFT) [1]. AuGFT generalizes the traditional GFT by splitting the graph Laplacian into symmetric and skew-symmetric components and applying the canonical decompositions separately so that a redundant real GFT is defined for a digraph. AuGFT enables practical real-valued filtering for digraph signals. This generalized method ensures real-valued output with filters designed with little constraint. In parallel to this study, Kwak *et al.* proposed to apply polar decomposition to the adjacency matrix and provides a new interpretable filtering technique in the vertex domain [13]. The method in [13] decomposes the operators into product forms, while our proposed method decomposes them into sum forms. Although both of them can realize practical real filters, in this study, we adopt the latter as the base, which can directly utilize existing filters for undirected graph signals.

Despite the advantage of AuGFT, there remain challenges in establishing practical procedures to design filters efficiently. To further advance the method, we propose a novel filter design method that uses dictionary learning technique. We demonstrate its effectiveness through comparative analysis with some conventional filters for undirected graph signals.

The organization of this paper is as follows. In Section 2, we review digraph signal processing. In Section 3, we propose a digraph filter design method using a dictionary learning tech-

[1] https://github.com/msiplab/AuGFT

nique. The effectiveness of our proposed method is validated in Section 4. Finally, Section 5 concludes our contributions.

II. Review of Digraph Signal Processing

In this section, we introduce graph adjacency matrices and the Hermite symmetric graph Laplacian [12], as a graph operator that plays an important role in digraph signal processing. We also introduce the AuGFT based on them. Note here that the definition of AuGFT is generalized from that in the article [12] so that design parameters can be introduced to control the skew intensities.

A. Review of Digraph

Let us assume a finite digraph $\mathcal{G} = (\mathcal{V}, \mathbf{A})$, where $\mathcal{V} \subset \mathbb{N}_0$ is a set of vertices, $\mathbf{A} = (a_{m,n}) \in [0, \infty)^{N \times N}$ is an adjacency matrix, and $a_{m,n}$ is a weight from the vertex m to the vertex n. If $a_{m,n} = a_{n,m}$ holds for all $(m, n) \in \mathcal{V} \times \mathcal{V}$, then $\mathcal{G}$ is not directed and the adjacency matrix $\mathbf{A}$ is symmetric, that is, $\mathbf{A}^\intercal = \mathbf{A}$, where the superscript $\intercal$ denotes transposition. Unless otherwise stated, $N = |\mathcal{V}| < \infty$ denotes the number of vertices. Since the adjacency matrix $\mathbf{A}$ of a digraph $\mathcal{G}$ is not symmetric, we use *the Hermitian adjacency matrix* $\mathbf{C}$ defined as

$$\mathbf{C} \coloneqq \mathbf{A}_+ + \mathrm{j}\mathbf{A}_-, \tag{1}$$

which is symmteric and reflects the vertice connectivity of digraph where $\mathbf{A}_+ \coloneqq \frac{1}{2}(\mathbf{A} + \mathbf{A}^\intercal)$ and $\mathbf{A}_- \coloneqq \frac{1}{2}(\mathbf{A} - \mathbf{A}^\intercal)$.

B. Para-graph Laplacian

The para-graph Laplacian $\mathbf{L}$ is defined by symmetric real part and skew-symmetric imaginary part as

$$\mathbf{L} = \mathbf{D} - \mathbf{C} = (\mathbf{D} - \mathbf{A}_+) - \mathrm{j}\mathbf{A}_-, \tag{2}$$

where $\mathbf{D}$ is the degree matrix of $\mathbf{C}$ defined as the diagonal matrix having the m-th diagonal element:

$$d_{m,m} \coloneqq \sum_{n \in \mathcal{V} \setminus \{m\}} \sqrt{c_{m,n} c_{n,m}}, \ m \in \mathcal{V}. \tag{3}$$

Since $\mathbf{C}$ is Hermitian symmetric, $c_{n,m} = \bar{c}_{m,n}$ and $c_{n,m}c_{m,n} = \bar{c}_{m,n}c_{m,n} = |c_{m,n}|^2$ hold. This leads to $d_{m,m} = \sum_{n \in \mathcal{V} \setminus \{m\}} |c_{m,n}|$. For an undirected graph, $|c_{m,n}| = a_{m,n}$.

We further have a decomposition

$$\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal + \mathrm{j}\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\intercal \tag{4}$$

from canonical forms for real symmetric and real skew-symmetric matrices [14, Corollary 2.5.11(b)], where $\mathbf{U} \in \mathbb{R}^{N \times N}$ is an orthonormal matrix, $\mathbf{Q} \in \mathbb{R}^{N \times r}$ is a matrix consisting of r orthonormal columns,

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_0, \lambda_1, \cdots, \lambda_{N-1}) \tag{5}$$

with $\lambda_n \in [0, \infty)$, $\lambda_n \leq \lambda_m$ for $n < m$, and

$$\mathbf{\Sigma} = s_0 \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus s_1 \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \cdots s_{\frac{r}{2}-1} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{6}$$

with $s_n \in [0, \infty)$, $s_n \geq s_m$ for $n < m$, and r denotes the rank of $\Im\{\mathbf{L}\} = -\mathbf{A}_-$. The symbol $\oplus$ denotes the direct sum of the matrices to construct a block diagonal matrix.

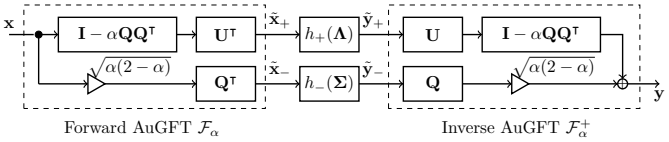


Fig. 1: Digraph filtering $\mathbf{y} = \mathbf{H}\mathbf{x}$, where $\mathbf{U} \in \mathbb{R}^{N \times N}$ and $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$ are the orthonomal matrix and diagonal matrix obtained by eigenvalue decomposition of $\Re\{\mathbf{L}\}$, and $\mathbf{Q} \in \mathbb{R}^{N \times r}$ and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ are the submatrices of the orthonormal matrix and block diagonal matrix obtained by canonical decomposition of $\Im\{\mathbf{L}\}$, respectively, where $\mathbf{L}$ is a para-graph Laplacian, r is the rank of $\Im\{\mathbf{L}\}$, $h_+(\cdot)$ and $h_-(\cdot)$ are para-graph frequency and skew intensity responses respectively. Note that $r = 0$ for an undirected graph. Therefore, AuGFT reduces to the ordinary GFT. The top channel takes the orthogonal complement space of $\mathrm{span}(\mathbf{Q})$ into account.

$\mathbf{L}$ is a generalization of the conventional graph Laplacian and satisfies the Hermitian symmetric property, i.e., $\mathbf{L}^\mathsf{H} = \mathbf{L}$, and

$$[\mathbf{L}\mathbf{x}]_m = \sum_{n \in \mathcal{V} \setminus \{m\}} |c_{m,n}| \cdot (x_m - e^{\mathrm{j}\angle c_{m,n}} x_n), m \in \mathcal{V}, \tag{7}$$

where the superscript H denotes the Hermitian transposition and $[\cdot]_m$ denotes the m-th element. The para-graph Laplacian quadratic form as a measure of graph signal variation (GSV) is obtained by

$$\Delta_{\mathbf{L}}(\mathbf{x}) \coloneqq \mathbf{x}^\mathsf{H}\mathbf{L}\mathbf{x} = \sum_{m=0}^{N-1} \sum_{n=m+1}^{N-1} |c_{m,n}| \cdot |x_m - e^{\mathrm{j}\angle c_{m,n}} x_n|^2, \tag{8}$$

where $\mathbf{x} \in \mathbb{C}^N$ (or $\mathbf{x} \in \mathbb{R}^N$).

C. Augmented Graph Fourier transform (AuGFT)

Let $x_n \in \mathbb{R}$ be a signal at the vertex n. A digraph signal $\{x_n\}_n$ is denoted by $\mathbf{x} = (x_0, x_1, \cdots, x_{N-1})^\intercal \in \mathbb{R}^N$. The AuGFT and its inverse are defined by

$$\tilde{\mathbf{x}} = \mathcal{F}_\alpha \mathbf{x} \coloneqq \left((\mathbf{I} - \alpha\mathbf{Q}\mathbf{Q}^\intercal)\,\mathbf{U} \quad \sqrt{\alpha(2-\alpha)}\mathbf{Q} \right)^\intercal \mathbf{x}, \tag{9}$$

$$\mathbf{x} = \mathcal{F}_\alpha^+ \tilde{\mathbf{x}} \coloneqq \left((\mathbf{I} - \alpha\mathbf{Q}\mathbf{Q}^\intercal)\,\mathbf{U} \quad \sqrt{\alpha(2-\alpha)}\mathbf{Q} \right) \tilde{\mathbf{x}}, \tag{10}$$

where $\tilde{\mathbf{x}} \in \mathbb{R}^{N+r}$ and $r = \mathrm{rank}(\Im\{\mathbf{L}\})$. $\alpha \in [0, 1]$ is a parameter that controls the intensity of the skew components.

D. Digraph filtering

We realize practical digraph filtering

$$\mathbf{y} = \mathbf{H}\mathbf{x} \tag{11}$$

for digraph signal $\mathbf{x}$, where $\mathbf{y} \in \mathbb{R}^N$ is the filtered digraph signal, and $\mathbf{H} \in \mathbb{R}^{N \times N}$ is a vertex-domain filter through AuGFT. The graph frequencies $\{\lambda_n\}_n$ and the skew intensities $\{s_n\}_n$ are all real and non-negative.

Fig. 1 shows the digraph filtering procedure. The vertex-domain filter $\mathbf{H}$ in (11) is formulated as

$$\mathbf{H} = \mathcal{F}_\alpha^+ \left(h_+(\mathbf{\Lambda}) \oplus h_-(\mathbf{\Sigma}) \right) \mathcal{F}_\alpha, \tag{12}$$

where

$$h_+(\mathbf{\Lambda}) \coloneqq \mathrm{diag}\left(h_+(\lambda_0), h_+(\lambda_1), \cdots, h_+(\lambda_{N-1})\right), \quad (13)$$

$$h_-(\mathbf{\Sigma}) \coloneqq h_-(s_0)\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix} \oplus h_-(s_1)\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix} \oplus \cdots \cdots \oplus h_-(s_{\frac{r}{2}-1})\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}. \quad (14)$$

$\mathbf{H}$ is always real if we design both $h_+(\cdot)$ and $h_-(\cdot)$ to be real, thus the output $\mathbf{y}$ is guaranteed to be real for a real input $\mathbf{x}$.

III. Digraph Dictionary Learning

The innovative AuGFT has made it possible to design digraph filters by decomposing them into responses on graph frequencies $\{\lambda_n\}_n$ and skew intensities $\{s_n\}_n$, however, challenges remain in how to design filters or choose parameters, especially in determining the skew intensity responses. In this section, we propose parameterizing the filters with AuGFT's α and optimizing the parameter with a dictionary-learning technique. We consider constructing a P-channel graph filter bank (GFB) $\mathbf{G} = \begin{bmatrix}\mathbf{H}_0 & \mathbf{H}_1 & \cdots & \mathbf{H}_{P-1}\end{bmatrix}$ through dictionary learning. Learning a P-channel GFB from graph signals can be reduced to solving the following optimization problem:

$$\{\hat{\mathbf{G}}, \hat{\mathbf{X}}\} = \arg\min_{\{\mathbf{G},\mathbf{X}\}} \frac{1}{2}\|\mathbf{V} - \mathbf{G}\mathbf{X}\|_{\mathrm{F}}^2 + \beta\mathscr{R}(\mathbf{X}), \quad (15)$$

where $\mathbf{V}$ is a data matrix of graph signals, $\mathbf{X}$ is a coefficient matrix for $\mathbf{G}$, $\beta \in [0, \infty)$ is the regularization parameter, $\mathscr{R}(\mathbf{X})$ is regularizer imposed to the coefficient matrix and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

A. Regularizer

Various types of regularizer $\mathscr{R}$ have been studied in the field of signal processing. For the purpose of dictionary learning, we propose to adopt a typical regularizer, the ℓ_1-norm of the coefficient matrix, as follows

$$\mathscr{R}(\mathbf{X}) = \sum_k \|\mathbf{x}_k\|_1, \quad (16)$$

where $\mathbf{x}_k$ is the k-th column vector of $\mathbf{X}$.

B. Optimization

We consider the design of filters through (15) and (16). The minimization problem is alternatively iterated with the following coefficient sparsification problem (Coefficient update) and the dictionary optimization problem (Dictionary update) to obtain optimally learned digraph filters.

- Coefficient update step: given dictionary $\mathbf{G}_{\hat{\theta}}$ in the initial state or trained at the previous step, obtain the optimal coefficients $\hat{\mathbf{X}}$ such that

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{V} - \mathbf{G}_{\hat{\theta}}\mathbf{X}\|_{\mathrm{F}}^2 + \beta\sum_k \|\mathbf{x}_k\|_1, \quad (17)$$

  where $\mathbf{G}_{\hat{\theta}}$ is the GFB determined by design parameter $\hat{\theta}$.

TABLE I: First experiment setup in IV-A

Number of data for dictionary learning	470
Filter for h_+	Half cosine
Number of filters P	4
Regularization parameter β	0.5
Standard derivation σ	5

- Dictionary update step: given the previous optimal coefficient $\hat{\mathbf{X}}$, learn the design parameter θ of the dictionary such that

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{2}\|\mathbf{V} - \mathbf{G}_{\theta}\hat{\mathbf{X}}\|_{\mathrm{F}}^2. \quad (18)$$

IV. Performance Evaluation

In this section, we conducted two experiments involving the design of the denoising filter to evaluate the feasibility of the proposed dictionary learning method. These experiments entailed adding noise $\mathbf{w}$ to a digraph signal $\mathbf{u}$ and subsequently attempting to denoise the observation

$$\mathbf{v} = \mathbf{u} + \mathbf{w} \quad (19)$$

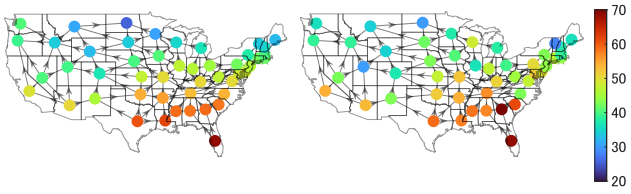
through sparse modeling. The noise is assumed to be additive white Gaussian noise (AWGN) with a standard deviation of σ.

Let $h_{p,+}(\cdot), h_{p,-}(\cdot)$ and α_p be the frequency response, skew response, and AuGFT parameter α of the p-th filter $\mathbf{H}_p$, respectively. In the following experiments, $\{h_{p,+}(\cdot)\}$ is determined as a half-cosine GFB [15], which is developed for an undirected graph signal, and $\{h_{p,-}(\cdot)\}$ is fixed to a set of identities, i.e., $h_{p,-}(s) = s$. We control the skewness parameter α_p for each channel independently. That is, $\{\alpha_p\}$ are the design parameters of the following GFBs.

We examine the feasibility of designing filters. To validate the performance of the designed digraph filters, we evaluated the RMSE of the denoised graph signal and the GSV in (8), compared them to the undirected graph case ($\alpha_p = 0$) and to the method for extracting the real part of the filter output using Furutani's digraph GFT [11]. The coefficient update, involving the ℓ_1-norm regularization, was addressed with the iterative shrinkage-thresholding algorithm (ISTA) for the coefficient update step, and we employed MATLAB's `fmincon` for the dictionary update step. In these experiments, the weights of the adjacency matrix are set as $a_{m,n} = 1$ for the edge $m \to n$.

A. Experiment for Digraph of Contiguous United States

First, we utilized a digraph with 48 US states (excluding Alaska and Hawaii) as vertices and monthly average temperature data as graph signals for each state. The edges in the graph were connected from lower to higher latitude. Fig. 2 shows the temperature signal on the directed graph of the continuous US used in this experiment. Generally, states closer to the equator, i.e. at lower latitudes, have higher average temperatures. Therefore, we think it is reasonable to assign directed vertices based on the latitude of the state to capture temperature flow. The experimental setup is detailed in Table I, with empirically determined parameters.



(a) (b)

Fig. 2: (a) Graph signal of the average temperature in Fahrenheit of March 2024 in the directed contiguous US. This graph has 48 US states excluding Alaska and Hawaii as vertices. (b) Example of noisy graph signal (RMSE: 5.056).

TABLE II: Cities used for the experiment and their characteristics

City	Country	vertices	Total Edges	Directed Edges
Sakarya	Turkey	103	150	114
Chon Buri	Thailand	108	140	112
Chaoyang (Guangdong)	China	108	206	88
Kanpur	India	117	208	98

B. Experiment for GSP-Traffic-Dataset

Second, we utilized the GSP-Traffic-Dataset[2] [16]. The data set includes virtual measurements of traffic volumes in 465 cities around the world, and the graph represents an actual road network with intersections as vertices and roads as edges. The signals in the data set represent the total number of vehicles passing through intersections in more than 500 seconds, obtained by simulations. The data set provides time-series data from a 50,000-second simulation.

Given that real roads include one-way streets, the graph is directed with more than 100 vertices in the graph. Consequently, we can conduct the experiments with a graph that has more vertices than the first experiment and, unlike an arborescence, has partially directed edges.

We used data from four cities for the experiment. The cities used and their characteristics are presented in Table II. We utilized the first 99 data for dictionary learning and the rest for validation. The experimental setup is detailed in Table III, with empirically determined parameters.

C. Experimental results

The results of the first experiment in IV-A and the second experiment in IV-B are presented in Tables IV and V, respectively. These tables show the average of 100 trial simulations in RMSE and GSV for the original data $\mathbf{u}$, noisy data $\mathbf{v}$, denoised data $\hat{\mathbf{u}}_u$ through undirected graph filters, and denoised data $\hat{\mathbf{u}}_d$ through the proposed digraph filteres.

[2]https://github.com/rukumagai/GSP-Traffic-Dataset

TABLE III: Second experiment setup in IV-B

Filter for h_+	Half cosine
Number of filters P	6
Regularization parameter β	0.2
Standard derivation σ	5

TABLE IV: Average of 100 simulation results in the first experiment of IV-A

	RMSE	GSV
original data	-	93113
noisy data	5.012	96720
undirected graph filtered	4.968	96290
Furutani's method (real part) [11]	5.006	96428
Digraph filtered (Proposal)	**4.961**	**95466**

TABLE V: Average of 100 simulation results for each city in the second experiment in IV-B

(a) Sakarya, Turkey

	RMSE	GSV
original data	-	62929
noisy data	5.002	67783
undirected graph filtered	4.816	62280
Furutani's method (real part) [11]	4.825	60358
Digraph filtered (Proposal)	**4.810**	**61664**

(b) Chon Buri, Thailand

	RMSE	GSV
original data	-	38553
noisy data	4.999	42936
undirected graph filtered	4.708	37654
Furutani's method (real part) [11]	4.711	39038
Digraph filtered (Proposal)	**4.703**	**37323**

(c) Chaoyang, China

	RMSE	GSV
original data	-	50338
noisy data	4.999	56923
undirected graph filtered	4.909	53300
Furutani's method (real part) [11]	4.898	52479
Digraph filtered (Proposal)	**4.866**	**51945**

(d) Kanpur, India

	RMSE	GSV
original data	-	45399
noisy data	4.990	51618
undirected graph filtered	4.842	48646
Furutani's method (real part) [11]	4.804	48191
Digraph filtered (Proposal)	**4.800**	**47207**

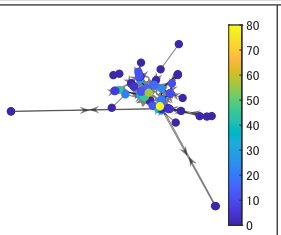
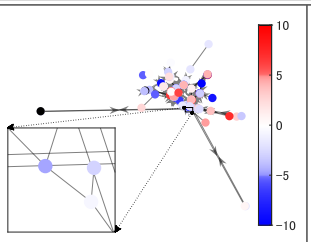
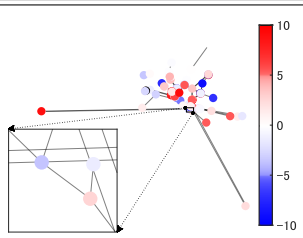
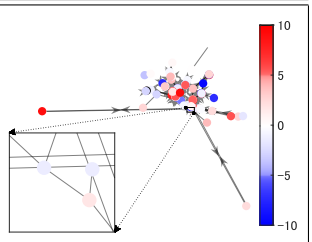
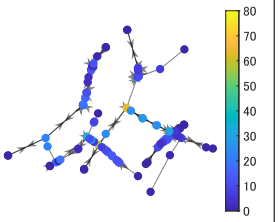
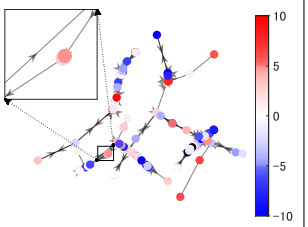
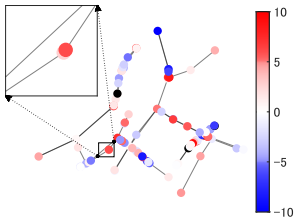
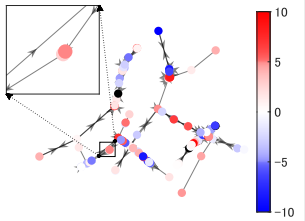
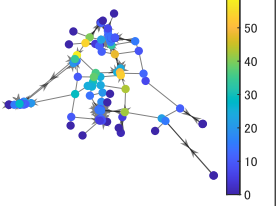
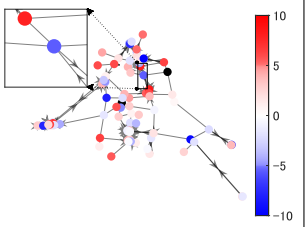
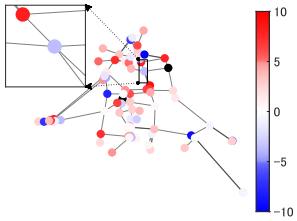
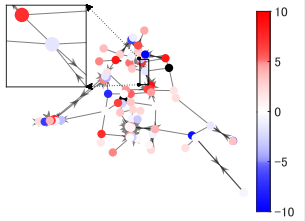
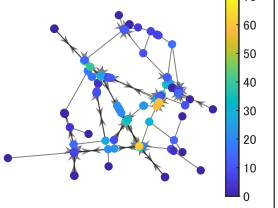
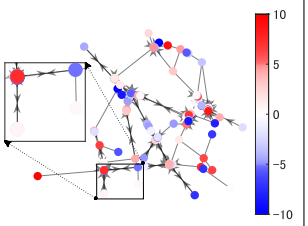
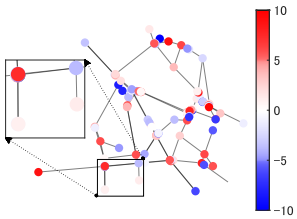
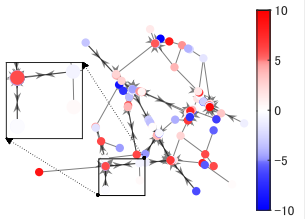
D. Discussion

In both experiments, compared to the undirected graph filters with $\alpha_p = 0$, the RMSEs of our proposal were improved, and the GSV decreased. The results show that digraph denoising filters can be effectively realized by controlling the intensity of the skew components on the graphs. An example of the results from the second experiment in IV-B is shown in Table VI. The graph signals are more smoothed with its adjacent vertex and vertices with directed edges show improvement in the results.

V. Conclusion

In this paper, we proposed a method for designing digraph filters that ensure real-valued outputs for real-valued inputs, through dictionary learning techniques with AuGFT. We realized the digraph filter design by optimizing the GFB parameterized by skewness parameters. The proposed method

TABLE VI: Examples of traffic volume graph signal on road for each city. The 'original' column gives the graph signal $\mathbf{u}$ used for validation. The 'noise' column gives $\mathbf{w} = \mathbf{u} - \mathbf{v}$, the 'undirected graph filtered' column gives $\mathbf{u} - \hat{\mathbf{u}}_u$, and the 'digraph filtered' column gives $\mathbf{u} - \hat{\mathbf{u}}_d$. The directed edges, expressed by arrows, represent one-way streets, and the undirected edges represent two-way streets.

	original $\mathbf{u}$	noise $\mathbf{w} = \mathbf{u} - \mathbf{v}$	undirected graph filtered $\mathbf{u} - \hat{\mathbf{u}}_u$	directed graph filtered $\mathbf{u} - \hat{\mathbf{u}}_d$
Sakarya, Turkey				
Chon Buri, Thailand				
Chaoyang, China				
Kanpur, India				

overcomes the challenges in conventional digraph signal processing and allows the design of appropriate filters for digraph signals. The effectiveness of the proposed method was verified through experiments with temperature data from the contiguous United States and traffic volume data in four cities provided by GSP-Traffic-Dataset. The introduction of control of the skewness parameter confirmed that effective digraph filters can be designed compared to undirected graph filters. Future work includes further refinement of the filter design method by extending the parameterization in AuGFT domain to design a more flexible GFB and exploring applications to other types of digraph signals.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP22H00512 and JP24K21314.

REFERENCES

[1] A. Ortega, *Introduction to Graph Signal Processing*. Cambridge University Press, 2022.

[2] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018. DOI: 10.1109/JPROC.2018.2820126.

[3] A. G. Marques, S. Segarra, and G. Mateos, "Signal processing on directed graphs: The role of edge directionality when processing and learning from network data," *IEEE Signal Processing Magazine*, vol. 37, no. 6, pp. 99–116, 2020. DOI: 10.1109/MSP.2020.3014597.

[4] Y. Li and Z.-L. Zhang, "Digraph Laplacian and the Degree of Asymmetry," *Internet Mathematics*, vol. 8, no. 4, pp. 381–401, 2012.

[5] F. Chung, "Laplacians and the cheeger inequality for directed graphs," *Annals of Combinatorics*, vol. 9, no. 1, pp. 1–19, Apr. 2005, ISSN: 0219-3094. DOI: 10.1007/s00026-005-0237-z. [Online]. Available: https://doi.org/10.1007/s00026-005-0237-z.

[6] R. Shafipour, A. Khodabakhsh, G. Mateos, and E. Nikolova, "A directed graph fourier transform with spread frequency components," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 946–960, 2019. DOI: 10.1109/TSP.2018.2886151.

[7] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3042–3054, 2014. DOI: 10.1109/TSP.2014.2321121.

[8] J. Domingos and J. M. F. Moura, "Graph fourier transform: A stable approximation," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4422–4437, 2020. DOI: 10.1109/TSP.2020.3009645.

[9] Y. Chen, C. Cheng, and Q. Sun, "Graph fourier transform based on singular value decomposition of the directed laplacian," *Sampling Theory, Signal Processing, and Data Analysis*, vol. 21, no. 2, p. 24, Aug. 2023, ISSN: 2730-5724. DOI: 10.1007/s43670-023-00062-w. [Online]. Available: https://doi.org/10.1007/s43670-023-00062-w.

[10] V. Satuluri and S. Parthasarathy, "Symmetrizations for clustering directed graphs," Mar. 2011, pp. 343–354. DOI: 10.1145/1951365.1951407.

[11] S. Furutani, T. Shibahara, M. Akiyama, K. Hato, and M. Aida, "Graph signal processing for directed graphs based on the hermitian laplacian," Jun. 2019.

[12] H. Kitamura, H. Yasuda, Y. Tanaka, and S. Muramatsu, "Realization of digraph filters via augmented gft," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 2885–2889. DOI: 10.1109/ICIP49359.2023.10222618.

[13] S. Kwak, L. Shimabukuro, and A. Ortega, "Frequency analysis and filter design for directed graphs with polar decomposition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 9661–9665. DOI: 10.1109/ICASSP48485.2024.10446657.

[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge University Press, 2012.

[15] N. Perraudin, J. Paratte, D. Shuman, *et al.*, "GSPBOX: A toolbox for signal processing on graphs," *ArXiv e-prints*, Aug. 2014. arXiv: 1408.5781 [cs.IT].

[16] R. Kumagai, H. Kojima, H. Higashi, and Y. Tanaka, "Gsp-traffic dataset: Graph signal processing dataset based on traffic simulation," in *Graph Signal Processing Workshop 2024, Delft, The Netherlands*, Jun. 2024.

Can We Estimate Purchase Intention Based on Zero-shot Speech Emotion Recognition?

Ryotaro Nagase*†, Takashi Sumiyoshi*, Natsuo Yamashita*, Kota Dohi*, and Yohei Kawaguchi*
* Hitachi, Ltd., Japan
E-mail: yohei.kawaguchi.xk@hitachi.com
† Ritsumeikan University, Japan

***Abstract*—This paper proposes a zero-shot speech emotion recognition (SER) method that estimates emotions not previously defined in the SER model training. Conventional methods are limited to recognizing emotions defined by a single word. Moreover, we have the motivation to recognize unknown bipolar emotions such as "I want to buy - I do not want to buy." In order to allow the model to define classes using sentences freely and to estimate unknown bipolar emotions, our proposed method expands upon the contrastive language-audio pre-training (CLAP) framework by introducing multi-class and multi-task settings. We also focus on purchase intention as a bipolar emotion and investigate the model's performance to zero-shot estimate it. This study is the first attempt to estimate purchase intention from speech directly. Experiments confirm that the results of zero-shot estimation by the proposed method are at the same level as those of the model trained by supervised learning.**

***Index Terms*—Speech emotion recognition, contrastive language-audio pre-training, zero-shot, purchase intention**

I. Introduction

Speech emotion recognition (SER) is a technique to estimate emotions conveyed by speech. This technique can be applied to the speech analysis of call centers [1], spoken dialogue agent [2], e-learning system [3], and mental health analysis [4], etc.

Many researchers have proposed various methods to improve the performance of SER. Especially, in recent years, there are many methods using deep learning such as SER with self-supervised learning model [5], [6]. In these methods, the model of SER is often trained by supervised-learning. Therefore, the trained model can recognize predefined seen emotions, while it is difficult to recognize unseen emotions. To alleviate this limitation, previous studies have proposed methods of zero-shot SER that can recognize undefined emotions. For example, X. Xu, et al. proposed the method using auditory affective descriptors (AAD) to transfer knowledge from the seen domain to the unseen domain [7], and the method using reconstructed semantic prototypes and data augmentation for the unseen domain [8].

The challenge of zero-shot SER is how to define the classes of emotions. Conventional methods do not assume the estimation of emotion classes that cannot be defined by a single word. Therefore, it is difficult to recognize the emotions expressed by text, such as "the feeling of willing to buy (purchase intention)." In order to realize SER capable of zero-shot recognizing various emotions, it is necessary to have a framework that can freely define classes during the training and estimation phase. We also have the motivation to realize the zero-shot estimation of bipolar emotions, such as "I want to buy - I do not want to buy."

To meet these requirements, in this study, we propose a new contrastive language-audio pre-training (CLAP) method that can perform zero-shot estimation of bipolar emotions. CLAP can decide the emotion freely because it can represent the classification class by the sentence. In our method, we define the multi-class emotions with a bipolar sub-class and train the model with CLAP for each emotion. We use six basic emotions with a bipolar sub-class as the multi-class emotions. We expect this model trained by the proposed method to correctly classify the emotions using the knowledge of six basic emotions, even if the estimated class is an unknown bipolar emotion. We also focus on purchase intention as a bipolar emotion and experiment on whether the model trained by the proposed method can zero-shot estimate purchase intention. The contributions of this study are as follows.

- This is the first method for zero-shot SER capable of estimating bipolar emotions, such as purchase intention.
- This study is the first time to directly estimate purchase intention from speech.

This paper proceeds as follows. In Section2, we describe the method of the basic CLAP, proposed method with CLAP, and the data augmentation. In Section3, we explain the setup for the experiment and show the results. Finally, in Section4, we present our conclusions and future work.

II. Methodology

A. CLAP

CLAP is the training scheme that calculates the similarity or dissimilarity between acoustic and linguistic embeddings [9]. This method makes it possible for us to define the categories during the estimation phase and classify speech into unseen categories. Moreover, the model can estimate more various categories because CLAP can define them with the sentence.

The overviews of CLAP during the training or estimation phase are shown in Figures 1 and 2, respectively. The inputs are audio and text of categories. Let $\mathbf{x} \in \mathbb{R}^{N \times T}$, $\mathbf{y} \in \mathbb{R}^{N \times L}$ be the audio and the text, respectively, where T and L are the length of them, and N is the size of mini-batch. Note that $\mathbf{x}$, $\mathbf{y}$ have correspondence in mini-batch.

The inputs are converted to the embeddings by the audio and text encoders. The audio encoder $f_{\mathrm{a}}(\cdot)$ and text encoder $f_{\mathrm{t}}(\cdot)$

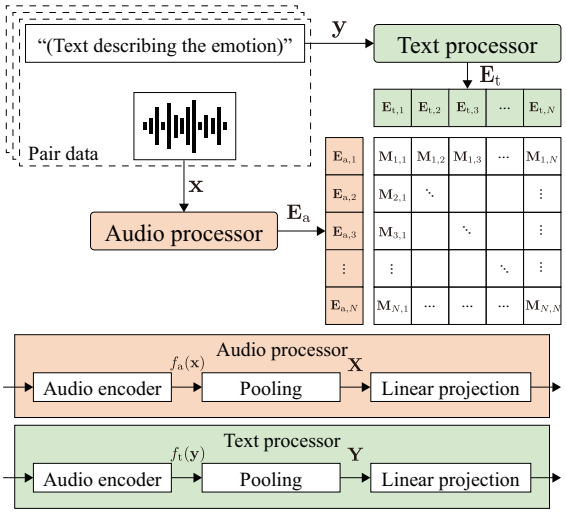


Fig. 1. Overview of clap during the training phase

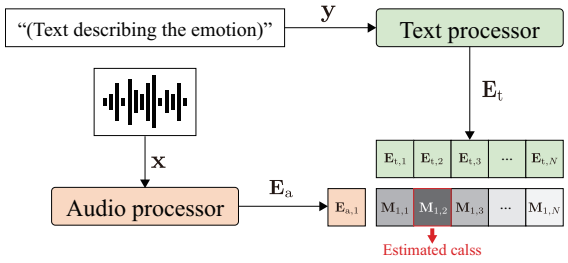


Fig. 2. Overview of clap during the estimation phase

are given in Equation 1 and 2, respectively. Let $\mathbf{X} \in \mathbb{R}^{N \times D_a}$, $\mathbf{Y} \in \mathbb{R}^{N \times D_t}$ be the audio embeddings and text embeddings, where D_a, D_t are the number of dimensions of each encoder's output. Moreover, Pooling is applied to average or select the embeddings from each encoder.

$$\mathbf{X} = \text{Pooling}(f_a(\mathbf{x})) \quad (1)$$
$$\mathbf{Y} = \text{Pooling}(f_t(\mathbf{y})) \quad (2)$$

Those embeddings are input into the linear projection, which are given in Equation 3 and 4. Let $\mathbf{E}_a \in \mathbb{R}^{N \times D}$, $\mathbf{E}_t \in \mathbb{R}^{N \times D}$ be the fixed embeddings of D dimensions from each linear projection. Also, $\text{Linear}_a(\cdot)$, $\text{Linear}_t(\cdot)$ are the linear projection for audio and text, respectively.

$$\mathbf{E}_a = \text{Linear}_a(\mathbf{X}) \quad (3)$$
$$\mathbf{E}_t = \text{Linear}_t(\mathbf{Y}) \quad (4)$$

During the training phase, we calculate the similarity matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ given in Equation 5. Let τ be the temperature parameter.

$$\mathbf{M} = \tau(\mathbf{E}_t \cdot \mathbf{E}_a^\top) \quad (5)$$

After that, we calculate and optimize the symmetric loss function [10] $\mathcal{L}(\cdot)$ given in Equation 6, where $\text{CE}(\cdot)$ is the function of cross entropy, and $\hat{\mathbf{M}} \in \mathbb{R}^{N \times N}$ is the ground truth matrix.

$$\mathcal{L}(\mathbf{M}, \hat{\mathbf{M}}) = \frac{1}{2}(\text{CE}(\mathbf{M}, \hat{\mathbf{M}}) + \text{CE}(\mathbf{M}^\top, \hat{\mathbf{M}})) \quad (6)$$

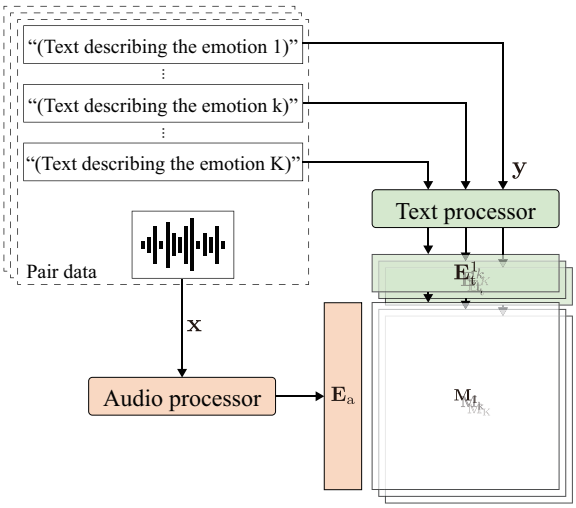


Fig. 3. Overview of the proposed method

Optimizing this loss function, we can train the model to minimize the loss between the ground truth and the estimated values and maximize the loss between the incorrect answer and the estimated values.

During the estimation phase, we defined the categorical classes by words or sentences. Then, we input the speech and texts represented the categories into the encoder and linear projection corresponding to each modalities, and obtain fixed-length vectors. Finally, the similarity between speech and text is calculated using the fixed-length vectors, and the class indicated by the text of the highest similarity is the estimation result.

This method has been applied to SER. For example, in the original paper on CLAP, zero-shot SER was performed using the CLAP model trained with environmental sound datasets [9]. Y. Pan et al. also proposed GEmo-CLAP for gender and emotion classification and improved the accuracy rate of four emotion classifications that were not zero-shot [11]. However, there is still no research on recognizing emotion-related classes such as purchase intention using the model of zero-shot SER trained with emotional speech datasets.

B. Zero-shot SER with multi-class multi-task CLAP

We propose the training method of the zero-shot SER for multi-class emotional speeches, which is called multi-class multi-task CLAP. The overview of the proposed method is shown in Figure 3.

During the training phase, the model outputs the similarity matrix of the multi-class prepared for each emotion. After calculating the contrastive loss with the ground truth matrix in each emotion, we summarize and optimize them. Equation 7 formulates the calculation of the loss $\mathcal{L}_{\text{all}}$. Let K be the number of emotion classes, k be the index of an emotion class, and $\mathbf{M}_k \in \mathbb{R}^{N \times N}$, $\hat{\mathbf{M}}_k \in \mathbb{R}^{N \times N}$ be the similarity matrix and ground truth matrics in each emotion, respectively.

$$\mathcal{L}_{\text{all}} = \sum_{k=1}^{K} \mathcal{L}(\mathbf{M}_k, \hat{\mathbf{M}}_k) \quad (7)$$

TABLE I
TEMPLATE OF THE PROMPT FOR PARAPHRASING

 # Prompt for paraphrasing
You are an imaginative assistant. Given the emotion description, please generate ten paraphrase sentences in Japanese. Note that you cannot add the prohibited word. ### Input and output Emotion description: [e.g. 私の気持ちは覚醒しています (My emotion is aroused.)] Prohibited word: [e.g. 覚醒 (aroused)] Paraphrase sentences:

During the estimation phase, we also define the categories and classify speeches as before. The model trained by the proposed method can recognize multi-class emotions with the sub-class. Therefore, we expect the proposed method will enable the zero-shot SER to estimate emotions such as unknown bipolar emotions.

C. The data augmentation of multi-class by paraphrasing

We extend the dataset by paraphrasing the text indicating the multi-class to increase the vocabulary and grammar of text associated with speech. Paraphrasing was performed using GPT4, a large-scale language model provided by OpenAI[1]. The template of prompts given to GPT4 is shown in Table I. GPT4 performs better in English than Japanese [12], so the prompt is written in English. The prompt includes the text of multi-class and the forbidden word. If the forbidden word is included in the output, we modify it by hand.

III. EXPERIMENTS

We experimented with the binary classification task of whether purchase intention from speech. In the following, we explain the experimental setup and results.

A. Dataset

Before the experiment, we created the internal dataset which has annotated emotional utterances in Japanese. First, we collected emotional utterances by role-playing conversations. We prepared six scenarios related to conversations between salespeople and customers. Five Japanese speakers talked freely for each scenario. Note that we did not prepare sentences for reading. The sampling rate of the recorded utterance was 44,100 Hz. The recorded data were separated into sessions for each scenario, and we obtained the dataset consisting of six sessions. Afterward, three annotators listened to collected utterances and scored the intensity values of the six emotions (pleasant-unpleasant, aroused-sleepy, dominant-submissive, credible-doubtful, interested-indifferent, positive-negative) on a seven-point scale (1.very low–7.very high). These emotions were assigned according to the related work [13]. In addition, these annotators scored the intensity values indicating the purchase intention on a seven-point scale for only customer's data at corrected utterances.

[1]https://openai.com/gpt-4

Finally, we arranged the labels for the classification task. The distribution of scores of each annotator was normalized by the mean and variance of the distribution of scores given by all annotators. The normalized labels exclude the most frequent value and sets it as a threshold. The score below the threshold are relabeled as negative labels (0) and the score above the threshold are relabeled as positive labels (1). The number of utterances in two classes of six emotions was 2,598, and the number of utterances with the two classes of the purchase intention was 940. We converted the sampling rate of each utterance to 16,000Hz.

During the training and estimation phase using the proposed method, we used descriptions that explain the emotions corresponding to the scores aggregated above. Table II shows the correspondence between the class scores of each emotion and the descriptions of each emotion. In addition, we experimented with the augmented data by paraphrasing, which was 11 times the number of the original dataset. In the evaluation, we used one of the six sessions as the testing data and others as the training data.

B. Models and metrics

We explain the model structure of the proposed method. The audio encoder consisted of Japanese HuBERT [14] and linear projection. The output dimensions of them were 768 and 512, respectively. Before passing embeddings into linear projection, we averaged the audio embeddings obtained from HuBERT in the time direction. The text encoder consisted of Japanese DistilBERT [15] and linear projection. The output dimensions of them were also 768 and 512, respectively. Before passing embeddings into linear projection, we only used class token from the embeddings obtained from DistilBERT. We used the pre-trained parameters of HuBERT[2] and DistilBERT[3] provided by Hugging Face [16].

The baseline in this paper was the model to estimate purchase intention by supervised learning instead of CLAP. This model structure was a combination of HuBERT, linear projection, and one fully connected layer, and we used the pre-trained HuBERT in the same way as the proposed method. The threshold for binary classification on the baseline is defined by the Youden index of the receiver operatorating characteristic (ROC) curve. The number of epochs was 300, the batch size was 64, the learning rate was 0.000001, and the optimization method was Adam [17]. The loss functions of the baseline and the proposed method were binary cross entropy loss and symmetric cross entropy loss, respectively.

We used weighted accuracy (WA), unweighted accuracy (UA), and recall of each category as the evaluation metrics. Note that WA is the total number of correct answers to the total data, and UA is the average recall of each category. The higher these metrics are, the more correctly the performance can estimate purchase intention. Moreover, we calculated the frequency rate of matching the random value and the correct

[2]rinna/japanese-hubert-base
[3]laboro-ai/distilbert-base-japanese

TABLE II
CORRESPONDENCE BETWEEN TWO CLASSES OF EACH EMOTION AND THE DESCRIPTIONS OF EACH EMOTION

	Negative label (0)	Positive label (1)
Unpleasant-pleasant	私は不快な気持ちです (My emotions are uncomfortable.)	私は快な気持ちです (My emotions are pleasant.)
Sleepy-aroused	私の気持ちは睡眠しています (My emotions are asleep.)	私の気持ちは覚醒しています (My emotions are awake.)
Submissive-dominant	私の気持ちは従属的です (My emotions are subservient.)	私の気持ちは支配的です (My emotions are dominant.)
Doubtful-credible	私は不信に思います (I distrust.)	私は信頼しています (I trust.)
Indifferent-interested	私は無関心です (I am indifferent.)	私は関心があります (I am interested.)
Negative-positive	私は否定的です (I am negative.)	私は肯定的です (I am positive.)

TABLE III
WA, UA, AND THE RECALL OF PURCHASE INTENTION

	Text indicating classes (ja)	No augmentation				Augmentation by paraphrasing			
		WA	UA	Purchase intention		WA	UA	Purchase intention	
				No	Yes			No	Yes
Random	-	49.5	49.4	48.8	50.0	-	-	-	-
Supervised [Baseline]	-	**74.0**	**69.2**	**78.5**	60.0	-	-	-	-
Zero-shot [Ours]	(1) 私は買う気 { なし, あり } です (I am {unwilling / willing} to buy.)	60.2	69.1	51.6	**86.7**	61.8	62.3	61.3	**63.3**
	(2) 私は購買意欲 { なし, あり } (I {do not have / have} the purchase intention.)	64.2	60.5	67.7	53.3	65.0	63.3	66.7	60.0
	(3) 私は { 欲しくない, 欲しい } です (I {do not want / want} to buy it.)	61.0	61.8	60.2	63.3	**73.2**	**69.8**	**76.3**	**63.3**

answer as a chance rate. In the evaluation of the proposed method, we used the three texts. They all indicated the purchase intention, while the words used in the sentence differed. We used these texts to compare the proposed method with the chance rate and baseline.

C. Results

Table III shows WA, UA, and recall of each category. The result of the random was about 50% for both WA and UA. Comparing the result of the random with the baseline, WA was 24.5%, and UA was 19.8% higher in the baseline result than in the random. These results indicate that the baseline model does not make a random estimation. It also shows that it is possible to estimate purchase intention from speech using deep learning. Comparing the result of the random with the proposed methods, WA and UA were higher in all proposed methods than in the random. These results show that the zero-shot estimation by the proposed method is not random, and these methods can train the model to estimate unknown bipolar emotions.

Without data augmentation, the results show that the recall with purchase intention "Yes" using text (1) is higher in the proposed method than in the baseline. The UA using text (1) and the recall with purchase intention "Yes" using text (3) were similar to the baseline result. On the other hand, the recalls with purchase intention "No" using text (1) and "Yes" using text (2) were lower than 60%. It indicates that the model trained by the proposed method cannot represent the correspondence well between the semantic information of text and speech under some conditions.

All evaluation scores with data augmentation are improved overall than those without data augmentation. Those scores equal the baseline, especially when using text (3). This indicates that data augmentation make it possible for the model to train the correspondence between text and speech semantic information relatively more easy. These results indicate that the zero-shot SER trained by the proposed method can estimate purchase intention that is unseen class by data augmentation of paraphrasing and appropriate creation of text indicating classes.

We also investigated the significant difference between the evaluation scores of the baseline and the proposed method using text (3) through a sign test. If there was no significant difference between them, we could assume that the number of samples that were correct by the baseline and incorrect by the proposed method was the same as those that were correct by the proposed method and incorrect by the baseline. We performed a two-tailed sign test based on the above hypothesis.

The p-value p about the baseline and the proposed method using text (3) without data augmentation was less than 0.05 ($p < 0.05$). On the other hand, the p-value p about the baseline and the proposed method using text (3) with data augmentation was more than 0.05 ($p > 0.05$). These results indicated that the significant difference between the baseline and the proposed method decreased with data augmentation. In other words, when using text (3) along with data augmentation, the proposed method can provide zero-shot estimates comparable to the supervised method.

IV. Conclusion

This study proposed a new method for zero-shot speech emotion recognition. The proposed method is a multi-class multi-task CLAP that can recognize unknown bipolar emotions. We found that zero-shot SER trained by the proposed method with paraphrasing data augmentation can be used to estimate purchase intention with the same accuracy as a supervised learning model. In the future, we would like to expand speech data with multi-class labels and perform zero-shot inference for generously related classes other than purchase intention.

References

[1] M. Bojanić, V. Delić, and A. Karpov, "Call redistribution for a call center based on speech emotion recognition," *Applied Sciences*, vol. 10, no. 13, 2020, ISSN: 2076-3417. DOI: 10.3390/app10134653.

[2] Y. Gao, Z. Pan, H. Wang, and G. Chen, "Alexa, My love: Analyzing reviews of Amazon Echo," in *Proc. SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI – 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, Guangzhou, China, Oct. 2018, pp. 372–380. DOI: 10.1109/SmartWorld.2018.00094.

[3] W. Li, Y. Zhang, and Y. Fu, "Speech emotion recognition in e-learning system based on affective computing," in *Proc. ICNC 2007 – Third International Conference on Natural Computation*, vol. 5, Haikou, China, Aug. 2007, pp. 809–813. DOI: 10.1109/ICNC.2007.677.

[4] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "LSSED: A large-scale dataset and benchmark for speech emotion recognition," in *Proc. ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, Jun. 2021, pp. 641–645. DOI: 10.1109/ICASSP39728.2021.9414542.

[5] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *Proc. ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May. 2022, pp. 6922–6926. DOI: 10.1109/ICASSP43922.2022.9747870.

[6] L. W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *Proc. ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, Jun. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095036.

[7] X. Xu, J. Deng, N. Cummins, Z. Zhang, L. Zhao, and B. W. Schuller, "Autonomous emotion learning in speech: A view of zero-shot speech emotion recognition," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 949–953. DOI: 10.21437/Interspeech.2019-2406.

[8] X. Xu, J. Deng, Z. Zhang, Z. Yang, and B. W. Schuller, "Zero-shot speech emotion recognition using generative learning with reconstructed prototypes," in *Proc. ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, Jun. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10094888.

[9] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, Jun. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095889.

[10] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML'21 – 38th International Conference on Machine Learning (ICML)*, vol. 139, Online, Jul. 2021, pp. 8748–8763.

[11] Y. Pan, Y. Hu, Y. Yang, *et al.*, *GEmo-CLAP: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition*, arXiv.preprint, arXiv:2306.07848, 2023. arXiv: 2306.07848 [cs.CL].

[12] OpenAI, *GPT-4 technical report*, arXiv.preprint, arXiv:2303.08774, 2023. arXiv: 2303.08774 [cs.CL].

[13] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, 2011, ISSN: 0167–6393. DOI: https://doi.org/10.1016/j.specom.2010.08.002.

[14] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 3451–3460, Oct. 2021, ISSN: 2329-9290. DOI: 10.1109/TASLP.2021.3122291.

[15] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *DistilBERT, a distilled version of BERT: Smaller, faster,*

cheaper and lighter, arXiv.preprint, arXiv:1910.01108, 2019.

[16] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP 2020 – 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, Oct. 2020, pp. 38–45.

[17] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv.preprint, arXiv:1412.6980, 2014. arXiv: 1412.6980 [cs.CL].

Text-guided Visual Prompt Tuning with Masked Images for Facial Expression Recognition

Rongkang Dong, Cuixin Yang, and Kin-Man Lam.
The Hong Kong Polytechnic University
E-mail: {rongkang97.dong, cuixin.yang}@connect.polyu.hk, enkmlam@polyu.edu.hk

***Abstract*—Facial expression recognition (FER) has significantly advanced through the application of deep learning techniques for visual content classification. Recent research has explored the use of pre-trained language-image models, such as CLIP, which leverage natural language supervision to enhance image backbone training and facilitate the learning of general visual representations. Concurrently, visual prompt tuning has emerged as a method to minimize tuning overhead for downstream tasks by freezing the pre-trained backbone models and incorporating additional learnable parameters, known as visual prompts, into the model input. This strategy circumvents the need to update the entire neural network, focusing instead on optimizing visual prompts for specific tasks. In this study, we propose a novel tuning scheme, namely Text-guided Visual Prompt Tuning with Masked facial images (T-VPT-M), for both basic and compound FER. Our method utilizes natural language supervision for visual prompt learning and employs a random masking mechanism to adapt visual prompts to diverse informative facial regions. Experimental results on three real-world datasets, encompassing both basic and compound facial expressions, demonstrate the efficacy of the T-VPT-M scheme.**

I. Introduction

Facial expressions are the most essential non-verbal component in human communication [1]. Automatic facial expression recognition (FER) plays a significant role in facilitating human-machine interaction [2]. Over the past several decades, facial expression analysis and recognition have been studied and realized through machine learning techniques. Early research predominantly utilized traditional machine learning methods [3]. However, recent advancements in FER have been dominated by deep-learning-based methods [4], [5] due to their remarkable performance in computer vision tasks [6]–[8]. Previous studies [9], [10] primarily focused on recognizing a neutral face and six basic facial expressions defined by Ekman and Friesen [11]: happiness, sadness, anger, disgust, fear, and surprise. Du et al. [12] found that a facial expression can convey more than one basic emotion, introducing the concept of compound facial expressions. Consequently, several works [13], [14] have explored the compound FER task using deep-learning-based methods.

Research on deep learning for FER mainly relies on adapting and fine-tuning pre-trained models. The resource consumption for full model training escalates with an increase in trainable parameters, especially when employing modern vision transformer (ViT) architectures [5]. One alternative approach is to fine-tune the classification head (e.g., a linear layer) only, which significantly reduces training costs but may not achieve the same performance as full model fine-tuning. Recent advancements in visual prompt tuning [15], [16] offer a more efficient and effective training method by fixing the backbone models while tuning both the classification head and the visual prompts. Visual prompts are additional learnable parameters appended to the input images or embeddings. However, existing visual prompt tuning methods primarily extract features from visual inputs using an image encoder, overlooking the potential of incorporating other modalities that can augment the semantic understanding of an image.

Natural language plays a crucial role in enhancing human comprehension of visual information within images, a principle that extends to machine understanding as well. Research has demonstrated that pre-trained image models, when supervised by natural language descriptions, exhibit strong performance across various tasks [17]. For instance, a simple text description, such as “a photo of a happy face”, effectively conveys the visual content of a face image depicting a happy facial expression. Vision-language learning enables deep models to harness useful semantic information from image-text pairs.

In this paper, we introduce a novel approach, called Text-guided Visual Prompt Tuning with Masked Images (T-VPT-M), for FER. Building upon the concept of supervising vision models with natural language, this scheme allows a learnable visual prompt to be guided by text descriptions, thereby enhancing the semantic understanding of deep vision models. We further adopt a masking approach that randomly masks regions of the facial images [18], [19]. Previous masking methods were used for self-supervised learning, enabling backbone models to achieve high capacity and generalization by fully utilizing the non-masked regions. Our text-guided visual prompt tuning scheme, combined with the masking method, enables the visual prompt to learn general facial expression knowledge from diverse informative facial regions, rather than focusing solely on explicit facial appearances. The contributions of this study are summarized as follows:

- We introduce a simple yet effective text-guided visual prompt tuning scheme for both basic and compound FER tasks. This scheme adopts natural language supervision for visual prompt learning, allowing the visual prompts to incorporate the semantic knowledge learned from text descriptions.
- We combine a random masking method with the tuning

scheme. Masking regions of facial images forces visual prompts to encode more effective and general knowledge of facial expressions from diverse informative facial appearances.
- We conduct several experiments to demonstrate the effectiveness of the text-guided visual prompt learning scheme using two popular pre-trained deep models, namely ResNet-50 and ViT-B/16, on three facial expression datasets.

II. Related Work

A. Vision-language models

Visual-language models have become a prominent research topic in various deep-learning applications. These models can be used to generate images from text descriptions (i.e., text-to-image models) [20] and generate text descriptions from images (i.e., image-to-text models) [21], [22]. Additionally, by leveraging a large number of image-text pairs, visual-language models, such as the Contrastive Language-Image Pre-training (CLIP) model [17], perform contrastive training between image features, encoded by an image encoder (e.g., ResNet [4], ViT [5]), and corresponding text features, extracted by a text encoder (e.g., transformer [23]). These visual-language models aim to learn general visual representations. Contrastive learning supervised by text description is effective and enables vision models to achieve superior zero-shot performance. Subsequent works have further improved the performance of vision-language models and adapted them for other tasks [19], [24], [25]. In this study, we introduce a text-guided visual prompt tuning scheme with random masking for basic and compound FER, utilizing text embeddings to supervise the learning of visual prompts.

B. Facial expression recognition

In recent years, deep learning has significantly accelerated the development of FER [2]. Research has primarily focused on static and dynamic basic FER [9], [10], with a few studies exploring compound FER [13], [14], which represents a further step towards real-world FER. In addition to using ResNet [4], studies [26], [27] have also adopted ViT structures [5] to improve FER performance.

Beyond unimodal analysis, multi-modal facial behavior analysis [28]–[32] has emerged, utilizing modalities other than vision (e.g., language and audio). Linguistic and audio modalities provide complementary information that fosters a deep understanding of facial expressions. Praveen et al. [31] proposed an incongruity-aware cross-attention model to harness the benefits of complementary relationships between audio and visual modalities for video emotion recognition, capturing inter-modal interactions. Zhang et al. [30] employed a text-embedded framework for facial behavior analysis and introduced a two-stage contrastive learning scheme. Li et al. [29] devised CLIPER, a unified vision-language model for FER, which introduces a set of different learnable text descriptions for each facial expression. Zhou et al. [32] proposed a cross-modal emotional-aware prompting framework for FER, which considers multi-modal prompting. Differently, we introduce masked facial images for text-guided visual prompt learning, which enables the visual prompts to adapt to different facial regions, thereby improving FER performance.

C. Prompt tuning

Existing deep learning methods typically require fine-tuning the entire pre-trained model to adapt it to downstream tasks, which demands substantial computational resources. Prompt learning has extended to the computer vision community from the natural language processing community. Visual prompt tuning [15], [16] mitigates high training overhead by introducing additional negligible learnable parameters, i.e., visual prompts, during fine-tuning, while keeping the backbone models frozen. This approach can even achieve superior performance compared to the full fine-tuning method. Zhou et al. [33] proposed CoOp, an approach designed to combat the difficulty of performance variance caused by word changes in text descriptions. In contrast to our T-VPT-M scheme, CoOp optimizes the learnable text contents from text descriptions.

III. Text-guided Visual Prompt Tuning with Masked Images

A. Preliminaries

1) CLIP: The Contrastive Language-Image Pre-training (CLIP) method enables deep neural networks (DNNs) to learn general visual representations from natural language supervision [17]. This method utilizes a dataset comprising 400 million image-text pairs for contrastive training. An image encoder is employed to extract the image feature $\boldsymbol{i}$, while a text encoder is adopted to extract the text feature $\boldsymbol{t}$. The text encoder processes descriptions, such as "A photo of a {class}", where "{class}" represents the category name. Given a batch of image-text pairs, the training objective of CLIP is to optimize the visual and text encoders by computing the cross-entropy loss between the image features and the text features, which can be written as follows:

$$\mathcal{L}_i = -\frac{1}{B}\log\sum_{j=1}^{B}\frac{\exp(S(\boldsymbol{i}_j,\boldsymbol{t}_j)/\tau)}{\sum_{k=1}^{B}\exp(S(\boldsymbol{i}_j,\boldsymbol{t}_k)/\tau)}, \tag{1}$$

$$\mathcal{L}_t = -\frac{1}{B}\log\sum_{j=1}^{B}\frac{\exp(S(\boldsymbol{t}_j,\boldsymbol{i}_j)/\tau)}{\sum_{k=1}^{B}\exp(S(\boldsymbol{t}_j,\boldsymbol{i}_k)/\tau)}, \tag{2}$$

$$\mathcal{L}_{CLIP} = \frac{1}{2}(\mathcal{L}_i + \mathcal{L}_t), \tag{3}$$

where B is the batch size, $S()$ denotes the cosine similarity, j and k are the batch index, and τ represents the learned temperature parameter.

2) Visual Prompt Tuning: Visual prompt tuning [15], [16] transfers the tuning concept from fine-tuning the backbone model to tuning visual prompts, with the backbone model frozen. This approach can significantly reduce training resources, as the number of additional parameters is much smaller than that of the entire model. Jia et al. [15] introduced visual prompts that are prepended to the image patch features.

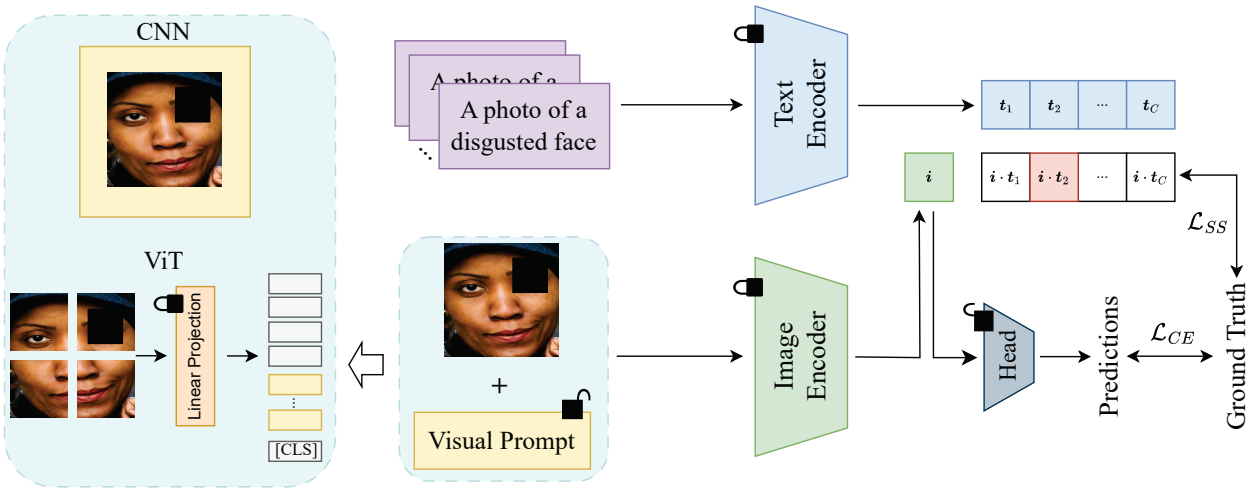


Fig. 1. Framework of text-guided visual prompt tuning with image masking for FER.

They proposed two variants of visual prompt tuning (VPT) for ViTs: VPT-Shallow and VPT-Deep. VPT-shallow only inserts visual prompts into the input of the first Transformer layer, while VPT-Deep inserts visual prompts into the input of every Transformer layer. Both variants of VPT introduce negligible additional storage and computational cost. For Convolutional Neural Networks (CNNs), visual prompts are evenly padded around the input images [15], [16].

B. T-VPT-M for FER

Fig. 1 illustrates our framework of text-guided visual prompt tuning for FER. In our study, we adopt two commonly used DNN models as the image encoder: ResNet [4] and Vision Transformer [5]. Consequently, we employ two types of visual prompts for model tuning. We pad the visual prompts around the input images for ResNet, while we prepend the visual prompts to the input of the first Transformer layer for ViT. We utilize the VPT method [15], which employs text prompts with a template of "A photo of a {class} face", where "{class}" represents the category name of the facial expression, such as happy or sadly disgusted.

The visual inputs to the framework comprise masked facial images and learnable visual prompts. The visual prompts, with a width of n_c, are padded evenly around the input images for CNN, while n_v visual prompts are prepended to the patch embeddings for ViT. Additionally, a masking method enables the visual prompts to encode more general knowledge for FER. This method randomly masks an $m \times m$ region for each facial image with a probability of p. By extracting a visual feature $\boldsymbol{i}'$ of each masked image, the training objective of our framework can be written as follows:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{SS}, \tag{4}$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss between the predicted probability distribution and the one-hot ground truth, and $\mathcal{L}_{SS}$ is the softmax loss of the similarity score computed in the feature space, which can be written as follows:

$$\mathcal{L}_{SS} = -\frac{1}{B}\sum_{j=1}^{B} \log \frac{\exp(S(\boldsymbol{i}'_j, \boldsymbol{t}_y)/\tau)}{\sum_{c=1}^{C} \exp(S(\boldsymbol{i}'_j, \boldsymbol{t}_c)/\tau)}. \tag{5}$$

In Eq. (4), λ_1 denotes the hyperparameter for balancing the two loss terms. Note that $\mathcal{L}_{SS}$ does not work for optimizing classification head. In Eq. (5), $\boldsymbol{t}_y$ represents the text feature of the ground truth class y, C is the number of classes, and c is the class index. Learnable parameters in our framework only include the visual prompt and classification head, which introduce negligible additional parameters and reduce the training resource.

IV. Experiment

A. Datasets

We conducted experiments on the RAF-DB [34] basic subset (RAF-DB_B) and compound subset (RAF-DB_C), as well as the SFEW-2 dataset [35]. RAF-DB [34] contains 29,672 real-world facial images collected from the Internet with manual annotations. RAF-DB_B includes 12,271 training samples and 3,068 test samples, labeled with the neutral emotion and six basic emotions. RAF-DB_C contains eleven categories of compound facial expression images, with 3,162 samples used for training and 792 samples used for testing. The SFEW-2 database [35] consists of over 1,500 facial images annotated with the neutral emotion and six basic emotions. The database is divided into three sets: 958 samples for training, 436 samples for validation, and 372 samples for testing. We used the validation set for evaluation since the ground-truth labels for the test set were not provided. The effectiveness of our scheme was evaluated using overall accuracy as the performance metric.

B. Implementation Details

Our model was implemented in PyTorch [36]. We used pre-trained ResNet-50 [4] and ViT-B/16 [5] as the image

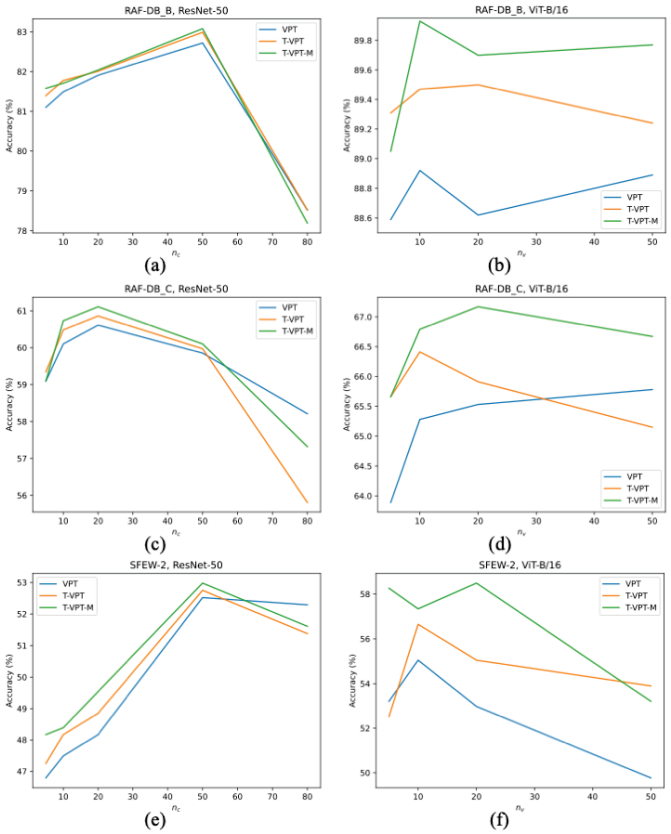


Fig. 2. Experiments on visual prompts in VPT-based methods with different n_c and n_v for ResNet-50 and ViT-B/16.

TABLE I
EXPERIMENTAL RESULTS ON COMPARISON AMONG DIFFERENT METHODS. "TRAIN PARAMS" DENOTES THE TOTAL NUMBER OF TRAINABLE PARAMETERS WITH RESPECT TO THE BACKBONE MODELS. FOR ALL VPT-BASED METHODS, n_c IS SET TO 50 FOR RESNET-50, AND n_v IS SET TO 20 FOR VIT-B/16. THE NUMBERS HIGHLIGHTED IN BOLD ARE THE BEST RESULTS AMONG ALL THE METHODS.

ResNet-50				
Methods	Train Params	RAF-DB_B	RAF-DB_C	SFEW-2
Full fine-tune	1.000187×	82.99	49.12	34.17
Linear	0.000187×	79.27	57.95	48.62
VPT	0.002912×	82.72	59.85	52.52
T-VPT (Ours)	0.002912×	82.99	59.97	52.75
T-VPT-M (Ours)	0.002912×	**83.08**	**60.10**	**52.98**
ViT-B/16				
Methods	Train Params	RAF-DB_B	RAF-DB_C	SFEW-2
Full fine-tune	1.000042×	89.21	63.38	53.21
Linear	0.000042×	85.43	61.11	54.82
VPT	0.000220×	88.62	65.53	52.98
T-VPT (Ours)	0.000220×	89.50	65.91	55.05
T-VPT-M (Ours)	0.000220×	**89.70**	**67.17**	**58.49**

encoders, along with the text encoders from CLIP [17]. Before conducting the experiments, MTCNN [37] was employed to align and crop the facial images in all the datasets. For data augmentation, we applied resizing, cropping, as well as random horizontal flipping. Facial images were resized to 256×256 ($(256 - 2n_c) \times (256 - 2n_c)$) and cropped to 224×224 ($(224 - 2n_c) \times (224 - 2n_c)$) before uniting visual prompts for ViT-B/16 (ResNet-50). We optimized the models using

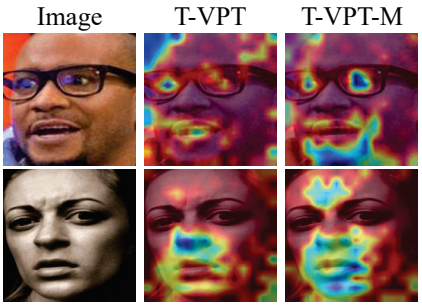


Fig. 3. Visualization of attention maps using GradCAM++ [38], [39]. The experiments used ViT-B/16 as the backbone on the RAF-DB_C dataset.

TABLE II
EXPERIMENTS ON RANDOM MASKING WITH DIFFERENT MASKING SIZE m AND PROBABILITY p. THE EXPERIMENTS WERE CONDUCTED WITH VIT-B/16 ON THE RAF-DB_C DATASET. THE BEST RESULT IS HIGHLIGHTED IN BOLD.

$p \backslash m$	32	64	128
0.3	66.04	66.79	66.67
0.5	67.05	67.17	**67.80**
0.7	66.54	65.66	67.05
0.9	65.91	65.28	65.78

stochastic gradient descent with an initial learning rate of 1.0, a momentum of 0.9, and a weight decay of 0.0001. The total number of training epochs was 100, with the initial 10 epochs used for warming up. We adopted a cosine annealing strategy to decrease the learning rate. The batch size was set to 128. We set λ_1 and τ to 1 for the training loss in the experiments. Unless other specified, the masking probability p was set to 0.5, and the mask size m was set to 32 and 64 for ResNet-50 and ViT-B/16, respectively.

C. Experimental Results

1) Baseline Comparison: We compare several methods, including full fine-tuning, linear, visual prompt tuning [15] (VPT), text-guided visual prompt tuning (T-VPT), and text-guided visual prompt tuning with masking (T-VPT-M), on different facial expression datasets to demonstrate the effectiveness of the proposed methods in FER. The full fine-tuning method updates all the parameters of the backbone and the classification head, while the linear method only tunes the head (i.e., a linear layer) with the backbone parameters frozen. VPT, T-VPT, and T-VPT-M introduce new learnable parameters, i.e., visual prompts, and therefore update both the head and the visual prompts. T-VPT employs natural language to supervise the learning of visual prompts. Based on T-VPT, T-VPT-M utilizes a random masking method that randomly masks a region of the facial image and incorporates visual prompts for visual feature extraction.

Tab. I illustrates the comparison results. We observe that the linear method underperforms compared to the full fine-tuning method, but it requires significantly fewer trainable parameters. In contrast, VPT provides a substantial accuracy increase over the linear method in both backbone models by inserting lightweight learnable parameters into the input. This

method still updates far fewer model parameters compared to the full fine-tuning method. Notably, VPT achieves over a 3% accuracy increase over the linear method in most experiments. By adopting the text-guided mechanism for VPT, i.e., the T-VPT method, the performance in terms of accuracy is further improved. Additionally, T-VPT-M, which incorporates a random masking mechanism, achieves even higher accuracy than T-VPT, especially with the ViT backbone. Compared to the VPT baseline method, our T-VPT-M method achieves over a 1% accuracy improvement in ViT-B/16, while showing a smaller accuracy improvement in ResNet-50.

2) Visual prompts: We conducted experiments to investigate how the number of visual prompt parameters influences the performance of VPT-based methods. The experiments were done with different n_c in $\{5, 10, 20, 50, 80\}$ and n_v in $\{5, 10, 20, 50\}$ for ResNet-50 and ViT-B/16, respectively. From Fig. 2, we can observe that the accuracy initially increases and then decreases as n_c or n_v becomes larger. A small number of visual prompt parameters results in insufficient transfer capacity from the pre-trained tasks to the target tasks, i.e., FER in this study. Conversely, if too many visual prompt parameters are introduced, the visual prompts will dominate the visual inputs, resulting in weakening the model's ability to leverage image content information. Another key observation in these experiments is that T-VPT-M consistently achieves the highest accuracy when the number of visual prompt parameters is relatively small. This demonstrates the superiority of our natural language supervision and random masking strategy for VPT with an appropriate number of visual prompt parameters.

3) Random masking: To evaluate the effectiveness of this method, we first compare the attention maps of T-VPT without random masking (i.e., T-VPT) and with random masking (i.e., T-VPT-M) in Fig. 3. We can observe that the random masking method makes the learned visual prompt to focus more on crucial and informative facial regions for FER.

Moreover, to investigate the impact of mask size and masking probability, we conducted experiments with mask sizes $m \in \{32, 64, 128\}$ and masking probabilities $p \in \{0.3, 0.5, 0.7, 0.9\}$. Larger mask sizes enable the visual prompts to learn facial expression knowledge from diverse informative facial regions by focusing on different facial appearances rather than the obvious ones. The probability p controls the input balance between the whole images and the masked images.

These experiments utilized the ViT-B/16 backbone on RAF-DB_C, with the results shown in Tab. II, with the findings indicating that an appropriate combination of mask size and masking probability can significantly improve accuracy compared to the non-masking T-VPT method. For instance, a mask size of 128 and a masking probability of 0.5 achieves an accuracy of 67.80%, which is 1.89% higher than the T-VPT method without image masking. Tab. II demonstrates that accuracy initially increases and then drops as the masking probability becomes larger, with the best results observed at a probability of 0.5 across all three mask sizes. Additionally, performance can improve when a larger mask size is paired with a smaller masking probability.

V. Conclusions

This paper presents a novel method called text-guided visual prompt tuning with random image masking (T-VPT-M) for facial expression recognition (FER). Our method utilizes text descriptions to supervise the learning of visual prompts. To ensure the visual prompts capture various facial regions, we employ a facial masking method that randomly masks facial regions in face images. Experimental results demonstrate that the combination of text-guided visual prompts and the masking method enhances FER performance. Additionally, we observe that FER accuracy initially increases but then decreases as the number of visual prompt parameters increases. Our results suggest that a modest mask size and masking probability can lead to optimal classification performance.

References

[1] A. Mehrabian *et al.*, *Silent messages*. Wadsworth Belmont, CA, 1971, vol. 8.

[2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[3] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[5] A. Alexey Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021, pp. 1–12.

[6] C. Yang, J. Xiao, Y. Ju, G. Qiu, and K.-M. Lam, "Improving robustness of single image super-resolution models with monte carlo method," in *IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 2135–2139.

[7] C. Yang, R. Dong, J. Xiao, *et al.*, "Geometric distortion guided transformer for omnidirectional image super-resolution," *arXiv preprint arXiv:2406.10869*, 2024.

[8] Z. Lyu, J. Xiao, C. Zhang, and K.-M. Lam, "Ai-generated image detection with wasserstein distance compression and dynamic aggregation," in *ICIP*, IEEE, 2024, pp. 3827–3833.

[9] R. Zhao, T. Liu, Z. Huang, D. P. Lun, and K.-M. Lam, "Spatial-temporal graphs plus transformers for geometry-guided facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2751–2767, 2023.

[10] R. Zhao, T. Liu, Z. Huang, D. P. Lun, and K.-M. Lam, "Geometry-aware facial expression recognition via attentive graph convolutional networks," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1159–1174, 2023.

[11] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124, 1971.

[12] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *National Academy of Sciences*, vol. 111, no. 15, E1454–E1462, 2014.

[13] R. Dong and K.-M. Lam, "Bi-center loss for compound facial expression recognition," *IEEE Signal Processing Letters*, pp. 641–645, 2024.

[14] R. Dong and K.-M. Lam, "Semi-supervised learning for compound facial expression recognition with basic expression data," in *International Workshop on Advanced Imaging Technology*, vol. 13164, 2024, pp. 314–319.

[15] M. Jia, L. Tang, B.-C. Chen, *et al.*, "Visual prompt tuning," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 709–727.

[16] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola, "Exploring visual prompts for adapting large-scale models," *arXiv*, 2022.

[17] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16 000–16 009.

[19] X. Dong, J. Bao, Y. Zheng, *et al.*, "Maskclip: Masked self-distillation advances contrastive language-image pretraining," in *CVPR*, 2023, pp. 10 995–11 005.

[20] D. Valevski, D. Lumen, Y. Matias, and Y. Leviathan, "Face0: Instantaneously conditioning a text-to-image model on a face," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–10.

[21] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, vol. 162, 2022, pp. 12 888–12 900.

[22] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023, pp. 19 730–19 742.

[23] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.

[24] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "Slip: Self-supervision meets language-image pre-training," in *ECCV*, 2022, pp. 529–544.

[25] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *ECCV*, 2022, pp. 696–712.

[26] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *ICCV*, 2021, pp. 3601–3610.

[27] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, "Vision transformer with attentive pooling for robust facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3244–3256, 2022.

[28] Y. Zheng, H. Yang, T. Zhang, *et al.*, "General facial representation learning in a visual-linguistic manner," in *CVPR*, 2022, pp. 18 697–18 709.

[29] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Cliper: A unified vision-language framework for in-the-wild facial expression recognition," *arXiv*, 2023.

[30] X. Zhang, T. Wang, X. Li, H. Yang, and L. Yin, "Weakly-supervised text-driven contrastive learning for facial behavior understanding," in *ICCV*, 2023, pp. 20 751–20 762.

[31] R. G. Praveen and J. Alam, "Incongruity-aware cross-modal attention for audio-visual fusion in dimensional emotion recognition," *IEEE Journal of Selected Topics in Signal Processing*, 2024.

[32] H. Zhou, S. Huang, F. Zhang, and C. Xu, "Ceprompt: Cross-modal emotion-aware prompting for facial expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[33] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[34] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *CVPR*, 2017, pp. 2852–2861.

[35] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *2011 IEEE International Conference on Computer Vision Workshops*, 2011, pp. 2106–2112.

[36] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NIPS*, vol. 32, 2019.

[37] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[38] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 839–847.

[39] J. Gildenblat and contributors, *Pytorch library for cam methods*, https://github.com/jacobgil/pytorch-grad-cam, 2021.

Noise-Robust Estimation of Early-part Room Impulse Responses based on Physics-Informed Neural Network with Dynamic Pulling Method

Ken Kurata*, Gen Sato*, Izumi Tsunokuni* and Yusuke Ikeda*
* Tokyo Denki University, JAPAN
E-mail: {24fmi09, 24udc03, 21udc02}@ms.dendai.ac.jp, yusuke.ikeda@mail.dendai.ac.jp

***Abstract*—Recently, physics-informed neural networks (PINN) have been applied to the problem of estimating the early room impulse responses (RIR) from a small number of microphones. PINNs use two types of loss functions: one corresponding to physical laws and the other to data errors. A challenge with PINNs is that when the gradients associated with these two loss functions conflict, learning does not progress properly. In this study, we propose a method for estimating early-part RIRs by introducing the dynamic pulling method (DPM) into the SIREN architecture with residual connections, using noisy microphone signals. From the simulation experiments in a two-dimensional sound field, the proposed method achieves higher accuracy in estimating RIRs in noisy environments compared to conventional methods.**

I. Introduction

The room impulse response (RIR) represents the sound propagation in a room from a loudspeaker to a microphone under the assumption of linear time invariance to the sound field. In particular, the early part of the RIR is important because it represents the early reflected sound components, which have more energy than the late reflections and have a significant effect on the sound localization and timbre impression of a sound source [1].

In applications such as sound field control and sound field visualization, the variation in the early part of the RIRs between different microphone positions is significant. As a result, measurements must be taken at multiple points rather than at a single point. However, when RIR measurements are required over large regions or at high spatial densities, obtaining measurements at many points can be challenging.

In recent years, many methods have been proposed to estimate RIRs at more points using signals obtained from a limited number of microphones. In particular, non-parametric methods for numerically estimating the sound field have been studied. For example, compressed sensing (CS) methods [2], [3] have been proposed using plane waves [4], spherical harmonic functions [5], [6], and the modal expansion [7], which are solutions to the wave equation. The equivalent source method [8]–[11] represents the sound field by the superposition of point sources, selected and weighted under the assumption of sparsity in time, space and other factors.

On the other hand, RIR estimation methods using deep learning have also attracted much attention in recent years. Convolutional neural networks (CNN) [12] and generative adversarial network (GAN) [13] can model complex relationships between inputs and outputs and have been applied to RIR estimation. Lluís *et al*. proposed a U–net architecture trained on the sound field data obtained using the Green function in a rectangular room [14]. Similarly, E. Fernandez-Grande *et al*. employed a GAN-based approach to estimate RIRs from limited measurements, achieving improved estimation accuracy compared to conventional plane wave regression methods [15].

Furthermore, a deep learning model called physics-informed neural networks (PINN) [16], which introduces physical laws into the loss function, has been proposed. This approach allows the model to be trained such that the output follows the partial differential equations governing the system of interest. And, SIREN [17], a network of multi layer perceptrons with sinusoidal activation functions, has proven to be an effective architecture for learning neural implicit representations of various signals and solving wave equations. Additionally, PINN and SIREN were applied to the inverse problem, and a method for estimating RIRs using them was proposed [18], [19]. In [18], a comparison of the PINN and CS methods [20] showed that the PINN method had improved estimation accuracy.

PINN generally uses two types of loss functions: one for data errors and another for physical laws. This leads to the problem of conflicting gradient vectors from the two loss functions, resulting in insufficient learning. To address this problem, several methods have been proposed. Dynamic pulling method (DPM) [21] adjusts the balance of the two gradients when they conflict. In [22], the self-adaptive loss balancing method has been proposed to determine the composition of two gradients at each learning step. In [19], the self-adaptive loss-balanced method was adapted to the RIR estimation problem using SIREN. When solving inverse problems, such as RIR estimation, overfitting the data can degrade estimation accuracy if the data contain significant noise. Moreover, if the learning process fully leverages physical laws, it can effectively remove noise from the data. In [23], an RIR estimation method has been proposed that is robust to noise in the microphone signals by dynamically switching the loss functions to tolerate data errors. However, further research is required to improve RIR estimation accuracy when large amounts of noise are present

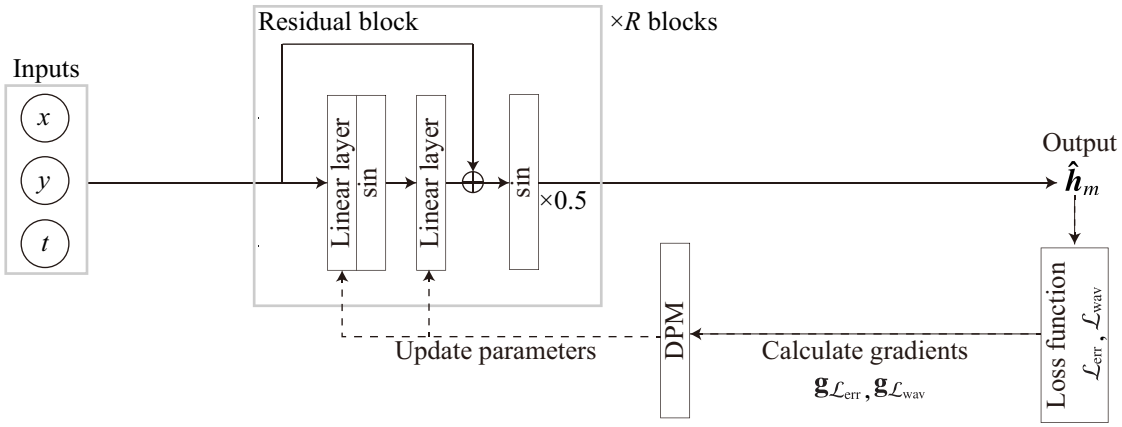


Fig. 1. Network architecture. Inputs are coordinates (x, y) and time t. The network is composed of R layers of residual blocks. The loss function is calculated by the output RIRs $\hat{\boldsymbol{h}}_m (m \in \mathcal{M})$ and measured RIRs $\boldsymbol{h}_m$, and gradients $\mathbf{g}_{\mathcal{L}_{\text{err}}}$ and $\mathbf{g}_{\mathcal{L}_{\text{wav}}}$ are computed to use DPM. For updating the network parameters, the gradients are modified by DPM. $\mathcal{M}$ is the location of all estimated points.

in the microphone signals.

In this study, to improve the RIR estimation accuracy in a noisy environment, we propose the estimation method of early RIRs based on PINN by introducing DPM and residual connections [24]. The proposed method uses DPM to prevent overfitting to the data and take advantage of losses due to the wave equation to improve estimation accuracy. In addition, this study uses residual connections to improve estimation accuracy by utilizing deeper network layers not considered in conventional RIR estimation methods.

The remainder of this paper is organized as follows. Section II describes the network architecture and DPM in the proposed method. Section III presents the simulation experiments in a two-dimensional sound field to compare the estimation accuracies between the conventional methods and the proposed methods. Finally, we conclude the paper in secttion IV.

II. Proposed methods

A. Physics-informed neural network using residual connection

A deep learning method using physical laws in loss functions is known as PINN. When solving acoustical problems, the wave equation with second derivatives is used as the physical laws. However, general activation functions such as ReLU did not allow second-order derivatives in the automatic differentiation. In this study, a SIREN network [17] based on a multi-layer perceptron (MLP) that allows for higher-order differentiation by using a sine function as an activation function. In addition, a residual connection is employed to enhance learning stability in deeper networks [24].

The proposed network is illustrated in the Fig. 1. The network is represented as a composite function of each network layer ϕ as follows

$$f(\mathbf{w}, \mathbf{x}) = (H_R \circ H_{R-1} \circ \cdots H_r \circ \cdots H_1)(\mathbf{x}), \tag{1}$$

$$H_r = \frac{1}{2} \sin(\omega_0(\phi_{2r-1}(\mathbf{x}_{2r-1})\mathbf{w}_{2r} + \mathbf{b}_{2r}) + \mathbf{x}_{2r-1}), \tag{2}$$

$$\phi_{2r-1}(\mathbf{x}_{2r-1}) = \sin(\omega_0(\mathbf{x}_{2r-1}^{\top}\mathbf{w}_{2r-1} + \mathbf{b}_{2r-1})), \tag{3}$$

where H_r is the residual block in the r-th layer $(r = 1, \ldots, R)$. The $\mathbf{w}$ and $\mathbf{b}$ are the weights and biases of the network, respectively. The training proceeds with the input $\mathbf{x}$. ω_0 is the SIREN weights.

In this study, the early-part RIR is estimated using the coordinate (x, y) and time information t as input $\mathbf{x} = (x, y, t)$. Note that SIREN multiplies the residual computed output by 1/2 to ensure that the output remains within the range of $[-1, 1]$ [17].

The loss function can be represented using the wave equation as follows

$$\mathcal{L} = \mathcal{L}_{\text{err}} + \lambda \mathcal{L}_{\text{wav}}, \tag{4}$$

$$\mathcal{L}_{\text{err}} = \frac{1}{\tilde{M}} \sum_{m \in \tilde{\mathcal{M}}} \|\hat{\boldsymbol{h}}_m - \boldsymbol{h}_m\|_2^2, \tag{5}$$

$$\mathcal{L}_{\text{wav}} = \frac{1}{M} \sum_{m=1}^{M} \|\frac{1}{c}\frac{\partial^2 \hat{\boldsymbol{h}}_m}{\partial t^2} - \nabla^2 \hat{\boldsymbol{h}}_m\|_2^2, \tag{6}$$

where λ is a balance parameter between the loss function of data $\mathcal{L}_{\text{err}}$ and the loss function of wave equation $\mathcal{L}_{\text{wav}}$, c is the speed of sound, t is time, and ∇ is the gradient at x and y of $\hat{\boldsymbol{h}}_m$. $\hat{\boldsymbol{h}}_m$, $\boldsymbol{h}_m$ denote m-th estimated RIRs and measured RIRs, respectively. $\tilde{\mathcal{M}}$ is the set of microphone indices, and M is the total number of measurement positions including microphone positions and evaluation points.

B. Dynamic pulling method

When using a loss function that combines multiple types of losses, the gradients from each loss may conflict, disturbing the learning process. Thus, in this study, we use the following algorithm based on the DPM [21], which dynamically adjusts the combination of gradients from the loss functions.

$$\mathbf{g}^{(k)} = \begin{cases} \mathbf{g}_{\mathcal{L}}^{(k)}, \text{ if } \mathbf{g}_{\mathcal{L}_{\text{err}}}^{(k)} \cdot \mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)} \geq 0 \\ v + \mathbf{g}_{\mathcal{L}}^{(k)}, \text{ otherwise,} \end{cases} \tag{7}$$

$$v = \frac{-\mathbf{g}_{\mathcal{L}}^{(k)} \cdot \mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)} + \delta^{(k)}}{\|\mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)}\|_2^2} \mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)}, \tag{8}$$

$$\delta^{(k+1)} = \begin{cases} w\delta^{(k)}, & \text{if } \mathcal{L}_{\text{wav}}^{(k)} > \epsilon \\ \frac{\delta^{(k)}}{w}, & \text{if } \mathcal{L}_{\text{wav}}^{(k)} \leq \epsilon, \end{cases} \tag{9}$$

where $\mathbf{g}^{(k)}$ is the gradient at the k-th step, and $\mathbf{g}_{\mathcal{L}}^{(k)}$, $\mathbf{g}_{\mathcal{L}_{\text{err}}}^{(k)}$, and $\mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)}$ are the gradients of the losses $\mathcal{L}$, $\mathcal{L}_{\text{err}}$, and $\mathcal{L}_{\text{wav}}$, respectively. w, ϵ, and δ are hyperparameters. And v is the gradient manipulation vector to bring $\mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)}$ closer to $\mathbf{g}_{\mathcal{L}_{\text{err}}}^{(k)}$, which is updated for each learning.

This algorithm first computes the inner product of the obtained gradient vectors. Then, according to (7), if the inner product is greater than or equal to 0, i.e., the gradient vectors point in the same direction, the parameters of the network are updated as is; otherwise, the overall gradient is manipulated using the gradient manipulation vector v and the parameters are updated.

That is, as shown in Fig. 1, after obtaining the gradient from each loss function, this algorithm is used to adjust the gradient and update the network.

III. Simulation experiment

A. Simulation conditions

Simulation experiments were conducted to estimate the early part of RIRs from a small number of microphone signals in a two-dimensional sound field. The estimation accuracies of four methods was compared: two conventional methods, CS [10] and PINN [18], and two proposed methods, PINN with DPM (DPM) and PINN with DPM and residual connections (DPM–Res).

Fig. 2 shows the arrangement of measurement points (12×12) and estimation points (20×20) . The Gaussian noises were added to the microphone signals with $\text{SNR} = 10$ dB. The sound source was a line source at $(0, 1.5)$ m. The length of microphone signal was 0.02 s. The sampling frequency was 8 kHz. The early-part RIRs were analytically simulated using SFS toolbox [25], including up to second-order reflections.

The network is trained for 20000 iterations using Adam optimizer with learning rate equal to 1.0×10^{-5}. For DPM-Res, the number of blocks was $R = 2$. For DPM, there were four hidden layers. For simple PINN, the number of hidden layers was three. For initial parameters, the balance parameter was $\lambda = 1.0 \times 10^{-6}$. ω_0 = 7 for DPM and DPM-Res and ω_0 = 12 for PINN.

The estimation accuracy was evaluated by the normalized mean square error (NMSE) as follows.

$$\text{NMSE} = 10 \log_{10} \frac{1}{M} \sum_{m=1}^{M} \frac{\|\hat{\boldsymbol{h}}_m - \boldsymbol{h}_m\|_2^2}{\|\hat{\boldsymbol{h}}_m\|_2^2}. \tag{10}$$

B. Results

We compared NMSEs of the estimated sound fields by the two conventional methods and two proposed methods. Fig. 3 shows the estimated sound field at $\text{y} = 0.03$ m.

From Fig. 3(c), the amplitude of the estimated sound field by CS was significantly small. The microphone signal was noisy with an SNR of 10 dB, making it difficult to estimate the sound field using CS. From Fig. 3(d)–(f), it can be observed that both the amplitude and phase estimated by deep learning methods (PINN, DPM, DPM-Res) are significantly close to the ground truth. In addition, Fig. 3(d), (e), and (f) show that the estimation errors decrease in the order of PINN, DPM, and DPM-Res. This indicates that DPM and DPM-Res are more robust to microphone noise than the conventional PINN.

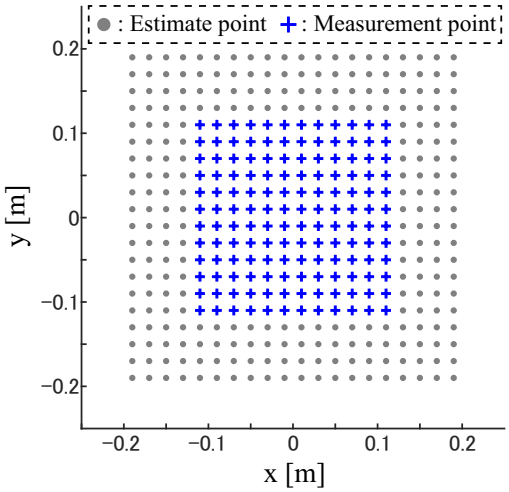


Fig. 2. Arrangement of measurement points (crosses) and estimation points (circles). Each point is 0.02 m apart in the center of a 6 m × 4 m rectangular room. The target sound field is two-dimensional and enclosed by the four rigid walls.

Next, the NMSEs were compared between the conventional and proposed methods. Fig. 4 shows the distributions of NMSE for CS, PINN, DPM, and DPM–Res. Fig. 4(a) and (b) show the results of the conventional methods. PINN improved the estimation accuracy by approximately 4.0 dB NMSE compared to CS. Fig. 4(b) and (c) show that the proposed DPM method achieved an improvement of approximately 3.0 dB in NMSE estimation accuracy compared to PINN. Furthermore, DPM-Res improves the estimation accuracy by approximately 1.2 dB in NMSE compared to DPM and by approximately 4.2 dB in NMSE compared to PINN.

Table I shows comparison of the mean NMSEs at the estimation points only, excluding the microphone positions i.e., points outside the microphones. PINN improved by approximately 2.5 dB in mean NMSE compared to CS. DPM and DPM-Res improved estimation accuracy by approximately 5.6 dB and 6.8 dB in the average NMSE, respectively, compared to CS.

Furthermore, Fig. 5 shows NMSE at each training step using deep learning methods. From the figure, it is evident that PINN exhibits overfitting to noise. In contrast, DPM effectively mitigates overfitting. Moreover, DPM-Res demonstrates a faster convergence rate compared to DPM.

These results show that the proposed methods (DPM and DPM-Res) significantly improve the estimation accuracy over conventional methods for noise removal for RIRs at microphone points and for RIR estimation at points outside the microphones. In particular, DPM-Res achieves the most

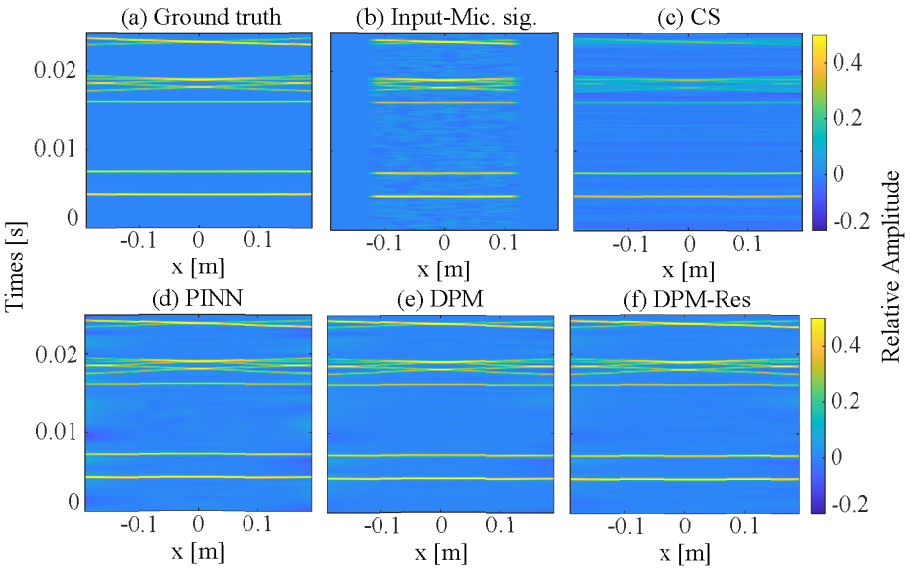


Fig. 3. Estimated RIR signals at y = 0.03 m (a) Desired sound field (b) Microphone signals for estimation (c)–(f) Estimated RIRs by CS, PINN, DPM and DPM-Res, respectively.

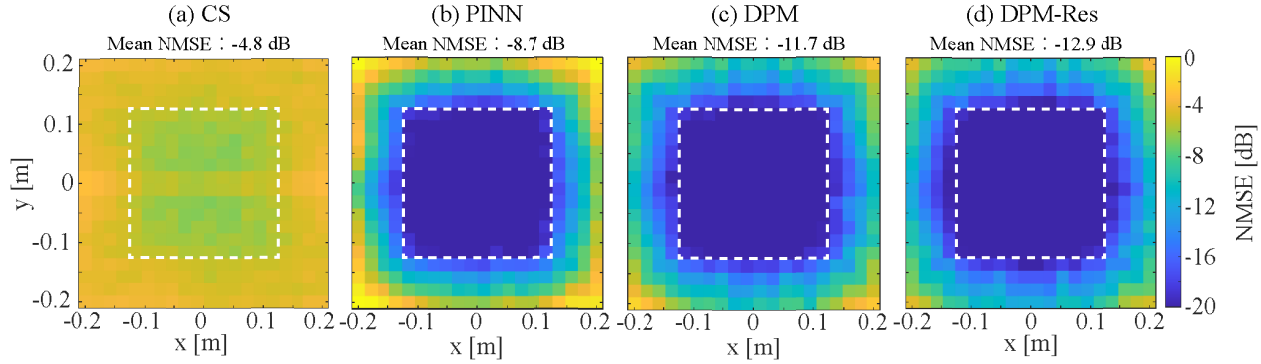


Fig. 4. Time-averaged NMSE maps with CS, PINN, DPM and DPM-Res. The white dot-lined borders indicates the region of microphone positions. (a)–(b) Conventional methods (CS, PINN) (c)–(d) Proposed methods (DPM, DPM-Res)

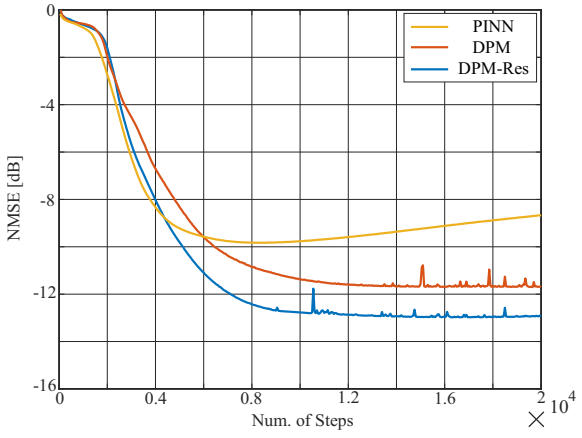


Fig. 5. Comparison of NMSEs in each learning step with PINN, DPM, and DPM-Res.

TABLE I
MEAN NMSE AT THE ESTIMATION POINTS ONLY, EXCLUDING THE MICROPHONE POSITIONS

Method	CS	PINN	DPM	DPM–Res
Mean NMSE [dB]	-4.3	-6.8	-9.9	-11.1

IV. CONCLUSIONS

In this study, we proposed the estimation of early-part RIR based on PINN using DPM and residual connections. From the simulation experiments, the proposed method improved estimation accuracy over conventional PINN and CS. In future work, we will study the configuration of the microphone array and further improve the PINN to reduce the number of microphones and improve the estimation accuracy of the early-part RIR estimation.

ACKNOWLEDGMENT

This work was partially supported by Research Institute for Science and Technology of Tokyo Denki University Grant Number Q24D-03 / Japan.

accurate estimation even in noisy environments.

REFERENCES

[1] T. Gotoh, Y. Kimura, A. Kurahashi, and A. Yamada, "A consideration of distance perception in binaural hearing," *The Journal of the Acoustical Society of Japan*, vol. 33, no. 12, pp. 667–671, 1977.

[2] D. L. Donohoh, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[4] S. A. Verburg and E. Fernandez-Grande, "Reconstruction of the sound field in a room using compressive sensing," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3770–3779, 2018.

[5] M. Pezzoli, M. Cobos, F. Antonacci, and A. Sarti, "Sparsity-based sound field separation in the spherical harmonics domain," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1051–1055.

[6] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "Sound field separation in a mixed acoustic environment using a sparse array of higher order spherical microphones," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 151–155.

[7] O. Das, P. Calamia, and S. V. Amengual Gari, "Room impulse response interpolation from a sparse set of measurements using a modal architecture," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 960–964.

[8] G. H. Koopmann, L. Song, and J. B. Fahnline, "A method for computing acoustic fields based on the principle of wave superposition," *The Journal of the Acoustical Society of America*, vol. 86, no. 6, pp. 2433–2438, Dec. 1989.

[9] E. Fernandez Grande, A. Xenaki, and P. Gerstoft, "A sparse equivalent source method for near-field acoustic holography," *Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 532–542, 2017.

[10] I. Tsunokuni, K. Kurokawa, H. Matsuhashi, Y. Ikeda, and N. Osaka, "Spatial extrapolation of early room impulse responses in local area using sparse equivalent sources and image source methods," *Applied Acoustics*, vol. 179, p. 108 027, 2021.

[11] N. Antonello, E. D. Sena, M. Moonen, P. A. Naylor, and T. V. Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1929–1941, 2017.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, 2012.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[14] F. Lluís, P. Martínez-Nuevo, M. B. Møller, and S. E. Shepstone, "Sound field reconstruction in rooms: Inpainting meets super-resolution," *The Journal of the Acoustical Society of America*, vol. 148, no. 2, pp. 649–659, 2020.

[15] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft, "Generative models for sound field reconstruction," *The Journal of the Acoustical Society of America*, vol. 153, no. 2, pp. 1179–1190, 2023.

[16] M. Raissi, P. Perdikaris, and G. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[17] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2020.

[18] M. Pezzoli and a. A. S. F. Antonacci, "Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, 2023, pp. 2177–2184.

[19] X. Karakonstantis and E. Fernandez-Grand, "Room impulse response reconstuction using physics-informed neural networks," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, vol. 378, 2023, pp. 3181–3188.

[20] E. Zea, "Compressed sensing of impulse responses in rooms of unknown properties and contents," *Journal of Sound and Vibration*, vol. 459, p. 114 871, 2019.

[21] J. Kim, K. Lee, D. Lee, S. Y. Jhin, and N. Park, "Dpm: A novel training method for physics-informed neural networks in extrapolation," in *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, Association for the Advancement of Artificial Intelligence, 2021, pp. 8146–8154.

[22] Z. Xiang, W. Peng, X. Liu, and W. Yao, "Self-adaptive loss balanced physics-informed neural networks," *Neurocomputing*, vol. 496, pp. 11–34, 2022.

[23] I. Tsunokuni, G. Sato, Y. Ikeda, and Y. Oikawa, "Spatial extrapolation of early room impulse responses with noise-robust physics-informed neural network," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E107-A, no. 9, 2024.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Con-*

ference on Computer Vision and Pattern Recognition (CVPR), 2016.

[25] H. Wierstorf and S. Spors, "Sound field synthesis toolbox," *132nd Convention of the Audio Engineering Society*, 2012.

Assessment and Improvement of Customer Service Speech with Multiple Large Language Models

So Watanabe*, Chee Siang Leow*, Junichi Hoshino†, Takehito Utsuro†, and Hiromitsu Nishizaki*
* University of Yamanashi, Japan
E-mail: s_watanabe@alps-lab.org, {leow,hnishi}@yamanashi.ac.jp Tel/Fax: +81-552208361
† University of Tsukuba , Japan
E-mail: jhoshino@esys.tsukuba.ac.jp, utsuro@iit.tsukuba.ac.jp

***Abstract*—This paper introduces a framework that utilizes multiple large language models (LLMs) to assess and enhance customer service interactions in the service industry. Effective communication with customers is pivotal for improving customer satisfaction. To enhance these skills, customer service staff require precise and constructive feedback. This study utilizes multiple LLMs within a round-table discussion framework, named "ReConcile,'' to evaluate and suggest improvements for customer service dialogues. The proposed method scores customer service speech and suggests suitable responses. A subjective experiment was conducted in which human subjects compared the effectiveness of customer service speech assessments and response suggestions generated by both a single LLM and the ReConcile method. Results demonstrated that the scores generated by ReConcile aligned more closely with human judgment compared to those from a single LLM, indicating the efficacy of ReConcile in providing suggestions for improving customer service staff's speech.**

I. Introduction

In the service industry, particularly in sectors such as hospitality, face-to-face communication skills are crucial not only for providing quality service and enhancing customer satisfaction but also for mitigating potential problems [1], [2]. However, acquiring proficient customer service skills through traditional training methods is often time-consuming and costly. Conventional approaches, such as role-playing exercises and service manuals, frequently fall short in replicating the nuances of actual customer interactions.

To address these limitations, researchers have proposed various innovative training methods to enhance customer service skills. For instance, Furuno et al. [3] developed a virtual reality (VR) system to evaluate specific behaviors, such as bowing accuracy. Similarly, Nishio et al. [4] introduced a system that provides feedback on the use of filler words and speech fluency during customer service interactions. While these systems represent significant advancements, they primarily focus on specific behavioral aspects and do not adequately address the modification and evaluation of response content—a critical component of effective service communication.

To address these limitations, Sano et al. [5] developed an advanced spoken dialogue system for scenario-based customer service training. While this system could evaluate the appropriateness of predefined responses within given scenarios, it lacked the flexibility to adapt to the unpredictable nature of real-world customer service interactions.

Effective customer service communication requires the ability to adapt to specific customer situations, emphasizing both content accuracy and respect for the customer. This is particularly crucial when considering customer service in languages with complex honorific systems, such as Japanese. In Japanese, "*keigo*" (honorific language) plays a vital role in expressing respect and maintaining social hierarchies. The appropriate use of keigo by service personnel significantly influences the perceived quality of customer service communication in Japanese society. Particularly in the service industry, the adept use of honorifics and a customer-centric approach are indispensable for successful communication with Japanese customers.

Given these linguistic and cultural considerations, this paper proposes a novel method for appropriately modifying and presenting customer service language used by service staff. Our approach aims to address the unique challenges posed by languages with complex honorific systems while maintaining the flexibility required for real-world customer interactions.

In recent years, with the emergence of large language models (LLMs) and their effective use in various domains, research on LLMs has progressed, and prompt engineering has also developed to generate more accurate responses by designing instructions and sentences to be entered into LLMs. Wei et al. [6] introduced a Chain of Thought (CoT) approach that embeds a stepwise process in prompts to elicit better responses from LLMs. There are studies that apply the approach of CoT. For example, Zero-shot CoT (Kojima et al. [7]), CoT-SC (Wang et al. [8] and Fu et al. [9]) and Tree of Thought (Yao et al. [10] and Long [11]). Furthermore, Bsharat et al. [12] presented 26 prompting principles that are essential for improving the response accuracy of LLMs. However, no study has used LLMs to evaluate and modify the appropriateness of customer service speech content. It is also unclear whether LLMs can provide appropriate feedback on honorifics that are unique to Japanese.

We propose a method for analyzing the speech of customer service staffs and modifying it to more appropriate expressions using the ReConcile method [13], which uses a combination of multiple LLMs, specifically GPT-4[1], Claude2 [14], and Bard (PaLM2) [15]. In the original paper [13], the effectiveness of the method was demonstrated on tasks requiring

[1]https://openai.com/blog/chatgpt

common sense reasoning and mathematical reasoning, such as StrategyQA[16], ECQA[17], GSM8K[18], and AQuA[19]. In recent years, it has become clear that multiple LLMs can be used to solve various tasks more effectively [20]–[22]. This paper also aims to more appropriately modify and evaluate customer service speech through ReConcile, which is based on a council system of outputs from multiple LLMs. Since the modification and evaluation of customer service responses do not have a single definitive answer, diverse expressions may be appropriate depending on the customer's needs and the store's situation. Therefore, it is necessary to generate optimal responses tailored to each specific situation.

ReConcile consists of three phases. At each phase, different LLMs are used to generate and modify responses, and finally, the optimal response is derived. First, the customer service situation settings, customer speeches and actions, and the customer service staff's responses to them, including prompts, are entered into ReConcile, and the customer service staff's responses are modified to generate the optimal response (response modification part). Next, the situation setting, the customer's speech, and the staff's original and modified responses are re-entered into ReConcile, and the original response is rated on a 10-point scale, along with the generation of explanatory text on which the rating is based (scoring part).

To verify the effectiveness of ReConcile, an evaluation experiment was conducted with 20 subjects, 13 with customer service experience and seven with no customer service experience, to compare the results of response modification and scoring by a single LLM with those of ReConcile.

Subjects evaluated the response modification results from each method and selected the most appropriate response, with ReConcile rated as the most appropriate response modification in 57% of all evaluations. This was significantly higher than the second place score of 20% for GPT-4.

For both scoring and explanation, ReConcile ranked first with 55%, a 21 point improvement over GPT-4's 34%. These results indicate that ReConcile is able to modify responses and generate scores and explanations more appropriate to the customer service context than a single LLM. Moreover, ReConcile's modified responses and scores are at or near the same level as those generated manually, suggesting its ability to provide appropriate feedback in customer service training. The contribution of this study is that the ReConcile method, which uses multiple LLMs, has the potential to automatically analyze and correct the speech of customer service staffs and provide opportunities to learn appropriate customer service expressions, including nuanced honorific expressions. Through the generation of examples, persuasive scoring, and explanations, multiple LLMs such as ReConcile can be expected to contribute to the improvement of customer service staffs' communication skills.

The test dialogues data used in this paper are publicly accessible on Github[2].

[2]https://github.com/watanabe0402/Customer-service-speech-assessment-with-ReConcile/tree/main

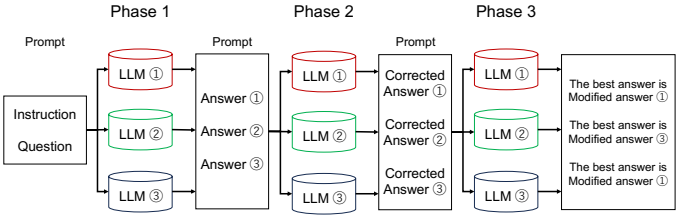


Fig. 1. A processing flow of ReConcile

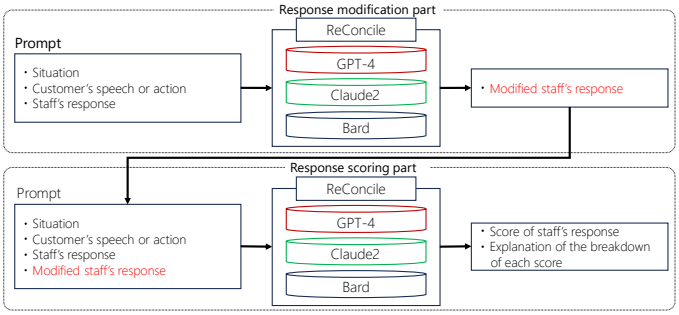


Fig. 2. Flow of customer service evaluation using ReConcile

II. Proposed Method

A. ReConcile

First, the flow of ReConcile is described: ReConcile performs three phases, which take advantage of the strengths of different large language models (LLMs) to obtain more accurate answers.The comprehensive process is illustrated in Fig. 1.

In Phase 1, a prompt consisting of instructions or queries is formulated and then submitted to multiple LLMs. Because LLMs are developed based on different data sets, their outputs embody unique perspectives and insights, even when responding to identical prompts. In Phase 2, the responses generated by each LLM in Phase 1 are merged into a new prompt, which is again input to each LLM. This process allows each LLM to output specific feedback for the multiple responses given, identifying the good and bad points of each response. In addition, each LLM generates responses again based on this feedback. In Phase 3, the responses modified in Phase 2 are compiled into a prompt and entered into each LLM. Each model is then tasked with selecting the most appropriate response from the revised options. If the optimal responses identified by each LLM differ, Phase 2 responses are further evaluated and adjusted. Conversely, if the most appropriate responses identified by all LLMs converge in Phase 3, that response is considered the final answer. However, if consensus among all LLMs' responses remains elusive after up to five iterations of Phases 2 and 3, the final response is determined by majority vote.

B. Customer service evaluation using ReConcile

In this study, we use ReConcile to evaluate the content of staff's responses by modifying their statements to generate more appropriate responses. In addition, we score the customer

service responses out of 10 and also generate explanations that serve as the basis for these scores. This entire process is illustrated in Fig. 2.

The prompts for each phase of ReConcile in both the response modification and scoring parts are shown in Fig. 3. First, the store situation, the customer's speech or action, and the staff's response to it are included in the prompt and entered into ReConcile. ReConcile modifies the staff's response to generate the most appropriate response. This is called the response modification part. Next, the store situation, the customer's speech or action, the staff's original response, and the previously generated modified response are all included in the prompt and re-entered into ReConcile. ReConcile scores the staff's original response on a scale of 1 to 10. This scoring is based on four criteria: correct use of keigo (3-point scale), respect for the customer (2-point scale), adaptation to the situation (3-point scale), and simplicity and clarity of speech (2-point scale). In addition, it simultaneously generates an explanation as to why it received such a score. This is called the scoring part.

III. Experiment

A. Experimental conditions

In the evaluation experiment, we assessed the effectiveness of ReConcile by comparing the responses from the individual GPT-4, Claude2, and Bard LLMs with those from the integrated ReConcile method. Furthermore, we evaluated the appropriateness of the response modified by ReConcile. The appropriateness evaluations were conducted through questionnaires administered to 20 subjects, 13 with customer service experience and seven without customer service experience.

1) Test dialogues data: A specific customer service scenario was created as evaluation data. The customer service scenarios were created by having ChatGPT play the role of the customer and a human play the role of the staff, interacting with the customer like real-life dialogues. In total, 12 customer service scenarios were created, including eight types of restaurants and four types of other stores. From the created scenarios, 30 dialogues that included characteristic interactions and complaint handling specific to the establishment were extracted, and an evaluation experiment was conducted using these dialogues.

2) Ranking and Appropriateness Evaluation: The test dialogue data presented in Section III-A1) is shown in the underlined part of Fig. 4. Each LLM and ReConcile are used to make corrections to this staff's responses, assign a score, and generate an explanation of why the score is the way it is.

In order to evaluate the modifications proposed by each LLM (an example is shown in Table I) and the scores and their explanations shown in Fig. 6, subjects were asked four questions in a questionnaire format. First, in Question 1, the subjects were asked to rank the GPT-4, Claude2, Bard, and ReConcile modifications in terms of their suitability for customer service situations, accuracy of salutations, and clarity of information. This will help us determine which method produced the best suggested revisions. Next, in Question 2, subjects were presented with only the score for the staff member's speech content by each method and asked to rate the appropriateness of the score with two choices, as shown in Fig. 5. In Question 3, as shown in Fig. 6, subjects were presented with both the scores from Question 2 and explanations for these scores, and were asked to rank the methods again. By comparing the rankings with and without the score explanations, we can determine whether the subjects were satisfied with the scores. The final question, Question 4, provides a binary scale of whether the proposed modifications of ReConcile, the scoring, and its explanation are appropriate. Question 4 allows us to indicate whether ReConcile is able to make appropriate modifications and correctly explain the rationale for appropriate scoring and scores.

TABLE I
Examples of staff's customer service response "Welcome! We're sorry to keep you waiting today. Is it all right?" as modified in each LLM and ReConcile

	Modified staff's speech
GPT-4	Welcome! We are currently quite busy, so you will have to wait a while. Thank you for your understanding
Claude2	Welcome! We apologize for the inconvenience, but due to the current high volume of customers, there will be a short wait before we can seat you.
Bard	Welcome! Today, we are exceptionally busy, and there may be a bit of a wait. Would that be alright with you?
ReConcile	Welcome! Today, we are experiencing significant congestion, and there may be an approximate **10 minutes** wait before we can guide you to your seat. We appreciate your patience and understanding.

Furthermore, we also analyzed how correctly LLMs could point out honorifics in Japanese. For example, in one dialogue example, while GPT-4 modified the staff's response from "分かりました" (I understand) to "了解しました" (roger), ReConcile further refined it to the more appropriate "かしこまりました" (certainly, sir/ma'am). The phrase "了解しました" is often used by a superior giving permission or acknowledgment to a subordinate, making it unsuitable for use when addressing a customer, who should be treated with the utmost respect. On the other hand, "かしこまりました" is a more formal and polite expression in Japanese, commonly used in customer service to show deference and respect to the customer. Thus, it was shown that ReConcile can be modified to use honorifics more suitable for customer service.

In conclusion, the results of this study demonstrate that ReConcile, a method utilizing multiple LLMs, is superior to single LLM-based approaches in evaluating and improving customer service speech. ReConcile's ability to generate more appropriate modifications, accurate scores, and well-reasoned explanations, as well as its sensitivity to the nuances of Japanese honorifics, makes it a promising tool for improving customer service quality of customer service and communication skills.

B. Results

Table II shows the results of Questions 1 through 3, and shows that ReConcile was selected as the most appropriate method for correcting staff response speeches (57% of total

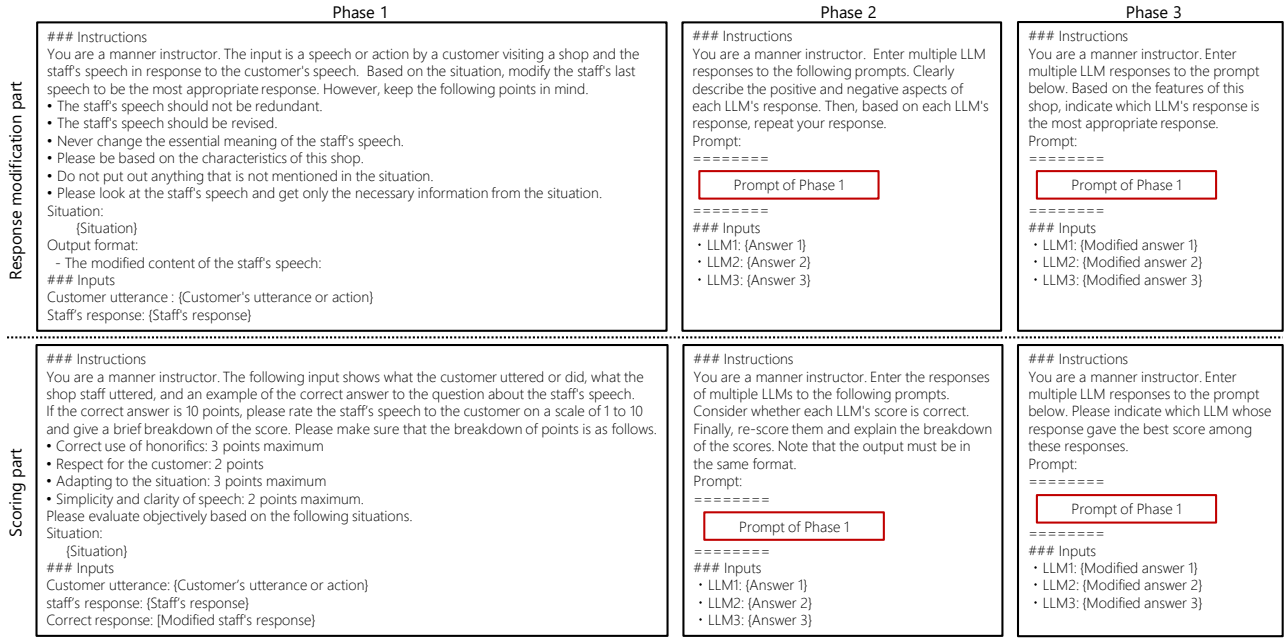


Fig. 3. Prompts given at each phase in ReConcile. The prompts are presented in English, but in the experiment, the prompts were given in Japanese with the exact same meaning.

Service scene setting
- Type of shop: Casual restaurant
- Type of seat: Table seats
- Num. of customers: 3 persons
- Status in the shop: Crowded and no seats available
- Reservation: No
- Order method: Order placed at the table
- Payment: Cash, credit card, electronic money

Dialogue between a service staff and a customer

Correct this speech

C: (Entering a casual restaurant)
S: (JP)「いらっしゃいませ！本日は少々お待ちいただいておりますが、よろしいですか？」
(EN) "Welcome! We're sorry to keep you waiting today. Is it all right?"
C: We are two adults and one child. How long do we have to wait?

Fig. 4. An example of a customer service scene setting and a customer service dialogue scenario. "C" is the customer's speech or action, and "S" is the service staff's response. In reality, the LLMs modify the underlined Japanese speech.

responses), 37 percentage points higher than GPT-4 (20%). This indicates that ReConcile, which uses multiple LLMs, could modify responses more appropriately than a single LLM. Comparing the results of the staff's response modification by each method (Table I), ReConcile's response clearly indicates a specific waiting time of "10-minute wait". While this "10" figure may be a hallucination, what is important regardless is that a quantitative waiting time was presented. The provision of a quantitative waiting time allows customers to decide whether to wait or go to another store. In customer service, providing such specific information properly manages customer expectations, demonstrates store transparency, and is related to building customer trust. Therefore, ReConcile's revised response was deemed most appropriate.

ReConcile's scores were rated as "appropriate" more often than those of a single LLM, with a 2-point lead over GPT-4

TABLE II
RESULTS OF SUBJECT EXPERIMENTS FOR QUESTIONS 1 THROUGH 3 IN EACH LLM AND RECONCILE. NUMBERS ARE RATES [%]

	GPT-4	Claude2	Bard	ReConcile
[Q1] Rate ranked 1st (modified response)	20	5	17	57
[Q2] Appropriateness of scores (binary scale)	72	68	18	74
[Q3] Rate ranked 1st (scores and their breakdowns)	34	6	4	55

alone, suggesting that ReConcile's scores are more appropriate. In addition, ReConcile was chosen first for scoring and commenting (55% of total responses), a 21-point improvement over GPT-4 (34%), suggesting that ReConcile scores more appropriately and generates more persuasive commenting than a single LLM.

Table III presents the results for Question 4, showing that ReConcile's responses are highly appropriate for modification (81%), scoring (78%), and commenting (82%), all exceeding those of a single LLM. These results show that ReConcile adapts to store situations, provides appropriate response modifications, and generates satisfactory explanatory text with appropriate scoring. Furthermore, 74% of the respondents found ReConcile's scoring to be "appropriate" when judged by the score alone (Table II). However, when the score and comment were considered together (Table III), 78% of subjects found the ReConcile score "appropriate", an improvement of four percentage points. This finding confirms that the explanations generated by ReConcile are more convincing in justifying the given scores.

GPT-4	Claude2	Bard	ReConcile
Total：8/10 Details: Correct use of keigo: 3/3 Respect for the customer: 2/2 Adapting to the situation: 2/3 Simplicity and clarity of speech: 1/2	Total：5/10 Details: Correct use of keigo: 2/3 Respect for the customer: 1/2 Adapting to the situation: 1/3 Simplicity and clarity of speech: 1/2	Total：7/10 Details: Correct use of keigo: 3/3 Respect for the customer: 2/2 Adapting to the situation: 2/3 Simplicity and clarity of speech: 2/2	Total：7/10 Details: Correct use of keigo: 3/3 Respect for the customer: 1/2 Adapting to the situation: 2/3 Simplicity and clarity of speech: 1/2

Fig. 5. Examples of scores of the customer service speech output from each LLM and the proposed method

GPT-4	Claude2	Bard	ReConcile
Total：8/10 Details: Correct use of keigo: 3/3 Reason: The staff uses proper keigo and polite language. Respect for the customer: 2/2 Reason: With the greeting, "Welcome," the staff guides the customer according to the customer's situation, showing the respect for the customer. Adapting to the situation: 2/3 Reason: The staff informs the customer of the crowded situation and responds appropriately to the situation, but does not give specific wait times. Simplicity and clarity of speech: 1/2 Reason: The speech is simple, but not completely clear, as the staff lack specificity about wait times and make it difficult for customers to clearly understand the actual wait time.	Total：5/10 Details: Correct use of keigo: 2/3 Reason: Some honorifics are used, such as “いらっしゃいませ” (please come in) and “お待ちいただく” (please wait for me), but the use of “よろしい” is somewhat inappropriate. Respect for the customer: 1/2 Reason: The staff informs that the customer will wait for some time, but he/she do not consider the customer's position at all. Adapting to the situation: 1/3 Reason: There is some mention of how crowded the restaurant is, but no clear explanation of estimated wait times. Simplicity and clarity of speech: 1/2 Reason: Simple utterance, but insufficient explanation of waiting time and in-store conditions.	Total：7/10 Details: Correct use of keigo: 3/3 Reason: All honorifics are used correctly. Respect for the customer: 2/2 Reason: The staff politely say "Welcome" to our customers and take care to ask for their approval. Adapting to the situation: 2/3 Reason: The staff clearly inform the customer that the restaurant is crowded, but they do not provide specific wait times. Simplicity and clarity of speech: 2/2 Reason: Simple and easy to understand.	Total：7/10 Details: Correct use of keigo: 3/3 Reason: The staff uses proper honorifics and are polite in greeting customers. Respect for the customer: 1/2 Reason: The greeting, "Welcome," is appropriate, but a little more care needs to be taken in the wording used to inform the customer to wait. “よろしいですか” does not show enough consideration for the customer's situation, and therefore does not show enough respect. Adapting to the situation: 2/3 Reason: Although the staff informs the customer of the crowded situation, he/she does not provide specific wait times, so their adaptation to the situation is not perfect. Simplicity and clarity of speech: 1/2 Reason: The speech is simple, but lacks clarity due to the lack of specific details regarding wait times and the degree of crowding. Providing more specific information would help customers understand the situation more accurately.

Fig. 6. Examples of scores and breakdown descriptions of the customer service speech output from each LLM and the proposed method

TABLE III
QUESTION 4: RECONCILE'S APPROPRIATENESS FOR RESPONSE SPEECHES [%]

Appropriateness of the modification	Appropriateness of scores	Adequacy of commentary
81	78	82

IV. CONCLUSIONS

This paper proposed the use of ReConcile as an approach to evaluate and improve customer service language using multiple LLMs. ReConcile modifies staff's responses to specific customer service situations, provides examples of appropriate responses, and generates scores and explanations for the original responses. Subjective experiments showed that ReConcile outperforms single LLM-based methods in terms of response modification, scoring, and explanation generation, confirming its effectiveness in evaluating and improving customer service speech. It was also shown that ReConcile can handle Japanese-specific honorifics.

In the future, we plan to incorporate ReConcile into a customer service training system as a tool for response modification, scoring, and commentary generation.

ACKNOWLEDGMENT

This research was supported by JSPS Grant-in-Aid JP23H00994.

REFERENCES

[1] R. Asriyani and I. W. Anggayana, “Mastering the language of service: English communication skills for food and beverage professionals,” *Jurnal Manajemen Pelayanan Hotel*, vol. 7, no. 2, pp. 1127–1139, 2023.

[2] S. Muda and S. M. Rashid, “Customer satisfaction towards communication skills of the franchise restaurant frontliners,” in *Proceedings of Breaking the Barriers, Inspiring Tomorrow, European Proceedings of Social and Behavioural Sciences*, vol. 110, 2021, pp. 223–230.

[3] T. Furuno, S. Fujita, D. Wang, *et al.*, “Multimodal vr customer service training system using conversational customer actors,” *IPSJ Journal*, vol. 63, no. 1, pp. 231–241, 2022.

[4] T. Nishio, S. Iida, Y. Sano, *et al.*, “Customer service training vr system that can train how to speak,” in *The Special Interest Group Technical Reports of IPSJ, Vol.2021-CVIM-224*, 2021, pp. 1–4.

[5] Y. Sano, C. S. Leow, S. Iida, *et al.*, "Spoken dialog training system for customer service improvement," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 403–408.

[6] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 24 824–24 837.

[7] T. Kojima, S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 22 199–22 213.

[8] X. Wang, J. Wei, D. Schuurmans, *et al.*, "Self-consistency improves chain of thought reasoning in language models," in *Proceedings of Eleventh International Conference on Learning Representations*, 2023.

[9] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot, "Complexity-based prompting for multi-step reasoning," in *Proceedings of Eleventh International Conference on Learning Representations*, 2023.

[10] S. Yao, D. Yu, J. Zhao, *et al.*, "Tree of thoughts: Deliberate problem solving with large language models," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 11 809–11 822.

[11] J. Long, "Large language model guided tree-of-thought," *arXiv preprint, arXiv:2305.08291*, 2023.

[12] S. M. Bsharat, A. Myrzakhan, and Z. Shen, "Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4," *arXiv preprint, arXiv:2312.16171*, 2024.

[13] J. C.-Y. Chen, S. Saha, and M. Bansal, "Reconcile: Round-table conference improves reasoning via consensus among diverse llms," *arXiv preprint, arXiv:2309.13007*, 2023.

[14] Anthropic, "Model Card and Evaluations for Claude Models," *Technical Report 2023*, 2023. [Online]. Available: https://x.gd/WIoLC.

[15] R. Anil, A. M. Dai, O. Firat, *et al.*, "Palm 2 technical report," *arXiv preprint, arXiv:2305.10403*, 2023.

[16] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant, "Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies," *arXiv preprint, arXiv:2101.02235*, 2021.

[17] S. Aggarwal, D. Mandowara, V. Agrawal, D. Khandelwal, P. Singla, and D. Garg, "Explanations for CommonsenseQA: New Dataset and Models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Aug. 2021, pp. 3050–3065.

[18] K. Cobbe, V. Kosaraju, M. Bavarian, *et al.*, "Training verifiers to solve math word problems," *arXiv preprint, arXiv:2110.14168*, 2021.

[19] W. Ling, D. Yogatama, C. Dyer, and P. Blunsom, "Program induction by rationale generation: Learning to solve and explain algebraic word problems," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 158–167.

[20] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," *arXiv preprint, arXiv:2305.14325*, 2023.

[21] T. Liang, Z. He, W. Jiao, *et al.*, "Encouraging divergent thinking in large language models through multi-agent debate," *arXiv preprint, arXiv:2305.19118*, 2023.

[22] J. Li, Q. Zhang, Y. Yu, Q. Fu, and D. Ye, "More agents is all you need," *arXiv preprint, arXiv:2402.05120*, 2024.

Robust Quantile Regression Under Unreliable Data

Yoshifumi Shoji, Masahiro Yukawa
Department of Electronics and Electrical Engineering, Keio University, Yokohama, Japan
E-mail: shoji.yoshifumi@ykw.elec.keio.ac.jp, yukawa@elec.keio.ac.jp

***Abstract*—This paper addresses the quantile regression task when some non-negligible portion of data are corrupted by accidental factors such as temporary sensor malfunctions. Here, the task is to find the empirical quantile of the "reliable" data with the "unreliable" ones excluded. For this task, we propose the MC-pinball loss which is the composition of the minimax concave (MC) penalty and the pinball loss. The simulation results show that the proposed approach yields reasonable estimates of the true quantile. A potential benefit of the proposed approach is also shown with respect to the parameter tuning.**

I. Introduction

Quantile regression [1] has been studied for more than half a century, and it has been applied to various problems, including wind power generation [2], [3], power forecast [4], wage estimation [5], [6], gold price estimation [7], to name just a few. It still remains an active research topic in many areas including signal processing [8] and statistics [9].

We consider the situation where the input vectors $(\boldsymbol{x}_i)_{i=1}^m$ and the outputs $(y_i)_{i=1}^m$ are given. The key assumption here is that a non-negligible portion of the outputs are corrupted by accidental factors such as temporary sensor malfunctions, measurement/transmission errors, and human errors. Those outputs will be referred to as *unreliable* data. The rest of the outputs are *reliable* data assumed to be modeled well by means of linear estimation. As such, we define the residual vectors

$$\boldsymbol{\epsilon} := [\epsilon_1, \epsilon_2, \cdots, \epsilon_m]^\top := \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_\star \in \mathbb{R}^m, \tag{1}$$

where $\boldsymbol{y} := [y_1, y_2, \cdots, y_m]^\top \in \mathbb{R}^m$, $\boldsymbol{X} := [\boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \cdots \ \boldsymbol{x}_m]^\top \in \mathbb{R}^{m\times n}$, and $\boldsymbol{\theta}_\star \in \mathbb{R}^n$ is the 'true' regression vector.

The deterministic $\boldsymbol{x}_i$s, y_is, and ϵ_is are samples of random vector X and rondom variables Y and ε, respectively. Here, ε obeys a mixture of distributions of which the probability density function is given by

$$h(\epsilon) := (1-\beta)f(\epsilon) + \beta g(\epsilon), \ \epsilon \in \mathbb{R}, \tag{2}$$

where f is the density of ε for reliable data, and g is that for unreliable data with its proportion $\beta \in [0,1]$. The set of ϵ_is for reliable data is denoted by $\mathcal{D}_{\rm r} = (\epsilon_i)_{i\in\mathcal{I}_{\rm r}} (\subset \mathcal{D} := (\epsilon_i)_{i\in\mathcal{I}})$, where $\mathcal{I}_{\rm r} \subset \mathcal{I} := \{1, 2, \cdots, m\}$. The conditional density of Y given $X = \boldsymbol{x}$ is given by $h_Y(y \mid X = \boldsymbol{x}) = h(y - \boldsymbol{x}^\top\boldsymbol{\theta}_\star)$, where $y - \boldsymbol{x}^\top\boldsymbol{\theta}_\star = \epsilon$. We also define $f_Y(y \mid X = \boldsymbol{x}) = f(y - \boldsymbol{x}^\top\boldsymbol{\theta}_\star)$ and $g_Y(y \mid X = \boldsymbol{x}) = g(y - \boldsymbol{x}^\top\boldsymbol{\theta}_\star)$.

To clarify the concept, the situation under consideration is illustrated in Fig. 1. The reliable data are concentrated, while the unreliable data reside in those regions which are apart from the region of reliable data. Suppose, for instance, that one is interested in the traffic of wireless communication, or more specifically, one would like to know such an interval in which the "actual" traffic falls with a prespecified probability. In this context, each "reliable" datum y_i is the actual traffic given $\boldsymbol{x}_i$, and ϵ_i is the mismatch from its 'best' linear approximation in some sense. In this case, y_i is regarded as a sample of Y generated from $f_Y(Y \mid X = \boldsymbol{x}_i)$. The "unreliable" data, on the other hand, deviate from the actual traffic, and those data may cause fatal errors in quantile regression.

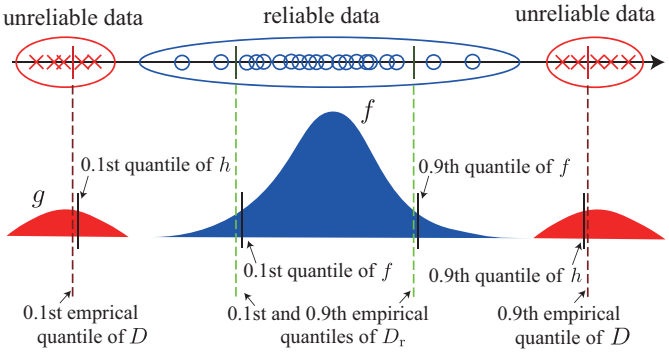


Fig. 1: An example of density of ε. Circle/cross indicates $\epsilon_i (= y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta}_\star)$ for reliable/unreliable data.

Keeping the above arguments in mind, the desirable interval would be estimated by finding the empirical quantiles using the dataset that solely contains the "reliable" data. Since separating those reliable and unreliable data is a nontrivial task, the goal of the current study is to build a loss function which is insensitive to the "unreliable" data so that its minimizer gives a reasonable estimate of the empirical quantiles of the "reliable" data. The so-called pinball loss (an asymmetric least absolute deviation) is typically used for quantile regression, and the question address in this work is how to make it insensitive to the "unreliable" data.

In this paper, we first consider the use of Moreau enhancement of the pinball loss function, motivated by the fact that the Moreau enhancement of the absolute-value function has been used successfully for robust signal recovery [10]. It then turns out to be undesirable for the present task. We therefore propose an alternative approach based on the MC-pinball loss which is a composition of the pinball loss with the minimax concave (MC) penalty [11], [12]. The proposed loss has a couple of nice properties leading to insensitivity to the unreliable data. Numerical studies are presented with f being the Gaussian

This work was partially supported by JST SICORP Grant Number JPMJSC20C6, Japan.

distribution and with g being three different distributions: the uniform distribution, the gamma distribution, and a point mass (Dirac's delta function). We demonstrate that the proposed approach yields a good estimate of the true quantile in terms of system mismatch (the quantile estimation error) and coverage rate even for a large β up to 0.5, while the performance of the conventional approach based on the pinball loss degrades sharply as β increases slightly from zero.

II. Robust Quantile Regression Problem under Unreliable Data

Throughout the paper, let $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{R}_{++}$ denote the sets of nonnegative integers, real numbers, and strictly positive real numbers, respectively. Let X be a real-valued random vector of dimension $n \in \mathbb{N}^* := \mathbb{N} \setminus \{0\}$, and Y be a real-valued random variable which is related to X. The αth conditional quantile function of Y given $X = \boldsymbol{x}$ is defined by $q_\alpha(\boldsymbol{x}) := \inf\{y \in \mathbb{R} : F(y \mid X = \boldsymbol{x}) \geq \alpha\}$, where X may be characterized by its (right-continuous) distribution function $F(y \mid X = \boldsymbol{x}) := \mathbb{P}\{Y \leq y \mid X = \boldsymbol{x}\}$ [1], [9].

In practical applications, the data could be changed into completely different values due to a variety of reasons during the data acquisition/transmission process. Such "unreliable" data do *not* obey the true distribution, and hence the empirical αth quantile using the whole dataset containing the unreliable data could be significantly different from of the data distribution. The task is therefore to estimate the αth empirical quantile of the set of "reliable" data, which is expected to be a reasonable estimate of the true quantile of $f_Y(y \mid X = \boldsymbol{x})$. Here, we assume implicitly that those unreliable data reside mostly in those regions which are apart from the bulk of reliable data, i.e., such unreliable data that are mixed up with the reliable one are supposed to give negligible impacts on the estimation. See Fig. 1.

The αth empirical quantile $\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_\alpha$ at $\boldsymbol{x}$ from the training dataset is obtained via minimization of the pinball loss function [1] $\sum_{i=1}^m \rho_\alpha(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})$, where (see Fig. 2)

$$\rho_\alpha : \mathbb{R} \to [0, +\infty) : z \mapsto \begin{cases} \alpha z, & \text{if } z \geq 0, \\ -(1-\alpha)z, & \text{otherwise.} \end{cases} \tag{3}$$

In the present scenario, the empirical quantile obtained with the pinball loss would be an estimate of the quantile of $h_Y(y \mid X = \boldsymbol{x})$, while what is desired to estimate is the quantile of the density $f_Y(y \mid X = \boldsymbol{x})$ of the reliable data. In the following section, we present a modified loss function to seek for the empirical quantile of the reliable data.

III. Proposed Approach to Robust Quantile Regression

Under the implicit assumption stated in the previous section, the idea is to modify the pinball loss in such a way that it saturates when $|z|$ exceeds some thresholds. A similar idea has been studied in the context of robust signal recovery [10]

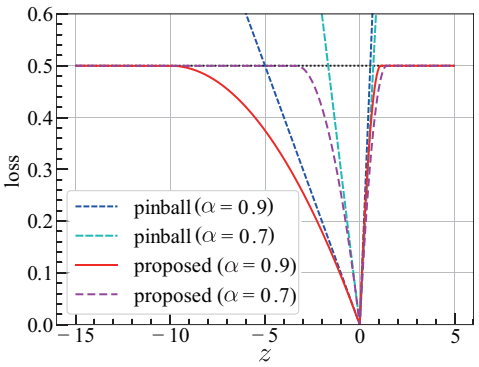


Fig. 2: Pinball and proposed losses ($\alpha = 0.7, 0.9$)

using the MC penalty [11], [12]

$$\phi_\gamma^{\mathrm{MC}} : \mathbb{R} \to [0, +\infty) : z \mapsto \begin{cases} |z| - \dfrac{1}{2\gamma} z^2, & \text{if } |z| < \gamma, \\ \dfrac{1}{2}\gamma, & \text{otherwise,} \end{cases} \tag{4}$$

where $\gamma \in \mathbb{R}_{++}$ is the saturation factor.

In fact, the MC penalty can be expressed as $\phi_\gamma^{\mathrm{MC}} = |\cdot| - {}^\gamma|\cdot|$. Here, for any convex function $f : \mathbb{R} \to \mathbb{R}$, its Moreau envelope of index $\gamma > 0$ is defined by

$${}^\gamma f(z) := \min_{x \in \mathbb{R}} \left(f(z) + \frac{1}{2\gamma}(z - x)^2 \right), \ z \in \mathbb{R}. \tag{5}$$

The function $\phi_\gamma^{\mathrm{MC}}$ has an enhanced sparsity-promoting property owing to the Moreau envelope ${}^\gamma|\cdot|$ of $|\cdot|$, and thus $\phi_\gamma^{\mathrm{MC}}$ is referred to as the Moreau enhancement [13] of $|\cdot|$.

We first consider a straightforward approach of applying the Moreau enhancement to the pinball loss. This approach, however, will turn out to be undesirable in the present context. Therefore, we propose an alternative approach by composing the two functions ρ_α and $\phi_\gamma^{\mathrm{MC}}$, and we finally compare it with the straightforward approach.

A. Moreau-enhanced Pinball Loss and Its Issues

The Moreau-enhanced pinball function is given by

$$\begin{aligned} \phi_{\alpha,\gamma} : \mathbb{R} \to [0, +\infty) : z &\mapsto \rho_\alpha(z) - {}^\gamma\rho_\alpha(z) \\ &= \begin{cases} \dfrac{\alpha^2}{2}\gamma, & \text{if } z \geq \alpha\gamma, \\ \alpha z - \dfrac{1}{2\gamma} z^2, & \text{if } 0 \leq z < \alpha\gamma, \\ -(1-\alpha)z - \dfrac{1}{2\gamma} z^2, & \text{if } -(1-\alpha)\gamma < z < 0, \\ \dfrac{(1-\alpha)^2}{2}\gamma, & \text{otherwise.} \end{cases} \end{aligned} \tag{6}$$

The process of generating $\phi_{\alpha,\gamma}$ from $|\cdot|$ is indicated in Fig. 3 by the blue arrow. Different weights are given to each side of $|\cdot|$ to obtain the pinball loss ρ_α (bottom left). The Huberized pinball (bottom middle, the Moreau envelope of ρ_α) is subtracted from ρ_α to obtain $\phi_{\alpha,\gamma}$. It can be seen that

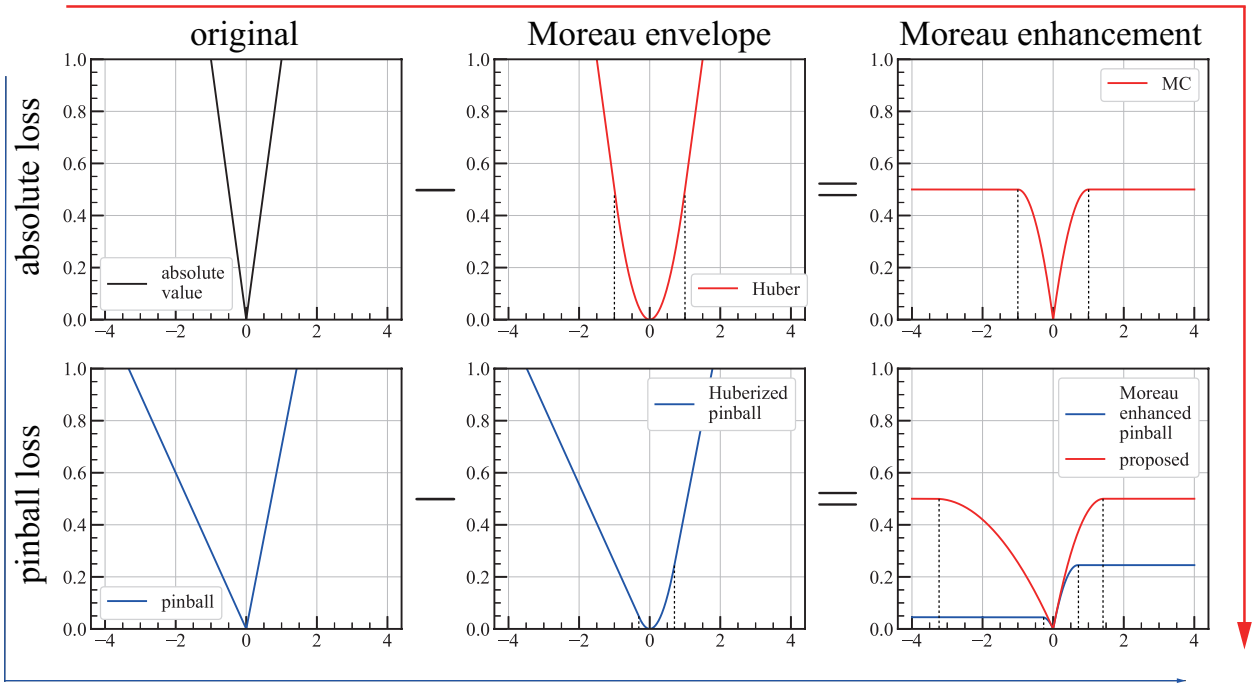


Fig. 3: The generation processes of the proposed loss $\phi_{\alpha,\gamma}^{\text{MC-pin}}$ and the Moreau-enhanced pinball loss $\phi_{\alpha,\gamma}$ for $\alpha = 0.9$.

the constants after saturation for $\phi_{\alpha,\gamma}$ significantly differ from each other on both sides of the real line. Because of this, $\phi_{\alpha,\gamma}$ encourages a larger portion of data than the desirable proportion α stay on the left side if $\alpha \approx 1$ (or on the right side if $\alpha \approx 0$).

Viewing Fig. 3 under (6), moreover, it can be seen that the ranges of nonzero gradient on the left and right sides are proportional to $1-\alpha$ and α, respectively. Indeed, the quantile estimate needs to be such that the number of data on each side is proportional to α and $1-\alpha$. If ϵ for reliable data is uniformly distributed, for instance, the nonzero-gradient ranges are desired to share the same proportion. In this specific case, the ranges for $\phi_{\alpha,\gamma}$ have the reciprocal proportion, implying that $\phi_{\alpha,\gamma}$ is undesirable to separate undesirable data from (the bulk of) reliable ones. To resolve those issues mentioned above, we propose an alternative approach in the following subsection.

B. Proposed Loss Function

We define the MC-pinball function as follows:

$$\begin{aligned}\phi_{\alpha,\gamma}^{\text{MC-pin}} &: \mathbb{R} \to [0,+\infty) : z \mapsto \phi_{\gamma}^{\text{MC}} \circ \rho_{\alpha}(z) \\ &= \begin{cases} \alpha z - \dfrac{\alpha^2}{2\gamma} z^2, & \text{if } 0 \le z \le \dfrac{1}{\alpha}\gamma, \\ -(1-\alpha) z - \dfrac{(1-\alpha)^2}{2\gamma} z^2, & \text{if } -\dfrac{1}{(1-\alpha)}\gamma < z \le 0, \\ \dfrac{1}{2}\gamma, & \text{otherwise,} \end{cases}\end{aligned} \tag{7}$$

where $\gamma \in \mathbb{R}_{++}$ governs the points at which the gradient vanishes. The task of estimating the empirical quantile of reliable data is cast as the following minimization problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \Phi_{\alpha,\gamma}^{\text{MC-pin}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) := \sum_{i=1}^{m} \phi_{\alpha,\gamma}^{\text{MC-pin}}(y_i - \boldsymbol{x}_i^{\top}\boldsymbol{\theta}). \tag{8}$$

The MC-pinball function $\phi_{\alpha,\gamma}^{\text{MC-pin}}$ is depicted in Fig. 2. It approximates the pinball loss well in the vicinity of the origin, while it saturates on both sides of the real axis. The generation process of the proposed loss is indicated by the red arrow in Fig. 3. As stated already, the MC loss (top right) is generated by subtracting the Huber function (top middle: the Moreau envelope of $|\cdot|$) from $|\cdot|$ (top left). It is then composed with ρ_α to obtain the proposed loss $\phi_{\alpha,\gamma}^{\text{MC-pin}}$ (bottom right).

Comparing the two graphs in the bottom-right panel of Fig. 3 under (4), $\phi_{\alpha,\gamma}^{\text{MC-pin}}$ possesses the following remarkable properties: (i) the constants $\gamma/2$ after saturation coincides on both sides with each other, and (ii) the nonzero-gradient ranges are proportional to α and $1-\alpha$. These are in sharp contrast to the case of $\phi_{\alpha,\gamma}$, and $\phi_{\alpha,\gamma}^{\text{MC-pin}}$ is free from the issues of $\phi_{\alpha,\gamma}$ raised in Section III-A. Indeed, the range proportion is desirable when ϵ for reliable data is uniformly distributed, and the proposed loss works well when it is normally distributed, as shown by simulations in Section IV.

Thanks to properties (i) and (ii), a solution of the problem in (8) tends to give a robust estimate of the true quantile. To obtain an intuition, suppose, under property (ii), that the reliable data reside in the vicinity of the empirical quantile of the reliable data, and that the unreliable data are scattered in the saturated region (see Fig. 1). In this case, the empirical quantile is likely a minimizer of the MC-pinball loss for the reliable data in an approximate sense. Importantly, moreover,

the unreliable data do not affect the minimizer, because a slight deviation from the empirical quantile of the reliable data does not change the losses at all for those unreliable data owing to property (i). Consequently, a solution of (8) would be a good approximation of the empirical quantile of the reliable data (and thus a reasonable estimate of the true quantile).

Remark 1. The proposed loss has the following asymptotic properties: (i) $\lim_{\gamma\to+\infty}\phi^{\text{MC-pin}}_{\alpha,\gamma} = \rho_\alpha$, and (ii) $\lim_{\gamma\downarrow 0}(\gamma/2)\phi^{\text{MC-pin}}_{\alpha,\gamma} = \|\cdot\|_0$. This implies that γ needs to be tuned appropriately. Fortunately, our numerical studies suggest that the proposed loss gives a good performance for γ chosen from a reasonably wide range.

IV. Numerical Examples

We conduct simulations to show that the proposed approach yields robust estimates of the true quantiles under a variety of situations. After presenting the simulation conditions, we show how the performance changes as the proportion β of unreliable data changes. We then show insensitivity of the proposed approach to the choice of the parameter γ.

A. Simulation conditions

Throughout, we consider the quantile regression for $\alpha = 0.1, 0.9$ to find the 80% interval with the top and bottom 10 percents excluded. The input matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ ($m = 2000$, $n = 11$) and the regression vector $\boldsymbol{\theta}_\star \in \mathbb{R}^n$ are i.i.d. with the uniform distribution $\mathcal{U}[0,1]$ and the standard Gaussian distribution, respectively.[1] The output vector is generated as $\boldsymbol{y} := \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^m$ is i.i.d. with the densities f and g described below. The residuals ϵ_i of reliable data are generated from the Gaussian mixture distribution $f(\epsilon) := \frac{1}{3}\frac{1}{\sqrt{2\pi\cdot 0.1^2}}\exp\left(-\frac{(\epsilon+0.5)^2}{2\cdot 0.1^2}\right) + \frac{2}{3}\frac{1}{\sqrt{2\pi\cdot 0.2^2}}\exp\left(-\frac{(\epsilon-0.5)^2}{2\cdot 0.2^2}\right)$. For the generation of ϵ_i of unreliable data, we consider three types of distributions (see Fig. 4): (i) the uniform distribution $g_1(\epsilon) := \begin{cases}\frac{1}{2a}, & -a\le\epsilon\le a,\\ 0, & \text{otherwise},\end{cases}$ for $a := 50$, (ii) the gamma distribution $g_2(\epsilon) := b\exp\big(-b\epsilon\big), \epsilon > 0$, for $b := 100$, and (iii) the point mass $g_3(\epsilon) := \delta(\epsilon - c)$ for $c := 100$, where δ is Dirac's delta.

Two experiments are conducted.

Experiment A: The performance is compared for different proportions β of unreliable data for the tuned parameter $\gamma := 2.0$.

Experiment B: The performance of the proposed approach is tested for different saturation factors γ for $\beta = 0.3$.

To seek for a stationary point of the proposed loss in (8), we employ Algorithm 1, where

$$\nabla_{\boldsymbol{\theta}}\Phi^{\text{MC-pin}}_{\alpha,\gamma}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}) = \sum_{i=1}^m \nabla_{\boldsymbol{\theta}}\phi^{\text{MC-pin}}_{\alpha,\gamma}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta}). \tag{9}$$

[1]Following the convention, the last component of each row vector is a constant so that the last component of $\boldsymbol{\theta}_\star$ represents the bias term.

Algorithm 1 generalized gradient descent

Set $\boldsymbol{\theta}_0 := \boldsymbol{0}$, $\gamma \in \mathbb{R}_{++}$, $\mu \in \mathbb{R}_{++}$.
For $k = 0, 1, 2, \cdots$
$\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k - \mu\nabla_{\boldsymbol{\theta}}\Phi^{\text{MC-pin}}_{\alpha,\gamma}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}_k)$

Here, $\nabla_{\boldsymbol{\theta}}\phi^{\text{MC-pin}}_{\alpha,\gamma}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta}) = -\nabla_z\phi^{\text{MC-pin}}_{\alpha,\gamma}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta})\boldsymbol{x}_i$, where

$$\nabla_z\phi^{\text{MC-pin}}_{\alpha,\gamma}(z) \in \partial_z\phi^{\text{MC-pin}}_{\alpha,\gamma}(z) = \tag{10}$$

$$\begin{cases}\left\{\alpha - \alpha^2\frac{z}{\gamma}\right\}, & \text{if } 0 < z \le \frac{1}{\alpha}\gamma,\\ \left\{-(1-\alpha) - (1-\alpha)^2\frac{z}{\gamma}\right\}, & \text{if } -\frac{1}{1-\alpha}\gamma \le z < 0,\\ [-(1-\alpha), \alpha], & \text{if } z = 0,\\ \{0\}, & \text{otherwise.}\end{cases} \tag{11}$$

Here, ∂_z is Clarke's generalized gradient operator with respect to z [14]. In simulations, we simply let $\nabla_z\phi^{\text{MC-pin}}_{\alpha,\gamma}(0) = 0 \in [-(1-\alpha), \alpha]$.

The stepsize is set to $\mu = 0.001$. The performance is evaluated with test datasets of sample size 1000 (generated in the same way as the training datasets) by two metrics: (i) the system mismatch $\frac{\|\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\alpha,\star}\|_2^2}{\|\boldsymbol{\theta}_{\alpha,\star}\|_2^2}$ with the ℓ_2 norm $\|\cdot\|_2$, where $\hat{\boldsymbol{\theta}}_\alpha$ is the estimate of the true αth quantile $\boldsymbol{\theta}_{\alpha,\star}$, and (ii) coverage rate $\frac{1}{|\mathcal{I}_{\text{r}}|}\sum_{i\in\mathcal{I}_{\text{r}}} 1_{S_y(\boldsymbol{x}_i)}(y_i)$, where $|\mathcal{I}_{\text{r}}|$ is the number of reliable data, and

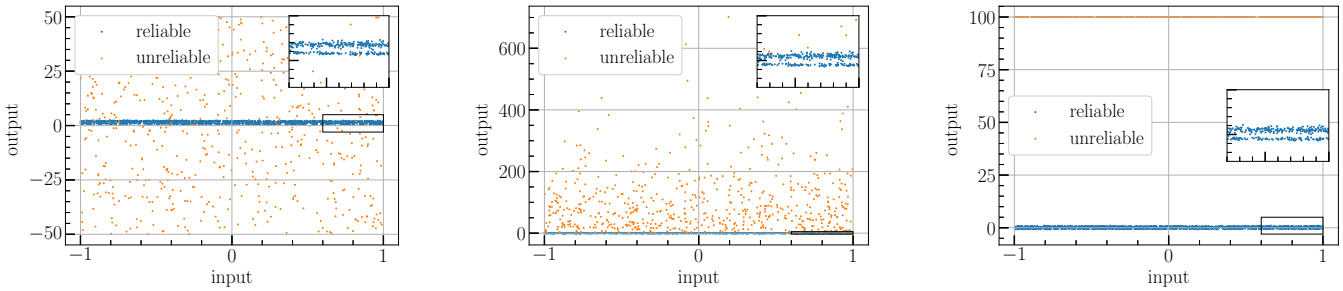
$$1_{S(\boldsymbol{x}_i)}(y_i) := \begin{cases}1 & \text{if } y_i \in S(\boldsymbol{x}_i) := [\boldsymbol{x}_i^\top\hat{\boldsymbol{\theta}}_{0.1}, \boldsymbol{x}_i^\top\hat{\boldsymbol{\theta}}_{0.9}],\\ 0 & \text{otherwise.}\end{cases} \tag{12}$$

The results are averaged over 300 independent trials.

B. Experiment A: Robustness under different proportions β of unreliable data

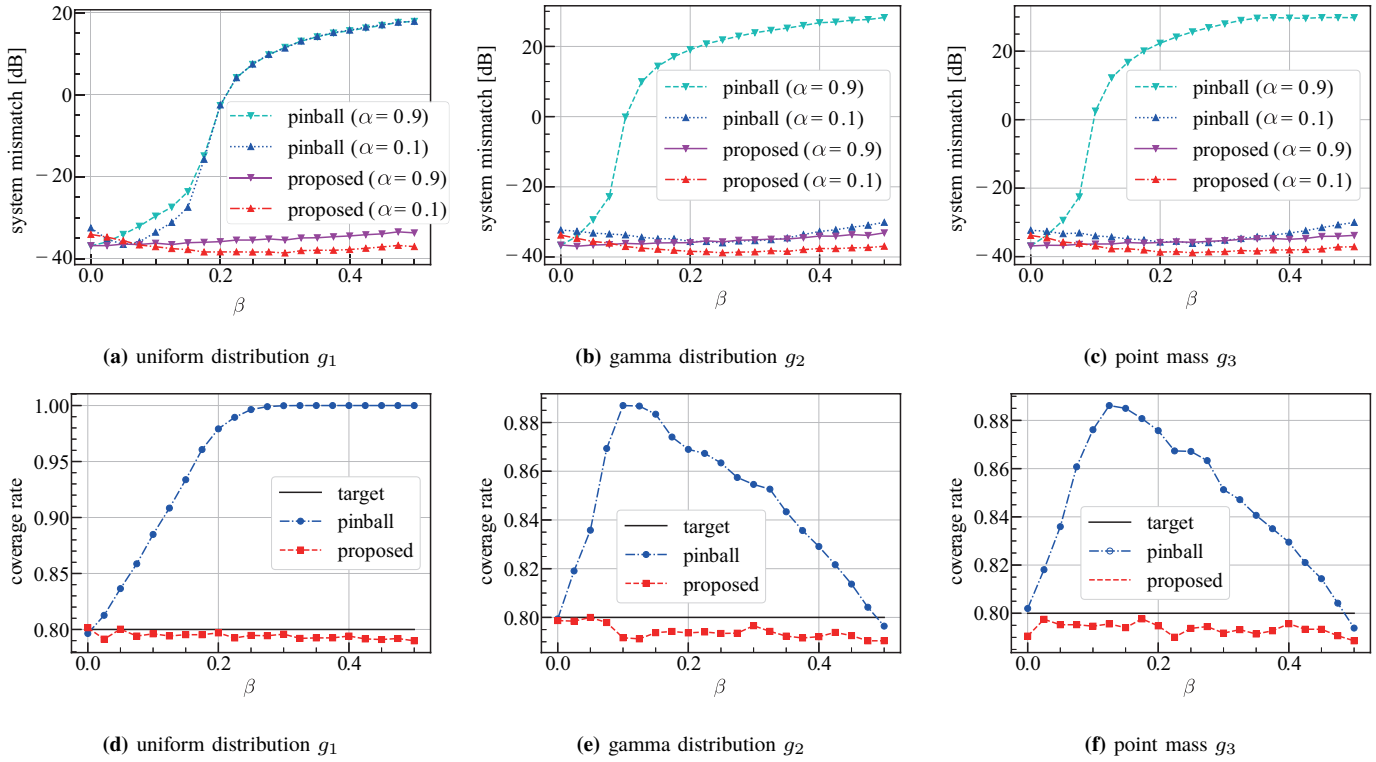
The results are plotted in Fig. 5 for each density of unreliable data. For all types of density, the proposed approach attains low system mismatches and coverage rates close to the target rate 0.8 $(= 0.9 - 0.1)$ over the whole range of β. (The average error rates of coverage rate for each distribution g_i were (d) 6.0×10^{-3}, (e) 6.0×10^{-3}, and (f) 6.5×10^{-3}.) This means that the proposed approach enjoys remarkable robustness for quantile regression even when the proportion of unreliable data is large. As expected, on the other hand, the performance of the pinball loss deteriorates significantly.

To be more specific, we focus on the case of uniform distribution g_1. In this case, unreliable data arise both above and below the region of reliable data, making the upper and lower quantile estimates be larger and smaller, respectively, so that the interval expands with β. This is the reason why both upper and lower mismatches (the system mismatches of both upper and lower quantiles) increase with β. Let us now turn our attention to the cases of g_2 and g_3. In contrast to the case of g_1, unreliable data arise only above the reliable-data region.



(a) uniform distribution g_1 **(b)** gamma distribution g_2 **(c)** point mass g_3

Fig. 4: Samples of ε generated from $0.7f + 0.3g_i$, $i \in \{1, 2, 3\}$, i.e., $\beta := 0.3$.



(a) uniform distribution g_1 **(b)** gamma distribution g_2 **(c)** point mass g_3

(d) uniform distribution g_1 **(e)** gamma distribution g_2 **(f)** point mass g_3

Fig. 5: Results of Experiment A: performance comparisons for different proportions of unreliable data.

Because of this, the upper mismatch increases sharply when β reaches 0.1 so that $y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}_{0.9}$ exceeds some part of ϵ_is of the unreliable data. Once $\boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}_{0.9}$ exceeds all y_is of the "reliable" data, the coverage rate starts to decrease due to the gradual increase of $\boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}_{0.1}$.

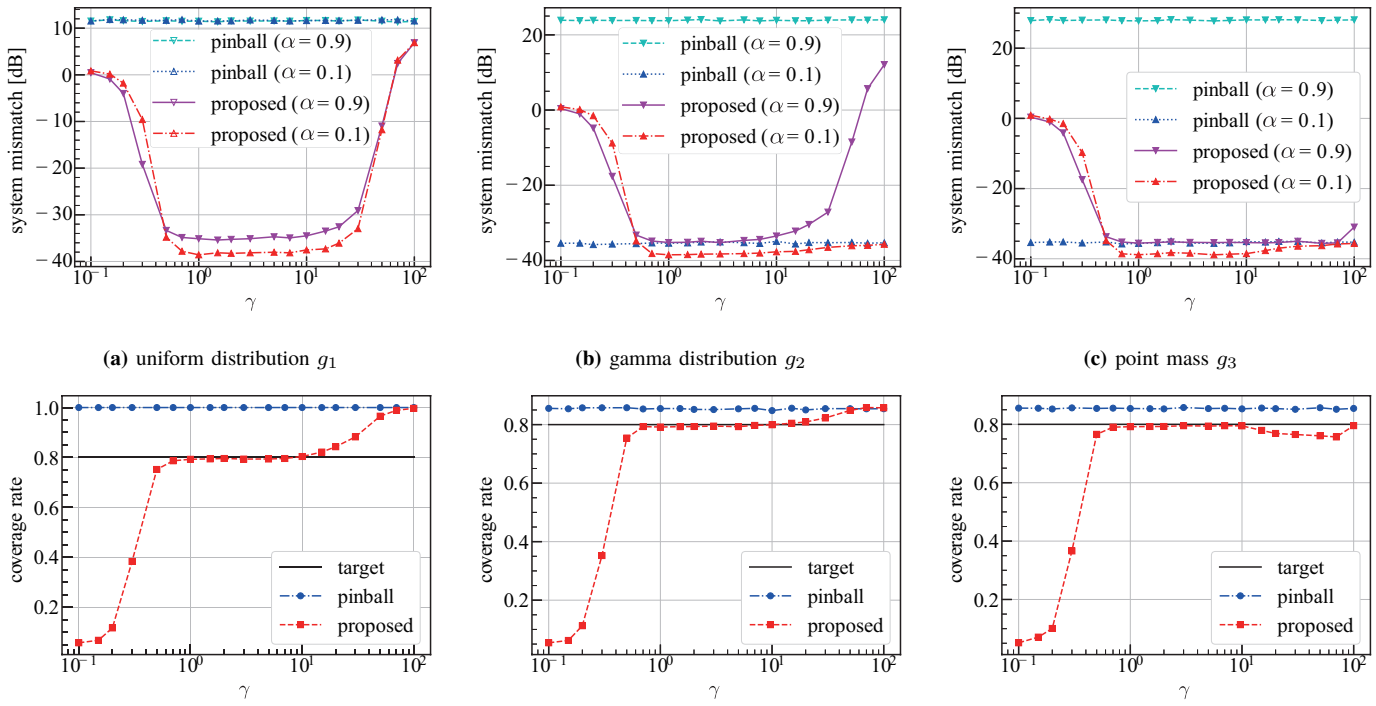
C. Experiment B: Insensitivity to the choice of γ

Fig. 6 shows the results for Experiment B. For all densities g_i, the performance is stable in the neighborhood of the best γ in both metrics. Remarkably, the range of good γ is common among the three cases; i.e., a $\gamma \in [1, 10]$ gives reasonably good performance for all g_is. This indicates the saturation parameter γ would be simple to tune. This is a potential advantage of the proposed approach. As expected from the arguments presented in Remark 1, a too large or too small γ gives deteriorated performance.

V. Concluding Remarks

We proposed the MC-pinball loss for the task of robust quantile regression in the presence of unreliable data. The proposed loss is the composition of the MC penalty and the pinball loss, enjoying the nice properties as opposed to the Moreau-enhanced pinball loss. The numerical examples suggested that the proposed approach yields robust estimates of the true quantile in the presence of unreliable data. The potential benefit of the proposed approach was also shown in terms of the parameter tuning.

Although the proposed loss is nonconvex, there was no



(a) uniform distribution g_1 **(b)** gamma distribution g_2 **(c)** point mass g_3

(d) uniform distribution g_1 **(e)** gamma distribution g_2 **(f)** point mass g_3

Fig. 6: Results of Experiment B: insensitivity to the choice of γ.

issue of dependency on the initialization nor local minima etc. in our simulations. This may suggest that the proposed loss could have benign nonconvex landscapes, which needs further detailed investigations. Sensor malfunctions or measurement errors (which could be the source of unreliable data) are typically tackled at the hardware level, and the human errors are handled by costly preprocessing. Those errors could be supported by the proposed approach at the software level, and this is a potential advantage from the practical viewpoint.

References

[1] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[2] C. Gallego-Castillo, R. Bessa, L. Cavalcante, and O. Lopez-Garcia, "On-line quantile regression in the RKHS (Reproducing Kernel Hilbert Space) for operational probabilistic forecasting of wind power," *Energy*, vol. 113, pp. 355–365, 2016.

[3] J. B. Bremnes, "Probabilistic wind power forecasts using local quantile regression," *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, vol. 7, no. 1, pp. 47–54, 2004.

[4] S. B. Taieb, R. Huser, R. J. Hyndman, and M. G. Genton, "Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2448–2455, 2016.

[5] P. S. Martins and P. T. Pereira, "Does education reduce wage inequality?: Quantile regression evidence from 16 countries," *Labour economics*, vol. 11, no. 3, pp. 355–371, 2004.

[6] G. L. Gaeta, G. Lubrano Lavadera, and F. Pastore, "Overeducation wage penalty among ph.d. holders: an unconditional quantile regression analysis on italian data," *International journal of manpower*, vol. 44, no. 6, pp. 1096–1117, 2023.

[7] F. Jareño, M. d. l. O. González, M. Tolentino, and K. Sierra, "Bitcoin and gold price returns: A quantile regression and nardl analysis," *Resources policy*, vol. 67, pp. 101 666–, 2020.

[8] T. Ichinose, M. Yukawa, and R. L. Cavalcante, "Online kernel-based quantile regression using huberized pinball loss," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 1803–1807.

[9] Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[10] M. Yukawa, H. Kaneko, K. Suzuki, and I. Yamada, "Linearly-involved Moreau-enhanced-over-subspace model: Debiased sparse modeling and stable outlier-robust regression," *IEEE Trans. Signal Process.*, vol. 71, pp. 1232–1247, 2023.

[11] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, Apr. 2010.

[12] I. Selesnick, "Sparse regularization via convex analysis," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4481–4494, Sep. 2017.

[13] J. Abe, M. Yamagishi, and I. Yamada, "Linearly involved generalized moreau enhanced models and their proximal splitting algorithm under overall convexity condition," *Inverse Problems*, vol. 36, no. 3, pp. 1–36, 2020.

[14] F. H. Clarke, "Generalized gradients and applications," *Transactions of the American Mathematical Society*, vol. 205, pp. 247–262, 1975.

Measurement of Relative Transfer Function for Own Voice in Head-Mounted Microphone Array

Kyoka Kazama, Taishi Nakashima and Nobutaka Ono
Tokyo Metropolitan University, Tokyo, Japan
E-mail: kazama-kyoka@ed.tmu.ac.jp, taishi@tmu.ac.jp, onono@tmu.ac.jp

***Abstract*—In this paper, we investigate a suitable vocalization method for measuring the relative transfer function (RTF) of a user's own voice in head-mounted microphone arrays (HMMAs). The realization of auditory devices that can extract a desired voice from a mixture is of great interest. Blind source separation can separate the target speaker's voice, but it also removes the user's own voice, making conversation difficult. Our aim is to estimate the RTF of own voice in advance to prevent such removal and improve dialogue quality. For RTF measurements using a loudspeaker, a broadband signal such as a time-stretched pulse (TSP) signal is typically used as the source signal. However, when measuring the RTF of own voice, it is not possible to produce the TSP signal. To address this, we investigate which vocalizations are suitable for RTF measurement. We evaluate the consistency of estimated RTF from three types of vocalization using a self-fabricated HMMA. Our results show that vocalization with smooth pitch variation is more suitable for RTF measurement than others. Furthermore, the results of individual dependency on RTFs obtained from male and female subjects showed little individual dependence in the low-frequency band.**

I. Introduction

Extracting the target signal from a mixture signal with noise and interference is a crucial task in acoustic signal processing. Classical methods include beamforming, which estimates a filter to enhance the target signal using prior information such as steering vectors and noise intervals. Popular beamformers include the minimum variance distortionless response beamformer and the linearly constrained minimum variance beamformer [1], [2]. Blind source separation (BSS) estimates filters to suppress interference by exploiting the statistical independence of the source signals [3]. Prominent methods in this area are independent vector analysis (IVA) [4], [5] and auxiliary function-based IVA [6]. Real-time applications of BSS have also been actively studied [7]–[9].

In recent years, acoustic signal processing for wearable devices has been actively researched [10]–[12]. For example, hearing aids require improvement of speech intelligibility in noisy environments [13], [14]. With the development of augmented reality (AR), acoustic signal processing on smart glasses has also been actively researched [15]–[17]. Real-world multimodal datasets using AR glasses has been published, and its challenges were also held [18], [19]. In situations where a wearable device enhances the target speaker's voice in real time during a conversation, it is crucial not only to emphasize the target speaker's voice but also to ensure that the user can clearly hear their own voice. Even though BSS is effective for enhancing the target speaker's voice, simply applying it may unintentionally suppress the user's own voice.

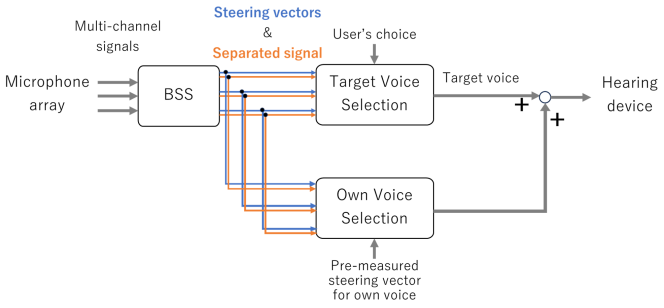


Fig. 1. Flow of the speech enhancement system without removing the users' own voice

One way to avoid the removal of the user's own voice is to pre-measure the RTF (Relative Transfer Function) of the user's voice and use it, as the RTF is expected to remain stable due to the consistent relative position between the microphones of the head-mounted device and the user's mouth. Fig. 1 illustrates the flow of the speech enhancement system we are considering using BSS. The multichannel signal from the microphone array is input into the BSS, which outputs a pair of steering vectors and separated signals. These are used to select both the target speaker's voice and the user's own voice. The Target Voice Selection block allows the user to choose the desired speaker's voice, while the Own Voice Selection block uses the pre-measured steering vector to automatically select the user's own voice. By summing these two signals and presenting them through the hearing device, a smoother conversation is expected.

The key issue here is how to measure the RTF. Typically, the transfer function is measured by playing known source signals from a loudspeaker [20]. These source signals often include broadband signals like white noise and time-stretched pulse (TSP) signals. However, for most practical applications, direct measurement of the transfer function is not feasible, thus the RTF is widely used [21]–[25]. When measuring the RTF, it is preferable to use broadband source signals, although the source signal does not need to be known. In cases where the source is the user's own voice, generating broadband signals like white noise by vibrating the human vocal cords is not practical.

Therefore, in this study, we investigate vocalizations for measuring the RTF using the user's own voice. A custom-made head-mounted microphone array (HMMA) is used to

record the user's voice and measure the RTF. Specifically, we evaluate the estimation accuracy and individual dependency for three types of vocalizations: reading, singing, and pitch sweep.

II. Relative Transfer Function of Own Voice

Let us consider an HMMA with M microphones in the short-time Fourier transform (STFT) domain, depicted in Fig. 2. Let $S(\omega, \tau)$ be the source signal, and $X_m(\omega, \tau)$ $(m = 1, \ldots, M)$ be the observed signal at mth microphone modeled as

$$X_m(\omega, \tau) = a_m(\omega) S(\omega, \tau), \tag{1}$$

where ω and τ are the frequency and time frame indices, respectively, and $a_m(\omega)$ is the transfer function from the source to the mth microphone.

Normalizing each transfer function with the one of the reference microphone r, (1) can be rewritten as:

$$X_m(\omega, \tau) = H_{rm}(\omega) X_r(\omega, \tau), \tag{2}$$

$$H_{rm}(\omega) = \frac{a_m(\omega)}{a_r(\omega)}, \tag{3}$$

$$X_r(\omega, \tau) = a_r(\omega) S(\omega, \tau), \tag{4}$$

where $H_{rm}(\omega)$ denotes the RTF between each microphone and the reference microphone, and $X_r(\omega, \tau)$ is the observed signal at the reference microphone. RTF can be estimated as the solution that minimizes the following least-squares error criterion using only observed signal:

$$J = \sum_{\tau} |X_m(\omega, \tau) - H_{rm}(\omega) X_r(\omega, \tau)|^2, \tag{5}$$

$$\hat{H}_{rm}(\omega) = \frac{\sum_{\tau} X_r^*(\omega, \tau) X_m(\omega, \tau)}{\sum_{\tau} |X_r(\omega, \tau)|^2}. \tag{6}$$

$\hat{H}_{rm}(\omega)$ minimizes the above least-squares error (5). However, even though the source signal does not need to be known, if there exists an ω such that $\sum_{\tau} |X_r(\omega, \tau)|^2$ in the denominator of (6) becomes close to zero, the estimation accuracy at that frequency will decrease. Therefore, similar to impulse response estimation, it is desirable to have a broad frequency band in the source signal. When the source signal is a voice, what types of vocalizations can yield a more broadband signal is an interesting research question. Thus, in this study, we investigate the suitable types of vocalization for measuring the RTF of own voice.

III. Experiments

A. Recording

We conducted recordings to investigate the suitable types of speech for measuring the RTF of own voice. First, we fabricated an HMMA with six MEMS microphones arranged in a circular array: two on the front of the glasses frame and two on each side, all positioned along the same circumference as showned in Fig. 3. The RASP-ZX by SYSTEM IN FRONTIER INC. was used to interface the microphones and the laptop.

Next, the subjects wearing the HMMA performed three types of vocalizations: 1) `reading`, 2) `singing`, and 3) `pitch sweep` in a soundproof room as shown in Fig. 4. The `reading` involved reading six sentences selected from the JVNV corpus [26], producing approximately 50 seconds of audio. Since `reading` includes unvoiced sounds such as consonants, it may be easier to cover high-frequency bands. The `singing` involved singing a phrase from a traditional Japanese children's song, "Momiji," producing approximately 10 seconds of audio. Since the phrase includes predetermined pitches corresponding to specific musical notes, certain frequencies may lack sufficient power. The `pitch sweep` involved three repetitions of smoothly raising the pitch from the lowest possible note to the highest, producing approximately 15 seconds of audio. Since the `pitch sweep` resembles a TSP signal, it is expected to cover a wide frequency range. However, it may be somewhat demanding for the subjects to perform. The sampling frequency was 16 kHz. Fig. 5 shows an example of a spectrogram of recorded voice. These three types of vocalizations were recorded in five sets for each subject by two persons; one female and one male. After each set, the subjects were removed and put again in the HMMA. Under these conditions, we investigated which type of vocalization was most suitable for measuring RTF and the individual independence of the RTF of own voice using the HMMA.

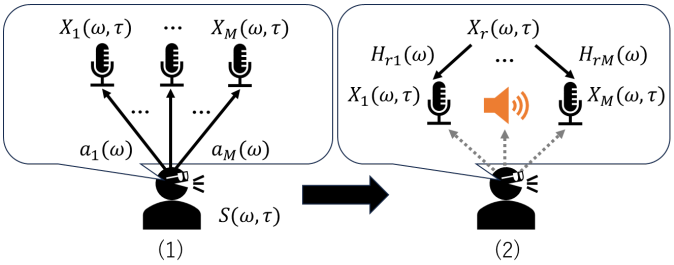


Fig. 2. Diagrams of the observed signal model: (1) model using transfer functions, (2) model using relative transfer functions

B. Relative transfer function measurement

We extracted 10 seconds of audio from each vocalization and estimated RTF based on (5). We set `FRONT-L` in Fig. 3 as the reference microphone on the glasses' front left side. The STFT parameters were a frame length of 512, a frame shift of 256, and a Hann window as the window function.

Since the actual RTF is unknown, it is not easy to directly evaluate the measurement accuracy of RTF. Therefore, in this study, we used the standard deviation (SD) to evaluate the degree to which the estimated RTF matches across multiple measurements. Precisely, it was calculated for each frequency bin ω from the logarithmic amplitude of the RTF measured five times under the three vocalizations. The logarithmic amplitude $A(\omega)$ was calculated as:

$$A(\omega) = 20 \log_{10} |\hat{H}_{rm}(\omega)|. \tag{7}$$

The smaller the SD value, the more stable RTF is and the higher the estimation accuracy. In addition, from the results of

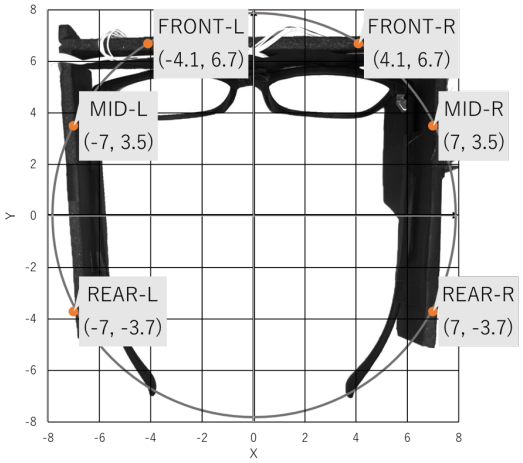


Fig. 3. Microphone geometry of our head-mounted microphone array.



Fig. 4. Female subject wearing the head-mounted microphone array in a soundproof room.

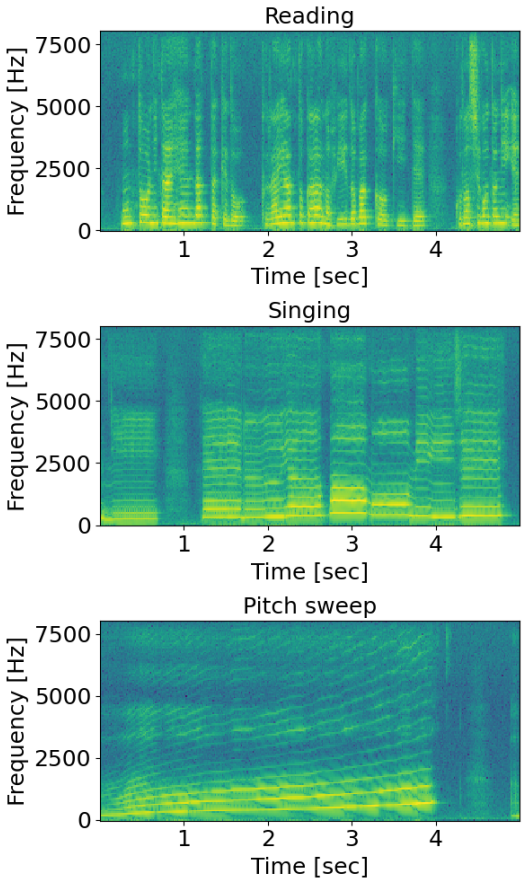


Fig. 5. Examples of spectrograms of recorded voices.

the SD evaluation, we select one vocalization that is suitable for measuring the RTF of speech and evaluate the individual independence of the RTF for that vocalization. Individual independence is evaluated by comparing the average logarithmic amplitude $A(\omega)$ and phase $\phi(\omega) = \angle \hat{H}_{rm}(\omega)$ of the RTF measured five times between two subjects.

C. Comparison of Power Spectra

Before running the experiment, we checked how much power is included in each frequency bin of the recorded data. The power $P(\omega)$ is calculated by adding up the amplitude values of each time frame of the observed signal at the reference microphone for each frequency bin:

$$P(\omega) = 10 \log_{10} \sum_{\tau} |X_r(\omega, \tau)|^2. \tag{8}$$

It is thought that the larger the power, the more stable RTF can be measured.

Fig. 6 shows the average power for each frequency obtained from the voices recorded five times under each speaking condition. The results show that the power is insufficient in the high-frequency band compared to the low-frequency band. This suggests that the signal-to-noise ratio (SNR) is insufficient in high-frequency bands, making it difficult to measure RTF accurately. Comparing each vocalization, there is a large variation in power in the low-frequency band for `reading` and `singing`. On the other hand, it can be confirmed that the power in the low-frequency band is stable in the `pitch sweep`. Comparing each speaker, female speakers have more power in relatively high-frequency bands, and male speakers have more power in low-frequency bands. However, there are no large differences in the bands with power, so the difference is insignificant in RTF measurement.

D. Results

Fig. 7 shows the SD of each vocalization. The results confirm that the SD of each vocalization is small in the low-frequency band and large in the high-frequency band. This is likely because the power in the high-frequency band of the voice is insufficient, and the SNR is not sufficient compared to the low-frequency band (6). It also suggests the possibility that the RTF itself has complex characteristics with peaks and dips. However, it can be seen that the RTF can be measured stably in the low-frequency band, which is the main band of speech. Focusing on the low-frequency band, it can be seen that the SD of `pitch sweep` is the smallest for both male speakers (Fig. 7 right) and female speakers (Fig. 7 left). This is thought to be because, as can be seen from Fig. 6, by raising

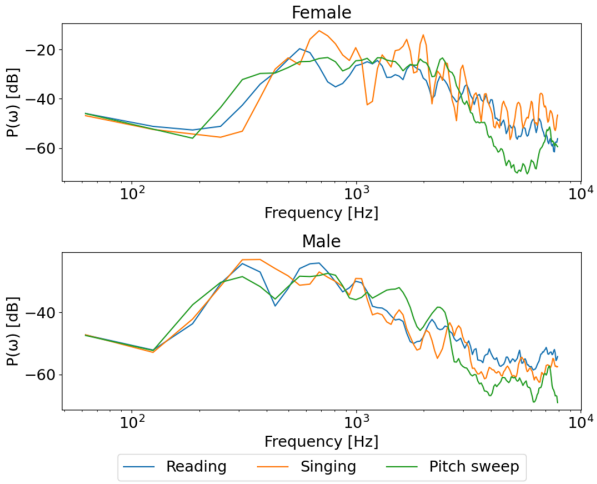


Fig. 6. Power spectra of two subjects

the pitch smoothly from bass to treble like the TSP signal, it approaches a sound source with constant power over a wide band. From this, `pitch sweep` is a suitable vocalization for measuring the RTF of own voice.

Next, Fig. 8 shows the results of comparing average the logarithmic amplitude and phase of the RTF measured five times with `pitch sweep` for two subjects. The results of logarithmic amplitude (Fig. 8 left) show that, although some deviations in peak positions are observed, the two subjects obtain approximately the same amplitude characteristics, especially in the low-frequency band. The phase results (Fig. 8 right) show that sound diffraction around the head affects the `REAR-L` and `REAR-R` microphones placed behind the glasses. However, overall, the two subjects obtained roughly the same phase characteristics. This may be due to individual differences such as head size, in the case of HMMA, the positional relationship between the microphone and the mouth does not vary significantly depending on the individual. In the future, it will be necessary to conduct experiments with more subjects and verify whether the RTF can be measured similarly.

IV. Conclusion

In this study, we investigated the suitable vocalization for measuring the RTF of own voice in HMMA and evaluated the individual independence of the RTF. The subject's own voice was recorded using an HMMA that was originally prototyped. As a result of the experiments, it was found that vocalizations that raise the pitch smoothly, like TSP signals, are suitable for measuring the RTF of own voice. In addition, it was suggested that the RTF of own voice in HMMA has little dependence on the individual in the low-frequency band. Although further experiments with a larger number of subjects is necessary, it is expected that the same RTF can be used for any speakers. In the future, we aim to achieve real-time BSS using the RTF of the HMMA, allowing for the separation of surrounding sounds without canceling out the user's own voice.

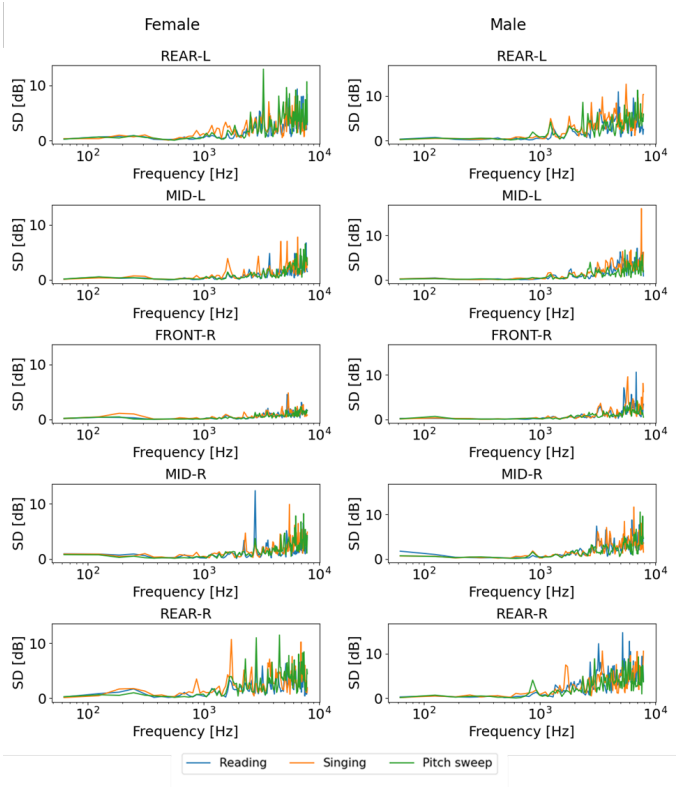


Fig. 7. Standard deviation of RTF for three vocalizations over five measurements.

Acknowledgment

This research was supported by JST CREST (JPMJCR19A3).

References

[1] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[2] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[3] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Transactions on Signal and Information Processing*, vol. 8, e12, 2019.

[4] T. Kim, T. Eltoft, and T. W. Lee, "Independent vector analysis: An extension of ica to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.

[5] A. Hiroe, "Solution of permutation problem in frequency domain ica using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.

[6] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

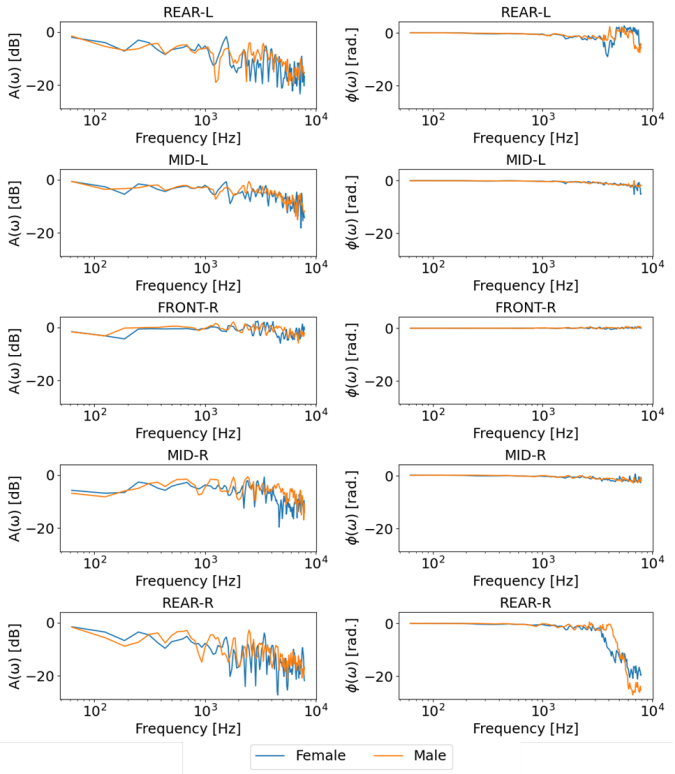


Fig. 8. Comparison of the relative transfer function obtained from male and female subjects.

[7] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. HSCMA*, IEEE, May 2014, pp. 107–111.

[8] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *Proc. ICASSP*, 2017, pp. 216–220.

[9] Y. Kuriki, T. Nakashima, K. Yamaoka, *et al.*, "Efficient low-latency convolution with uniform filter partition and its evaluation on real-time blind source separation," in *Proc. APSIPA ASC*, 2022, pp. 766–770.

[10] P. Hoang, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Multichannel speech enhancement with own voice-based interfering speech suppression for hearing assistive devices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 706–720, 2022.

[11] Y. Wakabayashi, K. Yamaoka, and N. Ono, "Sound field interpolation for rotation-invariant multichannel array signal processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2286–2298, 2023.

[12] T. Nakashima, Y. Wakabayashi, and N. Ono, "Self-rotation-robust online independent vector analysis with sound field interpolation on circular microphone array," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 1, 2024.

[13] C. Haruta and N. Ono, "Low complexity of DNN based speech enhancement for hearing aid systems," *The Journal of the Acoustical Society of Japan*, vol. 78, no. 5, pp. 227–237, 2022.

[14] M. Ohlenbusch, C. Rollwage, and S. Doclo, "Multi-microphone noise data augmentation for DNN-Based own voice reconstruction for hearables in noisy environments," in *Proc. ICASSP*, Apr. 2024, pp. 416–420.

[15] M. A. Cantu and V. Hohmann, "Enhancement of hearing aid processing via spatial spectro-temporal post-filtering with a prototype eyeglass-integrated array," in *Proc. IWAENC*, 2022, pp. 1–5.

[16] S. Hafezi, A. H. Moore, P. Guiraud, *et al.*, "Subspace hybrid beamforming for head-worn microphone arrays," in *Proc. ICASSP*, 2023, pp. 1–5.

[17] R. Arakawa, M. Parvaix, C. Lai, H. Erdogan, and A. Olwal, "Quantifying the effect of simulator-based data augmentation for speech recognition on augmented reality glasses," in *Proc. ICASSP*, 2024, pp. 726–730.

[18] J. Donley, V. Tourbabin, J.-S. Lee, *et al.*, *EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments*, 2021. arXiv: 2107.04174 [cs.SD].

[19] P. Guiraud, S. Hafezi, P. A. Naylor, *et al.*, "An introduction to the speech enhancement for augmented reality (spear) challenge," in *Proc. IWAENC*, 2022, pp. 1–5.

[20] F. Satoh, "Measurement techniques of room impulse-response," *The Journal of the Acoustical Society of Japan*, vol. 58, no. 10, pp. 669–676, 2002.

[21] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[22] Z. Koldovsky, J. Malek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1335–1347, Aug. 2015.

[23] Z. Koldovsky and P. Tichavsky, "Sparse reconstruction of incomplete relative transfer function: Discrete and continuous time domain," in *Proc. EUSIPCO*, IEEE, Aug. 2015, pp. 394–398.

[24] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[25] N. Gossling and S. Doclo, "Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field," in *Proc. IWAENC*, IEEE, Sep. 2018, pp. 146–150.

[26] D. Xin, J. Jiang, S. Takamichi, Y. Saito, A. Aizawa, and H. Saruwatari, "JVNV: A corpus of japanese emotional speech with verbal content and nonverbal expressions," *IEEE Access*, vol. 12, pp. 19 752–19 764, 2024.

Dictionary Learning Based Two-stage Near-lossless Video Compression

Zuhai Zhang[†], Luheng Jia[†]*, Li Song[‡], Shuyuan Zhu[§], Yuanfang Guo[¶] and Kebin Jia[†]
[†] Beijing University Of Technology, Beijing
[‡] Shanghai Jiao Tong University, Shanghai
[§] University of Electronic Science and Technology of China, Sichuan
[¶] Beihang University, Beijing
Corresponding email: luhengjia@bjut.edu.cn

***Abstract*—Traditional hybrid video coding framework using block based predictive coding and transform coding, such as the High Efficiency Video Coding (HEVC), cannot further dig out the redundancy remained in quantized transformed residual, causing extra bits consumption. Measured by rate-distortion (R-D) performance, the problem of higher bits consuming can be solved reversely by video quality enhancing. In this work, we proposed a video coding scheme that solve the problem by enhancing the reconstructed video quality using supplementary information from further compressed quantization error. Aiming at better R-D performance for near-lossless video coding, we propose a novel video coding scheme using a two-stage framework that extracts quantization error as complementary information which is compressed using dictionary learning and sparse representation. The employed over-complete dictionary is learned through K-SVD with orthogonal matching pursuit (OMP) for sparse representation. Statistically redundancy is further removed by a modified context-adaptive binary arithmetic coding (CABAC) with adaptive context models. This approach not only retains the advantages of the traditionally encoder for lossy compression but also exploits the redundancy in the quantization error to achieve high-quality near-lossless compression. Experimental results demonstrate that our method significantly outperforms traditional HEVC lossy encoder with over -20% BD-BR on average at high bitrate range for near-lossless coding, while the method is also proved to be efficient at low bitrate range achieving over -50% BD-BR on average with average PSNR over 41dB, which retains near-lossless performance.**

I. Introduction

Nowadays, the market share of the HEVC [1] is still rising with its ability to efficiently compress high-quality videos including 4K and 8K ultra-high-definition (UHD) videos. Besides the pursuits for higher resolution, lossless and near-lossless compression that provides higher data security and better quality have become a trend in recent years. Near-lossless video compression aims to preserve a significant amount of detail such that the loss of information is imperceptible to the human eye. Near-lossless compression offers a balanced trade-off between video bitrate and quality than fully lossless encoding and common lossy encoding in many applications such as medical video compression and surveillance video compression. However, most of the coding tools in HEVC are geared towards lossy compression. Merely using lower quantization parameters to achieve near-lossless compression often fails to fully leverage the encoder's potential for effective compression. There is still substantial room for improvement in optimizing near-lossless compression.

To ensure compression efficiency while preserving image details as much as possible, near-lossless compression was introduced and applied early on in JPEG-LS [2]. Subsequently, various near-lossless compression algorithms such as [3], [4] are incorporated into video encoding. Both approaches aim at video compression at high bitrate range. The former uses an adaptive strategy to select the optimal prediction algorithm to achieve smaller residuals, while the latter optimizes the quantization scheme at high bitrate range to achieve better compression results. While these methods attempt to modify traditional video encoders, we propose an optimization approach using a two-stage framework. This encoding process framework aims to enhance the encoder's efficiency while preserving its inherent advantages, allowing for flexible adjustments in compression rate. Moreover, the context models in context-adaptive binary arithmetic coding (CABAC) adopted in standard video encoder are not designed for lossless or near-lossless compressed data, causing reduced coding efficiency.

In addition to direct improvements to the encoder, another approach to lossless and near-lossless video compression adopts two-stage frameworks [5]–[8]. The two-stage framework decomposes video encoding process into two parts including a lossy compression stage and an subsequent error components compression stage. Heindel et al. [8] modified the two-stage encoding framework for near-lossless video compression. Bai et al. [9] introduced a learning-based coding strategy into the two-stage framework for near-lossless encoding. Both approaches have demonstrated the efficiency of the two-stage framework in near-lossless compression. Additionally, scalable video coding framework [7] with base-layer and enhancement-layer is considered as a two-stage coding framework. However, these approaches have not employed entropy encoder specifically designed according to the statistic

This work was supported by National Natural Science Foundation of China under Grant 61901012 and Science and Technology Commission of Shanghai Municipality under Grant 22DZ2229005. (*Corresponding author: Luheng Jia.*)

characteristics of the error components in the second encoding stage, and hence statistical redundancy has not been efficiently removed.

To address the aforementioned drawbacks of current methods, we propose a video coding scheme operates within a two-stage framework, employing dictionary learning and sparse representation to compress the residuals, followed by an specifically designed entropy encoder adapting the characteristics of sparse represented residual. The proposed modified two-stage framework retains efficiency of the standard lossy video encoder in the first stage, while the second stage leverages difference between original frame and reconstruction frames as complementary information to enhance the quality of reconstruction frame consuming small extra bits. And hence, better rate-distortion (R-D) performance is achieved. Additionally, the complementary information in the second stage is sparse represented using an patch-size adaptive dictionary learning method to characterize various distribution features of residuals at different bitrate. Finally, the sparse coefficients are entropy encoded using a modified CABAC-based entropy encoder with specificlly designed context models. The proposed scheme achieves significant coding efficiency gain for near-lossless coding scenario at high bitrate range, while the scheme outperforms standard lossy video encoder at lower birate range.

The remainder of the paper is organized as follows. Section II introduces the dictionary learning and sparse representation methods used. Section III describes the proposed framework and entropy encoding strategy. Section IV presents the experimental results. Section V concludes the paper.

II. Problem Formulation of R-D Optimized Two-Stage Near-lossless Video Coding

The proposed two-stage encoding framework is shown in Fig.1. We denote the original current frame as x. And the reconstructed version of x from the first stage is denoted as $\widehat{x}$. The difference between x and $\widehat{x}$ is denoted as r. And we have

$$r = x - \widehat{x} \tag{1}$$

which is adopted as complementary information to enhance the quality of final reconstructed frame $\widehat{x}'$. To reduce the extra bit consumption, r is sparse represented and entropy coded in the second stage. The reconstructed complementary information from the second stage is denoted as $\widehat{r}$. The final reconstructed frame $\widehat{x}'$ is obtained as follows.

$$\widehat{x}' = \widehat{x} + \widehat{r} \tag{2}$$

With equation (1) and (2), the final reconstruction error e is obtained as follows.

$$e = x - \widehat{x}' = r - \widehat{r} \tag{3}$$

Without using quantization as traditional lossy compression, the compression method adopted in the second stage controls R-D trade-off by sparsity setting, which can introduce very limited distortion. Therefore, $|e|$ is much smaller than $|r|$ leading to better reconstruction quality and near-lossless compression performance. To be noticed, r is the reconstruction error of traditional video encoder. The encoding distortion D of traditional video encoder is the variance of r, which is denoted as σ_r^2. While the encoding distortion D' of the proposed two-stage encoding framework is the variance of e, which is denoted as σ_e^2. The distortion difference can be considered as quality improvement. We assume e and r are of zero mean. We have

$$\begin{aligned} \Delta D &= D - D' = \sigma_r^2 - \sigma_e^2 \\ &= 2E(r \cdot \widehat{r}) - E(\widehat{r}^2) \\ &= 2 \cdot E(r^2) - 2 \cdot E(r \cdot e) - E(r^2) - E(e^2) + 2 \cdot E(r \cdot e) \\ &= E(r^2) - E(e^2) \end{aligned} \tag{4}$$

Under near-lossless coding scenario, $E(e^2)$ as the final reconstruction error of the two-stage coding scheme is small. By carefully adjusting the sparsity of sparse represented coefficients, we can always obtain $E(e^2)$ smaller than $E(r^2)$ and hence $\Delta D \geq 0$.

By employing the widely used logarithmic R-D model $R = \alpha \cdot log_2 \frac{\sigma^2}{D}$ [10], the D-R model for the traditional video encoder is as follows.

$$D = \frac{\sigma^2}{2^{\frac{R}{\alpha}}} \tag{5}$$

in which σ^2 is the quantized transformed coefficients of predictive residual before entropy coding. For easy comparison, we assuming the first stage of the proposed scheme using the same coding parameters set as standard video encoder, and hence the first stage has the same R-D characteristic as in (5). And the R-D characteristics of second stage in the proposed two-stage encoding framework is modeled as follows.

$$D' = \frac{\sigma'^2}{2^{\frac{R'}{\alpha}}} \tag{6}$$

where σ'^2 is the variance of the processed input difference from first stage before entropy coding. R' is the extra bits consumed in the second stage. We formulate the two-stage video coding as maximize the quality difference with given extra bits budget using the two-stage framework compared with the traditional standard encoder.

$$\max_{\theta} \ \Delta D \ \ s.t. \ \Delta R \leq R_T \tag{7}$$

where ΔD is the improved quality by introducing the complimentary information in the second stage as in (4). θ is the coding parameter sets. And R_T is the target extra bit budget. With the first stage encoder the same as the comparative standard encoder and D is considered as a constant, maximizing the distortion difference is equivalent to minimizing the final reconstruction distortion D'. Additionally, ΔR is the extra bits cost in the two-stage framework, which is R' if we use the same encoder in the first stage as the standard encoder. From (6), we can find that minimizing D' is equivalent to minimizing σ'^2. Finally, the problem is simplified as

$$\min_{\theta} \ \sigma'^2 \ \ s.t. \ R' \leq R_T \tag{8}$$

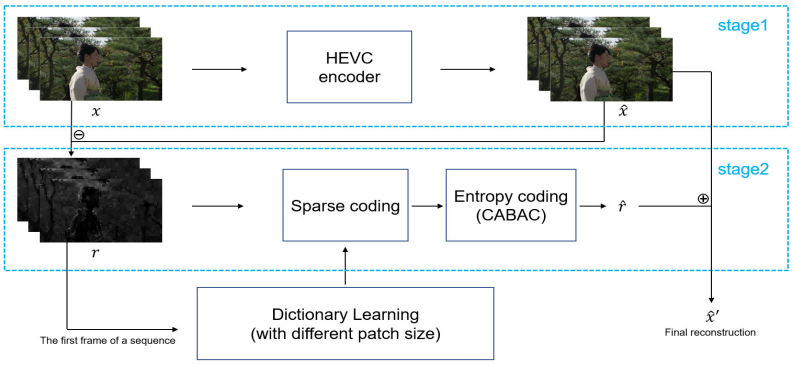


Fig. 1. The framework of proposed dictionary learning based two-stage video compress scheme

Since the input data of the second stage encoder r in (1) is the different between the original frame and reconstructed frame of the first stage, which is the quantization error. Under video coding scenario, quantization error is not fully decorrelated and contains structural information of original frame as shown in Fig.2, however, the difference frame r contains mainly high frequency noise. Therefore, traditional predictive coding and transform coding fail to compress the difference frame. To reduce the variance of the r and leading to small σ'^2, the dictionary learning based sparse representation method is an efficient approach, which is widely used in image demonising. The sparse representation not only removes high frequency noise in r, but also efficiently protects the remained structure information as patterns in learned dictionary. Finally, leading to small σ'^2 value and better overall R-D performance.

Moreover, to reduce the extra bits consumption, specifically designed CABAC encoder according to the statistic characteristic of sparse represented components is designed. The detailed algorithm design is introduced in the following section.

III. The Proposed Framework And Entropy Encoding Strategy

A. Dictionary Learning Based Two-stage Video Compression

Our proposed two-stage framework shown in Fig.1 consists of two main stages. The first stage is responsible for initially compressing the video to obtain a compressed lossy video bitstream. To ensure that the compressed video retains more details, it is necessary to restore the information lost in the first stage encoding. Therefore, in the second stage, we compress the different between the original video and the reconstructed video of first stage to restore high-quality near-lossless video. As is discussed in the first section, we adopt dictionary learning based sparse representation to compress the difference frame as complementary information to enhance reconstruction quality.

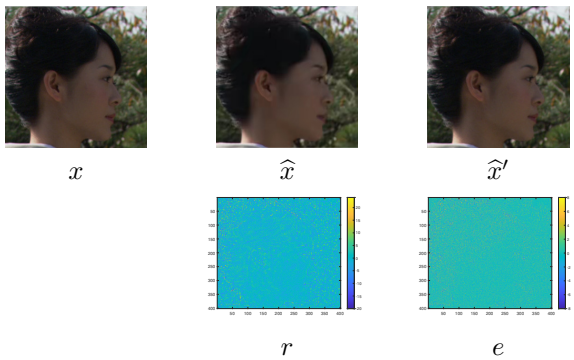


Fig. 2. Illustration of a segment of frames and residuals: x (original), $\widehat{x}$ (reconstructed), their residual r, and $\widehat{x}'$ (final reconstructed), and the final reconstruction error e.

In compressing the difference frame in the second stage, we use an over-complete dictionary, pre-trained with a large amount of data using the K-SVD algorithm, for sparse representation. The difference frame is divided into patches, the n_{th} patch $\mathbf{r_n}$ is decompose into a dictionary matrix Φ and a sparse coefficient matrix $\mathbf{s_n}$. The dictionary Φ stores the features of the original samples $\mathbf{r_n}$, and the sparse coefficient matrix s_n can represent r_n using fewer non-zero coefficient through the dictionary matrix Φ, as shown in the following equation.

$$\mathbf{r_n} = \Phi \cdot \mathbf{s_n} \tag{9}$$

where each column of $\mathbf{r_n}$ represents a $p \times p$ image patch, and each column of Φ represents a dictionary atom. We can represent an image block $\mathbf{r_{n,i}}$ in $\mathbf{r_n}$ using the sparse coefficients $\mathbf{s_{n,i}}$ corresponding to a column in $\mathbf{s_n}$ and the dictionary Φ, i.e., $\mathbf{r_{n,i}} = \Phi \cdot \mathbf{s_{n,i}}$.

To solve for Φ and $\mathbf{s_n}$ in order to represent the residual matrix R, an optimization problem needs to be addressed, as shown in the following equation:

$$\min_{\mathbf{\Phi},\mathbf{s_n}} ||\mathbf{r_n} - \mathbf{\Phi} \cdot \mathbf{s_n}||_F^2 \quad s.t.\ \forall i, ||\mathbf{s_{n,i}}||_0 \leq T_0 \tag{10}$$

where T_0 represents the number of nonzero elements. The notation $||\cdot||_F$ stands for the Frobenius norm and $||\cdot||_0$ stands for the zero norm. Using K-SVD, we first compute the sparse coefficient matrix $\mathbf{s_n}$ through sparse representation. Then, during the dictionary learning process, we sequentially update each atom of the dictionary $\mathbf{\Phi}$. This involves K iterations where each iteration requires performing singular value decomposition(SVD) to update the dictionary.

Furthermore, in our proposed DL-based two-stage framework, sparse representation is essential both during the learning of the K-SVD dictionary and when using the learned over-complete dictionary to represent image samples. This involves

computing a sparse solution s_i under the residual r and dictionary Φ, as shown in the following equation:

$$\min_{\mathbf{s_{n,i}}} ||\mathbf{s_{n,i}}||_0 \quad s.t. \ ||\mathbf{r_{n,i}} - \mathbf{\Phi} \cdot \mathbf{s_{n,i}}||_2^2 \leq \epsilon \tag{11}$$

ϵ represents a small allowable error. We use orthogonal matching pursuit (OMP) [11] to select the dictionary's basis vector that best matches the current difference patch $\mathbf{r_{n,i}}$ at each iteration. It updates the residual until the predetermined sparsity condition is met. The resulting sparse matrix is then entropy coded to obtain the a complementary bitstream.

Specifically, different patch sizes adopted in the dictionary learning and sparse representation show varying compression performance under different bit rates. We propose using two different patch sizes to handle video residuals across different bit rate ranges. When processing high bit rate videos using small quantization parameters (QP), the difference frame r to be compressed are relatively sparse. In this case, we use a small patch size of 8×8 for sparse encoding. Conversely, when handling low bit rate videos with large QP, where difference frame r contains more structural information, as shown in Figure 2, we use a larger patch size of 16×16.

To be noticed, the proposed scheme is different from traditional two-stage encoding methods that adjust compression ratio only by QP. In contrast, the method proposed in this paper allows for dynamic adjustment of multiple parameters including QP in the first stage compression, sparsity level and patch size during sparse representation based compression in the second stage. This multi-parameter controlled the compression scheme offers greater scalability compared to traditional methods, leading to better R-D performance. Through this framework, we can not only fully utilize the original encoder's algorithm for lossy compression but also supplement the lost information to flexibly enhance the video reconstruction quality and achieve a near-lossless performance.

B. Entropy Coding Designed for Sparse Represented Data

In this work, we modify the CABAC for compression and set reasonable context index to achieve the maximum compression efficiency. CABAC mainly consists of binarization, context modeling, and binary arithmetic coding. Our modified CABAC firstly iterates through each column of sparse coefficient matrix $\mathbf{s_n}$. For each column, we record the number of non-zero coefficients num, the values of the coefficients val, the sign of the coefficients $sign$, and the run lengths between coefficients run.

These three variables are then binarized in sequence. The number of non-zero coefficients (num) and the sign of non-zero coefficients ($sign$) are encoded using fixed-length coding, which ensures a straightforward and efficient representation of the count of significant values in each column. The coefficient values (val) and the run lengths (run) are encoded using exponential Golomb coding. Exponential Golomb coding is chosen for its efficiency in encoding a wide range of values with fewer bits, especially when dealing with sparse data. This method helps to minimize the bit rate by taking advantage of the distribution patterns in the sparse coefficient matrix, leading to more effective compression.

After binarization, we perform arithmetic coding on the four variables using different contexts. Unlike HEVC's CABAC, which establishes a table for coefficient probability changes and converts them by lookup, we simplify this step. We use traditional binary arithmetic coding, where the context probability is dynamically updated based on the previous symbol. However, for each variable to be encoded, we set different context models and mechanisms for selecting these models. This allows us to optimally encode according to the corresponding probability distribution of each variable. For example, for val, we use the preceding symbol to select different context models to achieve the optimal compression rate.

By utilizing context models specifically designed to each variable, we can more accurately predict the probability distribution of the symbols, thereby enhancing the efficiency of the arithmetic coding process. For num, a context model based on the distribution of previous counts can be employed. For val, the context is chosen based on the preceding coefficient, allowing the model to adapt to local variations in the data.

This context-adaptive approach enables the CABAC to better exploit the statistical properties of the sparse coefficient matrix, leading to a higher compression ratio and more efficient encoding.

IV. IMPLEMENTATION AND EXPERIMENTAL RESULTS

A. Implementation

The method proposed in this paper consists of two stages. The first stage involves a basic encoder using HM16.20 for initial lossy compression of the video. In the second stage, difference frames are extracted between the first frames of the reconstructed and original videos. These frames undergo dictionary learning for feature extraction, and following frames are sparsely represented based on the shared dictionary. The resulting sparse coefficient matrix is entropy encoded to generate the bitstream for the second part. The experiments use the official HEVC datasets, with Classes A through E representing different resolutions. At the decoder, residuals are reconstructed into near-lossless form using the sparse coefficient matrix and the dictionary.

B. Experimental Results

To demonstrate the high coding efficiency of the proposed dictionary learning based two-stage compression method in near-lossless video coding in high bitrate range, we use small QP values of 1, 3, 5 and 7 in the first stage encoding. For the second stage coding, we employ a fixed dictionary learning sparsity L of 6, a patch size $p \times p$ of 8×8. Standard test sequence of different resolution and characteristics are employed encoded under Low-Delay Main configuration. The standard HEVC test model HM16.20 with same coding configuration is used as the comparative method. The Bjontegaard Delta Bitrate (BD-BR) [10] is used to measure the R-D performance

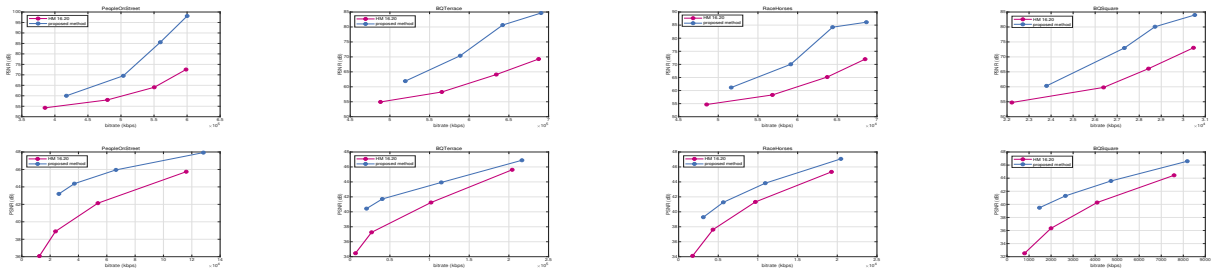


Fig. 3. TOP: High-bitrate R-D curves. Bottom: Low-bitrate R-D curves. From left to right: PeopleOnStreet (Class A), BQTerrace (Class B), RaceHorses (Class C), BQSquare (Class D).

TABLE I
BD-BR FOR HIGH-BITRATE CODING

Class	Name	QP	Bitrate(kbps)	PSNR(dB)	Bitrate(kbps)	PSNR(dB)	BD-BR
ClassA	Traffic	1	577236.76	76.04	668166.68	93.89	-14.20%
		3	535838.38	66.12	541805.17	93.50	
		5	483697.06	59.42	503812.08	73.68	
		7	372053.75	54.21	404999.16	60.34	
	PeopleOnStreet	1	598408.27	72.54	600148.11	98.11	-15.70%
		3	550358.25	64.06	559510.79	85.57	
		5	479540.71	58.03	503510.03	69.54	
		7	385080.43	54.26	417423.74	60.00	
ClassB	Cactus	1	649417.20	81.45	650635.68	95.68	-19.01%
		3	606358.76	69.81	609813.81	93.64	
		5	564251.79	60.07	580899.03	76.26	
		7	490673.41	55.20	517644.89	62.48	
	BQTerrace	1	687526.79	69.28	690565.55	84.66	-32.38%
		3	634135.23	64.10	642196.10	80.64	
		5	565464.34	58.27	588692.95	70.39	
		7	488142.68	54.94	519441.66	61.90	
ClassC	BQMall	1	130057.35	78.13	130143.79	88.72	-16.30%
		3	120625.53	67.56	121541.88	87.06	
		5	112504.52	60.02	116095.21	75.99	
		7	94904.86	54.92	101148.16	61.32	
	Racehorse	1	68508.28	71.99	68670.81	86.11	-25.95%
		3	63731.26	65.16	64413.56	84.20	
		5	56823.99	58.33	59134.56	70.03	
		7	48533.16	54.67	51630.90	61.16	
ClassD	BQSquare	1	30462.74	73.04	30521.32	83.51	-14.75%
		3	28421.90	66.07	28716.51	80.08	
		5	26381.74	59.80	27325.90	72.97	
		7	22218.88	54.75	23793.62	60.31	
	BlowingBublles	1	28431.05	76.06	28456.62	82.66	-9.70%
		3	26403.46	66.62	26625.27	80.42	
		5	24463.63	59.44	25293.60	73.30	
		7	21062.52	54.84	22371.51	60.56	
ClassE	Fourpeople	1	233402.96	74.89	233873.51	89.89	-31.33%
		3	217671.19	66.33	220293.33	88.22	
		5	192975.68	59.47	201753.80	72.45	
		7	135327.89	53.94	150377.98	59.31	
	Johnny	1	228459.63	74.70	190761.44	88.35	-43.60%
		3	213700.47	66.49	180062.38	86.49	
		5	189210.13	59.55	164890.23	71.54	
		7	130501.36	53.97	121242.64	59.47	
Average		1	323191.10	74.81	333009.58	89.16	-22.29%
		3	299724.44	66.23	303099.13	85.98	
		5	269531.36	59.24	280438.54	72.62	
		7	218849.89	54.57	235432.28	60.69	

TABLE II
BD-BR FOR LOW-BITRATE CODING

Class	Name	QP	Bitrate(kbps)	PSNR(dB)	Bitrate(kbps)	PSNR(dB)	BD-BR
ClassA	Traffic	17	70426.46	44.93	82743.45	46.76	-33.00%
		22	22852.28	41.73	35770.09	44.35	
		27	7813.48	38.95	21205.44	42.58	
		32	3166.67	36.39	16968.12	41.38	
	PeopleOnStreet	17	116064.08	45.74	128182.72	47.93	-56.40%
		22	53705.14	42.14	66462.02	45.95	
		27	23783.96	38.9	37111.35	44.37	
		32	12409.74	36.07	26159.39	43.20	
ClassB	Cactus	17	171672.82	44.45	182143.01	45.95	-59.10%
		22	45857.47	39.81	56948.36	42.67	
		27	10022.62	37.36	21482.60	40.99	
		32	4322.14	35.33	16065.08	40.01	
	BQTerrace	17	204200.25	45.62	216479.38	46.89	-61.81%
		22	101856.72	41.25	114826.91	43.93	
		27	26916.82	37.26	40449.17	41.72	
		32	6668.23	34.46	20574.82	40.43	
ClassC	BQMall	17	22682.86	44.32	25083.17	46.45	-36.43%
		22	7590.78	40.85	10093.14	44.04	
		27	3547.32	37.94	6147.50	42.26	
		32	1712.65	35.01	4405.24	40.75	
	Racehorse	17	19270.23	45.34	20453.38	47.08	-46.10%
		22	9677.00	41.32	10926.00	43.83	
		27	4330.79	37.62	5637.29	41.27	
		32	1763.41	34.11	3113.69	39.28	
ClassD	BQSquare	17	7581.41	44.44	8181.15	46.59	-68.10%
		22	4089.19	40.28	4716.88	43.57	
		27	1997.11	36.34	2650.36	41.29	
		32	796.38	32.52	1471.22	39.48	
	BlowingBublles	17	7091.50	44.17	7465.26	45.92	-42.35%
		22	3136.84	39.75	3613.39	42.56	
		27	1424.34	36.14	1981.21	39.89	
		32	646.98	33.03	1223.16	37.99	
ClassE	Fourpeople	17	17977.44	45.83	23432.35	49.25	-79.17%
		22	3654.89	43.25	9297.70	48.06	
		27	1228.06	40.93	7036.98	47.19	
		32	602.90	38.40	6574.01	46.46	
	Johnny	17	18119.69	46.22	23540.17	49.78	-32.70%
		22	4352.52	43.81	9932.94	48.67	
		27	1015.57	41.70	6704.56	47.84	
		32	389.91	39.64	6169.47	47.20	
Average		17	65508.67	45.11	71770.40	47.26	-51.52%
		22	25677.28	41.42	32258.74	44.76	
		27	8208.01	38.31	15040.65	42.94	
		32	3247.90	35.50	10272.42	41.62	

difference. Experimental results in Table 1 demonstrates that for near-lossless coding scenario in high bitrate range, our proposed scheme outperforms traditional HEVC encoder by -22.29% BD-BR on average. It is interesting to notice that in near-lossless video coding, the proposed method maintains similar bitrate while achieves significant quality improvement, for example, PSNR increase of 16.54 dB is achieved with only 1.27% bitrate increase for sequence BQTerrace encoded using QP of 3. The small increase of bitrate is due to that with small QP, the difference frame from the first stage is sparse. With small extra bits as complementary information, the lost details are well restored and the final reconstruction quality is significantly improved.

The proposed scheme also shows better R-D performance than traditional encoder in low bitrate coding scenario, and maintains near-lossless performance. The evaluation is conducted with a fixed dictionary learning sparsity L of 6, a patch size a patch size $p \times p$ of 16×16. The Low-Delay P Main configuration is used with QP values of 17, 22, 27, and 32 that follows the common test conditions. The standard HEVC test model HM16.20 with same coding configuration is used as the comparative method. The BD-BR is used to measure the R-D performance difference. As shown in Table 2, the proposed dictionary learning based two-stage compression method significantly outperforms standard HEVC lossy encoding in the low bit rate range that an average BD-BR of -51.52% is achieved. In video coding at low bitrate range, the difference frame generated from the first stage contains abundant noise and also structural information, which is sparse represented and lossy compressed in the second stage controlled by the sparsity level. Since the lost information complementary and quality restore in the second stage, the PSNR is always higher than that of standard encoder. It hence more reasonable to compare the bitrate reduction between the proposed scheme with higher QP and standard encoder with lower QP. For example, when encoding sequence Cactus, the proposed scheme using QP of 27 achieves PSNR of 40.99 dB, which is of 1.18dB higher than the standard encoder using QP of 22. However, the bitrate

TABLE III
R-D PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND ELC [7]

	BD-BR (%)	
	ELC [7]	Our method
ClassA	-10.4%	-17.0%
ClassB	- 7.8%	-28.1%
ClassC	-15.8%	-24.1%
ClassD	-21.2%	-14.2%
ClassE	- 7.4%	-32.5%
Average	-12.5%	-23.2%

reduction is 59.10%. This demonstrates the proposed schemes always works in common bitrate range. Moreover, we can find from Table 2 that the average PSNR values using QP from 22, 27, 32 and 37 are all larger than 41 dB. Commonly, a PSNR measure of 40 dB, or above, typically constitutes visually lossless coding, considered as near-lossless [12]. We also depict the R-D curves in Fig. 3 to clearly shows the R-D performance at different bitrates range. It can be observed that our proposed method achieves significant improvements in both high and low bitrate video coding scenarios.

Moreover, comparison is made between our proposed method and a multi-layer coding framework with enhancement layer (ELC) proposed in [7]. The experimental results in Table 3 demonstrate that with the same coding parameters, our proposed scheme outperforms the ELC approach by -10.6% BD-BR on average. To be noticed, our proposed method shows better R-D performance in high resolution test sequences including Class A, B, C and E, indicating that the dictionary learning based compression is more efficient for high resolution videos. While for Class D of low resolution sequences, the comprasion method ELC shows better R-D performance.

V. CONCLUSIONS

In this work, we propose a dictionary learning based two-stage video encoding framework to compress and leverage the reconstruction errors as complementary information to enhance final reconstruction quality. CABAC is modified to be adaptive to the sparse represented reconstructed error, leading to high coding efficiency. Experimental results demonstrate that the proposed method achieves over 20% BD-BR reduction dealing with near-lossless video coding at high bitrate range. Additionally, the proposed framework is efficient under normal bitrate range, leading to over 50% BD-BR reduction and retaining near-lossless coding with PSNR values above 41 dB.

The proposed method can be implemented in video coding systems beyond HEVC. During training stage of dictionary learning, apart from adjusting the QP in the first lossy compression stage with standard encoder, we can also fine-tune training parameters such as sparsity L and patch size p during the second encoding stage with sparse coding tools to find the optimal coding parameter set. Moreover, rate control employing bit allocation algorithm between two encoding stages can further improve the overall R-D performance. In our future work, well designed inter-stage rate control method and optimal coding parameters determination will yield even greater benefits.

REFERENCES

[1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS," *IEEE Transactions on Image processing*, vol. 9, no. 8, pp. 1309–1324, 2000.

[3] A. Antony and S. Ganapathy, "Highly efficient near lossless video compression using selective intra prediction for HEVC lossless mode," *AEU-International Journal of Electronics and Communications*, vol. 69, no. 11, pp. 1650–1658, 2015.

[4] N. Nazari, R. Shams, M. Mohrekesh, and S. Samavi, "Near-lossless compression for high frame rate videos," in *2013 21st Iranian Conference on Electrical Engineering (ICEE)*, IEEE, 2013, pp. 1–6.

[5] X. Cai and Q. Gu, "Improved HEVC lossless compression using two-stage coding with sub-frame level optimal quantization values," in *2014 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2014, pp. 5651–5655.

[6] S. Takamura and Y. Yashima, "Lossless scalable video coding with h. 264 compliant base layer," in *IEEE International Conference on Image Processing 2005*, IEEE, vol. 2, 2005, pp. II–754.

[7] A. Heindel, E. Wige, and A. Kaup, "Low-complexity enhancement layer compression for scalable lossless video coding based on HEVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1749–1760, 2016.

[8] A. Heindel and A. Kaup, "Scalable near-lossless video compression based on HEVC," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, IEEE, 2017, pp. 1–4.

[9] Y. Bai, X. Liu, W. Zuo, Y. Wang, and X. Ji, "Learning scalable ly=-constrained near-lossless image compression via joint lossy image and residual compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 946–11 955.

[10] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *ITU SG16 Doc. VCEG-M33*, 2001.

[11] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[12] E. Palma and I. Tabus, "Near-lossless coding of plenoptic camera sensor images for archiving light field array of views," in *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, IEEE, 2022, pp. 1–6.

EEND-EM: End-to-End Neural Speaker Diarization with EM-Network

Beom Jun Woo*, Ji Won Yoon†, Min Hyun Han*, Chan Yeong Moon*, Nam Soo Kim*
* Department of Electrical and Computer Engineering and INMC, Seoul National University, Seoul, Korea
E-mail: {bjwoo, mhhan, cymoon}@hi.snu.ac.kr, nkim@snu.ac.kr Tel/Fax: +82 2-884-1824
† Department of Artificial Intelligence, Chung-Ang University, Seoul, Korea
E-mail: jiwonyoon@cau.ac.kr Tel/Fax: +82 2-820-5549

***Abstract*—In recent years, various studies have been conducted to further enhance end-to-end neural speaker diarization (EEND) systems. However most of the methods increase model complexity by requiring additional modules during inference. In this paper, we introduce EEND-EM, a novel end-to-end neural speaker diarization model that integrates an EM algorithm-aware self-distillation method into the EEND framework. Our approach aims to enhance diarization performance by utilizing oracle guidance features derived from the EM algorithm, improving the model's prediction accuracy without increasing model complexity. Through experiments on the LibriMix dataset, EEND-EM demonstrated significant improvements in diarization error rates (DER), particularly in minimizing missed speech (MS) and confusion (CF) metrics, when compared to the baseline EEND-EDA model. Furthermore, attention weight visualizations indicate that the transformer encoders in EEND-EM are trained more effectively. The code is available at https://github.com/CARNIVAL-IITP/Speaker-Identification/tree/main/EEND-EM**

I. INTRODUCTION

Speaker diarization aims to assign segments of audio to distinct speakers, effectively aligning the audio data with the corresponding speaker identities to determine 'who spoke when' in a multi-speaker environment. Recent development of technology has increased the demand of speaker diarization in multiple fields, such as speech recognition [1] and speaker verification [2].

Classical cascaded methods approach speaker diarization by first detecting speaker-active frames and then clustering them using speaker embeddings. The number of clusters, which corresponds to the number of speakers, is determined during inference through mathematical techniques like eigenvalue analysis [3] or hierarchical clustering with a preset threshold. Advanced methods such as utilizing improved speaker embeddings [4] or better embedding feature extractor [5] led to increase of clustering-based diarization performance. However, these methods struggle with handling speaker overlaps, as each speech frame is usually assigned to only one speaker. To overcome the limitations, the end-to-end neural speaker diarization (EEND) method has been researched to consider speaker diarization as multi-label classification problem [6], [7] . However both [6] and [7] have limitation of processing situations where the number of speaker is flexible. To address this issue, Horiguchi et al. [8] introduced the EEND-EDA system, employing a sequence-to-sequence approach with an LSTM encoder-decoder network to derive speaker-wise attractors from frame-wise embeddings and estimate speaker existence probabilities from the attractors. This approach helps speaker diarization system to handle varying numbers of speakers.

Research efforts have been made to further enhance the performance of the EEND-EDA system. One approach utilizes speaker-specific prior information, such as voice activity patterns or speaker embeddings [9], [10]. Target Speaker Voice Activity Detection (TS-VAD) improves performance by handling overlapping speech and detecting the speech activities of a set of target speakers [11]. Another method integrates tasks such as speech separation or speech counting. [12], [13] The multitask learning approach suggests that training models together, which are related to the diarization task, can have a complementary effect. However, in terms of model complexity during inference, these methods require additional modules to achieve performance improvements.

In this paper, we adopt EM-Network [14] which proposes a novel self-distillation framework that leverages target information for supervised sequence-to-sequence learning tasks. This framework is designed to improve the prediction accuracy of sequence models by incorporating oracle guidance derived from the target sequence, creating a more accurate latent space. We integrate EM-Network to speaker diarization tasks and show that proposed model EEND-EM can achieve higher performance without additional modules.

II. BACKGROUND

A. EEND-EDA

Our proposed method utilizes the original EEND-EDA [8] where its speaker wise encoder-decoder attractor enables the system to deal with varying number of speakers. EEND-EDA architecture is composed of SA-EEND [7] and LSTM based encoder-decoder. First, the input speech is transformed into Mel-filterbanks. Then the acoustic feature is processed through 2 linear layers and 4 transformer encoder blocks to get frame-wise embeddings. Secondly the embedding goes into LSTM based encoder-decoder which calculates speaker-wise attractor existence probability and produces corresponding attractors. Finally, diarization result is calculated as dot products between the attractors and the frame-wise embeddings. EEND-EDA is

trained with 2 types of losses; diarization loss and attractor existence loss.

Given diarization result p_t and groundtruth labels y_t, permutation invariant training is done as follows.

$$\mathcal{L}_{\text{pit}} = \min_{s} \sum_{t=1}^{T} \mathcal{H}(p_{s,t}, y_t) \tag{1}$$

where t and s is frame and speaker set respectively. $\mathcal{H}$ refers to binary cross entropy (BCE) which is defined as follows.

$$\mathcal{H}(y_t, p_t) = \sum_{s=1}^{S} \left\{ -y_{s,t} \log p_{s,t} - (1 - y_{s,t}) \log (1 - p_{s,t}) \right\}. \tag{2}$$

Attractor existence loss is calculated with the binary cross entropy (BCE) loss between attractor existence loss probability q and binary label vector l, where S stands for the number of speakers.

$$\mathcal{L}_{\text{attractor}} = \frac{1}{S+1} \mathcal{H}(q, l) \tag{3}$$

B. EM-Network

EM-Network [14] proposes self-distillation method based on mathematical theory to maximize model performance. For current estimates of parameter $\theta^{(t)}$, Q-function calculates the expected value of log-likelihood with respect to the current estimates of the latent variables' z distribution.

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{z|x,\theta^{(t)}}[\log P(x, z|\theta)], \tag{4}$$

where x is input data, and θ is the parameters of the model. After the above E-step, M-step updates θ to maximize the Q-function.

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}). \tag{5}$$

Through iteration of above EM steps, parameters that maximize the likelihood of the input is determined.

Equation (6) from EM-Network indicates that the Q-function's minimization is equivalent to minimizing the KL-divergence between the distributions of the EM-Network and the sequence model, thereby making it analogous to the maximization step of the traditional EM algorithm.

$$\begin{aligned} Q(\theta|\phi^{(t)}) &= -D_{KL}\left(P(z|x, y; \phi^{(t)}) \parallel P(z|x; \theta)\right) \\ &\approx -\mathcal{L}_{\text{kd}}(\phi^{(t)}, \theta). \end{aligned} \tag{6}$$

As a result, EM algorithm can be directly applied to sequence to sequence model without any iteration steps, where EM algorithm originally requires. EM-Network proved that addition of distillation loss between teacher model's logit and student model's logit leads to direct optimization of original sequence model, which are speech recognition task and machine translation task. As mentioned in original paper, we also adopt L2 loss instead of KL-divergence loss for training stability.

III. PROPOSED METHOD

Our proposed method starts from the original EEND-EDA [8] framework. The proposed method integrates EM-Network [14] to the speaker diarization task. Firstly, we introduce oracle teacher model that generates oracle guidance feature that combines to original speaker diarization model. Secondly, we explain self distillation-like training scheme so that the original model can learn from oracle guidance.

A. EEND-EM

The most challenging part of designing the EM-Network is to avoid trivial solutions. Oracle decoder in the EM-Network copies the target directly instead of utilizing the output embedding from the last transformer encoder. To prevent the trivial solution problem, we analyze label aggregator which corresponds to the post-processing of speaker diarization model.

Diarization result p_t and ground-truth labels y_t are aggregated before calculating loss. Label aggregation divides labeled sequences into overlapping frames, aggregates the labels within each frame based on a majority rule. For this reason, we assumed that giving target speaker label into oracle encoder can be applied same as the original EM-Network.

From this point forward, we refer to the EM-Network as the oracle teacher model and the EEND-EDA part as the student model in the EM aware self-distillation framework. EM-Network is consisted of 3 parts; oracle encoder, sequence model and oracle decoder. Oracle encoder generates oracle guidance feature using speech labels. Unlike original EM-Network oracle encoder structure, EEND-EM oralcle encoder consists of a linear layer, max pooling and self-attention based transformer encoder layer. Oracle guidance feature is used as key K and value V for input of oracle decoder which is explained below.

$$K, V = OracleEncoder(y_t) \tag{7}$$

For sequence model, we adopt original EEND-EDA framework to our work. After the frame-wise embeddings are generated through stacked transformer blocks, oracle decoder is inserted. Oracle decoder merges the outputs of the sequence model and those of the oracle encoder. The decoder is a transformer decoder that utilizes cross-attention to effectively guide the source sequence information. The cross-attention layer receives the output embeddings of the final transformer encoder as queries and the oracle guidance features as key-value.

$$Q = SA\text{-}EEND(x_t) \tag{8}$$

$$e_{s,t}^{\phi} = OracleDecoder(Q, K, V) \tag{9}$$

Output $e_{s,t}^{\phi}$ is used to calculate the attractor existence probability and diarization result of EM-Network.

After forward propagation of EM-Network, EEND-EDA is forwarded using same input x_t. EEND-EDA shares its parameter with EM-Network. For each epoch, both model is forwarded to predict each diarization result and attractor existence probability.

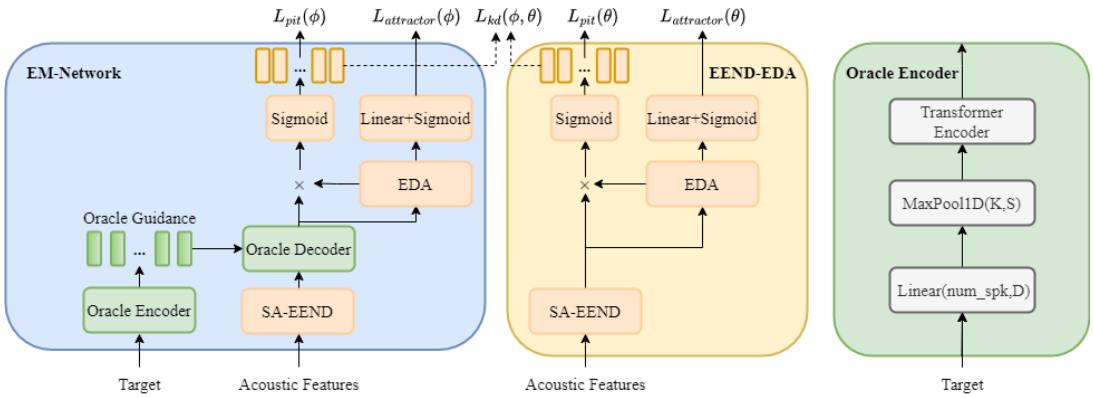


Fig. 1: Model architecture of EEND-EM. Left blue box is architecture of EM-Network. Middle yellow box is overview of EEND-EDA. Green box shows how oracle encoder is composed of. D stands for embedding number. K and S in MaxPool1D block stands for kernel and stride size respectively. All EEND-EDA modules (orange boxes) in EEND-EM share parameters.

$$p^{\phi}_{s,t}, q^{\phi}_{s,t} = EM\text{-}Network(x_t) \tag{10}$$
$$p^{\theta}_{s,t}, q^{\theta}_{s,t} = EEND\text{-}EDA(x_t) \tag{11}$$

where parameter has relationship of $\theta \in \phi$. During inference and validation, we exclusively utilize EEND-EDA model. EM-Network in EEND-EM is only used while training.

B. EM algorithm aware training

Let EM-Network's parameter and original speaker diarization model's parameter be ϕ and θ respectively. There are 3 types of losses for training the proposed model. Two out of three types adopt original PIT loss equation (1) and attractor loss equation (3) for both diarization model. PIT loss for each model is $\mathcal{L}_{\text{pit}}(\phi)$ and $\mathcal{L}_{\text{pit}}(\theta)$. Also attractor existence loss is expressed as $\mathcal{L}_{\text{attractor}}(\phi)$ and $\mathcal{L}_{\text{attractor}}(\theta)$ respectively. Total PIT loss and attractor existence loss is as followed.

$$\mathcal{L}_{\text{pit}} = \mathcal{L}_{\text{pit}}(\phi) + \mathcal{L}_{\text{pit}}(\theta) \tag{12}$$
$$\mathcal{L}_{\text{attractor}} = \mathcal{L}_{\text{attractor}}(\phi) + \mathcal{L}_{\text{attractor}}(\theta) \tag{13}$$

To leverage oracle guidance features as soft labels for training the original sequence model, we introduce an additional distillation loss. This loss is computed using the L_2 distance between the diarization outputs of the EM-Network and the original sequence model. Since the EM-Network is trained with ground truth labels, it is expected to have learned more accurate parameters. By using the distillation loss, we aim to transfer this learned knowledge from the oracle-guided EM-Network to the original model, thereby enhancing its diarization performance

$$\mathcal{L}_{\text{kd}}(\phi,\theta) = \sum_t (p^{\phi}_t - p^{\theta}_t)^2 \tag{14}$$

Total loss will be as follows. λ is scheduled distillation loss variable to warmup the training of each model parameter ϕ and θ.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pit}} + \mathcal{L}_{\text{attractor}} + \lambda * \mathcal{L}_{\text{kd}}(\phi,\theta) \tag{15}$$

IV. Experiments and Result

A. Training and test datasets

We tested the performance of EEND-EM with the LibriMix dataset [15], which contains training and test mixtures from LibriSpeech [16] train-clean100 and test-clean samples mixed with WHAM! [17] at a 16kHz sampling rate. The dataset includes two-speaker (Libri2Mix) and three-speaker (Libri3Mix) mixtures, consisting of 58 hours/11 hours and 40 hours/11 hours of training/test sets, respectively. We used the min mode during experiment to benchmark our results against previous studies.

B. Configurations

We used baseline model EEND-EDA from ESPnet [18] [1]. Acoustic feature is generated through 80-dimensional log-mel filterbanks with window size of 25ms and frame shift of 10ms. SA-EEND from Fig. 1 is consisted of 4 transformers with embedding size of 256. EDA module is 1 recurrent neural network with hidden unit size of 256. Oracle Encoder has 3 modules. Linear layer maps number of speakers to embedding size 256. Then transformer encoder is followed by maxpooling layer with kernel size of K and stride size of S. Oracle decoder is a transformer decoder with embedding size of 256. We optimize via Adam with batch size of 64. Warmup learning rate scheduler is applied where warmup step is 30000 and learning rate is 2e-3. The scheduled distillation loss variable λ is set to increase logarithmically from 1e-8 to 1 across epochs. Total epoch is set as 250. Training is done with one NVIDIA GeForce RTX 3090 GPU.

C. Evaluation metric

We assessed speaker diarization performance by measuring the diarization error rate (DER (%)), which includes speaker confusion (SC (%)), false alarms (FA (%)), and missed detections (MS (%)). The collar tolerance is set to 0 seconds.

TABLE I: Experimental results on a fixed 2-speaker scenario (Libri2Mix) for min mode in terms of DER, FA, and MI (%). † denotes our re-implementation.

Method	DER	FA	MS	CF
SA-EEND†[7]	6.13	3.54	2.11	0.48
EEND-EDA†[8]	5.93	3.48	2.07	0.38
EEND-EM$_{512,128}$	5.61	3.47	1.83	0.31
EEND-EM$_{128,64}$	5.23	3.67	1.34	**0.22**
EEND-EM$_{64,32}$	5.15	3.57	1.35	0.23
EEND-EM$_{32,16}$	**4.98**	**3.45**	**1.29**	0.24

TABLE II: Experimental results on a fixed 3-speaker scenario (Libri3Mix) for min mode.

Method	DER	FA	MS	CF
SA-EEND† [7]	9.05	5.92	2.72	0.41
EEND-EDA† [8]	8.81	5.81	2.62	0.38
EEND-EM$_{128,64}$	7.14	5.21	1.65	0.28
EEND-EM$_{64,32}$	6.93	5.15	1.51	0.27
EEND-EM$_{32,16}$	**6.82**	**5.14**	**1.44**	**0.24**

D. Results on the LibriMix dataset

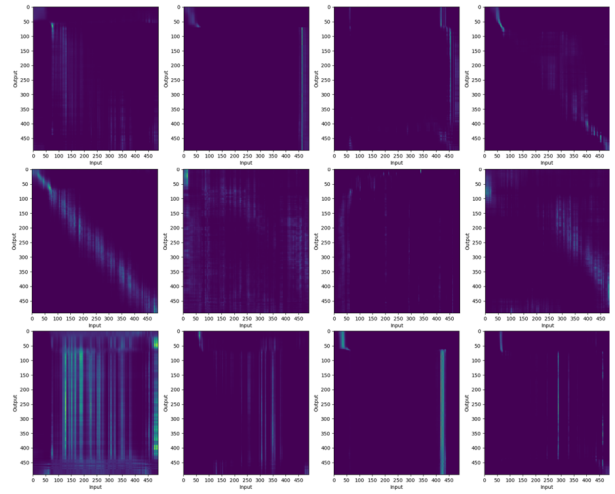
We performed experiments on scenarios with a set number of speakers. The proposed models were assessed under conditions involving 2 speakers and 3 speakers, utilizing the test sets min mode from Libri2Mix and Libri3Mix, respectively. We trained each SA-EEND [7], EEND-EDA [8] and EEND-EM based on ESPNet [18]. From Table I and II, we can see that application of EM-Network for speaker diarization proves to be effective. From the result, we analyze that the more the kernel and stride sizes differ from the original label aggregating settings, the better the model's performance. We can see that relative performance improvement is 19.07% and 29.17% for 2-speaker and 3-speaker scenarios, respectively.

Table I includes experiment of EEND-EM where kernel and stride size is equivalent to label aggregator setting, which is 512 and 128 each. The results indicate that the proposed model yields similar outcomes to the original model, suggesting that the EM-Network produces trivial solutions. As the authors assumed, setting the kernel and stride similar to the label aggregator is ineffective for generating oracle guidance features.

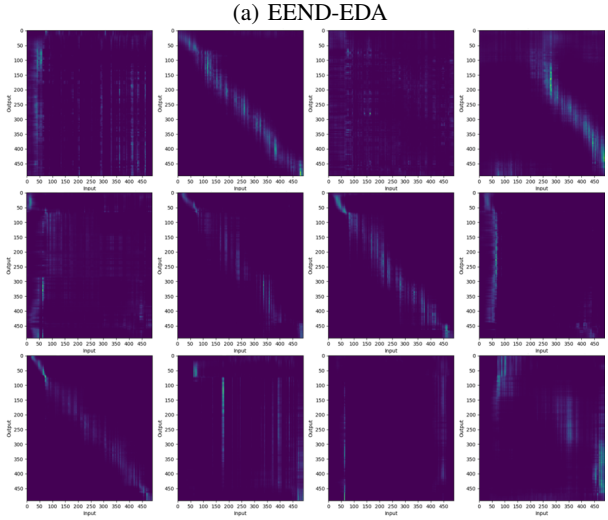
E. Visualization of self-attention heads

To further investigate the effect of EM algorithm-aware training, we plot attention weights for both the baseline and proposed models. Figure 2 shows the first three transformer encoder weights drawn from a Libri2Mix test dataset. The attention weights in the EEND-EDA exhibit varied distribution patterns. Some plots display sparse attention concentrated in specific regions, while others have scattered or diagonal patterns indicating sequential attention. In contrast, the EEND-EM$_{32,16}$ shows more coherent attention patterns, with the diagonal self-attention weights being more pronounced and consistent across different plots. From the results, we observe that improved alignment in attention weights correlates with better speaker-related outcomes.

[1]https://github.com/espnet/espnet/tree/master/egs2/librimix/diar1



(a) EEND-EDA



(b) EEND-EM$_{32,16}$

Fig. 2: Visualization of attention weight matrices in the first three encoder block. Each line corresponds to each encoder block with four heads.

V. Conclusions

In this paper, we propose EEND-EM, which integrates an EM algorithm-aware self-distillation method into the EEND framework. The results demonstrate the effectiveness of this approach, particularly in improving the MS and CF metrics. Additionally, the attention plots indicate that the transformer encoder in the proposed model is better trained compared to the baseline EEND-EDA. Finally, we can explore better structures for generating oracle encoder and decoder features to provide improved oracle guidance in the future.

VI. Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00456, Development of Ultra-high Speech Quality Technology for Remote Multi-speaker Conference System)

REFERENCES

[1] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5554–5558. DOI: 10.1109/ICASSP.2018.8462661.

[2] C. Xu, W. Rao, J. Wu, and H. Li, "Target speaker verification with selective auditory attention for single and multi-talker speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2696–2709, 2021. DOI: 10.1109/TASLP.2021.3100682.

[3] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2020. DOI: 10.1109/LSP.2019.2961071.

[4] J.-W. Jung, H.-S. Heo, B.-J. Lee, *et al.*, "In search of strong embedding extractors for speaker diarisation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096449.

[5] H.-S. Heo, Y. Kwon, B.-J. Lee, Y. J. Kim, and J.-W. Jung, "High-resolution embedding extractor for speaker diarisation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10097190.

[6] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Interspeech*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202572807.

[7] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 296–303, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202572979.

[8] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1493–1507, 2022, ISSN: 2329-9304. DOI: 10.1109/taslp.2022.3162080. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2022.3162080.

[9] D. Wang, X. Xiao, N. Kanda, T. Yoshioka, and J. Wu, *Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization*, 2022. arXiv: 2208.13085 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2208.13085.

[10] Z. Du, S. Zhang, S. Zheng, and Z. Yan, *Speaker overlap-aware neural diarization for multi-party meeting analysis*, 2022. arXiv: 2211.10243 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2211.10243.

[11] I. Medennikov, M. Korenevsky, T. Prisyach, *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Interspeech 2020*, ser. interspeech$_2$020, ISCA, Oct. 2020. DOI: 10.21437/interspeech.2020-1602. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1602.

[12] S. Maiti, Y. Ueda, S. Watanabe, *et al.*, *Eend-ss: Joint end-to-end neural speaker diarization and speech separation for flexible number of speakers*, 2022. arXiv: 2203.17068 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2203.17068.

[13] J. Ao, M. S. Yıldırım, R. Tao, *et al.*, *Used: Universal speaker extraction and diarization*, 2024. arXiv: 2309.10674 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2309.10674.

[14] J. Yoon, S. Ahn, H. S. Lee, M. Kim, S. Kim, and N. S. Kim, "Em-network: Oracle guided self-distillation for sequence learning," *ArXiv*, vol. abs/2306.10058, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259203399.

[15] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, *Librimix: An open-source dataset for generalizable speech separation*, 2020. arXiv: 2005.11262 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2005.11262.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[17] G. Wichern, J. Antognini, M. Flynn, *et al.*, *Wham!: Extending speech separation to noisy environments*, 2019. arXiv: 1907.01160 [cs.SD]. [Online]. Available: https://arxiv.org/abs/1907.01160.

[18] S. Watanabe, T. Hori, S. Karita, *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211. DOI: 10.21437/Interspeech.2018-1456. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456.

Multi-Task Learning Approaches for Music Similarity Representation Learning Based on Individual Instrument Sounds

Takehiro Imamura*, Yuka Hashizume* and Tomoki Toda*
* Nagoya University, Japan
imamura.takehiro@g.sp.m.is.nagoya-u.ac.jp, hashizume.yuuka@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

***Abstract*—Music similarity representation learning (MSRL) based on individual instrument sounds (InMSRL) is a potential technique to develop a new function in music recommendation and retrieval systems, allowing users to focus on multiple partial elements of music pieces. There have been proposed two main approaches, *Cascade* that sequentially performs music source separation (MSS) and music similarity feature extraction for each instrument sound and *Direct* that directly extracts disentangled music similarity features. Each approach has a specific problem; in *Cascade*, a separation error often causes adverse effects on the feature extraction; and in *Direct*, it is usually hard to learn accurately disentangled features. In this paper, we propose multi-task learning approaches that leverage MSS to address these issues. For *Cascade*, we propose end-to-end fine-tuning of the MSS model and the music similarity feature extractors using an auxiliary separation loss. For *Direct*, we propose multi-task learning based on the disentangled music similarity feature extraction and MSS based on reconstruction with the disentangled music similarity features. We conduct experimental evaluations and demonstrate that 1) the end-to-end fine-tuning for *Cascade* significantly improves InMSRL performance, 2) the multi-task learning for *Direct* is also helpful to improve disentanglement performance in the feature extraction, and 3) *Cascade* with the end-to-end fine-tuning outperforms *Direct* with the multi-task learning.**

I. Introduction

Music recommendation and retrieval systems are useful technologies under the current situation that the number of music pieces has already exceeded 1 billion[1] and further market expansion is expected[2]. Methods utilizing users' listening histories [1], [2] have been widely used in these systems although these methods cause several limitations, e.g., hard to handle music pieces with less listening records. To address this issue, content-based methods to extract content features from a music piece to capture its characteristics have been studied. Recently methods based on deep learning have attracted attention since they can extract more precise content features [3]–[6] than classical methods [7], [8]. In particular, music similarity representation learning (MSRL) capable of extracting music similarity features in an unsupervised manner without using any hand-crafted labels is a promising way to develop the music recommendation and retrieval systems to widely handle existing music pieces.

MSRL based on individual instrument sounds (InMSRL) [9], [10] is a potential technique to develop a new function allowing users to focus on multiple partial elements of music pieces, e.g., searching for music pieces with similar drum sounds. Hashizume et al. [9] proposed InMSRL model, which inputs clean individual instrument sounds into the music similarity feature extractors (*Clean*), and demonstrated its high performance in the music similarity feature extraction. However, in general, these clean individual instrument sounds are not publicly available, making it practically impossible to utilize them in general-purpose music recommendation and retrieval systems. Therefore, research on the InMSRL model, which inputs the music pieces themselves, has progressed and two main approaches have been proposed. The first approach sequentially performs music source separation (MSS) and music similarity feature extractions (*Cascade*) [9]. However, since an MSS model and music similarity feature extractors are independently trained, separation errors are likely to cause adverse effects on the music similarity feature extraction. On the other hand, the second approach directly extracts disentangled music similarity features (*Direct*) [10]. This method learns a disentangled feature space consisting of different subspaces for individual instrument sounds. However, such a learning is not straightforward, and InMSRL performance tends to degrade for some instruments.

In this paper, we propose multi-task learning approaches that leverage MSS to address issues of *Cascade* and *Direct* and aim to construct a universally applicable InMSRL model. For *Cascade*, we propose *Cascade-FT* that performs end-to-end fine-tuning (FT) of the pre-trained MSS model and the music similarity feature extractors using an auxiliary separation loss. For *Direct*, we propose *Direct-Reconst* that uses multi-task learning based on the disentangled music similarity feature extraction and MSS based on reconstruction (Reconst) with the disentangled music similarity features. We conduct experimental evaluations and demonstrate that 1) *Cascade-FT* can improve InMSRL performance compared to *Cascade*, 2) *Direct-Reconst* can improve disentanglement performance in the music similarity feature extraction, and 3) *Cascade-FT* have better performance than *Direct-Reconst*.

[1]https://go.pardot.com/l/52662/2023-10-23/ljk7xt/52662/169805013966KGzgtB/Spotify_2023_Culture_Next_Report_JP_v3.pdf
[2]https://www.ifpi.org/wp-content/uploads/2020/03/Global_Music_Report_2023_State_of_the_Industry.pdf

II. Conventional InMSRL Methods

A. Clean and Cascade

Hashizume et al. [9] proposed two InMSRL methods: one inputting clean individual instrument sounds into music similarity feature extractors (*Clean*) and the other inputting individual instrument sounds separated by the pre-trained MSS model of Spleeter [11] into them (*Cascade*).

The music similarity feature extractor of *Clean* and *Cascade* is both trained using a triplet loss. In the i-th triplet, the three types of sample segments, which is anchor $\mathbf{x}_i^{(a)}$ that serves as the basis, positive $\mathbf{x}_i^{(p)}$ defined as similar to the anchor and negative $\mathbf{x}_i^{(n)}$ defined as dissimilar to the anchor, are used. By denoting a distance function as $d(\cdot)$, a loss function can be formulated as follows:

$$\mathcal{L}_{\text{triplet}} = \max\{0, d(\mathbf{x}_i^{(a)}, \mathbf{x}_i^{(p)}) - d(\mathbf{x}_i^{(a)}, \mathbf{x}_i^{(n)}) + \delta\} \quad (1)$$

where δ is a margin that defines the minimum distance between the anchor-positive and anchor-negative pairs. To perform label-free learning, assuming that sample segments extracted from the same music piece are similar to each other, a triplet is constructed as follows:

- Anchor: Extracted from a randomly selected music piece
- Positive: Extracted from the same music piece as that of the anchor
- Negative: Extracted from a different music piece from that of the anchor.

In *Cascade*, it is inevitable to cause separation errors in MSS. The previous studies [9] have confirmed that the performance of *Cascade* significantly degrades compared with *Clean*. Therefore, it is crucial to optimize the MSS model for the instrument-dependent music similarity feature extractors.

B. Direct

Hashizume et al. [10] also proposed the other InMSRL method to extract a disentangled music similarity feature with a single feature extractor, where the disentangled music similarity feature consists of subspaces for individual instrument-dependent music similarity features, e.g., the first to 128-th dimensional components of the 640-dimensional feature vector are used to represent the music similarity focusing on drums.

The training process first involves pre-training. In this training, the single disentangled music similarity feature extractor is trained using the disentangled features as reference targets, which is formed by concatenating instrument-dependent music similarity features extracted by *Clean*.

Next, similar to *Cascade*, the disentangled music similarity feature extractor is further updated by using the triplet loss as shown in (1). However, unlike *Cascade*, it is not straightforward to train such a feature extractor. To develop the disentangled music similarity feature extractor working reasonably, the following two approaches are used.

- Conditioning the output of the disentangled music similarity feature extractor
- Using pseudo-music-pieces as inputs.

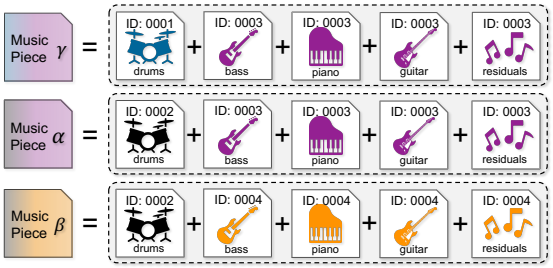


Fig. 1. Overview of pseudo-music-pieces. Instruments of the same color and the same ID indicate sample segments extracted from the same music-piece. This figure illustrates an example of the pseudo-music-pieces created for learning focusing on drums.

Conditioning process conduct a masking operation inspired by other disentangled representation learning [12], [13]. For example, when focusing on the bass feature, we leave only the dimensional components corresponding to a subspace for the bass feature and masks the other dimensional components to 0. By partially masking the feature vector, each subspace can model the music similarity feature depending on a specific instrument sound.

The Use of pseudo-music-pieces aims to improve the disentangled performance of *Direct*. Fig. 1 shows an overview of the pseudo-music-pieces. In Fig. 1, a music piece α and a music piece β are similar to each other in drums but dissimilar in the other instruments sounds. In contrast, the music piece α and a music piece γ are dissimilar in drums but similar in the other instruments. In the triplet loss-based learning, by using the music piece α as the anchor, the music piece β as the positive, and the music piece γ as the negative, the model can focus only on the drum features.

However, it is still challenging to accurately disentangle a music piece into the instrument-dependent subspace features. Consequently, the performance of InMSRL based on *Direct* tends to be insufficient.

III. Proposed InMSRL Methods Leveraging Multi-task Learning

A. Cascade-FT

To address the issue of *Cascade*, we propose *Cascade-FT* to optimize the MSS model by performing end-to-end fine-tuning (FT).

1) Network Architecture: The network architecture of *Cascade-FT* consists of the MSS model and the instrument-dependent music similarity feature extractor connected in series as shown in Fig. 2. The MSS model is based on the U-Net [14], [15] structure, similar to the Spleeter [11] used in *Cascade*. The network outputs a separation mask and the separated instrument sound is generated by Hadamard product of the input music spectrogram and the separation mask. In this paper, we develop the instrument-dependent MSS models to separately estimate the separation masks for individual instrument sounds. The instrument-dependent music similarity feature extractor is based on the U-Net encoder structure

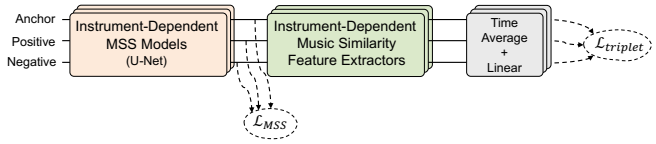


Fig. 2. Overview of *Cascade-FT* model.

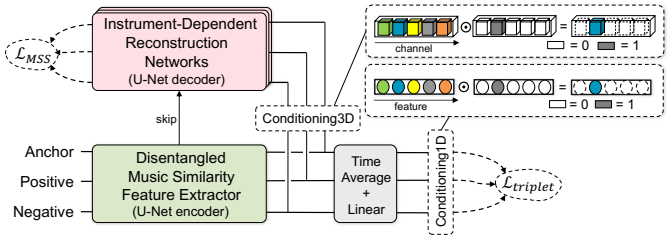


Fig. 3. Overview of *Direct-Reconst* model.

additionally using time-averaging and flattening operations and a fully-connected layer to output a 128-dimensional feature vector for each instrument sound.

2) Training: The training procedure consists of three stages: training of the MSS models, training of the instrument-dependent music similarity feature extractors and end-to-end fine-tuning. First, the MSS models are trained in the same manner as proposed by Jansson et al. [15]. The separation loss for each instrument sound (denoted as $\mathcal{L}_{MSS}$ in Fig. 2) is calculated as the L1 loss between the output separated instrument sound and a clean target instrument sound. Next, the music similarity feature extractors are trained using the triplet loss given by (1) (denoted as $\mathcal{L}_{triplet}$ in Fig. 2) in the same manner as in *Cascade*. During the training, the MSS models are frozen and their parameters are not updated. The L2 norm is employed as the distance function $d(\cdot)$ in the triplet loss. Finally, in the end-to-end fine-tuning stage, all parameters of the cascaded network consisting of the MSS models and the instrument-dependent music similarity feature extractors are updated by using a combined loss function given by the triplet loss for the instrument-dependent music similarity feature extractors and the separation loss for the MSS models as an auxiliary loss. Note that three inputs (anchor, positive, and negative) are required to compute the triplet loss, the auxiliary separation loss for the MSS models during fine-tuning is calculated for all three inputs. In the training, the pseudo-music-pieces segments (shown in Fig. 1) are also used as in *Direct*. Besides, we implement the data augmentation as described in Section III-B3.

B. Direct-Reconst

To address the issue of *Direct*, we propose *Direct-Reconst* incorporating MSS based on the reconstruction (Reconst) with the disentangled music similarity features for the training of the disentangled feature extractors.

1) Network Architecture: Fig. 3 shows the network architecture of *Direct-Reconst*. The *Direct-Reconst* network consists of three parts: the disentangled music similarity feature extractor, a reconstruction network to reconstruct each instrument sound from output sequences of the disentangled music similarity feature extractor, and a time-averaging and flattening operations and fully-connected layer to generate the disentangled music similarity feature vector from the output sequences. The disentangled music similarity feature extractor has a similar structure to the encoder of U-Net [15], and the reconstruction network has a similar structure to the decoder of U-Net [15]. The each layer of the disentangled music similarity feature extractor and those of the reconstruction network are connected by skip connections. The instrument-dependent reconstruction networks are developed for individual instrument sounds. As in the MSS models, the reconstructed instrument sound is generated by Hadamard product of the input music source spectrogram and the output separation mask.

2) Training: The training procedure consists of two stages: pre-training of the music similarity feature extractor and multi-task learning of the music similarity feature extractor and the instrument-dependent reconstruction network. In the pre-training of the music similarity feature extractor, we follow the same training procedure as in *Direct* [10]. We use 31 out of the 2^5 possible combinations of 5 musical instrument sources (drums, bass, piano, guitar, and residuals) as input, excluding the silent pattern. The training loss for the multi-task learning is a combination of the triplet loss given by (1) (denoted as $\mathcal{L}_{triplet}$ in Fig. 3) for the disentangled music similarity features and the reconstruction loss (denoted as $\mathcal{L}_{MSS}$ in Fig. 3) for the output reconstructed instrument sounds. The distance function $d(\cdot)$ in the triplet loss for the disentangled music similarity features is defined as the L2 norm. The reconstruction loss is defined as the L1 loss between the output instrument sound from the reconstruction network and the clean instrument sound in the same manner proposed by Jansson et al. [15]. As in *Direct*, we use the conditioning operation and the pseudo-music-pieces.

3) Disentanglement Enhancement: To enhance the disentangled music similarity feature extractor, we modify the conditioning process and utilize pseudo-music-pieces. The modified conditioning process applies the masking operation to not only the output of the time-averaging and flattening operations and fully-connected layer (Conditioning1D in Fig. 3) but also the input of the reconstruction network (Conditioning3D in Fig. 3). Conditioning1D is the same as the conditioning process used in *Direct*. Conditioning3D is its extension to apply the masking operation to a feature sequence. By Conditioning3D, the reconstruction network can focus only on the features corresponding to each target instrument sound. For the pseudo-music-pieces, we further introduce data augmentation (DA). While *Direct* generates a fixed set of triplet data of the pseudo-music-pieces beforehand and use it in the training, *Direct-Reconst* introduces a process of randomly generating triplet data of the pseudo-music-pieces each time to construct a mini-batch during training.

IV. Experimental Evaluations

A. Dataset

The dataset used for evaluation was the Slakh [16], which was also used in the previous study [9], [10]. The dataset consisted of MIDI-generated music pieces and MIDI instrument tracks in the music pieces. Each music piece contained drums, bass, piano, and guitar sounds. Following previous studies [9], [10], the other sounds in the music pieces were treated as residuals.

The dataset consisted of 2100 music pieces containing multiple groups of music pieces generated from the same MIDI file. In this experiment, we excluded music pieces generated from the same MIDI file, resulting in 1200 music pieces used for training, 270 music pieces used for validation, and 136 music pieces used for evaluation.

B. Evaluation Metrics

The proposed methods were based on the same learning paradigm as the previous studies [9][10], assuming that segments extracted from the same music piece were similar to each other. Therefore, as an evaluation metric, we used music ID estimation accuracy as used in the previous studies. In this experiment, we used the following two metrics, a music ID estimation score on normal-test-music-pieces (MES-Normal) and a music ID estimation score on pseudo-test-music-pieces (MES-Pseudo).

1) Music Estimation Score on Normal-Test-Music-Pieces (MES-Normal): To evaluate the performance of the feature representation, we used the accuracy of the music ID estimation with a simple method using the feature representation. Specifically, assuming that all test segments were embedded into the learned feature space beforehand and the music IDs of all segments were known except for a test segment to be estimated, we used the 5-nearest neighbors (5NN) method to estimate the music ID of the test segment. In the evaluation for each instrument sound, we only used feature dimensions corresponding to the target instrument in 5NN distance calculation while masking the other feature dimensions. The entire test dataset (136 music pieces) was used to calculate the music ID estimation accuracy.

2) Music Estimation Score on Pseudo-Test-Music-Pieces (MES-Pseudo): The proposed method aimed to learn the music similarity feature representations focusing on individual instrument sounds. However, in MES-Normal, the ground truth label for the 5NN method was the same over all instrument sounds as shown in the top part of Fig. 4. Therefore, it was essentially hard to evaluate the disentanglement performance of the learned representations by MES-Normal. To investigate the disentanglement performance, we used pseudo-test-music-pieces in MES-Pseudo. In MES-Pseudo, the ground truth label was different between the target instrument and the others; e.g., the label of the target instrument sound (i.e., drums label) was different from the others as shown in the bottom part of Fig. 4. Furthermore, we excluded all segments extracted from the same pseudo-music-piece as that of each test segment to prevent the music ID estimation focusing on the no-targeted instruments. The pseudo-test-music-pieces used for the test consisted of 50 music pieces in total, with 10 music pieces for each target instrument sound label and 5 music pieces for each non-target instrument sound label.

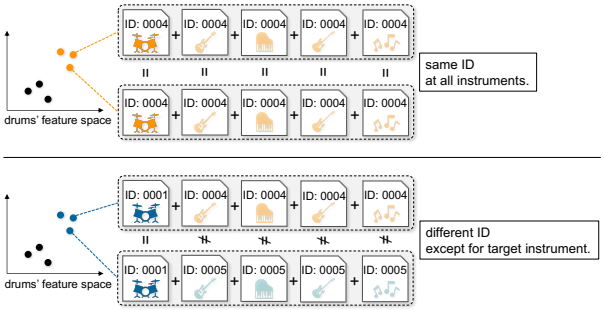


Fig. 4. Difference between MES-Normal and MES-Pseudo. The top part of the figure shows MES-Normal, and the bottom part shows MES-Pseudo. This is the example of evaluation for the drums. Instruments of the same color and the same ID indicate segments extracted from the same music piece.

C. Experimental conditions

Music segments used in the experiments were cut into 3-second segments for training and 10-second segments for validating and testing. The music segments where the target instrument was silent were excluded. The sampling rate was set to 44100 Hz, and a window size of 2048 and an offset of 512 were used for the short-time Fourier transform (STFT). The number of mel frequencies for the log Mel-spectrogram used as input to the *Cascade-FT* music similarity feature extractors was set to 259. The learning rate for the training of the MSS models and the music similarity feature extractors in *Cascade-FT* was set to 0.00005, and the learning rate for the fine-tuning was set to 0.00001. The learning rate for the pre-training of *Direct-Reconst* and the multi-task training of the disentangled music similarity feature extractor and the reconstruction network was set to 0.0001. Adam [17] was used to train both models. The maximum number of epochs was set to 400, and training was terminated if the minimum value of the loss function on the validation data was not updated over 100 epochs.

D. Experimental Results

Evaluation results of MES-Normal and MES-Pseudo are shown in Table I and Table II, respectively. We also show an evaluation result of MSS accuracy for the output separated sounds in *Cascade* methods in Table III.

1) Evaluation of Cascade-FT: It can be observed from Table I that *Cascade-FT* achieves higher evaluation scores than the previous method [9] for all instruments. This suggests that the proposed methods achieve higher InMSRL performance compared to the previous method. From a comparison between *Cascade* w/ from-scratch w/o pseudo-music-pieces and *Cascade* w/ Spleeter [9], the performance improvements can be seen in former. This is caused by the insufficient separation accuracy of the MSS model of *Cascade* [9], as shown in Table

TABLE I

EVALUATION RESULTS OF MES-NORMAL (%). THE EVALUATION SCORES OF *Clean* [9], *Cascade* W/ SPLEETER [9] AND *Direct* [10] ARE RESPECTIVELY QUOTED FROM [9] AND [10]. IN W/O PSEUDO-MUSIC-PIECES, THE MUSIC SIMILARITY FEATURE EXTRACTORS ARE SIMPLY TRAINED WITH NORMAL MUSIC PIECES. EXCLUDING ABLATION, THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	drums	bass	piano	guitar	residuals
Clean [9]	98.04	94.60	98.14	96.35	-
Cascade w/ Spleeter [9]	88.91	63.87	50.34	-	-
Cascade w/ from-scratch	90.98	73.39	80.77	79.53	-
w/o pseudo-music-pieces (ablation study)	92.71	90.20	93.62	90.90	-
Cascade w/ from-scratch, FT (*Cascade-FT*)	**93.03**	**74.96**	**81.96**	**82.78**	-
w/o pseudo-music-pieces (ablation study)	94.89	95.63	96.21	94.40	-
Direct [10]	89.69	**84.45**	**85.70**	**86.27**	84.86
Direct w/ DA	89.33	71.09	79.74	81.75	85.67
Direct w/ DA, Reconst (*Direct-Reconst*)	**91.14**	81.30	84.76	85.17	**88.84**

TABLE II

EVALUATION RESULTS OF MES-PSEUDO (%). THE EVALUATION SCORES OF *Direct* [10] ARE QUOTED FROM THE PREVIOUS STUDY [9]. IN W/O PSEUDO-MUSIC-PIECES, THE MUSIC SIMILARITY FEATURE EXTRACTORS ARE SIMPLY TRAINED WITH NORMAL MUSIC PIECES.

Method	drums	bass	piano	guitar	residuals
Cascade w/ from-scratch	98.68	93.02	91.73	92.19	-
w/o pseudo-music-pieces (ablation study)	95.09	77.30	81.02	77.60	-
Cascade w/ from-scratch, FT (*Cascade-FT*)	**98.91**	**94.80**	**93.55**	**93.89**	-
w/o pseudo-music-pieces (ablation study)	95.96	71.54	69.59	77.40	-
Direct [10]	85.5	37.1	31.3	44.7	74.7
Direct w/ DA	97.93	68.22	69.22	63.24	89.99
Direct w/ DA, Reconst (*Direct-Reconst*)	**98.25**	**77.74**	**79.20**	**82.47**	**94.62**

TABLE III

EVALUATION RESULTS OF THE MSS ACCURACY FOR THE OUTPUT SEPARATED SOUND IN *Cascade*. SDR (SIGNAL-TO-DISTORTION RATIO) WAS USED FOR THE EVALUATION. THE RESULTS OF *Cascade* W/ SPLEETER [9] ARE QUOTED FROM THE PREVIOUS STUDY [9]. MUSEVAL [18] WAS USED FOR CALCULATION OF SDR.

	SDR			
Method	drums	bass	piano	guitar
Cascade w/ Spleeter [9]	-13.7	-15.5	-14.7	-
Cascade w/ from-scratch	**15.50**	10.54	**7.81**	6.94

III. This poor performance of Spleeter is likely due to the fact that Spleeter is trained on music with raw-audio-songs, while the experiments in this paper and [9] use music pieces generated from MIDI. Moreover, we can also observe that the fine-tuning in the proposed method is effective for further performance improvements from a comparison between *Cascade* w/ from-scratch and *Cascade-FT*. This results demonstrates that the performance of the MSS model in *Cascade* methods strongly affects the accuracy of InMSRL.

The disentanglement performance of each InMSRL method can be compared in Table II. All evaluation scores of *Cascade-FT* exceed 90%. We also observe that the fine-tuning is helpful to further improve the performance.

These results suggest that the proposed method *Cascade-FT* can learn high-quality music similarity feature representations focusing on individual instrument sounds.

2) Evaluation of Direct-Reconst: Table I shows that *Direct-Reconst* does not outperform the previous method [10] for some instruments, i.e., bass, piano, and guitar. On the other hand, *Direct-Reconst* significantly outperforms previous method in the evaluation result of MES-Pseudo as shown in Table II. These results indicate that *Direct* [10] for MES-Normal is significantly affected by the leakage of the other instrument sounds and its disentanglement performance is actually low. On the other hand, the proposed method *Direct-Reconst* not only improves the evaluation scores of MES-Pseudo but also maintains the evaluation scores of MES-Normal at the same level as the previous method. Therefore, the proposed method can achieve better InMSRL performance than the previous method. Table II also shows that DA significantly improves the MES-Pseudo score, demonstrating the effectiveness of DA. Additionally, a comparison of *Direct* w/ DA and *Direct-Reconst* in Tables I and II shows that the multi-task learning of the disentangled music similarity feature extraction and the reconstruction is effective for improving the InMSRL performance.

3) Clean, Cascade and Direct Approaches Comparison: Tables I and II show that the *Direct* approach tends to have higher MES-Normal scores than MES-Pseudo scores for some instruments except for drums and residuals. Normally, MES-Normal score would be lower than or equal to the MES-Pseudo score because the MES-Normal uses 136 target labels compared to 10 for the MES-Pseudo at 5NN. This is considered to be due to the leakage of the other instrument sounds as discussed in Section IV-D2. In contrast, the *Cascade* approach can more precisely focus only on target instrument sound, as demonstrated by the higher MES-Pseudo scores than the MES-Normal scores. Therefore, it is demonstrated that the *Cascade* approach achieves higher InMSRL performance than the *Direct* approach. On the other hand, the *Direct* approach needs to use only the disentangled music similarity feature extractor in the inference step, and therefore, its computational cost is lower than the *Cascade* approach.

MES-Normal scores of *Clean* are the most highest in all

of InMSRL models and it is predicted that MES-Pseudo scores would be much higher scores than MES-Normal scores because of its less target labels. This result is to be expected because *Clean* utilizes clean individual instrument sounds as input, which are generally not publicly available, therefore explicitly providing distinct individual instrument features to the music similarity feature extractors. By utilizing multi-task learning, we were able to improve the performance of *Cascade* and *Direct* models, which input the music pieces themselves. However, there is still room for further improvement in *Cascade* and *Direct* approach considering *Clean* model.

4) The effectiveness of pseudo-music-pieces: In *Cascade* approach, the evaluation result of w/o pseudo-music-pieces showed in Table I and Table II indicates that by using pseudo-music-pieces, we can minimize the adverse effects of separation errors caused by the MSS model. Note that although the performance w/o pseudo-music-pieces looks higher than that w/ it in Table I, this result is caused by the leakage of the other instrument sounds, and therefore, the actual InMSRL performance is limited. The use of pseudo-music-pieces is also essential in *Direct* approach as reported in [10]. These results demonstrate that the use of pseudo-music-pieces is an important technique to improve InMSRL performance.

V. Conclusion

In this paper, we have proposed two methods to improve InMSRL performance by modifying two existing approaches, *Cascade* and *Direct*. For *Cascade*, we have proposed end-to-end fine-tuning of the MSS model and the music similarity feature extractors using an auxiliary separation loss, and for *Direct*, we have proposed joint training of the disentangled feature extraction and MSS based on the reconstruction with the disentangled music similarity features. We have conducted experimental evaluations and have demonstrated that the end-to-end fine-tuning for *Cascade* improves InMSRL performance, the multi-task learning for *Direct* is also helpful to improve disentanglement performance in the feature extraction and *Cascade* with the end-to-end fine-tuning outperforms *Direct* with the multi-task learning. Future work includes using raw-audio-songs and vocal sound.

Acknowledgment

This work was partly supported by JST CREST under Grant Number JPMJCR19A3.

References

[1] M. Balabanović and Y. Shoham, "Content-based, collaborative recommendation," *Association for Computing Machinery*, vol. 40, no. 3, pp. 66–72, 1997.

[2] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. in Artif. Intell.*, 2009, 19 pages.

[3] A. Elbir and N. Aydin, "Music genre classification and music recommendation by using deep learning," *Electronics Letters*, vol. 56, no. 12, pp. 627–629, 2020.

[4] M. Won, S. Oramas, O. Nieto, F. Gouyon, and X. Serra, "Multimodal metric learning for tag-based music retrieval," in *IEEE ICASSP*, 2021, pp. 591–595.

[5] J. Singh and V. K. Bohat, "Neural network model for recommending music based on music genres," in *ICCCI*, 2021, pp. 1–6.

[6] M. S. Fathollahi and F. Razzazi, "Music similarity measurement and recommendation system using convolutional neural networks," *International Journal of Multimedia Information Retrieval*, vol. 10, pp. 43–53, 2021.

[7] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech & Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[8] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Trans. Audio, Speech, & Language Processing*, vol. 16, no. 2, pp. 424–434, 2008.

[9] Y. Hashizume, L. Li, and T. Toda, "Music similarity calculation of individual instrumental sounds using metric learning," *APSIPA ASC*, pp. 33–38, 2022.

[10] Y. Hashizume, L. Li, A. Miyashita, and T. Toda, "Learning multidimensional disentangled representations of instrumental sounds for musical similarity assessment," in *arXiv e-prints: 2404.06682*, 8 pages, 2024.

[11] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and state-of-the art music source separation tool with pre-trained models," *The Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2019.

[12] A. Veit, S. Belongie, and T. Karaletsos, "Conditional similarity networks," in *IEEE CVPR*, 2017, pp. 1781–1789.

[13] J. Lee, N. J. Bryan, J. Salamon, Z. Jin, and J. Nam, "Disentangled multidimensional metric learning for music similarity," in *IEEE ICASSP*, 2020, pp. 6–10.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[15] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *ISMIR*, 2017, pp. 23–27.

[16] E. Manilow, G. Wichern, P. Seetharaman, and J. L. Roux, "Cutting music source separation some slakh: A dataset to study the impact of iraining data quality and quantity," in *IEEE WASPAA*, 2019, pp. 45–49.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 13 pages, 2014.

[18] F. R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*, 2018, pp. 293–305.

Enhancing Early Plant Disease Detection: 1D to 2D Spectral Transformations

Mas Ira Syafila Mohd Hilmi Tan, Lai-Kuan Wong, Yuen Peng Loh, Chih-Yang Pee
Multimedia University, Malaysia
E-mail: {masirasyafila, lkwong, yploh, cypee}@mmu.edu.my

***Abstract*—Early detection of plant diseases can be challenging due to the subtle differences in early-stage symptoms. Relying solely on the visible traits is insufficient for accurate diagnosis, especially for presymptomatic detection, emphasizing the significance of improved approaches for early disease identification. Spectroscopy serves as a reliable tool in diagnosing asymptomatic diseases, but interpreting raw 1D spectral data is complex due to the indistinguishable patterns among different disease condition. This study presents a novel approach by converting 1D spectral data of cassava leaves into 2D spectral images to reveal the distinguishing features of healthy cassava leaves from those affected with Cassava Mosaic Disease (CMD), and Cassava Brown Streak Disease (CBSD), aiming to enhance classification accuracy. Traditional machine learning techniques such as k-Nearest Neighbors (kNN) and Extremely Gradient Boosting (XGBoost) are compared with deep learning methods, including a 1D Convolutional Neural Network (CNN) based on EfficientNetB0 for direct spectral classification and several pretrained CNN models for 2D image data classification, including ResNet50V2, MobileNetV2, EfficientNetB0 and InceptionV3. Different padding strategies for 2D data are evaluated, including no padding, end padding, beginning padding, top-bottom padding, and centered padding. Our findings demonstrate that transforming 1D spectral data into 2D images and applying effective padding techniques can improve classification accuracy by 5.07% compared to 1D data classification. This study offers a promising alternative approach for early plant disease detection by converting 1D spectral data into 2D representations, enabling CNNs to capture spatial features that improve cassava disease classification accuracy.**

I. Introduction

Diagnosing presymptomatic plant diseases poses significant hurdles, particularly when depending on visual inspection alone [1]. This inspections often struggle to accurately identify pathogens at low prevalence levels due to the latency of symptom emergence [2].

To address this limitation, visible and near-infrared (VIS-NIR) spectroscopy provides a promising solution for for early detection of plant diseases. This technique measures spectral reflectance from visible (400 - 750 nm) to near-infrared (750 - 2500 nm), to assess changes in the plant's physiological and biochemical status. However, interpreting these spectra is tricky due to the overlapping attributes of various compounds [3]. This complexity emphasizes the importance of improved techniques for plant disease diagnosis.

While previous studies have primarily focused on extracting features from one-dimensional (1D) spectral data for plant disease classification, the complexity of interpreting these spectra remains a challenge. Therefore, to overcome this challenges, this study explores the novel approach of converting 1D spectral data into two-dimensional (2D) image formats. This transformation enables the use of advanced image-based deep learning models, leveraging pretrained convolutional neural networks (CNNs) by capturing spatial correlations and enhancing classification performance while reducing noise [4]. Recent advancements in deep learning (DL) techniques have shown that CNN models achieve state-of-the-art performance in 2D image classification tasks. Reference [5] compared 1D and 2D CNNs for land cover and crop classification, showing that 2D CNNs outperformed 1D CNNs. This highlights the importance of converting 1D spectral data into a 2D format for improved classification performance.

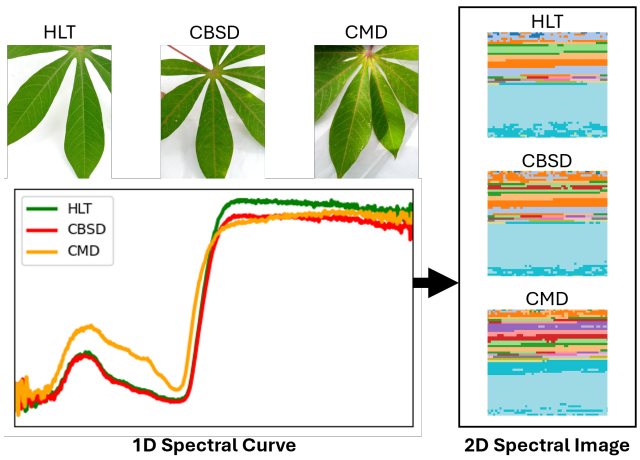


Fig. 1. Spectral data of three presymptomatic cassava leaf classes: Healthy (HLT), Cassava Brown Streak Disease (CBSD), and Cassava Mosaic Disease (CMD). The 1D spectral curve reveals overlapped and similar patterns among the classes. While the 1D spectral data tends to be localized and shows overlapping patterns among classes, the 2D representation stretches the spectrum, providing a broader view that captures spatial information and enhances class differentiation.

The significance of this study lies in its focus on enhancing early cassava disease detection by transforming 1D spectral data into 2D images. By converting the sequential arrangement of 1D spectral data into 2D spectral images, the approach facilitates the utilization of a wide range of complex and diverse spectral features. Fig. 1 shows the transformation outcome, where 1D spectral data from three different class samples is converted into corresponding 2D spectral images. The 2D format reveals spatial patterns and relationships between spectral bands that are not apparent in the linear 1D format.

Each pixel in the 2D image corresponds to specific spectral features, providing a more detailed view of the data. This 2D representation helps differentiate between samples, improving classification performance for early disease detection.

We hypothesize that by converting 1D spectral data into 2D images, CNNs will be able to better exploit spatial relationships among spectral bands. This approach leverages the spatial learning capability of CNNs, which is known to enhance feature extraction, leading to improved classification accuracy.

In addition, our study assesses several padding techniques to ascertain their impact on classification accuracy thus exploring the most effective padding techniques for the models. This innovative approach combines traditional machine learning techniques such as k-Nearest Neighbors (kNN) and Extreme Gradient Boosting (XGBoost) with deep learning methods, including a 1D CNN based on EfficientNetB0 for direct spectral classification and several pretrained CNNs for 2D image data classification, including ResNet50V2, MobileNetV2, EfficientNetB0 and InceptionV3. This comprehensive evaluation offers a fresh perspective on leveraging spectral data in higher dimensions for plant disease detection. The key findings of this study are as follows:

- An alternative approach is introduced to enhance plant disease detection by converting 1D spectral data into 2D spectral images. This method provides a novel perspective on leveraging transformed spectral data for plant disease classification.
- Evaluation of the impact of various padding techniques (No padding, Beginning padding, End padding, Between Padding, and Centered padding) on the classification performance of 2D spectral images. For example, EfficientNetB0 models achieves an accuracy of 88.64% with no padding, showing notable increments of accuracies ranging from 90% to 93% with different padding techniques.
- This study reveals a notable classification accuracy improvement by 5.07% for EfficientNetB0 on 2D data with Between padding. This finding demonstrates that transforming 1D spectral data into 2D spectral images, aided with appropriate padding techniques, successfully enhances the classification performance.

II. Related Works

A. VIS-NIR Spectroscopy for Early Detection of Plant Disease

Numerous studies have demonstrated the effectiveness of VIS-NIR spectroscopy for identification of plant diseases. Through the analysis of changes in spectral reflectance, this technique is a viable tool for early detection of plant disease symptoms [6]. This method facilitates the identification of stress or disease onset prior to the emergence of visible symptoms, thereby offering valuable insights for the early diagnosis and management of diseases.

Reference [7] demonstrates the capabilities of VIS-NIR spectroscopy for early detection of powdery mildew in grapevines. Similarly, a study employs VIS-NIR spectroscopy to detect powdery mildew in cucumber plants at a asymptomatic stage [8]. The study achieves over 78% accuracy in classifying healthy versus infected plants, illustrating that spectral data can be used to identify diseases before symptoms become visible.

Another study combined NIR spectroscopy with machine learning algorithms for presymptomatic detection of rice sheath blight and achieves high accuracy in detecting viral infections before visual symptoms appeared [9].

B. CNNs for Spectral Plant Disease Detection

Convolutional Neural Networks (CNNs) have demonstrated notable potential in processing data for the classification of plant diseases. CNNs, initially developed for image analysis [10], have been adapted to accommodate various data types allowing it identify complex patterns and features that are essential for accurate disease detection.

For instance, a 1D-CNN model is developed to capture and fuse multiple deep features from VIS-NIR spectra for the nondestructive identification of pesticide residues on Hami melon surfaces [11]. The model demonstrates an impressive classification accuracy of 99.17% for distinguishing pesticide residues. Another study focused on identifying bacterial blight (BB) resistant rice seeds, demonstrating that CNN models using 2D spectral images show superior classification accuracy compared to traditional machine learning methods [12].

Furthemore, comparison of 1D, 2D and three-dimensional (3D) CNNs are investigated for the detection of aphid infection in cotton leaves using a VIS-NIR hyperspectral imaging system. The results show that 2D CNNs excel in classifying RGB images, while 3D CNNs perform better for hyperspectral images, indicating the advantage of multidimensional approaches in enhancing classification performance [13]. This study highlights the importance of converting 1D spectral data into higher dimensions to capture spatial features.

III. Methodology

The proposed approach of this paper mainly consist of two key components: (1) conversion of 1D spectral data into a 2D format by reshaping it into 2D grid, and (2) applying zero padding techniques to enhance data representation. The first component focuses on transforming the data structure to enable more effective analysis, while the second component examines various padding strategies to improve the quality and usability of the 2D data.

A. Conversion of 1D Spectral Vector to 2D Spectral Image

Transformation of 1D spectral data to 2D spectral images facilitates the utilization of a wide range of complex and diverse spectral features. The sequential arrangement of 1D spectral bands can restrict the capacity to identify relationship between bands. Conversely, the 2D format arranges the bands into a grid, which offers a more comprehensive integration of spectral bands and a larger spectral coverage. The spatial arrangement of each pixel in the transformed 2D image is

associated with a specific spectral feature, which improves the ability to differentiate between different samples by offering a more detailed and in-depth view of the spectral data.

The original 1D spectral vector, denoted as 1×b (where b represents the number of bands), is transformed into a 2D matrix with dimensions $w \times h$, where w and h represent width and height, respectively. The spatial integration of spectral information is improved by this reconfiguration.

The 2D matrix is expanded to include a depth dimension for channels in CNNs. As a result, the input data format is transformed into $w \times h \times d$, where d represents the number of channels.

B. Zero Padding Techniques

Zero padding involves adding zero values to the 1D spectral data to achieve the desired 2D image dimensions. Zero-padding enables CNNs to maintain spatial information within their internal representations by preserving the dimensionality of the data during convolution. In contrast, the absence of padding prevents the network from encoding positional information. By applying zero-padding, CNN models are able to preserve spatial details, which helps the filters to identify specific features and ensures that this spatial information is consistently propagated across the entire image [14]. During the padding process, it is essential to manage the amount of padding applied to avoid loss of useful information.

To determine the amount of padding required, the following formula is used:

$$P_{\text{total}} = D_{\text{target}} - D_{\text{original}} \tag{1}$$

where D_{target} is the desired dimension of the 2D image and D_{original} is the dimension of the original 1D spectral data. This calculation helps ensure that the 2D spectral images are generated with the necessary dimensions. Padding techniques influence the spatial distribution of spectral data, which affects how spectral features are captured and interpreted by the CNN. By maintaining the spatial dimensions of the input data, padding allows models to learn more robust features from the spectral images, thereby enhancing learning and classification performance.

IV. Experiments and Results

A. Dataset

The dataset used for this study consists of publicly available spectral data labeled with several cassava disease conditions [15]. It includes spectral data and images for three classes: healthy (HLT), cassava mosaic disease (CMD), and cassava brown streak disease (CBSD). Data collection is stopped once symptoms appeared, resulting in a dataset that is mostly presymptomatic. Consequently, the image data do not accurately represent the condition as they mostly capture the presymptomatic stage.

Only the spectral dataset from weeks 14 to 19 of the screen house experiment is utilized. The chosen time frame is determined based on Real-Time Polymerase Chain Reaction (RT-PCR) chemical lab test results, which showing pronounced and distinctive differences between classes. Throughout this study, a filtered dataset comprising 2900 spectral bands ranging from 450 to 1000 nm is used.

B. Implementation Details

1) 1D Spectral Vector to 2D Spectral Images Conversion: The spectral data, originally in 1D format with 2900 spectral bands, are normalized between 0 and 1. 1D spectral data are transformed into 2D spatial dimensions, introducing additional context that can be leveraged by CNNs. The conversion involved two main approaches:

1) Non-Padding Conversion: The 1D spectral data are reshaped into 50×58 images.
2) Padding Conversion: The data are converted into 75×75 images. Four padding techniques are used as depicted in Fig. 2:
   - Beginning Padding: Zero padding at the beginning (top) of the data.
   - End Padding: Zero padding at the end (bottom) of the data.
   - Between Padding: Zero padding at both the beginning and end (top and bottom) of the data.
   - Centered Padding: Zero padding around all sides of the data.

For the non-padding conversion, the 1D spectral data containing 2900 bands is reshaped into a 2D image of size 50×58 without any additional padding. Although a 50×58 matrix is not a square number, this format is chosen to avoid excessive padding and to retain as much original spectral information as possible. This size is used for early comparisons but is found incompatible with InceptionV3 due to its minimum size requirement.

Zero padding is employed to reshape 2900 bands into a 75×75 image. The 75×75 size is specifically chosen to meet the minimum input requirement for the InceptionV3 model [16]. Different padding techniques are assessed for their effects on classification accuracy, with the goal of determining which method best preserved spectral information and optimized model performance.

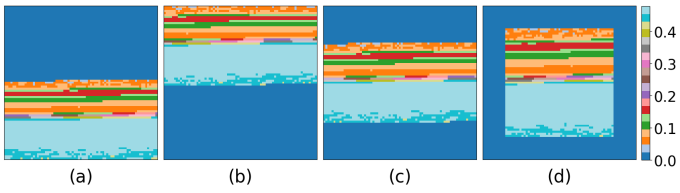


Fig. 2. 2D conversion of 1D spectral data (75×75) with zero-padding techniques: (a) Beginning, (b) End, (c) Between, and (d) Centered.

2) Cassava Disease Classification: The study explores classification techniques for cassava diseases using both traditional machine learning (ML) and deep learning (DL) approaches on 1D and 2D spectral data. The dataset is divided into 80% for training and validation and 20% for testing. Within

the training/validation set, a further split of 80% for training and 20% for validation is used to ensure balanced class representation.

Traditional machine learning techniques such as kNN and XGBoost, and an 1D-CNN are employed to classify cassava diseases on 1D spectral data. The kNN classifier is configured with k = 7 and the XGBoost classifier employs 200 estimators. The 1D-CNN architecture used in this study is adapted from the EfficientNetB0 model [17].

For 2D spectral images, pretrained 2D-CNN models such as ResNet50V2 [18], MobileNetV2 [19], EfficientNetB0 [17], and InceptionV3 [16] are employed for classification of cassava diseases. The models are fine-tuned, with three layers frozen for each model.

Training for 1D-CNN and 2D-CNN are performed using the Adam optimizer (learning rate of 0.0001) at 100 epochs, with a batch size of 32, employing early stopping and learning rate reduction techniques. Classification performance is evaluated using various metrics like loss, accuracy, precision, recall, and F1-score, with all metrics reported as percentages except for loss.

C. Comparison of 1D and 2D Spectral Data

The impact of data transformation can be examined through the comparison of 1D spectral data for each of the classes with the corresponding 2D spectral images as shown in Fig. 1. The spectral curves for different classes overlap significantly, with the profiles for all classes being particularly similar. This indicates that 1D spectral data may not capture sufficient information to distinguish between similar disease symptoms effectively.

Conversely, even though distinct visual patterns in the 2D spectral images are not always evident when examining all image samples, the effect of stretching the 1D spectrum into a 2D format, enables the model to extract more features from the entire set of data. On the other hand, 1D spectral data tends to localize features, which restricts the model's ability to extract features to a deeper level.

In addition to the classification taskes, Principal Component Analysis (PCA) is applied on both 1D and 2D spectral data solely for visualization purpose to illustrate the clusters between classes in the data [20]. For the 1D spectral data, PCA was performed directly on the spectral vector. For the 2D spectral data, the images were first reshaped into 1D vectors, then subjected to PCA. Fig. 3 illustrates the projection of both data types into a lower-dimensional space, highlighting the enhanced class separation achieved with 2D spectral data. For clarity, axis values have been omitted from the PCA plots. 2D spectral images produce more distinct class groupings than the 1D spectral data, which could contribute to improved classification performance. This improvement arises because in 2D spectral images, spatial relationships between adjacent spectral bands are preserved during PCA, allowing it to capture more complex features. By transforming 1D data into 2D images, we introduce an additional layer of spatial context, which helps to differentiate the classes more effectively. The CNN models are able to extract and apply broader features due to the improved class separability in the 2D representation. The benefits of 2D representation can be discerned in the following section, which presents the classification results for both 1D and 2D data, highlighting the superior performance of CNN models with 2D spectral images.

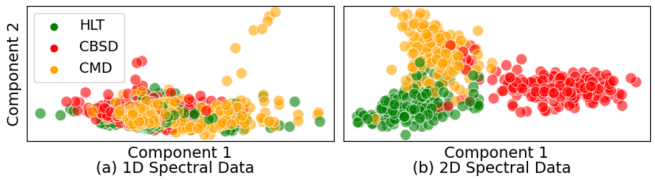


Fig. 3. PCA visualization comparing (a) 1D and, (b) 2D spectral data. The 2D spectral data shows improved class separation with more distinct cluster, implying that the transformation to 2D facilitates better discrimination among different classes compared to the 1D data.

D. Classification Results

The classification performance in this study is based on the evaluation of loss, accuracy, precision, recall, and F1-score. Our work is incomparable to the existing study by Owomugisha et al. [21] utilizing the same dataset, as they employed different models, settings and analyzed on two classes instead of three. Additionally, their work utilised different wavelength range and incorporated techniques such as dimensional reduction and wavelength selection for classification, which is not implemented in our study. The classification results for both 1D and 2D spectral data in this study is summarized in Table I and II, respectively.

The performance of 1D spectral data classification presented in Table I shows that XGBoost had the best performance with classification accuracy of 88.64%, precision of 88.89%, recall of 88.70%, and F1-score of 88.71% with lowest test loss of 0.3105 while kNN exhibited the worst performance with accuracy of 69.78%, precision of 69.83%, recall of 69.91%, and F1-score of 69.67% with significantly high test loss of 4.3207. The 1D-CNN based on EfficientNetB0 performs comparably well to XGBoost, with accuracy, precision, recall, and F1-score of 88.03%, 88.07%, 87.87%, and 87.90, respectively, but with a slightly higher loss of 0.5524. suggesting that both models are effective for 1D spectral data classification.

TABLE I
PERFORMANCE METRICS FOR 1D SPECTRAL DATA CLASSIFICATION

Model	Loss	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
XGBoost	**0.3105**	**88.64**	**88.89**	**88.70**	**88.71**
kNN	4.3207	69.78	69.83	69.91	69.67
1D-CNN	0.5524	88.03	88.07	87.87	87.90

Additionally, Table II presents the performance metrics for classification models ResNet50V2, MobileNetV2, EfficientNetV2, and InceptionV3 applied to the converted 2D spectral data, including conditions with no padding (None)

and paddings (Beginning, End, Between, Centered). EfficientNetB0 accomplishes the best accuracy, precision, recall and F1-score of 93.71%, 93.72%, 93.75% and 93.70%, respectively, with the lowest loss of 0.4350 when classified on 2D spectral data with Between padding. This showcases an notable enhancement up to 5.07% over the best accuracy achieved for 1D spectral data classification. Following this, ResNet50V2 classified on the 2D spectral data with Beginning padding demonstrated a competitive accuracy, precision, recall, and F1-score of 93.10%, 93.19%, 93.00%, and 93.04%, respectively, but with a slightly higher loss of 0.7497. Notably, EfficientNetB0 shows consistent high classification performance and low loss accross different padding techniques. For non-padded data, EfficientNet achieved an accuracy of 88.64%, whereas applying various padding strategies on 2D spectral data improve the accuracy to a range of 90% to 93%. Likewise, the impact of padding is significant across most classification models, as it leads to a notable increase in accuracy compared to non-padded data. However, without padding, CNN models may not fully leverage the potential of the 2D data, which can impact the effectiveness of 2D representation. Consequently, the best classification performance on 2D data without padding is achieved by EfficientNetB0, showing similar performances to the XGBoost classifier on 1D spectrad data, with both having 88.64% accuracy. Padding plays an important role in maintaining spatial consistency and enabling the CNN models to extract more robust features. Despite this, the underlying benefit of 2D data representation remains valid. It is mainly due to how the CNN processes the data which directly influences the performance outcomes. 2D spectral images provide a spatial representation of data, whereas 1D spectral data only provides a sequence of values.

TABLE II
PERFORMANCE METRICS FOR 2D SPECTRAL DATA CLASSIFICATION

Model	Padding	Loss	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ResNet50V2	None	1.0260	88.44	88.45	88.43	88.42
	Beginning	0.7497	93.10	93.19	93.00	93.04
	End	0.8231	91.48	91.48	91.42	91.44
	Between	1.0203	88.84	88.78	88.74	88.73
	Centred	1.0752	90.06	89.98	89.93	89.94
MobileNetV2	None	1.3131	83.77	84.57	83.74	83.31
	Beginning	3.7184	66.73	76.76	67.23	64.83
	End	1.0985	81.54	83.19	81.74	81.31
	Between	0.8142	86.41	87.05	86.52	84.40
	Centred	1.7963	82.35	84.37	81.79	82.03
EfficientNetB0	None	0.7613	88.64	88.67	88.51	88.55
	Beginning	0.6611	91.68	91.65	91.69	91.65
	End	0.6839	90.26	90.17	90.15	90.14
	Between	**0.4350**	**93.71**	**93.72**	**93.75**	**93.70**
	Centred	0.5899	91.48	91.46	91.49	91.46
InceptionV3	None	-	-			
	Beginning	1.2255	86.61	86.56	86.59	86.57
	End	1.3252	87.83	87.73	87.66	87.67
	Between	1.2581	88.24	88.23	88.14	88.17
	Centred	0.7938	91.28	91.27	91.21	91.22

To gain deeper insights into the classification performance, we further analyze the confusion matrices of the top-performance classifiers on 1D and 2D data. The comparison

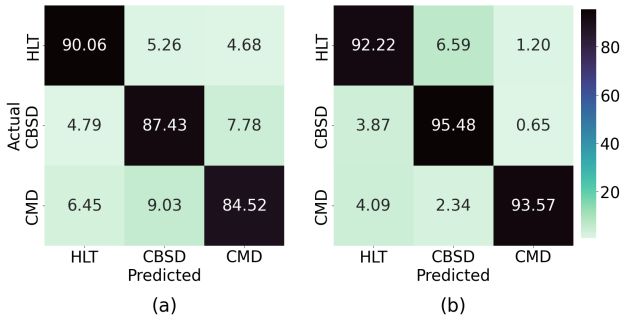


Fig. 4. Confusion matrix (%) illustrating classification results for (a) XGBoost on 1D spectral data, and (b) EfficientNetB0 with Between padding on 2D spectral data.

of confusion matrices between XGBoost on 1D spectral data and EfficientNetB0 with Between padding on 2D spectral data is depicted in Fig. 4.

XGBoost on 1D data achieves an accuracy of 88.64%. The confusion matrix in Fig. 4(a) shows that, 90.06% of class HLT are correctly classified, while 5.26% are misclassified as CBSD and 4.68% as CMD. For CBSD, 87.43% are correctly classified, with 4.79% misclassified as HLT and 7.78% as CMD. Class CMD has 84.52% correctly classified, with 6.45% misclassified as HLT abd 9.03% as CBSD. These results indicate that there is some misclassifications between classes, particularly between CBSD abd CMD.

EfficientNet on 2D spectral data with Between padding demonstrates a higher accuracy up to 93.71%. The confusion matrix in Fig. 4(b) displays 92.22% of class HLT are correctly classified, with 6.59% misclassified as CBSD and 1.20% as CMD. For CBSD, 95.48% are correctly classified, with 3.87% misclassified as HLT and less than 1% as CMD. CMD shows 93.57% correct classification with 4.09% misclassified as HLT and 2.34% as CBSD. This portrays the EfficientNetB0 great performance in discriminating between classes, with minimal misclassifications, utilizing 2D data representation with effective padding strategy.

V. CONCLUSION

This study demonstrates that converting 1D spectral data into 2D spectral images, with the application of appropriate padding techniques, allows CNNs to capture spatial relationships for early classification of cassava disease. Our approach yields a high classification accuracy of 93.71% for 2D spectral images with Between Padding, with about 5.07% accuracy enhancement compared to the 1D spectral classification result. The utilization of several padding techniques show that EfficientNetB0 consistently performs well across all padding methods, with Between padding yielding the best results. The results validate that utilizing higher-dimensional data and employing suitable padding approaches significantly enhance the accuracy of the model. Nonetheless, the findings of this study are limited to the cassava leaf spectral dataset, potentially restricting generalization to other datasets. Future research should explore further integration of advanced signal

processing techniques, such as sliding windows or continuous wavelet transforms, to complement this padded 2D approach. In addition, exploring other padding strategies and assessing other method for converting 1D to 2D data are highly recommended.

References

[1] A. J. Mastin, F. van den Bosch, Y. Bourhis, and S. Parnell, "Epidemiologically-based strategies for the detection of emerging plant pathogens," *Scientific Reports*, vol. 12, no. 1, p. 10 972, 2022.

[2] P. Negi and S. Anand, "Plant disease detection, diagnosis, and management: Recent advances and future perspectives," *Artificial Intelligence and Smart Agriculture: Technology and Applications*, pp. 413–436, 2024.

[3] M. I. S. Mohd Hilmi Tan, M. F. Jamlos, A. F. Omar, F. Dzaharudin, S. Chalermwisutkul, and P. Akkaraekthalin, "Ganoderma boninense disease detection by near-infrared spectroscopy classification: A review," *Sensors*, vol. 21, no. 9, p. 3052, 2021.

[4] Z. Wang and R. Zuo, "An evaluation of convolutional neural networks for lithological mapping based on hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[5] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778–782, 2017.

[6] J. C. M. Ortiz, L. M. H. Carvajal, and V. B. Fernandez, "Detection of significant wavelengths for identifying and classifying fusarium oxysporum during the incubation period and water stress in solanum lycopersicum plants using reflectance spectroscopy," *Journal of Plant Protection Research*, vol. 59, no. 2, 2019.

[7] S. Vélez, E. Barajas, J. A. Rubio, D. Pereira-Obaya, and J. R. Rodríguez-Pérez, "Field-deployed spectroscopy from 350 to 2500 nm: A promising technique for early identification of powdery mildew disease (erysiphe necator) in vineyards," *Agronomy*, vol. 14, no. 3, p. 634, 2024.

[8] S. Atanassova, P. Nikolov, N. Valchev, S. Masheva, and D. Yorgov, "Early detection of powdery mildew (podosphaera xanthii) on cucumber leaves based on visible and near-infrared spectroscopy," in *AIP conference proceedings*, AIP Publishing, vol. 2075, 2019.

[9] A. O. Conrad, W. Li, D.-Y. Lee, G.-L. Wang, L. Rodriguez-Saona, and P. Bonello, "Machine learning-based presymptomatic detection of rice sheath blight using spectral profiles," *Plant Phenomics*, 2020.

[10] M. A. Abdou, "Literature review: Efficient deep neural networks techniques for medical image analysis," *Neural Computing and Applications*, vol. 34, no. 8, pp. 5791–5812, 2022.

[11] G. Yu, B. Ma, J. Chen, X. Li, Y. Li, and C. Li, "Nondestructive identification of pesticide residues on the hami melon surface using deep feature fusion by vis/nir spectroscopy and 1d-cnn," *Journal of Food Process Engineering*, vol. 44, no. 1, e13602, 2021.

[12] J. Zhang, Y. Yang, X. Feng, H. Xu, J. Chen, and Y. He, "Identification of bacterial blight resistant rice seeds using terahertz imaging and hyperspectral imaging combined with convolutional neural network," *Frontiers in Plant Science*, vol. 11, p. 821, 2020.

[13] T. Yan, W. Xu, J. Lin, *et al.*, "Combining multi-dimensional convolutional neural network (cnn) with visualization method for detection of aphis gossypii glover infection in cotton leaves using hyperspectral imaging," *Frontiers in Plant Science*, vol. 12, p. 604 510, 2021.

[14] M. A. Islam, M. Kowal, S. Jia, K. G. Derpanis, and N. D. Bruce, "Position, padding and predictions: A deeper look at position information in cnns," *International Journal of Computer Vision*, pp. 1–22, 2024.

[15] G. Owomugisha, J. Nakatumba-Nabende, J. J. Dhikusooka, E. Taravera, E. Nuwamanya, and E. Mwebaze, "A labeled spectral dataset with cassava disease occurrences using virus titre determination protocol," *Data in brief*, vol. 49, p. 109 387, 2023.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[17] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, 2016, pp. 630–645.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[20] J. P. Bharadiya, "A tutorial on principal component analysis for dimensionality reduction in machine learning," *International Journal of Innovative Science and Research Technology*, vol. 8, no. 5, pp. 2028–2032, 2023.

[21] G. Owomugisha, E. Nuwamanya, J. A. Quinn, M. Biehl, and E. Mwebaze, "Early detection of plant diseases using spectral data," in *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, 2020, pp. 1–6.

A Two-Stage Method for 3D Architecture Wireframe Reconstruction from Airborne LiDAR Point Cloud

Jiahao Zhang, Qi Liu†, Le Hui, Yuchao Dai
School of Electronics and Information & Shaanxi Key Laboratory of Information Acquisition and Processing
Northwestern Polytechnical University, Xi'an, Shaanxi 710129, China

***Abstract*—The 3D roof reconstruction from airborne LiDAR point clouds is an important mid-level vision process in computer vision. Most of the existing three-dimensional roof reconstruction models are trained and tested on simulated roof datasets. Compared with the complex real roof point cloud environment, it is difficult to regress the precise roof wireframe. Based on the Building3D Tallinn dataset, we propose a deep learning-based two-stage approach to reconstruct the roof wireframe from airborne LiDAR point cloud. Initially, the method aims to identify the locations of vertices, followed by predicting the edges connecting these vertices. In the first stage, the primary concept involves offset regression from candidate corner points and utilizing clustering to label the actual corner of the point cloud. The corner points are fully connected between two pairs to execute the second classification of the projected edge, while the Graph Neural Network is employed to extract the feature for label regression, ultimately leading to the final result. In the experiments, we provide several mainstream backbone network feature extraction methods for comparison, and the top-performing method outperformed the best approach reported by the Building3D Tallinn dataset.**

I. Introduction

The utilization of 3D wireframe models has become increasingly prevalent in various fields such as computer vision, architecture, and geospatial analysis. One of the foremost advantages of 3D wireframe models is their efficiency in representing complex geometric shapes with minimal data. Unlike surface or solid models, wireframe models utilize lines and curves to delineate the edges of an object, which significantly reduces computational load and memory usage. Moreover, in the context of airborne LiDAR point cloud data, 3D wireframe models play a pivotal role in efficiently reconstructing architectural features. Their ability to accurately capture the geometric framework of buildings from sparse point cloud data makes them indispensable for applications in remote sensing and geospatial intelligence. This efficiency and accuracy are particularly valuable in scenarios where rapid and large-scale data processing is required. Although the 3D wireframe reconstruction of buildings has high application value and scientific significance, it still remains a challenging problem in computer vision tasks.

Point clouds are a versatile representation of three-dimensional (3D) data, widely utilized in various fields such as computer vision, robotics, and architectural analysis. They provide a rich source of spatial information by capturing the 3D coordinates of numerous points on object surfaces. Despite their advantages, the raw point cloud data often lack structural connectivity, making it challenging to directly infer the underlying geometric structures. Meanwhile, Wireframes offer a simplified yet informative representation by highlighting the essential edges and vertices of objects. They facilitate efficient geometric analysis and visualization, bridging the gap between raw point clouds and structured 3D models. The task of regressing wireframes from point clouds is thus a critical step in converting raw spatial data into meaningful geometric representations. The existing methods for wireframe regression from airborne LiDAR point cloud predominantly rely on synthetic datasets. These datasets, while useful for controlled experiments and method development, often fail to capture the complexity and variability of real-world scenarios. In particular, for applications involving rooftop point cloud datasets, the regularity and simplicity of synthetic data do not adequately reflect the irregular and intricate nature of actual rooftops. This mismatch significantly impairs the applicability and performance of these methods on real-world data.

The regression of wireframe models from rooftop point clouds is a significant task in the fields of computer vision and 3D modeling, with applications ranging from urban planning to augmented reality.In recent years, many traditional data-driven algorithms have been proposed. However, there are two fundamental disadvantages of data-driven algorithms. First, their accuracy is based on the precision of the proposed geometric primitives. Second, their operation sequence is progressive, which easily leads to accumulated errors.

To address the above issues, inspired by Point2Roof[1], we propose a deep learning-based two-stage approach to reconstruct the roof wireframe from airborne LiDAR point cloud. In terms of model training, we have tried several trunk feature extraction networks and adopted a new loss function. Moreover, the labeling strategy for points has been improved. Finally, we have achieved the best results on the largest real airborne point cloud dataset–Building3D[2]. The main contributions of our paper are as follows:

- We propose a new two-stage method for 3D Architecture Wireframe reconstruction from airborne LiDAR point cloud, which first regresses the corners and then regresses the wireframe models based on the corners. This method can be applied to the existing large-scale real datasets.

† Corresponding author (liuqi@nwpu.edu.cn)

This research was funded by the National Natural Science Foundation of China (62271410) and Key Laboratory of Smart Earth, NO. KF2023YB02-08.

- In our experiments, we compared the effectiveness of various current mainstream backbone feature extraction networks for corner feature extraction from point clouds, providing a reference to facilitate future research.

II. Related Work

To our knowledge, there are mainly three methods to reconstruct roof models from point clouds: model-driven methods, data-driven methods, and deep learning-based methods.

A. Model-driven methods

The model-driven method, as a top-down approach, is based on the principle of presupposing buildings as a combination of several simple parameterized shapes. The key issue of the model-driven methods is how to define a suitable roof shape and estimate the optimal parameters of the roof model. Li and Shan [3] proposed a multi-primitive reconstruction method for generating 3D building models from point clouds. This method initially employs the RANSAC plane segmentation algorithm to extract planar patches of each building. Subsequently, a metric method is defined to select potential architectural primitive types from predefined primitives. Finally, parameter estimation is conducted on multiple primitives to reconstruct a topologically consistent model. Although model-driven approaches can maintain the integrity of prediction models as much as possible when dealing with sparse and low-density point clouds, it may be challenging to propose a set of geometric shapes that can cover all buildings when confronted with complex real-world roof models.

B. Data-driven methods

The data-driven method, also known as the bottom-up approach, is generally divided into two stages. In the first stage, a point segmentation algorithm is used to extract major roof primitives from the input LiDAR point clouds. In the second stage, they reconstruct the 3D building model by analyzing the original topological structure. Nan and Wonka [4] proposed a framework named PolyFit for reconstructing manifold and watertight polygon surface models from extracted planar patches. It initially utilizes a RANSAC-based plane segmentation algorithm [5] to generate candidate planes. Subsequently, an optimal subset of planes is selected from these candidates through energy optimization to generate the final polygon surface. The data-driven approach can achieve remarkable results when dealing with complete and high-quality datasets, yet it still encounters two fundamental issues, which were discussed in Section 1.

C. Deep learning-based methods

With the continuous development of deep learning in point cloud processing in recent years, many deep learning based methods for wireframes reconstructing from point clouds have been proposed[1], [6]–[8]. Meanwhile, many methods for extracting deep features of point clouds have been proposed successively[9]–[11]. Such as PointNet++ uses multi-scale feature extraction for point clouds, PointTransformer uses attention mechanism to extract deep features, and SPVCNN uses sparse convolution to voxelize point clouds and then performs 3D convolution to obtain features. These backbone feature extraction networks provide diverse options for wireframe reconstruction. Among the existing methods, Point2Roof uses PointNet++ to extract deep features of point clouds, but it has poor performance in real sparse airborne LiDAR point cloud scenes of roofs. Thus, we propose a two-stage wireframe regression model based on PointTransformer.

III. Method

As we know, a wireframe is consisted of a group of corners and their connection relationship.So we use a two-stage end-to-end deep learning method to regress wireframe from input cloud points. Our method begins by using PointTransformer V2 to extract point-wise features from airborne LiDAR cloud points, classifying them according to a specific label strategy and generating offsets for each point relative to its corresponding corner. We then apply the DBSCAN algorithm to merge corners from these predicted classifications and offsets. The location and features of these corners are refined using point-wise features. In the second stage, we score the connection relationships between points based on corner features, determine connections between points based on these scores, and merge these connections with the corner points to reconstruct a complete wireframe. Next, a detailed introduction to the method will be provided based on three parts.The whole pipeline of the proposed method is shown in Fig. 1.

A. Point-wise feature extracting and cluster prediction

Our method employs PointTransformer V2 as the feature extraction backbone network. Compared with PointNet++, this network can combine global features with local features, making it more capable of acquiring contextual information, and it is highly robust and adaptive to sparse point cloud data.

Label generation.Because the wireframe ground truth cannot be directly used in model training. We supply a new label strategy, which will generate two types of labels. Firstly, the classification labels will determine whether there are true corner points within its spherical radius range, And the offset labels indicates the offset between it and the corresponding corner point.

For the classification labels, if a cloud point has a ground truth corner within its radius will be assigned label 1; otherwise, the label 0. And for the offset labels, if the classification label corresponding to this point is 0, the offset label is [0,0,0]. If it is 1, the offset label Δp is represented as follows:

$$\Delta p = \left[\frac{\mathbf{p}_x - \mathbf{c}_x}{r}, \frac{\mathbf{p}_y - \mathbf{c}_y}{r}, \frac{\mathbf{p}_z - \mathbf{c}_z}{r} \right] \tag{1}$$

Where $(\mathbf{p}_x, \mathbf{p}_y, \mathbf{p}_z)$ and $(\mathbf{c}_x, \mathbf{c}_y, \mathbf{c}_z)$ are the coordinates of the point and its corresponding nearest corner.

Feature extraction.We use PointTransformer v2 as the backbone network for feature extraction. Point Transformer V2 is an improved point cloud processing model that enhances

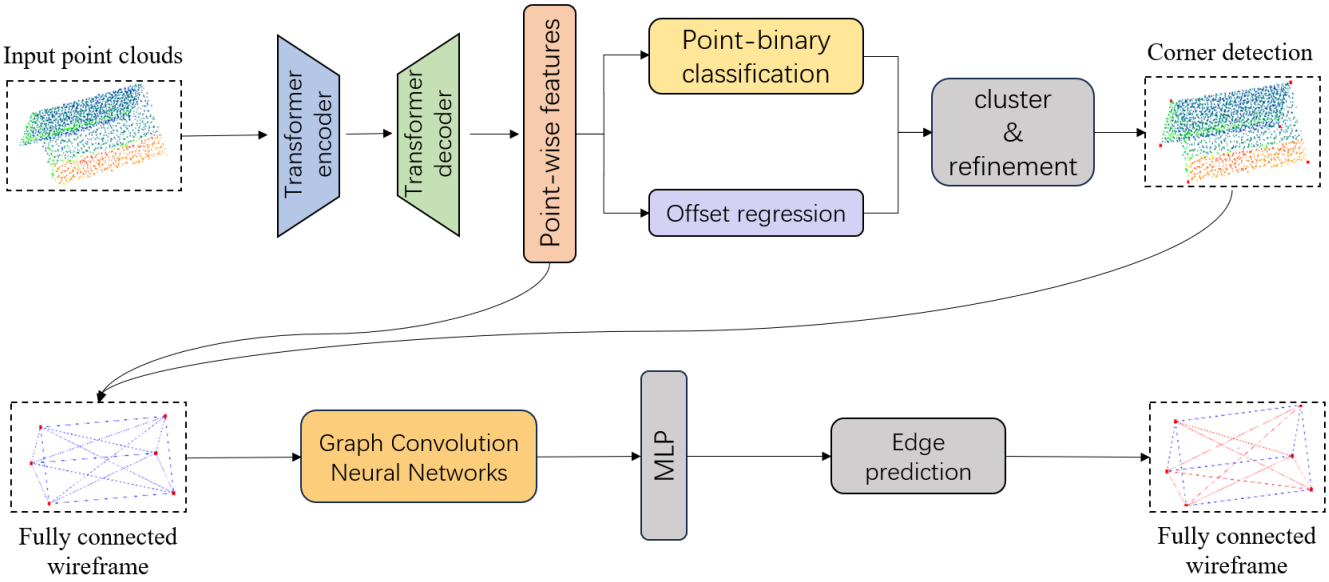


Fig. 1. The framework diagram of our usage method. The upper part is used to detect corners, while the lower part detects edges

the ability to capture local and global geometric information through a more efficient self attention mechanism and optimized feature representation methods, enhancing computational efficiency and robustness. It performs well in point cloud classification tasks and can more accurately recognize and classify complex 3D shapes, making it suitable for multiple fields such as autonomous driving and robot navigation.

In our method, we input the normalized coordinates and RGB information of the point cloud as features into the PointTransformer V2, and obtain deep features with 64 channels, which facilitate point-wise binary classification and offset prediction. The obtained positive points, combined with the predicted offset, will yield the predicted point clusters for subsequent clustering.

B. Candidate point clustering

After passing through the backbone network, we obtained the predicted candidate corners near the corresponding truth inflection points. For these points, we used DBSCAN clustering method to cluster them into one point, which is the predicted corner. However, there is often an offset gap between the predicted corners and the truth points. Therefore, for each clustered corner, an offset relative to the truth corner is labeled. Then, the feature of each point cloud point is weighted, and the prediction score is extracted at multiple scales to extract the final predicted corner feature, thereby obtaining the predicted offset of the predicted point.The visualization of corner point prediction results is shown in Fig. 2.

C. Edge feature extraction

By using the previous stage, the predicted corners and feature information for each point have been obtained. After pairwise connection of each scene's points and comparison with the truth values, the truth labels can be obtained. After passing the obtained features through a Graph convolutional neural network (GCN) module, the features for each edge can be obtained. The GCN model focuses on extracting features from the feature relationships between corners themselves and between corners. It can carefully select salient features for each edge. For each corner features $\mathbf{C} = \{c_1, c_2, \ldots, c_d\}$, we first directly form a feature matrix $\mathbf{H}^{(0)} \in \mathbb{R}^{n \times d}$. Then, we feed the added features into server GCN layers to obtain deeper corner features. Then we further use a feature concatenation operator to extract the final edge features. The detailed process of the GCN module can be expressed as:

$$\begin{aligned} \mathbf{H}^{(l+1)} &= \sigma\left(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \\ \mathbf{F}_{ij} &= \mathbf{H}_i^{(L)} \| \mathbf{H}_j^{(L)} \end{aligned} \tag{2}$$

Among them, σ represent ReLU function. In addition, $\mathbf{W}^{(l)}$ is the weights of the l-th layer of GCN, whereas $\hat{\mathbf{A}}$ is the normalized adjacency matrix. Subsequently, we use an MLP layer to further extract edge features and perform binary classification.

D. Loss function

Our method is end-to-end model training, so the overall loss function is also composed of three parts: L_{pw} for the point-wise classification and offset regression, $L_{cluster}$ for clustering refinement, and Edge Regression loss function L_{edge}

$$L = L_{pw} + L_{cluster} + L_{edge} \tag{3}$$

The backbone network loss function L_{pw} consists of two parts: a point-wise binary classification loss function L_{cls}

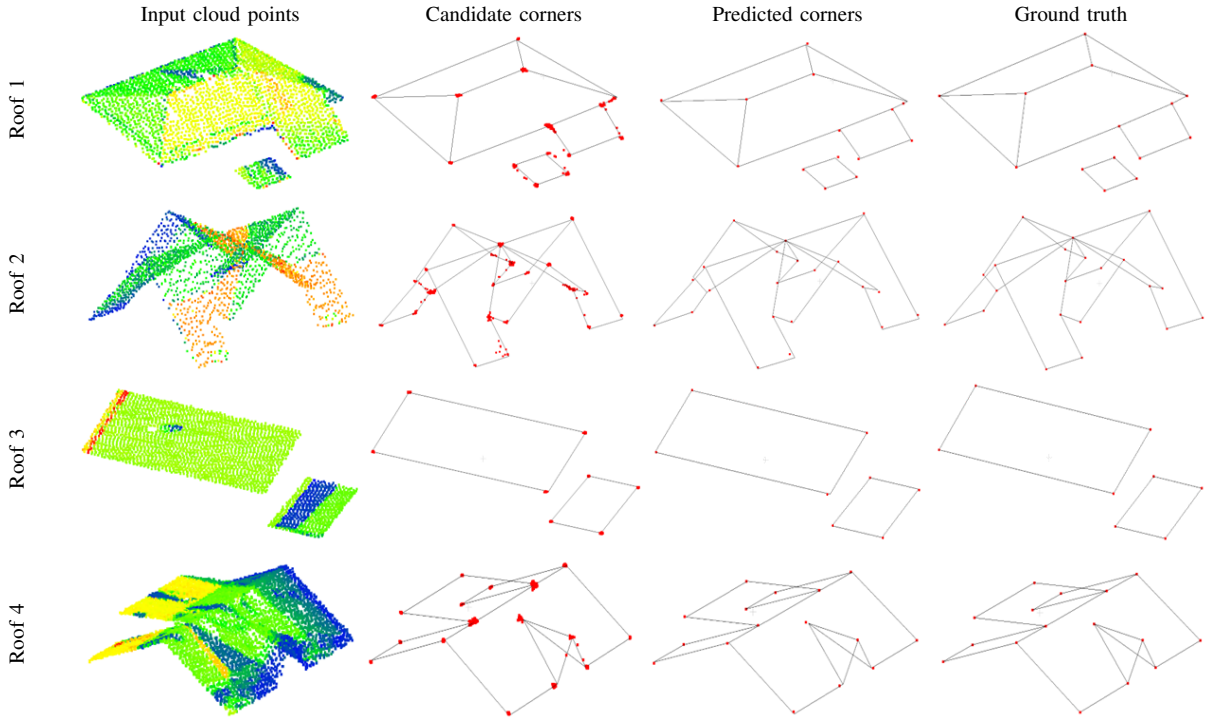


Fig. 2. Demonstrating the corner prediction process in stages.

and an offset regression loss function L_{reg}. According to our investigation, the number of negative points in the label of the model is much greater than the number of positive points. We use the commonly used focus loss function for the classification loss function and use smooth L_1 loss for the offset regression loss function. These two loss functions are defined as:

$$\begin{aligned} L_{pw} &= L_{\text{cls}} + L_{\text{offset}} \\ L_{\text{cls}} &= \sum_{p} L_{\text{bce}}(s_p, l_{\text{cls}}(p)) \cdot \frac{1}{\max(N_{pn}, 1)} \\ L_{\text{offset}} &= \frac{1}{N_{pn}} \sum_{\mathbf{n}} L_{\text{smooth-Ll}}(\Delta\hat{\mathbf{n}}, \Delta_{\text{reg}}(\mathbf{n})) \cdot \eta(s_{\mathbf{n}}) \end{aligned} \tag{4}$$

Among them, s_p and $l_{cls}(p)$ are the predicted classification labels and truth labels for each point n, N_{pn} is the number of labeled positive points, $\Delta\hat{\mathbf{n}}$ and Δ_{reg} are the predicted offset values and true offset values, respectively.$\eta(\cdot) = 1$ if $s_n = 1$; otherwise $\eta(\cdot) = 0$.

For the offset regression loss function of clustering refinement $L_{cluster}$ and the loss function of edge binary classification L_{edge}, they use smooth-L1 loss and focal loss[12], respectively.

$$\begin{aligned} L_{cluster} &= L_{\text{offset}} \\ L_{\text{edge}} &= \sum_{\mathbf{e}} L_{\text{bce}}(s_{\mathbf{e}}, l_{\text{edge}}(\mathbf{e})) \cdot \frac{1}{\max(N_{en}, 1)} \end{aligned} \tag{5}$$

where $s_{\mathbf{e}}$ and l_{edge} refer to the predicted confidence and the label of each edge e.N_{en} is the number of labeled positive edges.

IV. Experiment

A. Implementation details

For the PointTransformer module, the hyperparameters we use are consistent with the Point Transformer V2 Mode 2 provided by PointTransformer v2. After generating features, we added sequential models for classification and offset prediction, and finally output three channel offset prediction and one channel classification prediction. For the corner feature extraction module, we apply PointNet++ MSA module to extract the initial corner features. The multi-scale receiving radii are 0.1 and 0.2. The final number of channel corner features is 128. For the corner refinement module, the network structure is the same as the offset regression head. In the GCN module, we use two GCN layers for each corner feature. Then, apply the FC layer to generate one-dimensional labels. All experiments were conducted on 8 RTX 3090 machines. For self-supervised training, the model is optimized by AdamW with a basic learning rate of 0.0001 and a batch size of 64. The number of training points is 2048. Data denoising includes Gaussian perturbation and random rotation, with probabilities of 0.8 and 0.5, respectively. The learning rate scheduler uses MultiStepLR.

B. Evaluation metric

According to the evaluation indicators requested by the competition on the official website, we can submit our running results for testing to obtain the performance of our method. The introduction of the evaluation indicators are as follows: **Wireframe edit distance** (WED) is a variant of the graph edit distance (GED). It measures how many elementary graph edit

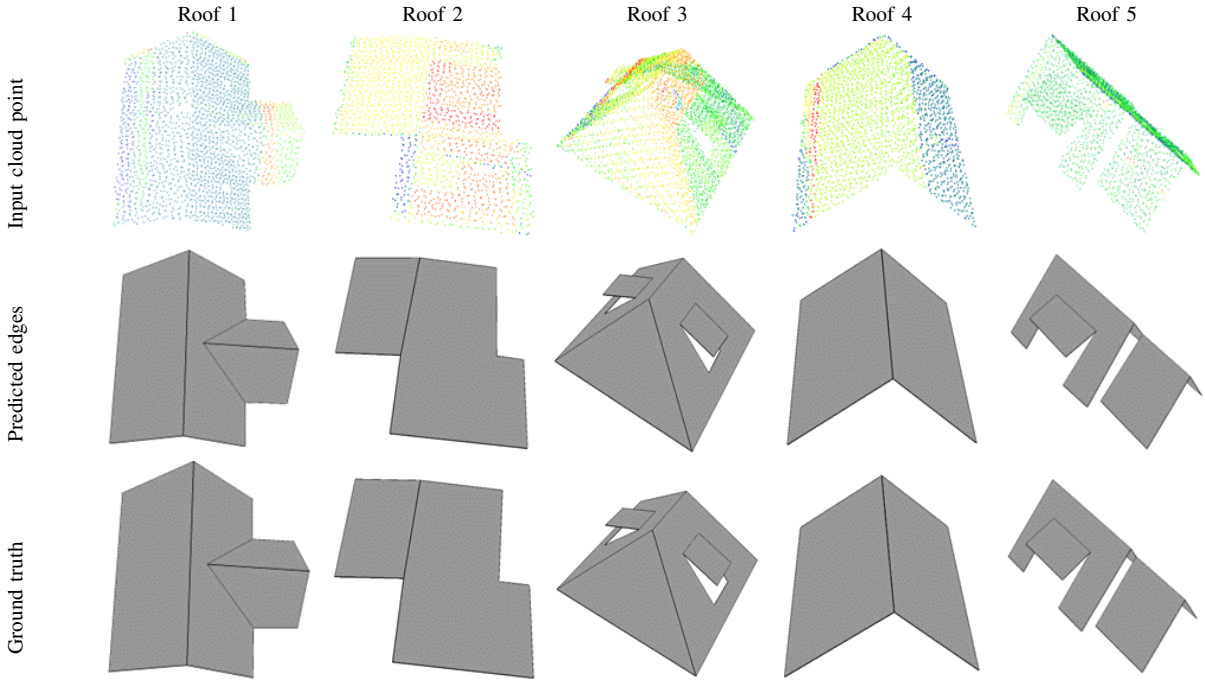


Fig. 3. Edge prediction results

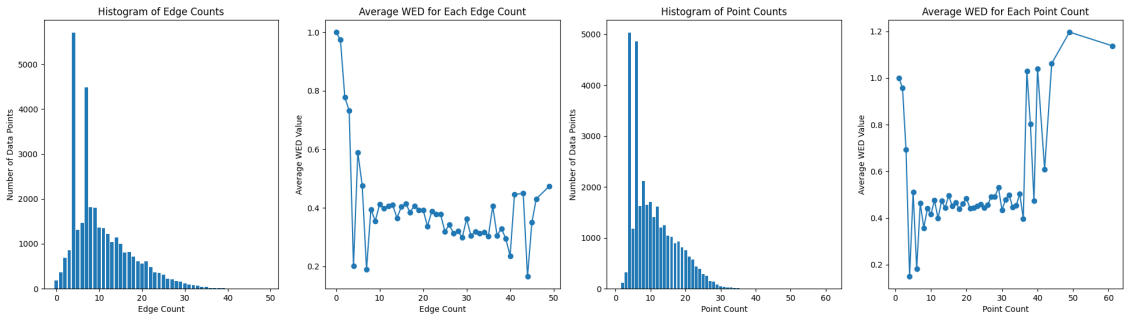


Fig. 4. The horizontal axis of the two graphs on the left represents the predicted number of edges or corners, the vertical axis represents the total number of data under that number, the horizontal axis of the two graphs on the right also represents the predicted number of edges or corners, and the vertical axis represents the average WED of the data under each number.

operators (corner insertion/deletion, edge insertion/deletion, etc.) are needed to transform predicted wireframe to ground truth wireframe, which firstly introduced in the PC2WF[7]. **Average corner offset** (ACO) is the average offsets between predicted corners and ground-truth corners. The smaller offsets indicate better quality of generated models. **Corner precision** (CP), **edge precision** (EP), **corner recall** (CR), and **edge recall** (ER) are calculated through confusion matrix to evaluate the accuracy of corner and edge classification. larger CP and EP values indicate more precise classification of corners and edges, while larger CR and ER values indicate lower rates of missing edge detection.

$$\begin{aligned} \mathrm{CP} &= \frac{C_{TP}}{C_{TP}+C_{FP}}, \quad \mathrm{CR} = \frac{C_{TP}}{C_{TP}+C_{FN}} \\ \mathrm{EP} &= \frac{E_{TP}}{E_{TP}+E_{FP}}, \quad \mathrm{ER} = \frac{E_{TP}}{E_{TP}+E_{FP}} \end{aligned} \tag{6}$$

CF1 and **EF1** represent the F1 score of corners and edges, respectively.

$$CF1 = \frac{2CP \times CR}{CP+CR}, \quad EF1 = \frac{2EP \times ER}{EP+ER} \tag{7}$$

C. Comparative experiments

During the competition, we conducted partial control experiments, and the research results showed that our method has adopted a better strategy. We used the backbone networks of PointNet++ and PointTransformer respectively, and used KNN query and ball query for candidate point annotation methods.

We tested Tallinn's train dataset on the optimal checkpoint model and classified the dataset based on the predicted number of edges or corners. Firstly, we separately counted the number of detected edges or corners, and then calculated the average WED value for each edge or corner. The statistical graph is

TABLE I
COMPARATIVE EXPERIMENTS BETWEEN DIFFERENT METHODS.

Methods	WED	ACO	CP	CR	CF1	EP	ER	EF1
PC2WF	-	0.45	0.24	0.13	0.16	0.02	0.14	0.04
Building3D	-	0.29	0.90	0.53	0.66	0.88	0.23	0.36
Ours(SPVCNN)	0.783	0.240	0.897	0.771	0.829	0.764	0.287	0.417
Point2Roof	0.447	0.282	0.938	0.634	0.756	0.865	0.423	0.568
Ours(PointTransformer)	**0.345**	**0.235**	**0.967**	**0.69**	**0.80**5	**0.875**	**0.556**	**0.68**

TABLE II
ABLATION STUDY OF THE GCN MODULE.

Edge extra model	WED	ACO	CP	CR	EP	ER
Concatenation	0.465	0.265	0.953	0.653	0.735	0.425
GCN	**0.345**	**0.235**	**0.967**	**0.69**	**0.875**	**0.68**

shown in Fig .3. According to the experimental results, it can be found that we will get better results when detecting an intermediate number of corners or edges. Errors only occur when facing poor quality point clouds of some roofs. The results of comparative experiments between different methods are shown in TABLE I, and the edge prediction results are partly shown in Fig. 3.

D. Ablation study

To validate the effectiveness of our graph convolutional neural network, we compared it with the concatenation layer. As observed, when using the graph convolutional neural network and the concatenation layer respectively, their performance on the indicators CP and CR were basically the same. However, the graph convolutional neural network demonstrated better results in terms of EP, ER and WED. The experimental results indicate that the proposed edge prediction network can effectively enhance the effectiveness of edge prediction.The results of the ablation study are shown in TABLE II.

V. CONCLUSIONS

To directly reconstruct the wireframe structure from airborne LiDAR point clouds, we devised a two-stage roof reconstruction approach based on deep learning. We divided the task into two parts: corner prediction and edge prediction. Furthermore, our method can directly generate models applicable to large-scale real-world roof point cloud datasets. Experimental results demonstrate that our method can accurately predict the positions of corners and edges on the Tallinn dataset of build3D, and our current approach significantly outperforms traditional methods trained on simulated datasets.

REFERENCES

[1] "Point2roof: End-to-end 3d building roof modeling from airborne lidar point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 193, pp. 17–28, 2022.

[2] R. Wang, S. Huang, and H. Yang, "Building3d: An urban-scale dataset and benchmarks for learning roof structures from point clouds," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20 019–20 029, 2023.

[3] Z. Li and J. Shan, "Ransac-based multi primitive building reconstruction from 3d point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 185, pp. 247–260, 2022, ISSN: 0924-2716.

[4] L. Nan and P. Wonka, "Polyfit: Polygonal surface reconstruction from point clouds," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[5] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," *Computer Graphics Forum*, vol. 26, no. 2, pp. 214–226, 2007.

[6] S. Huang, R. Wang, B. Guo, and H. Yang, *Pbwr: Parametric building wireframe reconstruction from aerial lidar point clouds*, 2023. arXiv: 2311.12062 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2311.12062.

[7] Y. Liu, S. D'aronco, K. Schindler, and J. D. Wegner, "Pc2wf: 3d wireframe reconstruction from raw point clouds," *ArXiv*, vol. abs/2103.02766, 2021.

[8] T. Jiang, Y. Wang, Z. Zhang, *et al.*, "Extracting 3-d structural lines of building from als point clouds using graph neural network embedded with corner information," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–28, 2023.

[9] C. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Neural Information Processing Systems*, 2017.

[10] X. Wu, Y. Lao, L. Jiang, X. Liu, and H. Zhao, "Point transformer v2: Grouped vector attention and partition-based pooling," *ArXiv*, vol. abs/2210.05666, 2022.

[11] H. Tang, Z. Liu, S. Zhao, *et al.*, "Searching efficient 3d architectures with sparse point-voxel convolution," in *European Conference on Computer Vision (ECCV)*, 2020.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

Performance Improvement of Single Plane-Wave Imaging Using U-Net and Discrete Wavelet Transform

Hiromi Shidara* Kanta Miura* Takuro Ishii† Koichi Ito* Takafumi Aoki* Yoshifumi Saijo‡ and Jun Ohmiya§
* Graduate School of Information Sciences, Tohoku University, Japan
E-mail: shidara@aoki.ecei.tohoku.ac.jp
† Frontier Research Institute for Interdisciplinary Sciences, Tohoku University, Japan
‡ Graduate School of Biomedical Engineering, Tohoku University, Japan
§ Konica Minolta, Inc., Japan

***Abstract*—Single Plane-Wave Imaging (SPWI), which transmits a single plane wave, can acquire ultrasound images at more than 1,000 fps, although it has poor lateral resolution and contrast. Some methods have been proposed to improve the quality of ultrasound images acquired by SPWI using deep learning, however, the quality is lower than that of compounded images, which are composed of multiple SPWI images. In addition, the RF signal is used as an input, which is computationally expensive. In this paper, we propose a method to improve the performance of SPWI using U-Net and the Discrete Wavelet Transform (DWT). The proposed method uses In-phase and Quadrature (IQ) data as the input and output of U-Net and loss functions that take into account the characteristics of the RF signal to improve the quality of images, and also uses IQ data after DWT to reduce the computational complexity and the inference time. Through a set of experiments using our ultrasound image dataset, we demonstrate the effectiveness of the proposed method.**

I. Introduction

Ultrasound imaging can acquire images in real time with a small and portable device and with low impact on the human body compared to Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Therefore, ultrasound imaging is utilized for analyzing and evaluating the cardiovascular system and observing muscular tissue [1]. In standard ultrasound imaging, ultrasound images are obtained by sequentially transmitting more than 100 focused ultrasound beams, resulting in speeds limited to a few tens of fps [2], [3]. Since detailed tissue observation [4]–[7] and blood flow evaluation [8]–[11] require ultrasound images at high frame rates, Plane-Wave Imaging (PWI), which uses plane waves to acquire ultrasound images, has been used [12].

PWI can acquire images at a higher frame rate than standard ultrasound imaging by transmitting plane waves of the same width as the field of view several times. Single Plane-Wave Imaging (SPWI), which acquires images by transmitting a single plane wave, can archive speeds of more than 1,000 fps, while there is a problem that the lateral resolution and contrast of the images are degraded [13]. Coherent Plane-Wave-Compounding (CPWC) is applied to SPWI to address the above problems [14]. CPWC acquires a compound image by transmitting a plane wave multiple times in different directions while controlling the delay of the ultrasound elements, and then coherently adding the received Radio Frequency (RF) signals. CPWC can improve the lateral resolution and contrast of images, although the frame rate is lower than that of SPWI due to the need to transmit a plane wave multiple times. To address the trade-off between frame rate and image quality in CPWI, several methods have been proposed to improve the quality of ultrasound images obtained by SPWI comparable to that of compound images by deep learning [15], [16]. Li et al. [15] proposed a method using U-Net [17], which is commonly used in image segmentation and image generation. U-Net in this method consists of VGG-13 [18] as an encoder and a network of four inverse convolution layers as a decoder. This U-Net is trained to minimize the L1 loss between the compound images obtained by CPWC and the ultrasound images generated by U-Net. Since the L1 loss between images is used, the high-frequency components of the compound image cannot be represented sufficiently, resulting in the generation of blurred ultrasound images. Perdios et al. [16] also proposed a U-Net based method. Mean Signed Logarithmic Absolute Error (MSLAE) is used as a loss function to take into account the wide dynamic range of ultrasound images and to preserve the characteristics of the RF signal. On the other hand, it is necessary to set an appropriate threshold to train U-Net with MSLAE, however, the details of the threshold determination are not available, making the method unreproducible. Both methods use RF signals as input, which have higher resolution than images, and are computationally expensive, therefore, the real-time capability of ultrasound imaging may be compromised.

To address the above problems, we propose a method to improve the quality of ultrasound images acquired by SPWI using U-Net and to reduce the computational cost by using Discrete Wavelet Transform (DWT). The proposed method uses In-phase and Quadrature (IQ) data, which consists of the real part of the beamformed RF signal and the imaginary part of the Hilbert transform of the RF signal, as the input and output of U-Net to represent the high frequency component

of the ultrasound signals. Since the effect of the point spread of the envelope signal, which is the amplitude of the IQ data, increases with the transmission distance of the ultrasound [19], we introduce loss functions that take into account the effect of the point spread, which is close to the envelope signal, and the frequency characteristics of the RF signals. We also reduce the computational complexity and inference time by inputting the IQ data after DWT into U-Net. Since 2D data can be decomposed into four sub-bands (LL, LH, HL, HH) by DWT, DWT has been applied to image coding and texture identification based on multiple resolution analysis [20]. In the field of ultrasound imaging, DWT has been also used to reduce speckle noise in ultrasound images [21] and to automatic diagnosis of Graves' disease [22]. Since DWT can reduce the data size by half while keeping the information in the 2-D data, the proposed method can reduce the computational complexity and inference time by inputting the IQ data after DWT. The effectiveness of the proposed method is demonstrated through a set of experiments using the ultrasound image dataset created by the authors.

II. Method

In this section, we describe the details of the network architecture and loss functions for the proposed method to improve the quality of ultrasound images acquired by SWPI.

A. Network Architecture

The proposed method employs an encoder-decoder model based on U-Net [17] as shown in Fig. 1. The proposed method replaces the encoder of U-Net with ResNet-34 [23] and uses a decoder consisting of inverse convolution layers as in U-Net. To clarify the difference between the original U-Net and U-Net used in the proposed method, U-Net used in the proposed method is denoted as ResU-Net in the following. We use IQ data as the input and output of ResU-Net instead of ultrasound images to take into account the high-frequency components of ultrasound images. The use of IQ data makes it possible to take into account point the effect of point spread of the envelope signal acquired by CPWC and frequency characteristics of the RF signal. For IQ data with $H \times W$, 2D DWT is applied to each of the real and imaginary parts, which is decomposed into four sub-bands, i.e.,LL, LH, HL, and HH, and the size of IQ data becomes $\frac{H}{2} \times \frac{W}{2}$ for each sub-band. The four sub-bands are concatenated in the channel direction and the data with $8 \times \frac{H}{2} \times \frac{W}{2}$ is used as the input to ResU-Net. The IQ data is output by applying 2D IDWT to the data with $8 \times \frac{H}{2} \frac{W}{2}$ obtained at the final layer of ResU-Net and reconstructing it to the original size of $2 \times H \times$ W.

B. Loss Functions

The training of the proposed method employs the three loss functions based on the IQ data, the envelope signal, and the frequency characteristics. In the calculation of the loss functions, the IQ data after compounding are used as the ground truth. In the following, we denote $r(n_1, n_2)$ and $i(n_1, n_2)$ as the real and imaginary parts of the ground truth IQ data of an $N_1 \times N_2$ matrix, respectively. Let $\hat{r}(n_1, n_2)$ and $\hat{i}(n_1, n_2)$ denote the real and imaginary parts of IQ data output by ResU-Net, respectively. The loss based on IQ data, L_{IQ}, is defined by

$$L_{\mathrm{IQ}} = \frac{1}{N_1 N_2} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \Big\{ |\hat{r}(n_1, n_2) - r(n_1, n_2)| + \Big| \hat{i}(n_1, n_2) - i(n_1, n_2) \Big| \Big\}. \tag{1}$$

The loss based on the envelope signal, L_{Env}, is defined by

$$L_{\mathrm{Env}} = \frac{1}{N_1 N_2} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \left\{ \left| \sqrt{\hat{r}(n_1, n_2)^2 + \hat{i}(n_1, n_2)^2} - \sqrt{r(n_1, n_2)^2 + i(n_1, n_2)^2} \right| \right\}. \tag{2}$$

The use of L_{Env} makes it possible to achieve a similar effect of point spread to the envelope signal obtained by CPWC. Let $A(n_1, k_2)$ denote the amplitude and $\theta(n_1, k_2)$ denote the phase obtained by 1D discrete Fourier transform of the real part of the IQ data in the column direction, i.e., n_2, respectively. The loss based on the frequency characteristics, L_{Fourier}, is defined by

$$L_{\mathrm{Fourier}} = \frac{1}{N_1 K_2} \sum_{n_1=0}^{N_1-1} \sum_{k_2=0}^{K_2-1} \Big\{ \lambda_{\mathrm{A}} \left| \hat{A}(n_1, k_2) - A(n_1, k_2) \right| + \lambda_\theta \left(1 - \cos\left(\hat{\theta}(n_1, k_2) - \theta(n_1, k_2) \right) \right) \Big\}, \tag{3}$$

where $A(n_1, k_2)$ and $\theta(n_1, k_2)$ are the amplitude and phase spectrum calculated from the real part of IQ data after compounding, respectively, $\hat{A}(n_1, k_2)$ and $\hat{\theta}(n_1, k_2)$ are the amplitude and phase spectrum calculated from the real part of IQ data output by ResU-Net, respectively, k_2 is the index of discrete frequency, and λ_{A} and λ_θ are hyperparameters that adjust the balance between amplitude and phase spectrum, respectively. The use of L_{Fourier} makes it possible to close the frequency spectrum of the output by ResU-Net to the frequency spectrum of the ground truth of the RF signal. The total loss function, L, used in training is defined by

$$L = L_{\mathrm{IQ}} + \lambda_{\mathrm{Env}} L_{\mathrm{Env}} + \lambda_{\mathrm{Fourier}} L_{\mathrm{Fourier}}, \tag{4}$$

where λ_{Env} and $\lambda_{\mathrm{Fourier}}$ are hyperparameters that adjust the balance among loss functions.

III. Experiments and Discussion

This section describes a performance evaluation of the proposed method for improving the quality of SPWI.

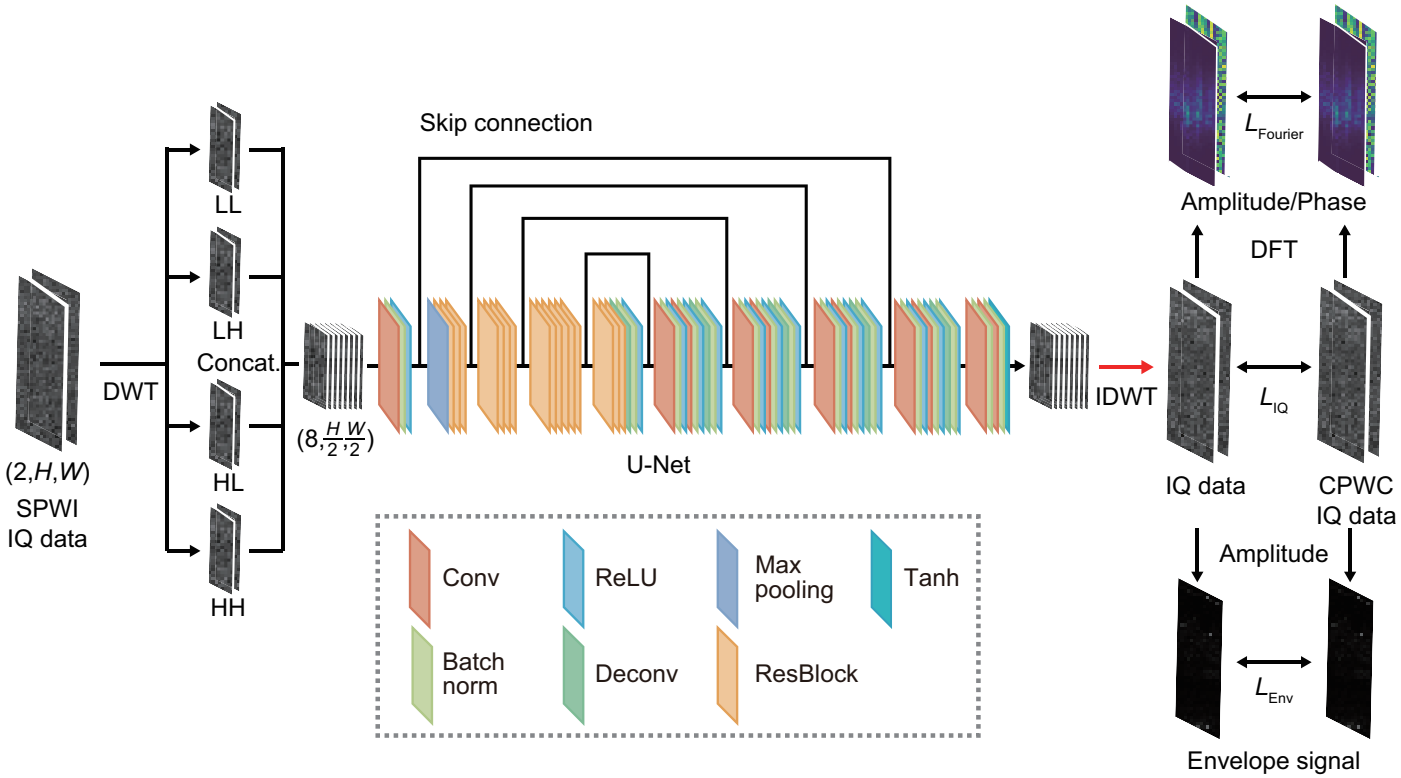


Fig. 1. Overview of the proposed method and loss functions used in training.

TABLE I
CONFIGURATION OF SPWI/CPWC PAIRS FOR TRAINING, VALIDATION, AND TEST IN OUR DATASET.

Target	Train	Val	Test	Total
Breast ultrasound examination phantom	2,800	400	800	4,000
Image quality assurance phantom	400	—	1,000	1,400
Subject #1	500	—	—	500
Subject #2	—	200	200	400
Subject #3	—	—	100	100
Total	3,700	600	2,100	6,400

A. Dataset

In the experiments, we use our SPWI/CPWC dataset[1]. This dataset contains 6,400 pairs consisting of IQ data acquired by SPWI and IQ data acquired by CPWC as shown in Table I. The targets are the breast ultrasound examination phantom (BP), the image quality assurance phantom (QAP), and the cervical regions of three healthy subjects (Subject). The IQ data of CPWC used as ground truth in the experiments are obtained by coherently adding 75 beamformed IQ data acquired by varying the transmission angle of the plane wave in the range of $-16 \sim 16$ degrees.

B. Experimental Condition

The dataset is separated into 3,700 pairs for training, 600 pairs for validation, and 2,100 pairs for test as shown in Table I. The IQ data acquired by SPWI is padded to 384×768 and normalized to have each element in the range $[-1, 1]$. We set $\lambda_{\mathrm{A}} = 0.014$, $\lambda_{\theta} = 0.026$, $\lambda_{\mathrm{Env}} = 0.560$, and $\lambda_{\mathrm{Fourier}} = 0.440$, which are optimized by Optuna [24]. AdamW [25] is used as the optimizer, the learning rate is 1e-4, the batch size is 16, and the number of epochs is 300.

We compare the performance of the proposed method ("ResU-Net") with that of the conventional method [15] ("Li et al."), the proposed method with the original U-Net ("U-Net"), and the proposed method without DWT ("ResU-Net w/o DWT"). Since the code for the conventional method is not publicly available, we implemented it according to [15] and trained it under the same experimental conditions as the proposed method We employ "U-Net" in the Pytorch implementation[2] and evaluate it in the same environment as the other methods. The experimental conditions for "U-Net" are the same as for the proposed method, except for the encoder architecture. The difference between "ResU-Net w/o DWT" and "ResU-Net" is whether DWT is applied to IQ data or not. We use the discrete Haar wavelet transform as DWT in the experiments. All methods are implemented using Pytorch 1.13.1 and evaluated on an AMD EPYC 7502 32-core Processor and an A100 (80GB).

C. Evaluation Metrics

In this experiment, we evaluate image quality, lateral resolution, and contrast to compare the quality improvement of each

[1]https://github.com/gsisaoki/Improvement_of_Ultrasound_Image_Quality

[2]https://pytorch.org/hub/mateuszbuda_brain-segmentation-pytorch_unet/

TABLE II
EXPERIMENTAL RESULTS FOR EVALUATING IMAGE QUALITY, WHERE THE VALUES IN BOLD INDICATE THE HIGHEST VALUES FOR EACH METRIC.

Method	PSNR [dB] ↑			SSIM ↑			LPIPS ↓		
	BP	QAP	Subject	BP	QAP	Subject	BP	QAP	Subject
SPWI (Input)	17.25	16.05	14.74	0.217	0.218	0.222	0.437	0.436	0.447
Li et al. [15]	**20.28**	**19.03**	**18.93**	0.310	0.313	**0.260**	0.635	0.590	0.635
U-Net [17]	14.50	13.13	14.80	0.115	0.086	0.110	0.625	0.624	0.628
ResU-Net w/o DWT	17.17	17.62	16.78	**0.314**	**0.351**	0.243	**0.461**	**0.430**	0.485
ResU-Net	17.68	17.76	17.24	0.301	0.336	0.223	0.469	0.435	**0.459**

TABLE III
EXPERIMENTAL RESULTS FOR EVALUATING LATERAL RESOLUTION AND CONTRAST, WHERE THE VALUES IN BOLD INDICATE THE HIGHEST VALUES FOR EACH METRIC.

Method	FWHM_t [%] ↑	ΔFWHM [μm] ↑	CR [dB] ↑	GCNR ↑
SPWI (Input)	76.38	—	4.95	0.379
CPWC (GT)	93.17	—	7.34	0.521
Li et al. [15]	83.27	-36.28	6.31	**0.612**
U-Net [17]	92.32	**217.841**	4.41	0.410
ResU-Net w/o DWT	**92.83**	133.375	**8.35**	0.477
ResU-Net	89.98	93.17	7.54	0.445

TABLE IV
COMPARISON OF FLOPS, THE NUMBER OF PARAMETERS, AND INFERENCE TIME OF EACH METHOD.

Method	FLOPS	# of parameters	Inference time [ms] GPU	CPU
Li et al. [15]	139.57B	13.40M	54.64	487.5
U-Net [17]	217.11B	31.04M	55.45	740.8
ResU-Net w/o DWT	32.49B	24.35M	54.49	183.0
ResU-Net	8.53B	24.37M	55.09	53.12

method. For evaluating the image quality, we employ Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) [26], and Learned Perceptual Image Patch Similarity (LPIPS) [27]. Higher values for PSNR and SSIM indicate higher image quality, while lower values for LPIPS indicate higher image quality. LPIPS is known as an image quality evaluation metric that is closer to human perception compared to PSNR and SSIM [27]. For evaluating the lateral resolution, we employ FWHM_t and $\Delta\mathrm{FWHM}$, which are based on the Full Width at Half Maximum (FWHM). For the 4,949 wire targets in QAP, FWHM is the width of the pixel that is -6dB from the peak value of the point spread function in the horizontal direction. FWHM_t is the ratio of the number of wire targets whose FWHM is below the threshold t [μm] out of 4,949 wire targets. In the experiments, we set $t = 1,580\mu$m to take into account the width of the adjacent wire targets. $\Delta\mathrm{FWHM}$ is the difference of FWHM between the single plane-wave image and the generated image for a wire target whose FWHM is less than the threshold t [μm] in the single plane-wave image, the compound image, and the generated image. For evaluating the contrast, we employ Contrast Ratio (CR) and Generalized Contrast-to-Noise Ratio (GCNR) [28]. In the experiments, CR and GCNR are calculated for the surrounding regions of 813 grayscale targets in QAP, and their average are used for evaluation.

D. Results and Discussion

Table II shows the results of quantitative evaluation of image quality for each method. The conventional method has the highest PSNR. Since PSNR evaluates image quality based on the root-mean-square of the differences in the pixel areas of the generated and compound images, the conventional method, which is trained using the L1 loss between the images, has the advantage. For SSIM, ResU-Net w/o DWT is the highest for BP and QAP, and the conventional method is the highest for Subject. For LPIPS, ResU-Net w/o DWT is the lowest for BP and QAP, and ResU-Net is the lowest for Subject. Comparing U-Net and ResU-Net w/o DWT, ResNet is suitable as an encoder for improving the quality of SPWI, since ResU-Net w/o DWT is better on all metrics. ResU-Net w/o DWT and ResU-Net exhibit similar performance on all metrics. Considering that LPIPS is the closest to human subjective evaluation among the three evaluation metrics [27], ResU-Net w/o DWT and ResU-Net can produce ultrasound images with quality close to that of compound images. Table III shows the results of the quantitative evaluation of the lateral resolution and contrast for each method. The conventional method is lower than the other methods in all metrics except for GCNR. ResU-Net w/o DWT performs slightly better than ResU-Net on all metrics. Table IV shows FLOPS, the number of parameters, and the inference time on GPU and CPU for each method. ResU-Net w/o DWT and ResU-Net achieve fewer FLOPs than the other methods. ResU-Net reduces FLOPS to about $\frac{1}{4}$ compared to ResU-Net w/o DWT. The inference time on GPU is similar for all methods, while ResU-Net achieves the shortest inference time on CPU at 53.12 ms. ResU-Net is particularly suitable for general ultrasound systems, which are often configured with a CPU. From the above results, we have demonstrated that the loss functions that take into account the characteristics of the RF signal can produce high-quality ultrasound images, and that the use of DWT can reduce the computational complexity and inference time.

IV. CONCLUSION

We proposed a method for improving the performance of SPWI using U-Net and DTW. The proposed method uses IQ data as the input and output of U-Net and to represent the high frequency component of the ultrasound signals and loss functions that take into account the effect of the point spread and the frequency characteristics of the RF signals. We also reduce the computational complexity and inference time by inputting the IQ data after DWT into U-Net. We demonstrated

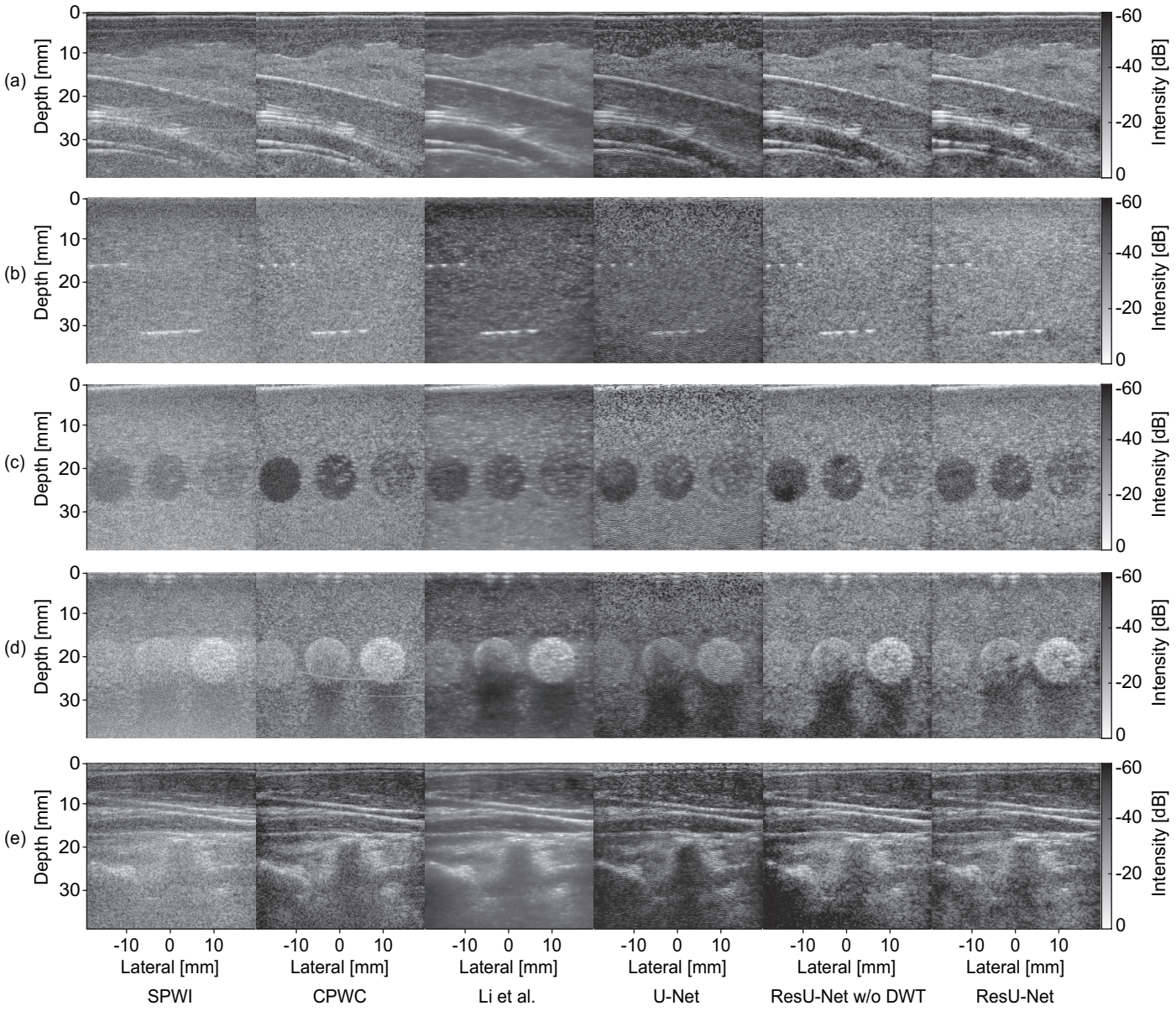


Fig. 2. Examples of ultrasound images generated by each method: (a) BP, (b) wire target in QAP, (c) grayscale target in QAP, (d) grayscale target in QAP, and (e) Subject.

the effectiveness of the proposed method through a set of experiments using our image dataset. In the future work, we will apply the ultrasound images whose quality is improved by the proposed method to the analysis of dynamic organs such as the heart.

REFERENCES

[1] T. L. Szabo, *Diagnostic Ultrasound Imaging: Inside Out*. Academic Press, 2004.

[2] P. Hoskins, K. Martin, and A. Thrush, *Diagnostic Ultrasound: Physics and Equipment*. CRC Press, 2019.

[3] J.-Y. Lu, H. Zou, and J. F. Greenleaf, "Biomedical ultrasound beam forming," *Ultrasound in Medicine and Biology*, vol. 20, no. 5, pp. 403–428, Apr. 1994.

[4] J. Bercoff, M. Tanter, and M. Fink, "Supersonic shear imaging: A new technique for soft tissue elasticity mapping," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 51, no. 4, pp. 396–409, Apr. 2004.

[5] M. Fink and M. Tanter, "Multiwave imaging and super resolution," *Physics Today*, vol. 63, no. 2, pp. 28–33, Feb. 2010.

[6] E. Macé, G. Montaldo, I. Cohen, M. Baulac, M. Fink, and M. Tanter, "Functional ultrasound imaging of the brain," *Nature Methods*, vol. 8, pp. 662–664, Jul. 2011.

[7] T. Deffieux, C. Demené, and M. Tanter, "Functional ultrasound imaging: A new imaging modality for neu-

roscience," *Neuroscience*, vol. 474, pp. 110–121, Oct. 2021.

[8] J. Udesen, F. Gran, K. L. Hansen, J. A. Jensen, C. Thomsen, and M. B. Nielsen, "High frame-rate blood vector velocity imaging using plane waves: Simulations and preliminary experiments," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 55, no. 8, pp. 1729–1743, Aug. 2008.

[9] K. L. Hansen, J. Udesen, F. Gran, J. A. Jensen, and M. Bachmann Nielsen, "In-vivo examples of flow patterns with the fast vector velocity ultrasound method," *Ultraschall in der Medizin*, vol. 30, no. 5, pp. 471–477, Sep. 2009.

[10] J. Bercoff, G. Montaldo, T. Loupas, *et al.*, "Ultrafast compound doppler imaging: Providing full blood flow characterization," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 58, no. 1, pp. 134–147, Jan. 2011.

[11] I. K. Elroll, A. Swillens, P. Segers, T. Dahl, H. Torp, and L. Lovstakken, "Simultaneous quantification of flow and tissue velocities based on multi-angle plane wave imaging," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 60, no. 4, pp. 727–738, Apr. 2013.

[12] O. Coutere, M. Fink, and M. Tanter, "Ultrasound contrast plane wave imaging," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 59, no. 12, pp. 2676–2683, Dec. 2012.

[13] M. Tanter and M. Fink, "Ultrafast imaging in biomedical ultrasound," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 61, no. 1, pp. 102–119, Jan. 2014.

[14] G. Montaldo, M. Tanter, J. Bercoff, N. Benech, and M. Fink, "Coherent plane-wave compounding for very high frame rate ultrasonography and transient elastography," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 56, no. 3, pp. 489–506, Mar. 2009.

[15] Z. Li, A. Wiacek, and M. A. L. Bell, "Beamforming with deep learning from single plane wave RF data," *Proc. IEEE Int'l Ultrasonics Symp.*, pp. 1–4, Sep. 2020.

[16] D. Perdios, M. Vonlanthen, F. Martinez, M. Arditi, and J.-P. Thiran, "CNN-based image reconstruction method for ultrafast ultrasound imaging," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 4, pp. 1154–1168, Nov. 2022.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, vol. 9351, pp. 234–241, Oct. 2015.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proc. Int'l. Conf. Learning Representations*, pp. 1–14, May 2015.

[19] C. Chen, H. H. G. Hansen, G. A. G. M. Hendriks, J. Menssen, J.-Y. Lu, and C. L. de Korte, "Point spread function formation in plane-wave imaging: A theoretical approximation in fourier migration," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 2, pp. 296–307, Feb. 2020.

[20] T. Guo, T. Zhang, E. Lim, M. Lopez-Benitez, F. Ma, and L. Yu, "A review of wavelet analysis and its applications: Challenges and opportunities," *IEEE Access*, vol. 10, pp. 58 869–58 903, 2022.

[21] P. Namsopa, S. Koonkarnkhai, P. Kovintavewat, H. Dinsakul, and S. Suwansawang, "A study and exploration of discrete wavelet transform for speckle noise reduction in ultrasound images," *Proc. Int'l Tecnical Conf. Circuits/Systems, Computers, and Communications*, pp. 1–5, Jun. 2023.

[22] S. .-. Yu, J. .-. He, Y. Pan, Y. Wang, J. Wen, and R. Yan, "The application of texture classification network with hybrid adaptive wavelet in graves' disease ultrasound diagnosis," *Proc. Int'l Conf. Wavelet Analysis and Pattern Recognition*, pp. 38–43, Jul. 2023.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770–778, Jun. 2016.

[24] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *Proc. ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, pp. 2623–2631, Jul. 2019.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *Proc. Int'l Conf. Learning Representations*, pp. 1–10, 2019.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[27] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 586–595, Jun. 2018.

[28] A. Rodriguez-Molares, O. M. H. Rindal, J. D'hooge, *et al.*, "The generalized contrast-to-noise ratio: A formal definition for lesion detectability," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 4, pp. 745–759, Apr. 2020.

Ensemble learning based head-related transfer function personalization using anthropometric features

Yih-Liang Shen†, and Tai-Shi Chi*
* Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan
† Institute of Communications Engineering, National Yang Ming Chiao Tung University, Taiwan
E-mail: dennis831209@gmail.com; tschi@nycu.edu.tw

***Abstract*—An ensemble learning based model to predict the logarithmic magnitude response of the head-related transfer function (HRTF) using anthropometric features is proposed. Subjects are first clustered based on relevant anthropometric features, then the ensemble learning algorithm is used on clustered results for predicting the log-magnitude response of HRTF. The model contains two levels of deep neural networks (DNNs). The first-level DNNs predict log-magnitude HRTFs in clustered groups, and the second-level DNN integrates estimates produced by the first-level DNNs. Experimental results show the proposed model outperforms compared DNN models in terms of the log-spectral distortion (LSD) measure with greater stability.**

I. Introduction

In virtual-reality applications, it is essential to accurately present spatial sounds to the user. The sound waves from a sound source interact with the head, torso, and pinna to cause scattering and diffraction. The acoustic path from the sound source to an ear can be considered an acoustic channel and is characterized by its impulse response, which is referred to as the head-related impulse response (HRIR). The frequency representation of the HRIR is called the head-related transfer function (HRTF). Using HRTFs, a sound can be spatialized as it comes from a specific location. HRTFs contain binaural cues such as interaural time difference (ITD) and interaural level difference (ILD), which are primarily used by humans to localize the sound source [1]. Since HRTFs are personalized measures, using HRTFs of a different person to spatialize the sound can result in confusion about the location of the sound [2][3]. However, directly measuring an individual's HRTFs, which involves complex, time-consuming, and expensive procedures, is impractical.

Many methods have been proposed to estimate HRTFs. For example, the model-based methods divide the acoustic path into several blocks and develop a physical model for each block [4]. In addition to developing analytic geometry models, manipulating HRTFs in anthropometric dimensions was also attempted due to the high correlation between body characteristics and HRTFs. For instance, the frequency scaling of HRTFs was investigated in [5] to find the optimal scaling factor between different subjects by minimizing the difference of HRTFs expressed in anthropometric dimensions. Some studies also suggest that approximated HRTFs can be obtained from available HRTF datasets by looking for people with similar body characteristics [6]. Recently, researchers use a three-dimensional ear model and simulated HRTFs to relate the shape of pinna to HRTFs [7]. Signal processing based methods were also developed for estimating HRTFs. For example, active sensory tuning (AST) was used to synthesize HRTFs of the target person from a generic HRTF by adjusting poles and zeros of the transfer functions through several optimization steps [8][9]. Principal component analysis (PCA) and regression analysis were combined for estimating HRTFs [10][11]. In recent years, deep neural networks (DNNs) were also proposed to estimate HRIRs using anthropometric features [12], and even to generate HRTFs from the image of the pinna [13]. Besides anthropometric features, spatial principal component analysis and spherical harmonic analysis were conducted on HRTFs to extract representative features as training targets of DNN models [14][15].

However, DNN-based methods require a large amount of training data to avoid over-fitting. To alleviate the problem, we have proposed an auto-encoder to reduce the dimension of HRTFs [16]. Although we showed the proposed auto-encoder outperforms the DNN model [12], the variance of its performance across target subjects is still large. In this paper, we aim to further improve the accuracy and the stability of the HRTF estimation model using ensemble learning (EL), which has been used in many research fields to reduce system's variance while increasing the system's generalization capability [17][18][19]. Ensemble learning is a machine learning method to combine multiple models. It leverages the strength of different models and reduces the impact of individual model's biases and errors, delivering more robust and reliable predictions. Inspired by the integrated deep and ensemble learning (IDEA) algorithm proposed in [20], we construct an EL-based model with groups of subjects. Each group contains subjects with similar anthropometric features such that the model can learn more representative characteristics of the group. As shown in [5][7], some anthropometric features are more relevant to HRTFs than others. The subjects in the dataset are first clustered by these relevant features, and then put through a two-stage training, consisting of an individual group training and an ensemble integration training. In this way,

we can build an EL-based HRTF estimator with improved accuracy and stability.

The rest of the paper is organized as follows. In Section II, we describe the pre-processing of training data, the clustering processes, and the EL model architecture. In Section III, experimental results and system comparisons are presented and discussed. Finally, Section IV concludes our work.

II. Proposed Method

In this section, we first introduce the pre-processing steps, and then demonstrate the initial grouping on subjects and the architecture of the proposed model.

A. Dataset and pre-processing

The CIPIC HRTF database contains HRIR data of 45 subjects with their anthropometric features [21]. For each subject, a total of 1250 HRIRs, combinations of 25 azimuth angles and 50 elevation angles, and their 37 anthropometric features, including 17 features related to the head and torso, and 10 features related to each pinna, are recorded. The definitions of the anthropometric features, and azimuth and elevation angles can be found in [21]. However, only 35 out of the 45 subjects have complete anthropometric features, hence, their features and HRIRs were used in this study. Before training DNN models, we normalized the anthropometric features and transformed the HRIRs to HRTFs.

When estimating the magnitude responses of HRTFs of the left/right ear, we only used the ten anthropometric features related to the left/right ear and the 17 head-torso features. Therefore, the input to the DNN models was a 27-dimensional feature vector. We followed the formula in [12] to normalize each feature as

$$a_i' = (1 + e^{-\frac{(a_i - \mu_i)}{\sigma_i}})^{-1} \quad (1)$$

where a_i is the i-th anthropometric feature, and μ_i and σ_i are the mean and standard deviation over all i-th features, respectively. Finally, $[a_1', a_2', ..., a_{27}']$ was used as the input vector to DNN models.

To derive the magnitude responses of HRTFs, we conducted 512-point discrete Fourier transform (DFT) on HRIRs and used the constant-Q filterbank (Q=8) to smooth the magnitude responses. Since sound characteristics between 0.2 kHz and 15 kHz are relevant to localization [22], we selected the corresponding frequency bins from the magnitude responses of HRTFs in the frequency domain and took the logarithm to obtain the spectra of HRTFs in decibel. After these steps, each log-magnitude spectrum was represented by a 173-frequency-bin vector. Since we used the sigmoid activation function in the output layer of the DNNs, we normalized the log-magnitude spectra of HRTFs to values between 0 and 1. To ensure fairness in evaluation, all compared models were trained using the same pre-processed data. Note that, the log-magnitude spectrum of the HRTF is the estimation target in this paper. For synthesizing the spatial sound, the phase response of the HRTF is derived using minimum-phase reconstruction as in [10][14].

B. Subject grouping for ensemble learning

It has been shown that the HRTFs are systematically different among subjects, and a scaling factor on the frequency axis can be found to minimize the difference between individuals' HRTFs [5]. The optimal scaling factor between the HRTFs of two subjects can be predicted by the relative physical sizes of the two subjects. The study demonstrated that the measurements of "Pinna-cavity height" and "Head width" are highly correlated with the optimal scaling factor. These two measurements can be explicitly or implicitly obtained from the anthropometric feature set of the CIPIC database. The head width ($x1$) is directly recorded in CIPIC, and the pinna-cavity height can be derived by summing the cavum concha height ($d1$), the cymba concha height ($d2$), and the fossa height ($d4$) of CIPIC. Therefore, we selected the "Pinna-cavity height" and "Head width" features as basis to group the 35 subjects for the proposed EL-based model. The proposed model is referred to as the Ensemble-Anthro (EA) model and the version using these two features for initial grouping is represented by $EA_{x1,d1+d2+d4}$.

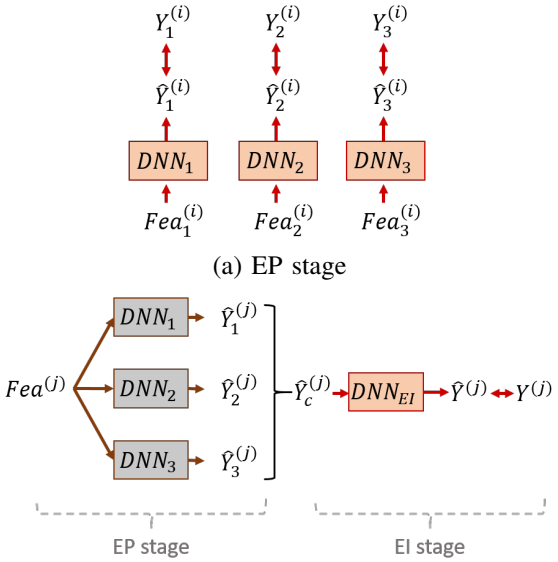
On the other hand, a more recent study analyzed the sensitivity of pinna morphology on HRTFs, and showed that the cavum concha width ($d3$) and the fossa height ($d4$) are related to the sensitive control points which are highly correlated to HRTFs [7]. Similar findings were also observed in other studies [23][24]. Therefore, we developed another version of the EA model, referred to as $EA_{d3,d4}$, using these two ear features for initial grouping. Considering findings in [5] and [7], we also tried the third version of the EA model using ear-related $d3$, $d4$ features and the head-related $x1$ feature for initial grouping, which is referred to as $EA_{x1,d3,d4}$. Note, $x1$, $d1$, $d2$, $d3$, and $d4$ are the corresponding feature labels used in CIPIC. For each version of the EA model, the k-means clustering method was used to initially divide the 35 subjects into three groups based on selected features. After clustering, $EA_{x1,d1+d2+d4}$, $EA_{x1,d3,d4}$, and $EA_{d3,d4}$ models relatively have (9, 10, 16), (8, 11, 16), and (9, 10, 16) subjects in three groups.

C. Architecture of the proposed EA model

Similar to the IDEA algorithm proposed in [20], we constructed an EL-based model for estimating log-magnitude spectra of HRTFs. In the training phase, the proposed model went through an ensemble preparation (EP) stage and an ensemble integration (EI) stage, as respectively shown in the top and bottom panels of Fig. 1.

In the EP stage, we trained three independent DNNs for the three groups clustered by the k-means method. The anthropometric parameters of the i-th subject in the three groups were respectively collected as $Fea_1^{(i)}$, $Fea_2^{(i)}$, and $Fea_3^{(i)}$, which were used as the input features to the three DNNs, DNN_1, DNN_2, and DNN_3. Each DNN was trained to minimize the mean square error between the estimated log-magnitude spectrum $\hat{Y}^{(i)}$ and the real log-magnitude spectrum $Y^{(i)}$.

In the EI stage, we first fixed DNN_1, DNN_2, DNN_3 trained in the EP stage. For subject j, his anthropometric



(a) EP stage

(b) EP stage followed by EI stage

Fig. 1. Architecture diagram of the proposed EL model, including the ensemble preparation (EP) and the ensemble integration (EI) stages.

parameters $Fea^{(j)}$ was put through the three DNNs to produce three estimated log-magnitude spectra $\hat{Y}_1^{(j)}, \hat{Y}_2^{(j)}$, and $\hat{Y}_3^{(j)}$. The three estimates were concatenated into a 519-dimensional vector $\hat{Y}_c^{(j)}$ as the input features to the following EI DNN (DNN_{EI}). As shown in the bottom panel of Fig. 1, the DNN_{EI} was trained by minimizing the mean square error between the estimated log-magnitude spectrum $\hat{Y}^{(j)}$ and the real log-magnitude spectrum $Y^{(j)}$.

In the inference phase, the anthropometric parameters of a test subject were put through the model, including DNN_1, DNN_2, DNN_3, and DNN_{EI}, to produce the estimated log-magnitude spectra of HRTFs. The overall model maps the subject's anthropometric features to the subject's log-magnitude spectrum of the HRTF via the mapping function $M(\cdot)$ as

$$M(\cdot) = g([f_1(\cdot), f_2(\cdot), f_3(\cdot)]) \tag{2}$$

where $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$, and $g(\cdot)$ are the mapping functions of DNN_1, DNN_2, DNN_3, and DNN_{EI}, respectively.

III. Experimental Results

A. Model settings and evaluation metric

For evaluations, we compared the proposed EA model with a DNN baseline model [12] and our previously proposed AutoEn+DNN model [16]. Pilot experiments were conducted to determine the optimal numbers of hidden layers and hidden units for each DNN in the EA model. When training the DNNs, the ReLU and the sigmoid function were used as the activation functions of hidden layers and the output layer, respectively, and the Adam optimizer was used with a learning rate of 0.0001 and a drop-out rate of 0.2. In EP stage, we trained DNNs with 100 epochs. In EI stage, we stopped training when the validation loss ceased to decrease for 10 epochs. Note that we used network parameters tuned in their original studies [12][16] to implement the compared DNN baseline model and the AutoEn+DNN model.

Two experiments were conducted to evaluate the proposed EA model. In the first experiment, we trained an individual model for each direction. In total, we trained 25 models for 25 azimuth angles at 0° elevation. Our focus on the horizontal plane was primarily to be consistent with the previous study [16] to have fair comparisons. In that study, authors aimed to combine right and left ear HRIRs using one mapping rule, thus only considered the horizontal plane. The 25 tested azimuth angles were $\{0°, \pm5°, \pm10°, \pm15°, \pm20°, \pm25°, \pm30°, \pm35°, \pm40°, \pm45°, \pm55°, \pm65°, \pm80°\}$ as described in [21]. In the second experiment, we embedded the angle information in the input features to build a universal model for all directions. The original 27-dimensional anthropometric feature was cascaded with the azimuth and elevation angles to have a 29-dimensional feature as the input to the universal model.

We adopted the leave-one-out cross validation approach, i.e., one subject was used for test while the remaining 34 subjects were used for training. As in many HRTF estimation studies [6][25][11][12][13][16][26][23], we used the log spectral distortion (LSD; in dB) measure for model evaluation. It is formulated as follows

$$LSD(H, \hat{H}) = \sqrt{\frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \left(20 log_{10} \frac{\mid H(k) \mid}{\mid \hat{H}(k) \mid} \right)^2} \tag{3}$$

where k is the index of frequency bin; H is the real spectrum and $\hat{H}$ is the estimated spectrum.

B. Experiment 1: an individual model for each direction

In Experiment 1, we trained a total of 25 models for 25 azimuth angles at 0° elevation. To tune the optimal numbers of hidden layers and hidden units, we tried two or three hidden layers with 16, 32, or 64 hidden units in each layer for DNN_1, DNN_2, and DNN_3. For DNN_{EI}, we tried three or four hidden layers with 32, 64, or 128 hidden units in each layer. In the end, we constructed DNN_1, DNN_2, and DNN_3 with three hidden layers and 16 hidden units in each layer. For DNN_{EI}, we used three hidden layers with 64 hidden units in each layer. Table I lists the structures of these models in details. Using the leave-one-out cross validation, we ended up with 875 (=25×35) LSD measures to calculate the mean and variance.

Table II shows the overall mean and variance of the LSD measures of all compared models. Note that we re-implemented and re-trained the DNN baseline model and the AutoEn+DNN model such that their scores might be different from the scores reported in their original studies [12][16]. The lower mean and lower variance of the proposed EA models indicate they can estimate log-magnitude spectra more accurately and stably than the Baseline and the AutoEn+DNN models. We performed t-tests between each pair of the three

TABLE I
MODEL STRUCTURE OF $DNN_{1,2,3}$ / DNN_{EI} IN EXPERIMENT 1

Layer	Units	Activation Function
Input	27 / 519	-
Dense Layer 1	16 / 64	ReLU
Dropout 1	-	-
Dense Layer 2	16 / 64	ReLU
Dropout 2	-	-
Dense Layer 3	16 / 64	ReLU
Dropout 3	-	-
Output Layer	173 / 173	Sigmoid

TABLE II
THE OVERALL MEAN AND VARIANCE OF LSD OF COMPARED MODELS IN EXPERIMENT 1

	mean	variance
Baseline DNN	3.52	1.37
AutoEn+DNN	3.56	1.16
$EA_{x1,d1+d2+d4}$	**3.33**	1.07
$EA_{x1,d3,d4}$	**3.33**	1.07
$EA_{d3,d4}$	**3.33**	**1.06**

EA models, but found no significant differences. However, we observed significant differences (t-values: 3.67/4.47, p-values: 0/0) in t-tests between the DNN baseline/AutoEn+DNN model and the $EA_{x1,d1+d2+d4}$ model. These results suggest that the proposed EA models with three grouping criteria perform quite comparably to each other and significantly better than the DNN baseline and AutoEn+DNN models.

C. Experiment 2: a universal model for all directions

In Experiment 1, we followed the approach in [12] to train an individual model for each direction, which is not very practical in our opinion. Hence, we directly added the azimuth and the elevation angles to the input 27-dimensional feature to train a universal model for all directions. In this experiment, we also tuned the numbers of hidden layers and hidden units for DNN_1, DNN_2, DNN_3 and DNN_{EI}. In the end, these four DNNs were set with three hidden layers, each of which had 128 hidden units. Table III shows the structures of these models in details.

In addition to directly concatenating the 27 anthropometric features with two angles, we also tried the embedding method, referred to as the Ensemble-Embed (EE) model, to integrate the two types of features. For each group, we first used two small NNs to embed the two types of features into their latent spaces separately, then combined the latent codes for training the DNNs. Each of the small NNs contained one hidden layer of 128 hidden units and an output layer of 32 units. For each group, the corresponding two 32-dimensional latent codes were then concatenated into a 64-dimensional vector for further training the DNN. In this EE approach, we allowed the three groups to update their two small NNs simultaneously through backpropagation in the EP stage. To have comparable computations with EA models, we implemented DNN_1, DNN_2, and DNN_3 in the EE approach only with two hidden layers of 128 hidden units to offset computations from two small NNs

TABLE III
MODEL STRUCTURE OF $DNN_{1,2,3}$ / DNN_{EI} IN EXPERIMENT 2 IN EA APPROACH

Layer	Units	Activation Function
Input	29 / 519	-
Dense Layer 1	128 / 128	ReLU
Dropout 1	-	-
Dense Layer 2	128 / 128	ReLU
Dropout 2	-	-
Dense Layer 3	128 / 128	ReLU
Dropout 3	-	-
Output Layer	173 / 173	Sigmoid

TABLE IV
MODEL STRUCTURE OF DNN_1, DNN_2, DNN_3 IN EXPERIMENT 2 IN EE APPROACH

Layer	Units	Activation Function
Input (Anthro., Angle)	(27, 2)	-
Anthro Embedding Layer 1	128	ReLU
Anthro Embedding Layer 2	32	ReLU
Angle Embedding Layer 1	128	ReLU
Angle Embedding Layer 2	32	ReLU
Concatenation Layer	64	-
Dense Layer 1	128	ReLU
Dropout 1	-	-
Dense Layer 2	128	ReLU
Dropout 2	-	-
Output Layer	173	Sigmoid

in feature embedding. Detailed structures of DNN_1, DNN_2, and DNN_3 in this EE approach are listed in Table IV.

For evaluations, we also performed leave-one-out cross validation. For each test subject, we estimated their log-magnitude spectra of HRTFs in all 1250 (=25×50) directions. Table V shows the overall mean and variance over 43750 (=1250×35) LSD measures from the compared models, including the Baseline, the EA models, and the EE models. To statistically analyze the overall mean, we performed t-test between any two compared models. Table VI shows pair-wise t-test results. Each t-test has the same degree of freedom and the "*" symbol is used to indicate the compared two models producing significant differences in LSD ($p < 0.05$). Note, the AutoEn+DNN model was not compared since it was originally proposed to estimate HRTFs only at 0° elevation.

The results in Table V and Table VI collectively show all ensemble-learning based EA and EE models outperform the Baseline model with statistical significance. The fact that the $EA_{x1,d3,d4}$ model outperforms the $EA_{d3,d4}$ model indicates the head-related feature $x1$ is necessary for grouping subjects in the universal HRTF estimator. Superior performance of the $EA_{x1,d3,d4}$ model to the $EA_{x1,d1+d2+d4}$ model implies the $d3$ and $d4$ ear-related features are more relevant to HRTFs than the $d1+d2+d4$ ear-related feature. Experimental results also show all EA models can be further improved using the feature embedding method. However, the further improvement from the best performing EA model $EA_{x1,d3,d4}$ to its EE version $EE_{x1,d3,d4}$ is not statistically significant.

In addition, we compared the performance of HRTF prediction in each of the four frequency subbands (0.2∼4 kHz, 4∼8

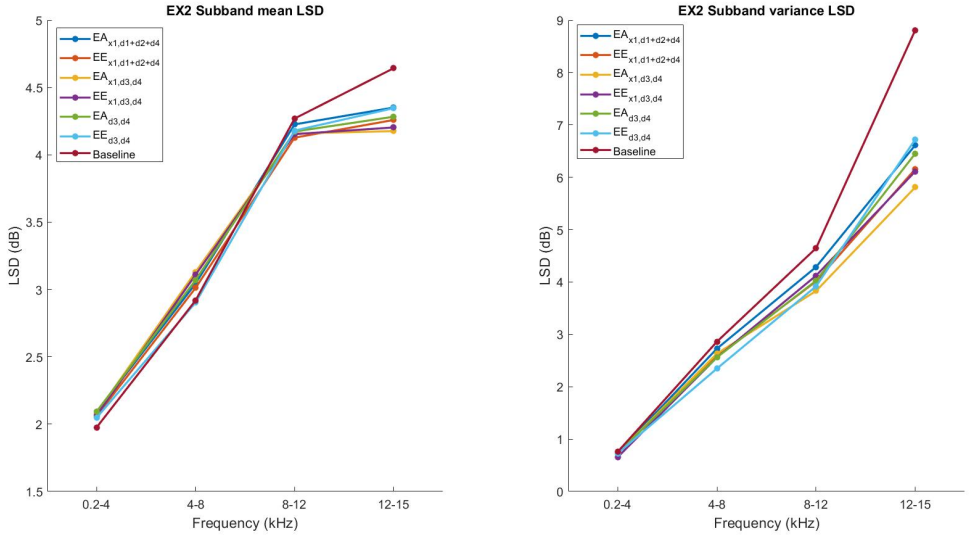


Fig. 2. Mean (left panel) and variance (right panel) of subband LSD of compared models in Experiment 2.

TABLE V
THE OVERALL MEAN AND VARIANCE OF LSD OF COMPARED MODELS IN EXPERIMENT 2.

	mean	variance
Baseline DNN	3.80	2.28
$\text{EA}_{x1,d1+d2+d4}$	3.73	1.89
$\text{EA}_{x1,d3,d4}$	3.68	1.62
$\text{EA}_{d3,d4}$	3.70	1.76
$\text{EE}_{x1,d1+d2+d4}$	3.66	1.70
$\text{EE}_{x1,d3,d4}$	3.67	1.75
$\text{EE}_{d3,d4}$	3.67	1.72

kHz, 8∼12 kHz, and 12∼15 kHz) in terms of LSD, as shown in Fig. 2. The left and right panels show mean and variance of subband LSD of compared models. The results clearly show the performance gains of the proposed models are more prominent in higher frequency bands (8∼12 kHz and 12∼15 kHz). This finding is quite reasonable since the proposed EL models use anthropometric features as the clustering basis and influences of variations of anthropometric features on HRTFs are mostly observed at high-frequency ranges [27].

IV. CONCLUSION

HRTFs are crucial for synthesizing spatial sounds. However, they vary widely among people and directions, and databases are usually too small to construct a high-fidelity DNN-based estimator. To tackle this problem, we propose an EL-based approach to estimate the log-magnitude spectrum of the HRTF. We choose several sets of anthropometric features, which were shown highly relevant to HRTFs in literature, for the initial grouping in the EL-based model.

We also embed the angles into the model to have an universal HRTF log-magnitude spectrum estimator for all directions. Simulation results demonstrate the best performing universal model uses the head-related feature $x1$ and the ear-related features $d3$ and $d4$ for initially clustering subjects. Results also show in two out of three clustering criteria tested, the EE model, which combines anthropometric features and angles in their latent forms, produces better estimates than the EA model, which directly combines anthropometric features and angles in their raw formats. Subband analysis on compared models further demonstrates the proposed EE and EA models significantly outperform the baseline model in reducing LSD variance in high frequency range (8 $\sim$ 15 kHz). It's worth noting that the proposed models currently can only estimate spectra of HRTFs for 1250 specific directions recorded in CIPIC. In the future, we will try our model on larger datasets and extend our model to estimate spectra of HRTFs for arbitrary directions as in [28]. We will also investigate the impact of factors, such as the number of groups and the basis for initial grouping, on the performance of the proposed model.

V. ACKNOWLEDGMENTS

This research is supported by Ministry of Science and Technology, Taiwan under Grant No MOST 110-2221-E-A49-115-MY3.

REFERENCES

[1] B. C. Moore, An introduction to the psychology of hearing. Brill, 2012.
[2] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," The Journal of the Acoustical Society of America, vol. 94, no. 1, pp. 111–123, 1993.
[3] P. Zahorik, P. Bangayan, V. Sundareswaran, K. Wang, and C. Tam, "Perceptual recalibration in human sound localization: Learning to remediate front-back reversals," The Journal of the Acoustical Society of America, vol. 120, no. 1, pp. 343–359, 2006.
[4] N. A. Gumerov, A. E. O'Donovan, R. Duraiswami, and D. N. Zotkin, "Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation," The Journal of the Acoustical Society of America, vol. 127, no. 1, pp. 370–386, 2010.

TABLE VI

PAIR-WISE T-TEST RESULTS OF EXPERIMENT 2. ENTRIES ARE (T-VALUE, P-VALUE) OF T-TEST WITH SIGNIFICANCE LEVEL OF 0.05. THE '*' SYMBOL INDICATES THE COMPARED TWO MODELS PRODUCE SIGNIFICANT DIFFERENCES IN LSD ($p < 0.05$).

	$EA_{x1,d1+d2+d4}$	$EE_{x1,d1+d2+d4}$	$EA_{x1,d3,d4}$	$EE_{x1,d3,d4}$	$EA_{d3,d4}$	$EE_{d3,d4}$
Baseline DNN	(7.32, 0)*	(14.32, 0)*	(12.91, 0)*	(13.23, 0)*	(10.36, 0)*	(12.90, 0)*
$EA_{x1,d1+d2+d4}$		(7.17, 0)*	(5.62, 0)*	(6.07, 0)*	(3.07, 0.002)*	(5.70, 0)*
$EE_{x1,d1+d2+d4}$			(-1.68, 0.093)	(-1.08, 0.28)	(-4.16, 0)*	(-1.50, 0.13)
$EA_{x1,d3,d4}$				(0.57, 0.57)	(-2.54, 0.011)*	(0.16, 0.87)
$EE_{x1,d3,d4}$					(-3.06, 0.002)*	(-0.41, 0.68)
$EA_{d3,d4}$						(2.67, 0.007)*

[5] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," The Journal of the Acoustical Society of America, vol. 106, no. 3, pp. 1480–1492, 1999.

[6] X. Liu and X. Zhong, "An improved anthropometry-based customization method of individual head-related transfer functions," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 336–339.

[7] P. Stitt and B. F. Katz, "Sensitivity analysis of pinna morphology on head-related transfer functions simulated via a parametric pinna model," The Journal of the Acoustical Society of America, vol. 149, no. 4, pp. 2559–2572, 2021.

[8] P. Runkle, A. Yendiki, and G. H. Wakefield, "Active sensory tuning for immersive spatialized audio." Georgia Institute of Technology, 2000.

[9] A. Silzle, "Selection and tuning of hrtfs," in Audio Engineering Society Convention 112. Audio Engineering Society, 2002.

[10] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," The Journal of the Acoustical Society of America, vol. 91, no. 3, pp. 1637–1647, 1992.

[11] T. Nishino, N. Inoue, K. Takeda, and F. Itakura, "Estimation of hrtfs on the horizontal plane using physical features," Applied Acoustics, vol. 68, no. 8, pp. 897–908, 2007.

[12] C. J. Chun, J. M. Moon, G. W. Lee, N. K. Kim, and H. K. Kim, "Deep neural network based hrtf personalization using anthropometric measurements," in Audio Engineering Society Convention 143. Audio Engineering Society, 2017.

[13] R. Miccini and S. Spagnol, "A hybrid approach to structural modeling of individualized hrtfs," in 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 2021, pp. 80–85.

[14] M. Zhang, Z. Ge, T. Liu, X. Wu, and T. Qu, "Modeling of individual hrtfs based on spatial principal component analysis," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 785–797, 2020.

[15] J. Xi, W. Zhang, and T. D. Abhayapala, "Magnitude modelling of individualized hrtfs using dnn based spherical harmonic analysis," in 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2021, pp. 266–270.

[16] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, "Autoencoding hrtfs for dnn based hrtf personalization using anthropometric features," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 271–275.

[17] T. G. Dietterich, "Ensemble methods in machine learning," in International workshop on multiple classifier systems. Springer, 2000, pp. 1–15.

[18] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[19] Y. Tsao and C.-H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 5, pp. 1025–1037, 2009.

[20] W.-J. Lee, S.-S. Wang, F. Chen, X. Lu, S.-Y. Chien, and Y. Tsao, "Speech dereverberation based on integrated deep and ensemble learning algorithm," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5454–5458.

[21] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The Cipic HRTF database," in Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575). IEEE, 2001, pp. 99–102.

[22] R. B. King and S. R. Oldfield, "The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional audio displays," Human factors, vol. 39, no. 2, pp. 287–295, 1997.

[23] H. Hu, L. Zhou, J. Zhang, H. Ma, and Z. Wu, "Head related transfer function personalization based on multiple regression analysis," in 2006 International Conference on Computational Intelligence and Security, vol. 2, 2006, pp. 1829–1832.

[24] M. Zhang, R. A. Kennedy, T. D. Abhayapala, and W. Zhang, "Statistical method to identify key anthropometric parameters in hrtf individualization," in 2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays, 2011, pp. 213–218.

[25] N. Inoue, T. Kimura, T. Nishino, K. Itou, and K. Takeda, "Evaluation of hrtfs estimated using physical features," Acoustical science and technology, vol. 26, no. 5, pp. 453–455, 2005.

[26] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A manifold learning approach for personalizing hrtfs from anthropometric features," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 24, no. 3, pp. 559–570, 2016.

[27] S. Ghorbal, T. Auclair, C. Soladie, and R. Seguier, "Pinna morphological parameters influencing hrtf sets," in International Conference on Digital Audio Effects, Edinburgh, United Kingdom, 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01831314

[28] D. Yao, J. Zhao, L. Cheng, J. Li, X. Li, X. Guo, and Y. Yan, "An individualization approach for head-related transfer function in arbitrary directions based on deep learning," JASA Express Letters, vol. 2, no. 6, p. 064401, 2022.

Secure Moving Object Detection Transformer in Compressed Video with Feature Fusion

Yuru Song, Yike Chen, Peijia Zheng*, Yusong Du and Weiqi Luo
Guangdong Provincial Key Laboratory of Information Security Technology,
MoE Key Laboratory of Information Technology,
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
*Corresponding author, email: zhpj@mail.sysu.edu.cn

***Abstract*—Moving Object Detection (MOD) is essential for surveillance videos, but the large data volume necessitates significant computational resources. Cloud computing offers a solution, though it raises privacy concerns. In this paper, we propose a secure MOD framework based on the Detection Transformer. The user-side protects video data using format-compatible selective video encryption before storing it in the cloud. In the cloud, we design a method to extract encrypted domain motion information from encrypted videos. Due to the dual constraints of compression and encryption, the extracted information is highly sparse, and common deep learning methods do not perform well. To address this issue, we propose a Convolutional Neural Network (CNN)-Transformer feature enhancement-fusion method to achieve effective feature alignment and fully exploit deep-level motion information. Specifically, we modulate CNN features at multiple scales using Transformer features, where high-level semantic information and low-level spatial information are fused for accurate moving object localization. We evaluate our model on the VIRAT and DUKE-MTMC datasets, demonstrating better detection performance and greater robustness in challenging scenarios compared to previous secure MOD methods.**

I. Introduction

In recent years, visual tasks such as moving object detection (MOD) in videos have become a major focus of artificial intelligence research. Given the extensive and voluminous nature of surveillance video datasets, the necessity for automated MOD is imperative. Local processing of these large datasets can be sluggish, prompting many users to opt for cloud-based solutions. However, uploading unencrypted videos to the cloud poses potential privacy risks. Therefore, it is necessary to implement privacy-enhancing methods on video data before storing it in the cloud. Fig. 1 shows the process of cloud encrypted video processing.

Initial MOD methods [1]–[4] for encrypted domain videos relied on traditional machine learning techniques, resulting in low detection accuracy. With the advancement of deep learning, neural network-based MOD models have greatly outperformed traditional methods. However, due to the disturbance of pixels in encrypted frames, deep learning MOD cannot be directly applied to encrypted videos. Tian et al. [5] proposed the first deep learning-based encrypted domain compressed video MOD framework. Nevertheless, Tian's method has areas for improvement. Firstly, their use of Convolutional Neural Network (CNN) models for MOD, which depends on anchor

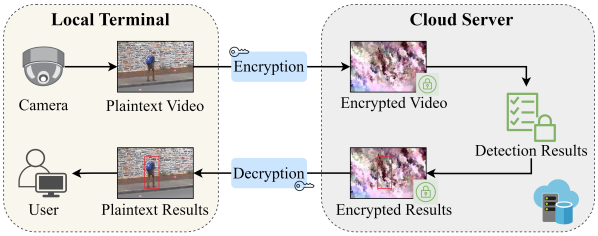


Fig. 1. The cloud-based surveillance system model.

boxes and complex post-processing steps. Secondly, CNNs primarily capture local object localization features without considering global features, which is inadequate for detecting complex scenes with long-range dependencies. Finally, the motion features extracted from encrypted domain compressed video bitstreams are sparser, and Tian's method did not fully utilize these sparse motion features.

Visual Transformers [6]–[8] have developed rapidly and significantly improved detection performance. However, no research has yet applied Transformers to MOD in encrypted compressed videos. Additionally, CNNs and Transformers learn different types of features—CNNs mainly capture localization features such as edges and lines around objects, while Transformers capture global pixel relationships and high-level semantic features. Many studies have shown that fusing these features benefits various visual recognition tasks [9], [10]. Given the sparsity of motion features extracted from encrypted compressed videos, feature fusion can fully leverage these sparse features, making it important to explore how to enhance model detection performance through this fusion.

In this paper, we design a novel Transformer-based secure MOD framework for compressed videos. Our approach employs the Selective Encryption (SE) technique to encrypt the compressed video bitstream, achieving an optimal balance between encryption efficiency and privacy. We extract encrypted domain motion information by leveraging entropy-encoded syntax elements from the compressed video bitstream and input it into the Deformable-DETR [11] model, using Swin-Transformer [12] as the backbone to extract multi-level features. These features are then fed into the encoder and our CNN-Transformer feature fusion module, with the encoder's

output also serving as input to the feature fusion module, and the final output being fed into the decoder.

The core design of our network is the CNN-Transformer feature enhancement and fusion module, which uses the object localization features from CNN to assist the high-level semantic features from Transformer, thereby increasing feature diversity. Our module has two sub-modules. The spatial enhancement module focuses on the position information of objects, using deformable convolutions to improve the ability to locate moving objects in occluded scenes and scenes with many small objects. The channel fusion module optimizes the channel distribution of features through global pooling and hyperparameter adjustment, dividing the features into multiple groups with semantic information characteristics captured by the encoder, and finally merging the features with the corresponding group's semantic information, achieving effective information fusion within each group. Thus, our module adeptly captures semantic and localization features, enabling comprehensive and efficient inter-domain feature alignment, fully exploiting the sparse information.

Our main contributions can be summarized as follows. **First**, we design a Transformer-based framework specifically for MOD in encrypted compressed videos, utilizing feature enhancement and fusion to fully exploit motion features and address the inherent sparsity of motion information in the encrypted domain. **Second**, we develop a CNN-Transformer feature enhancement-fusion module that captures and fuses high-level semantic and low-level localization features for effective inter-domain feature alignment. **Third**, experimental results on the H.264 and H.265 formatted datasets reveal that our model substantially improves accuracy in challenging scenarios with crowded moving objects.

II. Related Works

Deep-Learning-Based MOD [13], [14] employed CNNs to learn temporal and spatial features from video frames. TransVOD [15] introduced a temporal Transformer to aggregate spatial object queries and feature memories for each frame. However, these models are designed for video frames and are not suitable for the compressed domain. In encrypted videos, pixel values are scrambled and no longer represent accurate information. Existing deep learning MOD models cannot learn effective motion representations from scrambled pixels, making them unsuitable for encrypted videos.

Feature fusion is commonly used in various visual tasks. FPN [16] merges features at different scales to construct multi-layer feature maps, accommodating targets of different sizes. AFF [9] proposed a multi-scale channel attention module to better fuse features from different layers and scales. DA-DETR [17] blended CNN features using Transformer features at multiple scales for accurate object recognition and localization. Unlike these methods, we explore a feature enhancement-fusion method specifically for sparse encrypted compressed information, focusing on the effective utilization of the information. Moreover, we do not limit ourselves to the types of features extracted by the backbone.

Bitstream-encryption-based methods [1]–[4] typically employ format-compatible SE [18] to encrypt videos and then extract syntax elements such as Motion Vector Differences (MVD) from the encrypted compressed video bitstream. These methods, while fast since they do not require decoding compressed videos into frames, rely on traditional machine learning methods with empirical parameter settings for detection, which perform poorly in complex scenarios. To our knowledge, only one piece of research proposes a deep learning-based MOD framework for encrypted compressed video [5]. However, this method uses only a CNN architecture to operate on motion features, insufficiently utilizing sparse information, leaving room for improvement in both detection accuracy and speed.

III. Proposed Secure MOD Scheme with CNN-Transformer Feature Fusion Module

We describe the main procedure of our secure MOD scheme. First, compressed video bitstreams are encrypted locally and then uploaded to a cloud server. In the cloud server, the syntax elements are extracted from the encrypted bitstreams and restructured into motion feature maps. These feature maps serve as the input of our MOD Transformer, which consists of a backbone for extracting feature pyramids, an encoder for obtaining high-level semantic features, a CNN-Transformer feature enhancement-fusion module for feature alignment, and a decoder for annotating moving objects with bounding boxes. Finally, the cloud server sends the MOD results to the local terminal. We consider two popular attack models: Known Ciphertext Model and Known Background Model, and the security proof is the same as [5].

A. Privacy Enhancement on Video

Our secure MOD framework employs a format-compliant SE method [19], [20] for privacy enhancement on H.264 and H.265 videos. This video encryption scheme mainly encrypts intra-prediction modes (IPM), MVD, and residual data. Encrypting these syntax elements ensures significant spatial distortion in the reconstructed frames of the compressed encoding and causes error pixels to propagate from I-frames to reference frames, thereby protecting the visual content's privacy. We use H.265 encoding format as an example, but our framework is applicable to H.264 videos as well.

Intra_4×4 and Intra_16×16 IPM are related to intra-prediction data. For Intra_4×4 IPM, we encrypt the three fixed-length code bits representing the selected mode. For Intra_16×16 IPM, we encrypt the last bit of the codeword to maintain the block encoding mode [18]. MVD and ref_idx are both related to inter-prediction data. Since $suffix$ affects the sign and value of the MVD, we encrypt all $suffix$ bits of the MVD. To maintain format compatibility, only the last bit of ref_idx is encrypted under specific conditions.

Discrete Cosine Transform (DCT) coefficients representing residual data are quantized before entropy coding. We encrypt all $suffix$ bits of the Delta QP during the quantization process. In Context Adaptive Variable Length Coding (CAVLC), we encrypt all sign bits of the $sign_of_TrailingOnes$ and

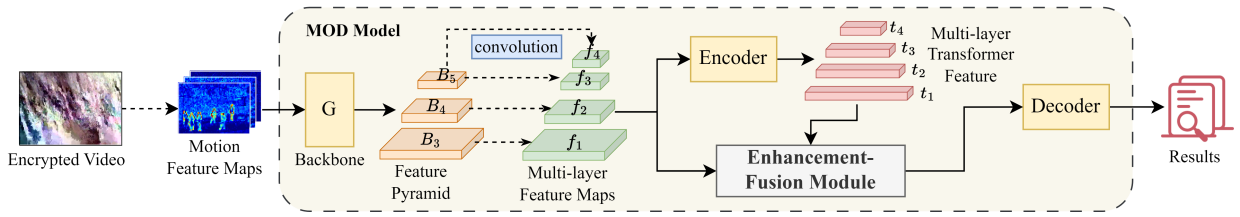


Fig. 2. Overview of the deep learning MOD part of the proposed framework.

$level_suffix$ codewords. This encryption process secures both the signs and values of the residual data.

B. Syntax Elements Extraction

We denote the t-th frame as $F(t)$, which is divided into a series of 4×4 minimum coding unit blocks. The i-th block is represented as $b_i(t)$. For each $b_i(t)$, the syntax elements contained are represented as $c_i(t)$, $m_i(t)$ and $d_i(t)$, corresponding respectively to the consumed bits, MVD magnitude, and residual density data. These syntax elements are not affected by SE and can reflect certain motion information.

The consumption of coding bits can reflect the degree of pixel changes caused by unpredictable motion, thereby reflecting motion information. Assuming that a basic coding unit block $B(t)$ contains θ 4×4 blocks, $c_i(t) = \lfloor \frac{\texttt{cit}(B(t))\times 4}{\sqrt{\theta}} \rfloor, i \in \mathbb{I}(t)$, where $\texttt{cit}(\cdot)$ is a function calculating the bit count of the coding unit and $\mathbb{I}(t) = \{i_1, i_2, \cdots, i_c\}$ represents the set of addresses of all 4×4 blocks within $B(t)$.

Motion Vector (MV) can reflect motion information and encoded as predictive MVs and MVD, but only MVD is entropy coded in video compression standards. Therefore, we use MVD to estimate motion information. $m_i(t)$ is defined as $m_i(t) = \lfloor \sqrt{L_x(t)^2 + L_y(t)^2} \rfloor$, where $L_x(t)$ and $L_y(t)$ are the lengths of the $suffix$ codeword for the horizontal and vertical components of the MVD, respectively. However, I-frames are encoded independently and lack MVs, which can cause confusion and temporal inconsistency in the model. To address this, we interpolate MVs into I-frames using Inverse Distance Weighted (IDW) temporal interpolation, leveraging the strong spatial-temporal correlation of adjacent frames.

Given that the contour areas of moving objects undergo a high degree of pixel substitution, these areas exhibit higher density compared to background blocks, enabling the density of non-zero DCT residuals $d_i(t)$ to highlight the edges of moving objects. We use $\texttt{dit}(\cdot)$ to denote the function that calculates the number of non-zero coefficients in a DCT block. Supposing a DCT block comprises δ 4×4 blocks, the residual density $d_i(t)$ is defined as $d_i(t) = \lfloor \frac{\texttt{dit}(B_{dct}(t))}{\sqrt{\delta}} \rfloor$.

Subsequent to the extraction of syntax elements, we reshape $c_i(t)$, $m_i(t)$, and $d_i(t)$ into two-dimensional matrices and collectively refer to them as motion feature maps.

C. Deep Learning MOD Model using Detection Transformer

Since Deformable-DETR contains a feature pyramid structure, it can better focus on targets of different scales and achieve feature fusion to a certain extent. In addition, Deformable-DETR uses the deformable attention mechanism to perform calculations in local areas of interest, which can more flexibly adapt to changes in target shape and improve the perception ability of the model in different scenarios. Therefore, we use Deformable-DETR as our basic model to extract the feature pyramid.

Our model consists of a backbone for feature extraction, an encoder containing 6 encoder-layers, a CNN-Transformer feature enhancement-fusion module, and a decoder with 6 decoder-layers for final predictions. Each encoder-layer and decoder-layer contains a multi-head deformable attention mechanism, residual connections and layer normalization, and a feed forward neural network. As shown in the Fig. 2, the feature pyramid B_3-B_5 extracted from the Swin-Transformer backbone is transformed into multi-scale feature maps f_i $(i = 1, 2, 3, 4)$ through a series of convolutions. Specifically, B_3-B_5 are transformed into f_1-f_3 via 1×1 convolutions, while the lowest resolution feature map f_4 is obtained from B_5 using a stride-2 3×3 convolution. Then, f_1-f_4 is used as the input of the encoder to obtain the Transformer features t_1-t_4. Next, the CNN-Transformer enhancement-fusion module takes the CNN features f_i $(i = 1, 2, 3, 4)$ and the Transformer features t_i $(i = 1, 2, 3, 4)$ as input. An overview of the processing on the encrypted video is shown in Fig. 2.

D. CNN-Transformer Feature Enhancement-Fusion Module

Syntax elements, extracted from encrypted domain compressed video bitstreams, contains sparser information compared to original RGB frames. Fully utilizing this motion information is challenging. Previous studies[9], [17] have shown that feature fusion can improve feature utilization and detection accuracy. Therefore, we add a CNN-Transformer feature enhancement-fusion module to Deformable-DETR. Notably, our module is plug-and-play and can be integrated into any DETR-like model. As illustrated in Fig. 3, the module takes multi-scale motion feature maps and multi-scale Transformer features as input. It comprises two parts: a spatial enhancement module for object localization and a channel fusion module for intra-group alignment.

1) Spatial-Enhancement Module: This module performs convolution operations on the multi-scale feature maps f_i $(i = 1, 2, 3, 4)$ to focus on regions related to moving objects, enhancing the ability to locate objects in occluded scenes and those with many small objects. We use deformable convolution v2 [21] for sparse learning of low-level localization features, allowing adaptive adjustment of convolution kernel sampling

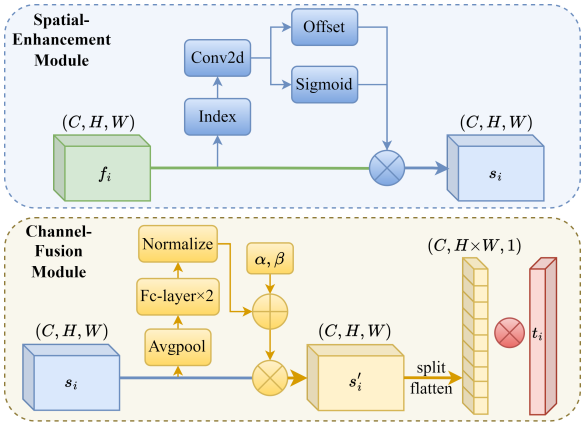


Fig. 3. Overview of proposed CNN-Transformer feature enhancement-fusion module.

positions to capture object location features more accurately. This enhances the ability to locate moving objects in occluded and densely populated scenes. We focus on the motion of each object, learning offsets based on their spatial positions, and then fusing the features of the same moving object:

$$s_i = \sum_{n=1}^{N} \omega_n \cdot f_i(p_n + O_{p_n}, c) \cdot \triangle m_n \tag{1}$$

where s_i is the new CNN feature obtained from this module, with each layer's feature dimension matching the original multi-scale feature map. N is the number of coefficient sampling positions, $p_n + O_{p_n}$ is the offset position of the target at p_n, and $\triangle m_n$ is the self-learned importance scalar of p_n.

2) Channel-Fusion Module: This module first adjusts the features of $s_i \in \mathbb{R}^{C \times H_i \times W_i}$ $(i = 1, 2, 3, 4)$ based on task importance within the channel, then groups CNN features according to the number of Transformer feature groups obtained from the encoder, and aligns and fuses each corresponding CNN-Transformer feature group.

To clarify, let's consider one layer of CNN feature s_1 and its corresponding Transformer feature t_1. We dynamically adjust the weights of s_1's channels in the context of the current detection task. Specifically, we first perform global average pooling to reduce complexity, then use two fully connected layers with ReLU activations and a normalization layer. Finally, we use a hard-sigmoid function to normalize the output weights to the range [-1,1] and adjust the channel features using four hyperparameters:

$$s'_1 = \max\left[\alpha^1(s_1) \cdot s_{1_c} + \beta^1(s_1), \alpha^2(s_1) \cdot s_{1_c} + \beta^2(s_1)\right] \tag{2}$$

where $\left[\alpha^1, \beta^2, \alpha^2, \beta^2\right]$ are hyperparameters controlling the adjustment of channel weights.

Next, we flatten the processed CNN features. Noting that the dimension of t_1 is $C \times (H_1 \times W_1) \times 1$, divided into k equal-sized image blocks along the $H_1 \times W_1$ dimension. We flatten s'_1 and perform feature fusion on the corresponding image blocks:

$$V_1 = \texttt{flatten}(s'_1) + t_1 \tag{3}$$

TABLE I
ABLATION EXPERIMENTS ON FEATURE ENHANCEMENT-FUSION MODULE.

SEM	CFM	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
×	×	64.5	95.9	77.9	6.6	50.5	68.1
✓	×	66.7	95.9	80.9	6.6	52.2	70.3
×	✓	66.7	95.9	80.9	6.6	52.9	70.3
✓	✓	**67.6**	**96.2**	**82.6**	**6.7**	**53.7**	**71.2**

Finally, the aligned features V_i $(i = 1, 2, 3, 4)$ are used as input for the decoder.

IV. EXPERIMENT

A. Experiment Setups

Consistent with previous works[5], we established datasets for our experiment from two large HD surveillance video datasets: VIRAT and Duke-MTMC. VIRAT consists of 680 video clips containing 400,806 frames, while Duke-MTMC includes 1,109 clips with 315,625 frames. We randomly selected 70% of the video clips for training, 5% for validation, and 25% for testing. We used the standard COCO Average Precision (AP) accuracy metrics [22] to evaluate performance. For loss calculation, we used L1 Loss with a weight of 5 for bounding box loss, GIoU Loss with a weight of 2 for IoU loss, and Focal Loss with a weight of 2 for bounding box classification. The number of object queries was set to 300. We used four NVIDIA GeForce RTX 3090 GPUs for model training and testing. The training of our two-stage model was optimized using the AdamW optimizer, with an initial learning rate of 0.0002 and a weight decay of 0.0001.

B. Ablation Study

The proposed CNN-Transformer feature enhancement-fusion module consists of the spatial-enhancement module (SEM) and the channel-fusion module (CFM). We conducted experiments on the Duke-MTMC dataset to examine their contributions to our MOD model's performance. Since SEM primarily enhances Transformer features using CNN methods, we used the fusion methods in CFM for feature integration during the experiments. As demonstrated in Table I, adding either module improved the model's detection results. The improvements from both modules were remarkably similar, but adding CFM provided a greater boost in AP_M, likely because channel fusion enhances the detection of medium-sized objects. The combination of SEM and CFM achieved the best 67.6% mAP, indicating that these structures are complementary.

C. Comparison Study

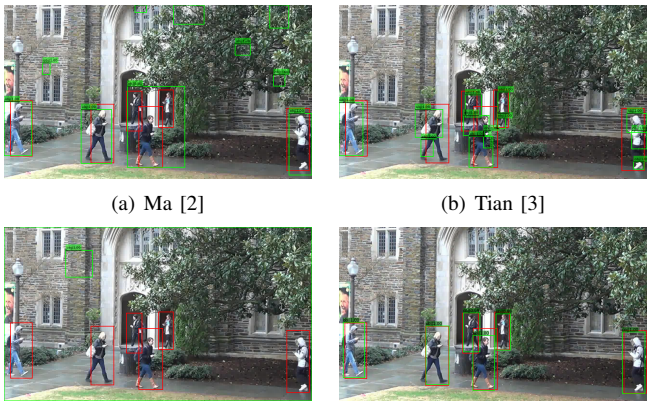
1) Comparison with Existing Secure MOD Methods: We compared our approach with secure MOD methods proposed by Guo [1], Ma [2], Tian [3], Liu [4], and Tian [5]. The first four methods use traditional machine learning for object detection, while the last one employs deep learning. To ensure fairness, we used scenario-independent testing, excluding *VIRAT_S_0400* from our training set. The results are shown in Table II. Due to the absence of large objects in the

TABLE II
THE COMPARISON WITH ALL EXISTING TRADITIONAL METHODS ON THE SCENE OF *VIRAT_S_0400*.

Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M
Guo [1]	11.0	44.1	0.90	5.80	18.3
Ma [2]	13.6	60.9	1.20	15.5	11.0
Tian [3]	16.4	54.9	5.70	15.3	18.0
Liu [4]	16.1	70.6	1.60	17.6	14.1
Tian [5]	50.9	96.9	46.7	**46.6**	54.5
Ours	**51.2**	**97.0**	**47.3**	46.2	**55.3**

TABLE III
THE COMPARISON OF OUR METHOD AND TIAN [5] ON THE FULL DATASET.

Dataset	Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Duke-MTMC	Tian [5]	66.1	95.0	79.9	6.5	51.3	69.7
	Ours	**67.6**	**96.2**	**82.6**	**6.7**	**53.7**	**71.2**
VIRAT	Tian [5]	36.2	78.5	28.1	22.1	41.3	**59.2**
	Ours	**36.4**	**79.1**	**28.4**	**22.4**	**41.8**	58.6



(a) Ma [2] (b) Tian [3]

(c) Liu [4] (d) Ours

Fig. 4. An example of visualization of results on the *camera_7_0022* from Duke-MTMC dataset. The red boxes represent the ground-truth and the green boxes indicate the results. The results from traditional methods are derived from their detection algorithms, and we use a confidence score threshold of 0.5 to display these bounding boxes.

scene, the result of AP_L is -1.000. Additionally, we compared our method with existing deep learning MOD methods on the complete VIRAT and Duke-MTMC datasets, as shown in Table III. Our approach achieved the highest detection performance on the *VIRAT_S_0400* video clip, except for AP_S, and showed significant improvements over traditional methods. On the complete Duke-MTMC and VIRAT datasets, our method outperformed Tian's [5] method, which used a CNN-based approach. Our DETR-based deep learning method, which integrates CNN and Transformer features, demonstrated superior overall performance.

To further showcase our method's superiority, we conducted comparative experiments in challenging scenarios, such as crowded and dynamic background scenes, and visualized the detection results on plaintext frames. As exemplified in Fig. 4, our method's detection performance in overlapping target scenarios is far superior to other methods.

2) Comparison with Existing MOD Models: Based on Tian's [5] comparative experiments, we compared our approach with MotionRec and T_RetinaNet proposed in [13] on the datasets used in the respective paper. We encrypted the dataset videos and split them into training and testing sets according to the paper's standards. We changed the detection categories to three (background, person, vehicle) and used the same evaluation metrics. We conduct experiments in all scenarios and present the results for three of them. As shown in Table IV, when MotionRec's depth was 30, the plaintext domain model achieved the best detection results. Although our method did not achieve the highest AP value compared to existing plaintext MOD methods, this is because plaintext MOD methods can directly access more motion information. In contrast, our method accesses limited motion information from encrypted video bitstreams due to privacy protection requirements.

TABLE IV
THE AP_{50} COMPARISON BETWEEN MOD MODELS. T_R REPRESENTS T_RETINANET, MR STANDS FOR MOTIONREC. D. DENOTES THE DATA DOMAIN WHICH INCLUDES P.(PLAINTEXT) AND E.(ENCRYPTED) DOMAIN.

Method	D.	*Sofa*	*Parking*	*Bungalows*	Overall
T_R_V1	P.	$72.0_{-2.2}$	$47.2_{-21.8}$	$88.7_{+2.5}$	$69.3_{-7.6}$
T_R_V2	P.	$35.1_{-39.1}$	$30.0_{-39.0}$	$42.6_{-43.8}$	$26.0_{-50.9}$
MR_V1	P.	$\mathbf{80.5}_{+6.3}$	$\mathbf{69.5}_{+0.5}$	$\mathbf{89.0}_{+2.8}$	$\mathbf{79.7}_{+2.8}$
MR_V2	P.	$70.3_{-3.9}$	$61.2_{-7.8}$	$84.6_{-1.6}$	$72.0_{-4.9}$
Tian[5]	E.	$74.2_{-0.0}$	$68.8_{-0.2}$	$86.0_{-0.2}$	$76.3_{-0.6}$
Ours	E.	74.2	69.0	86.2	76.9

However, it is noteworthy that our performance still surpassed three other plaintext MOD methods and was better than Tian's [5] deep learning MOD method. The slight difference between our method and the best plaintext method validates our framework's effectiveness and superiority in encrypted video MOD tasks. Additionally, MotionRec's prediction speed is only 2-5 fps, while our model achieves a detection speed of 42.2 fps, as it operates on compressed domain videos without needing to decode them into frames.

D. Adaptation to Other Video Codec and Encryption

Our method is not dependent on the video's encoding format or the selective encryption method used. We conducted experiments with H.264 and H.265 videos encrypted using different methods, as shown in the Table. The results indicate that the detection outcomes remain consistent regardless of the encryption method used for the same video encoding format. This consistency is because the encryption scheme does not affect the syntax elements in the same encoded video stream, resulting in identical extracted features. Additionally, detection results slightly differ between H.264 and H.265 videos due to compression efficiency differences, with H.265 yielding sparser information, leading to slightly lower detection performance compared to H.264.

V. CONCLUSIONS

We propose a Transformer-based deep learning secure MOD framework for encrypted domain compressed videos. Our approach uses SE to encrypt video bitstreams and then extracts motion information directly from the encrypted video,

TABLE V
PERFORMANCE COMPARISON OF THE PROPOSED FRAMEWORK USING VARIOUS ENCRYPTION SCHEMES APPLIED TO H.264/H.265 VIDEOS.

Encryption Scheme	Video Codec	VIRAT			Duke-MTMC		
		mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
Wang [23]	H.264	37.1	79.2	30.3	67.9	96.3	82.8
Xu [18]	H.264	37.1	79.2	30.3	67.9	96.3	82.8
Shahid [19]	H.265	36.4	79.1	28.4	67.6	96.2	82.6
Sallam [24]	H.265	36.4	79.1	28.4	67.6	96.2	82.6

which is used as input to the DETR model. We employ a CNN-Transformer feature enhancement-fusion method for effective feature alignment, improving the utilization of sparse information. Experimental results demonstrate that our method achieves advanced performance compared to other secure MOD techniques, with superior detection capabilities.

ACKNOWLEDGMENT

This work is supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515030087 and Grant 2022A1515011512, the National Natural Science Foundation of China under Grant 62272498, and the Guangdong Provincial Key Laboratory of Information Security Technology (No. 2023B1212060026). Thanks are due to Dr. Xianhao Tian for assistance with the experiments.

REFERENCES

[1] J. Guo, P. Zheng, and J. Huang, "An efficient motion detection and tracking scheme for encrypted surveillance videos," *ACM Trans. Multimedia Comput. Commun. & Appl.*, Sep. 2017.

[2] X. Ma, H. Peng, H. Jin, and B. Zhu, "Privacy-preserving cloud-based video surveillance with adjustable granularity of privacy protection," in *Proc. ICIP*, 2018.

[3] X. Tian, P. Zheng, and J. Huang, "Robust privacy-preserving motion detection and object tracking in encrypted streaming video," *IEEE Trans. Inf. Forensics Security*, pp. 1–1, 2021.

[4] C. Liu, X. Ma, S. Cao, J. Fu, and B. B. Zhu, "Privacy-preserving motion detection for HEVC-compressed surveillance video," *ACM Trans. Multimedia Comput. Commun. & Appl.*, p. 27, 2022.

[5] X. Tian, P. Zheng, and J. Huang, "Secure deep learning framework for moving object detection in compressed video," *IEEE Trans. Depend. Sec. Comput.*, 2023.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020, pp. 213–229.

[7] D. Han, X. Pan, Y. Han, S. Song, and G. Huang, "Flatten transformer: Vision transformer using focused linear attention," in *Proc. ICCV*, 2023, pp. 5961–5971.

[8] C. Zhao, Y. Sun, W. Wang, *et al.*, "Ms-detr: Efficient detr training with mixed supervision," in *Proc. CVPR*, 2024, pp. 17 027–17 036.

[9] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. WACV*, 2021.

[10] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation," in *Proc. CVPR*, 2017, pp. 3029–3037.

[11] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[12] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. ICCV*, 2021, pp. 10 012–10 022.

[13] M. Mandal, L. K. Kumar, M. S. Saran, and S. K. vipparthi, "MotionRec: A unified deep framework for moving object recognition," in *Proc. WACV*, Mar. 2020, pp. 2723–2732.

[14] A. Guzman-Pando and M. I. Chacon-Murguia, "DeepFoveaNet: Deep fovea eagle-eye bioinspired model to detect moving objects," *IEEE Trans. Image Process.*, pp. 7090–7100, 2021.

[15] L. He, Q. Zhou, X. Li, *et al.*, "End-to-end video object detection with spatial-temporal transformers," in *Proc. ACM MM*, 2021, pp. 1507–1516.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, 2017, pp. 2117–2125.

[17] J. Zhang, J. Huang, Z. Luo, G. Zhang, X. Zhang, and S. Lu, "Da-detr: Domain adaptive detection transformer with information fusion," in *Proc. CVPR*, 2023, pp. 23 787–23 798.

[18] D. Xu, R. Wang, and Y. Q. Shi, "Data hiding in encrypted H. 264/AVC video streams by codeword substitution," *IEEE Trans. Inf. Forensics Security*, pp. 596–606, 2014.

[19] Z. Shahid and W. Puech, "Visual protection of HEVC video by selective encryption of CABAC binstring," *IEEE Trans. Multimedia*, pp. 24–36, 2014.

[20] Q. Sheng, C. Fu, Z. Lin, *et al.*, "A fast selective encryption scheme for h. 264/avc video with syntax-preserving and zero bit rate expansion," *Signal Image & Video Process.*, pp. 1–15, 2023.

[21] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proc. CVPR*, 2019, pp. 9300–9308.

[22] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, Springer, 2014, pp. 740–755.

[23] Y. Wang, M. O'Neill, and F. Kurugollu, "A tunable encryption scheme and analysis of fast selective encryption for CAVLC and CABAC in H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1476–1490, 2013.

[24] A. I. Sallam, O. S. Faragallah, and E. M. El-Rabaie, "HEVC selective encryption using RC6 block cipher technique," *IEEE Trans. Multimedia*, pp. 1636–1644, 2018.

JAM: A Unified Neural Architecture for Joint Multi-granularity Pronunciation Assessment and Phone-level Mispronunciation Detection and Diagnosis Towards a Comprehensive CAPT System

Yue-Yang He[1], Bi-Cheng Yan[1], Tien-Hong Lo[1], Meng-Shin Lin[1], Yung-Chang Hsu[2], Berlin Chen[1]
[1]National Taiwan Normal University, Taipei, Taiwan
[2]EZAI, Taiwan
E-mail: {yueyanghe, bicheng, teinhonglo, 61147077s, berlin}@ntnu.edu.tw; mic@ez-ai.com.tw

***Abstract*—Computer-assisted pronunciation training (CAPT) systems are designed for second-language (L2) learners to practice their pronunciation skills by offering objective, personalized feedback in a stress-free, self-directed learning scenario. Both mispronunciation detection and diagnosis (MDD) and automatic pronunciation assessment (APA) are indispensable components in the CAPT systems. The former is responsible for pinpointing phonetic pronunciation errors and providing the corresponding diagnostic feedback, while the latter manages to evaluate oral skills across various linguistic levels with disparate aspects. Most existing efforts typically treat APA and MDD as independent tasks, where the correlations between the assessment scores and the phonetic pronunciation errors are nearly sidelined. In light of this, we introduce JAM (a Joint neural model for APA and MDD), a novel end-to-end neural model for CAPT that streamlines the components of APA and MDD into a unified structure with a parallel pronunciation modeling architecture. To capture fine-grained pronunciation cues from L2 learners' speech, electromagnetic articulography (EMA) features are introduced for the proposed model, which portrays the movement of articulatory structures, such as the jaw, lips, and tongue. A series of experiments conducted on the speechocean762 benchmark dataset demonstrate the feasibility and effectiveness of our approach compared to several competitive baselines. Additionally, an ablation study is presented to assess the contributions of different input features and training strategies in the proposed model.**

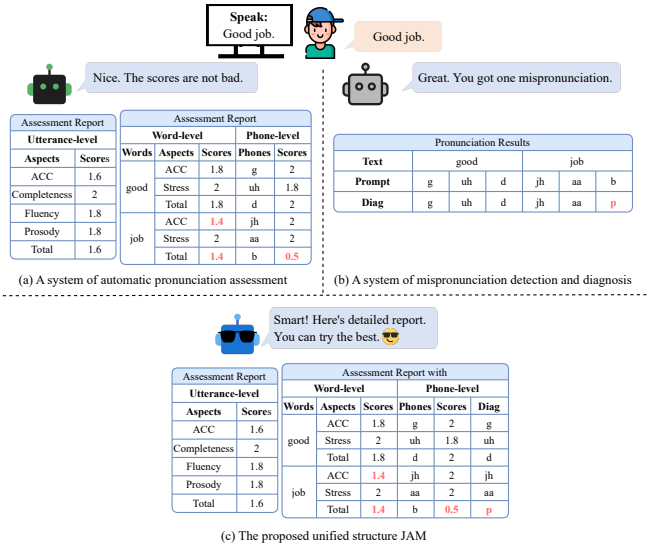


Fig. 1. A running example illustrates the two essential components of a CAPT system in a reading-aloud learning scenario. We present (a) phone-level mispronunciation detection and diagnosis, (b) automatic pronunciation assessment, and (c) the proposed unified structure, JAM. In this figure, '**Diag**' and '**Prompt**' represent the diagnosis result and the phone-level prompt sequence, respectively.

I. Introduction

In the tide of globalization, a growing number of people are willing or being demanded to learn foreign languages. This surging need fuels the development of computer-assisted pronunciation training (CAPT) systems, which open up new possibilities for L2 (second language) to practice pronunciation skills in a stress-free and self-directed learning scenario. Beyond the immense importance in the education and learning domain, CAPT systems have achieved remarkable success across various commercial platforms and assessment services, such as Duolingo [1] and Elsa speak [2].

A de-facto archetype system for CAPT is typically instantiated with a "reading-aloud" learning scenario, where an L2 learner is presented with a text prompt and instructed to pronounce it correctly. By working in conjunction with the input speech and the presented text prompt, CAPT systems can access the learner's speaking proficiency and immediately provide instructive diagnostic feedback [3][4][5]. To offer diagnostic feedback from different learning dimensions, as decipted in Figure 1(a) and 1(b), CAPT systems consist of two indispensable components: one is mispronunciation detection and diagnosis (MDD), and the other is automatic pronunciation assessment (APA). The former aims to pinpoint erroneous pronunciation segments at phone-level and provide the corresponding diagnostic feedback [6][7][8]. The latter, in contrast, draws attention on assessing and delivering various pronunciation scores to reflect the learner's pronunciation quality on some pronunciation aspects at various linguistic granularities [9][10][11]. Most existing efforts typically treat

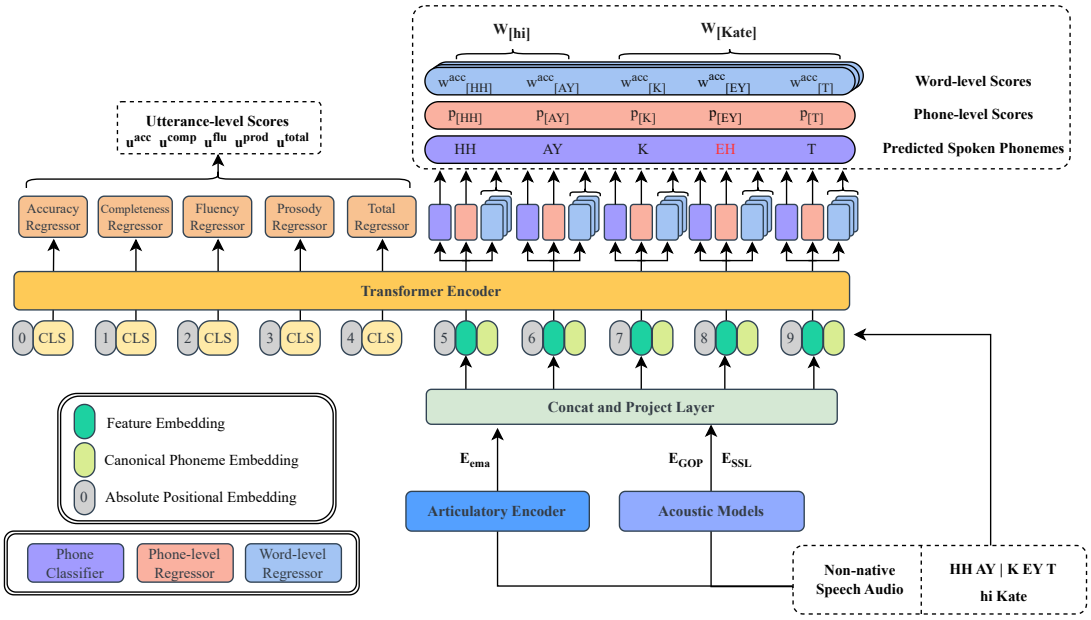


Fig. 2. Illustration of the proposed architecture of our APA method with phone-level mispronunciation detection and diagnosis. The red **'EH'** in the predicted phonemes sequence indicates that the speaker's pronunciation of this phoneme was diagnosed as a mispronunciation.

APA and MDD as independent tasks, where the correlations between the assessment scores and the phone-level pronunciation errors are nearly sidelined. However, empirical studies have revealed the negative correlations between assessment scores and the number of pronunciation errors across various aspects at diverse linguistic levels. For instance, phone-level pronunciation errors are closely related not only to utterance-level assessments, such as intelligibility [12] and fluency [13] in L2 English, but also to word-level assessments, such as stress [14] and intonation [15] in L2 Mandarin. In the follow-up work, Ryu et al. [16] ventured into integrating the tasks of MDD with APA, where the model architecture capitalized on a large-scale pre-trained speech encoder (i.e., Hubert), and then jointly optimized phone recognition and utterance-level assessment through the Connectionist Temporal Classification (CTC) [17] and cross-entropy objectives.

Towards a full-fledged CAPT system, as illustrated in Figure 1(c), this paper presents a novel end-to-end neural model, dubbed JAM (a **J**oint neural model for **A**PA and **M**DD), which leverages a unified architecture to streamline the phone dictation process and pronunciation assessments across multiple aspects (e.g., accuracy, stress, fluency) at various linguistic units (i.e., phone, word, and utterance). Specifically, JAM builds upon an iconic APA model, the goodness of pronunciation feature-based transformer (GOPT), and further integrates a phone-level predictor. When assessing a learner's speech, our model first creates timestamps for each phone segment in the phone-level text prompt using the Viterbi algorithm to align the text prompt with the learner's speech. Various pronunciation features are then extracted for each phone-level timestamp, and fed into a stack of transformer blocks to form contextualized pronunciation features. Based on these contextualized features, a predictor network generates a sequence of phonetic diagnostic feedback, while disparate regressors are employed to derive corresponding pronunciation scores from the phone-level to the utterance-level. To facilitate the modeling of fine-grained pronunciation cues and strengthen MDD performance, the proposed model incorporates electromagnetic articulography (EMA) features, which captures the movement of articulatory structures, such as the jaw, lips, and tongue. Comprehensive experiments conducted on the speechocean762 benchmark dataset show that our proposed method achieves significant and consistent improvements over several cutting-edge baselines. Additionally, an ablation study is conducted to examine the impact of different input features and training strategies on the performance of the proposed model.

II. METHODOLOGY

A. *Input Features*

The complete workflow is illustrated in Figure 2. We build our model based on GOPT, which processes an input utterance by first aligning the audio signals X with the corresponding text prompt T. This alignment step yields a sequence of phoneme-level goodness of pronunciation (GOP) features, denoted as $Emb_{GOP} = (Emb_{[p_1]}, Emb_{[p_2]}, ..., Emb_{[p_L]})$. Each individual GOP feature $Emb[p_l]$ is derived from a combination of log phone posterior (LPP) [19] and log posterior ratio (LPR) [20]. The resulting Emb_{GOP} is a matrix of size $84 \times L$, where L represents the length of the canonical phoneme sequence.

To enrich our model's input, we leverage pre-trained acoustic models (wav2vec 2.0 [21], Hubert [22], and WavLM [23]) to extract comprehensive acoustic features. Additionally, we

TABLE I
EXPERIMENTAL RESULTS OF DIFFERENT METHODS EVALUATED ON SPEECHOCEAN762.

Models		Phone-level Score		Word-level Score (PCC)			Utterance-level Score (PCC)					MDD Performance Score		
		MSE ↓	PCC ↑	ACC ↑	Stress ↑	Total ↑	ACC ↑	Completeness ↑	Fluency ↑	Prosody ↑	Total ↑	RE (%)	PR (%)	F1 (%)
APA methods	GOPT [9]	0.085 ±0.001	0.612 ±0.003	0.533 ±0.004	0.291 ±0.030	0.549 ±0.004	0.714 ±0.004	0.155 ±0.039	0.753 ±0.008	0.760 ±0.006	0.742 ±0.005	-	-	-
	HiPAMA [18]	0.084 ±0.001	0.616 ±0.004	0.575 ±0.004	**0.320 ±0.021**	0.591 ±0.004	0.730 ±0.002	0.276 ±0.177	0.749 ±0.001	0.751 ±0.002	0.754 ±0.002	-	-	-
	3M [3]	0.078 ±0.001	0.656 ±0.005	0.598 ±0.005	0.289 ±0.033	0.617 ±0.002	0.760 ±0.004	**0.325 ±0.141**	0.828 ±0.006	0.827 ±0.008	0.796 ±0.008	-	-	-
APA & MDD methods	Joint-CAPT-L1 [16]	-	-	-	-	-	0.719	-	0.775	0.773	0.743	**91.40**	26.70	41.40
	JAM	**0.076 ±0.002**	**0.664 ±0.001**	**0.622 ±0.012**	0.241 ±0.034	**0.638 ±0.005**	**0.773 ±0.007**	0.205 ±0.080	**0.831 ±0.004**	**0.829 ±0.004**	**0.805 ±0.004**	34.76	**64.10**	**45.01**

employ a pre-trained articulatory encoder [24] to obtain deep, EMA-like features from the last hidden layer, providing essential articulatory information for pronunciation assessment. All features are concatenated and projected into a 24-dimensional embedding Emb_{proj}.

B. *Transformer-based Architecture*

The projected acoustic features Emb_{proj}, canonical phone embeddings, and absolute positional embeddings are fed into a Transformer encoder. To obtain five aspects of utterance-level pronunciation scores in GOPT, five trainable [CLS] tokens are integrated with the phone-level input sequence, inspired by BERT [25].

Phoneme- and word-level regressors are added on top of the Transformer output of each corresponding phoneme. In addition, we use a classifier to predict each reference phoneme that the speaker pronounces.

C. *Optimization*

Our network is optimized using a Multi-Task Learning (MTL) framework. We employ Mean Squared Error (MSE) as the loss function for each granularity. The combined loss, denoted as $\mathcal{L}_{\text{APA}}$, is computed by directly summing the averaged losses at the utterance, word, and phoneme levels:

$$\mathcal{L}_{\text{APA}} = \mathcal{L}_{utterance} + \mathcal{L}_{word} + \mathcal{L}_{phoneme}, \tag{1}$$

where $\mathcal{L}_{utterance}$ and $\mathcal{L}_{word}$ are averaged utterance and word level losses of five utterance-level labels and three word-level labels, respectively; $\mathcal{L}_{phoneme}$ is the phoneme loss.

For the MDD task, the phone classifier is trained to increase the predicted probability for the correct phones by minimizing the cross-entropy loss, thus improving its accuracy in detecting mispronunciations:

$$\mathcal{L}_{\text{MDD}} = -\sum_{i=1}^{N}\sum_{p=1}^{P} y_{i,p}\log(\hat{y}_{i,p}), \tag{2}$$

where N is the number of samples, P is the number of phones, $y_{i,p}$ represents the true label for phone p in sample i (usually one-hot encoded), and $\hat{y}_{i,p}$ denotes the predicted probability of phone p for sample i.

The two loss are linearly combined as the overall objective for model learning:

$$\mathcal{L}_{CAPT} = \alpha\mathcal{L}_{MDD} + (1-\alpha)\mathcal{L}_{APA}, \tag{3}$$

where α are used to balance the two losses. α is chosen from the set of $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The value of α is set to be 0.3 based on the test set.

III. EXPERIMENTS AND RESULTS

A. *Dataset*

To evaluate our proposed method, we utilized the publicly available speechocean762 dataset [26] for the experiments. This dataset comprises 5,000 English speech recordings produced by 250 Mandarin L2 speakers, evenly split into training and test sets of 2,500 utterances each.

A key strength of speechocean762 is its comprehensive annotation framework. It provides human-rated scores at the utterance, word, and phoneme levels, each determined by five expert raters using consistent criteria. For training efficiency, these scores were normalized to a common 0-2 scale. Additionally, the dataset offers detailed mispronunciation transcriptions using a 39-phoneme set aligned with the CMU Pronunciation Dictionary [27]. This phoneme inventory was expanded to include <del> and <unk> symbols for omitted phonemes and non-categorial or distortion-like pronunciation, respectively.

B. *Experimental Setup*

Building upon the experimental setup detailed in [9], we employ the same DNN-HMM acoustic model to extract 84-dimensional GOP features. This model is based on a factorized time-delay neural network (TDNN-F) and trained on the Librispeech 960-hour dataset using the Kaldi recipe. To evaluate the efficacy of our proposed multi-view approach, we keep all training hyperparameters identical to the original GOPT. We conduct five independent experiments with different random seeds and report the mean and standard deviation of the Pearson Correlation Coefficient (PCC) and Mean Squared Error (MSE). For SSL features, we utilize pre-trained wav2vec 2.0, HuBERT, and WavLM models obtained from Hugging Face. Consistent with [3], we apply random dropout with probability *p* to each SSL feature before integration to mitigate overfitting.

TABLE II
ABLATION STUDY OF THE PROPOSED METHOD ON AUTOMATIC PRONUNCIATION ASSESSMENT.

Models	Phone-level Score		Word-level Score (PCC)			Utterance-level Score (PCC)				
	MSE ↓	PCC ↑	ACC ↑	Stress ↑	Total ↑	ACC ↑	Completeness ↑	Fluency ↑	Prosody ↑	Total ↑
JAM	**0.076**	**0.664**	**0.622**	0.241	**0.638**	**0.773**	**0.205**	**0.831**	**0.829**	**0.805**
w/o EMA feat	0.078	0.650	0.609	0.196	0.624	0.769	0.047	0.823	0.821	0.804
w/o SSL feat	0.083	0.624	0.550	**0.251**	0.568	0.724	0.155	0.783	0.780	0.754
w/o SSL, EMA feat	0.083	0.622	0.546	**0.251**	0.564	0.718	0.163	0.756	0.759	0.748

TABLE III
ABLATION STUDY OF PROPOSED METHOD ON MISPRONUNCIATION DETECTION AND DIAGNOSIS.

Models	RE (%) ↑	PR (%) ↑	F1 (%) ↑	FAR (%) ↓	FRR (%) ↓	DER (%) ↓	PER (%) ↓
JAM	**34.76**	64.10	**45.01**	**64.32**	0.58	**45.23**	2.81
w/o EMA feat.	33.61	65.48	44.38	66.39	0.51	46.47	**2.78**
w/o SSL feat.	19.23	**67.55**	29.87	80.77	**0.27**	51.54	2.79
w/o SSL, EMA feat.	20.12	66.89	30.91	79.88	0.29	52.43	2.80

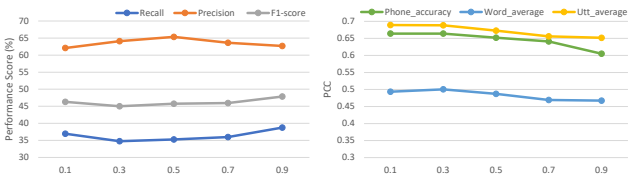


Fig. 3. Performance of different of loss weight α for MDD (left) and APA (right).

Following previous MDD work [28], the F1-score, a harmonic mean of precision (PR) and recall (RE), is commonly used as a measure of a model's accuracy. Specifically, phoneme mispronunciation detection performance is additionally evaluated by false rejection rate (FRR), false acceptance rate (FAR), and phoneme error rate (PER), as well as diagnostic error rate (DER).

C. Main Results

The performance of our framework and several state-of-the-art APA models is presented in Table I. Our proposed model JAM consistently competes with existing approaches across various evaluation metrics, demonstrating its superior ability to capture fine-grained phoneme-level errors and overall pronunciation proficiency. First, it is evident that our model consistently outperforms GOPT [9] and HiPAMA [18], which are recently proposed methods and can generalize well to the multi-aspect and multi-granular assessment, in almost metrics at the phone-level, word-level, and utterance-level, demonstrating its effectiveness on the speechocean762 dataset. Unfortunately, our proposed model exhibits poor performance in stress at the word-level and completeness at the utterance-level, likely due to data imbalance issues that need to be addressed. Secondly, while 3M [3] incorporates SSL and handcrafted features for prosody, our framework, which jointly trains APA and MDD and is augmented by EMA features, yields superior results. This underscores the benefits of integrating MDD and APA for pronunciation assessment. Finally, in the joint APA and MDD method, our proposed model JAM significantly exceeds Joint-CAPT-L1 [16] in APA performance and achieves higher F1-score. Although Joint-CAPT-L1 exhibits higher recall, its lower precision indicates a reduced ability to accurately detect correct pronunciations. Furthermore, the precision of 64% achieved by JAM indicates that our proposed model is more cautious when determining a pronunciation error. Only when the model is highly confident will it mark a pronunciation as incorrect.

D. Ablation Studies

To study the effectiveness of each feature involved in our model structure, we conducted an ablation study with the following settings: 1) removing EMA features, 2) removing self-supervised learning (SSL) features, and 3) removing EMA and SSL features simultaneously, leaving only GOP features for assessment. Table II and III shows the corresponding results. First, it can be observed that removing EMA features leads to a slight decrease in performance across most metrics in both automatic pronunciation assessment (APA) and mispronunciation detection and diagnosis (MDD). The decline indicates the importance of EMA features in capturing articulatory movements that contribute to the model's overall accuracy and effectiveness. Second, removing SSL features results in a more substantial drop in performance. This degradation is evident in the lower scores across both APA and MDD tasks. Third, when both EMA and SSL features are removed, leaving only GOP

features, the performance further declines. The results show significant decreases in various metrics, emphasizing the combined contribution of EMA and SSL features to the model's success. This highlights that relying solely on GOP features is insufficient for achieving optimal performance in both APA and MDD tasks. These findings underscore the significance of incorporating EMA and SSL features in our model to maintain high performance and accuracy in pronunciation assessment and mispronunciation detection and diagnosis.

Results on Mispronunciation Detection and Diagnosis. We further examined the diagnostic error rate (DER) in the ablation study. JAM achieved the lowest DER, indicating that EMA and SSL features enhance JAM's ability to detect the correct phoneme for erroneous pronunciation. However, JAM has a relatively lower FRR and higher FAR and PER. We trade off between FAR and FRR to increase the ability to correctly capture mispronunciations. These results indicate that the integration of EMA and SSL features is crucial for maintaining high performance and accuracy in pronunciation assessment and mispronunciation detection and diagnosis.

Effect of Loss Weight. Figure 3 illustrates the outcome of the MDD (left) and APA (right) under different loss weight α. A larger α indicates a greater loss weight assigned to the MDD. As shown in the figure, the F1-score of the MDD does not always increase monotonically with the increase of α. This suggests that simply increasing the loss weight of the MDD does not guarantee optimal performance on all metrics. In our main results, we selected $\alpha = 0.3$ as the optimal parameter, considering a trade-off between the PCC in phone accuracy and MDD effectiveness. At this setting, the model excels in phone accuracy while maintaining an acceptable performance in MDD.

IV. Conclusion

This paper presents JAM, an end-to-end neural model that integrates GOPT and a phone-level predictor, utilizing EMA features to significantly enhance phoneme recognition and pronunciation assessment performance. Experiments on the speechocean762 benchmark dataset show that JAM outperforms several advanced baselines, with an ablation study validating the impact of different features and strategies on its performance. In our future work, we plan to explore a hierarchical architecture and further enhance MDD performance as well as address the issues related to imbalanced data.

References

[1] P. Munday, "Duolingo. gamified learning through translation," *Journal of Spanish Language Teaching*, vol. 4, no. 2, pp. 194–198, 2017.

[2] A. Kholis, "Elsa speak app: Automatic speech recognition (asr) for supplementing english pronunciation skills," *Pedagogy: Journal of English Language Teaching*, vol. 9, no. 1, pp. 01–14, 2021.

[3] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 575–582. DOI: 10.23919/APSIPAASC55919.2022.9979979.

[4] B.-C. Yan, H.-W. Wang, Y.-C. Wang, and B. Chen, "Effective graph-based modeling of articulation traits for mispronunciation detection and diagnosis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[5] D. Korzekwa, J. Lorenzo-Trueba, T. Drugman, and B. Kostek, "Computer-assisted pronunciation training—speech synthesis is almost all you need," *Speech Communication*, vol. 142, pp. 22–33, 2022.

[6] B.-C. Yan, H.-W. Wang, and B. Chen, "Peppanet: Effective mispronunciation detection and diagnosis leveraging phonetic, phonological, and acoustic cues," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1045–1051. DOI: 10.1109/SLT54892.2023.10022472.

[7] W. Ye, S. Mao, F. Soong, *et al.*, "An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (apl) embeddings," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6827–6831.

[8] D. Y. Zhang, S. Saha, and S. Campbell, "Phonetic rnn-transducer for mispronunciation diagnosis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[9] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7262–7266.

[10] B.-C. Yan, H.-W. Wang, Y.-C. Wang, J.-T. Li, C.-H. Lin, and B. Chen, "Preserving phonemic distinctions for ordinal regression: A novel loss function for automatic pronunciation assessment," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7. DOI: 10.1109/ASRU57964.2023.10389777.

[11] W. Liu, K. Fu, X. Tian, *et al.*, "An asr-free fluency scoring approach with self-supervised learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[12] C. Zhu, T. Kunihara, D. Saito, N. Minematsu, and N. Nakanishi, "Automatic prediction of intelligibility of words and phonemes produced orally by japanese

learners of english," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 1029–1036.

[13] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 6234–6238.

[14] N. F. Chen, D. Wee, R. Tong, B. Ma, and H. Li, "Large-scale characterization of non-native mandarin chinese spoken by speakers of european origin: Analysis on icall," *Speech Communication*, vol. 84, pp. 46–56, 2016.

[15] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving mispronunciation detection of mandarin tones for non-native learners with soft-target tone labels and blstm-based deep tone models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2012–2024, 2019.

[16] H. Ryu, S. Kim, and M. Chung, "A joint model for pronunciation assessment and mispronunciation detection and diagnosis with multi-task learning," in *INTERSPEECH*, 2023.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[18] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[19] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[20] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[23] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[24] Y. M. Siriwardena, G. Sivaraman, and C. Espy-Wilson, "Acoustic-to-articulatory speech inversion with multi-task learning," *arXiv preprint arXiv:2205.13755*, 2022.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] J. Zhang, Z. Zhang, Y. Wang, *et al.*, "Speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[27] R. Weide *et al.*, "The carnegie mellon pronouncing dictionary," *release 0.6, www. cs. cmu. edu*, 1998.

[28] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2017. DOI: 10.1109/TASLP.2016.2621675.

Blind Self-Interference Analog Canceller with Differential Delay for Backscatter Communications

Koichi Nishikawa*, Shinsuke Ibi*, Takumi Takahashi†, and Hisato Iwai*
* Faculty of Science and Engineering, Doshisha University, 1-3 Tataramiyakodani, Kyotanabe, 610-0394, Japan
E-mail: ctwj0341@mail4.doshisha.ac.jp, {sibi, iwai}@mail.doshisha.ac.jp
† Graduate School of Engineering, Osaka University, Suita, Japan
E-mail: takahashi@comm.eng.osaka-u.ac.jp

***Abstract*—This paper demonstrates that the differential active self-interference canceller (DASIC), which does not require prior channel state information (CSI) of self-interference (SI), is effective in not only in-band full-duplex (IBFD) communications, but also backscatter communications. Unlike IBFD, backscatter communications have the characteristic that the desired signal is generated by nonlinear switching with rectangular pulse shape. Inspired by the characteristic, this paper proposes a method that enables cancellation in the analog domain by shortening the differential delay time of DASIC to a duration that can be implemented with a practical delay line. Additionally, focusing on the fact that the absolute value of the DASIC outputs remains constant regardless of the propagation delay time, an optimal maximum *a-posteriori* (MAP) detector is designed for scenarios in which the propagation delay time is unknown at the receiver. Computer simulations validate the efficacy of the proposed method by comparing the performance of MAP detectors when the propagation delay time is known and unknown, from the viewpoint of respect to bit error rate (BER) performance.**

I. Introduction

With the rapid proliferation of Internet of Things (IoT) communications in recent years, there has been an increasing demand for low-power wireless communications to achieve a sustainable society. Bluetooth low energy (BLE) is a representative low-power communication method. However, even the BLE requires the generation of a carrier wave during transmission, which consumes tens of milliwatts of power. On the other hand, passive backscatter device (BD) performs energy harvesting (EH) from the received signals and convey information data by reflecting or absorbing the received signals. Therefore, BD does not need to generate a carrier wave, enabling ultra-low-power wireless communications, consuming only tens of microwatts, which has garnered significant attention [1]–[3].

Fig. 1 shows a typical configuration of backscatter communications from node B to node A. In this scenario, node B (BD) receives the signal sent from node A to node C, and sends its own information to node A by reflecting or absorbing the received signal. However, node A receives not only the desired signal from the BD but also its own transmission signal at a much higher power level as self-interference (SI). To address this issue for backscatter communications, the utilization of in-band full-duplex (IBFD) is a promising technique that allows simultaneous transmission and reception of signals within the same frequency band [4]–[6].

Typically, SI is mitigated through a combination of passive and active cancellation techniques [7]–[9]. Passive cancellers enhance the path loss of the SI signal using specially designed antenna structures. On the other hand, active cancellers are divided into analog cancellers, which operate before the analog-to-digital converter (ADC), and digital cancellers, which work after the ADC [10]. The active canceller minimizes SI by subtracting SI replica from the received signal. Ideally, in the active digital SI canceller (ASIC), the replica is generated flawlessly as the signal at the receiver is completely known. However, generating this replica requires periodic transmissions of orthogonal pilot signals to acquire estimate channel state information (CSI). For example, in wireless IoT communications which are random and intermittently requested, pilot signals must be transmitted even in the absence of communication requests, leading to inefficient wireless communication systems. Differential active SI canceller (DASIC) has been proposed, which eliminates the need for prior channel estimation of SI [11], to circumvent the increased overhead required for the acquisition of the CSI.

Additionally, when DASIC are implemented as a digital canceller, there is a problem where the desired signal cannot be extracted if the power level of the signal before the ADC is not within the ADC's dynamic range. However, to address the dynamic range issue by operating DASIC in the analog domain, a long delay line corresponding to the symbol period is required, which presents a significant challenge for implementations. In this paper, we propose a novel method that does not impair the effectiveness of DASIC operated in the analog domain, even with extremely short differential delays, by taking advantage of the characteristics of backscattered signals.

The key contributions of this study can be summarized as follows:

- After formulating the SI structure, we propose a method for applying DASIC to the analog domain by shortening the differential delay time.
- The probability density function (PDF) for the unknown propagation delay time from the BD is defined, and an optimal maximum *a-posteriori* probability (MAP) detector that operates even when the delay time is unknown is designed.
- The efficacy of the proposed method is illustrated by

comparing the bit error rate (BER) performance for known and unknown propagation delay times.

Notations: The symbol discrete time and the continuous time are denoted by $[k]$ and (t), respectively. The real and complex number fields are represented by $\mathbb{R}$ and $\mathbb{C}$, respectively. $(\cdot)^*$ represents complex conjugate. $\Re\{\cdot\}$ represents the real part of complex numbers. The imaginary unit $\sqrt{-1}$ is represented by j. $\mathbb{E}\{\cdot\}$ is the expected value.

II. System Model

A. Configuration of IBFD transceiver

Fig. 1 illustrates the configuration of the IBFD transceiver. On the transmitter side, Gaussian minimum shift keying (GMSK) modulates information bits, and then the signals are converted into analog signals by a digital-to-analog converter (DAC). Subsequently, an IQ mixer up-converts the analog signals into radio frequency (RF) signals. The RF signals are filtered through a band-pass filter (BPF). The resultant signals are transmitted to the receiver. Conversely, on the receiver side, incoming signals are filtered through a BPF, down-converted using an IQ mixer, and then passed through a low-pass filter (LPF). The SI components present in analog signals are blindly removed by the DASIC circuit. The DASIC outputs are then transformed into digital signals by ADC. Eventually, the information bits are detected by the demodulator based on the DASIC outputs.

B. Signal model of BLE

The RF signal $s_{\mathrm{A}}(t) \in \mathbb{R}$ transmitted from node A is represented as

$$s_{\mathrm{A}}(t) = \sqrt{2}\alpha\Re\{x_{\mathrm{A}}(t)\exp[\mathrm{j}\omega_{\mathrm{c}}t]\}, \tag{1}$$

where α represents the amplitude of RF signals, and ω_{c} denotes the central angular frequency. The complex-valued baseband signal $x_{\mathrm{A}}(t) \in \mathbb{C}$ is expressed as

$$x_{\mathrm{A}}(t) = \exp\left[\mathrm{j}\phi_{\mathrm{A}}(t)\right], \tag{2}$$

where $\phi_{\mathrm{A}}(t)$ is an information-bearing variable.

In this paper, assuming IoT communications between BLE devices, GMSK is used [12]. Denoting the symbol period by T_{s}, within the interval $kT_{\mathrm{s}} \le t\ (= t' + kT_{\mathrm{s}}) < (k+1)T_{\mathrm{s}}$, the phase variable $\phi_{\mathrm{A}}(t)$ of GMSK can be expressed as [13], [14]

$$\phi_{\mathrm{A}}(t) = \pi \sum_{l=k-L+1}^{k} a_{\mathrm{A}}[l]q(t' + (k-l)T_{\mathrm{s}}) + \frac{\pi}{2}\sum_{l'=0}^{k-L} a_{\mathrm{A}}[l'], \tag{3}$$

where $a_{\mathrm{A}}[k] \in \{-1, 1\}$ represents an information bit in a bipolar form, and L denotes the memory size depending on the 3 dB bandwidth B_{G} of the Gaussian filter and the bandwidth-time product $B_{\mathrm{G}}T_{\mathrm{s}}$. The function $q(t)$ is a phase pulse, which is defined by

$$q(t) = \int_{-\infty}^{t} g(\zeta)d\zeta, \tag{4}$$

under a constraint of

$$q(t) = \begin{cases} 0 & t \le 0, \\ 1/2 & t > LT_{\mathrm{s}}. \end{cases} \tag{5}$$

The function $g(\zeta)$ is a frequency pulse given by

$$g(\zeta) = \frac{1}{2T_{\mathrm{s}}}\left[Q\left(2\pi B_{\mathrm{G}}\frac{\zeta - T_{\mathrm{s}}/2}{\sqrt{\ln 2}}\right) - Q\left(2\pi B_{\mathrm{G}}\frac{\zeta + T_{\mathrm{s}}/2}{\sqrt{\ln 2}}\right)\right], \tag{6}$$

where $Q(\cdot)$ is Q function.

C. Signal model of BD

Fig. 2 exhibits the configuration of the BD systems. The BD performs EH from the RF signal $s_{\mathrm{A}}(t)$ transmitted from node A to node C. Using a micro controller, the impedance of the antenna is switched, reflecting $s_{\mathrm{A}}(t)$ when the BD's information bit $a'_{\mathrm{B}}[k] = 1$, and absorbing it when $a'_{\mathrm{B}}[k] = 0$, where $a'_{\mathrm{B}}(t) \in \{0, 1\}$ is a pulse signal that can take values of 1 or 0. Consider the scenario in which node A receives a message from the BD. Let the propagation time for the signal transmitted from node A to be reflected by the BD and returned to the receiver of node A be τ $(0 \le \tau \le T_{\mathrm{s}})$. The received signal after passing through the LPF at node A is expressed as

$$r_{\mathrm{A}}(t) = h_{\mathrm{AA}}\alpha x_{\mathrm{A}}(t) + a'_{\mathrm{B}}(t)h_{\mathrm{AB}}\alpha x_{\mathrm{A}}(t-\tau) + z_{\mathrm{A}}(t), \tag{7}$$

where h_{AA} and h_{AB} represent the channel coefficients for the path A–A and A–B, respectively. $z_{\mathrm{A}}(t)$ denote a noise term that follows a complex Gaussian random process with a mean of 0 and a variance of $\beta = N_0/T_{\mathrm{s}}$, where N_0 is the noise power spectral density in a symbol duration. In (7), the first term represents the SI component, whereas the second term exhibits the desired signal component.

III. DASIC for Backscatter Communications

A. Principle of DASIC

Let us consider operations of DASIC at node A when BD sends backscatter signal to node A. Fig. 3 shows the schematic of DASIC circuit with designated differential delay parameter d $(0 < d + \tau < T_{\mathrm{s}})$. The delay can be caused by using delay line, for example 10 ns[1]. With the known baseband signals $x_{\mathrm{A}}(t)$ and delayed $x_{\mathrm{A}}(t-d)$, the phase shift $\varphi_{\mathrm{A}}(t)$ used in the DASIC process is defined as

$$\varphi_{\mathrm{A}}(t) = \frac{x_{\mathrm{A}}(t)}{x_{\mathrm{A}}(t-d)}. \tag{8}$$

To obtain the DASIC output, $r_{\mathrm{A}}(t-d)$ is multiplied by the phase shift $\varphi_{\mathrm{A}}(t)$, and the result is subtracted from $r_{\mathrm{A}}(t)$ as

$$\begin{aligned} \tilde{r}_{\mathrm{A}}(t) &= r_{\mathrm{A}}(t) - r_{\mathrm{A}}(t-d)\varphi_{\mathrm{A}}(t) \\ &= h_{\mathrm{AB}}\alpha\tilde{x}_{\mathrm{A}}(t-\tau) + \tilde{z}_{\mathrm{A}}(t), \end{aligned} \tag{9}$$

[1]This differs from the conventional DASIC of [11], which uses the symbol time T_{s} as the differential delay time for continuous GMSK signaling from node B to node A, making it difficult to implement with a delay line.

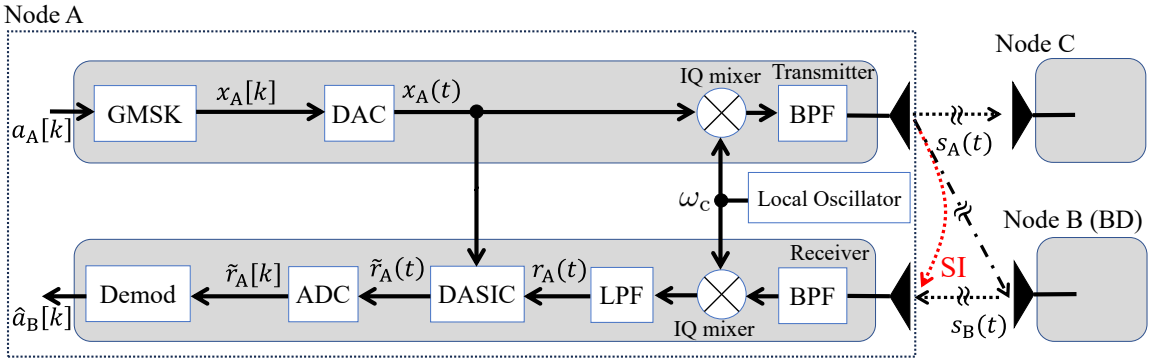


Fig. 1. Schematic of DASIC-IBFD transceiver.

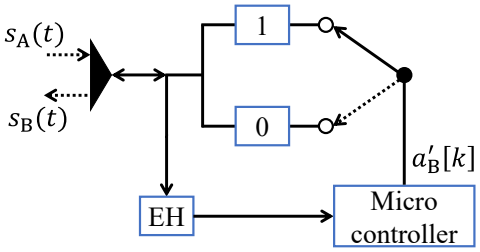


Fig. 2. Block diagram of BD.

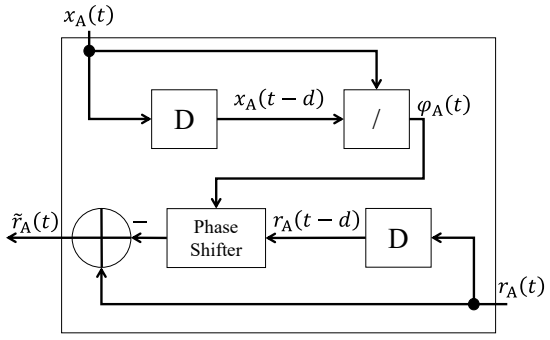


Fig. 3. Block diagram of DASIC process.

where the first term represents the desired signal component of the BD information, which is defined as

$$\tilde{x}_{\mathrm{A}}(t-\tau) = a'_{\mathrm{B}}(t)x_{\mathrm{A}}(t-\tau) - a'_{\mathrm{B}}(t-d)x_{\mathrm{A}}(t-\tau-d)\varphi_{\mathrm{A}}(t), \quad (10)$$

and the other term is noise component, which is expressed as

$$\tilde{z}_{\mathrm{A}}(t) = z_{\mathrm{A}}(t) - z_{\mathrm{A}}(t-d)\varphi_{\mathrm{A}}(t). \quad (11)$$

Note that nonlinear switching with rectangular pulse shape of BD signaling enables the avoidance of shrinking of the signal $\tilde{x}_{\mathrm{A}}(t-\tau)$. If $a'_{\mathrm{B}}(t) \approx a'_{\mathrm{B}}(t-d)$ as in the conventional DASIC, it is obvious that (10) disappears. By utilizing the DASIC for BD systems, the SI component can be eliminated completely in (9) without shrinking $\tilde{x}_{\mathrm{A}}(t-\tau)$.

Assuming the ideal sampling time at $t = kT_{\mathrm{s}} + \Delta t$ that maximizes the eye aperture and $a'_{\mathrm{B}}(t-d) = a'_{\mathrm{B}}[k-1]$, the ADC of (9) yields

$$\tilde{r}_{\mathrm{A}}[k] = h_{\mathrm{AB}}\alpha\tilde{x}_{\mathrm{A}}[k,\tau] + \tilde{z}_{\mathrm{A}}[k], \quad (12)$$

where we have

$$\tilde{x}_{\mathrm{A}}[k,\tau] = a'_{\mathrm{B}}[k]x_{\mathrm{A}}[k,\tau] - a'_{\mathrm{B}}[k-1]x_{\mathrm{A}}[k,\tau+d]\varphi_{\mathrm{A}}[k], \quad (13)$$

$$\tilde{z}_{\mathrm{A}}[k] = z_{\mathrm{A}}[k] - z_{\mathrm{A}}[k,d]\varphi_{\mathrm{A}}[k], \quad (14)$$

and $x_{\mathrm{A}}[k,\tau]$ is the baseband signal which is sampled in the presence of the influence of the fractional propagation delay τ. Note that $\mathbb{E}\{|\tilde{z}_{\mathrm{A}}[k]|^2\}$ is2β because (14) has two noise variables.

When the memory size is $L = 3$ as an example, the constituent components $x_{\mathrm{A}}[k,\tau]$, $x_{\mathrm{A}}[k,\tau+d]$, and $\varphi_{\mathrm{A}}[k]$ of $\tilde{x}_{\mathrm{A}}[k,\tau]$ can be represented as

$$x_{\mathrm{A}}[k,\tau] = \exp\Big\{\mathrm{j}\omega_0(\tau)a_{\mathrm{A}}[k] + \mathrm{j}\omega_1(\tau)a_{\mathrm{A}}[k-1] + \mathrm{j}\omega_2(\tau)a_{\mathrm{A}}[k-2] + \mathrm{j}\sigma_{\mathrm{A}}[k-3]\Big\}, \quad (15)$$

$$x_{\mathrm{A}}[k,\tau+d] = \exp\Big\{\mathrm{j}\omega_0(\tau+d)a_{\mathrm{A}}[k] + \mathrm{j}\omega_1(\tau+d)a_{\mathrm{A}}[k] + \mathrm{j}\omega_2(\tau+d)a_{\mathrm{A}}[k-2] + \mathrm{j}\sigma_{\mathrm{A}}[k-3]\Big\}, \quad (16)$$

$$\varphi_{\mathrm{A}}[k] = \exp\Big\{\mathrm{j}(\omega_0(0) - \omega_0(d))a_{\mathrm{A}}[k] + \mathrm{j}(\omega_1(0) - \omega_1(d))a_{\mathrm{A}}[k-1] + \mathrm{j}(\omega_2(0) - \omega_2(d))a_{\mathrm{A}}[k-2]\Big\}, \quad (17)$$

where we have

$$\omega_l(\tau) = \pi q(\Delta t - \tau + lT_{\mathrm{s}}), (l \in \{0,1,2\}), \quad (18)$$

$$\sigma_{\mathrm{A}}[k] = \frac{\pi}{2}\sum_{l=0}^{k} a_{\mathrm{A}}[l]. \quad (19)$$

B. Threshold for simple successive hard decision

Hereafter, to simplify the discussion, the case of MSK ($B_{\mathrm{G}}T_{\mathrm{s}} = \infty$) is considered[2]. In this case, sampling at the

[2]It has already been confirmed that DASIC is effective for SI cancellation even with modulation schemes other than MSK, such as QAM.

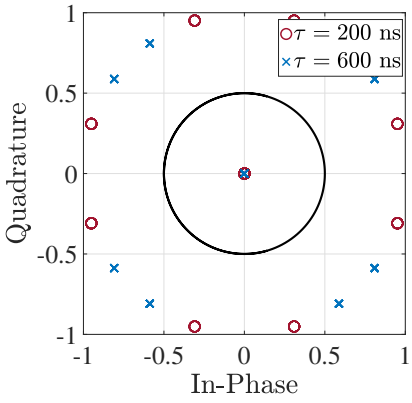


Fig. 4. Constellation of $\tilde{r}_{\mathrm{A}}[k]$: $h_{\mathrm{AB}}\alpha = 1, \tilde{z}_{\mathrm{A}}[k] = 0$.

timing $t' = \Delta t$ when the eye aperture is maximized results in

$$\omega_0(\tau) = 0,\ \omega_1(\tau) = \frac{T_{\mathrm{s}} - \tau}{2T_{\mathrm{s}}}\pi,\ \omega_2(\tau) = \frac{\pi}{2}. \quad (20)$$

Using (20), (15)∼(17) can be rewritten as

$$x_{\mathrm{A}}[k,\tau] = \exp\left\{\mathrm{j}\frac{T_{\mathrm{s}} - \tau}{2T_{\mathrm{s}}}\pi a_{\mathrm{A}}[k-1] + \mathrm{j}\sigma_{\mathrm{A}}[k-2]\right\}, \quad (21)$$

$$x_{\mathrm{A}}[k,\tau+d] = \exp\left\{\mathrm{j}\frac{T_{\mathrm{s}} - \tau - d}{2T_{\mathrm{s}}}\pi a_{\mathrm{A}}[k-1] + \mathrm{j}\sigma_{\mathrm{A}}[k-2]\right\}, \quad (22)$$

$$\varphi_{\mathrm{A}}[k] = \exp\left\{\mathrm{j}\frac{d}{2T_{\mathrm{s}}}\pi a_{\mathrm{A}}[k-1]\right\}. \quad (23)$$

By defining $\tilde{a}_{\mathrm{B}}[k] = a'_{\mathrm{B}}[k] - a'_{\mathrm{B}}[k-1] \in \{0, \pm 1\}$, (13) can be interpreted as

$$\begin{aligned}\tilde{x}_{\mathrm{A}}[k,\tau] &= a'_{\mathrm{B}}[k]\exp\left\{\mathrm{j}\frac{T_{\mathrm{s}} - \tau}{2T_{\mathrm{s}}}\pi a_{\mathrm{A}}[k-1] + \mathrm{j}\sigma_{\mathrm{A}}[k-2]\right\} \\ &\quad - a'_{\mathrm{B}}[k-1]\exp\left\{\mathrm{j}\frac{T_{\mathrm{s}} - \tau - d}{2T_{\mathrm{s}}}\pi a_{\mathrm{A}}[k-1]\right. \\ &\quad \left. + \mathrm{j}\sigma_{\mathrm{A}}[k-2] + \mathrm{j}\frac{d}{2T_{\mathrm{s}}}\pi a_{\mathrm{A}}[k-1]\right\} \\ &= (a'_{\mathrm{B}}[k] - a'_{\mathrm{B}}[k-1]) \\ &\quad \cdot \exp\left\{\mathrm{j}\frac{T_{\mathrm{s}} - \tau}{2T_{\mathrm{s}}}\pi a_{\mathrm{A}}[k-1] + \mathrm{j}\sigma_{\mathrm{A}}[k-2]\right\} \\ &= \tilde{a}_{\mathrm{B}}[k]x_{\mathrm{A}}[k,\tau] \quad (24) \\ &= \begin{cases} \pm x_{\mathrm{A}}[k,\tau] & (\tilde{a}_{\mathrm{B}}[k] \neq 0), \\ 0 & (\tilde{a}_{\mathrm{B}}[k] = 0). \end{cases}\end{aligned}$$

As shown in Fig. 4, because $|\tilde{x}_{\mathrm{A}}[k,\tau]| = 1$ does not depend on τ, if $a'_{\mathrm{B}}[k-1]$ is known by successive detection, simple hard decision is possible even when τ is unknown. Therefore, the threshold value for hard decision is set to $|h_{\mathrm{AB}}|\alpha/2$, and decisions are successively made from $k = 1$ as

$$\hat{a}'_{\mathrm{B}}[k] = \begin{cases} \hat{a}'_{\mathrm{B}}[k-1] & (|\tilde{r}_{\mathrm{A}}[k]| \leq |h_{\mathrm{AB}}|\alpha/2), \\ -\hat{a}'_{\mathrm{B}}[k-1] + 1 & (|\tilde{r}_{\mathrm{A}}[k]| > |h_{\mathrm{AB}}|\alpha/2). \end{cases} \quad (25)$$

Thus, unlike IBFD of [11], signal detection is possible in backscatter communications even with an extremely short differential delay d.

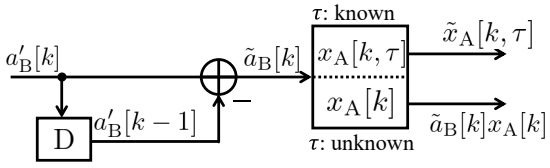


Fig. 5. Structure of IIR filter in DASIC when τ is known and unknown.

C. MAP detector with knowledge on τ

In the case of successive hard detection, the problem is that performance degrades due to the propagation of decision errors. Therefore, let us examine MAP detector, which can achieve optimum performance according to probabilistic theory to deal with the error propagation problem. Within the MAP framework, the log-likelihood ratio (LLR) is defined as

$$\lambda_{\mathrm{B}}[k] = \ln\frac{\Pr[a'_{\mathrm{B}}[k] = 1 \mid \{\tilde{r}_{\mathrm{A}}[k]\}]}{\Pr[a'_{\mathrm{B}}[k] = 0 \mid \{\tilde{r}_{\mathrm{A}}[k]\}]}, \quad (26)$$

where $\{\tilde{r}_{\mathrm{A}}[k]\}$ denotes a sequence of $\tilde{r}_{\mathrm{A}}[k]$. Assuming that τ is known, the structure of the infinite impulse response (IIR) filter used for generating $\tilde{x}_{\mathrm{A}}[k,\tau]$ is illustrated in Fig. 5 with memory size of 1. In this case, the trellis diagram has two states because $a'_{\mathrm{B}}[k-1]$ has zero or one. Therefore, the complexity is not so high.

Based on the trellis diagram of the IIR filter, the LLR can be calculated using the BCJR algorithm [15], without requiring a significant computational burden. The transition (joint) probability required for the BCJR process is given by

$$\Pr[\tilde{a}_{\mathrm{B}}[k] \mid \tilde{r}_{\mathrm{A}}[k]] \propto \exp\left\{-\frac{|\tilde{r}_{\mathrm{A}}[k] - h_{\mathrm{AB}}\alpha\tilde{a}_{\mathrm{B}}[k]x_{\mathrm{A}}[k,\tau]|^2}{2\beta}\right\}. \quad (27)$$

D. MAP detector without knowledge on τ

In the previous subsection, the MAP detector is designed by assuming that the propagation delay time τ is known. However, if node A or BD is moving, it is difficult to stably estimate τ, making τ unknown. Therefore, in this subsection, a MAP detector capable of detection even when τ is unknown is designed.

Using (21) and (24), $\tilde{x}_{\mathrm{A}}[k,\tau]$ can be rewritten as

$$\tilde{x}_{\mathrm{A}}[k,\tau] = \tilde{a}_{\mathrm{B}}[k]x_{\mathrm{A}}[k]\exp\left\{-\mathrm{j}\theta_\tau a_{\mathrm{A}}[k-1]\right\}, \quad (28)$$

where we have $\theta_\tau = \pi\tau/2T_{\mathrm{s}}$ $(0 < \theta_\tau < \pi/2)$. Then assuming that θ_τ is uniformly distributed, the PDF of $\tilde{r}_{\mathrm{A}}[k]$ when τ is

unknown is given by

$$
\begin{aligned}
&p(\tilde{r}_{\mathrm{A}}[k] \mid \tilde{a}_{\mathrm{B}}[k], h_{\mathrm{AB}}, x_{\mathrm{A}}[k]) \\
&= \int_0^{\frac{\pi}{2}} p(\theta_\tau) p(\tilde{r}_{\mathrm{A}}[k] \mid \tilde{a}_{\mathrm{B}}[k], h_{\mathrm{AB}}, x_{\mathrm{A}}[k,\tau]) d\theta_\tau \\
&= \int_0^{\frac{\pi}{2}} \frac{2}{\pi} \cdot \frac{1}{2\pi\beta} \exp\left\{ -\frac{|\tilde{r}_{\mathrm{A}}[k] - h_{\mathrm{AB}}\alpha\tilde{a}_{\mathrm{B}}[k] x_{\mathrm{A}}[k,\tau]|^2}{2\beta} \right\} d\theta_\tau \\
&= \frac{1}{\pi^2\beta} \exp\left[-\frac{1}{2\beta} \left\{ |\tilde{r}_{\mathrm{A}}[k]|^2 + |h_{\mathrm{AB}}\alpha\tilde{a}_{\mathrm{B}}[k]|^2 \right\} \right] \\
&\quad \cdot \int_0^{\frac{\pi}{2}} \exp\left[\frac{1}{\beta} \Re\left\{ \tilde{r}_{\mathrm{A}}^*[k] h_{\mathrm{AB}} \alpha x_{\mathrm{A}}[k,\tau] \tilde{a}_{\mathrm{B}}[k] \right\} \right] d\theta_\tau. \qquad (29)
\end{aligned}
$$

Then, using (28), the component of (29) can be rewritten as

$$
\begin{aligned}
&\Re\left\{ \tilde{r}_{\mathrm{A}}^*[k] h_{\mathrm{AB}} \alpha x_{\mathrm{A}}[k,\tau] \tilde{a}_{\mathrm{B}}[k] \right\} \\
&\quad = \Re\left\{ |\tilde{r}_{\mathrm{A}}^*[k] h_{\mathrm{AB}} \alpha \tilde{a}_{\mathrm{B}}[k]| \exp\left[\mathrm{j}(-\theta_\tau a_{\mathrm{A}}[k-1] + \psi) \right] \right\} \\
&\quad = |\tilde{r}_{\mathrm{A}}^*[k] h_{\mathrm{AB}} \alpha \tilde{a}_{\mathrm{B}}[k]| \cos(-\theta_\tau a_{\mathrm{A}}[k-1] + \psi), \qquad (30)
\end{aligned}
$$

where we have ψ $(-\pi < \psi \leq \pi)$ as

$$
\psi = \arg\left\{ \tilde{r}^*[k] h_{\mathrm{AB}} \alpha \tilde{a}_{\mathrm{B}}[k] x_{\mathrm{A}}[k] \right\}. \qquad (31)
$$

Finally, applying (30) to (29) yields

$$
\begin{aligned}
&p(\tilde{r}_{\mathrm{A}}[k] \mid \tilde{a}_{\mathrm{B}}[k], h_{\mathrm{AB}}, x_{\mathrm{A}}[k]) \\
&\qquad \propto \xi(\psi) \exp\left\{ -\frac{|h_{\mathrm{AB}}\alpha\tilde{a}_{\mathrm{B}}[k]|^2}{2\beta} \right\}, \qquad (32)
\end{aligned}
$$

where we have

$$
\begin{aligned}
&\xi(\psi) = a_{\mathrm{A}}[k-1] \\
&\quad \cdot \int_{-\frac{\pi}{2} a_{\mathrm{A}}[k-1] + \psi}^{\psi} \exp\left[\frac{|\tilde{r}_{\mathrm{A}}^*[k] h_{\mathrm{AB}} \alpha \tilde{a}_{\mathrm{B}}[k]|}{\beta} \cos\theta'_\tau \right] d\theta'_\tau, \qquad (33)
\end{aligned}
$$

with $\theta'_\tau = -\theta_\tau a_{\mathrm{A}}[k-1] + \psi$.

The transition (joint) probability for BCJR is given as

$$
\Pr[\tilde{a}_{\mathrm{B}}[k] \mid \tilde{r}_{\mathrm{A}}[k]] \propto \xi(\psi) \exp\left\{ -\frac{|h_{\mathrm{AB}}\alpha\tilde{a}_{\mathrm{B}}[k]|^2}{2\beta} \right\}. \qquad (34)
$$

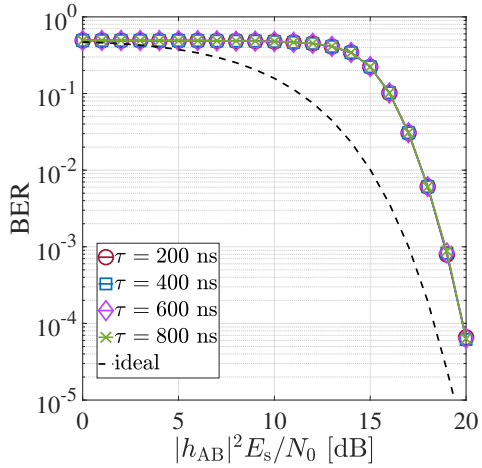
In (34), $\xi(\psi)$ is computed by piece-wise quadrature method[3].

IV. Simulation Results

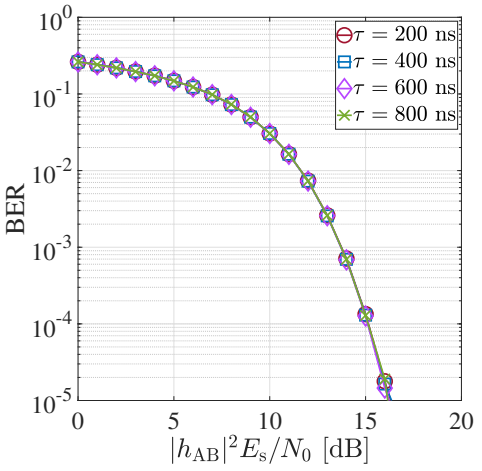
Computer simulations were conducted to validate the efficacy of the proposed method. The modulation scheme is GMSK ($B_{\mathrm{G}}T_{\mathrm{s}} = 0.5, L = 3$), assuming that BLE is used. Each information block consists of 64 bits. The symbol period and the differential delay are $T_{\mathrm{s}} = 1$ μs and $d = 10$ ns, respectively. The channel coefficient h_{AB} is assumed to be known at the receiver, whereas h_{AA} is unknown. Let η be signal-to-interference power ratio (SIR), which is defined as

$$
\eta = 10 \log_{10} \frac{|h_{\mathrm{AB}}|^2}{|h_{\mathrm{AA}}|^2} \quad [\mathrm{dB}]. \qquad (35)
$$

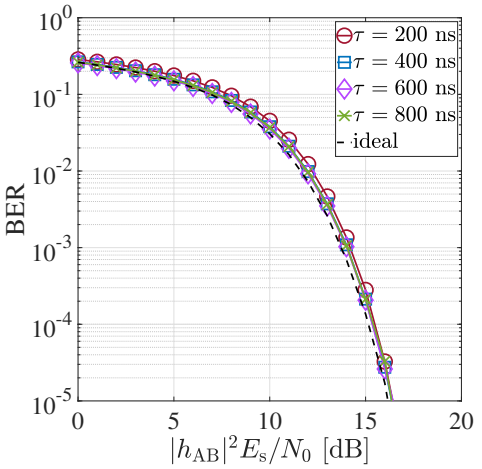
Fig. 6 shows the BER versus $|h_{\mathrm{AB}}|^2 E_{\mathrm{s}}/N_0$ performance. Although the channels follow Rayleigh fading, the horizontal axis represents the instantaneous $|h_{\mathrm{AB}}|^2 E_{\mathrm{s}}/N_0$ rather than the average. As a result, it is equivalent to a channel influenced by phase rotation without amplitude fluctuations.



(a) Successive hard detection.



(b) MAP decision: τ is known.



(c) MAP decision: τ is unknown.

Fig. 6. BER performance of DASIC: $\eta = -50$ dB.

Fig. 6(a) shows the BER performance of the successive hard decision. The "ideal" curve represents the performance assuming that all successive estimates are ideally correct

[3]Since it is possible to prepare a look-up-table of the integration results and refer to LUT during calculations, the computational load can be reduced.

($\hat{a}'_{\rm B}[k-1] = a'_{\rm B}[k-1]$). The performance remains BER ≥ 0.34 until $|h_{\rm AB}|^2 E_{\rm s}/N_0 = 14$ dB, but starts to decrease at $|h_{\rm AB}|^2 E_{\rm s}/N_0 = 15$ dB, reaching BER $< 10^{-4}$ at $|h_{\rm AB}|^2 E_{\rm s}/N_0 = 19.8$ dB, whereas "ideal" achieves BER $< 10^{-4}$ at $|h_{\rm AB}|^2 E_{\rm s}/N_0 = 18.3$ dB. This degradation is due to error propagation during successive detection. However, in the case of hard decision, the performance is consistent regardless of the propagation delay time τ, indicating that signal detection is possible even when τ is unknown.

Fig. 6(b) shows the performance of the MAP detection using (27) when τ is known. Because MAP detector is designed assuming τ is known, the performance is consistent regardless of τ. As expected, the performance degradation caused by successive hard decision is improved. Remarkably, the required $|h_{\rm AB}|^2 E_{\rm s}/N_0$ to satisfy BER $< 10^{-4}$ is 15.1 dB, improving the performance in hard decision by 4.7 dB. However, such a good performance, using (27), cannot be achieved unless τ is known.

On the other hand, Fig. 6(c) shows the performance of MAP detection using (34) when τ is unknown. The "ideal" curve represents the performance in Fig. 6(b) when τ is known. The required $|h_{\rm AB}|^2 E_{\rm s}/N_0$s to satisfy BER $< 10^{-2}$ at $\tau = 200$, $\tau = 400$, $\tau = 600$ and $\tau = 800$ ns are 12.2 dB, 12.0 dB, 12.0 dB, 11.9 dB, respectively, exhibiting slightly different performance depending on τ. This is because (34) is designed assuming MSK, and the Euclidean distance for $\tilde{a}_{\rm B}[k] = \pm 1$ varies depending on τ. Remarkably, BER $\leq 10^{-5}$ is achieved at $|h_{\rm AB}|^2 E_{\rm s}/N_0 = 16.4$ dB compared with $|h_{\rm AB}|^2 E_{\rm s}/N_0 = 16.2$ dB in the "ideal" case, indicating that while the performance degrades when τ is unknown, the degradation is limited to within 0.2 dB. Therefore, it can be concluded that MAP detection when τ is unknown is the most effective method for practical environments.

V. Conclusion

To address the problem of SI cancellation being required in backscatter communications just as in IBFD, this paper demonstrated that DASIC, which has been effective in IBFD, is also effective in backscatter communications. Additionally, by leveraging the characteristics of backscatter communications, where the desired signal is the delayed path of the terminal itself, the differential delay is shortened, enabling the application of DASIC in the analog domain. Furthermore, a MAP detector that operates even when τ is unknown is designed. The results show only a 0.2 dB degradation compared to the case where τ is known, concluding that this is the most practical approach.

Acknowledgment

A part of this work was supported by JSPS KAKENHI Grant Number JP23K20935 and JP23K22754, Japan.

References

[1] M. A. Alim, S. Saruwatari, and T. Watanabe, "Backscatter mac protocol for future internet of things networks," in *2017 IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2017, pp. 1–7.

[2] M. Zhang, J. Zhao, S. Chen, and W. Gong, "Reliable backscatter with commodity BLE," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020, pp. 1291–1299.

[3] B. Kellogg, A. Parks, S. Gollakota, J. R. Smith, and D. Wetherall, "Wi-fi backscatter: Internet connectivity for RF-powered devices," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 607–618, Aug. 2014.

[4] D. Bharadia, E. McMilin, and S. Katti, "Full duplex radios," *ACM SIGCOMM 2013*, pp. 375–386, Aug. 2013.

[5] Z. Zhang, K. Long, A. V. Vasilakos, and L. Hanzo, "Full-duplex wireless communications: Challenges, solutions and future research directions," *in Proc.IEEE*, vol. 104, pp. 1369–1409, Jul. 2016.

[6] K. E. Kolodziej, B. T. Perry, and J. S. Herd, "In-band full-duplex technology: Techniques and systems survey," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 7, pp. 3025–3041, 2019.

[7] Z. Zhang, X. Chai, K. Long, A. V. Vasilakos, and L. Hanzo, "Full duplex techniques for 5G networks: Self-interference cancellation, protocol design, and relay selection," *IEEE Communications Magazine*, vol. 53, no. 5, pp. 128–137, 2015.

[8] A. Sahai, G. Patel, C. Dick, and A. Sabharwal, "On the impact of phase noise on active cancelation in wireless full-duplex," *IEEE Trans. Veh. Commun.*, vol. 62, no. 9, pp. 4494–4510, Nov. 2013.

[9] E. Everett, A. Sahai, and A. Sabharwal, "Passive self-interference suppression for full-duplex infrastructure nodes," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 680–694, Feb. 2014.

[10] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 9, pp. 1637–1652, 2014.

[11] S. Ibi, T. Takahashi, and H. Iwai, "Differential active self-interference cancellation for asynchronous in-band full-duplex GFSK," *IEICE Trans. Commun.*, vol. E107-B, no. 8, pp. 552–563, Aug. 2024.

[12] Core Specification Working Group, *Bluetooth core specification*, 2021.

[13] R. H.-H. Yang, C.-K. Lee, and S.-J. Chern, "Performance improvement of the catastrophic CPM scheme with new split-merged MNSED," *IEICE Trans. Commun.*, vol. E102-B, no. 11, pp. 2091–2103, Nov. 2019.

[14] B. E. Rimoldi, "A decomposition approach to CPM," *IEEE Trans. Inf. Theory.*, vol. 34, no. 2, pp. 260–270, Mar. 1988.

[15] L. Hanzo, T. Liew, and B. Yeap, *Turbo Coding, Turbo Equalisation and Space-Time Coding*. New York: Wiley, 2002.

Personal Voice Activity Detection With Ultra-Short Reference Speech

Longting Xu*, Mingjun Zhang*, Wenbin Zhang†, Tianyi Wang†, Jiawei Yin†, and YuGao†‡
* College of Information Science and Technology, Donghua University, China
E-mail: xlt@dhu.edu.cn, zmj@mail.dhu.edu.cn
† AI Research Center, Midea Group (Shanghai) Co.,Ltd., Shanghai 201702, China
E-mail: gaoyu11@midea.com
‡ Corresponding Author

***Abstract*—Personal Voice Activity Detection (PVAD) is widely used in applications such as voice assistants. To accurately detect the voice activity of the target speaker, PVAD typically requires pre-registering the target speaker's speech as a reference. However, the excessively long voice enrollment process tends to reduce user motivation. To address this problem, we explore the possibility that PVAD can maintain good performance even with short reference speech. We propose a PVAD network that supports Ultra-Short reference speech, namely US-PVAD. Unlike traditional methods that rely on pre-trained speaker verification models to extract speaker embeddings, US-PVAD allows the direct input of the original reference speech. Since RNN states can memorize historical information and use it to guide subsequent time steps, we employ a Dual-Path RNN (DPRNN)-based network and use its RNN states as target speaker embedding. This approach eliminates the need for an external speaker embedding extractor with a large number of parameters. Additionally, the RNN states can be continuously updated during voice activity detection, allowing PVAD to obtain sufficient target speaker feature attributes from ultra-short reference speech. Experimental results show that US-PVAD exhibits better performance when using speech under 2 seconds or even as short as 0.2 seconds as the reference speech.**

***Index Terms*—personal VAD, voice activity detection, speaker verification**

I. Introduction

Voice Activity Detection (VAD) [1]–[3] is a technique that classifies audio segments as either speech or non-speech. It is typically a crucial front-end step for various tasks such as speaker verification (SV) [4]–[6], emotion estimation, and automatic speech recognition (ASR). A well-trained VAD system can detect all speech segments regardless of the speaker, which is known as standard VAD. The benefits of using standard VAD are multifaceted. Firstly, downstream systems may be highly sensitive to noise, and filtering out these signals can improve system performance. Secondly, in terms of required computational resources, running SV and ASR systems is often quite expensive, especially when dealing with mobile personal devices. Therefore, triggering downstream systems only when necessary through VAD can help limit resource and energy consumption to some extent. However, in scenarios involving multiple users, standard VAD systems often produce false positives, causing the system to be erroneously activated when not needed. Hence, there is a need for Personal Voice Activity Detection (PVAD) systems.

PVAD [7]–[13] is a technique used to identify the speech activity of a target speaker in multi-speaker scenarios. Compared to standard VAD, PVAD models can distinguish between the speech signals of different speakers, making them more suitable for responding to specific user downstream tasks. While traditional cascaded VAD and SV models can handle PVAD tasks, PVAD systems typically require lightweight designs for widespread real-time deployment. However, SV systems have high resource demands, operational costs, and detection delays, making them unsuitable for such applications.

The PVAD system proposed in [7] represents a significant advancement, offering a lightweight solution for robust target speaker speech detection through the Embedding Conditioned Training (ET) architecture. It eliminates the need for heavy SV systems for frame scoring during runtime, simply combining the target speaker's registered embedding with the acoustic features of the input speech. In subsequent research, the authors extended the PVAD architecture, proposing PVAD 2.0 [8] and evaluating its effectiveness in downstream ASR tasks. To address the issue of missing user-specific enrollment information, [9] proposed a PVAD training method with less enrollment. Ref. [14] explored target speaker detection in diarization contexts.

However, these methods tend to favor longer durations when selecting reference speech. While longer reference speech can be used to obtain better speaker embeddings through pre-trained SV models, accurately reflecting the attributes of the target speaker, in practice, excessively long enrollment times can reduce user motivation and result in a poor user experience. Therefore, it is crucial for PVAD to maintain good performance even under shorter enrollment conditions. Ref. [12] proposed a PVAD network using wakeup words as reference speech and achieved good performance. Ref. [13] investigated the effect of reference speech longer than 1 second on their proposed networks, but neither explored how to maintain good performance in PVAD networks when using ultra-short speech as reference.

Furthermore, as mentioned in [12], the performance of pre-trained SV models does not necessarily correlate with the

This work was supported in part by Donghua University Language and Text Application Research Project (No. Y2024-8).

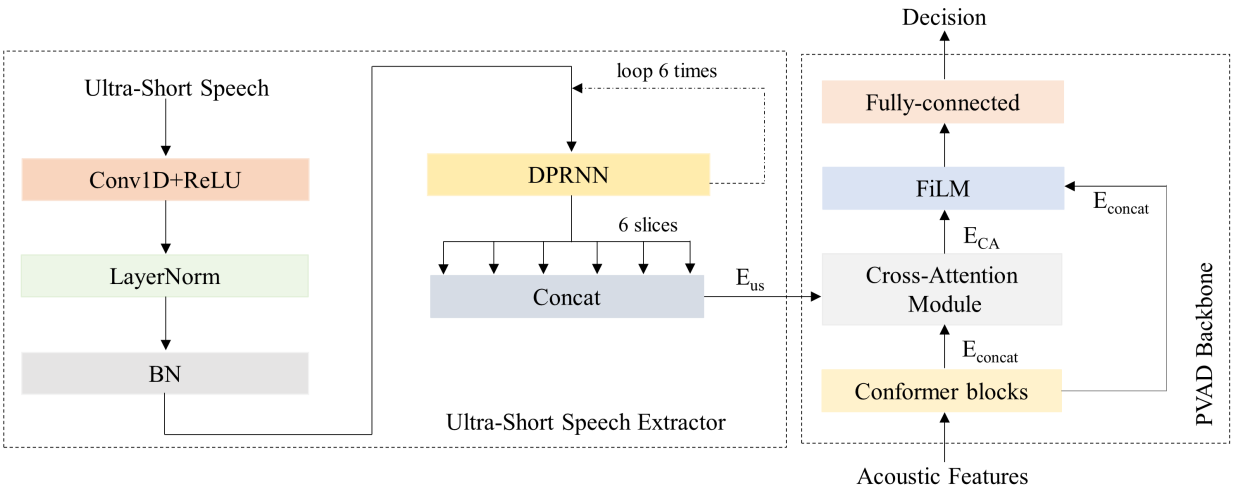


Fig. 1. The Proposed US-PVAD model.

performance of PVAD systems, as they are optimized using different loss functions. Additionally, the quality of embeddings extracted by different SV models for the same reference speech can directly impact PVAD system performance [15]. Based on the above, we propose a PVAD system capable of handling ultra-short reference speech without requiring external SV models to extract target speaker embeddings, namely US-PVAD. Since ultra-short speech contains too little speaker information, directly using its acoustic features or speaker embeddings for training may not be optimal. Inspired by [16], we employ a Dual-Path RNN (DPRNN)-based [17] network to process the raw reference speech, extracting the target speaker's attributes from the continuously updated RNN states during training. This approach avoids dependency on SV models, eliminates the impact of varying embedding quality from different SV models on PVAD performance, and ensures good performance even with very short reference speech.

II. Proposed Methods

The proposed US-PVAD takes acoustic features and the original reference speech as inputs, and provides frame-wise PVAD results as outputs. In this section, we will provide a detailed introduction to it.

A. Ultra-Short Speech Extractor

As shown in Fig. 1, US-PVAD mainly consists of ultra-short speech extractor and PVAD Backbone, which are jointly trained. Inspired by [16], we extract feature information in ultra-short speech using a network based on DPRNN, a well-known network for speech separation [17] and target speaker extraction [18]. It splits the input speech into small chunks for modeling long sequences. Its structure, shown in Fig. 2, consists of two main BiGRU layers, the local BiGRU for extracting intra-chunk features and the global BiGRU for extracting inter-chunk features. Since RNN states can memorize historical information and use it to guide the operation of the next time step, and also update it continuously while the network is running. Using RNN states as speaker features means that the feature information of the target speaker can be continuously updated in the VAD step after enrollment. Therefore, only a reference speech with a short duration is needed as a starting condition in the enrollment step, and the final RNN state can preserve its feature information and guide the network in detecting the voice activity of the target speaker.

The initial states of both local and global BiGRUs are set to zero. Since the layer output after the FC layer and Layer norm is more robust and also contains RNN states, we only use the slices of the layer output that contain RNN states, and both global BiGRU and local BiGRU states are discarded. In addition, to obtain more feature attributes from the ultra-short speech, we iteratively update the input 6 times via DPRNN, take the slices of the layer output obtained from each new iteration, and concatenate them to form the final target speaker embedding. The input and output are as follows:

$$Seq_{out}^{K\times P\times R} = DPRNN_i(Seq_{in}^{K\times P\times R}) \tag{1}$$

where Seq_{in} and Seq_{out} are the inputs and outputs of the lth $(i = 0...5)$ DPRNN iteration. The input reference speech is divided into R chunks, each of length P. Since this is a BiGRU layer with two directions, the output in each direction is K dimensional. Each iteration takes a slice of the layer output in the second dimension to get $Seq_{out}^{K\times R}$, and finally the 6 slices are stitched together to get $Seq_{out}^{K\times 6R}$, which serves as the speaker embedding of the ultra-short reference speech.

B. PVAD Backbone

The PVAD backbone first processes the acoustic features of the concatenated speech through a four-layer stacked Conformer block to generate E_{concat}. Next, we input the speaker feature E_{us}, obtained from the ultra-short speech extractor, as the key and value, and E_{concat} as the query, into the cross-attention (CA) module [19]. The output of the CA module is then used to perform a feature-wise affine transformation

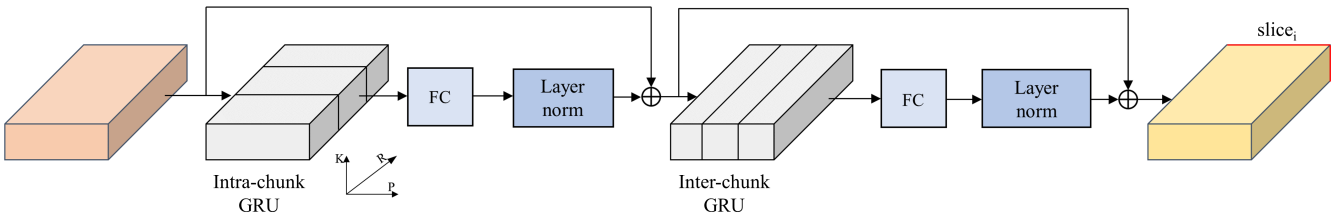


Fig. 2. The DPRNN block structure in US-PVAD. The red line in the output represent the extracted slice, with the slice being concatenated in the P direction.

(FiLM) [20] on E_{concat} to fuse these two sets of features. FiLM includes scaling and shifting operations, where E_{CA} is the output of CA, $\gamma()$ and $\beta()$ are the scaling and shifting vectors of FiLM, respectively. The FiLM process can be represented as:

$$FiLM(E_{concat}, E_{CA}) = \gamma(E_{CA}) \cdot E_{concat} + \beta(E_{CA}) \quad (2)$$

Finally, the fused features are fed into a fully connected layer to generate the decision p for the PVAD model. The entire process can be formulated as follows:

$$p = FC(FiLM(Con(F_{concat}), CA(Con(F_{concat}), E_{us})) \quad (3)$$

where Con denotes the Conformer blocks and F_{concat} is the acoustic feature of the concatenated speech.

III. Experimental Setup

A. Datasets

Referring to the setup described in [7] for simulated conversational speech, we created a multi-speaker training set using three main subsets of LibriSpeech [21]: the 100-hour, the 360-hour, and the 500-hour subsets totaling 960 hours and encompassing 2,338 different speakers. This resulted in 312,020 concatenated utterances. Each audio file contains three speakers, with one randomly selected as the target speaker in the concatenated utterance. We used the Montreal Forced Aligner [22] to create VAD ground truth annotations for the concatenated utterances. The PVAD labels for each frame were modified according to the target speaker: "non-speech" frames remained unchanged, while "speech" frames were relabeled as either "target speaker speech" or "non-target speaker speech" depending on whether the speech originated from the target speaker. Additionally, we generated 5,350 concatenated utterances as a test set using LibriSpeech's test-clean and test-other subsets, created in the same manner as the training set. We added noise and reverberation to the created training and test sets using the Musan [23] and RIRs [24] datasets to simulate real-world environments.

B. Implementation details

In the DPRNN block of the ultra-short speech extractor, the chunk size is set to 250 and the hop size is 125. The 1-D convolutional layer has 1 input channel and 256 output channels, with a kernel size of 2 and a stride of 1. The input and output sizes of the batch normalization (BN) are both 256. For the PVAD backbone, the LSTM-based model consists of 2 layers of LSTMs, each layer having 256 units, and the last layer is a fully connected layer that acts as a classifier. The Conformer-based model contains 4 Conformer layers in the block, each having a dimension of 256, an attention head count of 8, causal 7 × 7 convolution kernel, and 31 left context.

We use the pre-trained ResNet34 model provided by SpeechBrain [25] as the SV model in all baseline systems, which has about 15.5 million parameters, and the extracted speaker embedding vectors have a size of 256. In all models, we input acoustic features with the same settings, i.e., 40-dim log Mel-filterbank energy with a frame length of 25 ms and a frame shift of 10 ms. The loss function uses Weighted Pairwise Loss [7], where the weights are set to $w\langle \text{tss}, \text{ns} \rangle = w\langle \text{tss}, \text{ntss} \rangle = 1$ and $w\langle \text{ns}, \text{ntss} \rangle = 0.5$. During the training process, we use the Adam optimizer [26] with an initial learning rate of 5e-4. We use Recall, Precision, F1-score, and average precision (AP) for each class as metrics.

IV. Results

To illustrate that US-PVAD can better handle ultra-short reference speech, we compare our proposed US-PVAD model with several baseline models on reference speech of different lengths in Table 1. **LSTM+Concat** denotes the ET structure proposed in [7], while **Conformer+Concat** and **Conformer+FiLM** are the structures proposed in [8]. The configuration of all models follows the previous description.

A. Comparison of results

From the results, it can be seen that US-PVAD outperforms the baseline systems in all metrics for a reference speech duration of 0.2 seconds. In particular, in the extreme case of additional noise, the REC of the baseline systems are all around 50% (E13-E15), whereas US-PVAD still achieves a level of 60% (E16: 60.91%), even higher than the baseline system with a 2-second speech reference (E1: 57.01%). In terms of overall performance (F1 and AP), US-PVAD not only outperforms the baseline systems but also maintains high detection efficacy for NON-SPEECH (NS) even after adding noise (E16: 96.29%).

Moreover, US-PVAD demonstrates great detection efficacy for TARGET SPEAKER SPEECH (TSS), with a 0.2-second

TABLE I

RESULTS COMPARISON WITH SHORT REFERENCE SPEECH (%). REF DENOTES REFERENCE SPEECH OF VARYING DURATIONS. 3 SPEAKERS INDICATE CONCATENATED UTTERANCES FROM THREE DIFFERENT SPEAKERS. NOISE DENOTES ADDING NOISE FROM MUSAN AND RIRS. REC: RECALL. PRE: PRECISION. F1: F1 SCORE. AP REPRESENTS THE AVERAGE PRECISION ACROSS THREE CLASSES: NON-SPEECH (NS), NON-TARGET SPEAKER SPEECH (NTSS), AND TARGET SPEAKER SPEECH (TSS).

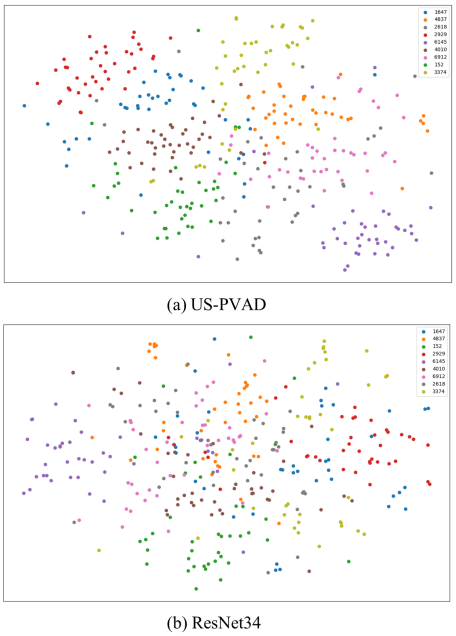
Ref	Exp	Model	Params(M) (PVAD/SV)	3 Speakers				3 Speakers + Noise			
				REC	PRE	F1	AP [NS NTSS TSS]	REC	PRE	F1	AP [NS NTSS TSS]
2s	E1	LSTM+Concat	1.09/15.53	60.45	72.89	78.04	[89.08 88.98 72.62]	57.01	71.27	75.29	[85.59 87.95 69.75]
	E2	Conformer+concat	1.16/15.53	**77.05**	78.65	83.48	[89.55 94.49 86.51]	74.22	78.81	83.08	[92.51 94.02 85.56]
	E3	Conformer+FiLM	1.16/15.53	74.94	81.99	84.26	[90.47 95.33 87.78]	**74.61**	79.48	84.28	[94.34 94.93 86.50]
	E4	US-PVAD	1.93/0.00	76.72	**83.19**	**86.98**	**[97.17 96.19 90.64]**	74.33	**82.64**	**85.75**	**[95.54 96.03 88.81]**
1s	E5	LSTM+Concat	1.09/15.53	52.64	72.07	76.22	[89.80 87.48 68.56]	45.70	70.41	72.55	[86.72 85.53 64.43]
	E6	Conformer+concat	1.16/15.53	73.90	76.01	83.65	[94.83 93.33 83.06]	68.49	73.11	80.58	[93.16 91.45 79.21]
	E7	Conformer+FiLM	1.16/15.53	73.99	80.58	85.38	[95.81 95.37 85.11]	65.95	78.45	81.60	[93.49 93.45 82.49]
	E8	US-PVAD	1.93/0.00	**74.41**	**81.48**	**85.48**	**[96.86 95.39 88.39]**	**70.45**	**78.65**	**83.52**	**[95.54 94.39 84.06]**
0.5s	E9	LSTM+Concat	1.09/15.53	48.01	66.62	73.48	[88.07 84.15 62.73]	44.31	67.52	72.31	[87.04 83.45 61.32]
	E10	Conformer+concat	1.16/15.53	63.29	67.40	77.22	[88.94 88.47 71.12]	59.87	64.14	75.29	[91.13 86.73 67.68]
	E11	Conformer+FiLM	1.16/15.53	67.14	77.02	81.14	[90.43 92.87 81.39]	63.95	75.31	79.42	[91.38 92.23 79.47]
	E12	US-PVAD	1.93/0.00	**73.01**	**79.74**	**84.09**	**[96.75 93.67 85.26]**	**70.27**	**77.59**	**81.88**	**[95.35 92.99 83.23]**
0.2s	E13	LSTM+Concat	1.09/15.53	40.98	58.22	69.20	[86.93 81.73 53.93]	40.77	52.46	67.01	[85.36 76.90 48.51]
	E14	Conformer+Concat	1.16/15.53	52.60	59.95	73.40	[95.53 85.62 65.17]	48.98	54.90	69.67	[85.77 80.34 55.69]
	E15	Conformer+FiLM	1.16/15.53	52.22	70.83	78.60	[96.38 89.16 72.20]	50.36	69.17	76.28	[93.72 88.24 70.57]
	E16	US-PVAD	1.93/0.00	**62.74**	**77.42**	**82.31**	**[97.05 92.24 79.30]**	**60.91**	**76.91**	**80.11**	**[96.29 91.97 77.65]**

reference speech (E16: 77.65%) outperforming one baseline using more than 1 second of reference speech (E1: 69.75%, E5: 64.43%) and closely matching two other baselines using 1 second of reference speech (E6: 79.21%, E7: 82.49%).

Furthermore, we evaluated the performance of US-PVAD with 1-second and 0.5-second reference speech. Especially for 0.5 seconds, which is a very short duration and can be used as a reference speech in practical applications, US-PVAD performs significantly better than the baseline systems. The results show that the baseline systems underperform US-PVAD in all metrics, especially in the AP metric of TSS, where US-PVAD (E12) using a 0.5-second reference speech outperforms all the baseline systems (E5-E7) using a 1-second reference speech. This suggests that US-PVAD can support ultra-short reference speech well due to the constant updating of the target speaker embedding during the ultra-short speech extraction step.

Additionally, the impact of changing the enrollment duration from 2 seconds of longer speech to 0.2 seconds of ultra-short speech varies across different models. The performance metrics of US-PVAD exhibit a smaller change compared to the baseline model, particularly in noise-free conditions. Specifically, the PRE of US-PVAD only decreases by 5.77% (E4: 83.19% to E16: 77.42%), whereas the baseline model experiences a maximum drop of 18.70% (E2: 78.65% to E14: 59.95%) and a minimum drop of 11.16% (E3: 81.99% to E15: 70.83%). This indicates that US-PVAD maintains more consistent performance when faced with shorter enrollment duration compared to the baseline models.

It is worth noting that US-PVAD (E4) outperforms the baseline models in terms of PRE and overall performance when using a 2-second reference speech, but its efficacy in terms of REC is lower compared to the two baselines (E2, E3). This may be because the baseline systems rely on an external SV model to extract speaker embeddings, and the embeddings extracted from the 2-second reference speech contain more speaker information, resulting in a slightly higher REC than US-PVAD. Since US-PVAD does not rely on an SV model, and the ultra-short speech extractor and the PVAD backbone network are co-trained, the observed performance difference aligns with our study's expectations. We intend to further investigate this issue in our future research. In addition, compared to the baseline model that relies on the SV model, US-PVAD has significantly fewer total parameters (1.93/0.00),



(a) US-PVAD

(b) ResNet34

Fig. 3. t-SNE results for aggregated embedding vectors of 0.2s speech from 9 speakers. (a) US-PVAD's ultra-short speech extractor; (b) ResNet34.

making it more suitable for lightweight deployments. Overall, US-PVAD is able to better handle ultra-short speech while maintaining good performance for normal-length reference speech.

B. Comparison of t-SNE

The quality of target speaker embeddings extracted from ultra-short speech is critical for the performance of the PVAD backbone. We compared the t-SNE visualizations of embedding vectors extracted by ultra-short speech extractor of US-PVAD and ResNet34. As illustrated in Fig. 3, for 0.2-second speech, the speaker features produced by US-PVAD exhibit clear clustering and distribution according to different speakers, achieving better separation of embedding vectors from different speakers compared to ResNet. This is a key factor contributing to US-PVAD's great performance in handling ultra-short speech.

V. Conclusions

This work explores the PVAD model that maintains good performance even with ultra-short reference speech. Specifically, we propose a PVAD model named US-PVAD, which can handle ultra-short reference speech and operates independently of a pre-trained SV model. US-PVAD utilizes RNN states generated by DPRNN, which can be continuously updated during the PVAD process, as features of the target speaker. Experimental results show that under conditions of short reference speech with varying lengths of less than 2 seconds, our proposed method outperforms other baseline methods and demonstrates greater robustness to variations in enrollment duration.

References

[1] S.-Y. Chang, B. Li, G. Simko, *et al.*, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5549–5553. DOI: 10.1109/ICASSP.2018.8461921.

[2] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4466–4469. DOI: 10.1109/ICASSP.2010.5495611.

[3] J. Ramirez, J. M. Górriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," *Robust speech recognition and understanding*, vol. 6, no. 9, pp. 1–22, 2007.

[4] L. Xu, K.-A. Lee, H. Li, and Z. Yang, "Co-whitening of i-vectors for short and long duration speaker verification.," in *INTERSPEECH*, 2018, pp. 1066–1070.

[5] L. Xu, R. K. Das, E. Yılmaz, J. Yang, and H. Li, "Generative x-vectors for text-independent speaker verification," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 1014–1020.

[6] L. Xu, K. A. Lee, H. Li, and Z. Yang, "Generalizing i-vector estimation for rapid speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 749–759, 2018.

[7] S. Ding, Q. Wang, S.-y. Chang, L. Wan, and I. L. Moreno, "Personal vad: Speaker-conditioned voice activity detection," *arXiv preprint arXiv:1908.04284*, 2019.

[8] S. Ding, R. Rikhye, Q. Liang, *et al.*, "Personal vad 2.0: Optimizing personal voice activity detection for on-device speech recognition," *arXiv preprint arXiv:2204.03793*, 2022.

[9] N. Makishima, M. Ihori, T. Tanaka, A. Takashima, S. Orihashi, and R. Masumura, *Enrollment-less training for personalized voice activity detection*, 2021. arXiv: 2106.12132 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2106.12132.

[10] W. Wang, X. Qin, and M. Li, "Cross-channel attention-based target speaker voice activity detection: Experimental results for the m2met challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9171–9175.

[11] Z. Kang, J. Wang, J. Peng, and J. Xiao, "Svvad: Personal voice activity detection for speaker verification," *arXiv preprint arXiv:2305.19581*, 2023.

[12] B. Zeng, M. Cheng, Y. Tian, H. Liu, and M. Li, "Efficient personal voice activity detection with wake word reference speech," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 241–12 245.

[13] F. Liu, F. Xiong, Y. Hao, K. Zhou, C. Zhang, and J. Feng, "As-pvad: A frame-wise personalized voice activity detection network with attentive score loss," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 476–11 480. DOI: 10.1109/ICASSP48485.2024.10446581.

[14] M. Cheng, W. Wang, Y. Zhang, X. Qin, and M. Li, "Target-speaker voice activity detection via sequence-to-sequence prediction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[15] A. Aloradi, M. Elminshawi, S. R. Chetupalli, and E. A. Habets, "Target-speaker voice activity detection in multi-talker scenarios: An empirical study," in *Speech Communication; 15th ITG Conference*, VDE, 2023, pp. 250–254.

[16] L. Yang, W. Liu, L. Tan, J. Yang, and H.-G. Moon, "Target speaker extraction with ultra-short reference speech by ve-ve framework," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[17] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-

channel speech separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50. DOI: 10.1109/ICASSP40776.2020.9054266.

[18] Y. Hao, J. Xu, J. Shi, P. Zhang, L. Qin, and B. Xu, "A unified framework for low-latency speaker extraction in cocktail party environments," in *Interspeech*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:226202888.

[19] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[22] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:12418404.

[23] D. Snyder, G. Chen, and D. Povey, *Musan: A music, speech, and noise corpus*, 2015. arXiv: 1510.08484 [cs.SD]. [Online]. Available: https://arxiv.org/abs/1510.08484.

[24] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224. DOI: 10.1109/ICASSP.2017.7953152.

[25] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, *Speechbrain: A general-purpose speech toolkit*, 2021. arXiv: 2106.04624 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2106.04624.

[26] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1412.6980.

Data Augmentation Methods and Influence of Speech Recognition Performance for TED Talk's English to Japanese Speech Translation

Kento Masuda* , Kazumasa Yamamoto† and Seiichi Nakagawa ‡
Department of Computer Science, Chubu University, Japan
E-mail: {*tp22021-6334@sti, †kazumasayamamoto@isc, ‡nakagawa@isc}.chubu.ac.jp

***Abstract*—In this paper, we compare uni-directional and bi-directional translation models for English to Japanese and various data augmentation methods to improve the performance of translation models using the Transformer. In the augmentation methods, we report on the performance improvements in text translation using pseudo parallel corpus based on the Japanese translation of English monolingual corpus, pseudo parallel corpus based on the English translation of Japanese monolingual corpus, pseudo parallel corpus based on translations of one side of the English-Japanese parallel corpus, addition of auto-encoder part to the translation model, and use of sentences generated by the auto-encoder as pseudo corpus, and training multilingual translation models based on English-Korean parallel corpus. Next, a speech translation model was constructed by cascade connecting machine translation model and speech recognition model, and evaluated with various models of Kaldi and Whisper speech recognizers to analyze the relationship between speech recognition rate and speech translation performance.**

I. Introduction

For TED talks, English speech recognition has been frequently presented at the IWSLT (International Workshop on Spoken Language Translation), but there are fewer on English to Japanese translations. In 2021, simultaneous text translation for English to Japanese [2] and in 2022, simultaneous speech translation for English to Japanese was included as a task [3].

For TED speech recognition, Ueno et al. achieved a word error rate (WER) of 8.56% using a 12 layers Conformer encoder and a uni-directional LSTM decoder with Attention function, trained on TEDLIUM2 (211 hours of training data)[4]. Currently, it seems effective to incorporate self-supervised learning (SSL) feature vectors into Conformer. In this paper, we use Kaldi, a typical recognition tool based on the DNN-HMM method, as a speech recognizer. Additionally, we utilize Whisper, which has been widely used recently. Whisper, though not using SSL technology, is a large-scale model trained on over 600,000 hours of data, achieving extremely high recognition rates. Whisper offers models like tiny, small, medium, and large; in this paper, we use medium and large models to analyze the relationship between various speech recognition rates and translation performance.

For speech translation, there are two types of methods: a hybrid (cascade) method that connects speech recognition and text translation and an end-to-end method. Recently, the latter has made remarkable progress, closing the gap between the two [3,5]. For TED English to Japanese speech translation, Fukuda et al. reported BLEU scores of 11.6 for offline method and 10.6 for online simultaneous translation method [6]. The best result at IWSLT 2022 was the USTC (University of Science and Technology of China) system using large external speech and language resources, with a WER of about 5%, a text input BLEU score of 22, and an offline speech input BLEU score of about 19 using an emsemble method [11]. At IWSLT 2023, offline speech translation, simultaneous speech translation, and new TED test data were evaluated, and the ACL presentation speech was also included in the test [7]. Once again, the cascade method of speech recognition and machine translation outperformed the end-to-end method. Three teams participated in the offline English to Japanese speech translation task for TED talks, with BLEU scores of 10.6, 16.5, and 18.7, respectively. The best method from HW-TSC (Huawei Translation Service Center) involved fine-tuning Whisper (conformer) for speech recognition and using a one-to-many (1-to-3) Deep-Transformer model for translation [7].

To compensate for the lack of parallel data, there is a data augmentation method using monolingual corpus [8]. This method creates a pseudo parallel corpus by machine translating monolingual corpus and mixes them with the base parallel corpus for training. Yamagishi et al. used transcription data of academic conference lectures from the Corpus of Spontaneous Japanese (CSJ) for monolingual corpus in this method [9]. Another method converts the base parallel corpus into pseudo corpora based on certain rules, increasing the training data several times (parallel phrase model) [10]. Bao et al. generated multiple translations by adding keywords from source sentences to obtain high-quality pseudo corpus [11]. There are also various methods using auto-encoders. Cheng et al. used monolingual data and English to Chinese & Chinese to English translation models to generate target English sentences from source English sentences, that was, English sentences into Chinese sentences and then the Chinese sentences into English sentences, learning the translation models to make the source and target sentences identical, achieving higher effectiveness than standard back translation approach [12].

The method we attempted is to learn so that the source side sentence and target side sentence are identical, which is

a similar method as Currey et al [13]. However, Currey et al. did not evaluate the method with the Transformer or the bi-directional model. Nor did they compare their method with a pseudo parallel corpus using sentences generated by an auto-encoder.

TABLE I: Data sets

(a) Training dataset for machine translation model

Data set	Training # sentence pairs	Validation # sentence pairs
IWSLT2016 English-Japanese parallel corpus	233,108 pairs	871 pairs
	1,863 lectures	8 lectures

(b) Data set for data augmentation

Data set	# sentences (#lectures)
IWSLT2018 (new-English monolingual)	80,222 sent. (862 lec)
CSJ (new-Japanese monolingural)	219,229 sent. (1,565 lec)
IWSLT2016 (English-Korean pairs)	230,240 pairs (1,920 lec)
ASPEC (written English-Japanese pairs)	1,000,000 pairs

(c) Test data

Data set	Test Sentences (English)
IWSLT2016 English-Japanese	963 sentences
	10 lectures

II. Speech Recognition Models

A. Kaldi

Kaldi provides recipes for various corpora, and in this paper, we used the TEDLIUMv3 corresponding recipe for speech recognition. Kaldi employs a DNN-HMM architecture. In this paper, the DNN has 13 hidden layers, each with 1024 units. The number of units in the output layer corresponds to the shared tied state number of triphone HMMs. The input features to the network are 40-dimensional MFCCs (Mel-frequency cepstral coefficients). For feature preprocessing, we applied fMLLR, LDA, and SAT to extract speaker-independent features. We trained the DNN-HMM on 450 hours of TED talk data. The language model was created from the text of 450 hours of TED talks with a vocabulary size of about 150,000 words.

B. Whisper

Whisper, provided by OpenAI, is an open-source software for speech recognition. It is designed as an end-to-end model, meaning it directly converts input audio into text without needing intermediate representations. This architecture simplifies the processing pipeline and can potentially lead to higher accuracy and efficiency. It uses about 680,000 hours of multilingual audio data scraped from the web for training, enabling high-accuracy speech recognition. Using large and diverse training data is expected to improve robustness against background noise and specialized terminology. The basic processing involves converting input audio into acoustic features and segmenting the data. Specifically, input audio is transformed into a log-Mel spectrogram. The hyperparameters for this feature extraction are hardcoded in sorce code, resulting in 80-dimensional features, with a 10ms stride, resampled to 16kHz mono audio. Whisper processes audio in 30-second segments, padding with zeros if necessary, to ensure consistent segment length. Whisper offers models of various sizes, with transcription accuracy improving as the number of parameters increases. In this experiment, we used the medium model with 769M parameters and the large-v3 model with 1550M parameters. We used these pre-trained models without fine-tuning on TED data.

III. Translation Models

A. Transformer

The Transformer [14] model consists of encoder and decoder layers. The encoder is composed of stacked identical encoders, each consisting of a self-attention mechanism and a feed-forward neural network (FFNN). The decoder is similarly composed of stacked identical decoders. In this experiment, we used fairseq ver.0.12.0 [15].

B. Number of Encoder-Decoder Layers

A standard Transformer model has 6 encoder layers and 6 decoder layers, but the optimal configuration varies depending on the amount of training data. Preliminary experiments in previous research [1] showed a BLEU score of 0.80 with 6 encoder and 6 decoder layers and a score of 13.22 with 3 encoder and 3 decoder layers. In this experiment, we compared various data augmentation methods with different encoder-decoder layer numbers.

C. Uni-directional and Bi-directional Translation Models

In an uni-directional translation model, the translation model is created by training on pairs of source (English) and target (Japanese) languages. In a bi-directional translation model, the model is created by training on datasets consisting of pairs of source and target languages and pairs where the source and target languages are swapped. This model performs both English to Japanese and Japanese to English translations for improving the robustness.

IV. Data Augmentation Methods

A. Data Augmentation Using Forward/Backward Translation of Monolingual Corpus

(a) Utilization from other than base corpus

Due to the limited amount of English and Japanese parallel corpora in the IWSLT base corpus of TED lectures used in this experiment, we created pseudo-parallel corpora by translating English or Japanese monolingual corpora forward (English to Japanese) / backward (Japanese to English) using the base model, thereby performing data augmentation [1]. As the monolingual corpora, we used the monolingual corpora used the English side of the IWSLT2018 English-Spanish parallel corpus and the simulated lectures from the CSJ Japanese corpus.

(b) Utilization from base corpus

Translating the English or Japanese side of the base corpus forward/backward using the model with the highest translation accuracy from previous research [1], the translated sentence with the other original sentence of the base parallel corpus is combined to create pseudo parallel corpora.

B. Data Augmentation Using Auto-Encoders

English to English and Japanese to Japanese parallel corpora for training (autoencoder [14]) are also included. During this process, language pair tags were added to the source language side to discriminate language pairs (such as EJ, JE, EE, JJ) as shown in Figure 1. Figure 1 shows the schematic illustration of the traing data setup using original parallel data and auto-encoder method.

(c) Parallel corpus with identical source and target sentences
For the base corpus, we added an autoencoder component with identical source and target sentences for training. In the case of an English to Japanese uni-directional translation model, we simultaneously trained the model to translate English into Japanese and added parts where English sentences were translated into the same English sentences or Japanese sentences into the same Japanese sentences. For the English to Japanese/Japanese to English bi-directional translation model, in addition to bi-directional translation between English and Japanese, we trained the model with the autoencoder functions of English to English and Japanese to Japanese.

(d) Using pseudo source sentences generated by auto-encoder
A pseudo-parallel corpus was created using an autoencoder trained with the same source and target sentences of the base parallel corpus and used for training as a data augmentation, that is, generated English-Japanese or English-generated Japanese in addition to English-Japanese.

(e) Addition of auto-encoder part from monolingual corpus other than the base corpus
As similar to the method (c), but using a monolingual corpus other than the base corpus (b), an auto-encoder component was added to the translation model.

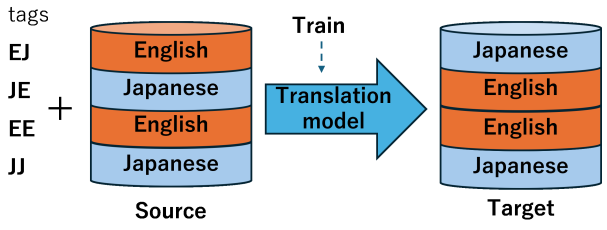


Fig. 1: Schematic illustration of the training data setup using the auto-encoder method

C. Data Augmentation Using English-Korean Parallel Data

(f) Considering that multilingual models are effective for low-resource learning data [14], we applied this approach to English to Japanese and Japanese to English translation tasks by training with a parallel corpus of English and Korean, which has a similar grammar to Japanese, in addition to the base parallel corpus.

D. Transfer Learning from ASPEC Corpus Model

(g) Although the ASPEC (Asian Scientific Paper Excerpt Corpus) corpus consists of written language rather than spoken language like TED, it contains a large-scale parallel corpus of 1 million sentence pairs. We trained uni-directional and bi-directional models using the ASPEC corpus, and used these parameters as initial values to train the translation model with the IWSLT base corpus and the data-augmented dataset. Note that the vocabulary of TED's spoken language was set as the vocabulary during the translation model training using the ASPEC corpus [1].

V. Experimental results

A. Data used

The IWSLT2016 English-Japanese parallel corpus was used as the base corpus for this study. The English side of the IWSLT2018 English-Spanish parallel corpus and the Japanese spoken corpus CSJ, a monolingual corpus of Japanese, were used as a monolingual corpus for data augmentation, and the IWSLT2016 English-Korean parallel corpus was used for the multilingual model. Table I shows the number of training, development, and test sentences.

B. Speech Recognition Results

The TED lectures vary in difficulty depending on their content, affecting translation performance. Additionally, the speakers differ for each lecture, resulting in significant variations in speech recognition performance by speaker. Whisper performed speech recognition for each lecture as a whole, rather than for each utterance. Whisper sometimes recognizes meaningless word sequences during silent intervals and may output repeated word sequences. Table IVshows the word error rate (WER=100 - error rates of insertion, deletion, and substitution) for each lecture and speaker using Kaldi and Whisper. In Kaldi, the WER ranged from about 10% to 33%, with an average of 15.37%. With Whisper's Medium model, the WER ranged from 3.4% to 13.34%, with an average of 8.85%. The Large model had a WER ranging from 3.14% to 11.76%, with an average of 7.21%.

C. Text Translation Results

The number of encoder and decoder layers was set to 3 or 6 layers, and both un-idirectional and bi-directional models were trained and evaluated. Optimization was performed using Adam with a learning rate of 0.0005, and training was conducted for up to 20 epochs. BLEU (BiLingual Evaluation Understudy) was used to calculate the similarity between

TABLE II: Results of the 10 TED talks English-to-Japanese translation evaluation experiment (BLEU)
The parenthesis () denotes result by fine-tuning model
"not well done" means thath the translation model was not well learned

model		Uni-direction		Bi-direction	
method \ #layers		3 layers	6 layers	3 layers	6 layers
Base line		**13.78**	10.67	13.53 (**14.40**)	not well done
b-1	+ Pseudo English	15.72	16.04	15.88 (16.11)	
b-2	+ Pseudo Japanese	**16.52**	16.19	15.63 (**16.23**)	
b-3	+ Pseudo English & Japanese	16.01	16.36	15.30 (15.79)	
c-1	+ English-English	13.73	15.47	13.70 (13.89)	15.81 (16.21)
c-2	+ Japanese-Japanese	13.48	15.19	13.80 (12.78)	16.14 (16.37)
c-3	+ English-English & Japanese-Japanese	13.66	15.22	13.26 (13.63)	15.67 (16.17)
d-1	+ Generated English	13.69	12.13	14.39 (14.70)	not well done
d-2	+ Generated Japanese	12.89	11.97	14.12 (14.39)	
d-3	+ Generated English + Generated Japanese	13.54	12.12	14.56 (14.79)	
e-1	+ new English-new English	13.50	14.81	13.65 (14.09)	14.76 (15.13)
e-2	+ new Japanese-new Japanese	12.87	14.65	13.44 (13.71)	14.14 (14.56)
e-3	+ New English-New English & New Japanese-New Japanese	13.98	15.03	13.88 (14.36)	15.59 (16.18)
f	+ English-Korean			14.66 (15.15)	not well done
h-1	+New English (a) + New Japanese (a) + ASPEC (g)			14.76 (15.48)	16.16 (**16.50**)
h-2	(h-1)+(b-3)+(c-3)			16.13 (**16.77**)	not well done

TABLE III: Multi-reference translation evaluation results for 2 TED talks (BLEU). ***() denotes fine-tuing model***

model		Uni-direction		Bi-direction	
method	#references \ #layers	3 layers	6 layers	3 layers	6 layers
Base line	single	**11.16**	8.75	11.69 (**12.18**)	not well done
	3 references	**24.47**	18.35	**25.94** (24.94)	
b-3	single	**12.41**	11.19	10.54 (12.90)	
	3 references	**27.34**	26.97	25.55 (27.32)	
c-3	single	9.96	10.94	11.37 (11.08)	12.19 (12.30)
	3 references	23.24	26.45	24.88 (24.59)	28.08 (26.62)
d-3	single	10.38	10.02	11.47 (11.71)	not well done
	3 references	23.20	22.85	24.55 (24.99)	
e-3	single	9.93	10.31	10.91 (11.16)	11.61 (11.86)
	3 references	22.74	25.98	24.29 (23.30)	27.71 (27.44)
f	single			12.06 (13.39)	not well done
	3 references			26.88 (27.04)	
h-2	single			12.44 (**13.09**)	
	3 references			28.15 (**30.94**)	

machine translation and reference, with values up to 4-gram used in this study. The English-Japanese text translation results for 10 lectures are shown in Table Ⅱ. Additionally, human translations by two translation agencies were obtained for two lectures, resulting in three references including the original reference, as shown in Table Ⅲ.

Fine-tuning of the bi-directional translation model involves additional re-training of the model with the base parallel corpus after model training using data augmentation. The baseline refers to the model trained with the IWSLT2016 English-Japanese base parallel corpus using both uni-directional and bi-directional translation models. "+Pseudo-English" refers to the pairs of translated English sentences and original Japanese sentences of the base parallel corpus using the method of Section Ⅳ(b) as a pseudo-parallel corpus for training. The same applies to "+Pseudo-Japanese." The English part of IWSLT2018 used as a monolingual corpus is referred to as "new-English," and the CSJ corpus as "new-Japanese."

The baseline results were BLEU values of 13.78 for the uni-directional model and 14.40 for the bi-directional model. The method of data augmentation of pseudo-corpora by machine translation of the base corpus (b) exceeded the baseline BLEU values for both pseudo-English and pseudo-Japanese. The uni-directional translation model improved the BLEU value by +2.74 (to 16.52). The bi-directional translation model also showed further improvement in translation accuracy with fine-tuning, with a BLEU value improvement of +1.83(to 16.23) compared to the bi-directional baseline. The augmentation method of adding an autoencoder part with identical source and target sentences to the base parallel corpus showed high translation accuracy for all models fine-tuned with 6 layers, achieving BLEU values of 16.00 or higher in text translation. This suggests that the translation model obtained robust internal representations through the added autoencoder part.

The method of using sentences generated by the autoencoder as pseudo-parallel corpora (d) did not show significant differences from the baseline. This is because the performance of the autoencoder was too good, resulting in high similarity between the generated pseudo-parallel corpus and the base parallel corpus (generated English: BLEU = 99.47, generated Japanese: BLEU = 98.84), and thus, generated sentences including errors was not mostly achieved.

The method of adding monolingual corpora other than the base corpus as the autoencoder part (e) showed higher improvement for the model (e-3) that added both the English part of IWSLT2018 and the CSJ corpus. However, the degree of translation accuracy improvement was lower compared to the model (c) that added the autoencoder part to the base

TABLE IV: Results of WER and BLEU by speech translation/text translation using Kaldi and Whisper Medium/Large models.
The parenthesis () denotes BLEU by baseline-model ; BLEU by best model (BLEU by baseline model)

Lecture	Kaldi WER	Medium WER	Large-v3 WER	Kaldi BLEU	Medium BLEU		Large-v3 BLEU		Text BLEU
1	9.53	10.87	9.58	13.37	13.57	(11.13)	13.89	(12.41)	16.82 (12.64)
2	11.00	11.19	5.10	11.58	9.87	(8.38)	10.09	(9.67)	12.55 (12.01)
3	11.30	9.78	8.68	10.46	11.54	(9.87)	11.98	(11.16)	12.57 (11.59)
4	11.49	3.40	3.14	14.14	15.02	(14.12)	15.44	(15.33)	15.54 (14.63)
5	11.71	4.27	4.18	16.77	18.54	(17.19)	18.65	(18.21)	21.37 (18.84)
6	12.93	10.43	6.70	8.69	8.75	(9.02)	9.22	(10.27)	11.87 (11.61)
7	14.79	8.41	7.61	11.24	10.90	(7.92)	11.20	(9.54)	11.76 (11.10)
8	16.38	13.34	11.76	17.79	15.03	(13.52)	15.87	(15.03)	21.21 (17.54)
9	21.42	9.79	9.67	11.12	14.01	(12.28)	14.07	(13.71)	16.49 (15.37)
10	33.17	6.99	5.67	11.27	15.30	(11.25)	15.75	(12.44)	15.71 (12.43)
Ave.	15.37	8.85	7.21	12.64	13.25	(11.47)	13.61	(12.77)	15.58 (13.81)

parallel corpus.

The use of the English-Korean parallel corpus (f) improved English to Japanese translation accuracy with a multilingual model incorporating the parallel corpus of English-Korean, a language with similar grammar to Japanese.

Finally, as various combinations, Table Ⅳ(h-2) shows the translation evaluation results of models that performed data augmentation using the method of adding pseudo-parallel corpora from monolingual corpora (a) and combining initial parameters from the ASPEC corpus model (g), with the method of using pseudo-corpora from the base corpus (b) and adding the autoencoder part to the base corpus (c). By combining the proposed method, which showed a high effect of translation accuracy improvement, with the best model from previous research [1], a high evaluation result of BLEU 16.77 was obtained in text translation.

The original reference is close to free translation of spoken language, limiting the BLEU improvement by machine translation. Therefore, when the reference is made multi (three) references, the BLEU value improves significantly (Table Ⅲ).

D. Speech Translation Results

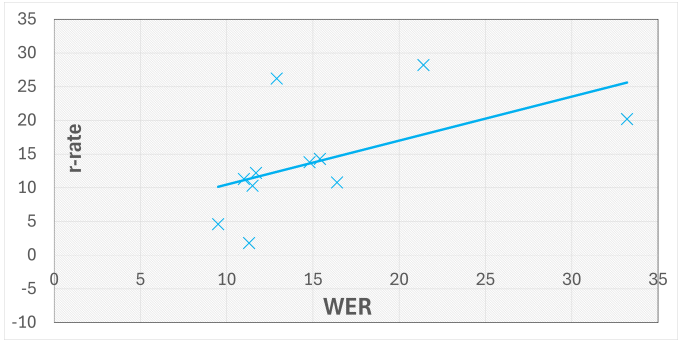
The speech translation results are shown in Table Ⅳ. The translation models are the baseline model and the best model (h-2). The speech recognizers are Kaldi and Whisper's medium model and large model. We should notice that the BLEU scores for Table Ⅳare slightly different from those of Table Ⅱ, because the BLEU in Table Ⅲis calculated for all lectures as a whole, but the BLEU in Table Ⅳis calculated for each lecture and then averaged. Compared to the BLEU values of text translation, the BLEU values of speech translation with speech input decreased by 1 to 5 points. Figure 2 shows the relationship between speech recognition performance (word error rate; WER) and the reduction rate in translation performance. The r-rate indicates the reduction rate of the BLEU value of speech translation results from the BLEU value of text translation results and is calculated by $(\frac{\text{Text BLEU}-\text{ASR BLEU}}{\text{Text BLEU}}) \times 100$. For Kaldi's speech recognition results, only the best translation model was evaluated. The correlation between WER and the reduction rate of BLEU in the best model was 0.545 for Kaldi model, 0.47 for Whisper's Large-v3 model, 0.75 for the Medium model, and 0.63 for Whisper overall, respectively. In the baseline model, the correlation between WER and the reduction rate of BLEU was 0.4 for the Large-v3 model, 0.73 for the Medium model, and 0.63 for Whisper overall, respectively. As seen in Figure 2 (b) and (c), both the baseline and best translation models show that the overall linear approximation almost coincides with recognition results of Whisper. This means that regardless of the translation model, the relationship between speech recognition accuracy (if the recognizer is the same) and the reduction in translation performance shows the same trend. For example, if the WER is about 7%, the reduction in translation performance due to speech input compared to text input is about 10% in BLEU value. However, for Kaldi's speech recognition results, if the WER is 10%, the reduction rate in translation performance is about 10%. This different trend may be caused by different ASR architectures which yield proper misrecgonition results.

VI. Conclusions

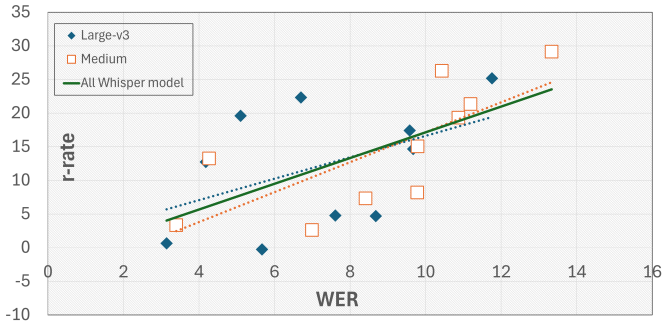
For the English-Japanese TED talk, using Transformer based machine translation models trained by the corpus consisting of 220,000 sentence pairs, translation performance was improved for both unidirectional and bidirectional models by incorporating auto-encoder components and augmenting the base corpus with machine-translated pseudo parallel corpora. For the 10 lectures, the best data augmentation method improved the BLEU score for text translation from a baseline of 14.40 to 16.77, and for speech translation from a baseline of 12.03 to 14.32.

For speech translation, we used Kaldi and Whisper as speech recognizers in a cascade setup with the text translation model. Whisper's word error rate ranged from 3% to 14%, and the translation performance degradation for speech input compared to text input was 1 to 3 BLEU points, or a 0% to 20% decrease. We demonstrated that with a speech recognition accuracy of around 95% (word error rate of 5%), the BLEU score degradation compared to text translation performance is approximately 10%. In Appendix, examples of ASR results and translation results are shown. In future work, we will develop simultaneous speech translation system.

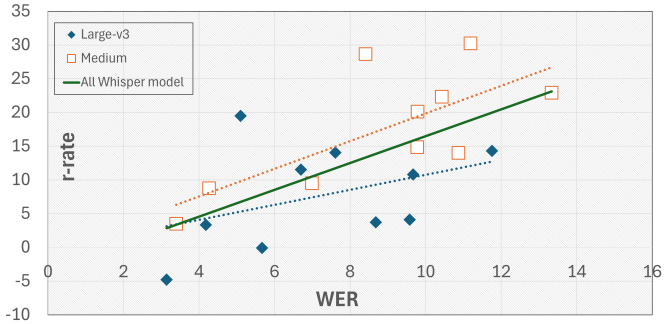
t



(a) Relation of WER of Kaldi ASR and reduction rate of BLEU



(b) Relation of WER of Whisper ASR and best translation model



(c) Relation of WER of Whisper ASR and baseline translation model

Fig. 2: Relation of WER of ASR and reduction rate of BLEU by speech translation to text translation

ACKNOWLEGMENT

This work was supported by JSPS KAKENHI Grant Numbers JP18H01062, JP1919K12027, JP22K12084.

REFERENCES

[1] H. Banno, H. Sakurai, J. Adachi, K. Yamamoto, S. Nakagawa, Consideration on Speech Recognition, Speech Translation and Speech Summarization for TED English Lectures, Proc. NLP, pp.1417–1422, 2023.3 (in Japanese)
[2] A. Anastasopoulos, O. Bojar, et al. : Findings of the IWSLT 2021 Evaluation Campaign, Proc. IWSLT-2021, pp.1–29, 2021.
[3] A. Anastasopoulos, L. Barrault, et al. : Findings of the IWSLT 2022 Evaluation Campaign, Proc. IWSLT-2022, pp.98–158, 2022.
[4] S. Ueno, A. Lee, T. Kawahara, Enhancing Synthesized Speech using Speaker Information and Phone Masking for Data Augmentation of Speech Recognition, Proc. Autumn Meeting of ASJ, pp.1149–1150, 2022.8 (in Japanese)
[5] W. Zhang, Z. Ye, et al. : The USTC-NELSLIP Offline Speech Translation Systems for IWSLT 2022, Proc. IWSLT, pp.198–207, 2022.
[6] R. Fukuda, Y. Ko, Y. Kano, et al. : NAIST Simultaneous Speech-to-Text Translation for IWSLT 2022, Proc. IWSLT 2022, pp.286–292 , 2022.
[7] M.Agarwal, et al. : Findings of the IWSLT 2023 Evaluation Campaign, Proc. IWSLT-2023, pp.1–61, 2023.
[8] Rico Sennrich, Barry Haddow, Alexandra Birch: Improving Neural Machine Translation Models with Monolingual Data, Proc. ACL, pp.86–96, 2016.
[9] Y. Yamagishi, T. Akiba, H. Tsukada, English–Japanese Machine Translation for Lecture Subtitles using Back-translation and Transfer Learning, Proc. 9-th GCCE, 2022
[10] J.E. Hu, et al. : Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting, Proc. Conf. NAACL, pp.839–850, 2019.
[11] Guangsheng Bao, Zhiyang Teng, Zhang, Yue: Target-SideAugmentation for Document-Level Machine Translation, Proc. ACL, pp.10725–10742, 2023.
[12] Y. Cheng, et al. : Semi-Supervised Learning for Neural Machine Translation, Proc. ACL, pp.1965–1974, 2016.
[13] A. Currey, et al. : Copied Monolingual Data Improves Low-Resource Neural Machine Translation, Proc. ACL, pp.148–156, 2017.
[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob, Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin: Attention is All You Need, Conference on Neural Information Processing Systems (NIPS2017), 30, 2017.
[15] M. Ott, et al. : fairseq: A Fast, Extensible Toolkit for Sequence Modeling, Proc. NAACL 2019, pp.48–53, 2019.

APPENDIX

Appendix: Example of ASR Result and Translation Result

Input English	I am learning that it's a genius idea to use a pair of barbecue tongs to pick up things that you dropped. I'm learning that nifty trick where you can charge your mobile phone battery from your chair battery.
Japanese Reference	私 が 学ん で いる の は 天才 的 な アイデア です 落とし た 物 を 拾う の に トング を 使う と か ね 他 に は 車 椅子 の バッテリー で 携帯 の 充電 を する 粋 な アイデア
Text Translation (BLEU=38.78)	私 は この アイデア を 学び まし た バーベキュー ・ トング を 使っ て 自分 が 落とし た もの を 手 に れる こと が できる と いう アイデア です 私 は 素晴らしい 技 学び まし た 皆 さん の 椅子 から 携帯 電話 を 充電 できる と いう の です
Whisper-Medium ASR (WER=5.13)	I am learning that it's a genius idea to use a pair of barbecue tongs to pick up things that you drop. I'm learning that nifty trick where you can charge your mobile phone battery from your chair battery.
Whisper-Medium Translation (baseline model) (BLEU=8.27)	私 は トング と いう 天才 的 な アイデア を 学び まし た バーベキュー ・ トング の 2 つ の バーベキュー セット を 使っ て あなた が 椅子 から 携帯 電話 を 充電 できる よう に する ため です
Whisper-Large-v3 ASR (WER=17.95)	I am learning that it's a genius idea to use a pair of paws, and to use your own barbecue tongs to pick up things that you drop. I'm learning that nifty trick where you can charge your mobile phone battery from your chair battery.
Whisper-Large-v3 Translation (best model) (BLEU=10.52)	私 が 学ん だ の は 義足 を 使う の は 天才 だ と いう こと そして 自分 の バーベキュー トング を 使っ て 落とし た もの を 手 に 入れる こと です 私 が 学ん だ の は スマート に 製 の トリック で 椅子 の バッテリー で 充電 できる と いう こと です

KhmerFormer: Multi-Scale CNNs-Transformer with External Attention for Ancient Khmer Palm Leaf Isolated Glyph Classification

Nimol Thuon* and Jun Du*†
* National Engineering Research Center of Speech and Language Information Processing (NERC-SLIP)
University of Science and Technology of China, Hefei, Anhui, 230027, China
E-mail: tnimol@mail.ustc.edu.cn, jundu@ustc.edu.cn

***Abstract*—Ancient Khmer palm leaf manuscripts are invaluable cultural artifacts in Southeast Asia, especially in Cambodia. The preservation and study of these manuscripts are hindered by their complex glyph structures and the scarcity of resources for ancient languages. This paper introduces KhmerFormer, a hybrid model that combines Multi-Scale Convolutional Neural Networks (MS-CNNs) and Vision Transformers (ViTs) with External Attention (EA) mechanisms to enhance glyph classification tasks. Our approach includes a preprocessing algorithm for Khmer glyph enhancements (IEPalmV2) and leverages multi-scale feature extraction with EfficientNet, followed by integration into ViTs where traditional self-attention is replaced by external attention. This model is tailored to address the unique and intricate challenges presented by ancient Khmer manuscripts. Our study is evaluated in ICFHR 2018 and newly extracted datasets show significant performance enhancements, particularly in handling low-resource and imbalanced datasets. Our results highlight the potential of hybrid architectures in advancing the analysis of historical documents and providing robust solutions for their preservation.**

I. Introduction

Ancient palm leaf manuscripts hold immense historical, cultural, and linguistic value, particularly in Southeast Asia. These manuscripts, containing intricate scripts and illustrations, present significant challenges for digital preservation and analysis due to their complex glyph structures and unbalanced datasets [1], [2]. The complexity, low quality, and unique characteristics of the Khmer palm leaf manuscript datasets introduce new challenges for analysis [3], [4]. Furthermore, degradation over time, such as fading ink and physical damage, further complicates the digitization and analysis of these documents [5], [6].

Advanced techniques, including Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and hybrid CNN-ViT models , have demonstrated strong performance in image classification tasks such as handwritten character recognition and historical document analysis [1], [7], [8]. However, challenges persist, especially when applying these models to Khmer palm leaf datasets, which include well-known architectures like ViTs, and hybrids CNN-Tranformer; despite their potential, ViTs often underperform compared to CNNs like EfficientNet and ResNet, largely due to the unique challenges posed by these datasets, such as varied characteristics, scales, and quality issues, which complicate the training process [5]. While Vision Transformers (ViTs) possess sophisticated self-attention mechanisms, they often fail to capture fine-grained details in certain contexts, leading to suboptimal results, particularly when dealing with unbalanced data where precise attention and feature extraction are critical [9], [10]. Recent studies have shown that external attention mechanisms can successfully handle complex datasets for classification tasks, enhancing the model's ability to focus on relevant features and thereby improving performance [11]. Therefore, exploring attention models, particularly those incorporating external attention mechanisms, could significantly benefit the analysis of palm leaf manuscripts, which often contain intricate details and imbalanced data.

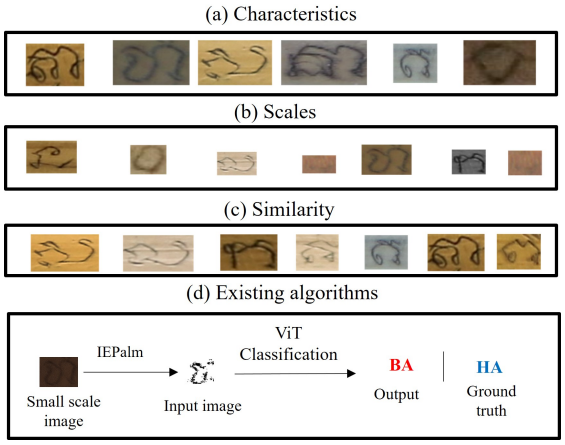


Fig. 1. Challenges in ancient Khmer script analysis: (a) Variability in glyph characteristics, (b) Diverse scales and low quality of some glyphs, (c) Similarity among characters, and (d) Illustration of challenges encountered in previous works [5] related to preprocessing and classification.

To address these challenges, we introduce KhmerFormer, which comprises two main components: preprocessing and

†Corresponding author

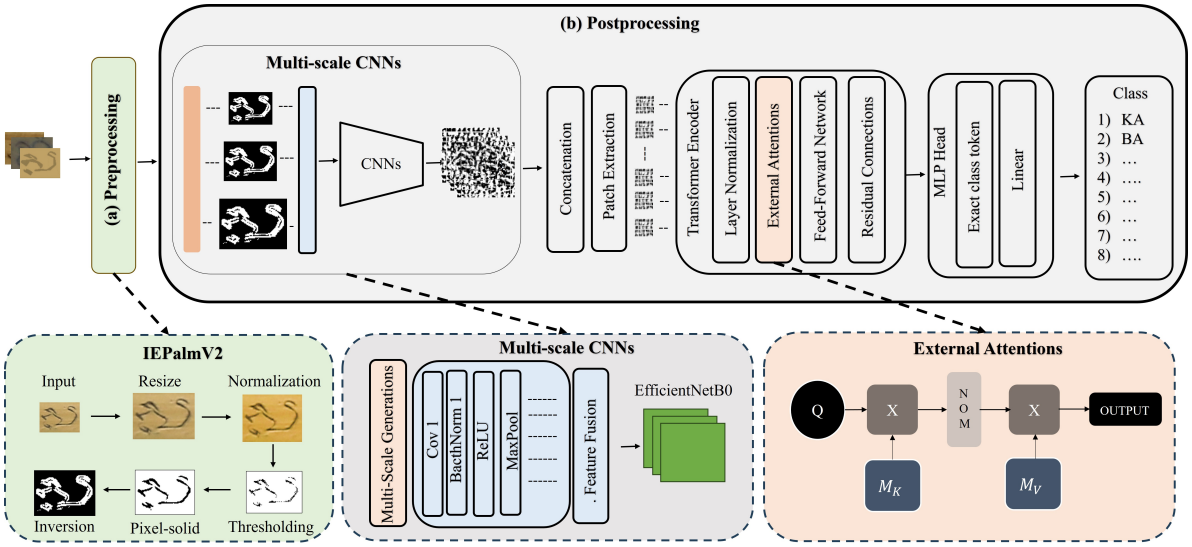


Fig. 2. This figure presents KhmerFormer's architecture for ancient Khmer isolated glyph classification, divided into two main components: (a) Preprocessing, which involves a straightforward algorithm named IEPalmV2 for enhancing palm leaf image quality; (b) Postprocessing, which includes a Vision Transformers (ViTs) [12] network enhanced with multi-scale CNNs and utilizes external attentions (EA) [11] instead of traditional self-attention (SA).

postprocessing. In preprocessing, we use a straightforward and effective algorithm to enhance low-quality training data by cleaning and solidifying the pixels. For postprocessing, we present a hybrid model that combines multi-scale CNN-Transformer architectures with external attention (EA) [11]. Extensive experiments on the ICFHR 2018 dataset [4] and newly extracted datasets [5] demonstrate significant improvements, especially in low-resource and imbalanced settings, underscoring the model's potential in advancing historical document analysis and preservation.

This paper makes the following significant contributions:

1) Enhanced the Khmer isolated glyph enhancement algorithms from previous approaches by integrating more refined cleaning and pixel enrichment techniques, significantly improving readability and quality.
2) Developed an effective CNN-Transformer network for Khmer palm leaf glyph classification, employing a multi-scale hybrid CNN-Transformer architecture enhanced with an external attention mechanism, which optimizes the capture of local and global features.
3) Evaluated the KhmerFormer model across multiple benchmarks, comparing its performance with state-of-the-art models and analyzing the impact of incorporating both ICFHR 2018 [4] and newly extracted datasets [5].

II. Khmer Palm Leaf Isolated Characters: Datasets and Challenges

The Khmer language, spoken by over 16 million people in Cambodia, consists of approximately 80 characters, including consonants, vowels, and special symbols. The ancient Khmer, found in palm leaf manuscripts, differs significantly from the modern Khmer and contains valuable educational, historical, and cultural content. The preservation of these manuscripts is crucial, especially given the substantial losses during civil wars [2], [3].

For isolated glyph datasets, we utilize the ICFHR 2018 challenge dataset [4], supplemented with manually extracted characters from text line datasets by Cambodian student [5]. These datasets include 111 classes of characters and short words. Khmer words can be formed from various combinations of symbols, consonants, and vowels. The dataset presents challenges such as varying quality conditions, scales, similarities, and imbalances (see details in Fig. 1). Previous research has identified several issues, including inadequate identification due to the preprocessing and postprocessing of small-scale, low-quality glyphs [6], [13]. For example, we tested the methods from the previous study [5] on the enhancement and classification of small glyphs often led to misidentifications or the loss of critical features, as illustrated in Fig. 1(d). Additionally, variations in quality, such as faded or damaged manuscripts, further complicate the postprocessing process.

Therefore, preprocessing and advanced postprocessing are essential to improve the accuracy and reliability of glyph identification and classification. Our proposed research study aims to address these challenges, enhancing the robustness and effectiveness of classifying ancient Khmer scripts.

III. Methodology

This section outlines the KhmerFormer frameworks for classifying isolated Khmer glyphs in palm leaf manuscripts, including preprocessing steps to enhance input image quality and postprocessing steps for classification tasks.

A. Preprocessing

In the preprocessing stage, we introduce a straightforward yet effective algorithm designed to enhance the readability and quality of characters extracted from palm leaf manuscripts,

referred to as IEPalmV2. Building upon the IEPalmV1 algorithm [5], we incorporate several key steps aimed at optimizing character enhancement. The preprocessing workflow consists of five primary steps, as illustrated in Fig. 2(a). The process is described as follows:

1) Resize: To ensure uniformity in input data, characters are resized to a standard dimension of 244×244 pixels. This resizing standardizes input sizes, thereby improving consistency in subsequent processing steps.
2) Normalization: We employ an improved variant Contrast Limited Adaptive Histogram Equalization (CLAHE) technique, augmented with Adaptive Histogram Clipping (AHC) [14]. This method dynamically adjusts clipping levels to prevent over-enhancement, thus reducing noise while preserving essential spatial frequency content.
3) Thresholding: We utilize a recently modified version of the Sauvola method, which is well suited to the unique challenges of palm leaf manuscripts [15]. The adaptive threshold, $T(x, y)$, is defined as:

$$T(x, y) = m(x, y) \left(1 + k \left(\frac{s(x, y)}{R} - 1\right)\right) \quad (1)$$

where $m(x, y)$ represents the local mean intensity, $s(x, y)$ denotes the local standard deviation, k is a sensitivity factor (usually between 0.2 and 0.5), and R indicates the dynamic range of standard deviations. This adaptive thresholding technique accommodates varying text densities and degradation levels, with integral images facilitating efficient computation.
4) Pixel Solidity: A dilation process using a 3×3 kernel is applied to enhance the continuity and legibility of black characters. The operation is defined as:

$$D(x, y) = \max_{(i,j) \in \text{Kernel}} \text{Image}(x + i, y + j) \quad (2)$$

Repeated iterations of this process increase pixel solidity, significantly enhancing text clarity.
5) Inversion: The binarized image is inverted, switching from a black-on-white to a white-on-black representation, as described by:

$$\text{Inverted_Image}(x, y) = 255 - \text{Image}(x, y) \quad (3)$$

Therefore, IEPalmV2 multi-task algorithm is designed to enhance the visibility of isolated glyphs before feeding them into the postprocessing stage, significantly improving the accuracy and effectiveness of classification/recognition tasks.

B. Postprocessing

*1) **Multi-scale convolutional neural networks (MS-CNNs):*** In the postprocessing stage, we employ MS-CNNs to extract features from enhanced images across multiple scales. This multi-scale approach enables the model to capture a wide range of detailed features, enhancing its capability to recognize and classify isolated Khmer glyphs. EfficientNetB0 [16] is incorporated for further processing, leveraging its compound scaling technique to deliver a balanced trade-off between depth, width, and resolution, resulting in robust feature representation and improved classification accuracy.

The procedure begins with the resizing of the input image I into various scales. Specifically, we utilize three scales: $s_1 = 1\times$ (original size), $s_2 = 0.5\times$, and $s_3 = 2\times$, represented as $I_{s_1}, I_{s_2}, I_{s_3}$. For each scaled image, a series of convolutional layers are applied to extract distinct features:

$$F_s = \text{ConvLayer}(I_s) \quad (4)$$

The convolutional operation at each scale involves applying multiple convolutional filters, followed by Batch Normalization and ReLU activation functions. This setup ensures that feature extraction is stable and efficient, mitigating issues like internal covariate shift and vanishing gradients:

$$F_s^{(l)} = \sigma\left(W^{(l)} * I_s + b^{(l)}\right) \quad (5)$$

Here, $W^{(l)}$ represents the convolutional filters, $*$ denotes the convolution operation, $b^{(l)}$ is the bias term, and σ is the ReLU activation function. The inclusion of Batch Normalization after each convolutional layer normalizes the feature maps, improving convergence rates and training stability.

Once the features are extracted from all scales, they are fused to create a comprehensive feature representation. This fusion can be performed either by concatenation, which stacks the feature maps along the depth dimension, or by averaging, which computes the mean of corresponding feature maps:

$$F = \text{Fusion}(F_{s_1}, F_{s_2}, F_{s_3}) \quad (6)$$

In our approach, we employ concatenation as it preserves more information from each scale. The fused features are then processed by EfficientNetB0 [16], which utilizes a compound scaling method to optimize the network's architecture. This includes proportionally adjusting the network's depth, width, and resolution to optimize both its capacity and efficiency:

$$F_{\text{EffNet}} = \text{EfficientNetB0}(F) \quad (7)$$

EfficientNetB0 [16] performs further high-level feature extraction, generating a final set of features suitable for classification tasks. The compound scaling is mathematically represented by:

$$\begin{aligned} d &= \alpha^k, \\ w &= \beta^k, \\ r &= \gamma^k, \end{aligned} \quad (8)$$

where d, w, and r represent the depth, width, and resolution scaling factors, respectively, and α, β, and γ are coefficients determined through grid search optimization.

*2) **ViTs with external attentions**:* After feature extraction using MS-CNNs, Vision Transformers (ViTs) equipped with external attention mechanisms are employed to enhance the representation of spatial relationships and local dependencies within the data. In this study, we opt for the standard ViT [17] and Swin Transformer [18] architectures, selected for their robust performance in handling complex visual tasks.

The feature map obtained from the MS-CNNs is first divided into non-overlapping patches. Each patch is then flattened and projected into a high-dimensional space via a linear transformation. Positional embeddings are added to retain spatial information, ensuring that the Transformer encoder can understand the spatial structure of the patches.

The Transformer encoder is composed of multiple layers, each featuring normalization, an external attention mechanism, and a feed-forward neural network (FFN). External attention [11], replaces traditional self-attention by focusing on a localized neighborhood of patches. This is particularly advantageous for minimizes overfitting and improves generalization:

$$\alpha_{ci} = \text{softmax}\left(\text{query}(c) \cdot \text{key}(p_i)\right) \tag{9}$$

In this mechanism, the query, key, and value for each patch are computed as follows:

$$\text{query}(c) = W_q c, \quad \text{key}(p_i) = W_k p_i, \quad \text{value}(p_i) = W_v p_i \tag{10}$$

Here, W_q, W_k, and W_v are learnable weight matrices. The local neighborhood N_c for a central patch c consists of the k nearest patches, determined by their spatial proximity. The parameter k is crucial as it controls the size of the receptive field, balancing the trade-off between capturing local interactions and computational efficiency.

The output from the external attention mechanism, v_c, aggregates contextual information from the neighboring patches:

$$v_c = \sum_{p_i \in N_c} \alpha_{ci} \text{value}(p_i) \tag{11}$$

This output is then combined with the original input via a residual connection, followed by layer normalization. The normalized output is subsequently processed through an FFN, which includes two linear transformations separated by a non-linear activation. The inclusion of residual connections and normalization layers ensures that the network remains stable and avoids issues such as exploding or vanishing gradients.

This sequence of operations, repeated across multiple Transformer layers, progressively refines the feature representations. The localized focus provided by external attention significantly enhances the model's ability to capture fine-grained details, thereby improving the classification performance, particularly in challenging low-resource and imbalanced data scenarios.

IV. Experimental setup

A. Experimental design

In this study, we assess the effectiveness of the proposed KhmerFormer architecture on Khmer palm leaf datasets [4], [5]. The evaluation includes comparisons with state-of-the-art methods and an analysis of the contributions from individual components, such as IEPalmV2, MS-CNNs, and the integration of external attention mechanisms (EA).

B. Dataset

Our evaluation is conducted using datasets from the ICFHR 2018 contest [4], specifically focusing on the Khmer dataset, along with additional manually extracted datasets [5]. The datasets are systematically categorized into classes and partitioned for training and testing, as detailed in Table I. The datasets are organized into two main categories:

- Kh_ICFHR2018: Comprises isolated Khmer glyphs from the ICFHR 2018 competition, serving as a standard benchmark for classification challenges [4].
- Kh_Mixed: A composite dataset that includes data from Kh_ICFHR2018, supplemented with additional manually extracted data (Kh_Additional), offering a richer and more diverse set of glyphs [5].

TABLE I
List of Khmer isolated glyph datasets

Dataset	Class	Training	Testing
Kh_ICFHR2018 [4]	111	113,206	90,669
Kh_Additional [5]	111	7,000	2,000
Kh_Mixed [4], [5]	111	120,206	92,669

C. Experiment setting

The experiments were conducted using NVIDIA 3090 GPUs with 24 GB RAM, leveraging both TensorFlow and PyTorch frameworks for model training. The EfficientNetB0 model [16] was initialized with pre-trained weights from the ImageNet dataset to enhance initial feature extraction. The palm leaf dataset images were uniformly resized to 224x224 pixels during preprocessing, as detailed in section III-B. The experimental workflow comprised a multitask preprocessing phase aimed at standardizing and enhancing the input data, followed by feature extraction using multiscale convolutional neural networks (MS-CNNs). The classification phase employed ViTs with external attention mechanisms, which were essential for capturing intricate patterns and dependencies within the data. The training process was consistently conducted over 100 epochs, allowing sufficient iterations for model convergence and performance stabilization.

D. Evaluation metrics

Model performance was evaluated using standardized metrics, with a primary focus on accuracy, as outlined in previous studies [1], [5].

V. Results and Analysis

A. Performance comparison with leading models

In this section, we present a comparative analysis of our proposed methods against state-of-the-art approaches, including models based on CNNs, ViTs, and hybrid models. We

define two variants of our architecture: KhmerFormer 1, which uses the standard ViT architecture, and KhmerFormer 2, which incorporates the custom Swin transformer models. As shown in Table II, our proposed KhmerFormer 2 demonstrates superior performance compared to previous methods for both the Kh_ICFHR2018 and Kh_Mixed datasets. This variant takes advantage of the hierarchical structure of the Swin Transformer, providing better representation and precision of features. On the other hand, KhmerFormer 1, while still competitive, shows performance comparable to EfficientNet models (B0, B1) [16], highlighting the effectiveness of incorporating external attention mechanisms with ViTs. The results indicate that our approaches, particularly KhmerFormer 2, offer significant improvements in classification accuracy over traditional and existing methods. The integration of multi-scale feature extraction and external attention mechanisms (EA) has proven to be effective in handling the complex characteristics of Khmer palm leaf manuscripts, especially in scenarios with imbalanced and low-quality datasets.

TABLE II
COMPARATIVE EVALUATION OF OUR PROPOSED KHMER GLYPH CLASSIFICATION MODEL AGAINST STATE-OF-THE-ART APPROACHES

Model	Kh_ICFHR2018	Kh_Mixed
VGG16 [5]	88.98	89.08
ResNet50 [5]	89.04	90.45
ResNet101 [5]	91.23	90.97
EfficientNetB0 [5]	90.56	91.85
EfficientNetB1 [5]	**91.67**	91.21
EfficientNetB2 [5]	91.05	91.55
EfficientNetB3 [5]	91.09	91.08
ViT-S 16 [5]	87.12	86.25
ViT-B 16 [5]	88.25	87.44
DeiT-S [5]	88.04	88.45
CvT-13 [5]	90.28	90.45
CvT-21 [5]	90.28	90.36
KhmerFormer 1	91.39	**92.19**
KhmerFormer 2	**93.65**	**94.86**

B. Ablation studies

To evaluate the contribution of each component in our model, we conducted ablation studies. We assessed the performance impact by systematically removing or modifying different components, such as IEPalmV2, MS-CNNs, and ViTs with external attention (EA).

TABLE III
EVALUATING EFFECTIVENESS OF DIFFERENT PREPROCESSING AND POSTPROCESSING TECHNIQUES

Preprocessing	Postprocessing	Kh_ICFHR2018	Kh_Mixed
RGB	EfficientNetB0 [5]	90.56	91.85
	KhmerFormer 1	89.05	89.81
	KhmerFormer 2	90.85	91.17
Binary (IEPalmV1)	EfficientNetB0 [5]	92.91	93.55
	KhmerFormer 1	90.87	91.47
	KhmerFormer 2	93.04	93.78
Binary (IEPalmV2)	EfficientNetB0	**93.03**	**93.87**
	KhmerFormer 1	**91.39**	**92.19**
	KhmerFormer 2	**93.65**	**94.86**

TABLE IV
COMPARATIVE EVALUATION OF MS-CNNS AND VIT MODELS WITH SELF-ATTENTION (SA) AND EXTERNAL ATTENTION (EA).

CNNs	ViTs	Kh_ICFHR2018	Kh_Mixed
ResNet50	ViT + SA	89.18	**90.98**
	ViT + EA	**89.65**	90.86
MS-ResNet50	ViT + SA	90.96	91.05
	ViT + EA	**91.35**	**91.49**
EffcientNetB0	ViT + SA	90.78	91.47
	ViT + EA	**91.06**	**91.95**
MS-EfficientNetB0	ViT + SA	90.91	91.84
	ViT + EA	**91.39**	**92.19**
	Swin + SA	92.08	92.90
	Swin + EA	**93.65**	**94.86**

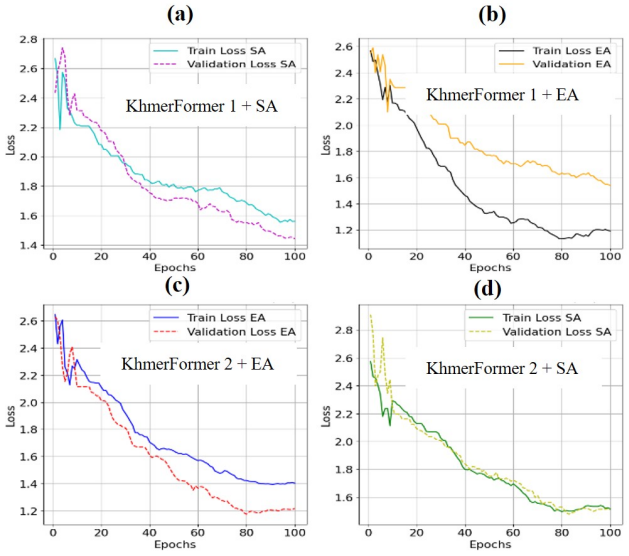


Fig. 3. Visualizing training performance of our proposed KhmerFormer 1 & 2 models with self-attention (SA) and external attention (EA)

*1) **Evaluating the impact of IEPalmV2**:* We evaluated the effect of the IEPalmV2 preprocessing pipeline by comparing the model's performance with and without this stage. The absence of the IEPalmV2 preprocessing steps, particularly enhancements such as normalization, thresholding, and pixel solidity, resulted in reduced model accuracy. As shown in Table III, the combination of IEPalmV2 preprocessing with EfficientNetB0 [16], KhmerFormer 1 & 2 demonstrated better results compared to previous studies using binary datasets IEPalmV1 [5] and original color images [4]. This underscores the significant impact of high-quality preprocessing before the postprocessing stage.

*2) **Assessing MS-CNNs performances**:* To demonstrate the effectiveness of MS-CNNs in combination with EfficientNet [16], we conducted experiments with and without the multi-scale feature extraction layer. The removal of MS-CNNs led to a noticeable decline in classification accuracy, highlighting the importance of capturing features at multiple resolutions. As illustrated in Table IV, we equipped different configurations, such as ResNet50 [19] and EfficientNetB0 [16], with ViTs

architectures like ViT [12] and Swin Tranformer [18], incorporating both self-attentions (SA) [20] and external attentions (EA) [11]. The results show the significant improvement and differences made by the inclusion of MS-CNNs.

3) ***External attention mechanism efficiency analysis****:* We performed a detailed comparative analysis between models using self-attention (SA) and those using external attention (EA) within the Vision Transformer (ViT) architectures. As shown in Table IV, external attention demonstrated an improved capability in handling complex datasets, such as palm leaf manuscripts. The results indicated that models incorporating external attention not only achieved higher classification accuracy but also exhibited better robustness in feature extraction step. These findings underscore the potential benefits of integrating external attention mechanisms into ViT frameworks for more effective processing of intricate and unbalanced datasets.

4) ***Training Dynamics and Performance****:* As shown in Fig. 3, we visualize the training and validation loss throughout the training process. The analysis reveals that KhmerFormer 1 & 2, incorporating external attention (EA) and self-attention (SA), achieved excellent results, demonstrating consistent training and validation loss performance across epochs. These models successfully captured complex features, resulting in more accurate and stable learning. This underscores the advantages of our hybrid models in managing localized information and reducing overfitting. These findings emphasize the effectiveness of our proposed approaches in enhancing model performance on complex datasets, such as ancient Khmer palm leaf manuscripts.

VI. Conclusion

In conclusion, this study proposed effective methods for preprocessing and postprocessing to improve classification tasks for ancient Khmer isolated glyphs in palm leaf manuscripts. The IEPalmV2 preprocessing method, though simple, significantly enhances image quality and preserves important information. For postprocessing, the hybrid MS-CNNs and the integration of Vision Transformers (ViTs) with external attention mechanisms led to substantial improvements in classification performance. However, the study faced limitations, including the quality and quantity of available datasets, which may affect the generalizability of the models, and the substantial computational resources required for training. Future research should aim to incorporate deeper linguistic features into the models, develop more resource-efficient architectures, and expand the dataset to enhance the robustness and applicability of this technology in real-world scenarios.

Acknowledgment

This work is support by CAS-TWAS academy science and National Natural Science Foundation of China under Grant No.62171427.

References

[1] M. W. A. Kesiman, D. Valy, and Burie, "Benchmarking of document image analysis tasks for palm leaf manuscripts from southeast asia," *Journal of Imaging*, vol. 4, no. 2, p. 43, 2018.

[2] F. E. Huffman, "The register problem in fifteen mon-khmer languages," *Oceanic Linguistics Special Publications*, no. 13, pp. 575–589, 1976.

[3] D. Valy, M. Verleysen, S. Chhun, and J.-C. Burie, "A new khmer palm leaf manuscript dataset for document analysis and recognition: Sleukrith set," in *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, 2017, pp. 1–6.

[4] M. W. A. Kesiman, "Icfhr 2018 competition on document image analysis tasks for southeast asian palm leaf manuscripts," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, IEEE, 2018, pp. 483–488.

[5] N. Thuon, J. Du, and J. Zhang, "Improving isolated glyph classification task for palm leaf manuscripts," in *International Conference on Frontiers in Handwriting Recognition*, Springer, 2022, pp. 65–79.

[6] N. Thuon, J. Du, and J. Zhang, "Syllable analysis data augmentation for khmer ancient palm leaf recognition," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 1855–1862.

[7] D. Parres and R. Paredes, "Fine-tuning vision encoder–decoder transformers for handwriting text recognition on historical documents," in *International Conference on Document Analysis and Recognition*, Springer, 2023, pp. 253–268.

[8] S. Inunganbi, "A systematic review on handwritten document analysis and recognition," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5387–5413, 2024.

[9] Y. Xu, C. Zhu, S. Wang, *et al.*, "Human parity on commonsenseqa: Augmenting self-attention with external attention," *arXiv preprint arXiv:2112.03254*, 2021.

[10] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," *arXiv preprint arXiv:2112.13492*, 2021.

[11] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5436–5447, 2022.

[12] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] N. Thuon, J. Du, Z. Zhang, J. Ma, and P. Hu, "Generate, transform, and clean: The role of gans and transformers in palm leaf manuscript generation and enhancement," *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 1–18, 2024.

[14] Y. Chang, C. Jung, P. Ke, H. Song, and J. Hwang, "Automatic contrast-limited adaptive histogram equalization with dual gamma correction," *Ieee Access*, vol. 6, pp. 11 782–11 792, 2018.

[15] P. Bannigidad and S. Sajjan, "Restoration of ancient kannada handwritten palm leaf manuscripts with modified sauvola method using integral images," in *International Conference on ICT for Sustainable Development*, Springer, 2023, pp. 39–47.

[16] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

DDPMVC: Non-parallel any-to-many voice conversion using diffusion encoder

Ryuichi Hatakeyama*,Kohei Okuda† and Toru Nakashika‡
* The University of Electro-Communications, Tokyo
E-mail: r.hatakeyama@uec.ac.jp
† The University of Electro-Communications, Tokyo
E-mail: okuda-k@dym.jp
‡ The University of Electro-Communications, Tokyo
E-mail: nakashika@uec.ac.jp

***Abstract*—In this paper, we propose DDPMVC, a voice conversion (VC) model for non-parallel data that utilizes the diffusion model. The distinctive feature of diffusion models lies in their high expressive power for high-dimensional data and their ability to learn more stably compared to traditional generative models. Although several VC methods using diffusion models have been proposed, only VoiceGrad meets the non-parallel and any-to-many condition, which makes it easy to handle in training and inference. DDPMVC is regarded as an extension of VoiceGrad that incorporates a rule-based diffusion process into the encoder. By using the encoder to convert speech into latent variables that have less speaker information, it is expected to improve the accuracy of the non-parallel VC. Experimental results demonstrated that the performance of DDPMVC surpassed that of VoiceGrad in terms of the mel cepstral distortion and speaker similarity.**

I. Introduction

Voice conversion (VC) is a technology that enables the transformation of a speaker identity into that of another without altering the speech content. It is categorized based on the type of data utilized: parallel data, which involves aligning the same spoken content across different speakers, and non-parallel data, which involves the arbitrary alignment of spoken content. Conversion targets include those from one specific speaker to another (`One-to-One`) and from specific multiple speakers to other specific multiple speakers (`Many-to-Many`). However, converting any speaker to specific multiple speakers (`Any-to-Many`) is deemed the ideal scenario.

Currently, research on VC predominantly utilizes deep learning models, with variational autoencoders (VAE) [1] and generative adversarial networks (GAN) [2] being notable examples. However, recent advancements have introduced diffusion models for voice conversion, demonstrating superior performance over traditional deep learning approaches such as VAE and GAN [3]–[5]. The diffusion-based VC [4] utilizes the continuous-time denoising diffusion probabilistic model (DDPM) [6], [7], distinguished by its use of an average voice extractor in the encoder. Similarly, DiffSVC [3] utilizes DDPM for VC, specifically targeting the conversion of singing voices. However, the diffusion-based VC requires parallel data for Many-to-Many VC models, and DiffSVC, while aimed at singing voices, does not support multiple speakers in Any-to-One conversions. In contrast, VoiceGrad [5] utilizes a score-based generative modeling (SBM) [8] within diffusion models for VC, achieving Any-to-Many conversions using non-parallel data. Non-parallel VC methods generally assume that the speech is generated by speaker-independent (phonetic) information accompanied with speaker information, and consist of an encoder that extracts phonetic information from speech and a decoder that generates speech from it given a target speaker's label [9]. VoiceGrad performs sampling without the use of an encoder. Consequently, it attempts the conversion to the target speaker without separating the speaker characteristics of the source speaker. This approach can result in suboptimal conversion accuracy, particularly when there are significant differences between the characteristics of the source and target speakers. Therefore, there is potential for improvement in the conversion accuracy of VoiceGrad under these conditions. In this study, we propose DDPMVC, a voice conversion model based on the diffusion model that incorporates a rule-based diffusion process into the encoder to extract speaker-independent information, building upon VoiceGrad's conversion process.Using a diffusion model for both the encoder and decoder in voice conversion simplifies the process and reduces training effort. We conducted a comparative analysis of the quality of conversion accuracy between DDPMVC and VoiceGrad.

II. Diffusion Model

A. Score-based generative modeling (SBM)

The score-based generative modeling (SBM) [8] utilizes scores to efficiently compute maximum likelihood estimates in high-dimensional data, addressing the challenges associated with these calculations. Given a likelihood $p(\boldsymbol{x})$ for an input variable $\boldsymbol{x}$, the gradient of the log likelihood $\log p(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, referred to as the score $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$, is denoted by the function $s(\boldsymbol{x})$. The common training approach for SBM is denoising score matching (DSM) [10], which focuses on learning scores related to the conditional probability during perturbation, rather than directly learning the scores. Let $\boldsymbol{s}_\theta(x)$ represent the score estimated by the model parameterized by

θ. The loss function for denoising score matching in SBM is then expressed as follows:

$$L_{\mathrm{SBM}}(\theta) = \sum_{l=1}^{L} w_l \mathbb{E}_{\boldsymbol{x},\tilde{\boldsymbol{x}}} \left[\left\| \frac{\boldsymbol{x} - \tilde{\boldsymbol{x}}}{\sigma_l^2} - \boldsymbol{s}_\theta(\tilde{\boldsymbol{x}}, \sigma_l) \right\|^2 \right], \quad (1)$$

where $\boldsymbol{x} \sim p(\boldsymbol{x})$, $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{x}, \sigma_l^2 \boldsymbol{I})$, and w_l represents the weight of the loss for each noise step. This approach utilizes perturbed distributions, disturbed by adding noise of various magnitudes to the data distribution, to calculate the score on these distributions. To align with Eq. (1), sampling is performed for each perturbed distribution using various noises through the Langevin Monte Carlo method. This is referred to as annealed Langevin dynamics [8].

B. Denoising diffusion probabilistic model (DDPM)

The denoising diffusion probabilistic model (DDPM) [7] consists of a diffusion process that gradually adds noise to the data $\boldsymbol{x}_0$ and a reverse diffusion process that traces back the diffusion process to remove the added noise. Viewing the noise-added data as latent variables, maximum likelihood estimation is performed as a latent variable model, and data sampling is conducted by sequentially sampling the latent variables through the reverse diffusion process. The diffusion process considers a Markov process that gradually adds noise to the data $\boldsymbol{x}_0$ and obtains a sequence of noise-added data $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T\}$. Specifically, this is expressed by

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \quad (2)$$

where $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})$ and $\{\beta_1, ..., \beta_T\}$ are noise parameters that control the variance magnitude. The reverse diffusion process is a generative process that starts from pure noise $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{0}, \boldsymbol{I})$ and gradually removes noise to generate data. Each step is normally distributed, and their means and variances are characterized by a model with parameters θ. This process is described by

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \quad (3)$$

where $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t))$. The parameters of this model are obtained by performing maximum likelihood estimation. The objective function is given by Eq. (4), where w_t represents the weight at each time step:

$$L(\theta) = \sum_{t=1}^{T} w_t \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|^2 \right]. \quad (4)$$

Here, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ represents the model that estimates the noise. With real data $\boldsymbol{x}_0$, $\boldsymbol{x}_{1:T}$ are considered latent variables. Since this involves maximum likelihood estimation with latent variables, it is performed by maximizing the variational lower bound (ELBO) [1] of the log-likelihood.

C. Continuous diffusion model

Song et al. (2021) [11] extended discrete-time models, such as SBM and DDPM, to continuous-time formulations by defining them as stochastic differential equations (SDEs). The continuous-time formulation of SBM, shown in Eq. (5), is referred to as the VE SDE, while the continuous-time formulation of DDPM, given in Eq. (6), is referred to as the VP SDE.

$$d\boldsymbol{x} = \sqrt{\frac{d[\sigma(t)^2]}{dt}} d\boldsymbol{w} \quad (5)$$

$$d\boldsymbol{x} = -\frac{1}{2}\beta(t)dt + \sqrt{\beta(t)}d\boldsymbol{w} \quad (6)$$

Here, $\boldsymbol{w}$ represents a standard Wiener process, or Brownian motion, and $d\boldsymbol{w}$ denotes a normal distribution with mean 0 and variance τ over an infinitesimal time interval τ. In continuous time, t is defined as $t \in [0, 1]$, and small time intervals are treated as $\Delta t = 1/N$, with the limit taken as $N \to \infty$. In this limit, the sequence $\{\sigma_i\}_{i=1}^N$ becomes a function $\sigma(t)$, and the sequence $\{\beta_i\}_{i=1}^N$ becomes a function $\beta(t)$. The difference between VE SDE and VP SDE lies in the method of adding noise. In VE SDE, only the noise increases while keeping the input the same, with the variance diverging. In VP SDE, the input gradually disappears, and the noise gradually increases to a certain magnitude, keeping the variance constant. According to [12], the SDEs for the reverse diffusion processes of SBM and DDPM are respectively given by Eqs. (7) and (8):

$$d\boldsymbol{x} = \left[-\frac{d[\sigma(t)^2]}{dt} \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \right] dt + \sqrt{\frac{d[\sigma(t)^2]}{dt}} d\boldsymbol{w}, \quad (7)$$

$$d\boldsymbol{x} = \left[-\frac{1}{2}\beta(t)\boldsymbol{x} - \beta(t)\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \right] dt + \sqrt{\beta(t)} d\boldsymbol{w}. \quad (8)$$

Learning utilizes the conditional probability of the diffusion process (diffusion kernel) $p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))$. The diffusion kernel can also be expressed as a normal distribution [13]. The diffusion kernels for VE SDE and for VP SDE are respectively given by

$$p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0)) = \mathcal{N}(\boldsymbol{x}(t); \boldsymbol{x}(0), [\sigma(t)^2 - \sigma(0)^2]\boldsymbol{I}), \quad (9)$$

$$p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0)) = \mathcal{N}(\boldsymbol{x}(t); \boldsymbol{x}(0)e^{-\frac{1}{2}\gamma(t)}, \boldsymbol{I} - \boldsymbol{I}e^{-\gamma(t)}), \quad (10)$$

where $\gamma(t) = \int_0^t \beta(s)ds$. By estimating the score using this diffusion kernel, data can be sampled in the reverse diffusion process using the estimated score. The objective function is given by Eq. (11), where $\lambda(t)$ represents the weighting at each time step:

$$\begin{aligned} \theta^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathbb{E}_t \big\{ \lambda(t) \mathbb{E}_{\boldsymbol{x}(0)} \mathbb{E}_{\boldsymbol{x}(t)|\boldsymbol{x}(0)} \\ \left[||\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}(t), t) - \nabla_{\boldsymbol{x}(t)} \log p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))||_2^2 \right] \big\}. \end{aligned} \quad (11)$$

Sampling for the SDE is performed using the Euler-Maruyama method (EMM) [14] and predictor-corrector (PC) sampling [11]. The Euler-Maruyama method is a technique for obtaining samples by utilizing the score $\boldsymbol{s}_\theta(\boldsymbol{x}, t)$, estimated through score matching, in the first-order approximation of the reverse diffusion process (as in Eqs. (7) and (8)). Additionally,

the PC sampling is a sampling technique composed of two parts: a predictor, which utilizes a numerical solver for the discretized reverse diffusion SDE, and a corrector, which utilizes a score-based Markov Chain Monte Carlo (MCMC) method. The algorithm executes the predictor process N times, and for each execution, the corrector process is executed M times. The predictor involves the use of reverse diffusion samplers [11], which are the discretized equations of the reverse diffusion SDE. The corrector is executed to correct the results of the predictor. Techniques for the corrector include sampling by the Langevin Monte Carlo method and annealed Langevin dynamics.

III. Conventional Method: VoiceGrad

In this section, we briefly review VoiceGrad [5], which is related to our proposed method. VoiceGrad is a VC model based on SBM and DDPM. It based on SBM learns scores using Eq. (12) adapted for multiple speakers in DSM, and performs acoustic feature sampling through iterative updates of annealed Langevin dynamics (ALD) using the estimated scores. Additionally, It based on DDPM learns noise using Eq. (13) adapted for multiple speakers, and performs acoustic feature sampling through the reverse diffusion [5] using the estimated noise.

$$L_{\mathrm{DSM}}(\theta) = \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}_{k,\boldsymbol{x},\tilde{\boldsymbol{x}}}\left[\left\|\frac{\boldsymbol{x}-\tilde{\boldsymbol{x}}}{\sigma_l} - \sigma_l \boldsymbol{s}_\theta(\tilde{\boldsymbol{x}},\sigma_l,k)\right\|^2\right], \tag{12}$$

$$L_{\mathrm{DPM}}(\theta) = \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_l,l,k) - \boldsymbol{\epsilon}\right\|^2\right]. \tag{13}$$

where let $k \sim p(k), \boldsymbol{x} \sim p(\boldsymbol{x}|k), \tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{x}, \sigma^2\boldsymbol{I})$. The target speaker is conditioned on the speaker index k. The acoustic feature used is the mel spectrogram. The key characteristic of this model is the absence of an encoder. It is trained on non-parallel data and supports Any-to-Many conversions. However, during sampling, since the input speech retains not only phonetic information but also speaker information, there is a possibility that the speaker characteristics of the source speech may remain after conversion.

IV. Proposed Method: DDPMVC

In this work, we propose a VC method that utilizes a rule-based diffusion process as an encoder (diffusion encoder) and reverse diffusion sampling as a decoder, inspired by the VC process of VoiceGrad. By employing a rule-based diffusion process in the encoder, it is expected that it will be possible to remove the speaker's information from the input speech, leaving only the phonetic information. The noise added during the diffusion process is intended to be mild enough to disrupt the fine structures that represent the speaker's information in the input speech without destroying the phonetic information. Traditional VC methods using diffusion models [3], [4] remove the speaker identity from the input speech using a different technique in the encoder, and then input the speech, which retains only phonetic information, into a decoder that uses a diffusion model to produce a voice converted to the target speaker's identity. On the other hand, DDPMVC integrates the encoder and decoder within a unified diffusion model framework, allowing for holistic optimization and thus, it is expected to improve conversion accuracy. Furthermore, by utilizing continuous-time diffusion models, it is believed that the discretization errors that occur in SBM and DDPM can be reduced, potentially leading to better conversion accuracy compared to VoiceGrad. Moreover, we anticipate that the use of continuous-time diffusion models will reduce the discretization errors encountered in SBM and DDPM.

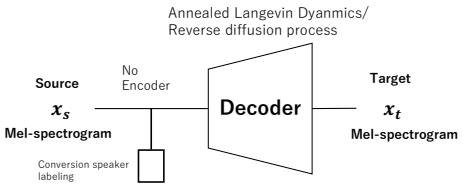


Fig. 1. Schematic diagram of VoiceGrad.

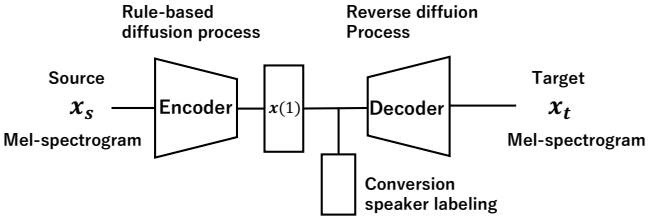


Fig. 2. Schematic diagram of DDPMVC.

The acoustic features in DDPMVC are 80-dimensional mel-spectrograms. The model is trained using these features, and during conversion, it outputs the acoustic features converted to the target speaker. Additionally, the mel-spectrogram $x_{d,m}$, where d is the d-th mel filter bank channel and m is the m-th frame index, is normalized using the mean ϕ_d and variance ξ_d as $x_{d,m} \leftarrow \frac{x_{d,m}-\phi_d}{\xi_d}$. The conversion process is shown in Fig. 2.

The mel-spectrograms $\boldsymbol{x}_s$ of the source speech is subjected to noise addition through a rule-based diffusion process (diffusion encoder) in the encoder. This involves adding noise to the extent that the content of the speech is somewhat preserved. The purpose of this is to minimize the possibility of incorrect exploration during sampling, thereby preventing alterations in the speech content as much as possible. The diffusion encoder utilizes the diffusion kernels given by Eqs. (9) and (10). The noise schedule for the VE SDE is set to $\sigma(t) \in [0.01, 0.85]$, and for the VP SDE, it is set to $\beta(t) \in [0.1, 2]$. Under this noise schedule, the clarity of the speech after adding noise was evaluated using Short-Time Objective Intelligibility (STOI) [15]. STOI is an objective evaluation method that takes clean speech and noisy speech as inputs and outputs a score between 0.0 and 1.0, with higher scores indicating better intelligibility. The clarity of the noisy speech for all test data was assessed using STOI, resulting in a score of 0.68 ± 0.10 for VE SDE and 0.52 ± 0.10 for VP SDE. These

TABLE I
COMPARISON OF MEL CEPSTRAL DISTORTION (MCD) WITH 95% CONFIDENCE INTERVALS BETWEEN VOICEGRAD AND DDPMVC. "PREDICTOR" INDICATES THAT REVERSE DIFFUSION SAMPLERS ARE USED FOR SAMPLING.

Type	SBM-VG	DDPM-VG	DDPMVC				
			DDPM	Predictor		PC	
				VE SDE	VP SDE	VE SDE	VP SDE
M→M	7.42 ± 0.07	7.06 ± 0.08	6.50 ± 0.07	7.25 ± 0.08	6.50 ± 0.07	7.02 ± 0.06	$\mathbf{6.48 \pm 0.07}$
F→F	6.71 ± 0.09	6.19 ± 0.09	$\mathbf{5.86 \pm 0.08}$	6.40 ± 0.08	5.90 ± 0.09	6.64 ± 0.08	6.11 ± 0.09
M→F	7.62 ± 0.09	7.03 ± 0.11	6.59 ± 0.08	7.57 ± 0.10	6.60 ± 0.08	7.31 ± 0.07	$\mathbf{6.50 \pm 0.08}$
F→M	7.15 ± 0.07	6.76 ± 0.10	$\mathbf{6.24 \pm 0.07}$	6.99 ± 0.07	6.27 ± 0.08	6.90 ± 0.06	6.31 ± 0.07
All	7.28 ± 0.05	6.81 ± 0.06	$\mathbf{6.34 \pm 0.05}$	7.13 ± 0.06	6.36 ± 0.05	7.02 ± 0.04	6.37 ± 0.04

scores indicate that a moderate level of noise was added, leaving the speech content partially intelligible. The speaker-independent mel-spectrograms $\boldsymbol{x}(1)$, to which noise has been added, undergoes reverse diffusion sampling of DDPM, VE/VP SDE in the decoder, ultimately yielding the target speaker's mel-spectrograms $\boldsymbol{x}_t$. The objective function for training is defined by Eq. (13) for DDPM and by Eq. (11) for VE/VP SDE. Finally, the transformed acoustic features are utilized to generate speech using HiFi-GAN[16].

V. EXPERIMENTS

A. Experiment setup

The dataset utilized for this study was the CMU Arctic voice corpus [17], which contains recordings of English-speaking individuals. Four speakers were selected for training, comprising two male speakers (`aew`, `ksp`) and two female speakers (`clb`, `lnh`), with 128 utterances for training and 30 utterances for testing each. Additionally, to assess the accuracy of VC for arbitrary speakers, 30 utterances from one male speaker (`bdl`) and one female speaker (`slt`) were added for sampling. The audio data was downsampled to 16 kHz and converted into 32-dimensional mel cepstrums as acoustic features. The WORLD vocoder [18] was used for voice analysis and synthesis. The network architecture utilized for training was based on U-Net [19], consisting of approximately 44.72 million parameters.

B. Objective evaluation

To assess the effectiveness of DDPMVC's encoder and the continuous diffusion model, experiments were conducted in stages. First, we compared VoiceGrad using SBM (SBM-VoiceGrad) with VoiceGrad using DDPM (DDPM-VoiceGrad). The settings for each VoiceGrad model were based on [5], with SBM having 11 time steps ($L = 11$) and DDPM having 20 time steps ($L = 20$). Next, to verify the effectiveness of the diffusion encoder in DDPMVC, a model was added that incorporates the diffusion encoder based on VP SDE into DDPM-VoiceGrad. By directly comparing it with DDPM-VoiceGrad, we could assess the impact of the diffusion encoder. This additional model is also treated as DDPMVC by definition. Finally, we conducted comparative experiments using DDPMVC. It was set to $N = 1000$ time steps, with the noise schedule for VE SDE set to $\sigma(t) \in [0.01, 0.85]$ and for VP SDE to $\beta(t) \in [0.1, 2]$. The sampling methods used were Reverse diffusion sampling and PC sampling. All these experiments used the trained speakers' speech as test data. For PC sampling, it used reverse diffusion samplers as the predictor and annealed Langevin dynamics as the corrector.

Additionally, experiments were conducted on unknown speakers who were not part of the training dataset. The specifications for SBM-VoiceGrad and DDPM-VoiceGrad remained as previously described. For DDPMVC, both VESDE and VPSDE were utilized, with Reverse diffusion sampling and PC sampling used as the sampling methods.

C. Subjective evaluation

In the subjective evaluation experiment, 13 participants conducted mean opinion score (MOS) tests to assess the naturalness of the converted speech and degradation mean opinion score (DMOS) tests to assess their similarity to the target speakers. The methods compared included Ground Truth, VoiceGrad using SBM and DDPM, and DDPMVC utilizing reverse diffusion samplers of VP SDE. For the evaluation of naturalness, 50 test questions were presented, comprising generated speech of the trained speakers from the three methods compared, along with the ground truth. The naturalness of the sound quality was rated on a 5-point scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). For the evaluation of speaker similarity, participants were given 50 test questions to assess the similarity between the trained speakers, converted using the three methods, and the target speakers. The similarity between the target and the converted speech was rated on a 5-point scale (1: Unlikely, 2: Not very likely, 3: Fairly likely, 4: Likely, 5: Definitely). Participants were asked to evaluate the results ignoring noise.

D. Results

Mel cepstral distortion (MCD) was used as an objective evaluation metric. The patterns of VC conducted were male-to-male (M→M), female-to-female (F→F), male-to-female (M→F), and female-to-male (F→M), forming four pairs. All results were calculated with a 95% confidence interval. First, when comparing SBM-VoiceGrad (SBM-VG) and DDPM-VoiceGrad (DDPM-VG), it can be seen from Table I that DDPM-VoiceGrad outperforms SBM-VoiceGrad in all conversion patterns. This indicates that DDPM achieves higher accuracy in voice conversion during the VoiceGrad transformation process. Next, the effectiveness of the diffusion encoder is examined. According to Table I, comparing DDPM-VoiceGrad with the version that includes a diffusion encoder (DDPMVC, DDPM) shows that the latter achieves higher

TABLE II
COMPARISON OF VOICE CONVERSION FOR UNKNOWN SPEAKERS.

Type	SBM-VG	DDPM-VG	DDPMVC				
			DDPM	Predictor		PC	
				VE SDE	VP SDE	VE SDE	VP SDE
M→M	7.14 ± 0.09	6.79 ± 0.10	$\mathbf{6.25 \pm 0.10}$	6.79 ± 0.10	6.33 ± 0.10	6.95 ± 0.08	6.45 ± 0.10
F→F	6.65 ± 0.08	6.00 ± 0.08	5.75 ± 0.07	6.24 ± 0.08	$\mathbf{5.74 \pm 0.08}$	6.53 ± 0.06	6.01 ± 0.08
M→F	7.40 ± 0.07	6.73 ± 0.08	$\mathbf{6.26 \pm 0.08}$	7.27 ± 0.08	6.30 ± 0.08	7.07 ± 0.07	6.38 ± 0.09
F→M	7.36 ± 0.08	6.91 ± 0.11	6.45 ± 0.09	7.15 ± 0.10	$\mathbf{6.44 \pm 0.10}$	7.05 ± 0.07	6.53 ± 0.09
All	7.14 ± 0.06	6.60 ± 0.06	$\mathbf{6.18 \pm 0.05}$	6.86 ± 0.07	6.20 ± 0.06	6.90 ± 0.05	6.34 ± 0.05

TABLE III
RESULTS OF SUBJECTIVE EVALUATION.

	MOS	DMOS
Ground Truth	4.39 ± 0.18	–
SBM-VG	1.95 ± 0.15	1.90 ± 0.13
DDPM-VG	$\mathbf{3.12 \pm 0.13}$	2.98 ± 0.18
DDPMVC	2.62 ± 0.15	$\mathbf{3.29 \pm 0.16}$

accuracy in all conversion patterns. This demonstrates the effectiveness of the diffusion encoder utilized in DDPMVC. It is thought that the diffusion encoder helps to separate speaker characteristics to some extent, and by performing reverse diffusion from a state distanced from the original speaker, it becomes easier to approach the target speaker. Next, the results of DDPMVC with continuous diffusion model are reviewed. When compared to SBM-VoiceGrad, it is evident that DDPMVC outperforms SBM-VoiceGrad in all sampling methods. Compared to DDPM-VoiceGrad, DDPMVC utilizing VESDE showed lower accuracy, but DDPMVC utilizing VPSDE demonstrated higher accuracy. When compared to Enc+DDPM-VG, the DDPMVC with VPSDE utilizing PC sampling outperformed in conversions from male to male and male to female. However, it showed lower accuracy in female-to-female, female-to-male conversions, and overall evaluation, indicating that the effectiveness of the continuous diffusion model is limited.

The results for unknown speakers are presented in Table II. DDPM-based DDPMVC showed the best overall performance. Furthermore, considering DDPMVC as a whole, it is clear that it surpasses the performance of SBM-VoiceGrad. Compared to DDPM-VoiceGrad, DDPMVC with VPSDE and DDPMVC with DDPM showed higher accuracy in each sampling method. Although the overall MCD values are lower than those for known speakers, this is because the pre-conversion voices were already similar to the target voices, resulting in lower initial MCD values.

The results of the subjective evaluation are shown in Table III. In terms of naturalness evaluation (MOS), DDPMVC showed better naturalness compared to SBM-VoiceGrad but was inferior to DDPM-VoiceGrad. However, in the similarity evaluation (DMOS), which is more important than naturalness in VC, DDPMVC outperformed VoiceGrad, indicating that DDPMVC's voices are closer to the target voices. These results are presumably due to the fact that the converted speech by VoiceGrad remained closely similar to the input speech with minimal conversion, whereas the converted speech by DDPMVC underwent a degree of conversion. As a result, more artifacts occurred with DDPMVC during conversion, which leads to less naturalness in converted speech by DDPMVC than VoiceGrad. It is worth noting that some techniques used in state-of-the-art VC methods such as phoneme posteriorgram (PPG) [20], [21], self-supervised learning (SSL) [22]–[24] could improve the overall MOS; our aim is to compare the accuracy of mel-spectrogram conversion by DDPMVC with a baseline, VoiceGrad.

VI. CONCLUSIONS

In this paper, we proposed DDPMVC, an Any-to-Many voice conversion model for non-parallel data that utilizes a rule-based diffusion process as the encoder and a continuous-time diffusion model's reverse diffusion process as a decoder. Compared to the conventional method VoiceGrad, DDPMVC, especially with VPSDE, showed superior results in MCD evaluation. However, in terms of MOS evaluation, it resulted in an inferior performance compared to VoiceGrad. In the evaluation of DMOS, DDPMVC achieved closer scores to the target speech, suggesting that while the conversion became closer to the target speaker, this might have resulted in a reduction in naturalness. Going forward, we aim to improve the model such that it can enhance the naturalness of the converted speech while maintaining the similarity to the target speaker. Furthermore, by utilizing methods such as PPG to make the intermediate representations even more speaker-independent, it is expected that naturalness can be increased, leading to further improvements in accuracy.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI under Grant 24H00715.

REFERENCES

[1] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2022. arXiv: 1312.6114 [stat.ML].

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014.

[3] S. Liu, Y. Cao, D. Su, and H. Meng, "DiffSVC: A diffusion probabilistic model for singing voice conversion," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 741–748. DOI: 10.1109/ASRU51503.2021.9688219.

[4] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, *Diffusion-based voice conversion with fast maximum likelihood sampling scheme*, 2022. arXiv: 2109.13821 [cs.SD].

[5] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and S. Seki, *Voicegrad: Non-parallel any-to-many voice conversion with annealed langevin dynamics*, 2024. arXiv: 2010.02977 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2010.02977.

[6] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, Lille, France: PMLR, Jul. 2015, pp. 2256–2265.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 6840–6851.

[8] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[9] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," *Interspeech*, pp. 5289–5293, 2017.

[10] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011. DOI: 10.1162/NECO_a_00142.

[11] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2021. arXiv: 2011.13456 [cs.LG].

[12] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982, ISSN: 0304-4149.

[13] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations* (Institute of Mathematical Statistics Textbooks). Cambridge University Press, 2019. DOI: 10.1017/9781108186735.

[14] X. Mao, "The truncated Euler–Maruyama method for stochastic differential equations," *Journal of Computational and Applied Mathematics*, vol. 290, pp. 370–384, 2015, ISSN: 0377-0427. DOI: https://doi.org/10.1016/j.cam.2015.06.002.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217. DOI: 10.1109/ICASSP.2010.5495701.

[16] J. Kong, J. Kim, and J. Bae, *Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis*, 2020. arXiv: 2010.05646 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2010.05646.

[17] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[18] M. Masanori, Y. Fumiya, and O. Kenji, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016. DOI: 10.1587/transinf.2015EDP7457.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[20] S.-w. Park, D.-y. Kim, and M.-c. Joe, "Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data," *arXiv*, 2020. eprint: 2005.03295 (eess.AS).

[21] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6. DOI: 10.1109/ICME.2016.7552917.

[22] W.-C. Huang, Y.-C. Wu, T. Hayashi, and T. Toda, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," *arXiv*, 2020. eprint: 2010.12231 (eess.AS).

[23] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-y. Lee, and L.-s. Lee, "FragmentVC: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," *arXiv*, 2021. eprint: 2010.14150 (eess.AS).

[24] J.-h. Lin, Y. Y. Lin, C.-M. Chien, and H.-y. Lee, "S2VC: A framework for any-to-any voice conversion with self-supervised pretrained representations," *arXiv*, 2021. eprint: 2104.02901 (eess.AS).

A Study on Variable Embedding Locations of Reversible Spectral Speech Watermarking

Xuping Huang[†*] and Akinori Ito[†]
[*] Interdisciplinary Faculty of Science and Engineering, Shimane University, Matsue 690-8504, Japan
E-mail: huang@cis.shimane-u.ac.jp
[†] Department of Communications Engineering, Graduate School of Engineering, Tohoku University, Sendai 980-8577, Japan
E-mail: akinori.ito.a2@tohoku.ac.jp

***Abstract*—Rapid advancements in machine learning and speech synthesis have made it feasible to manipulate and tempering audio content. Although watermarking is a useful technique for guaranteeing the integrity of evidential audio data, distortion of sound quality must be prevented. A reversible digital watermarking technique based on the human auditory system was conventionally proposed to embed feature data for tamper detection in the high frequency coefficients. However, this technique has the drawback that spectral analysis can reveal the embedding locations through the borderline. This research explored variable embedding locations while preserving good sound quality. The experimental results show the effectiveness of the proposed to conceal the embedding location, and the distortion are controlled to be imperceptible.**

I. Introduction

With the speed at which digital technologies are developing these days, it is becoming increasingly important to prevent tampering with digital materials. Tampering issues can be effectively resolved by using the information concealing technique, which allows us to insert the payload information for integrity check to the content data itself, including audio [1-7], image, and video data [8-19]. For the information concealment approach, high capacity and minimal distortion are often preferred. Furthermore, original data must be available for a number of uses, including military, medical, and investigative records, as permanent degradation is unacceptable [16]. For these reasons, reversible information hiding has been proposed for such purposes.

To achieve reduced distortion, the domain in which the payload is integrated is an important consideration. In the audio field, time domain linear predictive coding [3,4] and inaudible domain based on cochlear delay property [5] have been presented. In the imaging area, the integer discrete cosine transform (intDCT) is well-known for its ability to produce high capacity and low distortion for reversible information concealing [14-18]. Motivated by them, we looked on intDCT-based reversible information hiding.

Because picture and audio have different statistical features and perceptual qualities, the method of embedding payload into DCT coefficients should have been adjusted differently. When it comes to images, embedding the payload into higher DCT coefficients leads in poor resilience, while integrating it into lower DCT coefficients results in poor invisibility. Thus, the intermediate area of DCT coefficients has been expanded for embedding. Humans are less sensitive to differences in high frequency coefficients in audio than they are in images. As a result, in our previous work [7], we embedded the payload data into higher DCT coefficients.

However, due to the concentration of concealing locations, which is a half of high DCT regions in the previous work [7], the border line generated by the expansion of DCT coefficients would be visible on the spectrum, which could be a hint for attackers to estimate a watermarking method. To avoid hostile attacks, hiding places should be unobtrusive. As a result, a sophisticated algorithm with complex concealing places is necessary, as well as maintaining high-quality stego data.

Lee recommended that the watermark be adaptively implanted into the image [16] for improved stego data quality and complex concealing locations. The findings of image quality with the Peak Signal-to-Noise Ratio (PSNR) criterion are superior to the no-adaptation scenario. In this research, we study various embedding locations, including random addition of expansion locations, and the efficiency of adaptive embedding of payload to specified DCT coefficients via distortion estimation in order to keep good sound quality as well as conceal the embedding locations.

The overview of our previous work is summarized in Section II. The proposed algorithm is presented in Section III. The experimental evaluation on audio quality is described in Section IV, and the conclusion and future work are described in Section V.

II. Overview of previous work: intDCT-based reversible audio information hiding

In our previous work [7], a modified integer DCT coefficients expansion based reversible digital watermarking method is proposed and implemented for tampering detection. Suitable coefficients are explored according to the audio feature. As a general trend of integer DCT Type IV, the higher the frequency domain is, the lower the amplitude of the audio data will be. Thus, in our previous work, to control the distortion of stego data to be imperceptible, we expanded the DCT coefficients in the high frequency domain to obtain the hiding locations for payload embedding. In order to detect tampering precisely, a given original audio data is first divided into frames with a fixed sample-length N (typically, 512, 1024, etc.) for tampering localization. Let x_n $(1 \leq n \leq N)$ be an audio signal

within a frame. It is transformed from time domain to DCT domain with integer DCT by our previous work[7]. The integer discrete cosine transform (intDCT) is a transform similar to the discrete cosine transform (DCT). Let

$$\mathbf{h} = (h(1)\ h(2)\ \ldots\ h(N))^T \quad (1)$$

$$\mathbf{H} = (H(1)\ H(2)\ \ldots\ H(N))^T \quad (2)$$

be a time-domain signal at an N-point frame and its DCT coefficients, respectively. In a continuous case, we can obtain DCT coefficients $\mathbf{H}$ from a time-domain signal $\mathbf{h}$ by DCT-IV matrix as

$$\mathbf{H} = C_N^{IV}\mathbf{h} \quad (3)$$

where the (i,t)-th ($1 \le i \le N$, $1 \le t \le N$) elements of the DCT matrix C^{IV} are represented as

$$C_N^{IV}(i,t) = \sqrt{\frac{2}{N}}\left[\cos\left(\frac{(t+\frac{1}{2})(i+\frac{1}{2})\pi}{N}\right)\right] \quad (4)$$

In this paper, for DCT coefficients expansion based digital watermarking, Let X_n ($1 \le n \le N$) be DCT coefficients. When one bit of payload B is embedded to the n-th DCT coefficient X_n, the expansion technique is applied such as

$$X'_n = 2X_n + B. \quad (5)$$

After embedding, X'_n is transformed to time domain with the inverse integer DCT. The reversibility between x_n and X_n is guaranteed by the nature of the integer DCT. In our previous work [7], higher DCT coefficients are used because human is not so sensitive to the difference in high frequency coefficients. When payload is embedded into the upper half of DCT coefficients X_n ($N/2+1 \le n \le N$), the capacity about 0.5 bit per sample is achieved. However, since the DCT coefficients for expansion are concentrated on the high frequency domain, the borderline between the high frequency and the low frequency might be conspicuous to reveal the existence of the embedded payload by the spectral analysis, as Fig. 1 shown. We tried global steganalysis to check whether there is any signal embedded in another data. Figure 1 plots the a graphic about the original sound and stego signal of Ja_m5.wav, of some samples to compare how they change, where the horizontal axis represents the frame index and the vertical axis represents the order of DCT coefficients. It is obvious that when [9th, 16th] coefficients are expanded for embedding, by comparing the stego data and original data, the boundary for expansion and embedding might be visible.

III. Proposed Method: Variable Embedding Locations Exploration

On purpose to solve the boundary problem, and enhance the robustness towards to global steganalysis, we tried to explore variable embedding locations based on DCT coefficients expansion based digital watermarking. Our previous work [6] indicated that the random key of expansion may cause distortion to make the original data sounds noisy. Thus, distortion should be controlled to be imperceptible, since the original data is with probative importance, such as telephone recording used as evidence, and last will and testament, et al. Thus, we have to select appropriate locations for hiding the digest information for tampering detection, as well as considering the audio quality after expansion.

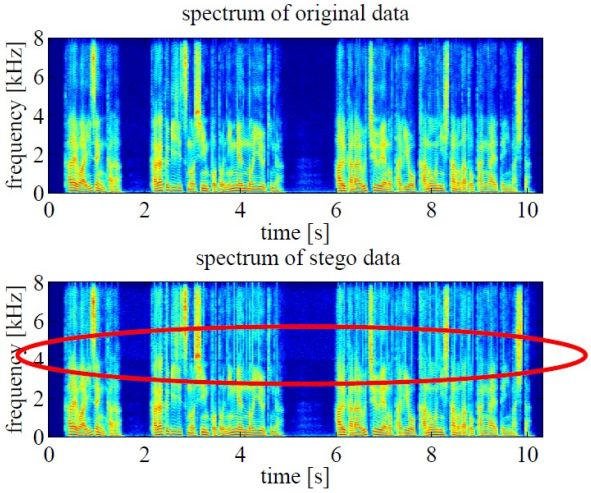


Fig. 1. Log of DCT coefficients difference between original data and stego data with a red mark on the borderline caused by expansion and embedding: Ja_m5.wav ($N = 2048$).

A. Solution 1: random addition of expansion blocks

Since the purpose is to keep the boundaries confidential, a random addition of expansion coefficients is supposed to be an efficient alternative solution. Figure 2 is an example of log amplitudes of DCT coefficients of the original speech data (Da_f2.wav, a female Germany speech data). The Y-axis is the DCT coefficients with a frame length N=2048. According to the feature of integer DCT type IV, amplitude tends to decrease as frequency increases. By utilizing this feature after transforming audio data from the time domain to the frequency domain, we could expand the amplitudes in the higher frequency domain, to reserve adequate hiding locations for payload to achieve imperceptibility. Our previous work [7] expanded the DCT coefficients from 1025-th to 2048-th coefficients to reverse embedding locations for embedding payload. In order to conceal the borderline, we hereby propose a solution to expand additional coefficients for hiding.

In order to extracted the data and reconstructed the original data, a location map is necessary here to mark the coefficients with expansion. We first segment N DCT coefficients to M blocks with an index i, ($1 \le i \le M$), indicating each block in order. The larger M is, the larger capacity for embedding the location map is required. Dividing N coefficients into M blocks can shorten the location map from N to M.

As the preliminary work, we tried to expand 512 coefficients, counting from the highest coefficients, out of 2048 coefficients in each frame, and we divided each 2048 coefficients into $M = 16$ blocks. This means 1/4 of coefficients (4 blocks) from the highest coefficients are expanded. With the

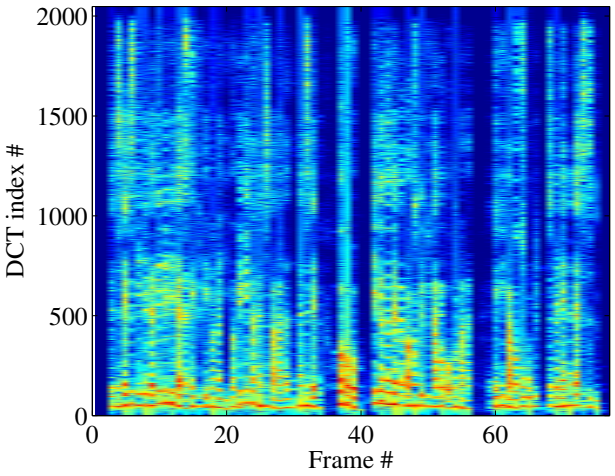


Fig. 2. The log amplitudes of DCT coefficients of the original speech data

motivation to conceal the boundary line, we hereby generated a random number with [4,5] as elements to decide each frame to expand 4 blocks or 5 blocks from the highest coefficients. We could also generated a random number with [8,9] for different capacity for hiding.

B. Solution 2: adaptive hiding with distortion estimation

The purpose of distortion estimation is to select the DCT coefficients with smaller distortion, and the policy to estimate distortion is discussed here. Stego DCT coefficients which are to be selected and expanded to hide payload are suppose to be: $X'_n = 2X_n + B, (1 \le n \le N)$, with a bit of embed data $B, B \in \{0,1\}$ in it. In order to find the proper DCT coefficients with smaller distortion in an effective way, we use a statistics way to estimate the possible distortion with payload embedded into each expanded DCT coefficients. Suppose each DCT coefficient X_n is a constant, then it is possible to estimate the distortion by calculate the error of mean square between stego data and original data. Theoretically, the distortion of stego data is supposed to be smaller if the hiding locations are adaptively selected by the error of mean square in ascending order. This work is an extension work of our previous work [6], and we explored variable embedding locations and compare the audio quality of each solution.

Let X'_n be the n-th DCT coefficient of stego data, X_n be the n-th DCT coefficient of original data, and E_n be the estimated distortion (error of mean square of difference between X'_n and $X_n, 1 \le n \le N$) after hiding a bit of data B in each DCT coefficient. Since $B \in \{0,1\}$, the probability of B is $P(B = 0) = 0.5$, and $P(B = 1) = 0.5$. Distortion estimation of each DCT coefficient can be represented by:

$$\begin{aligned} E\left[(X'_n - X_n)^2\right] &= X_n^2 + 2X_nB \sum_{B\in\{0,1\}} P(B) \\ &+ B^2 \sum_{B\in\{0,1\}} P(B^2) \\ &= X_n^2 + X_n + \frac{1}{2} \end{aligned} \quad (6)$$

According to formula (6), it is obvious that that the distortion of each coefficients has an up-and-down trend as the same as X_n value.

Length of each block is $\frac{N}{M}$, and the average distortion of DCT coefficients estimated in each block could be calculated by

$$\overline{E_i} = \frac{\sum_{n=\frac{(i-1)N}{M}+1}^{\frac{iN}{M}} E_n}{\frac{N}{M}}, \quad (1 \le i \le M, 1 \le n \le N) \quad (7)$$

Then the distortion estimated of each block $\overline{E_i}$ is sorted in ascending order to get the sorting result $ind, (1 \le ind \le M)$, indicating blocks number index i. We expand DCT coefficients in blocks where $\overline{E_i}$ is smaller, for example DCT coefficients blocks with $ind \le 8$.

The embedding steps could be summarized as follows:

Step 1 Divide original data into frames with length N and transform data $x_n, (1 \le n \le N)$ from time domain to DCT coefficients X_n using integer DCT IV algorithm;

Step 2 Calculate the estimated distortion E_n of difference between DCT coefficients of stego data and DCT coefficients of original data.

Step 3 Segment N DCT coefficients into M blocks, and calculate the average estimated distortion $\overline{E_i}$ and then sort them in ascending order to have the blocks index number i where $(1 \le ind \le M)$ to generate the location table T.

Step 4 Expand DCT coefficients X_n, where $(i-1)\frac{N}{M}+1 \le n \le i\frac{N}{M}$ and embed data to generate stego DCT coefficients of stego data $X'_n = 2X_n + B$ where $1 \le ind \le 8$.

Step 5 Hide table $T_i, (1 \le i \le M)$ into stego DCT coefficient X'_n where $N - M + 1 \le n \le N$

Step 6 Apply inverse integer DCT IV algorithm to X'_n to get stego data x'_n in time domain.

The extraction and reconstruction steps are summarized as follows:

Step 1 Transform stego data x'_n into DCT coefficients X'_n

Step 2 Calculate $\lfloor X'_n/2 \rfloor$ where $N - M + 1 \le n \le N$ and table information $T_i, (1 \le i \le M)$ equals to $X'_n - 2\lfloor X'_n/2 \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function.

Step 3 Read table T_i to get block index i where $T_i = 1$.

Step 4 Calculate $\lfloor X'_n/2 \rfloor$ where $\frac{N(i-1)}{M} + 1 \le n \le i\frac{N}{M}$, then original data after embedded data is removed equals to $\lfloor X'_n/2 \rfloor$ and embedded data equals to $X'_n - 2\lfloor X'_n/2 \rfloor$.

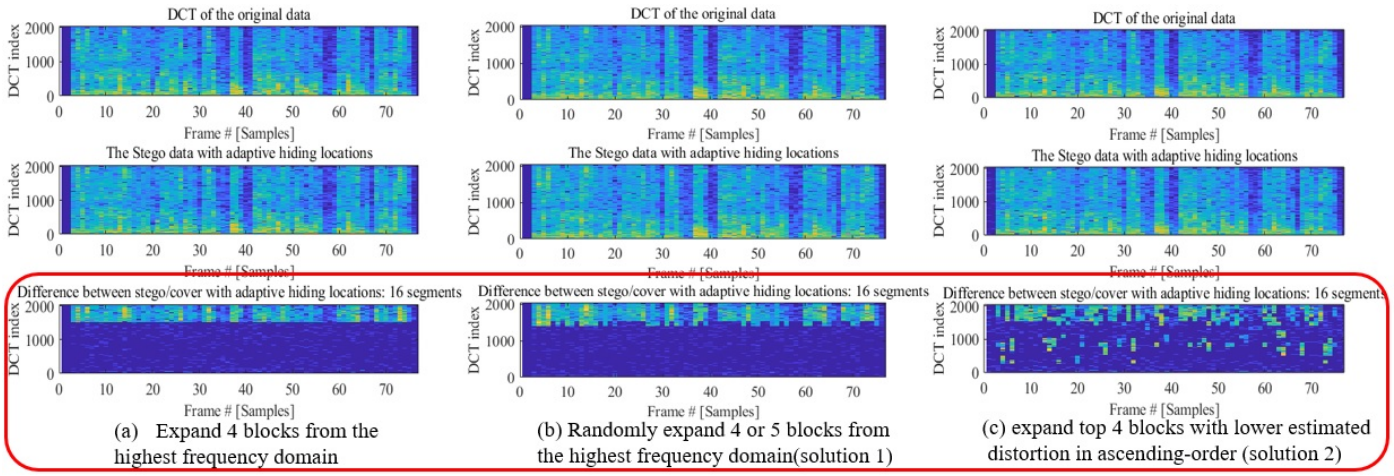


Fig. 3. Comparison on stego data generated of our previous work, randomly expansion (solution 1) and adaptive expansion (solution 2) with different expansion locations

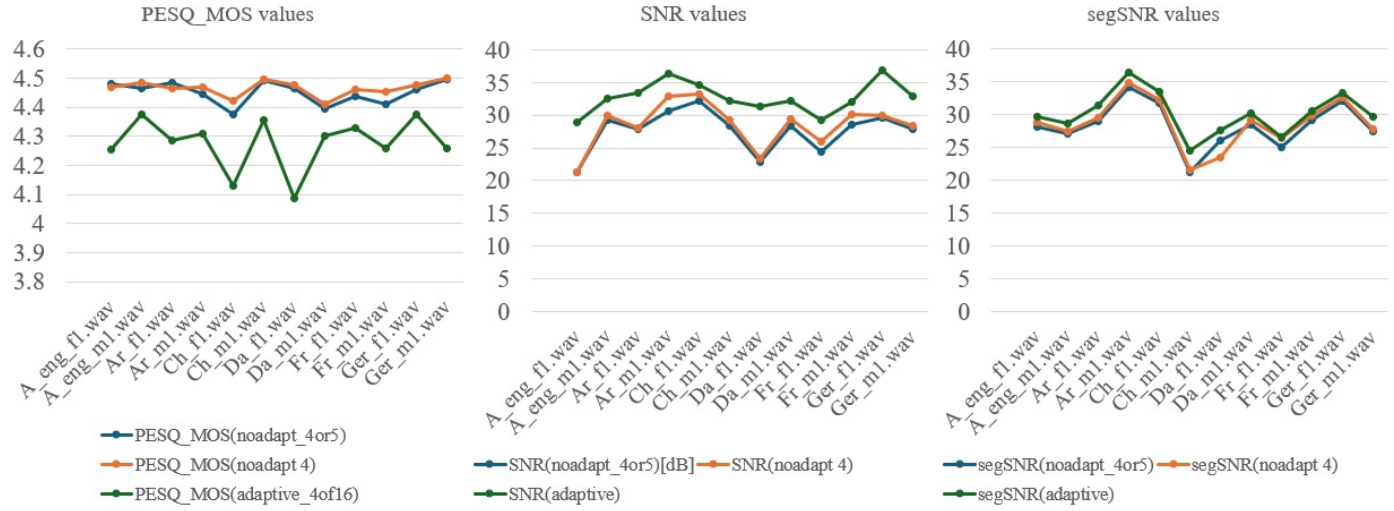


Fig. 4. Comparison on audio quality of our previous work, randomly expansion (solution 1) and adaptive expansion (solution 2) with different expansion locations

Step 5 Reconstruct original DCT coefficients $X_n, 1 \leq n \leq N$

Step 6 Transform X_n into time domain to get original data $x_n, 1 \leq n \leq N$.

IV. Experimental results and evaluation

A. Audio quality evaluation

When evaluating audio quality, we used segmental Signal-to-Noise Ratio (segSNR) and Perceptual Evaluation of Speech Quality (PESQ), which have been extensively used for evaluating the sound quality objectively in conventional works. In order to evaluate the listening quality of speech data, we used Mean Opinion Score Listening Quality Objective (MOS-LQO) score, which is defined by the ITU-T Recommendation P.862.1 [20]. We used AFsp packages version 9.0 to calculate SNR, segSNR scores and PESQ version 1.2 to calculate MOSLQO scores.

We evaluated our method using the dataset from ITU-T Test Signals for Telecommunication Systems–Test Vectors Associated to Rec. ITU-T P.50 Appendix I. We used 12 speech data (12 speakers by 6 languages: American English, Arabic, Mandarin Chinese, Danish, French, Germany, the first track of female and male speakers). The datasets are with a monophonic waveform with 16-kHz sampling and 16-bit quantization. The number of DCT coefficients in each frame is set to be 2048.

As the experimental results, we compared three alternative solutions for expansion: (1) expand 4 blocks from the highest coefficients (our previous work); (2) randomly expand 4 or 5 blocks from the highest coefficients (solution 1); and (3) expand top 4 blocks with lower estimated distortion in ascending-order (solution 2). Figure .3 plots the different expansion locations of (a) our previous work with no adaption, the proposed method of (b) solution 1 randomly 4 or 5 blocks

TABLE I
COMPARISON ON AUDIO QUALITY OF THE CONVENTIONAL WORK, PROPOSED SOLUTION 1, AND PROPOSED SOLUTION 2

Track	PESQ_MOS			SNR (dB)			segSNR (dB)		
	solution 1	conventional [7]	solution 2	solution 1	conventional [7]	solution 2	solution 1	conventional [7]	solution 2
A_eng_f1.wav	4.48	4.468	4.253	21.244	21.363	28.936	28.088	28.806	29.68
A_eng_m1.wav	4.466	4.483	4.374	29.208	29.926	32.568	27.151	27.497	28.733
Ar_f1.wav	4.484	4.466	4.285	27.806	27.991	33.447	28.934	29.465	31.404
Ar_m1.wav	4.447	4.468	4.307	30.699	32.847	36.289	34.201	34.845	36.346
Ch_f1.wav	4.374	4.423	4.131	32.16	33.173	34.622	31.792	32.305	33.461
Ch_m1.wav	4.494	4.495	4.354	28.353	29.272	32.171	21.153	21.639	24.43
Da_f1.wav	4.464	4.476	4.085	22.925	23.424	31.35	26.056	23.424	27.595
Da_m1.wav	4.393	4.412	4.302	28.423	29.362	32.15	28.516	29.226	30.283
Fr_f1.wav	4.438	4.462	4.327	24.476	26.005	29.222	24.957	26.36	26.572
Fr_m1.wav	4.411	4.454	4.26	28.6	30.094	32.005	29.11	29.869	30.474
Ger_f1.wav	4.461	4.475	4.377	29.532	29.925	36.903	32.048	32.581	33.402
Ger_m1.wav	4.496	4.499	4.258	27.886	28.39	32.963	27.36	27.866	29.686

expansion and (c) solution 2 adaptive expansion. By comparing Figure 3(a) and (b), the boundary line might be concealed, and Figure 3(c) has a complex key for expansion, and the location map for each frame is different, which enhance the complexity for global steganalysis attacks.

The audio quality of MOS values, SNR[dB] values, and segSNR values are plotted in Fig. 4. According to the results, the previous work and the proposed randomly expansion with 4 or 5 blocks (solutions) has considerable MOS values, and the adaptive expansion (solution 2) has generally lower MOSLQO values. However, all the data are better than 4.0, with the minimum MOSLQO value 4.085, which means all of the proposed method results an imperceptible distortion. For SNR values, the adaptive expansion (solution 2) has the best values, with is better than the conventional work with no adaption, and the randomly 4 or 5 blocks expansion (solution 1) have considerable SNR values with the conventional work with no adaption. For segSNR values, results of solution 2 is better than solution 1, and both of the proposed works have better segSNR values than the previous work. Since segSNR values indicate the differential in time domain, the difference between the stego data and the original data of the proposed methods are supposed to be lower. Table I lists up the audio quality results in details.

B. Discussion on borderline for expansion

The original purpose of the proposed method is to improve the vulnerability of detection on embedding locations. In our previous work [7], due to the concentration of hiding locations, the border line around 4 kHz is conspicuous as shown in Figure 1. To make the border line inconspicuous, we proposed solution 1 with random additional expansion blocks, and solution 2 with an adaptive expansion algorithm. The result to conceal the borderline of embedding locations is as shown in Figure 5. It is the result of an adaptive hiding with distortion sorting in ascending-order when block size $M = 64$, where the borderline is concealed as Figure 5 shows.

Though concealing the borderline is proposed as a countermeasure against spectral statistical analysis to hide the suspected watermarked data, it is still challenging to consider potential attacks techniques by global steganalysis. Global steganalysis employs various techniques to detect the presence of embedded payload by analyzing the statistical properties and structural characteristics of audio data without cover data. The frequency components, temporal patterns, and perceptual attributes can be indicative of watermarked payload by analyzing the stego data. Analyzing the robustness and resistance and evaluate the effectiveness of the proposed methods against global steganalysis is listed as a future challenge of this work.

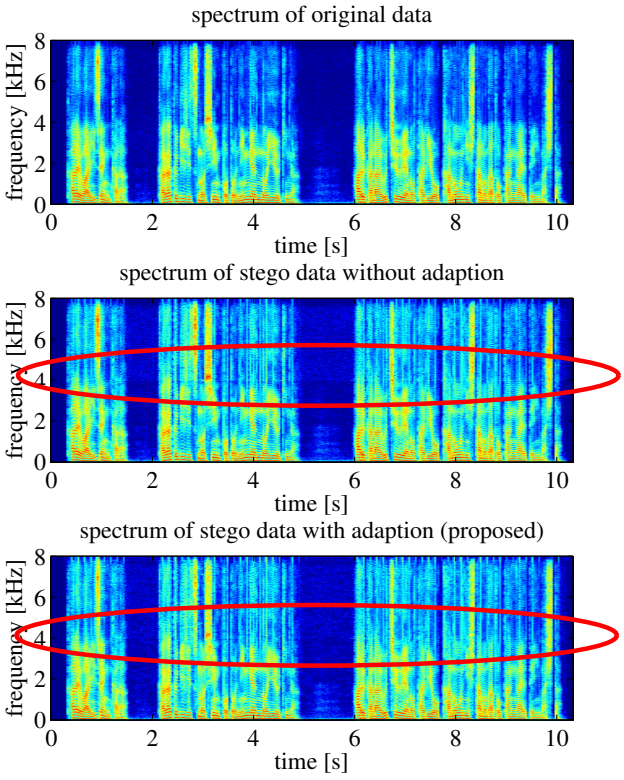


Fig. 5. Comparison of spectrum between stego data generated by method without adaption and with adaption (64 segments): Ja_m5.wav

V. Conclusions

To conceal the borderline of expansion for reserving embedding locations, we proposed solutions to explore the appropriate expansion locations. A random addition of expansion blocks solution, and an adaptive expansion by sorting the expected distortion are proposed. We evaluated the audio quality and the effectiveness to conceal the borderline, and the results show that comparably good quality of stego data as our previous work is achieved, besides, the hiding locations are inconspicuous to avoid malicious attack. Even the vulnerability of borderline is fixed, we still have the challenging to conceal the expansion locations against statistical spectral analysis by global steganalysis by objective evaluation as a future work.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP18K18052.

This work was supported by "Shimane University Grants for Joint Research Project led by Female Researchers" under the MEXT "Initiative for Realizing Diversity in the Research Environment (Collaboration Type).

References

[1] X. Zhang, et al.: Robust Reversible Audio Watermarking Scheme for Telemedicine and Privacy Protection. Computers, Materials & Continua 71.2 2022.

[2] M. Charfeddine, et al.: Audio watermarking for security and non-security applications, IEEE Access 10, pp: 12654-12677, 2022

[3] D.Q.Yan, and R.D.Wang.: Reversible Data Hiding for Audio Based on Prediction Error Expansion, Proc. of Intelligent Information Hiding and Multimedia Signal Processing, 249–252, 2008

[4] A. Nishimura, Reversible Audio Data Hiding Using Linear Prediction and Error Expansion, Proc. of Intelligent Information Hiding and Multimedia Signal Processing, 318-321, 2011

[5] Unoki, M.: Construction of auditory media signal processing infrastructure to prevent media clone attacks. Impact, vol. 2020(2), 21-23, 2020.

[6] X. Huang, et al.: Reversible Audio Information Hiding Based on Integer DCT Coefficients with Adaptive Hiding Locations. In: Shi, Y., Kim, HJ., Pérez-González, F. (eds) Digital-Forensics and Watermarking. IWDW 2013. Lecture Notes in Computer Science(), vol 8389. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-43886-2_27, 2014

[7] X. Huang, A. Ito.: Imperceptible and Reversible Acoustic Watermarking Based on Modified Integer Discrete Cosine Transform Coefficient Expansion. Applied Sciences. 2024; 14(7):2757. https://doi.org/10.3390/app14072757, 2024

[8] A. Anand, et al.: An improved DWT-SVD domain watermarking for medical information security. Computer Communications, vol.152, pp: 72-80, 2020

[9] Y. Yu, J. Gao, X. Mu, et al.: Adaptive LSB quantum image watermarking algorithm based on Haar wavelet transforms, Quantum Inf Process vol.22(180) , https://doi.org/10.1007/s11128-023-03926-1, 2023

[10] P. Garg, P.:A robust technique for biometric image authentication using invisible watermarking, Multimedia Tools and Applications, vol. 82(2), pp: 2237-2253, 2023

[11] Hernández-Joaquín, et al.: A secure DWT-based dual watermarking scheme for image authentication and copyright protection, Multimedia Tools and Applications, vol.82(27), pp: 42739-42761, 2023

[12] M. Roy, et al.: A perceptual hash based blind-watermarking scheme for image authentication, Expert Systems with Applications, vol. 227, https://doi.org/10.1016/j.eswa.2023.120237, 2023

[13] Sharma, Sunpreet, et al.: A review of image watermarking for identity protection and verification, Multimedia Tools and Applications, vol.83(11), pp.31829-31891, 2024

[14] H. R. Chennamma, et al.: A comprehensive survey on image authentication for tamper detection with localization, Multimedia Tools and Applications, vol.82(2), pp: 1873-1904, 2023

[15] D. Singh, et al.: An efficient self-embedding fragile watermarking scheme for image authentication with two chances for recovery capability, Multimedia Tools and Applications, vol. 82(1), pp: 1045-1066, 2023

[16] L. De, et al.:A reversible watermarking for image content authentication based on wavelet transform, Signal, Image and Video Processing, pp. 1-11, 2024

[17] G. Gao, M. Wang and B. Wu.:Efficient Robust Reversible Watermarking Based on ZMs and Integer Wavelet Transform, IEEE Transactions on Industrial Informatics, vol. 20(3), pp. 4115-4123, doi: 10.1109/TII.2023.3321101, 2024

[18] C. Zhan, et al.:Reversible Image Fragile Watermarking with Dual Tampering Detection, Electronics, vol.13(10), 1884. https://doi.org/10.3390/electronics13101884, 2024

[19] L. Tanwar, Lavi, et al.: Hybrid reversible watermarking algorithm using histogram shifting and pairwise prediction error expansion, Multimedia Tools and Applications, vol.83(8), pp: 22075-22097, 2024

[20] Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, ITU-T Recommendation P.862.1, International Telecommunication Union, 2001

Normalizing Flows-Based Latent Variable Rearrangement for Generative Image Steganography

Sifan Wu[1], Li Dong [1*], Teng Sio Hong[2], Jiale Chen[3], Diqun Yan[1] and Rangding Wang[1*]
[1] Ningbo University, Ningbo, China
E-mail: {2211100283, dongli, yandiqun, wangrangding}@nbu.edu.cn
[2] Faculty of Humanities and Social Sciences, Macao Polytechnic University, Macau, China
E-mail: p1308598@mpu.edu.mo
[3] Beijing Institute of Technology, Beijing, China
E-mail: 3120245681@bit.edu.cn

Abstract— **Image steganography is a technique for concealing secret messages within images. Traditional steganography methods typically involve modifying the cover image, making them vulnerable to detection and attacks. Recently, generative image steganography, which embeds messages during image generation, has emerged as a more robust solution to this challenge. However, these methods often face a trade-off between the visual quality of the generated images and the effectiveness of message concealment. Current approaches are frequently criticized for their low steganographic payloads and poor extraction accuracy. To address these issues, we propose a generative steganography method based on Normalizing Flows. Our approach rearranges the values of Gaussian latent variables, used for image generation, according to the secret messages and predefined mapping rules. The secret messages are embedded in the adjusted Gaussian latent variables, while the image generation process remains unaffected, as the latent variable values and their distribution are preserved. Experimental results demonstrate the superior performance of our method in terms of imperceptibility, steganographic capacity, and extraction accuracy.**

I. INTRODUCTION

Image steganography is a technique of hiding the secret message into the image without being detected [1]. Currently, the main challenge of steganography is to maintain a trade-off in steganographic capacity, imperceptibility, and robustness [2].

Traditional steganography relies on the cover image to embed messages. Ideally, the message is embedded by minimizing distortion to the cover image. Initially, Least-Significant Bit (LSB) algorithm embeds a secret message by modifying the least significant bit of the image [3]. This method is simple and useful, but it is easy to be detected and attacked. Subsequently, adaptive steganography methods were proposed. Adaptive steganography based on Syndrome-Trellis Codes (STC) improves the security of steganography by embedding messages in a statistically secure manner [4]. Messages are not only embedded in the spatial domain of cover images but also embedded in the frequency domain [5]. With the development of deep learning, steganography methods based on deep neural networks have developed. These methods achieved high steganographic payload and undetectable [6]. However, the steganography method with cover images usually selects an existing image as the cover image. Then, the image is distorted to embed the secret message. However, the image distortion makes them susceptible to detection and attack.

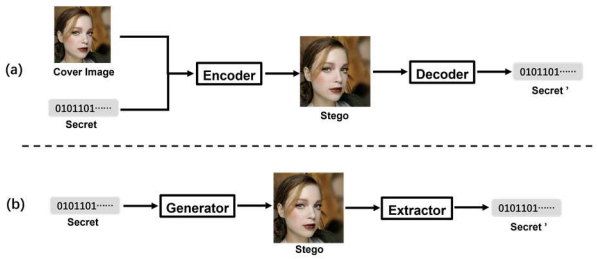


Fig. 1 Existing steganography frameworks: (a) Steganography with cover image. (b) Generative image steganography.

To address this problem, generative image steganography conceals messages during image generation and is difficult to detect. The workflows of the traditional and generative steganography methods are shown in Fig. 1. In comparison with traditional image steganography, the distribution of secret messages is mapped to the distribution of the latent variable used to generate the image. These methods not only improve the steganographic payload but also enhance the anti-detection ability and the visual quality of the generated images.

Generative Adversarial Networks (GANs), as a generative model, are widely used in generative steganography. The generator is trained to generate stego images that contain hidden messages, while a discriminator ensures that the generated stego images are imperceptibility from the images generated by latent variables without any mapping for the secret messages [7]. In addition, Normalizing Flows (NF) is also used as a generative model in the generative steganography. These methods embed secret messages into images using an invertible network

that allows for both embedding and extraction of the hidden messages [8]. They perform lossless mapping between the secret messages and the stego image. These methods retain high exact accuracy and are undetectable.

In this paper, a new generative steganography method based on NF is proposed. The performance and effectiveness of the proposed method are verified through experiments. The main contributions of this paper are as follows:

1). We propose a coverless steganography scheme based on Real-valued non-volume preserving (Real NVP).

2). A novel mapping method from secret information to latent variables is introduced.

3). Experimental results demonstrate that our scheme achieves higher undetectability compared to existing methods.

The rest of this paper is as follows: the second part introduces the related work. The third part describes the details of the proposed generative steganography framework. The fourth part shows the experimental results and analysis. The last part summarizes the research findings and discusses future research directions.

II. Related Work

A. Normalizing Flows

NF is a generative model that can map simple distributions into complex distributions through a series of invertible functions [9] [10] [11] [12] [13] [14]. NF was first proposed as a way to enhance the flexibility of variational inference [9]. Real NVP transformations are implemented using a series of coupled affine layers [12]. Real NVP offers efficient forward and inverse computations, making it suitable for tasks requiring reversible mappings, such as density estimation and generative modeling, with enhanced scalability and flexibility.

The fundamental of NF is the change of variable theorem. A series of neural network blocks f_i are connected sequentially to constitute the flows. The latent variable $\mathbf{z}_0$ is transformed to the latent variable $\mathbf{z}_k$ by the flows. The probability density function of the latent variable $\mathbf{z}_k$ can be modeled by using the change of variables theorem:

$$p_k(\boldsymbol{z}_k) = p_0(\boldsymbol{z}_0)\left|det\left(J_{f^{-1}}\right)\right|, \tag{1}$$

where $J_{f^{-1}}$ is the Jacobian matrix of the inverse transformation. The determinant $det(\cdot)$ quantifies the volume change caused by the transformation, ensuring that the probability density adjusts appropriately to reflect this change.

The goal of the normalized flows model is to generate images $\boldsymbol{x_i}$ that obey the distribution p_{data} of a given image dataset. The model as an image generator $\mathcal{G}$ can be expressed as:

$$\mathcal{G}^* = \underset{\mathcal{G}}{argmax} \sum_{i=1} log p_{\mathcal{G}}(\boldsymbol{x}_i) \tag{2}$$

NF model is trained to fit the log-likelihood function of the probability distribution of the dataset. Combine (1) and (2):

$$log\ p_{\mathcal{G}}(\boldsymbol{x}_i) = log\ p_z\left(\mathcal{G}^{-1}(\boldsymbol{x}_i)\right) + log\left|det\left(J_{\mathcal{G}^{-1}}\right)\right| \tag{3}$$

B. Generative Steganography Based on Normalizing Flows

NF is a powerful generative model, and the network is completely invertible. It is suitable for the task of image steganography.

NF can be used for steganography with cover images. It can implement high-capacity steganography which hides images in images [8]. Robust steganography can be implemented by adding some auxiliary network modules in the network framework [15]. Meanwhile, the Normalizing flows model can also be used to implement generative steganography. Some methods embed messages into the Gaussian latent variables which are used to generate the images. For example, the Gaussian variable values are encoded as float numbers, and the messages are hidden in the lower bits [16]. Gaussian variables are artificially constructed with the guidance of secret messages instead of random sampling [17]. Using conditionally invertible neural networks to guide image coloring, and the messages are hidden in the colors [18].

III. Proposed Method

We propose a novel mapping scheme from secret messages to latent variables, utilizing a latent variable rearrangement method. First, latent variables are sampled, and then they are rearranged based on the secret messages. This method ensures that the data distribution of the latent variables remains unchanged, thereby preserving high image generation quality. The overall framework details are outlined as follows.

A. Generating the Image with Embedded Message

The generating process is shown at the top of Fig. 2. It is organized into two steps: secret message embedding and stego image generation.

A binary secret message of length L bits is $\mathbf{S} \in \{0,1\}^L$. After preprocessing, encoding and encryption operations, the secret message $\mathbf{S}$ is converted to $\mathbf{x_s}$ of size $H \times W \times C$, and it obeys uniform distribution:

$$\mathbf{x_s} = \mathcal{F}(\mathbf{S}) \tag{4}$$

A Gaussian variable $\mathbf{z_0}$ of size $H \times W \times C$ randomly sample from the prior distribution $\mathbf{z} \sim N\left(\boldsymbol{\mu_0}, \boldsymbol{\sigma_0^2}\right)$. Subsequently, the values of $\mathbf{z_0}$ are rearranged by the mapping module $\mathcal{M}$ under the guidance of $\mathbf{x_s}$. The latent variable $\mathbf{z_s}$ which contains a secret message is obtained:

$$\mathbf{z_s} = \mathcal{M}(\mathbf{x_s}, \mathbf{z_0}) \tag{5}$$

The $\mathbf{z_s}$ is input to the image generative network $\mathcal{G}$ based on NF, and we can generate the stego image $\mathbf{y_{stego}}$ of size $H \times W \times C$:

$$\mathbf{y_{stego}} = \mathcal{G}(\mathbf{z_s}) \tag{6}$$

B. Extracting Process

The process of extracting the secret message is shown at the bottom of Fig. 2. It is the inverse process of the generating process, including Gaussian latent variable recovery and secret message inverse mapping.

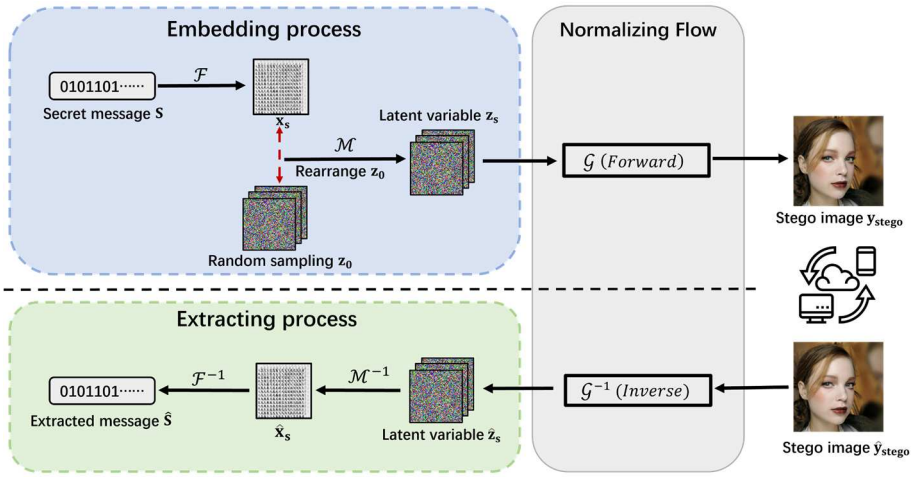


Fig. 2 The overview of the proposed method.

The stego image $\hat{\mathbf{y}}_{\mathbf{stego}}$ is received after transmitting. We input $\hat{\mathbf{y}}_{\mathbf{stego}}$ into the NF and inverse $\mathcal{G}^{-1}$. The recovered Gaussian variable $\hat{\mathbf{z}}_{\mathbf{s}}$ is obtained:

$$\hat{\mathbf{z}}_{\mathbf{s}} = \mathcal{G}^{-1}(\hat{\mathbf{y}}_{\mathbf{stego}}) \tag{7}$$

According to the mapping rule of the mapping module $\mathcal{M}$, we inverse map $\hat{\mathbf{x}}_{\mathbf{s}}$ with $\hat{\mathbf{z}}_{\mathbf{s}}$:

$$\hat{\mathbf{x}}_{\mathbf{s}} = \mathcal{M}^{-1}(\hat{\mathbf{z}}_{\mathbf{s}}) \tag{8}$$

Finally, the secret message is extracted by $\hat{\mathbf{x}}_{\mathbf{s}}$:

$$\hat{\mathbf{S}} = \mathcal{F}^{-1}(\hat{\mathbf{x}}_{\mathbf{s}}) \tag{9}$$

The whole processes of the embedding and extraction are organized in Algorithm 1 and 2.

Algorithm 1 Embedding process

Input: Secret message $\mathbf{S}$, Random Gaussian variable $\mathbf{z_0}$, Encoding operations $\mathcal{F}$, Mapping module $\mathcal{M}$, Normalizing Flows $\mathcal{G}$.

Output: Stego image $\mathbf{y_{stego}}$.

1: Encoding operations, $\mathbf{x_s} = \mathcal{F}(\mathbf{S})$.

2: Embedding secret message, $\mathbf{z_s} = \mathcal{M}(\mathbf{x_s}, \mathbf{z_0})$.

3: Generate the stego image, $\mathbf{y_{stego}} = \mathcal{G}(\mathbf{z_s})$.

4: Return $\mathbf{y_{stego}}$.

Algorithm 2 Extracting process

Input: Received stego image $\hat{\mathbf{y}}_{\mathbf{stego}}$, Decoding operations $\mathcal{F}^{-1}$, Inverse mapping module $\mathcal{M}^{-1}$, Inverse Normalizing Flows $\mathcal{G}^{-1}$.

Output: Extracted secret message $\hat{\mathbf{S}}$.

1: Normalizing Flows inverse, $\hat{\mathbf{z}}_{\mathbf{s}} = \mathcal{G}^{-1}(\hat{\mathbf{y}}_{\mathbf{stego}})$.

2: Inverse mapping, $\hat{\mathbf{x}}_{\mathbf{s}} = \mathcal{M}^{-1}(\hat{\mathbf{z}}_{\mathbf{s}})$.

3: Decoding operations, $\hat{\mathbf{S}} = \mathcal{F}^{-1}(\hat{\mathbf{x}}_{\mathbf{s}})$.

4: Return $\hat{\mathbf{S}}$.

C. *Message Mapping Module*

In the embedding process, the secret message $\mathbf{S}$ needs to be pre-processed firstly. Depending on the different steganographic payloads, it is encoded and resized to $\mathbf{x_s}$ of size $H \times W \times C$. When the payload is less than 3 bit/pixel, we encode $\mathbf{S}$ redundantly using BCH codes, repetition or neural network modules. When payload is more than 3 bit/pixel, state compression coding is required. The n-bits binary numbers are encoded into an integer of $[0, 2^n - 1]$. However, in practical scenarios, $\mathbf{S}$ is not ideally uniformly distributed after encoding. Hence, the $\mathbf{x_s}$ needs to be encrypted before the embedding, such as ChaCha20 [19].

The main idea for the mapping module $\mathcal{M}$: in order to embed the secret message and ensure that $\mathbf{z_s}$ and $\mathbf{z_0}$ have the same data distribution, the value of the Gaussian variable does not change. But the values of the Gaussian variables can be rearranged. Thus, the secret message cloud be embedded into the rearranged Gaussian variable $\mathbf{z_s}$. Details of how to realize the rearranged mapping are inspired by the different encoding methods of digital signals transmitted in physical layer, and we have two different mapping schemes:

1) Real-Valued Rearrangement Mapping (RVRM)

Referring to Non-Return-Zero Code (NRZ) code in signal transmission: a high-level signal represents 1 and a low-level signal represents 0. We do a similar mapping between $\mathbf{x_s}$ and $\mathbf{z_0}$. To preserve the distribution unchanged, the values of $\mathbf{z_0}$ should not be modified. Based on the integer values $[0, n]$ of $\mathbf{z_0}$, we divide $\mathbf{z_0}$ into n parts according to the probability density distribution. The values of $\mathbf{x_s}$ are mapped to correspond to the values of the n regions of $\mathbf{z_0}$. Thus, the values of $\mathbf{z_0}$ would be rearranged to obtain $\mathbf{z_s}$ according to the guidance of $\mathbf{x_s}$. In this way, the information of $\mathbf{x_s}$ is embedded into the Gaussian variable $\mathbf{z_s}$, and $\mathbf{z_s}$ have the same data distribution of $\mathbf{z_0}$.

However, the randomly sampled Gaussian variable is not a standard normal distribution. Dividing $\mathbf{z_0}$ into n parts based

on the probability density distribution will cause problems. The number of $\mathbf{z_0}$ values in the corresponding region does not match the number of integer values in the corresponding $\mathbf{x_s}$. Therefore, the values of $\mathbf{z_0}$ need to be slightly modified to match the number of values in the corresponding region.

Additionally, instead of dividing regions based on the probability density distribution, you can first sort the values in $\mathbf{z_0}$. Then we find the threshold value for each region based on the corresponding element amount in each region.

2) Difference Rearrangement Mapping (DRM)

The DRM is simple and accessible. However, it is vulnerable to rounding and precision loss. Thus, with reference to the Non-Return-Zero-Inverted Code (NRZI), the level signal flips to represent 0, and the level signal remains unchanged to represent 1. We similarly can map the values according to the difference between the i-th and $(i+1)$-th element values of $\mathbf{z_0}$. If the difference of adjacent element is more than the threshold α, it represents 1. Otherwise, it is less than the threshold β, it represents 0. Thus, we rearrange the values of $\mathbf{z_0}$ to get $\mathbf{z_s}$ with the guidance of $\mathbf{x_s}$. In this case, the $\mathbf{x_s}$ is embedded based on the difference of adjacent elements of $\mathbf{z_s}$.

At the same time, this scheme based on the change in the differences between adjacent elements can be extended. For example, we can map the values based on the differences between the elements at the same position but in different channels. It is even possible to map through the differences between the elements at the same position of the different variables.

In summary, two different schemes do not affect the quality of the generated images. It mainly affects the capacity of steganography and extraction accuracy. The real-valued rearrangement scheme is easier to implement mapping and has better steganography capacity. The difference rearrangement scheme has better extraction accuracy. We choose the suitable scheme according to different scenarios.

D. Image Generation Module

The image generation model $\mathcal{G}$ is based on Normalizing Flows. The model $\mathcal{G}$ is trained to use a randomly sampled Gaussian variable to generate a high-quality image. The generative image obeys the data distribution p_{data}.

When training the NF, the image dataset is used as the input of the model $\mathcal{G}^{-1}$. The model maximizes the likelihood of the prior distribution. The loss function of the model is:

$$\mathcal{L} = \log\ p_{\mathbf{z_0}}\left(\mathcal{G}^{-1}(\mathbf{y_{data}})\right) + \log\ |\det(J_{\mathcal{G}^{-1}})| \tag{10}$$

During the experiments, we use the pre-trained Real NVP model as the generative model $\mathcal{G}$ [18].

IV. Experiments

To evaluate the performance of the proposed generative steganography method, this section includes details of the experimental setup, performance evaluation, and comparison with baseline methods.

A. Implementation Details

All the experiments are conducted on an NVIDIA RTX 2080Ti GPU platform using Pytorch with Python.

1) Training Details

In our experiments, the generative model is Real NVP. The model trained on the CelebA dataset [20] to generate 64×64 images with 200 epochs. The batch size is set to 64. Features in residual blocks of the first few layers are set to 32. Number of residual blocks is set to 2. The model uses bottleneck, skip architecture, weight normalization, and affine coupling. The model does not use batch-norm coupling layer output. Adam is the optimizer, with the initial learning rate of 1e-3.

2) Evaluation metrics

In the experiments, we adopt the following evaluation metrics to evaluate the steganographic capacity, extraction accuracy, and anti-detectability.

To evaluate the visual quality of the generative image, we use no-reference image quality assessment, including FID [21] and Brisque score [22].

The steganographic capacity is measured by BPP, which represents the bits of information embedded in per pixel (bit/pixel) in the stego image:

$$\mathrm{BPP} = \frac{len(\mathrm{S})}{H \times W \times 3} \tag{11}$$

Where $len(\cdot)$ is the length of the hidden secret message, and H and W are the height and width of the steganographic image.

ACC is used to measure the extraction accuracy:

$$\mathrm{ACC} = 1 - \frac{||\hat{\mathrm{S}} - \mathrm{S}||_1}{\mathrm{len(S)}} \tag{12}$$

B. Comparison to Baselines

We evaluate the performances of the proposed methods in the aspects of extraction accuracy, anti-detectability, and imperceptibility under different hiding payloads, and compare with the following baselines: UDH [23], HiNet [24], SWE [25], and S2IRT [17].

1) Extraction accuracy

The information extraction accuracy is an important metric to measure the performance of steganography methods. TABLE I shows the information extraction accuracy at different hiding payloads.

The proposed steganography method with real-valued rearrangement mapping scheme (RVRM): when the hiding payloads are in the range of 1-2 BPP, the secret message is preprocessed and encoded.

The extraction accuracy can be achieved 100% by redundant coding and error correction code like BCH. At 3 BPP, the extraction accuracy can achieve 98%. As the hiding payloads continue to increase, a compression coding process is required during pre-processing. However, the extraction accuracy will be

reduced. If the recovery accuracy of the secret message is not required to be high, the steganography capacity can achieve 6-9 BPP.

TABLE I

The Extraction Accuracy with Different Hiding Payloads

Methods	0.25bpp	1 bpp	2 bpp	3 bpp
UDH	**100%**	98.21%	-	-
HiNet	-	92.94%	76.85%	74.31%
SWE	73.01%	71.34%	71.20%	71.22%
S2IRT	**100%**	**100%**	**100%**	99.43%
RVRM	**100%**	**100%**	**100%**	98.23%
DRM	**100%**	**100%**	**100%**	**99.98%**

The proposed steganography method with DRM has better anti-perturbation performance through difference mapping. It can effectively resistance to rounding and precision loss, and the extraction accuracy will be higher. Compared with RVRM, this method is more robust, but the steganography payload will be reduced.

2) Visual quality

To evaluate the quality of stego images, we use no-reference image quality evaluation strategies with FID [21] and Brisque score [22]. Our method generates 1,000 clean images by using randomly sampled latent variables and the corresponding stego images. The image quality scores are calculated in TABLE II. The lower scores indicate better performance. The visual quality scores of the stego images are compared with those of some existing steganography methods. As we can see from the results, the stego images generated by the proposed method have good visual quality. The Fig. 3 has some examples of randomly generated images.

The quality of the generated stego images by the proposed steganography method is mainly determined by the generation ability of the NF. The secret message embedding process only changes the order of the Gaussian variable values, which ensures that the Gaussian variables embedded with the secret message will not change the original data distribution.

TABLE II

The Scores of Stego Images with No-reference Image Quality Evaluation

Methods	FID↓	BRISQUE↓
SWE	-	9.07
S2IRT	-	**7.69**
Clean	108.35	19.56
Ours	**115.48**	17.45

We can see that the visual quality scores of the stego images are almost no different from the clean images. Therefore, the quality of the generated images will not be affected by the embedded information and the mapping schemes. The model used in our experiments is Real NVP. It is also feasible to change the generative model to the more advanced Normalizing flows model. And the generated images will have better visual quality.

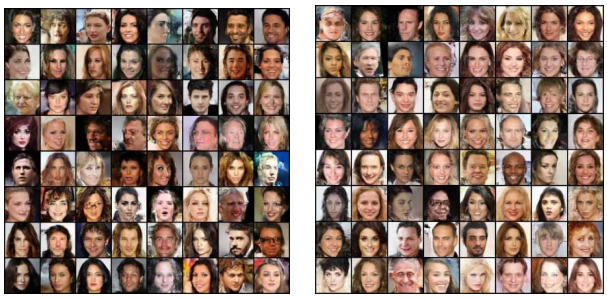


Fig. 3 The generated clean images (left) and stego images (right).

3) Steganalysis

The security is important to steganography algorithms. In order to evaluate the anti-detection ability of the proposed steganography method, we use the steganalysis methods Xu-Ne [26], SR-Net [27] to detect the stego images generated by our method. The dataset includes the clean images and corresponding stego images. TABLE III shows the statistical results of steganalysis and the comparison with existing methods. The detection accuracy is optimal when it approaches 50%. The statistics show that the proposed steganography method has higher security, and the detection accuracy is close to 65%. The stego image generation process of our proposed method is almost imperceptibility from the random generation process of clean images.

TABLE III

The Detection Accuracy (%) with Different Steganalysis Methods

Methods	Xu-Net	SR-Net
UDH	100%	100%
HiNet	100%	100%
SWE	73.53%	72.41%
S2IRT	74.90%	73.05%
Ours	**65.41%**	**61.11%**

In summary, the generated stego images using the proposed method are almost indistinguishable from clean images. The generated stego images can be considered the process of random image generation.

V. CONCLUSIONS

We propose a generative steganography method based on Normalizing flows. There are two mapping schemes to embed secret messages into random latent variables. The embedding and extraction will not change the data distribution of the stego images. Then the security of steganography can be guaranteed. The experiment results show that the quality of the stego images is determined by the NF, and it is not influenced by the information embedding. The generative stego images are almost imperceptibly from the clean images. In terms of extraction accuracy and payloads, the method in this paper has a good

comprehensive performance. In the future, we will continue to work on coverless steganography with the more advanced generative model and improve the robustness of this method.

VI. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 62171244, 61901237), Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F020011), Ningbo Natural Science Foundation- Young Doctoral Innovation Research Project (Grant No. 2022J080) (Corresponding author: Li Dong and Rangding Wang).

REFERENCES

[1] Filler T, Judas J, Fridrich J. Minimizing additive distortion in steganography using syndrome-trellis codes[J]. IEEE Transactions on Information Forensics and Security, 2011, 6(3): 920-935.

[2] Zhang C, Lin C, Benz P, et al. A brief survey on deep learning based data hiding[J]. arXiv preprint arXiv:2103.01607, 2021.

[3] Anderson R J, Petitcolas F A P. On the limits of steganography[J]. IEEE Journal on selected areas in communications, 1998, 16(4): 474-481.

[4] Filler T, Judas J, Fridrich J. Minimizing additive distortion in steganography using syndrome-trellis codes[J]. IEEE Transactions on Information Forensics and Security, 2011, 6(3): 920-935.

[5] Lian S. Multimedia content encryption: techniques and applications[M]. Auerbach Publications, 2008.

[6] Baluja S. Hiding images in plain sight: Deep steganography[J]. Advances in neural information processing systems, 2017, 30.

[7] Volkhonskiy D, Nazarov I, Burnaev E. Steganographic generative adversarial networks[C]//Twelfth international conference on machine vision (ICMV 2019). SPIE, 2020, 11433: 991-1005.

[8] Jing J, Deng X, Xu M, et al. Hinet: Deep image hiding by invertible network[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 4733-4742.

[9] Rezende D, Mohamed S. Variational inference with normalizing flows[C]//International conference on machine learning. PMLR, 2015: 1530-1538.

[10] Kingma D P, Salimans T, Jozefowicz R, et al. Improved variational inference with inverse autoregressive flow[J]. Advances in neural information processing systems, 2016, 29.

[11] Papamakarios G, Pavlakou T, Murray I. Masked autoregressive flow for density estimation[J]. Advances in neural information processing systems, 2017, 30.

[12] Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real nvp[J]. arXiv preprint arXiv:1605.08803, 2016.

[13] Kingma D P, Dhariwal P. Glow: Generative flow with invertible 1x1 convolutions[J]. Advances in neural information processing systems, 2018, 31.

[14] Grathwohl W, Chen R T Q, Bettencourt J, et al. Ffjord: Free-form continuous dynamics for scalable reversible generative models[J]. arXiv preprint arXiv:1810.01367, 2018.

[15] Lan Y, Shang F, Yang J, et al. Robust image steganography: hiding messages in frequency coefficients[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(12): 14955-14963.

[16] Wei P, Luo G, Song Q, et al. Generative steganographic flow[C]//2022 IEEE international conference on multimedia and expo (ICME). IEEE, 2022: 1-6.

[17] Zhou Z, Su Y, Li J, et al. Secret-to-image reversible transformation for generative steganography[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 20(5): 4118-4134.

[18] Ren Y, Liu T, Zhai L, et al. Hiding data in colors: Secure and lossless deep image steganography via conditional invertible neural networks[J]. arXiv preprint arXiv:2201.07444, 2022.

[19] Bernstein D J. ChaCha, a variant of Salsa20[C]//Workshop record of SASC. 2008, 8(1): 3-5.

[20] Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3730-3738.

[21] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.

[22] Mittal A, Moorthy A K, Bovik A C. No-reference image quality assessment in the spatial domain[J]. IEEE Transactions on image processing, 2012, 21(12): 4695-4708.

[23] Zhang C, Benz P, Karjauv A, et al. Udh: Universal deep hiding for steganography, watermarking, and light field messaging[J]. Advances in Neural Information Processing Systems, 2020, 33: 10223-10234.

[24] Jing J, Deng X, Xu M, et al. Hinet: Deep image hiding by invertible network[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 4733-4742.

[25] Hu D, Wang L, Jiang W, et al. A novel image steganography method via deep convolutional generative adversarial networks[J]. IEEE access, 2018, 6: 38303-38314.

[26] Xu G, Wu H Z, Shi Y Q. Structural design of convolutional neural networks for steganalysis[J]. IEEE Signal Processing Letters, 2016, 23(5): 708-712.

[27] Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2018, 14(5): 1181-1193.

Empower Typed Descriptions by Large Language Models for Speech Emotion Recognition

Haibin Wu*[1], Huang-Cheng Chou†[1], Kai-Wei Chang*[2], Lucas Goncalves‡[2], Jiawei Du*,
Jyh-Shing Roger Jang*, Chi-Chun Lee†[3], and Hung-yi Lee*[3]
* National Taiwan University, Taiwan
E-mail: f07921092@ntu.edu.tw, kaiwei.chang.tw@gmail.com, r11922185@ntu.edu.tw, jang@mirlab.org, hungyilee@ntu.edu.tw
† National Tsing Hua University, Taiwan
E-mail: huangchengchou@gmail.com, cclee@ee.nthu.edu.tw
‡ The University of Texas at Dallas, USA
E-mail: goncalves@utdallas.edu
[1]equal first author, [2]equal second author, [3]equal corresponding author, order is random

***Abstract*—Training speech emotion recognition (SER) requires human-annotated labels and speech data. However, emotion perception is complex. The pre-defined emotion categories are not enough for annotators to describe their emotion perception. Devoted annotators will use natural language rather than traditional emotion labels when annotating data, resulting in typed descriptions (e.g., "Slightly Angry, calm" to notify the intensity of emotion). While these descriptions are highly valuable, SER models, designed as classification models, cannot process natural languages and thus discard them. To leverage the valuable typed descriptions, we propose a novel way to prompt ChatGPT to mimic annotators, comprehend natural language typed descriptions, and subsequently adjust the given label of the input data. By utilizing labels generated by ChatGPT, we consistently achieve an average relative gain of 3.08% across all settings using 15 speech self-surprised learning models on the SUPERB, which provides a potential way to integrate the power of LLMs to improve the performances of SER.**

I. Introduction

Speech Emotion Recognition (SER) aims to discern emotional cues from speech inputs, representing a pivotal technology for human-computer interaction (HCI) systems. Over the years, significant advancements have been made in SER, making it a promising field of research and application [1]. One of the critical factors in SER is the annotation of ground truth labels used to train SER models. Typically, emotional corpora are annotated through perceptual evaluations. In these evaluations, annotators provide their ratings by completing questionnaires with a fixed set of options after listening to or watching a stimulus (speech or video). However, this conventional approach has its limitations. The fixed number of options can lead to bias in the labels [2], as annotators can only select the best available option even when it doesn't precisely describe the emotional content of the stimulus. This constraint can result in labels that do not fully capture the nuances of genuine human emotions.

To overcome this limitation, modern emotional datasets [3]–[5] have adopted an alternative approach by allowing annotators to select "other" and provide their descriptions of the emotions in their own words (natural language), resulting in **typed descriptions** (some examples are illustrated in Figure 2). This method can attenuate the problems caused by forced-choice paradigms and give a richer, more accurate representation of the emotional content. Typed descriptions, such as "slightly hopeful," can offer valuable insights that go beyond the pre-defined categories. Despite their potential, these typed descriptors are often discarded by most prior studies due to the limitations of current SER systems, which are designed as classification models that cannot process natural language inputs. As the typed descriptions are highly valuable, the inclusion and processing of these typed descriptions could significantly enhance the performance and accuracy of SER systems. Unfortunately, existing SER models ignore this data due to their inability to handle natural language effectively [6], [7].

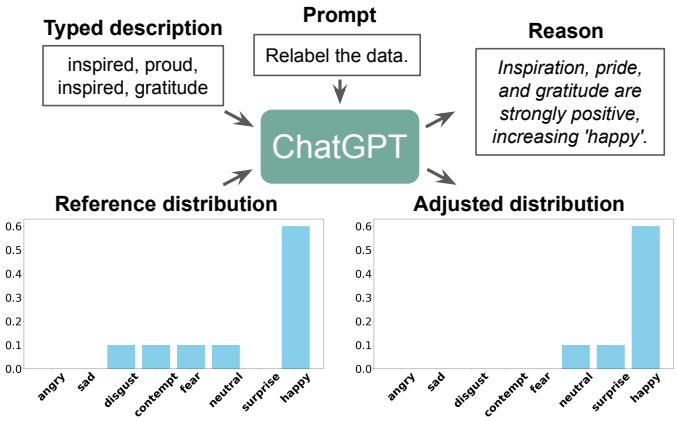


Fig. 1: Labeling process using ChatGPT. Three inputs are **Typed description**, **Reference distribution**, and **Prompt**. Two outputs are **Reason** and **Adjusted distribution**. Notice that the reference distribution is calculated by the number of votes for emotion classes. In the raw annotations of an example, there are instances of disgust, contempt, fear, neutrality, and happiness (*6), resulting in values of 0.6 for happiness and 0.1 for each of the remaining appearing emotions.

LLMs have demonstrated promising performance across a wide range of NLP tasks, including sentiment analysis [8]. Inspired by this success, we propose a novel approach to improve SER systems by utilizing LLMs to process and integrate typed descriptions. Specifically, we prompt ChatGPT to mimic annotators, comprehend the natural language typed descriptions, and adjust the given labels of the input data accordingly as illustrated in Figure 1. This approach harnesses the sophisticated understanding and text-generation capabilities of ChatGPT to enrich the emotional labels used in SER tasks.

In this study, we validate our proposed method on the MSP-PODCAST dataset, as it is the largest public English emotion

TABLE I: The Prompt for ChatGPT.

Objective:
As a knowledgeable assistant psychologist, your role is to analyze the given words and reference labels. You generate emotion label distributions. The emotions to consider are: 'angry,' 'sad,' 'disgust,' 'contempt,' 'fear,' 'neutral,' 'surprise,' and 'happy.' The order of emotions is very important. Please provide 8-dimensional emotion distributions for these 8 emotion classes based on the user input.
Input format:
The user input has two parts separated by #: The first part is the description. The second part is 8-dimensional reference emotion distribution, 'angry,' 'sad,' 'disgust,' 'contempt,' 'fear,' 'neutral,' 'surprise,' and 'happy.' The order of reference emotion is very important. The input has the format "descriptions#reference emotion distribution". Also give the reason for each data point why you want to change the reference emotion distribution. When given the answer, you should focus 25% on the "descriptions" and 75% on the "reference emotion distributions".
Example:
I will give you one example: User Input: Concerned,Interest#0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0. Generated Labels: {'angry': 0.1, 'sad': 0.2, 'disgust': 0.2, 'contempt': 0.3, 'fear': 0.0, 'neutral': 0.2, 'surprise': 0.0, 'happy': 0.0, "reason": ""}
Output format:
Reminder for the given data: It's very important to output the JSON format with an index.
Refine and Iterate :
I will give you 30 data points each time. Each data is separated by "—". It's very important. It's very important to make sure that you complete every response for 30 data points each time. Please reminder it. Output the JSON file that contains adjusted emotion label distributions based on reference distributions and detailed reasons why you adjust the reference emotion distributions for each word by each word. It's very important. It's very important that the JSON output file must contain the reference distributions and reasons. It's very important that do not contain the reference distributions and words. It's very important that use 15 to 20 words to explain the reason you want to change the reference distributions. It's very important that the sum of label distributions equals 1. It's very important to make sure that you explain the reasons for each word in descriptions.

dataset, which contains 34.37% of utterances with natural language descriptions. By leveraging labels generated by ChatGPT, we have consistently achieved an average relative improvement of 3.08% across almost all settings using 15 speech self-supervised learning models on the SUPERB benchmark. This improvement demonstrates the potential of integrating LLMs like ChatGPT to enhance the performance of SER systems. To the best of our knowledge, this study is the first to utilize ChatGPT to process typed descriptions to improve SER systems. Our findings suggest a new direction for future research and applications in the field of SER, showcasing the value of natural language descriptions and the power of LLMs in enhancing emotion recognition from speech.

II. Related Works

The integration of natural language descriptions into Speech Emotion Recognition (SER) has been a relatively unexplored area. One of the most closely related works is Chou et al. [9]. In their study, Chou et al. [9] combined polarity information derived from typed descriptions with the pre-defined emotions (e.g., happy or frustrated) provided by individual annotators. This was done within a label distribution framework to create a more comprehensive representation of the emotional content in spoken sentences. Finally, The authors then trained multi-task learning SER models using established three-label learning methods (soft-label, multi-label, and distribution-label) and demonstrated improved performance when typed descriptions were incorporated and evaluated on the MSP-Podcast corpus. However, the Chou et al. [9] approach has its limitations. The LIWC 2015 toolkit relies on predefined dictionaries to identify and interpret words. These dictionaries, although robust, are inherently limited in scope, often omitting many of the nuanced words found in typed emotional descriptions. For example, "Haaapy", indicating really happy, is not in the dictionaries. These limitation prevents the full utilization of the rich emotional content present in these natural language descriptions.

Different from the study by Chou et al. [9], our approach leverages the capabilities of large language models (LLMs) like ChatGPT to directly process and understand typed descriptions. By guiding the LLMs to consider both the natural language descriptions and the distributional labels calculated from predefined emotion categories, our method dynamically adjusts the labels. This approach not only captures a wider range of emotional nuances but also provides better generalization capabilities compared to the dictionary-based methods used in the LIWC toolkit.

III. Empower typed descriptions by ChatGPT

A. Resource

The MSP-PODCAST [4] collected spontaneous and diverse emotional speech from various real-world podcast recordings with a commercial license. In this work, we focus on primary emotion. Raters choose from nine categorized emotions in the primary emotion: angry, sad, happy, surprised, fear, disgusted, contempt, neutral, and "other." Each utterance has at least 5 raters, and the raters can type their own words to describe their emotion perception, named *typed description*. We exclude the other class and formulate the task as an 8-class emotion recognition task. We use the release version 1.11 of the database, including 84,030 utterances in the train set, 19,815 in the development set, 30,647 in the test1 set, and 14,815 in the test2 set. We combine the test1 and test2 as the test set. The database collects voices from more than 2,404 unique speakers.

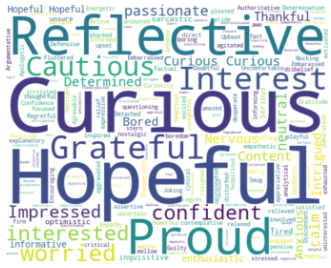


Fig. 2: The figures show typed descriptions of the MSP-PODCAST for an example.

B. Typed Descriptions

Fig. 2 displays a word cloud generated from typed descriptions collected in the MSP-PODCAST dataset. Word clouds are visual representations that highlight the frequency of words in a text corpus, with larger font sizes indicating higher frequencies. In this context, the word cloud provides insight into the types of words or phrases annotators use to describe their emotional responses in the dataset.

Incorporating typed descriptions can be valuable for understanding the nuances of human emotion. However, SER, primarily based on classification models, cannot process natural language and consequently overlooks these valuable typed descriptions.

C. Why Using ChatGPT for Relabeling

ChatGPT [10] exhibits a remarkable ability to comprehend and analyze natural language. [11] had used ChatGPT to do effective sentiment analysis. Hence, we utilize ChatGPT to mimic annotators, summarizing their thoughts to re-label the data with typed descriptions. While GPT models have been previously utilized for data labeling tasks, our approach stands out due to its innovative application in generating a distribution of labels instead of assigning a single label. We show that this approach leads to consistent improvements across all experimental settings, as shown in Table IV.

D. Prompt ChatGPT

We design a prompt to transform the released version of GPT-4 Turbo, a variant of ChatGPT, into a knowledgeable assistant **psychologist**. Its primary function is to generate a distribution across emotion labels based on the input-typed descriptions from annotators.

As shown in Figure 1, three inputs are provided to ChatGPT: the typed descriptions, reference distributions, and a well-designed prompt. When we prompt ChatGPT to refer to the distribution label, it fails to provide the distribution unless we supply the reference distribution. The format of the output emotion label is also a distribution. Guided by the prompt, the ChatGPT can adjust or maintain the reference distribution based on the typed descriptions. In the prompt, we also let ChatGPT explain why it changes or doesn't change the reference distributions. Without this, ChatGPT might default to laziness, consistently avoiding modifying the reference distributions. Table I in the supplementary material. shows the well-designed prompt, which contains five parts: objective, input format, example, output format, and refine and iterate. In the objective part, we clearly describe the goal of the task. Then, we define the input format, including descriptions and reference emotions. Afterward, we provide an example that provides the template the ChatGPT can follow. Finally, we ask the ChatGPT to output the file in JSON format. Notice that the current version of the prompt is the 14th version. In the refine and iterate part, we show more rules that can enhance the accuracy of the output of the ChatGPT. We encourage the community to provide the designed prompt.

We choose the MSP-PODCAST (P) dataset to verify the efficacy of our proposed prompt method in utilizing typed descriptions to improve SER, as it is the largest dataset and has the highest percentage

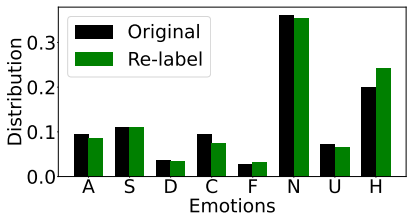


Fig. 3: The figures show comparisons of original and re-labeled label distribution. Emotion includes anger (A), sadness (S), happiness (H), surprise (U), fear (F), disgust (D), contempt (C), and neutral (N).

Model	Loss
Autoregressive Predictive Coding (APC) [12]	Generative loss
VQ-APC [13]	Generative loss
Non-autoregressive Predictive Coding (NPC) [14]	Generative loss
Mockingjay [15])	Generative loss
TERA [16]	Generative loss
DeCoAR 2 [17]	Generative loss
WavLM [18]	Discriminative loss
Hubert [19]	Discriminative loss
wav2vec 2.0 (**W2V2**) [20]	Discriminative loss
Data2Vec [21]	Discriminative loss
XLS-R [22]	Discriminative loss
VQ wav2vec (**VQ-W2V**) [23] [17]	Discriminative loss
wav2vec (**W2V**) [24]	Discriminative loss
Contrastive Predictive Coding (CPC) (**M CPC**)[25])	Discriminative loss

TABLE II: Summary of SSLMs

(6.08%) of typed descriptions among all other datasets. Figure 3 and 4 show the changes in label distributions between the original labels and the re-label one. The ChatGPT increased the number of fear and happiness and decreased the other emotions. In addition, Table III shows ten examples, including typed descriptions and reasons provided by ChatGPT.

IV. Experimental Setup

A. SSLM-based Codebase

1) Framework: Self-supervised learning (SSL) is a promising direction for developing speech models. This approach entails training a large model with large-scale unlabeled data to obtain robust and general representations. After pre-training, one can achieve nearly SOTA performance on downstream tasks by employing the fixed SSLMs alongside task-specific lightweight prediction heads [26]. Furthermore, SSLMs significantly enhance SER and demonstrate SOTA performance, as evidenced in [1].

We develop a comprehensive codebase highly depending on S3PRL [1] [26] to leverage 15 speech-supervised learning models as feature extractors and trains lightweight heads for exhaustive evaluation.

2) Self-supervised learning models: We leverage two mainstream categories of SOTA SSLMs (in S3PRL), pre-trained using generative losses and discriminative losses, summarized in Table II.

3) Label Representation: Inspired by **Semantics Space Theory** [27], we gather numerous annotations and compute a distribution-like (soft label) representation, to capture the high-dimensional nature of emotion perception more accurately. Notice that these distribution-like labels are the same as the **reference distribution** used for ChaptGPT as the reference label. Let's assume we gather five annotations from five raters for one sample. These annotations comprise neutral (N), anger (A), anger (A), sadness (S), and sadness (S). Subsequently, we compute the label distributions, which in this instance are represented as (N, A, S, H) = (0.2, 0.4, 0.4, 0.0) for training SER systems. Additionally, in order to enhance SER performance, we employ the label smoothing technique proposed by [28] to refine the vector, utilizing a smoothing parameter of 0.05. This approach assigns a small probability to emotional classes with zero values.

B. Evaluation Metric

We use the macro-F1 score [29] to evaluate the SER performance, considering recall and precision rates simultaneously. For the distribution-like multi-label training target, we select target classes by applying thresholds on the ground truth. A prediction is deemed successful if the proportion for a class surpasses $1/C$, where C represents the number of emotional classes, aligning with the settings

[1]https://github.com/s3prl/s3prl

TABLE III: Relabeled examples of typed descriptions with ChatGPT. ChatGPT also provides the reason for changing the reference distribution. **Change** denotes whether the reference distribution label is changed or not.

Change	Index	Typed Descriptions	Reason
Yes	01	calm,Slightly Angry,calm	Increased disgust to reflect slight anger.
No	01	Tranquil	Maintained high neutral for tranquil's peacefulness without strong emotions.

employed in prior research [30]. For instance, consider a four-class emotion recognition task, and the emotion classes contain neutral, anger, sadness, and happiness. Assume we consider the predictions for three different models: (0.2,0.35,0.35,0.1), (0.1,0.45,0.45,0.0), and (0.45,0.1,0.0,0.45). The three predictions are transformed into (0,1,1,0), (0,1,1,0), and (1,0,0,1), respectively, using the threshold. In these cases, only the first two predictions are fully corrected.

C. Training Details

We use the AdamW optimizer [31] with a 0.0001 learning rate, and the batch size is 32. We choose the best models according to the lowest value of the class-balanced cross-entropy loss on the development set. We use the Nvidia Tesla v100 GPUs with 32 GB memory for all results. The total of GPU hours is around 100 hours. According to [26], [32], [33], SSLMs usually result in consistent results and consume large computations. All results in the work are single-run. We also verify it by running experiments for small SSLMs, and the standard deviation is only less than 1% on average.

V. Results and Analysis

A. Main Results

We mainly use SSLMs as our backbone models to train SER systems in the work. Fig. 4 shows a randomly selected example to compare the original distribution and relabeled distribution by ChatGPT (data sample "PODCAST_1631_0043_0001.wav"). We observe that ChatGPT effectively comprehends the typed descriptions conveying positive emotions, thereby assigning greater weight to the "happy" emotion category. More examples can be found in Fig. F5a to Fig. F5d and Table III and Table F5 in the supplementary material.

Table IV presents the macro-F1 scores of the experiment along with the effects of incorporating data labeled by ChatGPT. We denote "W/O ChatGPT labels" and "W/ ChatGPT labels" to signify results without and with ChatGPT labels, respectively, while maintaining all other settings the same. We note the following observations: (1) The experiments involved 16 models, resulting in an average relative performance gain of 3.08%. (2) Particularly noteworthy is the case of CPC, which exhibits a substantial 9.45% relative improvement.

TABLE IV: The table presents macro-F1 scores using the integration of labels from ChatGPT.

SSLM	W/O ChatGPT Labels	W/ ChatGPT Labels	Relative Gain
WavLM	0.350	0.353	0.77%
W2V2 R	0.331	0.335	1.08%
XLS-R-1B	0.331	0.341	3.09%
Data2Vec-A	0.329	0.338	2.74%
Hubert	0.342	0.350	2.22%
W2V2	0.321	0.325	1.28%
VQ-W2V	0.292	0.300	2.74%
W2V	0.301	0.305	1.58%
CPC	0.265	0.290	**9.45%**
DeCoAR 2	0.308	0.317	3.14%
TERA	0.295	0.306	3.52%
Mockingjay	0.275	0.298	8.49%
NPC	0.275	0.290	5.75%
VQ-APC	0.296	0.310	4.94%
APC	0.298	0.307	3.19%
FBANK	0.186	0.186	0.00%
Average	0.298	0.307	3.08%

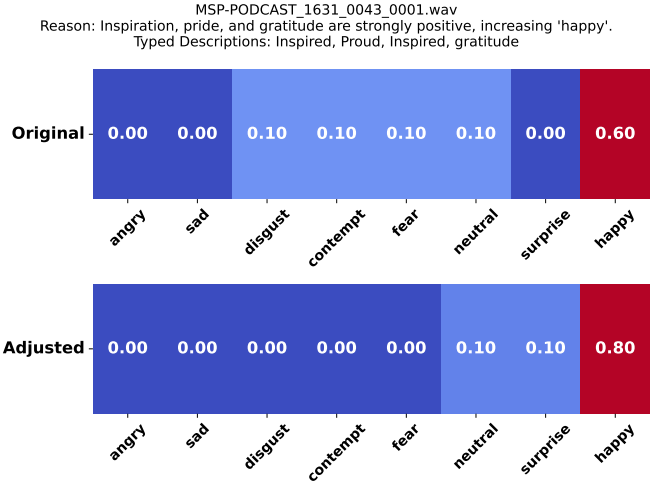


Fig. 4: Original and adjusted distributions. The original distribution, determined by tallying the votes for each emotion class, is compared with the adjusted distribution resulting from ChatGPT's re-labeling. Take the first one as an example; in the raw annotations of the example, there are instances of disgust, contempt, fear, neutrality, and happiness (*6), resulting in values of 0.6 for happiness and 0.1 for each of the remaining emotions.

B. Illustration for Re-labeled Data

Figure 4 shows one random example to compare the original distribution and relabeled distribution by ChatGPT (data samples "PODCAST_1631_0043_0001.wav"). We observe that ChatGPT effectively comprehends the typed descriptions conveying positive emotions, thereby assigning greater weight to the "happy" emotion category.

VI. Conclusion and Future Work

The findings of our study reveal a novel and effective approach to Speech Emotion Recognition (SER) by harnessing the capabilities of large language models (LLMs) such as ChatGPT to empower typed descriptions. By prompting ChatGPT to simulate annotators' cognitive processes and interpret typed descriptions of emotional perceptions, our method provides a more nuanced and accurate representation of emotional content compared to conventional classification models that rely solely on pre-defined categories. The consistent average relative gain of 3.08% in performance across 15 self-supervised learning models on the SUPERB codebase demonstrates the robustness and applicability of our methodology. Our research marks the first instance of leveraging ChatGPT to process and utilize typed descriptions to improve SER systems. In future work, we will use other databases containing typed descriptions, such as MSP-IMPROV [5] and NNIME [34] and other LLMs to investigate the generalization of the proposed method.

ACKNOWLEDGMENT

NSTC supported this research under Grants 112-2634-F-002-005. We also thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources. We also thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, *et al.*, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023. DOI: 10.1109/TPAMI.2023.3263585.

[2] J. A. Russell, "Forced-choice response format in the study of facial expression," *Motivation and Emotion*, vol. 17, no. 1, pp. 41–51, Mar. 1993. DOI: 10.1007/BF00995206.

[3] C. Busso and S. Narayanan, "Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, Sep. 2008, pp. 1670–1673.

[4] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, Oct. 2019. DOI: 10.1109/TAFFC.2017.2736999.

[5] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017. DOI: 10.1109/TAFFC.2016.2515617.

[6] H.-C. Chou, L. Goncalves, S.-G. Leem, A. N. Salman, C.-C. Lee, and C. Busso, "Minority Views Matter: Evaluating Speech Emotion Classifiers with Human Subjective Annotations by an All-Inclusive Aggregation Rule," *IEEE Transactions on Affective Computing*, pp. 1–15, 2024. DOI: 10.1109/TAFFC.2024.3411290.

[7] H. Wu, H.-C. Chou, K.-W. Chang, *et al.*, "Open-Emotion: A Reproducible EMO-SUPERB for Speech Emotion Recognition Systems," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2024.

[8] W. Zhang, Y. Deng, B. Liu, S. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," in *Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 3881–3906. [Online]. Available: https://aclanthology.org/2024.findings-naacl.246.

[9] H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. Busso, "Exploiting Annotators' Typed Description of Emotion Perception to Maximize Utilization of Ratings for Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7717–7721. DOI: 10.1109/ICASSP43922.2022.9746990.

[10] J. Achiam, S. Adler, S. Agarwal, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[11] K. Kheiri and H. Karimi, *SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning*, 2023. arXiv: 2307.10234 [cs.CL].

[12] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Interspeech 2019*, 2019, pp. 146–150. DOI: 10.21437/Interspeech.2019-1473.

[13] Y.-A. Chung, H. Tang, and J. Glass, "Vector-Quantized Autoregressive Predictive Coding," in *Proc. Interspeech 2020*, 2020, pp. 3760–3764. DOI: 10.21437/Interspeech.2020-1228.

[14] A. H. Liu, Y.-A. Chung, and J. Glass, "Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies," in *Proc. Interspeech 2021*, 2021, pp. 3730–3734. DOI: 10.21437/Interspeech.2021-349.

[15] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6419–6423. DOI: 10.1109/ICASSP40776.2020.9054458.

[16] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021. DOI: 10.1109/TASLP.2021.3095662.

[17] S. Ling and Y. Liu, *Decoar 2.0: Deep contextualized acoustic representations with vector quantization*, 2020. arXiv: 2012.06659 [eess.AS].

[18] S. Chen, C. Wang, Z. Chen, *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022. DOI: 10.1109/JSTSP.2022.3188113.

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021. DOI: 10.1109/TASLP.2021.3122291.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.

neurips . cc / paper_files / paper / 2020 / file / 92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

[21] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language," in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 17–23 Jul 2022, pp. 1298–1312. [Online]. Available: https://proceedings.mlr.press/v162/baevski22a.html.

[22] A. Babu, C. Wang, A. Tjandra, *et al.*, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Interspeech 2022*, 2022, pp. 2278–2282. DOI: 10.21437/Interspeech.2022-143.

[23] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rylwJxrYDS.

[24] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Interspeech 2019*, 2019, pp. 3465–3469. DOI: 10.21437/Interspeech.2019-1873.

[25] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[26] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*, 2021, pp. 1194–1198. DOI: 10.21437/Interspeech.2021-1775.

[27] A. S. Cowen and D. Keltner, "Semantic Space Theory: A Computational Approach to Emotion," *Trends in Cognitive Sciences*, vol. 25, no. 2, pp. 124–136, 2021, ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2020.11.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136466132030276X.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[29] J. Opitz and S. Burst, "Macro f1 and macro f1," *arXiv preprint arXiv:1911.03347*, 2019.

[30] P. Riera, L. Ferrer, A. Gravano, and L. Gauder, "No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems," in *Proc. SMM19, Workshop on Speech, Music and Mind 2019*, Graz, Austria, Sep. 2019, pp. 11–15.

[31] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7.

[32] H.-S. Tsai, H.-J. Chang, W.-C. Huang, *et al.*, "SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8479–8492. DOI: 10.18653/v1/2022.acl-long.580. [Online]. Available: https://aclanthology.org/2022.acl-long.580.

[33] T.-h. Feng, A. Dong, C.-F. Yeh, *et al.*, "Superb @ SLT 2022: Challenge on Generalization and Efficiency of Self-Supervised Speech Representation Learning," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1096–1103. DOI: 10.1109/SLT54892.2023.10022770.

[34] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, and C.-C. Lee, "NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 292–298.

An Investigation on the Speech Recovery from EEG Signals Using Transformer

Tomoaki MIZUNO* Takuya KISHIDA† Natsue YOSHIMURA‡ Toru NAKASHIKA*
* The University of Electro-Communications, Tokyo, Japan
† Aichi Shukutoku University, Aichi, Japan
‡ Institute of Science Tokyo, Kanagawa, Japan

***Abstract*—In recent years, brain-machine interfaces (BMI) have been researched to enable people who cannot physically speak or who are placed in situations where they cannot speak to engage in speech communication. However, synthesizing complete speech from ElectroEncephaloGraphy (EEG) signals remains a challenging task. In this paper, reconstructing speech from EEG signals, a Transformer-based model has been developed on the basis of data from listening to the speeches of two people, one male and one female. The objective of this study is to investigate the potential to reconstruct speech from EEG signals, including the characteristics of the corresponding speaker's speech and the potential to reconstruct speeches containing the corresponding linguistic content, by training a single model on both male and female speeches. Our findings reveal that our model can generate two distinctly different speakers' speeches from the EEG signals of the two speakers' speeches. Additionally, the EEG signals appear to contain information about speaker characteristics that can be reconstructed from the EEG signals.**

I. Introduction

Brain-Computer Interface (BCI) refers to computer technology that has the potential to enable people to operate a computer simply by thinking. Research on reading language information from brain activity and synthesizing it as speech is also advancing and could have significant implications for enabling those who are unable to speak due to illness or in environments where speech is not possible, such as underwater, or similar conditions, to communicate vocally.

BCI can be broadly classified into two types (invasive and noninvasive) depending on the method used to measure brain activity. Invasive methods involve the direct implantation of electrodes into the brain, providing high-resolution and precise data. This technique can directly measure the activity of specific neurons in the brain, resulting in minimal signal noise and very short temporal delays. However, it requires surgical procedures, which come with risks such as infections and inflammation. Non-invasive methods, on the other hand, allow for the measurement of brain activity without surgery, thereby posing relatively lower risks. These methods can measure electrical activity and changes in blood flow in the brain, but they offer lower resolution compared to invasive methods. Electrocorticography (ECoG), which is measured by placing electrodes on the surface of the brain in the invasive method, is used as a signal related to brain activity that enables highly accurate speech synthesis [1], [2]. Previous studies have successfully synthesized intelligible speech by combining brain activity and mouth movements, but complete speech has not yet been synthesized with signals without mouth movement information. In addition, the invasive type requires surgery to attach electrodes to the surface of the brain, which is costly. In contrast, electroencephalography (EEG) is a non-invasive method that measures electrical signals by placing electrodes on the scalp, making it a more convenient way to measure brain activity. However, synthesizing complete speech sounds with noninvasive EEG is a more challenging task than with ECoG. Nevertheless, a study synthesized two short vowels from EEG with high intelligibility, using a neural network introduced to infer phonetic features from EEG [3]. The structure of the neural network is relatively straightforward, which suggests that it may be a viable approach to infer speech features from EEG. It also suggests the potential for employing more complex network structures to infer speech features from EEG.

To propose this possibility, we attempted to synthesize speech from a single participant's EEG when listening to a single speaker's speech using a diffusion model first. We trained Diffwave [4], a vocoder using a diffusion model, by modifying the input to an EEG converted from a mel-spectrogram to a two-dimensional signal and assessed the quality of the synthesized speech. The results showed that most segments had only small noises and only one tone was mixed in a few times within the segment. The model possibly may not have handled time-series signals as well as it could have done.

Given that the auditory system is generally hierarchically organized, that low-level acoustic features (frequency, intensity, etc.) are processed faster than higher-order semantic processing [5], [6], and that the electrical signals generated from these features are collected using low spatial resolution methods, it seems that not all phonemes in speech are expressed as EEG while preserving their order. Additionally, although we refer to the EEG of one person when listening to the speech of a single speaker, it is important to note that individual differences in EEG expression do exist, such as the difference in reaction time between the elderly and the young [7]. We believe that a model that can encompass these characteristics would be beneficial. In this study, we focus on the Transformer [8].

Transformer, which is a deep learning model applied to the speech field, has been successfully used to synthesize speech with higher quality than conventional neural networks

in speech synthesis and voice conversion [9]. By using a deep learning model with Transformer to infer speech features from EEG, speech more complex than two types of short vowels may be possible to generate. In this study, we attempt to infer speech features from EEG using a deep learning model with Transformer. We use the speech of two speakers and the EEG of a person who listened to the speakers as training data, and then we perform inference to reconstruct the speech from the EEG. Another of our goals is to see if the difference in speaker characteristics between the two speakers can be recovered. We explore the possibility to recover the differences in speaker identity between the two speakers by learning from the two speakers' speeches and the speeches of the people who heard them and by performing inference to reconstruct speeches.

II. Architecture Selection

Voice conversion technology generates speech by converting one person's speech into another person's speech while retaining the linguistic information and speaking style of the speech. Since EEG is also a time-series signal, we thought that a deep learning model based on the voice conversion technique could be used to infer features that represent the speech from EEG. Transformer is an architecture used in deep learning models based onvoice conversion technology and has also been extensively applied beyond voice conversion technology.

Transformer excels at handling sequential data and capturing long-range dependencies. These are particularly effective when input data is transformed into dense vector representations that encapsulate semantic information. Positional encoding is added to maintain the order of data within the sequence. The attention mechanisms in Transformer enables the model to weigh the significance of different elements in the sequence, thereby enhancing contextual understanding.

This paper posits that the strengths of Transformer can be leveraged to infer complex patterns from sequential data. On the basis of these strengths, Transformer-based models used in voice conversion techniques are used to synthesize speech from EEG.

A. Transformer-based Text-To-Speech system

Fig. 1 shows the Transformer-based Text-To-Speech(TTS) system [9] and the Voice Transformer Network, which is a voice conversion system based on the TTS system network [10].

The Transformer model consists of an encoder and a decoder, and is used to find the mapping from a source sequence $\boldsymbol{x}_{1:n} = (x_1, ..., x_n)$ to a target sequence $\boldsymbol{y}_{1:m} = (y_1, ..., y_m)$. The lengths of the two sequences do not need to be the same.

The encoder Enc maps the source sequence $\boldsymbol{x}_{1:n} = (x_1, ..., x_n)$ to a sequence of hidden states $\boldsymbol{h}_{1:n} = (h_1, ..., h_n)$.

$$\boldsymbol{h}_{1:n} = \mathrm{Enc}\left(\boldsymbol{x}_{1:n}\right) \tag{1}$$

From the hidden states $\boldsymbol{h}_{1:n}$ and the decoder's previous hidden state $\boldsymbol{q}_{t-1}$, we obtain query, key and value.

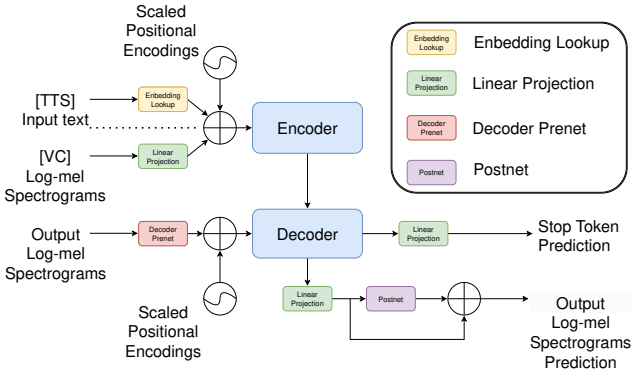


Fig. 1. Architecture of Transformer-based Text-To-Speech system and Voice Transformer Network.

$$\boldsymbol{Q} = W_Q \boldsymbol{q}_{t-1} \tag{2}$$

$$\boldsymbol{K} = W_K \boldsymbol{h}_{1:n} \tag{3}$$

$$\boldsymbol{V} = W_V \boldsymbol{h}_{1:n} \tag{4}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d_{model} \times d_{model}}$ are a learnable weight matrix. d_{model} is the input dimension. The weighted sum of these hidden states is the context vector $\boldsymbol{c}_t$.

$$\boldsymbol{c}_t = \sum_{k=1}^{n} a_t^{(n)} \boldsymbol{h}_k \tag{5}$$

Here, the attention weights $\boldsymbol{\alpha}_t$ are calculated using the hidden states $\boldsymbol{h}_{1:n}$ and the decoder's previous hidden state $\boldsymbol{q}_{t-1}$ through an attention mechanism [11], [12]. The attention weights, represented by the vector $\boldsymbol{\alpha}_t = (a_t^{(1)}, \cdots, a_t^{(n)})$, indicate which encoder hidden states the decoder should focus on to determine the output at time t. Each attention probability $a_t^{(k)}$ can be thought of as the importance of the hidden representation $\boldsymbol{h}_k$ at the tth time step.

$$\boldsymbol{a}_t = \mathrm{attention}\left(\boldsymbol{q}_{t-1}, \boldsymbol{h}_{1:n}\right) = \mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V} \tag{6}$$

d_k is the dimensionality of the key. Scaling by $\sqrt{d_k}$ is used to stabilize the gradient during training.

In practice, multi-head attention is often employed, which involves performing the above computation multiple times in parallel with different learned projections. For h heads:

$$\text{multi-head}\left(\boldsymbol{q}_{t-1}, \boldsymbol{h}_{1:n}\right) = (\text{head}_1, ..., \text{head}_h)W^O \tag{7}$$

$$where\ \text{head}_i = \text{Attention}(W_Q^i \boldsymbol{q}_{t-1}, W_K^i \boldsymbol{h_{1:n}}, W_V^i \boldsymbol{h_{1:n}}) \tag{8}$$

and $W^O \in \mathbb{R}^{d_{model} \times d_{model}}$ is a learnable weight matrix.

The decoder Dec takes the context vector $\boldsymbol{c}_t$, the previously generated output sequence $\boldsymbol{y}_{1:t-1} = (y_1, ..., y_{t-1})$, and the previous decoder hidden state as inputs and decodes the decoder hidden state $\boldsymbol{q}_t$ and $\boldsymbol{y}_t$.

$$\boldsymbol{y}_t, \boldsymbol{q}_t = \mathrm{Dec}\left(\boldsymbol{y}_{1:t-1}, \boldsymbol{q}_{t-1}, \boldsymbol{c}_t\right) \tag{9}$$

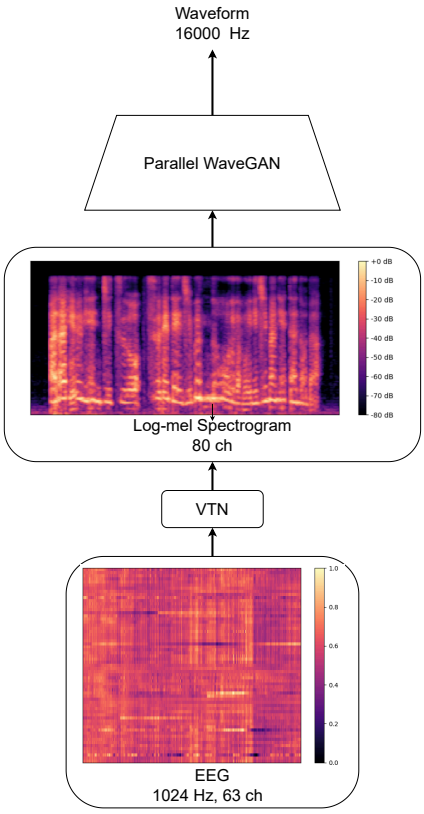


Fig. 2. Overview of the model for inferring speech from EEG: EEG is represented as a two-dimensional signal of channel and time, with normalized voltage. The x-axis of the EEG represents time and the y-axis shows the EEG electrode channels.

B. Voice Transformer Network: Transformer-based Voice Conversion System

A Voice Transformer Network is proposed as a voice conversion system based on a Transformer-based TTS system [10]. It adapts by passing a linear projection to convert the input from text to acoustic features (in this case, log-scale mel-spectrograms). Otherwise, the system is identical to the Transformer-based TTS system.

III. Method

Since EEG is also a time-series signal, we propose using the Voice Transformer Network to reconstruct speech from EEG. The speech Transformer Network is a model in which the input is replaced by acoustic features. The transformation process is shown in Fig. 2. This model consists of a Voice Transformer Network with 63 channels of EEG input and Parallel WaveGAN [14] as a vocoder, which is trained in an End-2-End fashion. Note that the Parallel WaveGAN has already been trained and the parameters of Parallel WaveGAN are not changed during training.

The following describes the dataset and model for the speech synthesis experiments from EEG using the Voice Transformer Network.

A. EEG measurements and dataset preparation

The EEG during Japanese speech listening was measured and processed to create a dataset of the EEG during speech listening. The speech sounds are the ATR 503 sentences [15] uttered by a male speaker (MMY) and a female speaker (FTK) respectively.

To create the EEG and speech dataset, the EEG was measured while the participants listened to the speech uttered by the two speakers. The Biosemi ActiveTwo system (Biosemi) was used for the measurement. The electrode configuration of the participants was 64 ch, based on the International 10–20 system. The EEG was initially sampled at a sampling frequency of 8,192 Hz. The signal was then bandpass filtered to retain signal frequency components between 1 and 40 Hz, and subsequently downsampled to a sampling frequency of 1,024 Hz to reduce computational cost. Then, independent component analysis (ICA) was applied to remove the components caused by eye movements. This method decomposes the EEG signals into statistically independent components, identifies those related to ocular artifacts based on their spatial and temporal characteristics, and reconstructs the EEG signal without these components, effectively removing eye movement-related artifacts from the data. The EEG electrode channel TP8, from which the eye movement-related component could not be removed, was not used in the analysis. Then, all data were normalized. To synchronize the 63 ch EEGs created by the above procedure with the audio, we cut the EEG from the trigger of the start of playback to 1024 samples (1 second) before the trigger of the start of playback of the next audio and created 503 EEG sets corresponding to the 503 sentences of audio. When creating the training sets, the sampling rate of the speech was changed to 16 kHz. The created dataset was divided into training data, validation data, and test data in the ratio of 480:12:11. MMY and FTK datasets were created and then merged, resulting in 960 sentences of training data, 24 sentences of validation data, and 22 sentences of test data in the end.

B. Model details

The parameters of the Voice Transformer Network[1] were used mostly as is. The input dimension was changed to use EEG as input. Also, this implementation did not include Guided Attention [16], but we used guided attention to reduce the risk of using brain wave information from periods when the subject was not listening during the speech reconstruction. Guided Attention means that the attention matrix A is approximately diagonal. The attention matrix A is a list of attention weights $\boldsymbol{a}_t$ at each time step. The loss function is defined as follows

$$L_{att}(A) = \mathbb{E}_{nt}[A_{nt}W_{nt}] \tag{10}$$

$$W_{nt} = 1 - \exp\left(-\frac{(n/N - t/T)^2}{2g^2}\right) \tag{11}$$

[1]https://github.com/unilight/seq2seq-vc/tree/main/egs/arctic/vc1

Here, n is the position of a character or word in the text, N is the total length of the text, t is the index of the time step, and T is the total length of the audio. A_{nt} refers to the entry of the attention matrix A, representing the correspondence between a specific text position n and a time step t. It indicates the weight of the attention that the text position n should give to the audio at time t, with higher values indicating stronger correlations and lower values indicating weaker correlations. W_{nt} is an element of the weight matrix used to calculate the guided attention loss, as defined in Equation (6). The parameter g controls the variance of the Gaussian function, adjusting the penalty strength on the basis of the distance between the text position n and the time step t. In this experiment, we set $g = 0.4$. For training, we designed and utilized the overall loss function as follows, incorporating the L1-based reconstruction loss L_{L1} used in the existing VTN, the binary cross-entropy loss L_{BCE} to enable the model to learn the timing to stop decoding, and the aforementioned guided attention loss.

$$L = L_{L1} + \alpha L_{BCE} + \beta L_{att} \tag{12}$$

Here, α and β are the weight applied to the binary cross-entropy loss and the guided attention loss. In this experiment, α and β were set to 10 and 1 respectively.

For Parallel WaveGAN [14] used as vocoder, a two-speaker model of MMY and FTK was trained and used for the experiment. The sampling rate was 16 kHz, and the default parameters of Parallel WaveGan [2] were used otherwise.

C. Validation Items

In conducting the experiments, the Voice Transformer Network was trained under the above conditions, and the following items were verified.

- Synthesis quality when inferring speech from EEG training data
- Synthesis quality when inferring speech from EEG test data

To investigate these, we investigated the following indices of speech inferred from EEG training data to examine the linguistic features of the speech.

- Word Error Rate (WER)
- Character Error Rate (CER)
- BertScore [17]

For WER and CER, a smaller value indicates fewer errors. The training data and ground-truth speech were recognized by the Large-v3 model of the speech recognition model Whisper [18] and compared with the text data of the original speech. The BertScore is an index that calculates the similarity of text using BERT, with higher values indicating greater similarity. For the P_{BERT}, R_{BERT}, and $F1_{\text{BERT}}$ values associated with BertScore, Precision (P_{BERT}) is the percentage of predicted correct answers that were actually correct, Recall (R_{BERT}) is the percentage of actual correct answers that were correctly predicted, and F1 score ($F1_{\text{BERT}}$) is the harmonic mean of Precision and Recall.

[2]https://github.com/kan-bayashi/ParallelWaveGAN

TABLE I
WER, CER, AND BERTSCORE OF THE GENERATED SPEECH FROM THE TRAINING DATA.

	WER (↓)	CER (↓)	BertScore(↑)		
			P_{BERT}	R_{BERT}	$F1_{\text{BERT}}$
Train	0.0976	0.0462	0.956	0.954	0.955
Test	2.16	1.13	0.622	0.629	0.625

TABLE II
MEAN AND VARIANCE OF F0, F1, AND F2 OF THE GENERATED MMY SPEECH FROM THE TEST DATA AND THE GROUND TRUTH SPEECH.

MMY	F0[Hz]		F1[Hz]		F2[Hz]	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
GT	154.6	38.17	773.6	483.2	1898	495.2
Predict	221.2	36.72	793.5	428.0	1940	455.0

TABLE III
MEAN AND VARIANCE OF F0, F1, AND F2 OF THE GENERATED FTK SPEECH FROM THE TEST DATA AND THE GROUND TRUTH SPEECH

FTK	F0[Hz]		F1[Hz]		F2[Hz]	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
GT	236.2	51.90	759.5	393.9	2010	498.6
Predict	293.1	36.72	730.8	370.2	2019	441.4

We also extracted the following data, computed frame by frame, to investigate the representation of talkativeness from the test data and the speech inferred from the ground-truth speech.

- Pitch (F0)
- 1st and 2nd formants (F1, F2)

We use support vector machines (SVM) to verify whether speaker information is reflected in the generated speech. A SVM was trained to calculate mel-frequency cepstral coefficients (MFCC) for each frame of 20 channels from the ground-truth speech to create a speaker identification model (94% accuracy). We use the SVM model to identify whether a male or female speech is used for the generated speech and to verify the accuracy.

IV. RESULTS AND DISCUSSION

Fig. 3 compares the mel-spectrograms of the ground-truth speech and the speech generated from the training data, and Fig.4 compares the mel-spectrograms of the ground-truth speech and the speech generated from the test data. The results are shown in Table reftext for the values of WER, CER, and BertScore for the speech synthesized from the EEG that was used for the training data, relative to the ground-truth speech.

The mean values and variances of F0, F1, and F2 calculated for every frame for all speech are shown in Table.II and Table.III. These two tables show the values for ground-truth (GT) speech and generated speech (Predict) for MMY and FTK respectively.

First, comparing the mel-spectrogram in Fig. 3, the speech generated from the training data clearly has a similar overall shape, with many of the wave timings and silence areas similar to the ground-truth speech. However, the speech generated from the test data in the mel-spectrogram in Fig. 4 differs from the truth speech in terms of the timing and characteristics of

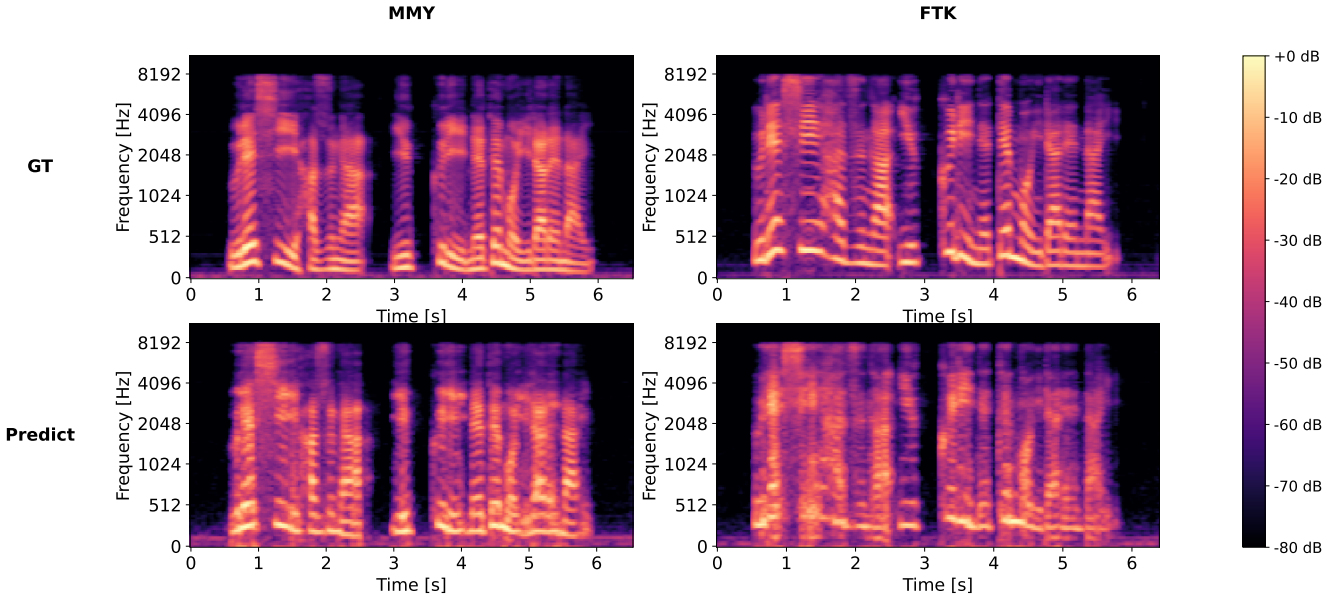


Fig. 3. Comparison of mel-spectrograms of ground-truth speech and speech generated from training data.

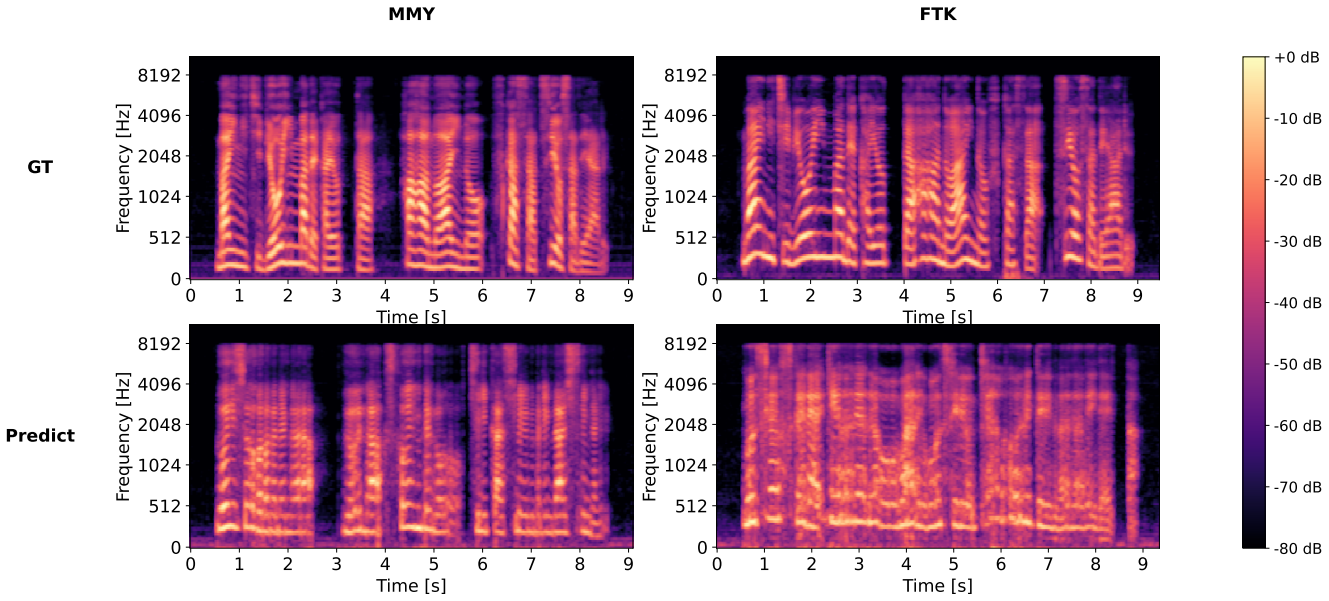


Fig. 4. Comparison of mel-spectrograms of ground-truth speech and speech generated from test data.

the silence. From Table I, the WER and CER values for speech inferred from the training data were 0.0976 and 0.0462, and the BertScore value was about 0.955 for both P_{BERT}, R_{BERT} and $F1_{\text{BERT}}$. From these values and the results in Fig. 3 and Fig. 4, we can say that the speech inferred from the training data is synthesized with a high degree of accuracy.

We attempted to perform the same experiment on speech generated from test EEG data, but the WER and CER values exceeded 1, and the BertScore value was about 0.625 for both P_{BERT}, R_{BERT} and $F1_{\text{BERT}}$. This can be described as having significantly low accuracy. In addition to the poor accuracy of generation from EEG signals, the generated speeches do not form coherent sentences, despite Whisper's attempts to infer them as meaningful sentences. This lack of coherence in the generated speech is considered to be a contributing factor. From the above, it can be seen that linguistic information could not be converted into speech.

To conduct the analysis for Tables II and III, we performed Steel-Dwass tests on the F0, F1, and F2 values calculated for each sentence from the original and generated speech

for both MMY and FTK. Steel-Dwass tests are a multiple comparison procedure. By utilizing this approach, we address the multiplicity issue that can occur with t-tests, ensuring a more robust statistical comparison across multiple groups. As a result, significant differences were observed in the mean values of F0 and F1 across all four comparisons. For F2, significant differences were also found between the two synthesized speeches, suggesting that the generated voices do not resemble each other closely. In addition, SVM was able to correctly identify the generated speeches as well as the ground-truth speeches with 85.34% accuracy. This indicates that different speeches can be recovered if the speakers of the speeches are different. The average value of MMY of the ground-truth speech and the generated speech increased from 154.6 to 221.2 Hz. The mean value of FTK of the ground-truth speech and FTK of the generated speech increased from 236.2 to 293.1 Hz. These results indicate that the F0 of the generated speech is higher than that of the ground-truth speech about 60 Hz. The values of variance do not change much, indicating that although the distribution of F0 is close to that of the heard sounds, the F0 is not correctly estimated as an average. However, although the generated speech has a different speaker characteristics from the ground-truth speech, MMY and FTK of the generated speech are clearly spoken by different speakers, and this can be seen from the F0 value and the accuracy of speaker identification by SVM. These facts suggest that features representing speaker characteristics can be estimated from EEG. However, we should note that the data in this experiment is only 480 sentences, and the high accuracy of the speech inferred from the training data suggests that the model may be over-trained on the training data. At the same time, the high accuracy of the speech inferred from the training data suggests that the model may be able to synthesize more accurate speech by increasing the training data.

V. Conclusion

As a model for the Transformer, we used the Voice Transfer Network to recover heard speeches from the EEG. This study showed the potential of the Transformer for speech decoding using EEG. Furthermore, using the speeches of two speakers and the EEG signals of the listeners, we were able to distinguish between the two speakers. This indicates that EEG signals contain information related to the vocal characteristics of the heard speeches. It also demonstrated the potential to reconstruct speeches from EEG signals.

As mentioned above, we could not examine the linguistic features of the speeches generated from the test data because the generated speeches were not saying meaningful sentences like the original speeches but rather were making a series of random sounds. Therefore, future work involves investigating in further detail the linguistic features of the speeches and improving the model structure to infer linguistic features from EEG signals.

Another issue that needs to be addressed in future experiments is to increase the amount of data in the experiments. First, the data needs to be increased to avoid over-fitting. We will also need to increase the number of speakers. In this experiment, we used the speeches of two speakers, a man and a woman, and the F0 values of these two speakers are very different. We will investigate the possibility of increasing the number of speakers to generate speeches with different characteristics from a wider variety of speakers.

VI. Acknowledgment

This work was supported by Hirokazu Matsuura, Kaya Chin, and Keito Nakamori of Institute of Science Tokyo for measuring and providing EEG data for the execution of this study. This research was supported by JSPS Grant-in-Aid for Scientific Research 23K11161 and 24H00715. This work was supported by JST FOREST Program (Grant Number JPMJFR216W).

References

[1] Anumanchipalli, G. K., et al. (2019). Speech synthesis from neural decoding of spoken sentences. Nature, 568, 493–498.

[2] Metzger, J. C., et al. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 620, 1037–1046.

[3] Akashi, W., et al. (2021). Vowel sound synthesis from electroencephalography during listening and recalling. Advanced Intelligent Systems, 2000164, 1-9.

[4] Kong, Z., et al. (2020). DiffWave: A Versatile Diffusion Model for Audio Synthesis. ArXiv, abs/2009.09761.

[5] Keshishian, M., et al. (2023). Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. Nature human behaviour, 7(5), 740–753. https://doi.org/10.1038/s41562-023-01520-0

[6] Zhang, Y., et al. (2011). Cortical Dynamics of Acoustic and Phonological Processing in Speech Perception. PLOS ONE, 6(7), e20963.

[7] Yang, W., et al. (2022). Auditory attentional load modulates the temporal dynamics of audiovisual integration in older adults: An ERPs study. Frontiers in aging neuroscience, 14, 1007954. https://doi.org/10.3389/fnagi.2022.1007954

[8] Zhang, L., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[9] Li, N., et al. (2019). Neural Speech Synthesis with Transformer Network. In Proceedings of the AAAI Conference on Artificial Intelligence, 33, 6706–6713.

[10] Huang, W.-C., et al. (2020). Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining. In Proc. Interspeech, 4676–4680.

[11] Bahdanau, D., et al. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[12] Luong, T., et al. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, Sep. 2015, 1412–1421.

[13] Shen, J., et al. (2018). Natural TTS Synthesis by Conditioning WaveNet on MEL Spectrogram Predictions. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4779–4783.

[14] Yamamoto, R., et al. (2019). Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. arXiv preprint arXiv:1910.11480.

[15] Akira, K., et al. (1990). ATR Japanese speech database as a tool of speech recognition and synthesis. Speech Communication, 9(4), 357–363.

[16] Tachibana, H., et al. (2018). Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4784–4788.

[17] Zhang, T., et al. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

[18] Radford, A., et al. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning, 28492-28518.

WavLM and Omni-Scale CNNs: Enhancing Boundary Detection in Partially Spoofed Audio

Menghan Li* and Zhihua Huang*†
* School of Computer Science and Technology, Xinjiang University, Urumqi, China
E-mail: 1131687972@qq.com Tel/Fax: +86-15599694888
† Key Laboratory of Signal Detection and Processing in Xinjiang, Umumgi, China
E-mail:zhhuang@xju.edu.cn Tel/Fax: +86-18699179000

***Abstract*—Partially spoofed/fake audio, in which segments of utterances are replaced with synthetic or natural audio clips, has emerged as a new form of deep audio forgery, posing potential severe threats to societal security. To address this issue, we employ a deep learning-based frame-level detection system, introducing a frame-level detection approach.We investigate the performance of WavLM in detecting waveform boundaries, utilizing WavLM as a feature extractor for original audio samples. Acoustic features and frame-level embeddings are concatenated, with an OS block embedded within the frame-level feature extraction process, in conjunction with CNN-1D to form the ResNet-OS network. This system can detect partially deceived audio and pinpoint the manipulated segments, effectively integrating multi-scale convolution to consider the interplay between local and global information in time series classification tasks. Experimental results show that this approach offers substantial generalization capabilities and robustness compared to traditional frame-level detection techniques.**

I. INTRODUCTION

Deepfake audio refers to sounds that are modified or created using deep learning technologies, aimed at deceiving both humans and machines. These technologies, particularly text-to-speech (TTS) and voice conversion (VC), have significantly enhanced the realism of synthetic voices, making them nearly indistinguishable from natural speech[1]–[3]. In this context, high-fidelity synthesis/conversion technologies invariably give thieves the tools they need to pose as someone else and conduct fraud. Therefore, the detection of audio forgeries is linked to societal security, privacy protection, and property safety, making it crucial to effectively detect deceptive speech to mitigate the potential threats posed by false information in audio content.

The ASVspoof Challenge is performed every two years to investigate defensive techniques against several kinds of spoofing attacks, such as synthetic speech, voice conversion, playback, and mimicry[4], in response to the concerns posed by audio spoofing assaults. However, many datasets and challenges have overlooked an important scenario. In this scenario, genuine speech is mixed with synthesized speech segments. This results in partial spoofing (PS).Attackers can use PS technology to modify key words in a sentence, such as time, place, and characters, thereby changing the semantic meaning of the sentence. This type of modification is low-cost and easy to perform. Therefore, defending against this PS scenario presents a significant challenge for defenders.

There have been notable developments in the field of Audio Deepfake Detection (ADD) with respect to PS scenarios in the last several years. Yi et al.[5]developed a dataset focused on altering several key words in semi-authentic audio, while Zhang et al.[6]designed the "PartialSpoof" voice database specifically for PS scenarios. These databases signify the commencement of PS scenario research within ADD tasks. Researchers have explored large-scale self-supervised pre-training models[7], [8],which have shown superior performance over traditional acoustic features such as MFCC and LFCC. Wu et al.[9] have developed a boundary detection system named "Fake Span Discovery", which is capable of identifying splicing boundaries. Additionally, Track 2 of ADD 2023 emphasizes the importance of accurately locating altered areas, making the research into the precise localization of spliced segments a more challenging and critical task. The performance of the boundary detection system still requires further enhancement.

This paper introduces a novel frame-level boundary detection system designed to identify partially spoofed audio. Unlike previous approaches that focused on utterance-level detection, our system exploits discontinuities between audio segments to accurately detect connection boundaries at the frame level. We employ WavLM as a feature extractor, together with the ResNet-OS architecture and a Transformer-based frame-level classifier for border identification. The HAD dataset was used to train the model, which was then evaluated on test sets. The results show that our system exceeds the performance of current partial fake audio detection methods in boundary detection.It achieved an Equal Error Rate (EER) of 0.06% on the HAD validation set and 0.023% on the HAD test set, demonstrating its cutting-edge ability to detect fabricated sections in audio streams.

II. METHOD

In this study, we focus on audio signals containing forged segments, with the objective to identify frames that contain discontinuous information. Given the acoustic feature input $X = (x_1, x_2, \ldots, x_T) \in \mathbb{R}^{D \times T}$, where D represents the dimension of features and T represents the number of frames, this task is defined as a frame-level binary classification problem. The classification labels $y = (y_{1,}y_{2,\ldots,}y_{T,}) \in \{0,1\}^T$.

Our system is capable of identifying specific frames that contain discontinuities. If a frame is located at the boundary

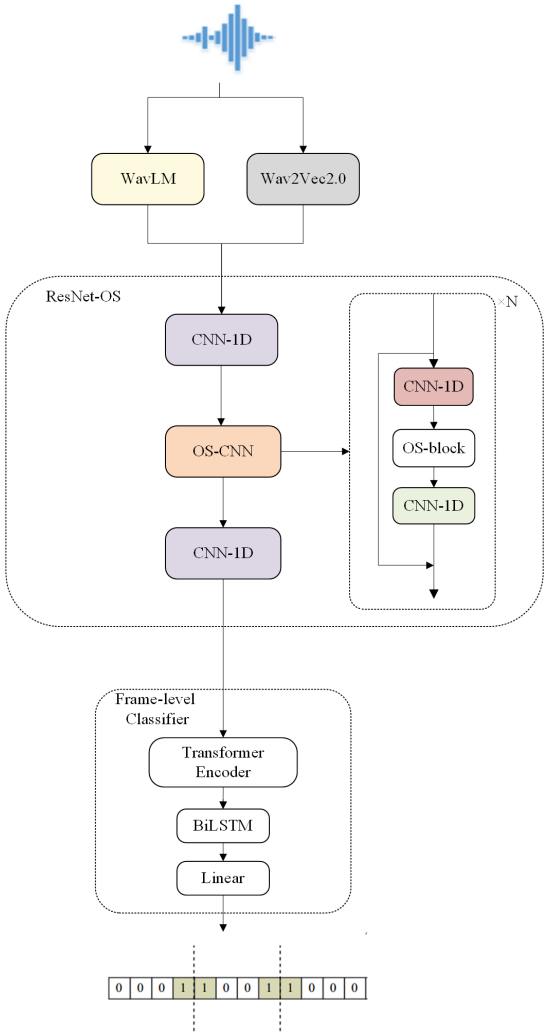


Fig. 1. The architecture of our proposed model

between forged and authentic audio, it is labeled as 1; otherwise, it is labeled as 0. To enhance the robustness of the system, frames close to the boundary are also marked as 1. This approach not only detects forged segments within the audio signal but also accurately pinpoints their locations.

A. Proposed Model

Figure 1 shows the design of our proposed model. We leverage the pretrained WavLM model[10] for feature extraction from the original audio. Subsequently, we utilize ResNet-OS to further extract frame-level embeddings. The combined Audio features and frame-level embeddings are sent into a Transformer-based classifier, which calculates the likelihood that each frame is a boundary.

B. Feature extraction

Wav2Vec[11]and WavLM[10]are two state-of-the-art self-supervised learning models that are primarily used for processing speech data. They have demonstrated exceptional performance in numerous downstream tasks, including automatic speech recognition (ASR)[12]. Consequently, we also employ these unsupervised models as feature extractors for original audio samples. We set the frame count T for both Wav2Vec and WavLM at 64, with an output feature frame rate of 20 milliseconds, culminating in a total feature dimension of 768.

C. Frame-level embedding extraction

After the feature extractor, we employ the ResNet-OS network to obtain frame-level embeddings, as illustrated in Figure 1. The ResNet-OS architecture includes two CNN-1D layers with N residual blocks sandwiched between them. Each residual block consists of two CNN-1D layers and an OS block, and features a residual connection that runs from the input to the output.

ResNet-OS is the core architecture of the model, consisting of N residual blocks, each configured as an OS-CNN. OS-CNN (Omni-Scale Convolutional Neural Network) is an architecture specifically designed for time series classification, based on the traditional 1D-CNN. This architecture incorporates OS block[13], enabling the model to effectively learn and extract features across various scales. Consequently, it excels in handling time series data characterized by complex patterns and dynamic changes. ResNet-OS includes two CNN-1D layers, between which N OS-CNN blocks are interspersed. Each OS-CNN block comprises two CNN-1D layers, with an OS block embedded between them. Moreover, the structure features a residual connection running from the input to the output to enhance learning capabilities and efficiency in information transfer. The architecture of the OS block, depicted in Figure 2, is a three-layer, multi-core convolutional network structure, where each convolutional layer employs multiple kernels and performs same-padding convolution operations. For configuring the kernel sizes, we use $p^{(i)}$ to represent the set of kernel sizes for the i-th layer:

$$P^{(i)} = \begin{cases} \{1, 2, 3, 5, \ldots, p^k\} & i \in \{1, 2\} \\ \{1, 2\} & i = 3 \end{cases} \tag{1}$$

This set can include various sizes, allowing each layer to capture features at different scales. Typically, smaller kernels are capable of capturing more detailed features, while larger kernels can capture broader features.The core concept of the OS block (Omni-Scale Block) revolves around how to cover all possible receptive field (RF) sizes by selecting convolutional kernels of different sizes. The receptive field size, which means the size of the input region that generates the feature, has always been an important component influencing the performance of 1D-CNN in time series classification tasks.

The set of receptive field sizes, S, represents the collective receptive field sizes of all paths. This is the description of how the receptive field set S is calculated:

$$S = \left\{ p^{(1)} + p^{(2)} + p^{(3)} - 2 \mid p^{(i)} \in \mathbb{P}^{(i)} \right\} \tag{2}$$

Based on Goldbach's conjecture, which states that any even integer can be stated as the sum of two primes, allows us to broaden the receptive field set:

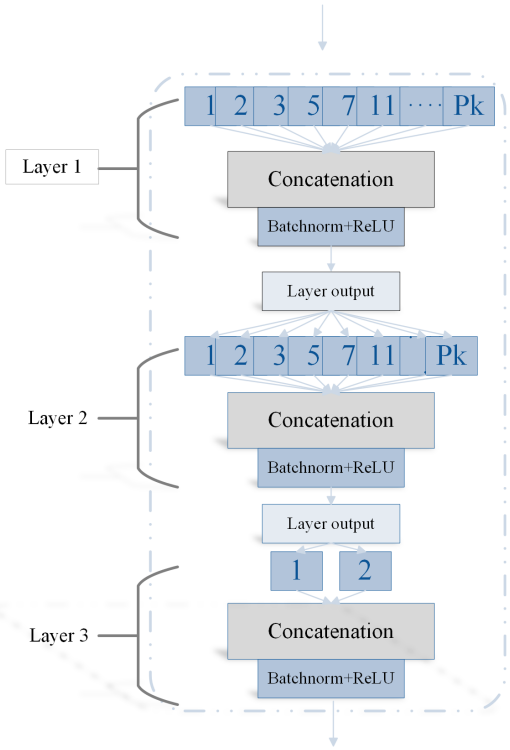


Fig. 2. Diagram of the OS block.

$$S = \left\{ e + p^{(3)} - 2 \mid p^{(3)} \in \mathbb{P}^{(3)}, e \in \mathbb{E} \right\} \tag{3}$$

where $\mathbb{E}$ is the set of all even receptive fields generated by the first two layers. Combining Equation (3) and Equation (1), we get:

$$S = \{e \mid e \in \mathbb{E}\} \cup \{e - 1 \mid e \in \mathbb{E}\} \equiv \mathbb{N}^{+} \tag{4}$$

$\mathbb{N}^{+}$ represents the set of all positive integers, indicating that through appropriate selection of convolutional kernels, it is possible to cover all receptive field (RF) sizes from 1 to the maximum possible size. Therefore, by properly choosing p^k, we can cover any range of integer receptive field sizes, thereby extracting more effective frame-level features.

D. Frame-level classifier

We use numerous Transformer encoders to record global context over vast distances across frames. We then further model the sequence embeddings produced by the Transformer encoders using BiLSTM. Ultimately, the boundary probability for every frame is estimated by a completely connected layer.

III. EXPERIMENT

A. Data Preparation

Our training data comes from the Half-truth Audio Detection (HAD) dataset[5]. This type of audio only modifies a few words in the original speech, posing a serious threat to audio verification systems, as these small-scale modifications are often difficult to detect. Table I provides statistics on the datasets used in our experiments. As shown in Table I,The HAD-train dataset consists of 26,554 real and 26,554 fake utterances, while the HAD-dev dataset contains 8,914 real and 8,914 fake utterances. All fake utterances in the HAD-train and HAD-dev datasets are generated using mainstream speech synthesis techniques. The HAD-test dataset includes 9,072 labeled utterances.

TABLE I
THE STATISTICS OF DATASETS (#UTTERANCES)

Name	Bona fide	Fake	ALL
HAD-train	26554	26554	53108
HAD-dev	8914	8914	17824
HAD-test	-	-	9072

The dataset for Half-truth Audio Detection (HAD) is derived from the AISHELL-3 corpus[14]. The Apache 2.0 license applies to this publicly accessible dataset. AISHELL-3 is a Mandarin voice corpus with many speakers that is utilized for text-to-speech (TTS) model training. We inserted audio segments into real speech according to the following strategy:

1. Named entity recognition and lexical annotation using jieba, followed by random replacement or substitution of these keywords with antonyms in order to change the semantics and emotional expression of the original utterance.

2. The (GST)-based Tacotron[15], [16] system is used to generate audio corresponding to the edited text, which is then processed by the neural vocoder LPCNet to enhance the naturalness and emotional expression of the synthesized audio and to ensure that its sound quality is consistent with the original recording.

3. The precisely processed synthesized audio is inserted into a specific position of the original audio, and the volume and sound quality are adjusted to ensure a seamless auditory integration and consistency between the synthesized part and the original audio.

As the final training data for the model, we merge some of the fake dataset with the actual discourse from the HAD-train dataset. Based on the HAD-test set, it is evaluated by combining it with the HAD-dev set as an adaptation dataset.

B. Parameter settings

ResNet-OS consists of 12 OS-CNN blocks, each containing a CNN-1D layer without bias, with a kernel size of 1 and both input and output sizes set to 512. The first CNN-1D operates without bias at a kernel size of 5, with an input size of 768 and an output size of 512. The final CNN-1D has a kernel size of 1, input size of 512, and outputs frame-level embeddings with an embedding size set to 128.

Four attention heads and two encoder layers make up the Transformer encoder of the frame-level classifier, which has a feed-forward network (FFN) size of 1024. The BiLSTM has 128 hidden units, and an activation function for ReLU comes after it. Lastly, each frame's probability is predicted using a 256-dimensional fully linked network.

C. Training Process

During the training process, genuine and fake utterances are considered identical since there is no boundary between them. We randomly select an audio sample from the partially fake dataset and a genuine sample from HAD-train, with a 0.5 probability of choosing a positive sample to maintain data balance. Each audio sample is trimmed to a fixed length,for instance, 0.64 seconds, 1.28 seconds, or 2.56 seconds. For online data augmentation, we leverage the MUSAN and RIR corpora. Positive samples' labels are set to all zeros. For negative samples, the label at the boundary between actual and fake audio clips is set to one, while the others are zero. Labels near the boundaries are likewise set to one.

In our experiments, the Adam optimizer and binary cross-entropy loss are used to train the model over 100 epochs. The learning rate is set at 10^{-4} and the batch size is set at 64. We assess the equal error rate (EER) on the test sets during training. For inference and assessment, we use the mean of the five models with the lowest EER for the wav2vec-based model. However, averaging models based on wavlm resulted in decreased performance. Therefore, for the wavlm-based model, we only consider the model with the best performance for evaluation.

IV. RESULTS AND DISCUSSION

Our research compared two systems: one trained with the Wav2Vec feature extractor and another with the WavLM feature extractor. The results are presented in Table II ,indicate that an audio length of 1.28 seconds offers the best performance for tasks using WavLM and Wav2Vec. This may be because this duration is sufficient to capture adequate contextual information without introducing noise or unnecessary details due to excessive length. Using Wav2Vec features, the model achieves an Equal Error Rate (EER) of 0.08% on the HAD-dev set and 0.067% on the HAD-test set at 1.28 seconds. For the WavLM model at the same duration, the EER on HAD-dev is 0.06%, and on HAD-test it is 0.023%, with performance degradation observed at audio lengths of 0.64 seconds and 2.56 seconds. These results suggest that the model trained with WavLM features exhibits superior generalization in this task compared to Wav2Vec.

TABLE II
SYSTEM PERFORMANCES REGARDING VARIOUS SEGMENT L, REPORTED IN EER.

Feature	**Test Sets**	**Wav length l (s)**		
		0.64s	1.28s	2.56s
Wav2Vec (w/o OS block)	HAD-dev	0.35%	**0.08%**	0.42%
	HAD-test	0.20%	**0.067%**	0.18%
WavLM	HAD-dev	0.19%	**0.06%**	0.28%
	HAD-test	0.12%	**0.023%**	0.10%

In this study, we adjusted the waveform length l to 1.28 seconds for models based on Wav2Vec and WavLM. We also conducted a series of ablation studies to analyze the performance impact of individual network components. As shown in Table III , We deleted specific components from the proposed network and documented their performance. Specifically, the"w/o ResNet-OS" model indicates the removal of the ResNet-OS component used for frame-level embedding extraction. For the Wav2Vec model, the equal error rate (EER) increased to 0.16% on the HAD-dev dataset and to 0.15% on the HAD-test dataset after removing ResNet-OS, demonstrating its critical role in model performance. Similarly, for the WavLM model, removing ResNet-OS resulted in an EER of 0.10% on HAD-dev and 0.05% on HAD-test. Removing the OS block lead to an EER of 0.073% on HAD-dev and 0.035% on HAD-test for the WavLM model, proving that the OS block also plays a significant role in enhancing model performance.

Overall, the WavLM model showed a better error rate performance compared to Wav2Vec, especially evident in the test set. These findings emphasize the critical role of the OS block and ResNet-OS structure in enhancing the accuracy of partial fake audio detection tasks, and removing these key components significantly reduces model performance.

TABLE III
PERFORMANCES (EER) REGARDING VARIOUS FEATURE EXTRACTOR AND ARCHITECTURE DESIGNS, W/O MEANS WITHOUT.

Model	**HAD-dev**	**HAD-test**
Wav2Vec	0.08%	0.067%
w/o ResNet-OS	0.16%	0.15%
WavLM	**0.06%**	**0.023%**
w/o ResNet-OS	0.10%	0.05%
w/o OS block	0.073%	0.035%

V. CONCLUSIONS

We designed a frame-level partial fake audio detection approach that not only delivers binary judgments at the utterance level for partially faked audio, but also accurately forecasts the insertion or replacement positions of false clips. By conducting ablation studies on our model components and evaluating various acoustic features, including Wav2Vec and WavLM, our experimental results demonstrate that this system outperforms existing boundary detection systems in detecting the boundaries of partially spoofed audio.

VI. ACKNOWLEDGEMENTS

This work was funded by the Natural Science Foundation of Xinjiang Uygur Autonomous Region of China(No.2022D01C59);NSFC(No.12464060).

REFERENCES

[1] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[2] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[3] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17 022–17 033.

[4] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: From the perspective of asvspoof challenges," *APSIPA*, vol. 9, 2020.

[5] J. Yi, Y. Bai, J. Tao, *et al.*, "Half-truth: A partially fake audio detection dataset," in *Proceedings of Interspeech*, 2023, pp. 1654–1658.

[6] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino, and N. Evans, "An initial investigation for detecting partially spoofed audio," in *Proceedings of Interspeech*, 2021, pp. 4264–4268.

[7] J. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 9241–9245.

[8] Z. Cai, W. Wang, and M. Li, "Waveform boundary detection for partially spoofed audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[9] H. Wu, H.-C. Kuo, N. Zheng, *et al.*, "Partially fake audio detection by self-attention-based fake span discovery," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 9236–9240.

[10] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[12] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 826–27 839, 2021.

[13] W. Tang, G. Long, L. Liu, T. Zhou, M. Blumenstein, and J. Jiang, "Omni-scale cnns: A simple and effective kernel size configuration for time series classification," 2022.

[14] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, *Aishell-3: A multispeaker mandarin tts corpus and the baselines*, arXiv preprint arXiv:2010.11567, 2020.

[15] R. J. Skerry-Ryan, E. Battenberg, X. Ying, Y. Wang, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[16] Y. Wang, D. Stanton, and Y. Zhang, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

NeRF-FCM: Attention-based Feature Calibration Mechanisms for 3D Object Detection Using NeRF

Hana Lebeta Goshu*, Jun Xiao*, Kin-Chung Chan*, Cong Zhang*, Mulugeta Tegegn Gemeda† and Kin-Man Lam*
* The Hong Kong Polytechnic University, Hong Kong
{hana-lebeta.goshu, jun.xiao, alfred.chen, cong-clarence.zhang}@connect.polyu.hk, enkmlam@polyu.edu.hk
† Jimma Institute of Technology, Ethiopia
mulugeta.geneda@ju.edu.et

***Abstract*—With the fast development of 3D vision, 3D object detection based on posed RGB images has become increasingly popular and attracted significant attention from researchers in recent years. Given the remarkable performance of Neural Radiance Field (NeRF) in modeling 3D scenes, recent 3D detection methods utilizing posed RGB images generated by NeRF models have achieved promising results. However, NeRF-based models often suffer from poor generalization and are prone to generating inconsistent image content for unseen views, which inevitably degrades the performance of existing NeRF-based 3D detectors. In this paper, we propose an effective feature calibration method to enhance the performance of 3D detection models based on posed RGB images produced by NeRF models. Specifically, our proposed method efficiently recalibrates the 3D features extracted from the backbone network, and adaptively computes the weights for fusion based on the statistical properties of the features. Experiments show that our method significantly outperforms the baseline model, achieving improvement of +8.6 AP@0.5, +5.5 AP@0.5, and +5.1 AP@0.5 on the Hypersim, 3D-FRONT, and ScanNet benchmarks, respectively, with anchor-free heads. Particularly, compared with the baseline model, our method can more accurately predict the 3D bounding boxes in 3D space, even when objects are poorly reconstructed by NeRF while keeping low computational costs with a minimal increase in model complexity.**

***Index Terms*—3D Object Detection, NeRF, Channel Attention, Multi-View**

I. Introduction

3D object detection is a foundational topic in the 3D vision tasks, attracting significant attention from researchers in recent years due to its substantial industrial value in applications such as robotics [1], autonomous driving [2], Virtual Reality (VR) [3], Augmented Reality (AR) [4], etc. Most existing 3D object detection methods rely on point clouds [5]–[8] and depth images [9] because these data types provide precise geometric information. However, obtaining highly accurate geometric data typically requires costly 3D sensors, including laser scanners, depth cameras, and stereo cameras, which can be impractical for real-world applications. To reduce high costs, 3D object detection methods based on posed RGB images have been proposed [10], [11]. These methods leverage monocular techniques and have shown promising results. However, monocular-based approaches are inevitably degraded when faced with limited fields of view, occlusion, and scale uncertainty. To address these issues, ImVoxelNet [12] adopts a multi-view image for 3D object detection, effectively learning feature representations of 3D voxel volumes from multi-view 2D images. Nonetheless, this method struggles with severe occlusion and complex geometric information.

Recently, Neural Radiance Field (NeRF) [13] has demonstrated remarkable capabilities in 3D scene representation and novel view synthesis. Inspired by this, several 3D object detection methods based on NeRF have been proposed, which achieve promising performance [14]–[17]. For example, NeRF-RPN [14] predicts the location of objects in 3D scenes using posed RGB images rendered by NeRF. Specifically, it takes NeRF outputs (i.e., color and density information) as input to produce 3D bounding boxes. During feature extraction, the feature maps received by the Region Proposal Network (RPN) head are highly biased towards capturing local contexts, neglecting global information. This leads to regions of interest that rely solely on local spatial information, further deteriorating detection accuracy in complex 3D scenes.

In this paper, we propose a robust method to incorporate global information into NeRF-based 3D detection methods, thereby enhancing detection performance. In the human visual system, global context (e.g., surroundings and background) is crucial for object recognition, but previous methods have predominantly focused on local features. In contrast, our method effectively fuses global and local features by employing a 3D Global Average Pooling (GAP) mechanism for a channel-wise attention mechanism. Additionally, inspired by ECA-Net [18], we propose an efficient method to learn the inter-dependency between global and local features with negligible additional cost. Consequently, our method can adaptively aggregate global and local features, learning effective 3D feature representations to enhance the accuracy of 3D object detection.

The contributions of this paper are summarized as follows:

- We demonstrate that incorporating global information can enhance the quality of NeRF-based 3D object detection from posed RGB images.
- We propose a novel framework for NeRF-based 3D object detection that utilizes a channel attention module to improve 3D feature representations and produce accurate 3D bounding boxes with minimal computational costs.
- Our experiments show that our proposed method can significantly enhance 3D object detection accuracy and suppress an existing baseline by +8.6 AP@0.5, +5.5 AP@0.5,

and +5.1 AP@0.5 on the Hypersim, 3D-FRONT, and ScanNet indoor datasets, respectively.

II. Related Works

A. Point Cloud-based 3D Object Detection

3D object detection from point clouds has achieved great results in outdoor and indoor scenes because point clouds provide reliable geometric structural information. It is divided into two types: grid-based and point cloud-based approaches. Grid-based approaches, such as VoxelNet [5] and PointPillars [6], typically transform point clouds into 3D grids and use 3D CNNs to process the grid representation. However, this approach is memory-intensive due to the dense volumetric representation and the use of 3D CNNs. To mitigate this problem, sparse fully-convolutional detection methods, such as FCAF3D [19] and [20], have been introduced, improving the quality of 3D object detection in both accuracy and efficiency. Conversely, point cloud-based methods detect 3D objects from the point clouds [7], [8]. These methods utilize deep Hough voting, as proposed in VoxelNet [7], to predict bounding box parameters from point features. However, voting-based 3D detection methods pose challenges when used with reconstruction networks, since they need extra data annotation on the point clouds. Recent point cloud-based methods have attracted significant attention owing to their strong performance in 3D detection, particularly with the emergence of various 3DGS methods [21], [22], eliminating the need for expensive 3D sensors. Methods that incorporate multi-views to enhance scene understanding have also become more prominent. ImVoxelNet [12] has achieved impressive results in indoor 3D detection using the 3D voxel-based feature volume [23]. Nevertheless, it fails to retain the inherent geometry of input scenes while constructing the feature volume.

B. Channel Attention and Object Detection

In the field of computer vision, channel attention methods are essential for improving the performance of deep CNNs across various tasks [18], [24]–[27]. The main challenge for convolutional networks is their limited ability to extract global features, as the convolution kernels mainly focus on local regions in space. To address this issue, channel attention methods have been investigated and widely used to enhance the representational power of networks by aggregating global information. The Squeeze-and-Excitation (SE) block in a SE-Net [24] has become a paradigm for channel attention, which assigns learned weights to different channels of convolutional layers and recalibrates the feature maps by capturing channel-wise dependencies, thus improving detection performance. CBAM [25] uses convolutions with large-size kernels to encode spatial information. ECA-Net [18], based on SE-Net, employs 1D convolutions to capture local cross-channel interaction while resolving the negative effects of channel dimension reduction. However, these techniques are mainly used for 2D object detection. To improve the quality of 3D object detection using NeRF from posed RGB images, this paper presents the 3D Squeeze and Excitation Channel Attention (3D SE-CA) and 3D Efficient Channel Attention (3D ECA) modules.

III. Methods

A. Overview

The architecture of our proposed method is illustrated in Fig. 1. Following NeRF-RPN [14], our method uses a pre-trained NeRF model and uniformly samples the volume density and RGB color information to generate a feature volume. To extract the feature maps, we employ a 3D backbone network. Before passing the extracted feature maps to the Feature Pyramid Network (FPN) [28], we integrate channel attention mechanisms to enhance feature representation. Finally, the RPN head uses the feature pyramid to produce a set of regional proposals.

B. Sampling Input and Extracting Features

To obtain accurate 3D geometric information, we leverage the capabilities of a pre-trained NeRF model. Thus, our network directly produces 3D bounding boxes of Regions of Interest (ROIs) in 3D space by using the extracted 3D volumetric information from NeRF as input. The extracted radiance and density volume field are uniformly sampled in order to build a feature volume. Each sample can be represented as (r, g, b, α), where c = (r, g, b) represents the emitted averaged color and α, the opacity, can be obtained as follows [14]:

$$\alpha = \text{clip}\left(1 - \exp(-\rho \cdot \delta), 0, 1\right), \tag{1}$$

where ρ is the volume density and δ represents the distance between two adjacent points, set to 0.01.

We employ a 3D backbone network to extract feature maps after sampling inputs on the 3D grid and extracting RGB and volume densities. Subsequently, we apply our 3D channel attention methods to further extract channel-wise global information and recalibrate the weight of each channel to boost feature representation. We embed 3D SE-CA and 3D ECA between the 3D backbone and 3D FPN to investigate the effect of adding these channel attention methods and determine the optimal placement for best performance.

The 3D FPN takes enhanced feature maps from the channel attention module to produce multi-scale feature volumes, addressing scale variation issues in images and improving the ability of the model to accurately detect objects of various sizes. The 3D FPN exploits all layers of the backbone network, i.e., $C2, C3, C4$, and $C5$, to produce a feature pyramid comprising $P2, P3, P4$, and $P5$. We apply the channel attention module to the output of each bottom-up feature map, i.e., $C2, C3, C4$, and $C5$, before passing them to each top-down feature map, i.e., $P5, P4, P3$, and $P2$, as illustrated in Fig. 1. Thus, the $P2, P3, P4$, and $P5$ feature maps acquire global information from the channel attention module, which is then passed to each layer of a top-down pathway of the 3D FPN. Finally, the 3D RPN takes the feature pyramid produced by the 3D FPN block as input to generate 3D bounding boxes.

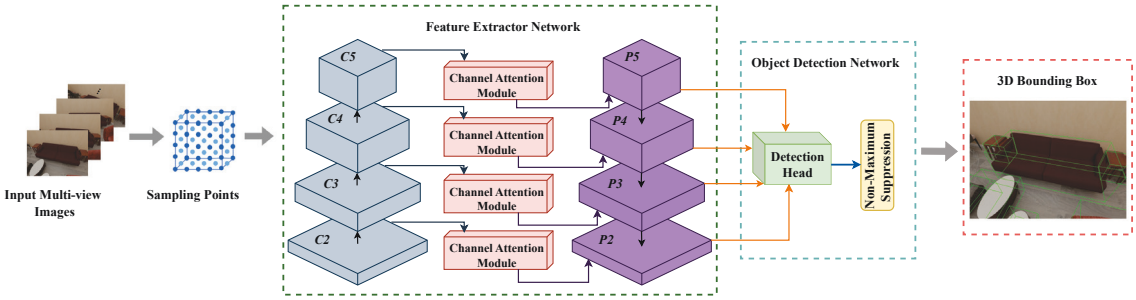


Fig. 1. Overview of our proposed method. Given multi-view images, NeRF retrieves 3D geometric information to improve the 3D detection pipeline. We use VGG19 to extract 3D feature maps and apply a channel attention module on each feature to learn channel importance with global information. The 3D FPN takes both global and local features to produce a feature pyramid, which is subsequently utilized by the detection head to produce accurate 3D bounding boxes.

C. 3D Squeeze and Excitation Channel Attention Module

The proposed 3D SE-CA module extends the 2D SE blocks presented in [24] to a 3D version (see Fig. 2), focusing on channels with more informative features in the 3D features map to improve 3D object detection performance. In the SE block, the transformation F_{se} is divided into the squeeze operation F_{sq} and the excite operation F_{ex}. In F_{sq}, a $1 \times 1 \times 1 \times C$ sized feature map is generated to obtain channel-wise global information statistics while keeping the number of channels C constant using 3D GAP. The output is produced by squeezing input U through the spatial dimensions of $D \times H \times W$ and indicated by $z_i = F_{sq}(u_i)$, which belongs to $Z \in R^C$. The F_{sq} is expressed as follows:

$$z_i = \frac{1}{(D \times H \times W)} \sum_{i=1}^{D} \sum_{j=1}^{H} \sum_{k=1}^{W} u_i(i,j,k), \tag{2}$$

where D, H, and W denote the depth, height, and width, respectively. We set $U = F_{tr}(X)$, where X is an input feature and $U = [u_1, u_2, \ldots, u_c]$, $u_i \in R^{D \times H \times W}$, is the output feature. The channel-wise output of F_{sq} modulates the inter-dependencies of all channels via excitation. The F_{ex} receives Z and learns the inter-channel interactions leveraging two fully connected layers, as follows:

$$s_i = \sigma(W_2 \delta(W_1 Z)), \tag{3}$$

where σ represents sigmoid activation function, δ denotes ReLU function [29], $W_1 \in R^{\frac{C}{r} \times C}$ includes parameters for dimensionality-reduction, and $W_2 \in R^C \times \frac{C}{r}$ includes parameters for dimensionality-increasing layers and $r = 16$. A channel-wised multiplication F_{scale} between each feature map u_i and excitation scalar s_i is ultimately employed to produce final re-scaling feature output U, as follows:

$$U = F_{scale}(u_i, s_i) = u_i \cdot s_i. \tag{4}$$

D. 3D Efficient Channel Attention Module

We extend the original 2D ECA-Net [18] to the 3D version and integrate it with NeRF-RPN [14] for 3D object detection. This module captures local cross-channel interactions while

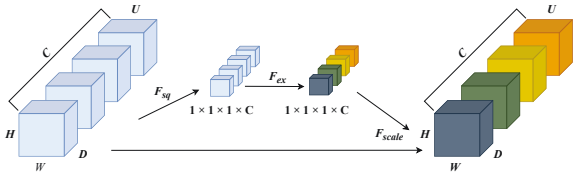


Fig. 2. The 3D Squeeze and Excitation Channel Attention Module.

alleviating the negative impact of channel dimension reduction by using a 1D convolution with adaptive kernel sizes, replacing traditional fully connected layers. This process includes GAP, followed by 1D convolution and sigmoid activation to modulate an input feature map (see Fig. 3). The output from a convolution block is denoted as $x \in R^{W \times H \times D}$, where W, H, and D are width, height, and depth, respectively. The feature map is forwarded to a 3D ECA block to enhance the network representation ability by applying the channel attention mechanism. In the 3D ECA module, the 3D GAP of the channel dimension is computed as follows:

$$g(x) = \frac{1}{(HWD)} \sum_{i=1,j=1,k=1}^{H,W,D} x_{i,j,k}, \tag{5}$$

where $g(x)$ represents features after global average pooling. Moreover, channel weights are calculated as follows:

$$\omega = \sigma(W_p g), \tag{6}$$

where σ denotes a sigmoid function and W_p is a parameter matrix for the channel attention, calculated as follows:

$$\begin{bmatrix} w^{1,1} & \ldots & w^{1,p} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,p+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-p+1} & \cdots & w^{C,C} \end{bmatrix}, \tag{7}$$

where W_p involves $p \times C$ parameters. Based on (7), the weight of g_i is computed as follows:

$$w_i = \sigma\Big(\sum_{j=1}^{p} w^j g^j\Big), \quad g^j \in \Omega_i^p, \tag{8}$$

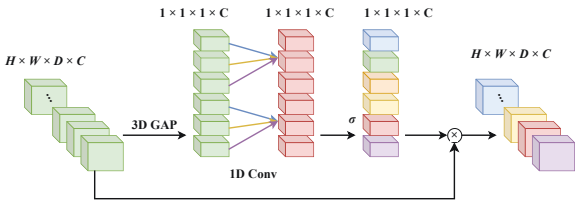


Fig. 3. The 3D Efficient Channel Attention Module.

where Ω_i^p represents the p-set of g_i adjacent channels. Alternatively, the above computation can be performed using $1D$ convolutional operation with the kernel size of p. Thus, the channel features in the g_x are derived as follows:

$$\omega = \sigma\left(\mathrm{C1D}_p(g_x)\right), \tag{9}$$

where C1D denotes one dimension convolution and $p = 3$.

E. 3D Detection Head and Loss Functions

We utilize a 3D RPN head to produce object proposals from the feature pyramid and adopt the loss functions from [14]. We conduct experiments on anchor-free and anchor-based methods to compare the results with the baseline method. We utilize the same loss function as in [30] for the anchor-based method:

$$\begin{aligned} L\left(\{p_i\},\{t_i\}\right) = \frac{1}{N_{cls}} \sum_i L_{cls}\left(p_i, p_i^*\right) \\ + \frac{\lambda}{N_{\text{reg}}} \sum_i p_i^* L_{reg}\left(t_i, t_i^*\right), \end{aligned} \tag{10}$$

where i represents an anchor index, t_i stands for box offsets, and t_i^* denotes the ground-truth bounding boxes. p_i and p_i^* correspond to predicted objectness and ground-truth label, respectively. L_{cls} and L_{reg} denote the classification loss and smooth L_1 loss, respectively, as presented in [31]. N_{cls} and N_{reg} represent the number of anchors utilized in the classification and regression loss computation, respectively, and λ is used to balance these two losses. For anchor-free methods, the centerness loss is added to obtain the final loss function [32]:

$$\begin{aligned} L\left(\{p_i\},\{\boldsymbol{t}_i\},\{c_i\}\right) = \frac{1}{N_{\text{pos}}} L_{cls}\left(p_i, p_i^*\right) \\ + \frac{\lambda}{N_{\text{pos}}} p_i^* L_{reg}\left(\boldsymbol{t}_i, \boldsymbol{t}_i^*\right) + \frac{1}{N_{\text{pos}}} p_i^* L_{ctr}\left(c_i, c_i^*\right), \end{aligned} \tag{11}$$

where N_{pos} represents the number of positive samples, λ denotes balance factor, L_{cls} is the focal loss, L_{reg} is the IoU loss for rotated boxes [33], and L_{ctr} is the centerness loss.

IV. Experiments and Analysis

A. Experimental Setup

In our experiments, we mainly follow NeRF-RPN [14], including backbones, detection head, training recipe, datasets *etc*. Three indoor datasets, Hypersim [34], 3D-FRONT [35], and ScanNet [36], which were constructed as NeRF datasets for 3D object detection by [14], are used. We utilize the NeRF implementation from Instant-NGP [37] to reconstruct them.

Hypersim is the photorealistic synthetic dataset for indoor scene understanding, containing 77,400 images of 461 indoor scenes. For NeRF training in 3D object detection, [14] thoroughly cleaned data based on NeRF reconstruction quality and object annotation usefulness, retaining about 250 scenes.

3D-FRONT is the large-scale dataset with professionally designed synthetic indoor scenes that contain room layouts and textured 3D models of furniture. From the dataset, 159 rooms and their bounding boxes were manually chosen, cleaned, and rendered to generate the NeRF dataset for 3D object detection.

ScanNet is the widely used dataset for indoor 3D detection, containing 2.5 million views from 1513 RGB scans. For each scene, video frames were divided into 100 bins, and bounding boxes were computed from annotated meshes [14]. Models were trained using a given depth and depth-guided NeRF [38].

We use AdamW [39] optimizer with the weight decay of 10^{-3} and the initial learning rate of 3×10^{-4}. For Equations (10) and (11), λ is set to 5.0 and 1.0, respectively. During testing, the top 2,500 proposals are chosen, and redundant proposals are eliminated using Non-Maximum Suppression (NMS), based on rotated IoU with a threshold of 0.1. We train our models using NVIDIA RTX 4090 GPUs with various model configurations on the three datasets. During inference, our method achieves high accuracy with a lower computational cost, with a run time of 61 seconds compared to 67 seconds for the baseline, on the 3D-FRONT dataset. On the Hypersism dataset, the run time of our method is 139 seconds, nearly equivalent to 138 seconds for the baseline on a single GPU, while it significantly enhances 3D object detection performance. We evaluate the methods mainly by average precision (AP) and recall (R) scores with 0.25 IoU and 0.5 IoU thresholds.

B. Experimental Results

We compare our proposed method with existing baseline methods [12], [14], [19], as depicted in Table I. Using VGG19 as the backbone network, with anchor-free and anchor-based RPN heads, we observe that the proposed channel attention methods, 3D SE-CA and 3D ECA, outperform the baseline NeRF-RPN by achieving higher recall scores and AP on all three datasets. This implies that channel attention is beneficial, and our method can effectively enhance 3D object detection performance. Specifically, 3D ECA has made significant improvements compared to the baseline in AP@0.5 on Hypersim (+8.6), 3D-FRONT (+5.5), and ScanNet (+5.1) with anchor-free head. The improvement, particularly in the AP@0.5 evaluation metric, shows that our method is capable of better integrating global information from multi-view and adaptively recalibrating channel weights, thus improving the quality of predicted 3D bounding boxes. Channel-wise interactions have significantly boosted feature representation.

We present visualizations of the predicted 3D bounding boxes generated by NeRF-RPN and our method with anchor-free head, as depicted in Fig. 4. We note that our method provides accurate detection predictions even on objects poorly

TABLE I

QUANTITATIVE COMPARISON WITH EXISTING METHODS ON 3D-FRONT, SCANNET AND HYPERSIM. THE BEST PERFORMANCES OF DIFFERENT EVALUATION METRICS ARE HIGHLIGHTED IN BOLD, WHILE DIFFERENT SETTINGS ARE IN BLUE. R INDICATES A RECALL SCORE AT AN IOU THRESHOLD.

Methods	3D-FRONT				ScanNet				Hypersim			
	R@0.25	R@0.5	AP@0.25	AP@0.5	R@0.25	R@0.5	AP@0.25	AP@0.5	R@0.25	R@0.5	AP@0.25	AP@0.5
ImVoxNet [12]	88.3	71.5	86.1	**66.4**	51.7	20.2	37.3	9.8	19.7	5.7	9.7	2.3
FCAF3D [19]	89.1	56.9	73.1	35.2	90.2	42.4	**67.7**	18.5	47.6	19.4	30.7	8.8
NeRF-RPN (anchor-based)	97.8	76.5	65.9	43.2	88.7	42.4	40.7	14.4	57.1	14.9	11.2	1.3
Ours (3D SE-CA)	**98.5**	**82.4**	71.8	52.6	86.2	38.9	41.2	15.5	60.3	18.1	12.1	2.4
Ours (3D ECA)	97.8	79.4	71.9	53.3	88.2	41.4	40.6	16.2	57.1	19.4	14.8	2.9
NeRF-RPN (anchor-free)	96.3	69.9	85.2	59.9	89.2	42.9	55.5	18.4	66.7	27.3	30.9	11.5
Ours (3D SE-CA)	97.8	75.7	86.4	64.5	89.7	43.8	56.9	**23.5**	**71.4**	**31.8**	33.4	**20.1**
Ours (3D ECA)	97.1	77.2	**86.9**	65.4	**91.6**	**44.3**	58.8	23.4	70.2	30.2	**35.4**	18.4



Fig. 4. Qualitative results of 3D object detection on the 3D-FRONT dataset. Compared to the baseline method NeRF-RPN, our method demonstrates superior performance in accurately predicting the 3D bounding boxes in 3D space.

reconstructed by NeRF in 3D space (see last row). 3D geometry in NeRF plays a significant role in region proposal tasks and poor NeRF reconstruction can seriously impair the prediction. However, our method can still detect them accurately compared to the baseline method.

C. Ablation Studies

We conducted the ablation study to examine the effect of adding a channel attention module to the output of each bottom-up feature map from the backbone network before passing them to corresponding top-down feature maps. To determine the optimal placement, we tested the module between various combinations of feature maps: $\{C5\}$ and $\{P5\}$, $\{C4, C5\}$ and $\{P4, P5\}$, $\{C3, C4, C5\}$ and $\{P3, P4, P5\}$, and finally among all feature maps, as shown in Table II. From the experiment results, it is evident that the best performance is attained when the channel attention module is located between all feature map levels $\{C2, C3, C4, C5\}$ and $\{P2, P3, P4, P5\}$. This configuration allows feature maps $\{P2, P3, P4, P5\}$ to effectively acquire channel-wise global information, thus enhancing detection accuracy.

TABLE II

ABLATION STUDY OF ADDING CHANNEL ATTENTION MODULE AT DIFFERENT LEVELS OF FEATURES BETWEEN THE 3D VGG19 BACKBONE AND 3D FPN ON 3D-FRONT DATASET. LEVELS 1, 2, AND 3 REPRESENT THE NUMBER OF LAYER COMBINATIONS WE UTILIZED.

Methods	3D-FRONT			
	R@0.25	R@0.5	AP@0.25	AP@0.5
3D ECA LEVEL1	97.0	73.5	83.6	60.0
3D ECA LEVEL2	97.0	74.3	86.0	63.6
3D ECA LEVEL3	96.3	75.7	86.0	63.8
Ours	**97.1**	**77.2**	**86.9**	**65.4**

V. CONCLUSIONS

In this paper, we proposed effective attention-based feature calibration mechanisms to enhance the performance of NeRF-based 3D detection from multi-view RGB images. Specifically, we investigated 3D SE-CA and 3D ECA modules, to incorporate global information into 3D feature maps extracted from the backbone network to enhance feature representation. The downstream detection network based on the enhanced features, with more discriminative power, achieves better performance in the challenging domain of 3D space. Our experimental results demonstrate that our method outperforms the baseline NeRF-RPN on all three datasets: Hypersim by 8.6 AP@0.5, 3D FRONT by 5.5 AP@0.5, and ScanNet by 5.1 AP@0.5, with an anchor-free RPN head. These results not only illustrate the importance of channel attention mechanisms in 3D object detection tasks, but also provide valuable insights into the critical factors to achieve optimal performance in this area.

REFERENCES

[1] J. Sun, Y. Xu, M. Ding, *et al.*, "Nerf-loc: Transformer-based object localization within neural radiance fields," *IEEE Robotics and Automation Letters*, 2023.

[2] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: A comprehensive survey," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 1909–1963, 2023.

[3] M. Martini, F. Solari, and M. Chessa, "Obstacle avoidance and interaction in extended reality: An approach based on 3d object detection," in *ICIAP*, 2023.

[4] W. Lee, N. Park, and W. Woo, "Depth-assisted real-time 3d object detection for augmented reality," in *ICAT*, vol. 11, 2011, pp. 126–132.

[5] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *CVPR*, 2018.

[6] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *CVPR*, 2019.

[7] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *ICCV*, 2019.

[8] C. R. Qi, X. Chen, O. Litany, and L. J. Guibas, "Imvotenet: Boosting 3d object detection in point clouds with image votes," in *CVPR*, 2020.

[9] X. Shen and I. Stamos, "Frustum voxnet for 3d object detection from rgb-d or depth images," in *WACV*, 2020.

[10] T. Wang, J. Pang, and D. Lin, "Monocular 3d object detection with depth from motion," in *ECCV*, 2022.

[11] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *ICCV*, 2019.

[12] D. Rukhovich, A. Vorontsova, and A. Konushin, "Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection," in *WACV*, 2022.

[13] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[14] B. Hu, J. Huang, Y. Liu, Y.-W. Tai, and C.-K. Tang, "Nerf-rpn: A general framework for object detection in nerfs," in *CVPR*, 2023.

[15] C. Xu, B. Wu, J. Hou, *et al.*, "Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection," in *CVPR*, 2023.

[16] C. Huang, Y. Hou, W. Ye, *et al.*, "Nerf-det++: Incorporating semantic cues and perspective-aware depth supervision for indoor multi-view 3d detection," *arXiv preprint arXiv:2402.14464*, 2024.

[17] C. Huang, X. Li, S. Zhang, L. Cao, and R. Ji, "Nerf-dets: Enhancing multi-view 3d object detection with sampling-adaptive network of continuous nerf-based representation," *arXiv preprint arXiv:2404.13921*, 2024.

[18] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *CVPR*, 2020.

[19] D. Rukhovich, A. Vorontsova, and A. Konushin, "Fcaf3d: Fully convolutional anchor-free 3d object detection," in *ECCV*, 2022.

[20] J. Gwak, C. Choy, and S. Savarese, "Generative sparse detection networks for 3d single-shot object detection," in *ECCV*, 2020.

[21] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[22] K.-C. Chan, J. Xiao, H. L. Goshu, and K.-m. Lam, "Point cloud densification for 3d gaussian splatting from sparse input views," in *ACM Multimedia 2024*.

[23] Z. Murez, T. Van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3d scene reconstruction from posed images," in *ECCV*, 2020.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018.

[26] J. Xiao, Z. Lyu, C. Zhang, Y. Ju, C. Shui, and K.-M. Lam, "Towards progressive multi-frequency representation for image warping," in *CVPR*, 2024.

[27] J. Xiao, Q. Ye, T. Liu, C. Zhang, and K.-M. Lam, "Deep progressive feature aggregation network for multi-frame high dynamic range imaging," *Neurocomputing*, vol. 594, p. 127 804, 2024.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

[29] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[31] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.

[32] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *ICCV*, 2019.

[33] D. Zhou, J. Fang, X. Song, *et al.*, "Iou loss for 2d/3d object detection," in *3DV*, 2019.

[34] M. Roberts, J. Ramapuram, A. Ranjan, *et al.*, "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in *ICCV*, 2021.

[35] H. Fu, B. Cai, L. Gao, *et al.*, "3d-front: 3d furnished rooms with layouts and semantics," in *ICCV*, 2021.

[36] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*, 2017.

[37] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[38] B. Roessle, J. T. Barron, B. Mildenhall, P. P. Srinivasan, and M. Nießner, "Dense depth priors for neural radiance fields from sparse input views," in *CVPR*, 2022.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

Predictive Analysis of Driver Drowsiness Progression: Multi-Level Drowsiness Classification Using Physiological Signals

Natchira Dachoponchai [1], Yodchanan Wongsawat [1], and Jetsada Arnin [1*]
[1] Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand
E-mail: natchira.dac@student.mahidol.edu Tel: +66-892165777
E-mail: yodchanan.won@mahidol.ac.th Tel: +66-24414255
[*] E-mail: jetsada.arn@mahidol.ac.th Tel: +66-835666542

Abstract— **Drowsiness poses a significant challenge to cognitive and motor functions, compromising safety in critical tasks such as driving and increasing the risk of traffic accidents. Existing driver drowsiness detection systems inadequately address the gradual progression of drowsiness, focusing solely on binary classifications of drowsiness. This study aims to develop a neural network model that utilizes physiological signals, including Electroencephalogram (EEG) and Electrocardiogram (ECG), to detect multiple levels of a driver's current drowsiness (Alert, Moderately Drowsy, and Extremely Drowsy). EEG and ECG data were collected from ten participants during 1-hour simulated driving experiments, supplemented by video recordings for once-per-minute drowsiness assessments through Observer Rating of Drowsiness (ORD) by two observers, which served as the ground truth. The neural network trained on 2-channel prefrontal EEG frequency-domain features, heart-rate variability (HRV) features, and driving time achieved accuracies of 92%, 77%, and 77% for Alert, Moderately Drowsy, and Extremely Drowsy, respectively. This high performance, with reliance on minimal electrodes and simple architecture, supports its feasibility for real-time applications as an early warning system for critical drowsiness in order to promote driver safety.**

I. INTRODUCTION

Drowsiness is a transitional state to falling asleep, characterized by fluctuating levels of alertness and the onset of sleep, manifested as difficulty in maintaining wakefulness and attentiveness during activities. It is a major concern in many daily-life scenarios, especially in attention-demanding tasks such as driving. Monotonous driving lowers physiological arousal, sensorimotor functions, and information processing, thereby reducing the driver's capacity to react to sudden and critical situations on the road and making drivers prone to drowsiness [1, 2]. Cognitive and behavioral changes resulting from drowsiness can reduce task efficiency and compromise safety, potentially leading to life-threatening consequences [3].

In recent years, intensive research has been conducted into techniques for detecting driver drowsiness to create a more objective evaluation. The current limitation of most driver drowsiness detection systems lies in emphasizing the binary classification of human mental states as *alert* or *drowsy* and issuing alerts to drivers only upon reaching severe levels of drowsiness [4, 5]. Since drowsiness is a dynamic process, defining a threshold that separates mental states into only two extreme categories is too simplistic and impractical. Increasing the number of detection levels allows for a more comprehensive capture of the gradual progression of drowsiness, which is essential for timely and effective intervention. However, overly fine-grained classification may make it difficult for users to interpret the result and make prompt decisions. Therefore, this study focuses on a three-level classification of driver drowsiness to identify the intermediate state between extremely alert and drowsy, enabling drivers to be aware of their mental state and decide when to rest before reaching severe drowsiness.

Driver drowsiness detection systems can be classified into four main types: biological-based, image-based, vehicle-based, and hybrid-based methods [4-7]. Physiological measures achieve the highest accuracy among these categories, directly representing our functional state. EEG and ECG are two commonly used physiological signals in drowsiness detection, offering insights into the underlying neurological and cardiovascular systems. EEG, a brain electrical signal, is considered the gold standard to indicate drowsiness as it represents the central nervous system activity [8]. The frequency shift of EEG signals is found to correlate with drowsiness. The study by Zhao *et al.* shows the rise in low-frequency power caused by drowsiness after driving for 90 minutes [9]. Various feature extraction methods are developed to obtain significant parameters underlying the complex and non-stationary dynamics of the brain [10]. The change in alertness affects the central and autonomous nervous systems, which is linked to HRV that can be extracted from ECG [11]. During drowsiness, our body becomes restful, where parasympathetic activity increases and sympathetic activity

decreases. This relaxing response lowers the heart rate and increases the intervals between heartbeats, allowing high-frequency HRV to occur [12]. The previous study suggests that combining EEG and ECG features enhances binary drowsiness detection model performance [13], making them the primary parameters for model development in this study.

Practicality is also important, apart from detection accuracy. This study aims to minimize the system's complexity to ensure feasibility for further real-world applications. We focus on reducing the electrode requirements to enhance the potential for developing wearable devices or gadgets and simplifying the detection model for real-time processing.

Thus, the main objective of this study is to develop the physiological-based multi-level driver drowsiness classification model using minimal EEG channels and individual ECG leads. Since NHTSA reports young adults aged 16-24 as a high-risk group for falling asleep while driving [14], young adult participants were chosen as representative EEG and ECG data acquisition groups in a controlled and safe simulated monotonous driving environment.

II. Materials and Methods

The development of a physiological-based machine learning model for multi-level driver drowsiness classification involves three phases: *Data Acquisition*, including driving simulator setup and data collection; *Data Preparation*, including data preprocessing, feature extraction and selection, and ground truth labeling; and *Model Development*, as shown in Fig. 1.

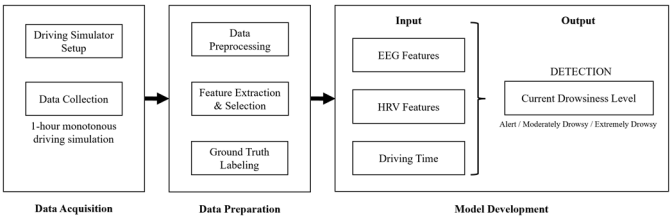


Fig. 1 Overview of Multi-Level Driver Drowsiness Classification Model Development

A. Driving Simulator Setup

Data Acquisition was performed using a 1-hour driving simulation in the Brain Computer Interface (BCI) Laboratory at Mahidol University with the 'City Car Driving' simulator and Logitech G27 USB Racing Wheel to mimic a monotonous driving environment. Physiological signals including EEG (12 channels - O1, O2, Fp1, Fp2, F3, F4, P3, P4, P7, P8, T7, T8) and ECG (lead II) were collected via a 16-channel g.USBamp RESEARCH device at a sampling rate of 512 Hz. Participant behaviors were monitored through video recordings using two cameras to capture overall body movements and facial expressions for drowsiness level labeling. Ethics approval for this experiment was granted by Mahidol University (MU-CIRB 2023/283.1209).

B. Data Collection

Data for model development was obtained from 10 healthy young adults aged 20-24 with valid driver's licenses and no known neurologic or cardiac conditions that could interfere with EEG and ECG signal interpretation. Individuals who took medications that may affect neurological or cardiac function, had a recent history of substance abuse in the past six months, consumed alcohol or caffeine within 8 hours before the study, or experienced recent traumatic events that might affect sleep patterns or mental state were excluded.

All participants were instructed to maintain a constant speed in a specified lane during a one-hour drive. To maintain a controlled environment and ensure the authenticity of participants' drowsiness progression, no more than three observers were present during the experiment, and their interaction with the subjects was minimized.

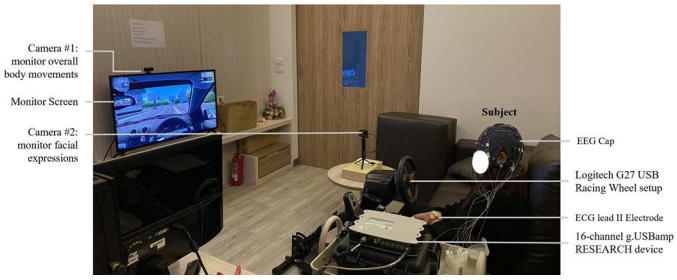


Fig. 2 Driving Simulator Setup at BCI Laboratory at Mahidol University

C. Data Preprocessing

Data preprocessing was executed using MATLAB's Signal Processing and EEGLAB Toolboxes. EEG preprocessing involves Butterworth band-pass filtering with a low-pass filter at 40 Hz (order 9) and a high-pass filter at 0.5 Hz (order 6), followed by Artifact Subspace Reconstruction (ASR) to remove artifacts. ECG preprocessing includes Butterworth band-pass filtering at 1-40 Hz (order 3) and QRS detection via the Pan & Tompkins Algorithm. Both EEG and ECG were then split into 1-minute segments with 50% overlap for further analysis. After preprocessing and excluding artifactual segments, a total of 1,132 segments each were generated from both EEG and ECG data.

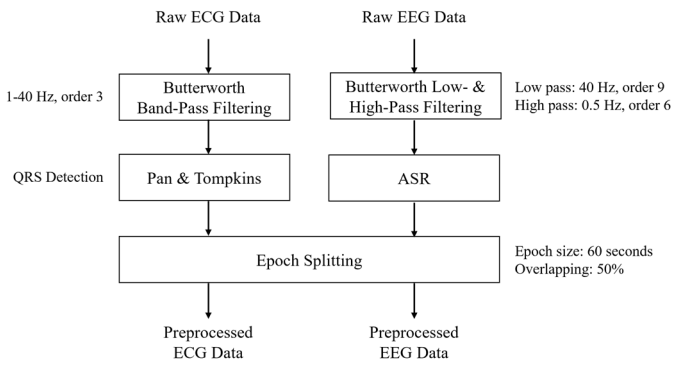


Fig. 3 Data Preprocessing Pipeline

D. Feature Extraction and Selection

This study focuses on the frequency domain features of both EEG and ECG signals. Features were extracted from each EEG and ECG data segment using MATLAB's Signal Processing and EEGLAB Toolboxes. The feature set includes:

EEG Features (from each channel): Relative Delta Power (0.5-4.0 Hz), Relative Theta Power (4.0-8.0 Hz), Relative Alpha Power (8.0-12.0 Hz), Relative Beta Power (12.0-30.0 Hz), Alpha/Beta Ratio, and (Theta+Alpha)/Beta Ratio

ECG Features: HRV features including Relative Low-Frequency (LF) Power (0.04-0.15 Hz), Relative High-Frequency (HF) Power (0.15-0.40 Hz), and LF/HF Ratio

Channels O1, O2, Fp1, and Fp2 were chosen as the primary EEG channels in this study due to their strong correlation with drowsiness, as indicated by previous research [13, 15] and their suitability for wearable device designs such as headbands. The feature sets from each channel were initially trained separately to determine the performance of single EEG channel usage. Subsequently, the combination of high-performing channels was examined to assess their integrated effects.

E. Ground Truth Labeling

ORD was applied for ground truth labeling, which involves assessing drowsiness levels by observing participants' behaviors and facial expressions from video recordings [16]. The 60-minute video data from the driving phase were divided into 60 segments, each one minute. For each video segment, two independent raters assigned a drowsiness level, categorized into five stages: Not Drowsy (scored as 1), Slightly Drowsy (2), Moderately Drowsy (3), Very Drowsy (4), and Extremely Drowsy (5). Then, the assessment scores from two raters were averaged, in order to minimize subjective bias. These labels for each minute of the sessions were aligned to the EEG and ECG segments at the same timestamps. In contrast, the segment that overlaps the labels was calculated by the mean of two consecutive labels that the segment overlaps. The average scores were used to categorize drowsiness into three simplified levels: Alert (average score ≤ 2), Moderately Drowsy ($2 <$ average score ≤ 3), and Extremely Drowsy (average score > 3). A score below two was classified as Alert, as the label 'Slightly Drowsy' is considered sufficiently alert for driving [17]. The cut-point of 3 for Extremely Drowsy was chosen because any segment with an average score above three must be labeled as 'Very Drowsy' or 'Extremely Drowsy' by at least one rater, which is critically harmful to driving. While not ideal for safe driving, the Moderately Drowsy stage is an early warning phase before progressing to the critical Extremely Drowsy stage.

F. Model Development

Multilayer Perceptron (MLP) was used to predict simplified drowsiness levels as Alert, Moderately Drowsy, and Extremely Drowsy from EEG and HRV features. Due to the progressive nature of drowsiness in a monotonous environment, driving time was included as one of the input features to enhance model accuracy. The dataset was split into an 80% Training set and a 20% Testing set, maintaining class proportions. The model architecture was simplified to ensure practicality for further real-time applications, consisting of an input layer, a batch normalization layer, two hidden layers with Rectified Linear Unit (ReLU) activation functions, each followed by a dropout layer and batch normalization layer, and a softmax layer for three-class prediction, as shown in Fig. 4. Hyperparameters were tuned using Random Search, varying the optimizer, learning rate, dropout rates, batch size, epochs, and units per layer as detailed in Table 1. 5-fold cross-validation was performed to minimize overfitting due to the limited sample size. Class weights were applied to handle class imbalance as some subjects might rarely experience extreme drowsiness under the simulated environment.

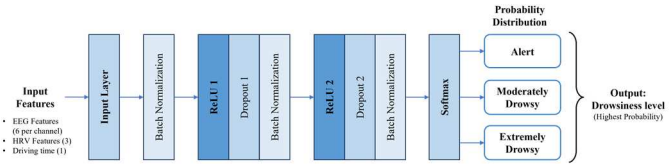


Fig. 4 Neural Network Architecture for Multi-level Drowsiness Classification

Table 1 Hyperparameter Search Space for Random Search

Hyperparameter	Search Space
Number of Units (2 hidden layers)	8, 16, 32
Dropout rates (2 dropout layers)	0.1 to 0.5
Optimizer	'adam', 'sgd', 'rmsprop'
Learning rate	0.001 to 0.1
Batch size	10, 20, 50
Epochs	50, 100, 150, 200

Several models were trained to compare performance across input feature sets to predict three drowsiness levels. EEG features were used in every model, with comparisons made based on different EEG channels (6 features from each channel) and the inclusion of HRV features (3 features). Additionally, the models with and without driving time as a feature were compared to investigate its impact on model performance. StandardScaler was used to rescale all features except driving time to preserve the consistent temporal properties, while vector concatenation was employed to integrate the input features.

III. Results

A. Comparison of Feature Sets

Table 2 shows the class-wise accuracy and F1 scores for models trained on different feature sets. The results demonstrate that including driving time and HRV features improves the models' performance across all EEG channels (O1, O2, Fp1, Fp2).

Table 2 Class-wise Accuracy and F1-scores for Models Trained on Different Feature Sets

EEG Channel	HRV features	Driving time	Class-wise Accuracy (%) 1	2	3	F1-score
O1	No	No	48.96	48.68	70.59	0.537
		Yes	92.71	53.95	20.59	0.635
	Yes	No	48.96	59.21	73.53	0.579
		Yes	78.13	61.84	73.53	0.720
O2	No	No	33.05	51.32	50.00	0.445
		Yes	79.66	59.21	47.06	0.687
	Yes	No	46.61	52.63	73.53	0.545
		Yes	73.73	78.95	50.00	0.725
Fp1	No	No	55.08	53.95	64.71	0.578
		Yes	83.05	65.79	79.41	0.772
	Yes	No	61.86	51.32	73.53	0.609
		Yes	83.05	75.00	85.29	**0.813**
Fp2	No	No	64.60	52.86	81.82	0.650
		Yes	84.07	57.14	86.36	0.759
	Yes	No	69.03	57.14	72.73	0.670
		Yes	84.07	52.86	72.73	0.728

Class 1: Alert, Class 2: Moderately Drowsy, Class 3: Extremely Drowsy

For all EEG channels, including driving time increases the class-wise accuracy across all classes, while adding HRV features results in less bias, more balanced class accuracies, and higher F1 scores. Among all EEG channels used, the Fp1 channel shows the most promising performance, with the model that includes HRV features and driving time achieving the highest overall F1-score of 0.813 and being the only one to reach an accuracy of at least 75% for all classes. When considering pairs of EEG channels, Fp1 and Fp2 generally provide higher accuracy than O1 and O2, representing a higher significant correlation with drowsiness classification. Thus, the combination of the former pair of EEG with both HRV features and driving time included is investigated.

B. Combination of Fp1 and Fp2 EEG channels

The final MLP is trained using an optimal combination of features, including EEG features from Fp1 and Fp2 channels, HRV features, and driving time. The optimized hyperparameters from the Random Search are 16 and 8 units in the first and second hidden layers, dropout rates of 0.12 and 0.46, the 'sgd' optimizer with a learning rate of 0.044, a batch size of 50, and 100 epochs. The predicted results, visualized in Fig. 5, indicate high true positive rates for Alert (92%), Moderately Drowsy (77%), and Extremely Drowsy (77%). The confusion matrix represents the model's high effectiveness in identifying the Alert state. For the Moderately Drowsy class, most errors are misclassified as a higher level of drowsiness, with a small proportion falsely classified as Alert. This is acceptable for a system designed to proactively and early detect higher levels of drowsiness rather than miss detections. Regarding the Extremely Drowsy class, 23% of samples are misclassified as Moderately Drowsy, indicating some confusion between these two classes. However, none of the actual Extremely Drowsy data points are detected as Alert, demonstrating the model's ability to distinguish between extreme classes and ensure safety in drowsiness detection during the extremely drowsy state.

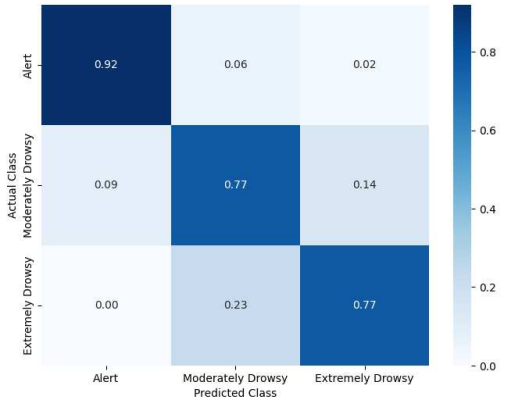


Fig. 5 Confusion Matrix for the Final Model Using Fp1 and Fp2 EEG Channels, HRV Features, and Driving Time

IV. Discussion

A. EEG channels

The Fp1 and Fp2 channels perform better in classifying drowsiness levels than O1 and O2 channels. This can be attributed to their placement over the prefrontal cortex, a brain region associated with executive functions and alertness. As drowsiness progresses, the frequency of EEG waves in this area tends to shift from higher frequencies, like beta waves, to lower frequencies, including alpha and delta waves, due to increased neuron synchronization [18]. Additionally, changes in blink rates, which affect EEG signals in Fp1 and Fp2, may also occur. These factors make frequency domain features from these complementary channels in the prefrontal cortex significant for neural networks to capture useful information from EEG and effectively classify drowsiness levels.

In contrast, the O1 and O2 channels are located over the occipital lobe and primarily process visual information. During relaxed states with eyes closed, sensory input reduction can lead to synchronous EEG and shifts to low-frequency ranges [19]. However, subjects must remain visually attentive during driving tasks despite cognitive declines. Therefore, Fp1 and Fp2 are more likely to reflect drowsiness-related changes during driving. Combining features from both channels can be more effective, especially when there is a loss of EEG signal or other issues.

B. Combination of EEG and HRV Features

ECG signals provide valuable information about heart-rate variability, which reflects autonomic nervous system activity

that changes as drowsiness progresses. Biological signals vary more distinctly between extreme states, such as being extremely alert and drowsy. The transition period, however, is challenging for the model to identify due to the gradual nature of drowsiness.

The study shows that integrating HRV features effectively enhances the model's ability to identify this intermediate state. This improvement is likely due to the more holistic view of physiological changes provided by ECG, enabling the model to detect subtle changes during the transition period between extreme states. Moreover, HRV features help mitigate some limitations of EEG signals, which are more susceptible to artifacts and noise that might obscure the neural patterns associated with drowsiness. Therefore, combining EEG and HRV features as a multimodal approach significantly enhances the drowsiness classification process by the neural network compared to using EEG features alone.

C. Driving Time

Adding driving time as a feature substantially impacts the model's performance. Drowsiness is a progressive condition, and including driving time helps the model understand the temporal context of the physiological signals. This temporal information allows the model to accurately capture the gradual transition between alertness and drowsiness and discern patterns and trends in the physiological data corresponding to different drowsiness stages. This is particularly important for detecting the Moderately Drowsy state, where transitions are more subtle and gradual. This finding also aligns with the previous study, which supports the positive impact of driving time on model performance [20].

Compared to HRV features, driving time has a greater influence on model accuracy, likely due to the similar pattern of drowsiness progression over time across individuals, while physiological responses can vary. Nevertheless, HRV remains a crucial complementary modality to EEG, as discussed earlier (see Discussion, *B. Combination of EEG and HRV Features*). The relative importance of HRV compared to driving time could be further investigated with greater consideration of circadian rhythms and other time-related factors that affect individual susceptibility to drowsiness [2].

D. Model Practicality

The practicality of the final model, trained using features from Fp1 and Fp2 EEG channels, frequency-domain HRV features, and driving time, is considered based on model performance and the feasibility of wearable technology.

In terms of model performance, the model effectively distinguishes between different levels of drowsiness, including the *Moderately Drowsy* state, with no instances of Extremely Drowsy being misclassified as Alert. This capability enables early warnings that allow users to be notified and aware of their declining mental readiness for driving before reaching a critical stage of drowsiness.

For application in wearable technology, the model's reliance on minimal electrodes, including two EEG channels in the prefrontal area, with HRV features, makes it feasible to design devices that capture those signals during driving. EEG monitoring devices can take the form of headbands for less intrusiveness and a more comfortable setup for continuous monitoring, while heart rate can be monitored via wristbands or smartwatches. Using minimal features and simple model architecture reduces the complexity, making real-time processing more feasible.

V. Conclusions

The study suggests that combining EEG and HRV features enhances the performance of 3-level drowsiness classification, particularly in identifying the intermediate state between alertness and critical drowsiness. Moreover, implementing driving time preserves the temporal pattern of drowsiness progression, leading to better overall accuracy. The final model based on 2-channel prefrontal EEG features, HRV features, and driving time yields high accuracy across all classes. Although there are some misclassifications, most errors in *Moderate Drowsy* detection are falsely predicted as *Extremely Drowsy,* which is acceptable for early warnings of critical drowsiness.

The model’s use of minimal electrodes and simple architecture supports its practicality for real-time applications, contributing to driver safety. Future developments could include expanding the variety of driving conditions and the participants’ age range to improve model generalization, developing wireless, real-time systems, and incorporating personalized participant data for improved accuracy. Additionally, pre-driving data could be applied to identify physiological markers correlating with drowsiness onset and progression, offering potential safety measures before driving.

VI. Acknowledgment

This work was technically supported by the BCI Laboratory at Mahidol University and funded by Tonkla Ramathibodi Program from the Faculty of Medicine Ramathibodi Hospital, Mahidol University.

References

[1] D. J. Mascord and R. A. Heath, "Behavioral and physiological indices of fatigue in a visual tracking task," *Journal of Safety Research,* vol. 23, no. 1, pp. 19-25, 1992/03/01/ 1992, doi: https://doi.org/10.1016/0022-4375(92)90036-9.

[2] H. Zhang, D. Ni, N. Ding, Y. Sun, Q. Zhang, and X. Li, "Structural analysis of driver fatigue behavior: A systematic review," *Transportation Research Interdisciplinary Perspectives,*

vol. 21, p. 100865, 2023/09/01/ 2023, doi: https://doi.org/10.1016/j.trip.2023.100865.

[3] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews,* vol. 44, pp. 58-75, 2014/07/01/ 2014, doi: https://doi.org/10.1016/j.neubiorev.2012.10.003.

[4] M. Ramzan, H. U. Khan, S. M. Awan, A. Ismail, M. Ilyas, and A. Mahmood, "A Survey on State-of-the-Art Drowsiness Detection Techniques," *IEEE Access,* vol. 7, pp. 61904-61919, 2019, doi: 10.1109/ACCESS.2019.2914373.

[5] Y. Albadawi, M. Takruri, and M. Awad, "A Review of Recent Developments in Driver Drowsiness Detection Systems," *Sensors,* vol. 22, no. 5, p. 2069, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/5/2069.

[6] M. Shaik, "A systematic review on detection and prediction of driver drowsiness," *Transportation Research Interdisciplinary Perspectives,* vol. 21, 06/23 2023, doi: 10.1016/j.trip.2023.100864.

[7] V. Saini and R. Saini, "Driver drowsiness detection system and techniques: a review," *International Journal of Computer Science and Information Technologies,* vol. 5, no. 3, pp. 4245-4249, 2014.

[8] S. S. Poorna, A. Deep, K. Hariharan, R. R. Jain, and S. Krishnan, "A Transfer Learning Approach for Drowsiness Detection from EEG Signals," in *Innovations in Computational Intelligence and Computer Vision*, Singapore, M. K. Sharma, V. S. Dhaka, T. Perumal, N. Dey, and J. M. R. S. Tavares, Eds., 2021// 2021: Springer Singapore, pp. 369-375.

[9] C. Zhao, M. Zhao, J. Liu, and C. Zheng, "Electroencephalogram and electrocardiograph assessment of mental fatigue in a driving simulator," *Accident Analysis & Prevention,* vol. 45, pp. 83-90, 2012/03/01/ 2012, doi: https://doi.org/10.1016/j.aap.2011.11.019.

[10] I. Stancin, M. Cifrek, and A. Jovic, "A Review of EEG Signal Features and Their Application in Driver Drowsiness Detection Systems," *Sensors,* vol. 21, no. 11, p. 3786, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/11/3786.

[11] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," (in English), *Frontiers in Public Health,* Review vol. 5, 2017-September-28 2017, doi: 10.3389/fpubh.2017.00258.

[12] K. Fujiwara *et al.*, "Heart Rate Variability-Based Driver Drowsiness Detection and Its Validation With EEG," *IEEE Transactions on Biomedical Engineering,* vol. 66, no. 6, pp. 1769-1778, 2019, doi: 10.1109/TBME.2018.2879346.

[13] M. Awais, N. Badruddin, and M. Drieberg, "A Hybrid Approach to Detect Driver Drowsiness Utilizing Physiological Signals to Improve System Performance and Wearability," (in eng), *Sensors,* vol. 17, no. 9, Aug 31 2017, doi: 10.3390/s17091991.

[14] *NHTSA Drowsy Driving Research and Program Plan*, 2016. [Online]. Available: https://www.nhtsa.gov/sites/nhtsa.gov/files/drowsydriving_strategicplan_030316.pdf.

[15] S. Chaabene, B. Bouaziz, A. Boudaya, A. Hökelmann, A. Ammar, and L. Chaari, "Convolutional Neural Network for Drowsiness Detection Using EEG Signals," *Sensors,* vol. 21, no. 5, p. 1734, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/5/1734.

[16] W. W. Wierwille and L. A. Ellsworth, "Evaluation of driver drowsiness by trained raters," (in eng), *Accid Anal Prev,* vol. 26, no. 5, pp. 571-81, Oct 1994, doi: 10.1016/0001-4575(94)90019-1.

[17] D. Wiegand, J. McClafferty, S. E. McDonald, and R. Hanowski, "Development and Evaluation of a Naturalistic Observer Rating of Drowsiness Protocol Final Report," 01/01 2005.

[18] T. Kirschstein and R. Köhling, "What is the Source of the EEG?," *Clinical EEG and Neuroscience,* vol. 40, no. 3, pp. 146-149, 2009, doi: 10.1177/155005940904000305.

[19] F. L. Britton JW, Hopp JLet al., F. L. St. Louis EK, Ed. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*. Chicago: American Epilepsy Society, 2016.

[20] C. Jacobé de Naurois, C. Bourdin, A. Stratulat, E. Diaz, and J. L. Vercher, "Detection and prediction of driver drowsiness using artificial neural network models," (in eng), *Accident Analysis & Prevention,* vol. 126, pp. 95-104, May 2019, doi: 10.1016/j.aap.2017.11.038.

ACE-Flow: Auto Color Encoding for Enhanced Low-Light Image Restoration

Jiachen Qiu, Yushen Zuo and Kin-Man Lam
Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong
E-mail: jiachen.qiu@connect.polyu.hk, yushen.zuo@polyu.edu.hk, enkmlam@polyu.edu.hk

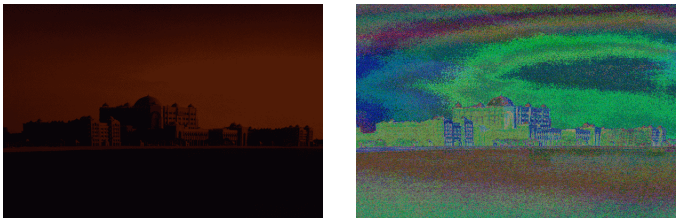
***Abstract*—Low-light image enhancement is a classic problem in low-level vision tasks, aiming to improve the quality of images captured in poor-lighting scenarios. Conventional deep enhancement models often produce distorted content (e.g., deviated lighting conditions, color bias) in extremely dark regions because they fail to capture comprehensive color information during the reconstruction process. To address these issues, we propose a novel normalizing flow-based model that incorporates an auto-color encoding method, called ACE-Flow, for low-light image enhancement. By leveraging auto-color encoding, our method can encode color information during feature extraction and effectively restore the corrupted image content in challenging regions. Furthermore, our approach can accurately learn the mapping from low-light images to high-quality ground-truth images, because the invertibility property of the normalizing flow implicitly regularizes the learning process. Experiments demonstrate that our method significantly outperforms other promising low-light enhancement models in terms of reconstruction and perceptual metrics. Additionally, the enhanced images produced by our model exhibit rich details with minimal distortion, resulting in superior visual quality.**

I. Introduction

Low-light image enhancement is an attractive topic in low-level vision tasks, with the aim of jointly improving luminance and removing undesirable noise caused by sensors and dim environments. Low-light image enhancement techniques have significant industrial values, including applications in modern imaging devices, surveillance, and autonomous driving, attracting considerable researcher attention.

In the past years, many deep learning-based low-light image enhancement models have been proposed and have achieved promising results. However, the methods in [1]–[4] often produce unacceptable artifacts when processing real-world Ultra-High-Definition images [5]. Wu *et al.* [6] and Xu *et al.* [7] proposed methods to simultaneously enhance luminance and remove noise in the spatial domain, resulting in degraded performance when the images are captured in challenging lighting conditions. Despite the promising results shown in [8], the heavy network structures highly limit its applications in real-world scenarios. All the mentioned methods largely ignore the profound color information in the reconstruction process, which inevitably leads to suboptimal solutions. In addition, after training, these methods cannot be generalized to other kinds of low-light images, such as infrared images.

Unlike previous studies, this paper focuses on more challenging image data, i.e., infrared images. When the infrared filter is removed in the imaging process, the captured images easily suffer from under-exposure in environments with inadequate lighting conditions. As a result, the main image content is unavoidably corrupted by the noise generated by sensors and other hardware devices. To handle these challenges, we propose a novel normalizing flow-based model that incorporate an auto-color encoding method, namely ACE-Flow. To restore the image content underlying the dark regions, our proposed method can effectively encode the color information of objects and image content by leveraging the proposed auto-color encoding method, providing beneficial prior information for restoration. As low-light image enhancement is intrinsically ill-posed, our method adopts the normalizing flow model to learn the mapping from low-light images to ground-truth images, because the invertible property of the normalizing flow implicitly regularizes the forward imaging process from ground-truth images to low-light images. Furthermore, due to the invertibility of normalizing flows, which allows for exact inference and reconstruction of input data, we have empirically observed that normalizing flows outperform diffusion models. Consequently, our method leverages normalizing flows to produce higher-quality images from low-light counterparts.



(a) Low light image (b) CE image

Fig. 1: A Low-Light Image (LLI) compared with its Color Encoding image (CE).

The main contributions of this paper are summarized as follows:

1) To address the challenging issues of Low-Light Infrared images (LLIR), we propose a novel normalizing flow-based model for low-light image enhancement.
2) Due to the ill-posed nature of low-light images, our proposed method leverages auto-color encoding (ACE) to capture the color information of low-light images, benefiting the restoration of image contents in challenging regions.

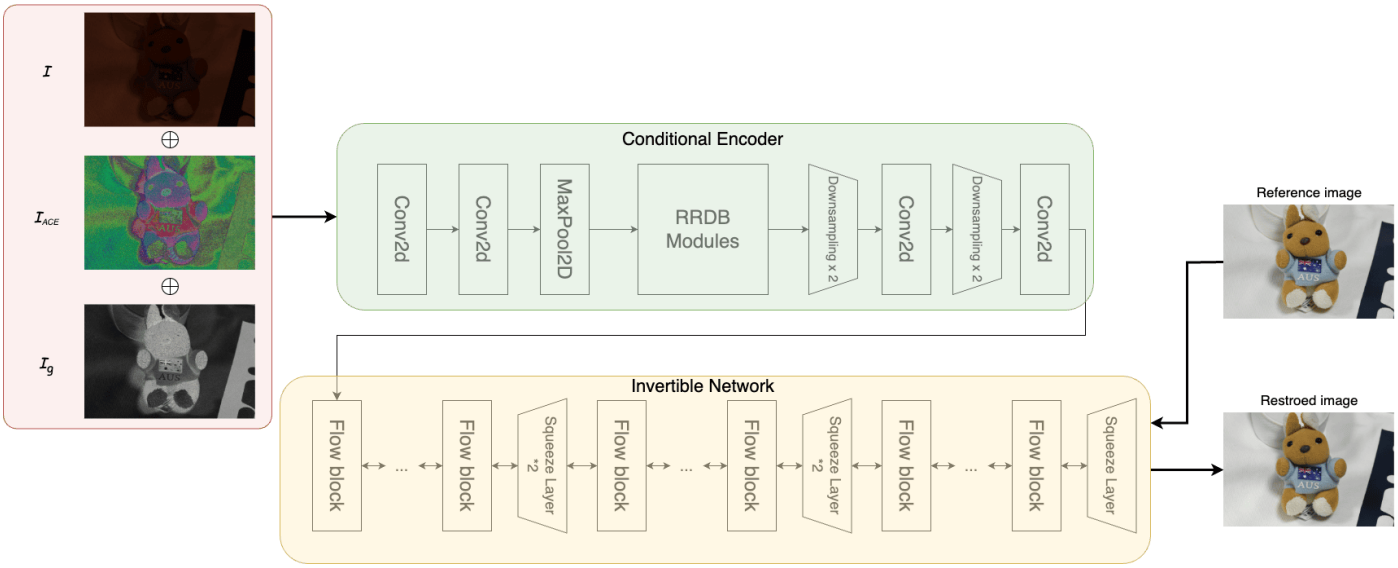


Fig. 2: Overview of our proposed framework (ACE-Flow). The inputs to the framework include the degraded input image (3 channels) I, the color encoding image (3 channels) I_{ACE}, and the inverted grayscale version of the low-light image I_g. $\oplus$ denotes concatenation. The RRDB modules inside the Conditional Encoder represent the Residual-in-Residual Dense Blocks (RRDB) [9].

3) We leverage the invertible property of the normalizing flow to effectively learn the mapping from low-light images to the corresponding ground-truth images, implicitly regularizing the forward imaging process.
4) Experiments show that our proposed method can achieve better performance than other promising deep low-light enhancement models for LLIR images, both in terms of reconstruction and perceptual metrics. Additionally, our method can effectively produce images with rich details and minimal distortion, resulting in superior visual quality.

II. Related Work

A. Low-light image enhancement

Low-Light Image Enhancement (LLIE) is a well-researched area in computer vision, with numerous methods proposed over the years to address the challenges of improving image quality under poor lighting conditions. Here, we categorize the related work into traditional methods, learning-based methods, and recent advancements in the field. LLIE is an active and attractive research area, with various models based on different architectures proposed to tackle this challenging topic [10]–[13]. Retinex-based models mainly focus on decomposing the image into illumination and reflectance components to suppress noise, improve dim lighting conditions, and remove artifacts. Deep learning-based models utilize the computational power of GPUs to train relatively big models with greater numbers of parameters, enabling them to tackle much more complex situations using larger datasets and achieve more robust image restoration [14].

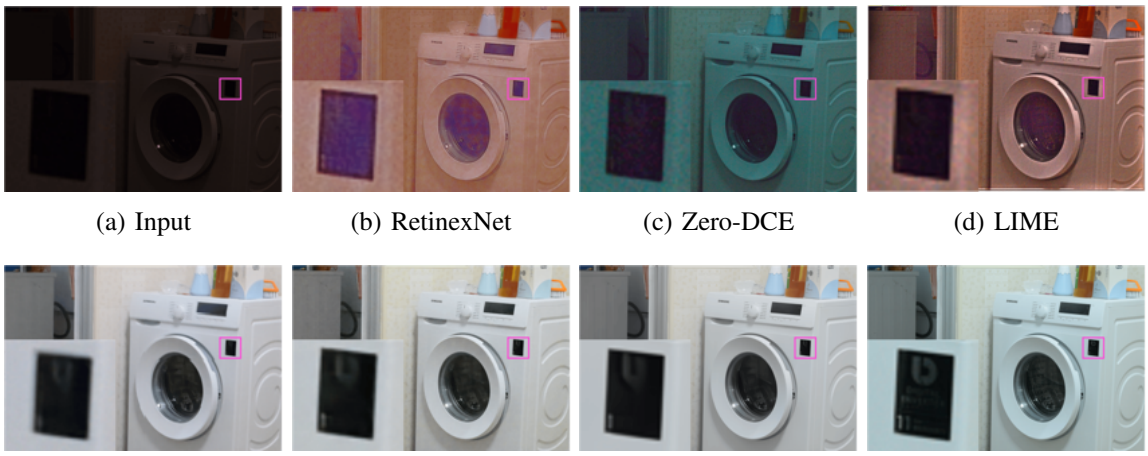
B. Normalizing flow

Normalizing flow is a powerful technique in deep learning for modeling complex data distributions. It transforms a simple probability distribution into a more complex one through a series of invertible transformations, making it highly suitable for tasks requiring detailed probabilistic modeling. Methods like Glow [15] and RealNVP [16] have demonstrated the efficacy of normalizing flow in generating high-quality images by learning the underlying data distribution. These models are particularly useful for tasks such as image synthesis and density estimation. Recently, normalizing flow has been applied to LLIE, as seen in LLFlow [17], which models the distribution of low-light images to effectively handle noise and artifacts. By leveraging the invertibility of flow-based models, LLFlow achieves superior performance in enhancing low-light images without introducing significant artifacts.

III. Methodology

A. The Overall Pipeline of Proposed Method

In this paper, we propose a novel low-light image enhancement technique based on a novel colour encoding method and using the normalizing flow model as the processing backbone. Fig. 2 shows the overall pipeline of the proposed method, consisting of three main parts: the new colour encoding block, a conditional encoder, and an invertible network. The colour encoding block draws inspiration from the powerful positional encoding technique in the Transformer models. Instead of coding positional information, we use a series of sine and cosine waves to encode the colour pixels directly. The rationale behind this design is to leverage this encoding scheme to extract powerful and discriminative representations that are



(a) Input (b) RetinexNet (c) Zero-DCE (d) LIME

(e) DiffIR (f) Restormer (g) Ours (h) Reference

Fig. 3: Perceptual quality comparison with SOTA methods in the field of low-light enhancement in the IR dataset. Our model performs better in noise reduction and restores more details with smooth surfaces. The small pink region within each image was magnified and repositioned to the lower-left corner for detailed comparison.

not easily discernible in low-light images to the human eye. We will demonstrate the effectiveness of this method later. In addition to this new colour encoding scheme, we also extract the gray map of the input low-light images, which, together with the sinusoidal wave-encoded colour map, serves as prior information in the restoration process. The conditional encoder is responsible for extracting features from the input and generating the input for the invertible network. The invertible network, which consists of flow blocks, aims to generate high-quality images based on the features of the input.

B. Preliminary

A normalizing flow model transforms a simple probability distribution into a more complex one through a series of invertible and differentiable transformations. Previous deep learning-based models primarily relied on pixel reconstruction loss. However, due to the challenge of separating undesirable artifacts from the true distribution of the reference images, these models often produce restored images with poor perceptual quality, resulting in noise and blurriness [8], [18].

Motivated by the good reconstruction performance of flow-based models [19]–[21], we realized that utilizing conditional probability distributions can address the aforementioned problem by encompassing various distributions of natural images. In particular, the state-of-the-art (SOTA) method LLFlow [17] demonstrated outstanding performance by using normalizing flow conditioned on low-light images to model the constrained distribution of the reference images. In our method, we adopt the core concept of conditional flow, along with the likelihood evaluation method proposed in LLFlow, as the backbone model of our framework. The conditional probability density function (PDF) for well-exposed images is formulated as follows:

$$f_{cnf}(x_{hq}|x) = f_z(\Theta(x_{hq};x)) \left| \det \frac{\partial \Theta}{\partial x_{hq}}(x_{hq};x) \right| \quad (1)$$

where $f_{cnf}(\cdot)$ represents the conditional PDF, and $\Theta(\cdot)$ is the bidirectional network composed of N invertible layers $\theta_1, \theta_2, \ldots, \theta_N$. The latent variable $z = \Theta(x_{hq};x)$ is derived by transforming the corrupted inputs x into normally exposed images x_{hq}. By employing maximum likelihood estimation, the model can be optimized using the negative log-likelihood loss function. The loss function is formulated as follows:

$$\begin{aligned} L_{nll}(x, x_{hq}) &= -\log f_{cnf}(x_{hq}|x) \\ &= -\log f_z(\Theta(x_{hq};x)) - \sum_{n=0}^{N-1} \log \left| \det \frac{\partial \theta_n}{\partial z_n}(z_n; g_n(x)) \right| \end{aligned} \quad (2)$$

where $g(\cdot)$ denotes the encoder that generates the conditional embedding of the layers θ_i from the bidirectional network.

C. Sinusoidal modulation color encoding

Inspired by the positional encoding scheme of the Transformer model [22], we generate three channels of colour modulated sinusoidal waves and create a color-encoded low-light input. Using a method similar to that used for sequential data in Transformer models, the color encoding for images is computed to derive a more powerful and discriminative representation. For a given input image, the color encoding is calculated as follows:

$$ACM(I_{(x,y,c)}, 2i) = \sin\left(\frac{I_{(x,y,c)}}{10000^{\frac{2i}{d}}}\right) \quad (3)$$

$$ACM(I_{(x,y,c)}, 2i+1) = \cos\left(\frac{I_{(x,y,c)}}{10000^{\frac{2i}{d}}}\right) \quad (4)$$

where $I_{(x,y,c)}$is the pixel value of the input image at spatial location (x, y) and $c \in \{0, 1, 2\}$ represents the red, green, and blue channels of the input image respectively, $i = 0, 1, ..., d/2$ is the modulation dimension index, and d

represents the modulation dimension, which corresponds to the total number of channels of the input image provided to the model. It is evident that Equations (3) and (4) map a $d-$D pixel vector at the position (x, y) to an auto colour encoding matrix $ACM(I_{(x,y,c)})$ in a specific channel $c \in \{0, 1, 2\}$. To obtain sinusoidal modulated color encoding, we follow a similar operation as in the positional encoding method of the Transformer model. Then, we sum over the modulation dimension indices, as follows:

$$ACE(x, y, c) = \frac{1}{d} \sum_{i=0}^{i=d} ACM(I_{(x,y,c)}, i) \tag{5}$$

where $ACE(x, y, c)$ is a sinusoidal modulated colour encoding at the spatial location (x, y) and channel c. Then, we concatenate $ACE(x, y, c)$, $c \in \{0, 1, 2\}$, to form a color encoding image ACE, as follows:

$$I_{ACE} = [ACE_r, ACE_g, ACE_b]. \tag{6}$$

where ACE_r, ACE_g and ACE_b denote the color encoding images of the red, green and blue channels.

D. Graymap

We also generate a graymap as an extra input image. The graymap I_g is calculated by taking the average of the red, green, and blue values at each pixel position, and then subtracting this average from 255, as follows:

$$I_g(x, y) = 255 - I_{mean}(x, y), \tag{7}$$

where I_{mean} represents the average value of the red, green and blue pixel values of the input image, i.e., the grayscale image. This operation provides an additional channel that highlights the inverse of the average intensity of the image. By emphasizing the areas of the image that have lower intensities, the graymap can guide the model to better understand the overall brightness and contrast. This additional channel can be particularly useful in enhancing the model's ability to detect and process features that might be more subtle or less obvious in the original image.

Finally, as illustrated in Equation (7), we concatenate the original image I, the sinusoidal modulated colour encoding image I_{ACE}, and the graymap I_g together to form the input I_{input} to the backbone network:

$$I_{input} = [I, I_{ACE}, I_g]. \tag{8}$$

Clearly, I_{input} has seven channels, each with the same resolution as the original input image. The first three channels are the original red, green, and blue channels, and the next three channels are the sinusoidal modulated colour encoding of the red, green, and blue pixels (see Equations (3), (4) and (5)),and the last channel is the graymap.

TABLE I: Quantitative comparison of our method with various state-of-the-art methods

Methods	PSNR(↑)	SSIM(↑)	LPIPS(↓)
RetinexNet [2]	11.14	0.628	0.586
LIME [23]	11.31	0.639	0.560
Zero-DCE [24]	11.40	0.592	0.443
DiffIR [18]	20.74	0.684	0.200
Restormer [8]	24.73	0.842	0.125
LLFlow [17]	25.42	0.866	0.118
Ours	**25.99**	**0.875**	**0.106**

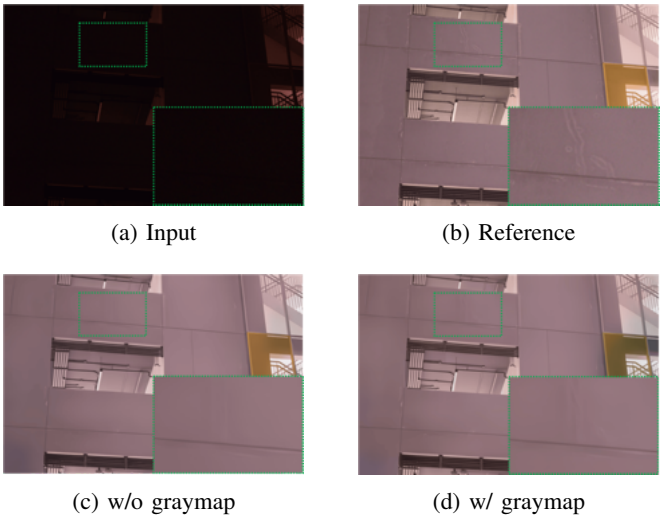
IV. Experiments

A. Experimental settings

To facilitate various experimental settings, we crop the original images of 400 × 600 resolution to 256 × 256 patches, optimizing I/O operations for better efficiency and time savings. Our training setup includes a total of 2,832 image pairs, while the evaluation set comprises 87 image pairs to assess the performance of our trained models. We evaluated our proposed model against several leading methods, including RetinexNet [2], LIME [23], Zero-DCE [24], DiffIR [18], Restormer [8], and LLFlow [17].

B. Experimental results

To evaluate the performance of different methods on the IR dataset, we retrained all the methods using the same training data, i.e., the training set of the IR dataset. For a fair comparison, we explored a wide range of hyperparameters for the compared methods and reported the best performance obtained. The experimental results are presented in Table I, and a visual comparison is shown in Fig. 3.



(a) Input (b) Reference

(c) w/o graymap (d) w/ graymap

Fig. 4: Visual comparison of different configurations of graymap. The green rectangular regions in the images are enlarged for easy comparison. (a) An input image, (b) the reference, (c) the restored image without using graymap and (d) the restored result using the graymap.



(a) Reference (b) CE of LLI (c) w/o ACE (d) w/ ACE

Fig. 5: Visual comparison of different configurations of ACE. We can see that by adding ACE, the model's performance in restoring low-light images is highly improved.

Based on our evaluation and analysis of the experimental results, we observe that RetinexNet exhibits limited generalization ability and produces unsatisfactory outputs, e.g., RetinexNet [2]. We conjecture that this limitation arises because the method assumes the existence of an invariant reflectance map across low-light inputs and ground-truth images, requiring a shared network to extract both illumination and reflectance maps, which is not feasible in our setting.

In contrast, methods based on other principles, such as Restormer [8] with a Transformer architecture and DiffIR [18] using a diffusion model, show better performance. However, our method achieves the best performance among all competitors in terms of both fidelity and perceptual quality.

C. Ablation Study

An ablation study was performed to evaluate the effectiveness of ACE and the graymap in enhancing model performance. The visual results, which highlight the impact of these components, are shown in Fig. 4 and Fig. 5.

The experimental results demonstrate that incorporating the graymap into the model significantly enhances its sensitivity to subtle image details. Specifically, in the green rectangular region of Fig. 4(c), it is evident that the model without the graymap struggles to capture edge information effectively. However, after integrating the graymap, the model is able to generate much clearer and more defined details that correspond to the edges present in the reference image, as illustrated in Fig. 4(b).

To further substantiate these findings, quantitative results are provided in Table II, where "A" and "G" represent the use of ACE and the graymap, respectively. The data in Table II clearly indicates that the combined use of ACE and the graymap yields substantial improvements in the model's performance, particularly in terms of PSNR. Notably, while the integration of the graymap alone results in a more pronounced increase in PSNR, it only brings about a slight enhancement in SSIM. On the other hand, the inclusion of ACE delivers a comparable improvement in SSIM, though the gain in PSNR is marginally lower compared to that achieved with the graymap. When both the graymap and ACE are employed together, the best performance is achieved.

TABLE II: Quantitative Results of Ablation Study. "A" means that ACE is used, while "G" means that graymap is used.

G	A	PSNR (↑)	SSIM (↑)
✗	✗	25.42	0.866
✓	✗	25.87	0.874
✗	✓	25.78	0.874
✓	✓	**25.90**	**0.876**

V. Conclusion

In this paper, we introduce a novel framework, the ACE-Flow model, designed to address the Low-Light Image Enhancement (LLIE) task using an infrared (IR) dataset. This dataset captures more photons by cutting off the infrared filter block inside the camera. By incorporating the ACE-Flow structure and a unique graymap, our model can more effectively interpret the color information and intricacies of low-light images, resulting in high-quality outputs. Extensive experiments demonstrate that our model excels in both quantitative evaluations and perceptual quality, showcasing its superior performance.

References

[1] L. Zhao, S.-P. Lu, T. Chen, Z. Yang, and A. Shamir, "Deep symmetric network for underexposed image enhancement with recurrent attentional learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 075–12 084.

[2] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018.

[3] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "Msr-net: Low-light image enhancement using deep convolutional network," *arXiv preprint arXiv:1711.02488*, 2017.

[4] S. Park, S. Yu, M. Kim, K. Park, and J. Paik, "Dual autoencoder network for retinex-based low-light image enhancement," *IEEE Access*, vol. 6, pp. 22 084–22 093, 2018.

[5] C. Li, C.-L. Guo, M. Zhou, *et al.*, "Embedding fourier for ultra-high-definition low-light image enhancement," *arXiv preprint arXiv:2302.11831*, 2023.

[6] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5901–5910.

[7] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 714–17 724.

[8] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[9] X. Wang, K. Yu, S. Wu, *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[10] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3291–3300.

[11] J. Xiao, W. Jia, and K.-M. Lam, "Feature redundancy mining: Deep light-weight image super-resolution model," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 1620–1624.

[12] H. Xie, Z. Huang, F. H. Leung, Y. Ju, Y.-P. Zheng, and S. H. Ling, "A structure-affinity dual attention-based network to segment spine for scoliosis assessment," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2023, pp. 1567–1574.

[13] Y. Ju, K.-M. Lam, J. Xiao, C. Zhang, C. Yang, and J. Dong, "Efficient feature fusion for learning-based photometric stereo," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[14] J. Xiao, Z. Lyu, C. Zhang, Y. Ju, C. Shui, and K.-M. Lam, "Towards progressive multi-frequency representation for image warping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2995–3004.

[15] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[16] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[17] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. Kot, "Low-light image enhancement with normalizing flow," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, pp. 2604–2612.

[18] B. Xia, Y. Zhang, S. Wang, *et al.*, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 095–13 105.

[19] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "Srflow: Learning the super-resolution space with normalizing flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, Springer, 2020, pp. 715–732.

[20] V. Wolf, A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "Deflow: Learning complex image degradations from unpaired data with conditional flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 94–103.

[21] M. Xiao, S. Zheng, C. Liu, *et al.*, "Invertible image rescaling," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, Springer, 2020, pp. 126–144.

[22] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[23] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on image processing*, vol. 26, no. 2, pp. 982–993, 2016.

[24] C. Guo, C. Li, J. Guo, *et al.*, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1780–1789.

Blind Estimation of Room Volume from Reverberant Speech Based on the Modulation Transfer Function

Nutchanon Siripool[1, 3], Suradej Duangpummet[2], Jessada Karnjana[2], Waree Kongprawechanon[1], and Masashi Unoki[3]
[1] Sirindhorn International Institute of Technology, Thammasat University, Thailand
E-mail: d6622300249@g.siit.tu.ac.th, waree@siit.tu.ac.th
[2] National Science and Technology Development Agency, Thailand
E-mail: suradej.dua@nectec.or.th, jessada.kar@nectec.or.th
[3] School of Information Science, Japan Advanced Institute of Science and Technology, Japan
E-mail: unoki@jaist.ac.jp, s2420404@jaist.ac.jp

***Abstract*—Estimating room volume is crucial in acoustic engineering, architectural acoustics, and emergency response. Traditional methods relying on the Sabine equation face challenges due to the difficulty in accurately measuring absorption coefficients. This paper introduces a method for blindly estimating room volume using reverberant speech on the basis of the modulation transfer function (MTF). Our approach processes reverberant speech through a seven-octave band filterbank, extracting temporal amplitude envelopes to form MTF features and input the MTF feature into a convolutional neural network to estimate room volume. Experimental results on various real-world and simulated room impulse responses demonstrate our method's effectiveness, achieving a mean square error of 0.14 in volume estimation. The mean absolute logarithm of the ratio between the estimated volume and the ground truth is 1.15. This approach significantly improves accuracy and robustness over the baseline method, offering a practical solution for scenarios where direct measurements are impractical. Potential applications include security monitoring, space management, and emergency evacuation planning.**

I. Introduction

Estimating room volume is essential in various fields, including acoustic engineering, where it is used to design sound systems [1], optimize room acoustics [2], and create soundproofing for any room. It also plays a critical role in emergency response scenarios, such as rescue operations during fires. Knowing the room volume (V) helps assess the capacity of the target room, estimate the oxygen requirements for firefighters, and determine the number of people potentially trapped inside a smoke-filled room. Additionally, room volume needs to be accurately estimated for architectural acoustics [2], influencing the design of concert halls [1], lecture rooms, and recording studios to enhance sound quality and speech intelligibility.

Traditional methods for calculating room volume often rely on the Sabine equation, as defined as

$$V \approx \left(\frac{RT_{60} \times 6.4\bar{\alpha}}{0.161} \right)^3, \tag{1}$$

which necessitates knowing the reverberation time (RT_{60}) and average absorption coefficient of room surface $(\bar{\alpha})$ [3]. However, measuring the absorption coefficient accurately in real-world situations can be challenging. Obtaining precise absorption coefficients involves finding out details of the materials and their properties within the room [3], which is often impractical and time-consuming [4]. Consequently, a more efficient and reliable approach is needed to estimate room volume.

Estimating room volume from reverberant speech is crucial in acoustics and speech processing. However, direct measurements or manual calibration are often impractical. Hence, many researchers have proposed methods for estimating room volume [5]–[9].

Recently, there have been many interesting developments in convolutional neural networks (CNNs) in the acoustic field. Genovese *et al.* [7] employed CNNs for blindly estimating room volume from single-channel noisy speech, achieving accurate results across various room sizes. Their approach focuses on critical parameters such as RT_{60} and direct-to-reverberant ratio (DRR). However, this model's performance decreased significantly on a separate, measured data corpus of unseen rooms, speech, and ambient noise. Vaswani *et al.* [9] introduced transformers, highlighting attention mechanisms for sequence modeling that enhance the analysis of reverberant speech characteristics, thereby improving room volume estimation accuracy. The paper employs data augmentation techniques such as SpecAugment [6]. This indicates that the base model may not perform optimally without these enhancements. Additionally, Shabtai *et al.* [5] explored room-volume classification from reverberant speech. Although the method is promising, it is less accurate and reliable than standard measurement techniques, requiring further investigation to ensure robustness in practical applications.

Previously, the use of modulation transfer function (MTF) based CNNs for blindly estimating room acoustic parameters [10] was demonstrated by simultaneously predicting various parameters, such as RT_{60} and speech transmission index (STI) [11], [12]. Although this method can estimate various room acoustic parameters, it still needs to catch up in determining the room volume, which is a crucial area of interest.

Therefore, we extend our previous methodologies by fo-

cusing on blind estimation of room volume from reverberant speech using MTF features extracted via a seven-octave band filterbank. The MTF feature extracted from reverberant speech signals can be explained by the concept of the MTF [13], [14]. We aim to provide a method that blindly estimate room volume robustly and accurately.

II. Background

A. Modulation Transfer Function

MTF, a transfer function of a linear system, characterizes a transmission channel on the basis of modulation frequency and the reduction in modulation depth [13]–[15]. It was introduced in room acoustics to evaluate how an enclosure affects speech intelligibility. In room acoustics, MTF quantifies the impact of reverberation; increased reverberation results in a lower modulation depth for modulated signals passing through the room. The ratios of modulation distortion between the input envelopes and their corresponding outputs are referred to as modulation indices. The MTF is defined as

$$m(f_m) = \frac{\int_0^\infty h^2(t) e^{-j2\pi f_m t} dt}{\int_0^\infty h^2(t) dt}, \tag{2}$$

where $h(t)$ denotes the room impulse response.

B. Temporal Amplitude Envelopes

The temporal amplitude envelope (TAE) can reveal information about the reverberation characteristics of a room, such as the RT_{60} [16], which are influenced by the room's volume. The TAE is a smoothed version of the input signal in reverberant environments. It provides essential information regarding the room's acoustic parameter [17]. In this work, we extracted the TAE using Eq. (3), which involves the Hilbert transform and a low-pass filter (LPF). The Hilbert transform is a mathematical operator used to analyze a real-valued signal, which is useful for extracting the amplitude and phase of the signal [18]. The LPF is a sixth-order Butterworth filter with a cut-off frequency of 30 Hz, and the signal is downsampled to 60 Hz to reduce complexity.

$$e_y(t) = \mathrm{LPF}\left[|\mathrm{y(t)} + j\mathrm{Hilbert(y)(t)}|\right]. \tag{3}$$

We conduct a preliminary study to establish a foundational understanding and verify the feasibility of estimating room volume using RT_{60} through the Sabine Eq. (1) and Eq. (4) from Kuster *et al.* [19], without relying on α or any specific coefficients. The goal was to determine whether using only RT_{60} could accurately map to room volume.

$$RT_{60} = 0.34 \ln(V) - 1. \tag{4}$$

This experiment uses publicly available data of 11 rooms (RT_{60} and room volume) [20][21], as shown in Table I.

The GT is defined as ground truth, and TableI shows the error between GT and the values calculated from Eqs. (1) - (4). Volume* represents the volume m^3 calculated from Eq. (1), while Volume** represents the volume calculated from Eq. (4).

TABLE I: Preliminary experiment result.

RT_{60} (GT)	Volume (GT)	Volume**	Volume*	Square Error**	Square Error*
0.08	11.88	23.96	0.57	128.02	145.98
0.21	15.00	35.12	10.22	22.81	404.89
0.33	47.30	50.28	40.40	47.59	8.89
0.37	12.00	56.23	55.92	1929.20	1956.10
0.37	246.00	56.39	56.38	35956.78	35950.65
0.39	48.30	59.63	65.49	295.50	128.47
0.47	99.60	76.12	116.83	296.98	551.16
0.64	202.00	123.67	286.71	7175.96	6135.00
0.65	72.90	126.62	297.63	50504.55	2885.65
0.70	18.00	148.41	378.68	130092.63	17007.59
1.22	370.00	685.00	2004.76	2672431.83	99221.95
			RMSE	154.78	36.86

The results show both the square error and root mean square error (RMSE). As shown in Table I, both Eqs. (1) - (4) yield very large errors. Therefore, we can conclude that we cannot accurately define room volume using the Sabine-based equation alone. This led us to utilize another method to estimate room volume without relying on the Sabine equation and α.

III. Proposed Method

In this section, the proposed framework is explained in detail, including the input features and the architecture of the CNN model.

A. Proposed Framework

In our paper, we present a method for estimating room volume through a blind experiment. The proposed framework uses reverberant speech to assess the room volume, leveraging advanced signal processing and machine learning techniques. Our proposed framework is shown in Fig. 1.

The reverberant speech is input into a seven-octave band filterbank, which covers frequencies of 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz. Each frequency band is processed to extract the TAEs. TAE 1 corresponds to the 125 Hz band, TAE 2 to the 250 Hz band, and so on, up to TAE 7 for the 8 kHz band.

Next, the seven TAEs are combined to form the MTF feature, which has dimensions of 7×240. This MTF feature encapsulates the essential characteristics of the reverberant speech across the different frequency bands, comprehensively representing the acoustic environment.

The MTF feature is fed into a CNN, which is trained to process this high-dimensional input. Then, the CNN estimates the room volume. The deep learning model leverages the rich feature set provided by the MTF, enabling it to learn complex patterns and relationships that correlate with the room's volume.

The proposed framework offers a sophisticated approach to estimate room volume, moving beyond traditional methods that rely heavily on specific coefficients or prior knowledge of the room's acoustic properties. By utilizing reverberant speech and advanced signal processing techniques, our method provides a robust solution for blind estimation of room volume, with potential applications in various real-world scenarios.

For instance, this methodology can be applied to predict a room's current occupancy. By calculating the difference between the actual room volume and the estimated volume

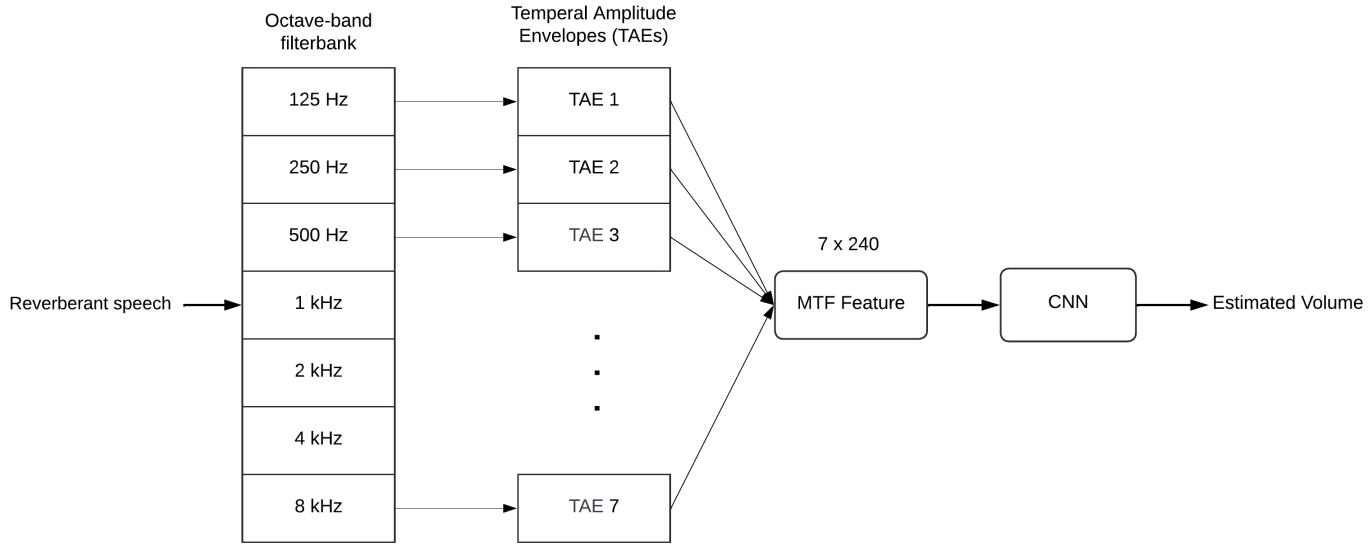


Fig. 1: Proposed framework for blind room volume estimation.

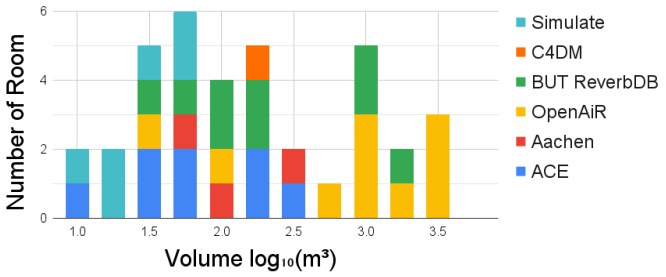


Fig. 2: Numbers of rooms for different room volumes.

(which is influenced by the presence of people), the number of individuals present can be inferred. This application is particularly beneficial for security monitoring, space management, and emergency evacuation planning, showcasing the practical utility of our proposed framework.

B. MTF Feature

The critical input feature in our framework is the MTF feature, a sophisticated representation derived from the temporal characteristics of reverberant speech. The MTF captures the modulation properties of the speech signal across different frequency bands, providing rich, multidimensional input for the subsequent processing stages.

The construction of the MTF feature begins with the extraction of TAEs from the reverberant speech signal. The reverberant speech is processed through a seven-octave band filterbank covering 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz. Each frequency band's TAE is computed, resulting in seven distinct TAEs.

These TAEs capture the amplitude fluctuations over time within their respective frequency bands, reflecting the dynamic characteristics of the reverberant speech signal. Combining these seven TAEs forms the basis of the MTF feature, encapsulating the modulation properties across a broad frequency spectrum.

The MTF feature combines the seven TAEs into a single multidimensional representation. Specifically, each TAE is represented as a vector of length 240, capturing the temporal modulation details at a high resolution. The concatenation of these seven TAE vectors results in the MTF feature with dimensions of 7×240. This high-dimensional feature vector comprehensively represents the temporal modulation characteristics of the reverberant speech signal across the different frequency bands.

The MTF feature is the input to our CNN, designed to process this rich, multidimensional input and estimate the room volume accurately. The CNN leverages the detailed modulation information encapsulated in the MTF feature, enabling it to learn complex patterns and relationships that correlate with the room's volume.

C. CNN Model Architecture

The CNN architecture has been utilized in this work because it can produce two-dimensional input signals. Figure 3 shows the CNN architecture with MTF input and one single output. The model has a total of four convolutional layers followed by an average pooling layer. The first two convolutional layers are followed by batch normalization. Batch normalization helps accelerate the training process by allowing the network to use higher learning rates. Also, batch normalization makes the training process more stable [22]. For more details on the CNN architecture model, visit our [1]GitHub repository.

IV. Experimental Setup

A. Dataset

As shown in Fig. 2, five publicly available datasets of real-world room impulse responses (RIRs), including 31 actual

[1]https://github.com/winnerrock/APSIPA2024

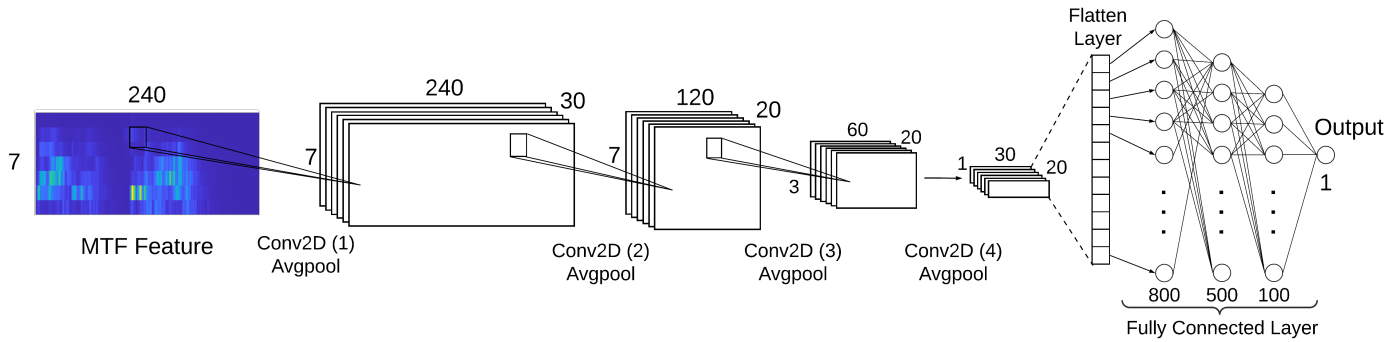


Fig. 3: Proposed CNN architecture model with the MTF feature.

rooms, are selected to cover a wide range of realistic acoustic room parameters. The RIR data collection locations include offices, meeting rooms, lecture rooms, churches, and more. These datasets contain the ACE challenge dataset [20], the Aachen Impulse Response (AIR) dataset [21], the Brno University of Technology Reverb database (BUT ReverbDB) [23], the OpenAIR dataset [24], and the C4DM dataset (C4DM) [25], which cover room volumes between 12 and 4500 cubic meters. All RIRs are resampled to have a sampling rate of 16 kHz. Moreover, we simulate six additional rooms and RIRs on the basis of the image-source method and head-related transfer function (HRTF) interpolation [26] to compensate for the lack of data range.

V. Experimental Setup and Results

In this experiment, we split the dataset into a training and evaluation set. The training set includes 25 rooms, with 4 rooms being simulated to address missing volume data. The evaluation set consists of 10 rooms selected from the 31 actual rooms, covering a volume range from 40 to 3500 cubic meters, ensuring comprehensive coverage of the overall dataset.

A. Evaluation Metrics

The room volume estimation problem was formulated as a regression task on the $\log_{10}$ of the room volume, ensuring that the estimation error was proportional to the magnitude of the room size. Due to the significant variability in room sizes, a logarithmic approach was deemed more appropriate than a linear one. The performance evaluation involved three widely used metrics: mean squared error (MSE), mean error (bias), and Pearson's correlation coefficient (ρ). Additionally, this study utilized a specific metric that is based on the mean absolute logarithm of the ratio between the estimated volume $\hat{V}$ in m^3 and the ground truth volume V. The MSE is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{5}$$

where N is the number of observations in the dataset, y_i is the actual value of the observation, and $\hat{y}_i$ is the predicted value of the i-th observation [27].

The mean absolute error (MAE) is determined by:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{6}$$

where $|y_i - \hat{y}_i|$ represents the absolute error.

Pearson's correlation coefficient (ρ) is computed using the formula:

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{7}$$

where ρ is the correlation coefficient, x_i are the values of the ground truth in a sample, $\bar{x}$ is the mean of the ground truth, y_i are the values of the estimate value in a sample, and $\bar{y}$ is the mean of the estimate value values [28].

The mean absolute logarithm of the ratio between the estimated volume and the ground truth is (MM), which is determined by:

$$\text{MM} = e^{\frac{1}{N}\sum_{i=1}^{N}\left|\ln\left(\frac{\hat{V}_i}{V_i}\right)\right|}, \tag{8}$$

where N denotes the number of test samples and $\ln$ is the natural logarithm. This metric summarizes the error by calculating the average multiple of the estimated volume relative to the actual volume. We use this evaluation to directly compare our results with those from the baseline study [7].

B. Results

In this section, we present a detailed analysis of the performance of our proposed method for blindly estimating room volume from reverberant speech. The results are compared with those of the method proposed by Genovese *et al.* [7] to highlight the improvements and efficacy of our approach.

Table II summarizes the performance metrics, including MSE, MAE, ρ, and MM. Our proposed method significantly reduced MSE, achieving a value of 0.139, compared to 0.190 by Genovese *et al.* This indicates that our method more accurately estimates volume with minor deviations from the actual values.

Additionally, our method achieves a correlation coefficient ρ of 0.792, which is considerably higher than the 0.39 obtained by Genovese *et al.*, suggesting a stronger linear relationship

between the estimated and actual volumes. The MM value of 1.148 further corroborates the accuracy of our estimates, being closer to 1, which implies less deviation from the actual room volumes.

TABLE II: Error comparison between the baseline and our proposed method.

Methods	MSE	MAE	ρ	MM
Genovese *et al.*[7]	0.190	-	0.390	2.280
Proposed Method	0.139	0.322	0.792	1.148

Figure 5 presents a scatter plot of the predicted room volumes versus the ground truth volumes. The distribution of points along the diagonal line indicates a high degree of correlation, reinforcing the quantitative findings. Most predictions are closely aligned with the actual values, showing the reliability of our method.

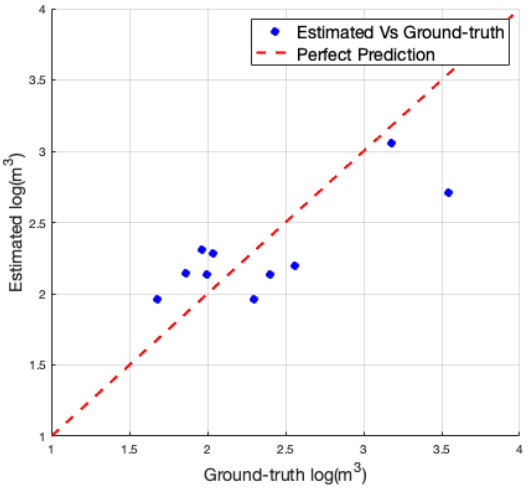


Fig. 4: Results of estimate volumes and ground truth volumes.

Figure 6 provides a bar chart comparing the logarithmic estimates of room volumes with the ground truth across various room sizes. The close match between the estimated and actual values across different volume ranges demonstrates the robustness of our method. This consistency across different room sizes indicates that our approach can generalize well to diverse acoustic environments.

VI. Discussion

The results demonstrate that our proposed method can effectively estimate room volume blindly from reverberant speech. The performance metrics, including MSE, ρ, and MM, indicate good results of 0.139, 0.792, and 1.148, respectively. Our approach outperforms the baseline method by Genovese *et al.* [7], although it is important to note that our work is currently based on reverberant speech only, not noisy speech. We plan to extend our method to handle noisy speech in the following research phase. We might also use other room acoustic parameters for the feature.

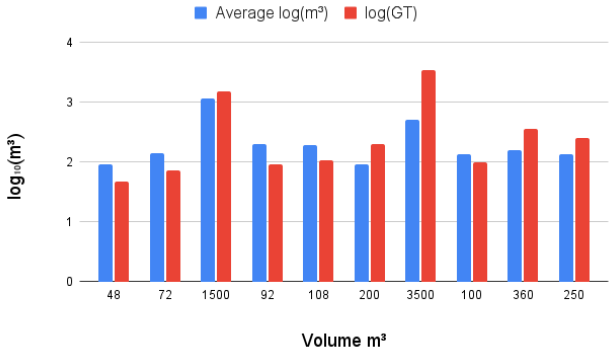


Fig. 5: Comparison between estimate and ground truth room volume.

VII. Conclusion

This paper presented a method for blindly estimating room volume from reverberant speech. By leveraging the modulation transfer function (MTF) feature in conjunction with a CNN, our method significantly improved accuracy and reliability compared to the baseline method according to the experiment results. This is evident from notable reductions in MSE, MM, and MAE and a higher ρ. Qualitative evaluations further reinforced these findings, showing a solid alignment between predicted and actual room volumes.

However, our method is based on reverberant speech rather than noisy reverberant speech. In future work, we will focus on noisy reverberant speech. Also, additional acoustic features could be added to improve the robustness and accuracy of room volume estimation. Expanding the dataset to include more diverse rooms could help generalize and improve the model. Integrating occupancy estimation techniques to assess the number of people in a room on the basis of deviations from the known empty room volume could provide valuable insights for building management and security systems.

In conclusion, by providing a framework for blindly estimating room volume from reverberant speech, our proposed method offers a significant step forward in room acoustic analysis. With continued research and development, this approach has the potential to estimate other room acoustic parameters simultaneously. This method will enhance the capabilities of acoustic scene analysis, sound field control, and architectural acoustics.

Acknowledgments

This work was supported by grants from the SIIT-JAIST-NSTDA Dual Doctoral Degree Program and Grant-in-Aid for Transformative Research Areas (A) (23H04344).

References

[1] R. Michael, *Model-based sound design and audio signal processing using State-space frequency responses*. Springer, 2019.

[2] F. Jacobsen, *Elements of Acoustical Engineering*. Springer, 2013.

[3] F. A. Everest and K. C. Pohlmann, *Master Handbook of Acoustics*. McGraw-Hill Education, 2014.

[4] O. Robin, A. Berry, O. Doutres, and N. Atalla, "Measurement of the absorption coefficient of sound absorbing materials under a synthesized diffuse acoustic field," vol. 136, no. 1, EL13–EL19, 2014.

[5] N. R. Shabtai, Y. Zigel, and B. Rafaely, "Towards room-volume classification from reverberant speech using room-volume feature extraction and room-acoustics parameters," *Acta Acustica United with Acustica*, vol. 99, no. 4, pp. 658–669, 2013.

[6] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[7] A. F. Genovese, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Blind room volume estimation from single-channel noisy speech," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 231–235.

[8] P. Srivastava, A. Deleforge, and E. Vincent, "Blind room parameter estimation using multiple multichannel speech recordings," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2021, pp. 226–230.

[9] C. Wang, M. Jia, M. Li, C. Bao, and W. Jin, "Attention is all you need for blind room volume estimation," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 1341–1345.

[10] S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response," *Applied Acoustics*, vol. 185, p. 108 372, 2022.

[11] T. Houtgast, H. J. M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. i. general room acoustics," *Acta Acustica united with Acustica*, vol. 46, no. 1, pp. 60–72, 1980.

[12] M. Unoki, A. Miyazaki, S. Morita, and M. Akagi, "Method of blindly estimating speech transmission index in noisy reverberant environments.," *J. Inf. Hiding Multim. Signal Process.*, vol. 8, no. 6, pp. 1430–1445, 2017.

[13] M. R. Schroeder, "Modulation transfer functions: Definition and measurement," *Acta Acustica united with Acustica*, vol. 49, no. 3, pp. 179–182, 1981.

[14] T. Houtgast and H. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acta Acustica United with Acustica*, vol. 28, no. 1, pp. 66–73, 1973.

[15] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[16] M. Unoki and Z. Zhu, "Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 233–244, 2020.

[17] L. Wang, S. Duangpummet, and M. Unoki, "Blind estimation of speech transmission index and room acoustic parameters by using extended model of room impulse response derived from speech signals," *IEEE Access*, vol. 11, pp. 49 431–49 444, 2023. DOI: 10.1109/ACCESS.2023.3276327.

[18] F. R. Kschischang, "The hilbert transform," *University of Toronto*, vol. 83, p. 277, 2006.

[19] M. Kuster, "Reliability of estimating the room volume from a single room impulse response," vol. 124, no. 2, pp. 982–993, 2008.

[20] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ace challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[21] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *2009 16th International Conference on Digital Signal Processing*, IEEE, 2009, pp. 1–5.

[22] W. Jung, D. Jung, B. Kim, S. Lee, W. Rhee, and J. H. Ahn, "Restructuring batch normalization to accelerate cnn training," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 14–26, 2019.

[23] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[24] D. T. Murphy and S. Shelley, "Openair: An interactive auralization web resource and database," in *Audio Engineering Society Convention 129*, Audio Engineering Society, 2010.

[25] R. Stewart and M. Sandler, "Database of omnidirectional and b-format room impulse responses," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 165–168.

[26] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," vol. 138, no. 2, pp. 708–730, 2015.

[27] GeeksforGeeks, *Mean squared error*, Accessed: 2024-07-13. [Online]. Available: https://www.geeksforgeeks.org/mean-squared-error/.

[28] N. University, *Strength of correlation*, Accessed: 2024-07-13. [Online]. Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/strength-of-correlation.html.

Semi-Supervised Far-Field Speaker Verification with Distance Metric Domain Adaptation

Han Wang*, Mingrui He*, Mingjun Zhang*, Longting Xu*
* School of Information Science and Technology, Donghua University, Shanghai, China
E-mail: wanghan65267@gmail.com, hemingrui2@outlook.com, zmj@mail.dhu.edu.cn, xlt@dhu.edu.cn

***Abstract*—Achieving high performance in far-field speaker verification is challenging due to the scarcity of data and prevalence of low-quality labels. A common approach is to use a semi-supervised learning framework. A pre-trained near-field speaker model extracts embeddings of unlabeled far-field speakers, and clustering methods generate pseudo-labels for fine-tuning. The accuracy of pseudo-label generation through clustering and the domain adaptation during fine-tuning are critical to this semi-supervised learning framework. In this paper, we propose a method that combines Uniform Manifold Approximation and Projection (UMAP) with a community detection algorithm for efficient dimensionality reduction and clustering. UMAP preserves the structural distribution of high-dimensional data, facilitating better clustering. We conduct domain alignment in the distance metric space to achieve domain adaptation during fine-tuning, referring to this approach as distance metric domain adaptation (DMDA). We evaluate the effectiveness of the dimensionality reduction community detection algorithm using the VoxCeleb (pre-trained) and FFSVC (fine-tuned) datasets. Experimental results demonstrate that the accuracy of generating pseudo labels is improved by using the reduced dimensional community detection algorithm. Additionally, the DMDA method effectively reduces domain mismatch issues during the fine-tuning process.**

I. Introduction

Speaker verification (SV) aims to verify whether two utterances are spoken by the same person [1], [2]. In recent years, deep learning has achieved significant success in speaker verification tasks [3], [4]. Speaker verification systems have shown remarkable improvement from the traditional i-vector method [5] to the DNN-based X-vector method [6]. Training X-vector-based speaker verification networks requires a large amount of well-labeled data. Due to difficulties in data recording and labeling, as well as noise interference, the performance of SV systems significantly degrades in far-field environments. Semi-supervised learning (SSL) [7] is commonly used to reduce the reliance on labeled data. A common SSL approach in far-field speaker verification involves first training an SV model using abundant near-field data. The well-trained near-field SV model is then used to extract speaker embeddings, which are clustered to obtain pseudo-labels. Finally, supervised fine-tuning is performed on the SV model using both pseudo-labeled far-field data and near-field data. However, the accuracy of the pseudo-labels largely depends on the robustness of the clustering algorithm and the reliability of the pre-trained model from the source domain. Additionally, domain mismatch between near-field and far-field data during fine-tuning also significantly affects the accuracy of the SV model.

Existing clustering methods typically use unsupervised approaches, such as K-means [8], Spectral Clustering [9], and Agglomerative Hierarchical Clustering (AHC) [10]. These methods rely on specific assumptions and require estimating the value of k in advance. For clustering large-scale datasets, traditional unsupervised methods and recent approaches like Graph Convolutional Networks (GCN) [11] and community detection algorithms [12] are commonly used. Chen et al. [13] propose a semi-supervised learning approach to improve speaker identification accuracy by label propagation on a graph encoding pairwise similarity for all labeled and unlabeled utterances. Tong et al. [14] apply graph convolutional networks to the constructed affinity graph, fully utilizing the local subgraph information to obtain the representation of speaker embeddings. Recently, community detection algorithms [15]–[17] have also been performing well in large-scale unsupervised speaker clustering tasks. Chen et al. [18] utilize Infomap to perform clustering and obtain the initial pseudo labels. The Leiden [17] algorithm also ensures that all communities are well-connected in speaker clustering.

To reduce the domain mismatch problem, recent approaches have been proposed in the field of speaker verification. One common approach is to improve the fine-tuning strategies. X. Qin et al. [19] use both source and target domain data to fine-tune the pre-trained model jointly. Y. Zheng et al. [20] reserve the speaker weights in the pre-training stage, which are subsequently utilized in the fine-tuning stage. Another approach focuses on enhancing the back-end processing techniques in the SV system. Unsupervised PLDA [21] adapts the PLDA model's covariance matrices to align with domain-invariant data. Aligning embedding vectors across multiple domains by minimizing inter-domain cosine distance or maximizing the mean discrepancy (MMD) [22] is also a common domain adaptation method. Recently, methods that align the distribution of distances within and between speakers have been shown to better match the data distribution for the SV task [23], [24]. Additionally, learning domain-invariant speaker representations [25]–[27] is another notable method for reducing domain mismatch.

In this paper, we employ the TAU community detection algorithm, combined with Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction to achieve the most efficient and accurate clustering results. To mitigate the issue of domain mismatch, we propose distance metric domain adaptation (DMDA), conducting domain alignment in

the distance metric space and to use Discriminative Joint Probability Maximum Mean Discrepancy (DJP-MMD) to measure the intra-speaker and inter-speaker differences. Experimental results on the FFSVC dataset demonstrate the improvement in clustering performance and the effectiveness of the DMDA method in reducing domain mismatch during the fine-tuning process.

II. Methods

A. Uniform Manifold Approximation

UMAP is a dimensionality reduction algorithm based on graph theory and manifold learning that maps high-dimensional data to low-dimensional space for visualization and analysis. The algorithm relies on a theoretical framework involving Riemannian geometry and algebraic topology. UMAP learns a mapping from the high-dimensional data distribution to a low-dimensional space, preserving the significant topological structures of the original high-dimensional space. We employ UMAP to learn the underlying manifold structure and construct a weighted k-nearest neighbor graph.Let $X = \{x_1, x_2, \ldots, x_n\}$ denote the extracted unlabeled speaker embeddings, where the connectivity of the neighborhood graph is determined by the edge weights. To construct the weighted k-nearest neighbor graph, we can calculate the weights between the speaker embeddings and their k nearest neighbors as follows:

$$w\left(\left(x_i, x_{i_j}\right)\right) = \exp\left(\frac{-\max\left(0, d\left(x_i, x_{i_j}\right) - \rho_i\right)}{\sigma_i}\right) \quad (1)$$

where $d(x_i, x_{i_j})$ denotes the distance between embeddings x_i and x_{i_j}, ρ_i is a distance threshold for embedding x_i, and σ_i is a scaling parameter for embedding x_i.

After approximating the manifold in the high-dimensional space, the next step in UMAP is to project it into a low-dimensional space. Once the manifold structure has been learned, it is necessary to define a minimum distance between the embedded points, to avoid multiple overlapping points. UMAP can find a good low-dimensional manifold representation by minimizing the cross-entropy cost function. Denoting the adjacency matrix of the graph as E, our ultimate goal is to find the optimal edge weights for the low-dimensional representation. These optimal weights are obtained by minimizing the cross-entropy function:

$$CE = \sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right) \quad (2)$$

where $w_h(e)$ and $w_l(e)$ denote the high-dimensional and low-dimensional edge weights, respectively.

B. TAU Community Detection Agorithm

We use the TAU community detection algorithm for clustering. This algorithm is an efficient method for modularity optimization, combining the greedy optimization approaches of Louvain and Leiden with genetic algorithms to better explore the solution space. We set each speaker embedding as a node in the community detection network, using the cosine similarity scores between embeddings as edge weights. The TAU algorithm utilizes the Leiden algorithm to detect communities in the input graph through modularity optimization. The modularity optimization function used by Leiden is expressed as follows:

$$\mathcal{Q} = \frac{1}{2m} \sum_i \sum_j \left(A_{ij} - \frac{k_i k_j}{2m}\right) \delta\left(C^{(i)}, C^{(j)}\right) \quad (3)$$

where $\mathcal{Q}$ is the modularity score, m is the total number of edges, A_{ij} represents the elements of the adjacency matrix, k_i and k_j are the degrees of nodes i and j, respectively, and $\delta(C^{(i)}, C^{(j)})$ is 1 if nodes i and j are in the same community, and 0 otherwise.

After the initialization step of generating initial individual partitions, the algorithm iteratively performs the following steps:

Step1. Applying Leiden to optimize each individual partition.

Step2. Evaluating fitness and selecting individuals.

Step3. Performing crossover and mutation on the population.

Step4. Migrating new individuals into the population.

C. Distance Matric Domain Adaptation

As shown in Fig. 1, We aim to reduce the domain mismatch issues in the SV system during the fine-tuning process.

1) Pre-training: In the pre-training stage, we train a deep speaker embedding model using out-of-domain data. To enhance the model's discrimination ability and improve generalization, we employ the angular margin softmax (AM-softmax) [28] loss function.

2) DJP-MMD for domain adaptation: The issue of label mismatch between the source domain and the target domain cannot be resolved merely by aligning distributions at the feature level, which may prevent the model from learning domain-invariant features. After obtaining the pre-trained SV model, we use the speaker labels from the source domain and the pseudo-labels from the target domain to construct sample pairs both within and between speakers. We then use cosine distance to estimate the distance distribution for these sample pairs. We measure the distance between the source domain and the target domain distributions using DJP-MMD.

$$d(\mathcal{D}_s, \mathcal{D}_t) = \mathcal{M}_T - \mu \mathcal{M}_D, \quad (4)$$

where $\mu > 0$ is a trade-off parameter. M_T measures the transferability between the same class in different domains, and M_D measures the discriminability between different classes in different domains.

We compute M_T using MMD, aiming to measure the overall distribution difference between the source domain $p(s)$ and the target domain $p(t)$.

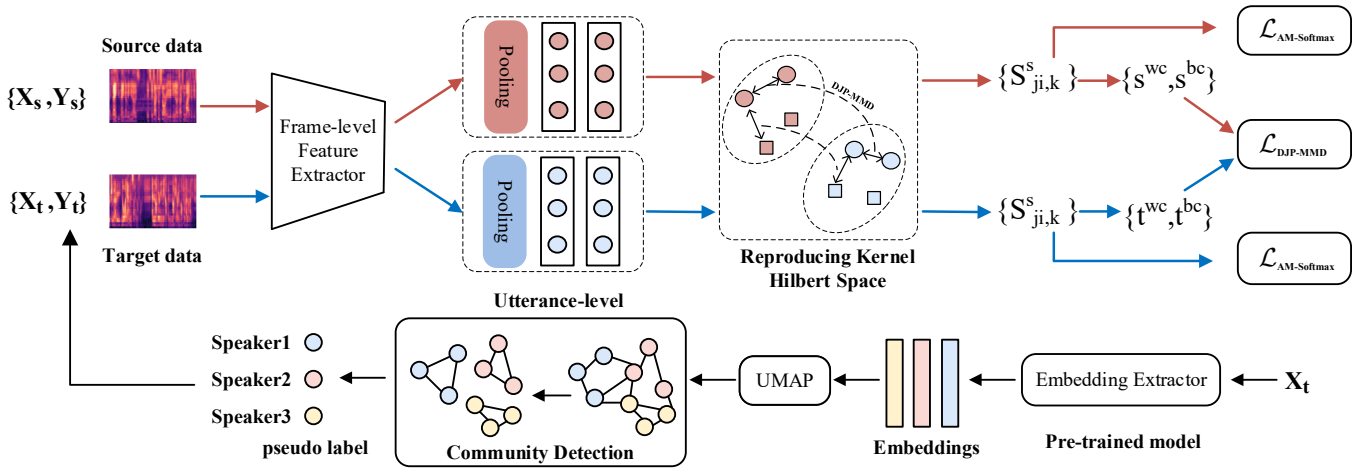


Fig. 1. Overview of the proposed DMDA method with pairwise-distance distribution alignment.

$$
\begin{aligned}
M_T(p(s),p(t)) =& \frac{1}{T_s^2}\sum_{n=1}^{T_s}\sum_{m=1}^{T_s} k(s_n,s_m) \\
&+ \frac{1}{T_t^2}\sum_{n=1}^{T_t}\sum_{m=1}^{T_t} k(t_n,t_m) \\
&- \frac{2}{T_sT_t}\sum_{n=1}^{T_s}\sum_{m=1}^{T_t} k(s_n,t_m)
\end{aligned} \tag{5}
$$

where T_s and T_t are the numbers of samples from the source and target domains in a mini-batch. $k(s,t)$ denotes the radial basis function (RBF) kernel.

$$
\begin{aligned}
M_T(p(s),p(t)) =& \frac{1}{C}\sum_{c=1}^{C}\Bigg(\frac{1}{T_{s,c}^2}\sum_{n=1}^{T_{s,c}}\sum_{m=1}^{T_{s,c}} k(s_{c,n},s_{c,m}) \\
&+ \frac{1}{T_{t,c}^2}\sum_{n=1}^{T_{t,c}}\sum_{m=1}^{T_{t,c}} k(t_{c,n},t_{c,m}) \\
&- \frac{2}{T_{s,c}T_{t,c}}\sum_{n=1}^{T_{s,c}}\sum_{m=1}^{T_{t,c}} k(s_{c,n},t_{c,m})\Bigg)
\end{aligned} \tag{6}
$$

where C is the total number of classes. $T_{s,c}$ and $T_{t,c}$ are the numbers of samples from the source and target domains in class. $s_{c,n}$ and $s_{c,m}$ are samples from the source domain in class. $t_{c,n}$ and $t_{c,m}$ are samples from the target domain in class.

The vectors s and t can represent either within-class or between-class distance representations. After computing the two DJP-MMD distances, we obtain the final DJP-MMD loss as follows:

$$
\begin{aligned}
\mathcal{L}_{\text{DJP-MMD}} =& \mathcal{L}_{\text{DJP-MMD}}(p(\mathbf{s}^{wc}),p(\mathbf{t}^{wc})) \\
&+ \mathcal{L}_{\text{DJP-MMD}}(p(\mathbf{s}^{bc}),p(\mathbf{t}^{bc}))
\end{aligned} \tag{7}
$$

where $\mathbf{s}^{wc}$ and $\mathbf{t}^{wc}$ are the within-class distance vectors from the source and target domains, respectively. $\mathbf{s}^{bc}$ and $\mathbf{t}^{bc}$ are the between-class distance vectors from the source and target domains, respectively.

The overall objective of DMDA is the weighted sum of the AM-softmax loss and the DJP-MMD losses:

$$
\mathcal{L}_{\text{DMDA}} = \mathcal{L}_{\text{DJP-MMD}} + \lambda\mathcal{L}_{\text{AM-softmax}} \tag{8}
$$

III. EXPERIMENTS

A. Datasets and Evaluation Metrics

We conduct experiments on the VoxCeleb 1&2 [29] and FFSVC [30] datasets. The VoxCeleb 1&2 dataset serves as the large-scale dataset for pre-training near-field models. The FFSVC2022 supplement is used as the unlabeled far-field target domain dataset. We evaluate the model on the development trials and evaluation trials of FFSVC2022. We strictly adhere to the competition requirements of FFSVC2022 Task 2, where adjustments to hyperparameters and testing model performance are only allowed on the development set. All data used is consistent with the official baseline for FFSVC2022. All trial pairs are single-channel speech segments. The evaluation set consists of a large trials file and anonymized audios. The experimental results are evaluated using the Equal Error Rate (EER) and the minimum of the normalized detection cost function (minDCF) with $P_{target} = 0.01$.

B. Implementation Details

We extract 80-dimensional log-mel filter bank features. The frame length for all features is set to 25ms with a frame shift of 10ms. Furthermore, we perform mean normalization within a sliding window of three seconds.

In the pre-training stage, the chunk-size is set to 200 for sampling. We utilize the AdamW optimizer [31] with a momentum of 0.9 and a weight decay of 5e-1. Additionally, the margin gradually increases from 0 to 0.2. We use 1 GPU with a mini-batch size of 512 and an initial learning rate of 0.01 to train all of our models.

For community detection, we first use UMAP to reduce the dimensionality by setting the number of neighbors in UMAP to 20, allowing it to find low-dimensional manifolds in the high-dimensional space. The UMAP algorithm reduces the dimensionality to 60 in the new feature space. We use cosine distance as the distance metric to compute sample similarities, applying it consistently across all clustering methods. We extract 40, 80, and 120 speakers from the supplement dataset to compare the clustering performance under different numbers of speakers.

For domain adaptation, we use a speaker-balanced sampling strategy to construct intra-speaker and inter-speaker pairs, ensuring that each speaker in a mini-batch contains the same number of utterances. The batch size is set to 128, with each speaker having 4 utterances. ResNet34 is chosen as the base backbone. Cosine similarity is used for backend scoring, and all experiments are conducted on ASV-subtools [32] and Wespeaker [33].

IV. RESULTS AND DISCUSSIONS

A. The Effectiveness of community detection

As shown in Table 1, we employ four different clustering methods with varying numbers of speakers and evaluate their performance based on average precision, recall, and F-score. When the number of speakers is 40, the traditional K-means method performs well, achieving accuracy surpassing that of the basic community detection algorithm. When the number of speakers increases to 80, the average precision of traditional K-means clustering decreases significantly. At 120 speakers, the advantage of the community detection algorithm becomes more apparent. The best clustering results are achieved when TAU and UMAP are used together for dimensionality reduction, and this method maintains good performance as the number of speakers increases.

TABLE I
COMPARISON OF SPEAKER CLUSTERING PERFORMANCE FOR VARIOUS NUMBER OF SPEAKERS

Nums	Methods	Precision	Recall	F-score
40	K-means	0.88	0.79	0.80
	Infomap	0.84	0.77	0.80
	Leiden	0.85	0.78	0.82
	TAU	0.86	0.79	0.80
	UMAP+TAU	0.90	0.83	0.87
80	K-means	0.85	0.77	0.81
	Infomap	0.83	0.74	0.80
	Leiden	0.84	0.76	0.82
	TAU	0.85	0.77	0.82
	UMAP+TAU	0.89	0.81	0.85
120	K-means	0.82	0.72	0.74
	Infomap	0.83	0.76	0.77
	Leiden	0.85	0.76	0.79
	TAU	0.87	0.78	0.82
	UMAP+TAU	0.89	0.80	0.84

B. The Effectiveness of DMDA

The experimental results are shown in Table 2. The table compares the performance of different methods on the FFSVC2022 development and evaluation datasets in terms of EER and minDCF. We compare the DMDA method with the FFSVC official baseline, the methods proposed by other participating teams, and the Unsupervised PLDA approach. Vanilla fine-tuning (using only the near-field dataset for fine-tuning) shows improvement over the pre-trained model. This indicates that fine-tuning the near-field speaker verification (SV) model with a far-field dataset is an effective approach. However, this process may lead to overfitting and catastrophic forgetting. The use of domain adaptation techniques further enhances the SV model's performance. Unsupervised PLDA effectively reduces domain mismatch problems. When we use our proposed DMDA method, the performance improves further. Notably, we achieve the best results with the DJP-MMD distance.

TABLE II
PERFORMANCE COMPARISON OF DMDA AND OTHER METHODS ON FFSVC2022 DATASETS

Method	FFSVC2022 dev		FFSVC2022 eval	
	EER%	minDCF	EER%	minDCF
Official baseline	7.79	0.746	7.64	0.739
rank2 HiMia	5.59	0.563	5.34	0.566
rank1 SPEAKIN	5.98	0.500	6.21	0.523
Pre-trained	8.96	0.751	-	-
Vanilla fine-tuning	6.74	0.749	7.14	0.695
Unsupervised PLDA	6.15	0.614	6.59	0.644
DMDA(MMD)	5.75	0.524	5.72	0.558
DMDA(DJP-MMD)	**5.24**	**0.507**	**5.21**	**0.535**

V. CONCLUSIONS

In this paper, we present a novel approach to enhance far-field speaker verification through a semi-supervised learning framework. By integrating Uniform Manifold Approximation and Projection (UMAP) with a community detection algorithm, we achieve efficient dimensionality reduction and clustering while preserving the structural distribution of high-dimensional data. Furthermore, we propose a distance metric domain adaptation (DMDA) method to address domain mismatch issues during fine-tuning by aligning domains in the distance metric space. We compare the accuracy of various clustering methods with different numbers of speakers and contrast the DMDA method with several domain adaptation techniques. Experimental results demonstrate that our proposed method effectively improves the performance of far-field speaker verification.

ACKNOWLEDGMENT

This work is supported by the Donghua University Language and Text Application Research Project (No. Y2024-8).

REFERENCES

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, *et al.*, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, pp. 1–22, 2004.

[2] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5115–5119.

[3] G. Bhattacharya, J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," in *Proc. Interspeech 2017*, 2017, pp. 1517–1521.

[4] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[7] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020.

[8] G. Hamerly and C. Elkan, "Learning the k in k-means," *Advances in neural information processing systems*, vol. 16, 2003.

[9] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[10] S. Zhou, Z. Xu, and F. Liu, "Method for determining the optimal number of clusters based on agglomerative hierarchical clustering," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 12, pp. 3007–3017, 2016.

[11] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "Simple and deep graph convolutional networks," in *International conference on machine learning*, PMLR, 2020, pp. 1725–1735.

[12] A. Lancichinetti and S. Fortunato, "Community detection algorithms: A comparative analysis," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 80, no. 5, p. 056 117, 2009.

[13] L. Chen, V. Ravichandran, and A. Stolcke, "Graph-Based Label Propagation for Semi-Supervised Speaker Identification," in *Proc. Interspeech 2021*, 2021, pp. 4588–4592.

[14] F. Tong, S. Zheng, M. Zhang, *et al.*, "Graph convolutional network based semi-supervised learning on multi-speaker meeting data," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6622–6626.

[15] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the national academy of sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[16] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, P10008, 2008.

[17] V. A. Traag, L. Waltman, and N. J. Van Eck, "From louvain to leiden: Guaranteeing well-connected communities," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[18] Z. Chen, J. Wang, W. Hu, L. Li, and Q. Hong, "Unsupervised speaker verification using pre-trained model and label correction," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[19] X. Qin, D. Cai, and M. Li, "Far-Field End-to-End Text-Dependent Speaker Verification Based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation," in *Proc. Interspeech 2019*, 2019, pp. 4045–4049.

[20] Y. Zheng, J. Peng, Y. Chen, *et al.*, "The speakin speaker verification system for far-field speaker verification challenge 2022," *arXiv:2209.11625*, 2022.

[21] N. Brummer, A. Mccree, S. Shum, D. Garcia-Romero, and C. Vaquero, "Unsupervised Domain Adaptation for I-Vector Speaker Recognition," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2014)*, 2014, pp. 260–264.

[22] J. Huang and T. Bocklet, "Intel Far-Field Speaker Recognition System for VOiCES Challenge 2019," in *Proc. Interspeech 2019*, 2019, pp. 2473–2477. DOI: 10.21437/Interspeech.2019-2894.

[23] L. Yi and M. W. Mak, "Cross-domain adaptation in distance space for speaker verification," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2023, pp. 2238–2243.

[24] J. Li, J. Han, and H. Song, "Cdma: Cross-domain distance metric adaptation for speaker verification," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7197–7201.

[25] Y. Wei, J. Du, H. Liu, and Z. Zhang, "Centriforce: Multiple-domain adaptation for domain-invariant speaker representation learning," *IEEE Signal Processing Letters*, vol. 29, pp. 807–811, 2022.

[26] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4889–4893.

[27] W. Huang, B. Han, S. Wang, Z. Chen, and Y. Qian, "Robust cross-domain speaker verification with multi-level domain adapters," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 781–11 785.

[28] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[29] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[30] X. Qin, M. Li, H. Bu, S. Narayanan, and H. Li, "The 2022 far-field speaker verification challenge: Exploring domain mismatch and semi-supervised learning under the far-field scenario," *arXiv preprint arXiv:2209.05273*, 2022.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.

[32] F. Tong, M. Zhao, J. Zhou, *et al.*, "Asv-subtools: Open source toolkit for automatic speaker verification," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6184–6188.

[33] H. Wang, C. Liang, S. Wang, *et al.*, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

IoT-based Smart Attendance System using Face Recognition and Motion Detection

Umi Syamimi Saadon*, Chern Hong Lim* and Lillian Yee Kiaw Wang *
* School of Information Technology, Monash University, Malaysia
E-mail: {umisyamimi.saadon, Lim.ChernHong, Lillian.Wang}@monash.edu

***Abstract*—Managing student attendance is crucial for effective classroom administration. A robust attendance system streamlines the attendance process and ensures accuracy. Many universities and research institutions currently use traditional attendance systems, which possesses flaws such as data inaccuracy, under-utilized attendance records, and vulnerabilities in smart device usage. This research proposes an IoT-based Smart Attendance System using Face Recognition and Motion Detection. The system will employ IoT devices which are the Intel® RealSense™ Depth Camera D455. The integration of facial recognition and motion detection will be explored to model student behaviors in class, including face detection and identification also primitive motion detection (Standing, Sitting, Walking and Raising Hand). Additionally, the research proposed a few pre-defined metrics to assess students' behavior such as Attendance rate, Participation rate, and Engagement level. This ongoing monitoring will link students' behavior with their academic performance. The preliminary outcomes provide encouraging results and showed the feasibility and effectiveness of our proposed IoT framework.**

I. Introduction

Nowadays, Internet of Things (IoT) technologies are being employed across various fields to enhance convenience and efficiency for users. IoT is increasingly being integrated into numerous industries to streamline processes and improve efficiency. Examples of widely used IoT devices include medical sensors, smart security systems, voice controllers, etc. In organisational and university attendance systems, IoT devices such as biometric identification, smart cards, and face recognition are commonly used. Transitioning from traditional attendance systems to automated ones enhances efficiency and reduces human error. IoT is a fundamental component of Industry 4.0, with significant potential to transform business and industrial processes. Its expansion makes devices smarter and more interconnected [1].

This research proposes an IoT-based Smart Attendance System , emphasizing the integration of facial recognition and motion detection for university students' attendance. Relying solely on face recognition for biometric identification limits the system's potential. It fails to provide comprehensive analysis of student behavior in class or generate personalized reports and guidance on student performance. Therefore, motion detection will be incorporated to capture and monitor student movements, such as leaving class early, taking extended breaks, or engaging in class activities. To achieve this, the proposed system will employ high-resolution cameras to capture data accurately. These cameras will provide detailed images and videos to ensure precise monitoring and data collection.

There were many works focused on improving accuracy and efficiency in attendance management systems. Rathinamala et al. [2] introduce the attendance tracking system based on the Viola-Jones Face Detection approach, which uses the HAAR cascade algorithm and the Local Binary Pattern Histogram (LBPH) algorithm to monitor and track the attendance. Chaitanya et al. [3] proposed an efficient attendance management approach using a Bluetooth-based Smart Attendance System coupled with an Android application. The technology aims to interact easily with existing infrastructures and Android handsets, ensuring accuracy and ease of use. Khang et al. [4] introduce a QR code-based attendance system for the purpose. It is designed for easy integration into a mobile app with phone camera. Nonetheless, Yi Yi [5] examine the development and implementation of a student attendance management system that involves fingerprint recognition technology. However, this approach has significant drawbacks, such as environmental factor. Dirt, moisture, and temperature can all have an impact on fingerprint sensor accuracy, potentially causing problems in some circumstances.

Although IoT devices have helped to enhance attendance systems, challenges including maintaining accuracy in a variety of environments, reducing user engagement, and relying too much on external devices still exist. Face recognition offers many advantages over other methods, particularly when integrated with motion detection. Facial recognition systems with an algorithms such as LBPH and HAAR cascade offer automatic monitoring without requiring students to interact with devices like fingerprint sensors or QR codes, which are easily influenced by external factors. Motion detection reduces false attendance and increases accuracy in attendance tracking by recognising student behaviour and engagement patterns. This combination works especially well in dynamic environments like universities, where user non-compliance or environmental conditions can make typical biometric systems unreliable [2][6]. As a result, combining these technologies offers a more reliable and flexible way to address the issues with student attendance. Moreover, Rajawat et al. [6] state that one of the best methods for tracking attendance is a facial recognition system which encouraged us to propose an attendance system that includes face and motion detection.

To enhance the above-mentioned systems, in this paper, we

proposed an integrated IoT-based Smart Attendance System using Face Recognition and Motion Detection. The contributions of this work are as follows: 1) We provide a comparative study on the feasible algorithms and camera setup for the proposed integrated IoT system, combining face recognition and motion detection in the attendance taking process. 2) We present a complete framework with the combination of YOLOv8 and MediaPipe for enhancing the accuracy of motion detection in the proposed IoT system.

II. Proposed Framework

The proposed IoT-based Smart Attendance System consist of four key stages which are input, detection, analysis, and output, as illustrated in Figure 1. In the input stage, the primary purpose is to collect raw data through the Intel® RealSense™ Depth Camera D455. This camera captures video footage of students in the classroom, based on predefined time intervals, which forms the foundational data that will be processed in subsequent stages. The detection module is where the system processes the raw video. It performs three critical tasks which are face extraction using YOLOv8 to isolate student's faces, face recognition to match these extracted faces against a pre-existing database for identification, and motion detection to monitor student's movements. This stage ensures that both attendance and student engagement are accurately detected and tracked. It is then followed by the analysis module to evaluate and interpret the data by calculating the predefined metrics such as attendance rate, class participation rates, and engagement level. Interpreting the patterns offer insights into the factors that influence student absenteeism and engagement. Finally, in the output stage, the processed information will be visualized and presented in a clear, interpretable format. This stage is designed to assist educators and administrators by providing actionable insights that could be helpful in decision-making processes for improving student attendance.

A. Input Stage

To facilitate effective data collection, input stage utilizes the Intel® RealSense™ Depth Camera D455 to capture students' facial structures and monitor movements. Instead of real-time processing, the system records video footage in predefined intervals based on class duration. For instance, a 120-minute class is divided into four 15-minute video segments. This approach reduces the performance challenges typically associated with real-time systems, such as lagging and slow processing speeds, which can hinder the accuracy and responsiveness of the system. By capturing and processing video in this manner, the system ensures a more stable and reliable operation, allowing for thorough analysis without the pressure of immediate processing. The decision to record video segments rather than implementing real-time detection is driven by several factors. Real-time systems often face performance bottlenecks, especially when handling complex tasks like simultaneous facial recognition and motion detection. This can result in delays, lower accuracy, and system instability. Furthermore, as classroom activities and human actions typically span longer

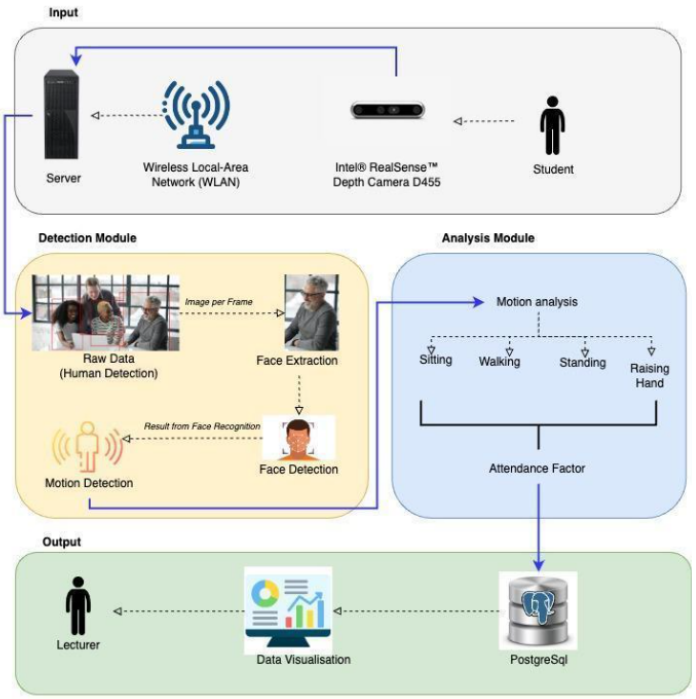


Fig. 1. Propose Framework - IoT-based Smart Attendance System

time intervals, timed video recording ensures that no critical data is lost, allowing for more reliable post-session analysis. This method strikes a balance between system complexity and processing efficiency, ensuring that the system operates within acceptable performance thresholds.

B. Detection Module

In the detection module, three primary processes occur: face extraction, face recognition, and motion detection. Face extraction initiates the module by using YOLOv8 (You Only Look Once) to extract faces from raw pictures or videos [9]. Following extraction, the detection process identifies and compares faces with a database. Subsequently, the module proceeds to perform motion detection, where it analyses changes in the scene to identify and track movements within the environment.

1) Face Extraction using YOLOv8: The two primary parts of YOLOv8 are the head and the backbone. Like previous YOLO models, the backbone's job is to extract features from the input image. To aid in the detection of objects with varied sizes, it is composed of multiple layers arranged into feature pyramid levels (P1 to P5), each of which captures features at a different scale [7].

2) Face Recognition: Face recognition occurs using a combination of the YOLOv8 model and MediaPipe's Holistic model. Initially, the YOLOv8 model predicts bounding boxes around objects, including faces, within each frame captured by the webcam. These bounding boxes and their corresponding classes are extracted from the YOLOv8 predictions. For each detected face, the frame is cropped to focus specifically on the face region using the coordinates from the bounding box.

After YOLOv8 identifies faces, the system utilises MediaPipe's holistic model to detect facial landmarks and potentially other body landmarks. This detection provides additional data about the position and movement of facial features. Once the features are extracted using MediaPipe, it further process the extracted key points from the facial landmarks to a sequence of key points data. This sequence is continually updated and maintained for the last 30 frames to track temporal patterns in the facial movements and expressions for recognition purpose.

3) Motion Detection: In this study, we again utilise MediaPipe for motion detection and employ a self-collected dataset to train the model. In this work, we limit our scope to specifically focused on four fundamental motions which are sitting, standing, walking, and raising a hand. By concentrating on these primitive actions, we aim to enhance the accuracy and reliability of the motion detection system, ensuring robust performances. MediaPipe will be used to detect and track landmarks (e.g: key points representing human poses) in each frame of the video sequence. The detected landmarks will be organised into sequences of fixed length (e.g., 30 frames per sequence). Each sequence represents a segment of continuous motion captured from the video stream.

The dataset comprising 30 sets of images, with each set containing 30 frames. These sequences are organised and labelled for supervised training. The LSTM (Long Short-Term Memory) model architecture is then defined, typically with multiple LSTM layers to capture temporal dependencies across the sequence of frames. During training, the model learns to predict the activity label for each sequence by optimising its weights through back propagation using an appropriate loss function and optimizer. Validation and testing phases assess the model's performance on unseen data to ensure robustness and generalisation. In the processing pipeline, the trained LSTM model processes incoming video streams, segments them into sequences of frames, and makes predictions instantaneously.

Finally, the results of the predictions are captured on the screen, allowing monitoring and analysis of student activities within the classroom environment. The application continues to process frames until the user exits. Overall, the integration of YOLOv8 for face detection and MediaPipe for facial and motion landmark detection, incorporated with machine learning enables comprehensive monitoring and analysis of student behaviour in the smart classroom setting.

To justify, besides MediaPipe which was used to facilitates landmark identification and keypoints in our work, YOLOv7 is found to be able to perform pose estimation, which has been successfully applied in educational environments. The technique is outlined in [10] which incorporates a feature extraction module that effectively extracts and ranks pertinent data by fusing self-attention and convolutional layers. However, in our proposed system, the combination of MediaPipe with YOLOv8 outperformed YOLOv7 (standalone) for pose estimation due to MediaPipe's pre-trained models designed specifically for detecting human movements and gestures. These models provide extensive information about the spatial configuration of body parts across time, which is useful for applications such as motion detection. Table 1 show the comparison of each detection.

TABLE I
COMPARISON OF MEDIAPIPE + YOLOV8 AND YOLOV7 FOR MOTION DETECTION

Feature	MediaPipe + YOLOv8	YOLOv7 (Standalone)
Purpose	Real-time perception tasks, including landmark detection (e.g., poses, hands, faces)	Object detection (identifying and localising objects in images/videos)
Performance	Optimised for real-time inference on mobile and edge devices	High accuracy and efficiency in object detection tasks
Model Available	Pose Estimation, Hand Tracking, Face Detection	Object Detection, Pose Estimation
Accuracy	Provides accurate landmark detection and pose estimation	High accuracy in detecting and localising objects in images/videos
Integration	Easily integrates with other machine learning models and pipelines	Typically integrated into computer vision pipelines for broader applications

C. Analysis Module

Understanding and analysing the factors that contribute to student absenteeism using data collected in detection module provides useful insights into enhancing educational results. Using face recognition and motion detection technology to evaluate attendance allows us to identify patterns and trends that influence student engagement. To efficiently assess student attendance, we defined metrics to calculate the attendance rate, participation rate, and engagement level of the students as shown in Table 2. The attendance rate metric is designed to accurately track and assess student presence throughout the entire class session. In our system, each 2 hours class is divided into four 15-minutes intervals, amounting to a total class time of 60 minutes. This is to ensure that student presence is measured consistently across different periods of the class. The attendance rate is calculated by assessing the time present for each student in each 15-minutes interval. Let *P1,P2,P3,* and *P4* represent the time a student is present in each of the four intervals and *D* refer to class duration which is the total time of four intervals.

The next metric focuses on participation, where it is evaluated by comparing their participation levels across different classes. This metric measures how often a student interacts in a class, with particular emphasis on activities such as raising a hand. In this context, the variable n represents the number of classes, i denotes the interaction count in the i-th class, and Ti refers to the total number of interactions in that class. By examining participation in this way, we can capture how students are engaging with the content and activities of the class. The rationale behind comparing participation across multiple classes is to provide insights into whether a student is consistently engaged or whether their participation fluctuates depending on the class. For example, a student may be passive in one subject but highly active in another. By identifying

TABLE II
UTILISATION OF DATA COLLECTED

Metric	Contribution	Equation	Motion Involve
Attendance	Identify and track students attendance	$A_r = (P_i + P_i + P_i + P_i/D) * 100$ (*Ar*):Attendance Rate (*Pi*): Total time students detected for each interval (*D*):Class Duration	Standing, Sitting, Walking and Raising Hand
Participation	Access students participation across different classes	$\sum_{i=1}^{n} interaction = T_i$ (*n*): Represents the number of classes (*i*): Denotes the interaction count in the i-th class (*Ti*): Total number of interactions	Raising Hand
Engagement Level	Detect movement to infer engagement of activity and monitor students movements in-class	$E_r = (A_m/T_m)$ (*Er*): Represents the engagement ratio (*Am*): Refers to attentive movement (*Tm*): Total movement	Standing, Sitting, Walking and Raising Hand

these patterns, educators can gain a deeper understanding of how different teaching methods, subjects, or classroom environments influence student behavior.

The engagement level metric is specifically designed to observe and measure student engagement within the classroom setting. By detecting movements such as standing, sitting, walking, and raising a hand, this metric aims to provide insight into how physically engaged students are during the class. The engagement ratio *(Er)* reflects how much of the detected movement is directly related to attentive behavior, helping to differentiate between active participation and general physical activity. This focus on in-class engagement allows educators to monitor how students physically respond to the teaching process, helping to gauge real-time participation levels and attention spans. Since classroom engagement is often linked to academic success, this metric provides a valuable tool for understanding and enhancing the student learning experience.

D. Output

In the output stages, the user is presented with the data that has been analysed by the analysis module. This system will provide complete data visualization, ensuring clarity and ease of interpretation. The visualisations will be designed to emphasize numerous factors that influence absenteeism, providing insights into patterns and trends. These insights could include tracking attendance rates over time and factor of absenteeism. Furthermore, the system will provide extensive reports that may be used for additional analysis or administrative purposes, allowing educators and administrators to make decisions about absenteeism prevention and reduction. Overall, the output module attempts to convey actionable information in an understandable and visually attractive style, hence increasing the system's effectiveness in monitoring and improving student attendance.

III. IOT PROTOTYPE DEVICE

To facilitate the IoT framework, a study on several IoT devices and camera which particularly on their functionalities and limitations has been done. Table 3 show a list of IoT devices that will be used in the IoT-based Smart Attendance System . The main purpose of the Intel® RealSenseTM Depth Camera D455 is to serve applications that need precise depth sensing and 3D scanning. Given that this research involves both face and motion detection, these cameras are therefore a perfect fit. Other than that, it is highly helpful in research and education, particularly when computer vision, depth sensing, and 3D technology are required in the classroom settings.

In conjunction, PostgreSQL is used to store data collection. Recognised for its resilience, scalability, and conformity to SQL standards, PostgreSQL is a powerful open-source relational database management system (RDBMS). Because of its durability and dependability, PostgreSQL is an ideal option for managing sensitive data and busy applications. It is also well known for its ability to handle huge quantities of data efficiently.

The Server with Windows 2010 and an Intel Core i9-7980XE processor is well-suited for this smart classroom monitoring project due to its high performance and robust capabilities. The Intel Core i9-7980XE processor, with its 18 cores and 36 threads, offers exceptional processing power and multitasking efficiency, which is essential for handling the intensive data processing requirements of real-time IoT data streams. Coupled with Windows Server 2010, this setup ensures reliable and stable operation for managing large volumes of data collected from multiple IoT devices. The server's high-performance configuration, including 32 GB of RAM and a 1 TB SSD, provides ample memory and fast storage, facilitating smooth data processing, quick retrieval, and efficient database management. This powerful combination ensures that the smart classroom monitoring system operates effectively.

Tableau and Power BI are both highly suitable for this smart classroom monitoring project due to their advanced data visualization and analytics capabilities. Tableau is renowned for its ability to create interactive and visually compelling dashboards that can display complex data in an easily understandable format. Power BI, on the other hand, integrates seamlessly with Microsoft products and offers customizable reporting and real-time data monitoring, making it an excellent choice for institutions already using Microsoft infrastructure. Both tools support a wide range of data sources and provide sophisticated

TABLE III
COMPARISON BETWEEN IOT DEVICES

Device Name	Type	Specification	Purpose	Features
Intel® RealSense™ Depth Camera D455	Depth Camera	Resolution: 1280x720 Depth Range: 0.5 to 6 meters Field of View: 87° x 58° Frame Rate: Up to 90 fps Connectivity: USB 3.1	Capturing depth information and monitoring student movements	High accuracy depth sensing, wide field of view, high frame rate
SQL PostgreSQL	Database	Type: Relational Database Management System (RDBMS) Version: Latest Features: ACID compliance, advanced indexing, complex queries	Storing, managing, and querying collected data	Robust data integrity, scalable performance, complex query support
Server: Windows 2010, Intel Core i9-7980XE	Server	OS: Windows Server 2010 Processor: Intel Core i9-7980XE RAM: 32 GB Storage: 1 TB SSD Network: Gigabit Ethernet	Processing and storing data from IoT devices	High processing power, large memory capacity, fast storage
Tableau or Power BI	Data Visualization	Tableau: Features: Interactive dashboards, real-time data analytics, extensive chart options Power BI: Features: Customise reports, real-time monitoring, integration with Microsoft products	Visualizing and analyzing data collected from the system	Interactive and customizable visualizations, real-time updates, easy integration with various data sources

data analysis features, enabling users to generate insightful reports and visualizations that can help in assessing various aspects of classroom performance.

IV. SYSTEM SETUP IN CLASSROOM

For effective face and motion detection, the room size and camera placement are pivotal. In small to medium rooms (10x10 feet to 15x15 feet), one or two cameras typically placed at a height of 7 to 8 feet, usually suffices. This height balances the angle for capturing faces and detecting motion while avoiding common obstructions. In larger rooms (over 15x15 feet), multiple cameras or one with a wide-angle lens may be necessary to cover the increased area comprehensively. Figure 2 show the propose IoT device setup used in this study.

The higher placement height ensures that the camera's field of view covers more of the room, enhancing detection accuracy. Additionally, using high-resolution cameras and ensuring adequate lighting or using cameras with low-light capabilities will further improve detection performance. Adjustments may be needed based on room layout and specific equipment features to optimise coverage and accuracy.

V. PRELIMINARY RESULT

A. Face Recognition and Motion Detection

Figure 3 presents the results of our activity recognition procedure, highlighting the successful identification of sitting and standing actions. The system demonstrates an operational speed of approximately 8 frames per second during real-time activity recognition. This performance metric ensures a smooth and responsive user experience. The accuracy of the recognition process is significantly improved when a complete human skeleton is detected, as this provides a more comprehensive set of data points for analysis.

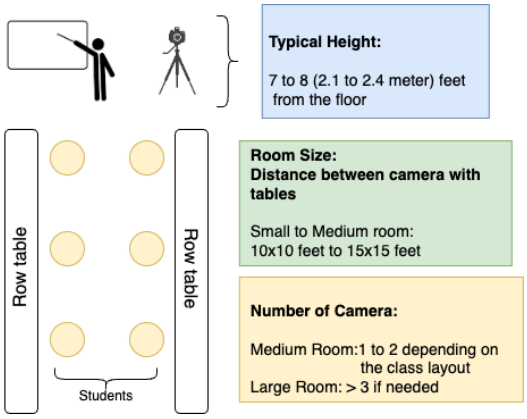


Fig. 2. Proposed Classroom Setting with IoT Devices

Moreover, the system does not only identifies the current activity of the individual but also accurately recognizes the person by name. This dual capability of person and activity identification is achieved through advanced algorithms that integrate skeletal tracking with biometric recognition. The outcomes illustrated in Figure 3 underscore the system's robustness and reliability in discerning both the identity and activities of individuals within the monitored environment.

B. Data Collection

The motion detection results for each student will be properly documented and saved to a PostgreSQL database. This data collection will be separated into phases, with each class session divided into four 15-minute intervals for a total of 120 minutes. During each phase, the system detects and records student activities such as sitting, standing, walking, and raising

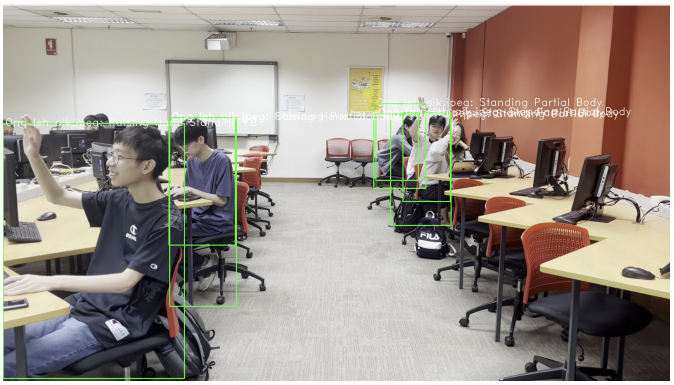


Fig. 3. Detection of Student Attendance

hands. By saving the phase-wise motion detection results, we ensure a precise and time-resolved analysis of student actions throughout the class, allowing for a more comprehensive knowledge of engagement patterns and their potential impact on attendance. Figure 4 is the example of the result for some data collected.

#	Student Name	Action	Date & Time
1	Marcus lim.jpeg	Standing Partial Body	2024-10-04 21:52:07
2	Marcus lim.jpeg	Standing Partial Body	2024-10-04 21:52:08
3	Marcus lim.jpeg	Standing Partial Body	2024-10-04 21:52:08
4	Chung yuan yi.jpeg	Raising Hand Standing FB	2024-10-04 21:52:08
5	Chung yuan yi.jpeg	Raising Hand Standing FB	2024-10-04 21:52:08
6	Chung yuan yi.jpeg	Raising Hand Standing FB	2024-10-04 21:52:08
7	Chung yuan yi.jpeg	Raising Hand Standing FB	2024-10-04 21:52:08
8	Chung yuan yi.jpeg	Raising Hand Standing FB	2024-10-04 21:52:08
9	Chung yuan yi.jpeg	Raising Hand Standing FB	2024-10-04 21:52:08
10	Marcus lim.jpeg	Raising Hand Standing FB	2024-10-04 21:52:09
11	Marcus lim.jpeg	Raising Hand Standing FB	2024-10-04 21:52:09
12	Marcus lim.jpeg	Raising Hand Standing FB	2024-10-04 21:52:09
13	Marcus lim.jpeg	Standing Partial Body	2024-10-04 21:52:09
14	Ong leh aik.jpeg	Standing Partial Body	2024-10-04 21:52:09

Fig. 4. Data Collection

C. Accuracy of the Detection Module

In addition to integrating face recognition and motion detection, we conducted extensive accuracy testing of the data. We utilized approximately 30 sets of images, each containing 30 frames, and trained for about 100 epochs. The results demonstrated significant improvements over time. The model's accuracy for both training and test data showed a consistent upward trend. As the training progressed and the number of epoch number increased, the graph for categorical accuracy approaching to 1.00, indicating a clear enhancement in data accuracy with extended training. Concurrently, the model loss decreased, reaching a value of 0.00 which further confirming the model's improved performance.

VI. Conclusion

In this project, we have integrated several advanced IoT devices, including the Intel® RealSense™ Depth Camera D455, to develop a smart attendance system. The selection

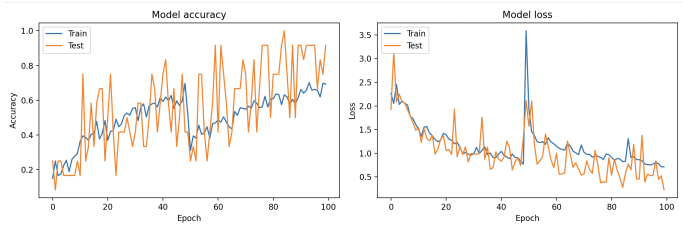


Fig. 5. Data Accuracy

of the D455 camera was based on a thorough comparison with other leading depth cameras. The D455 was chosen for its superior depth range, wide field of view, high frame rate, and robust SDK support, making it well-suited for capturing detailed and smooth video in larger classroom settings. The input module, utilizing the D455 camera, captures video data of students in real-time, which serves as raw data for the subsequent detection and analysis modules. This integration underscores the importance of IoT devices in enhancing the functionality and efficiency of modern attendance systems. While the analysis module is still under development, the potential benefits of the system are clear. The output stages will display the analyzed data to users through comprehensive data visualization. These visualizations will be designed to provide clear and actionable insights into various factors influencing absenteeism. Finally, the system aims to equip educators and administrators with the information they need to address and mitigate absenteeism effectively.

References

[1] Z. Mamatnabiyev, "Development of Attendance Monitoring System using IoT Technologies" in *16th International Conference on Electronics Computer and Computation (ICECCO)*, 2021.

[2] S. Rathinamala, A. Karthikeyan, R. Abishek and C. .K. Kannan, "Automatic Attendance Monitoring System Using LBPH and HAAR Algorithm" in *9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2023, pp.1604-1608.

[3] G. Ghaitanya, B. Ritesh, B. Tanish, M. Hanmant and S. Chudaman, "Efficient Attendance Management: A Smart Bluetooth-Based System for Seamless Monitoring and Automation" in *MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)*—, Apr 25-27, pp. 1-5.

[4] L. J. Khang and T. T. Hean, "QR Code-Based Student Attendance System" in *2nd Asia Conference on Computers and Communications (ACCC)*—, 2021, pp. 10-14.

[5] Y. Yi, "Student Attendance Management System Based on Fingerprint Identification Technology " in *International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, 2023.

[6] S. S. Rajawat and k. Saxena, "Face Recognition based Attendance System " in *1st IEEE International Conference on Industrial Electronics: Developments Applications (ICIDeA)*, 2022.

[7] Y. Liu, et al. "Gesture Recognition Model Based on YOLOv8 - AW" in *International Conference and Communications*, June 24,2023.

[8] L. H. Jian, G. H. Ming and L. H. Chern, "Automated Classroom Monitoring With Connected Visioning System " in *Proceedings of APSIPA Annual Summit and Conference 2017*, December 12-15,2017, pp. 386-393.

[9] L. Yiqing, Z. Linxio, L. Zijun, W Lin and Z. Lin, "Gesture Recognition Model Based on YOLOv8-AW" in *9th International Conference on Computer and Communications*—, 2023, pp. 2019-2023.

[10] Ziqi Ye,"Student Pose Detection Algorithm for the Classroom Environment Based on Improved YOLOv7 " in *6th International Conference on Information Systems and Computer Aided Education*—, 2023.

Robust Adaptive Filtering Based on Adaptive Projected Subgradient Method: Moreau Enhancement of Distance Function

Daiki Sawada and Masahiro Yukawa
Department of Electronics and Electrical Engineering, Keio University, Japan
E-mail: sawada@ykw.elec.keio.ac.jp, yukawa@elec.keio.ac.jp, Phone: +81 (0)45 566 1521

***Abstract*—This paper presents an outlier-robust adaptive filtering algorithm based on the adaptive projected subgradient method (APSM). The proposed algorithm is derived from a cost function involving the Moreau enhancement of the distance function to a hyperslab which permits a prespecified amount of estimation residuals. A data-dependent penalty is also exploited to preserve the overall convexity of the instantaneous cost with a minimal amount of estimation bias. Our instantaneous cost function involves a constant term to make the minimum of the cost be exactly zero so that the monotone approximation property of APSM holds with a fixed step size. Numerical examples show remarkable robustness of the proposed algorithm for both white and colored input signals.**

I. Introduction

Robustness against outlier has been a crucial issue in signal processing and machine learning [1] as well as in statistics. In the signal processing community, adaptive filtering has a long history [3], [4], and the classical algorithms such as the least mean square (LMS) algorithm [5] and the normalized LMS (NLMS) algorithm [6], [7] are known to be sensitive to outliers (impulsive noise). A number of robust algorithms have therefore been proposed, such as the sign algorithm [8], the least logarithmic absolute difference algorithm [9], least mean mixed-norm algorithm [10], robust mixed-norm algorithm [11], the hyperbolic cosine adaptive filter (HCAF) [12], and the logarithmic HCAF (LHCAF) [12], to name just a few.

Despite mathematical tractability, those approaches based on convex loss functions have limited robustness, and non-convex loss functions have also been studied, such as the sigmoid LMS algorithm [13], the generalized maximum correntropy criterion (GMCC) algorithm [14], and the generalized modified Blake-Zisserman algorithm [15]. Those nonconvex approaches, however, have an issue of local minima. Recently, the minimax concave (MC) loss function has been studied for robust signal recovery [16]–[18]. While the MC function was proposed originally for reducing the estimation bias in sparse modeling [19], [20], its use in robust signal recovery brings remarkable robustness owing to the so-called *redescending* property as well as overall convexity of the whole cost function. It is therefore of great interest to extend the same idea to robust adaptive filtering.

The adaptive projected subgradient method (APSM) [21] is a unified mathematical framework covering a variety of adaptive filtering algorithms including LMS, NLMS, the affine projection algorithms [22]–[26], the adaptive parallel subgradient projection algorithm [27], [28], their constrained versions [29]–[31], among many others. See also [32] for a comprehensive account of APSM. APSM exploits the subgradient projection to suppress instantaneous cost functions at every time instant, and it generates a sequence of adaptive filters approaching the set of minimizers of the instantaneous cost function monotonically at each iteration. Despite its nice mathematical properties as well as a number of empirical evidences [33]–[40] which support its practical advantages, APSM has not been utilized so far in the context of robust adaptive filtering.

In this paper, we present a robust adaptive filtering algorithm exploiting the MC loss under the APSM framework. Our MC-based loss function is derived as the Moreau enhancement of the distance function with respect to a bounded-instantaneous-error hyperslab which permits instantaneous errors below a prespecified threshold. As a result, the proposed loss function returns zero if the instantaneous error is below the bound, and it returns a constant value if the error exceeds another bound (threshold) to detect outliers. While the loss function is weakly convex, an introduction of the standard Tikhonov regularization may cause serious estimation biases. To preserve overall convexity of the whole cost function without causing such large biases, we focus on the fact that the instantaneous loss function is constant over the subspace orthogonal to the input vector. This property indicates that the loss can be convexified by adding the square of the filtered input multiplied by an appropriate scaling factor. This additional term can be regarded as a minimal possible data-dependent penalty to convexify the weakly convex loss function. Finally, we subtract a constant from the cost to make the minimum value of the whole cost function be zero so that the monotone approximation holds with a fixed step size. Numerical examples show that the proposed method outperforms the existing methods for both white and colored input signals.

This work was partially supported by JSPS Grants-in-Aid (22H01492).

II. Preliminaries

This section gives the notation and some mathematical tools used in this paper.

A. Notation and definition

The sets of nonnegative integers and real numbers are denoted by $\mathbb{N}$ and $\mathbb{R}$, respectively. Transposition of vector/matrix is denoted by $(\cdot)^\top$. Given a vector $\boldsymbol{x} \coloneqq [x_1, x_2, \cdots, x_n]^\top \in \mathbb{R}^n$, we define the ℓ_p norm $\|\boldsymbol{x}\|_p \coloneqq (\sum_{k=1}^n |x_k|^p)^{1/p}$ for $p \in [1, +\infty)$. The inner product between vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is defined as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \coloneqq \boldsymbol{x}^\top \boldsymbol{y} = \sum_{k=1}^n x_k y_k$. The Fenchel conjugate of a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is defined by $f^* : \mathbb{R}^n \to (-\infty, +\infty] : \boldsymbol{x} \mapsto \sup_{\boldsymbol{y} \in \mathbb{R}^n} (\langle \boldsymbol{x}, \boldsymbol{y} \rangle - f(\boldsymbol{y}))$.

B. Problem statement

Let $(\boldsymbol{u}_k)_{k \in \mathbb{N}} \subset \mathbb{R}^n$ be the sequence of input vectors $\boldsymbol{u}_k \coloneqq [u_k, u_{k-1}, \cdots, u_{k-n+1}]^\top$, where $k \in \mathbb{N}$ is the time index and $n \in \mathbb{N}$ is the tap length. Let $(n_k)_{k \in \mathbb{N}}$ and $(o_k)_{k \in \mathbb{N}}$ be the sequences of Gaussian noise and outlier, respectively. We consider the following linear system model:

$$d_k \coloneqq \boldsymbol{u}_k^\top \boldsymbol{h}^* + n_k + o_k, \qquad k \in \mathbb{N}, \tag{1}$$

where $\boldsymbol{h}^* \in \mathbb{R}^n$ is the estimand (the system to be estimated). The estimation residual (the error function) is defined by

$$e_k : \mathbb{R}^n \to \mathbb{R} : \boldsymbol{h} \mapsto \boldsymbol{u}_k^\top \boldsymbol{h} - d_k, k \in \mathbb{N}. \tag{2}$$

The task of adaptive filter is to estimate $\boldsymbol{h}^*$ by $\boldsymbol{h}_k \coloneqq [h_1^{(k)}, h_2^{(k)}, \cdots, h_n^{(k)}]^\top \in \mathbb{R}^n$ in a recursive fashion. Hereafter, we assume that $\boldsymbol{u}_k \neq \boldsymbol{0}$, because $\boldsymbol{u}_k = \boldsymbol{0}$ means that d_k only contains noise (and outliers) and such a data may not be used for adaptation.

III. Proposed Algorithm

To derive the proposed outlier-robust adaptive filtering algorithm, we first show that the Moreau enhancement of the distance function coincides with its composition with the MC penalty. Based on this, we present the proposed formulation to which APSM is applied to derive the proposed algorithm.

A. Moreau enhancement of distance function

We define the bounded instantaneous error hyperslab

$$S_k(\rho) \quad \coloneqq \quad \{\boldsymbol{h} \in \mathbb{R}^n \mid e_k^2(\boldsymbol{h}) \leq \rho\}, \tag{3}$$

where $\rho > 0$ is the error bound. The distance function from a point $\boldsymbol{h} \in \mathbb{R}^n$ to $S_k(\rho)$ is given by

$$d_{S_k(\rho)}(\boldsymbol{h}) \quad \coloneqq \quad \begin{cases} \dfrac{|e_k(\boldsymbol{h})| - \sqrt{\rho}}{\|\boldsymbol{u}_k\|_2}, & \text{if } |e_k(\boldsymbol{h})| > \sqrt{\rho}, \\ 0, & \text{if } |e_k(\boldsymbol{h})| \leq \sqrt{\rho}. \end{cases} \tag{4}$$

Given a convex function $f : \mathbb{R}^m \to \mathbb{R}$ defined on the Euclidean space $\mathbb{R}^m$ of arbitrary dimension m, its Moreau envelope of index $\gamma > 0$ is defined by

$${}^{\gamma} f(\boldsymbol{x}) \coloneqq \min_{\boldsymbol{y} \in \mathbb{R}^m} \left(f(\boldsymbol{y}) + \frac{1}{2\gamma} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right). \tag{5}$$

For instance, the Moreau envelope ${}^{\gamma}|\cdot|$ of the absolute-value function is the Huber loss function used widely in robust estimation. Subtracting the Moreau envelope ${}^{\gamma}|\cdot|$ from $|\cdot|$ yields the MC penalty [19], [20]:

$$\varphi_\gamma^{\mathrm{MC}}(x) = |x| - {}^{\gamma}|\cdot|(x), \quad x \in \mathbb{R}, \tag{6}$$

which is γ^{-1}-weakly convex; i.e., $\varphi_\gamma^{\mathrm{MC}} + (\gamma^{-1}/2)\|\cdot\|^2$ is convex. The function $\varphi_\gamma^{\mathrm{MC}}(x)$ is referred to as the *Moreau enhancement* of $|\cdot|$ [20], [42]. In fact, the Moreau enhancement of the distance function $d_{S_k(\rho)}$ can be written by using the MC penalty $\varphi_\gamma^{\mathrm{MC}}$.

Proposition 1. Let C be a nonempty closed convex subset of $\mathbb{R}^n$. Then, the Moreau enhancement of the distance function d_C of index $\gamma > 0$ is given by

$$(d_C)_\gamma(\boldsymbol{h}) \coloneqq d_C(\boldsymbol{h}) - {}^{\gamma} d_C(\boldsymbol{h}) = \varphi_\gamma^{\mathrm{MC}}(d_C(\boldsymbol{h})). \tag{7}$$

B. Proposed formulation

Based on Proposition 1, the Moreau enhancement of $d_{S_k(\rho)}$ is given by

$$(d_{S_k(\rho)})_\gamma(\boldsymbol{h}) = \varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h})). \tag{8}$$

Since $\varphi_\gamma^{\mathrm{MC}} \circ d_{S_k(\rho)}$ is γ^{-1}-weakly convex, an addition of $(\gamma^{-1}/2)\|\cdot\|^2$ as a regularizer makes the whole cost convex. This straightforward approach, however, causes a non-negligible bias, resulting in degraded performance. We therefore introduce the following alternative:

$$f_k(\boldsymbol{h}) \coloneqq \varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h})) + \frac{1}{2\gamma}(\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h})^2, \ \boldsymbol{h} \in \mathbb{R}^n, \tag{9}$$

where $\tilde{\boldsymbol{u}}_k \coloneqq \boldsymbol{u}_k / \|\boldsymbol{u}_k\|_2$. The loss function $\varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h}))$ is highly robust against large outliers owing to the saturation property of the MC function $\varphi_\gamma^{\mathrm{MC}}$. However, since $\varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h}))$ is concave over the subspace $\mathrm{span}\{\boldsymbol{u}_k\}$, it is difficult to guarantee the monotone approximation property of the global minimizer at each time instant, which is important to ensure stable performance of adaptive algorithms [21], [33]–[39]. Therefore, the second term $\frac{1}{2\gamma}(\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h})^2$ is introduced as a "minimal" possible regularizer, in some sense, to convexify $\varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h}))$. The point here is that $\frac{1}{2\gamma}(\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h})^2$ gives no impact on those components over the orthogonal compliment $\mathrm{span}^\perp\{\boldsymbol{u}_k\}$, thus yielding lower bias compared to the Tikhonov regularization $\|\boldsymbol{h}\|_2^2$. We mention here that $\boldsymbol{h}$ is implicitly assumed to be a random vector obeying a generalized Gaussian distribution with the precision matrix $\tilde{\boldsymbol{R}} \coloneqq E(\tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{u}}_k^\top)$, where $E(\cdot)$ stands for expectation. Our simulation will show that our method works well even though such an implicit assumption is violated (see Section IV).

We now focus on the minimum value

$$f_k^{\min} \coloneqq \min_{\boldsymbol{h} \in \mathbb{R}^n} f_k(\boldsymbol{h}) \tag{10}$$

$$= \begin{cases} \min\left\{ \dfrac{\gamma}{2}, \dfrac{(|d_k| - \sqrt{\rho})^2}{2\gamma \|\boldsymbol{u}_k\|_2^2} \right\}, & \text{if } |d_k| > \sqrt{\rho}, \\ 0, & \text{if } |d_k| \leq \sqrt{\rho}, \end{cases} \tag{11}$$

of the function f_k. Subtracting $f_k^{\min}$ from f_k, our instantaneous cost function $\Theta_k : \mathbb{R}^n \to [0, +\infty)$ is defined by

$$\Theta_k(\boldsymbol{h}) := \underbrace{\varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h}))}_{\substack{\text{robustifying the data} \\ \text{fidelity term}}} + \underbrace{\frac{1}{2\gamma}(\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h})^2}_{\substack{\text{convexifying} \\ \text{the whole cost}}} \underbrace{- f_k^{\min}}_{\substack{\text{making the} \\ \text{minimum be zero}}} \tag{12}$$

so that

$$\Theta_k^{\min} := \min_{\boldsymbol{h} \in \mathbb{R}^n} \Theta_k(\boldsymbol{h}) = 0. \tag{13}$$

The property in (13) allows the use of a constant step size in APSM (see Proposition 2 below). We remark that the first and second terms of (12) brings remarkable outlier-robustness while preserving the convexity of the whole cost.

C. Adaptive Projected Subgradient Method for proposed formulation

The nonnegative continuous convex function $\Theta_k(\boldsymbol{h})$ is minimized by APSM [21]:

$$\boldsymbol{h}_{k+1} = \boldsymbol{h}_k + \lambda_k \Big(T_{\mathrm{sp}(\Theta_k)}(\boldsymbol{h}_k) - \boldsymbol{h}_k \Big), \tag{14}$$

where $\lambda_k \in (0, 2)$ is the step size, and

$$T_{\mathrm{sp}(\Theta_k)}(\boldsymbol{h}) := \begin{cases} \boldsymbol{h} - \dfrac{\Theta_k(\boldsymbol{h})}{\|\Theta_k'(\boldsymbol{h})\|_2^2} \Theta_k'(\boldsymbol{h}), & \text{if } \Theta_k(\boldsymbol{h}) > 0, \\ \boldsymbol{h}, & \text{otherwise}, \end{cases} \tag{15}$$

with a subgradient

$$\Theta_k'(\boldsymbol{h}) := \begin{cases} \mathrm{sign}(e_k(\boldsymbol{h})) \Big(1 - \dfrac{1}{\gamma} \mathrm{clip}_\gamma(d_{S_k(\rho)}(\boldsymbol{h})) \Big) \tilde{\boldsymbol{u}}_k \\ \qquad + \dfrac{\tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{u}}_k^\top}{\gamma} \boldsymbol{h}, \quad \text{if } |e_k(\boldsymbol{h})| > \sqrt{\rho}, \\ \dfrac{\tilde{\boldsymbol{u}}_k \tilde{\boldsymbol{u}}_k^\top}{\gamma} \boldsymbol{h}, \quad \text{if } |e_k(\boldsymbol{h})| \le \sqrt{\rho}, \end{cases} \tag{16}$$

of Θ_k at $\boldsymbol{h}$. Here,

$$\mathrm{sign} : \mathbb{R} \to \{-1, 1\} : x \mapsto \begin{cases} 1, & \text{if } x \ge 0, \\ -1, & \text{if } x < 0, \end{cases} \tag{17}$$

$$\mathrm{clip}_\gamma : \mathbb{R}_+ \to [0, \gamma] : x \mapsto \begin{cases} \gamma, & \text{if } x \in (\gamma, +\infty), \\ x, & \text{if } x \in [0, \gamma]. \end{cases} \tag{18}$$

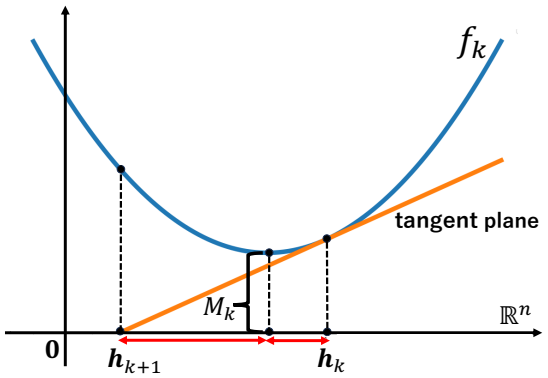
Proposition 2. (Monotone approximation [27, Theorem 2]) Assume that

$$\boldsymbol{h}_k \notin \Omega_k := \big\{ \boldsymbol{h} \in \mathbb{R}^n \mid \Theta_k(\boldsymbol{h}) = \Theta_k^{\min} = 0 \big\} \neq \emptyset. \tag{19}$$

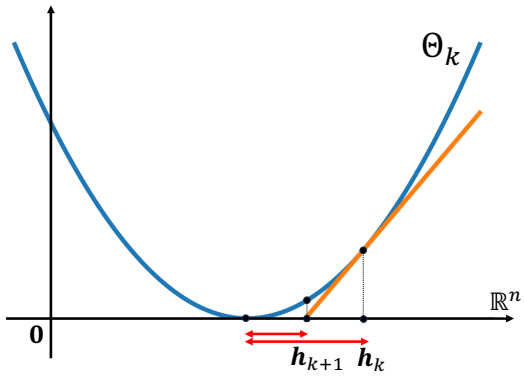
Then, for an arbitrary step size $\lambda_k \in (0, 2)$, it holds that

$$\|\boldsymbol{h}_{k+1} - \boldsymbol{h}_k^{(*)}\|_2 < \|\boldsymbol{h}_k - \boldsymbol{h}_k^{(*)}\|_2, \quad \forall \boldsymbol{h}_k^{(*)} \in \Omega_k. \tag{20}$$

Proposition 2 states that the sequence $(\boldsymbol{h}_k)_{k \in \mathbb{N}}$ monotonically approaches the solution set Ω_k at every time instant that is expected to contain the estimand. The role of $f_k^{\min}$ for the monotone approximation is illustrated in Fig. 1. Roughly speaking, the subgradient projection is the orthogonal projection onto the intersection of the tangent plane and $\mathbb{R}^n$. Viewing Fig. 1(a), one can see that $T_{\mathrm{sp}(f_k)}(\boldsymbol{h})$ is far from the minimizer of $f_k(\boldsymbol{h})$ than $\boldsymbol{h}$, meaning that the operation of $T_{\mathrm{sp}(f_k)}$ may put the current estimate away from the set Ω_k. This happens



(a) f_k



(b) Θ_k

Fig. 1: Illustrations of the subgradient projections $T_{\mathrm{sp}(f_k)}$ and $T_{\mathrm{sp}(\Theta_k)}$.

because the graph of f_k is distant (by $f_k^{\min}$) from the 'ground' $\mathbb{R}^n$, and, in this case, the step size must be smaller than a bound depending on $f_k^{\min}$ to ensure the monotone approximation property [21]. In contrast, the operator $T_{\mathrm{sp}(\Theta_k)}$ in Fig. 1(b) pushes $\boldsymbol{h}$ towards Ω_k so that the monotone approximation holds.

Proposition 3. For every $\gamma > 0$, $\Theta_k(\boldsymbol{h})$ is convex.

Proof: For the sake of accessibility, we present a primitive proof. It suffices to show convexity of $f_k(\boldsymbol{h})$. By virtue of the Moreau decomposition [41] and by the definition of $\varphi_\gamma^{\mathrm{MC}}$ in (6), $f_k(\boldsymbol{h})$ can be rewritten as

$$f_k(\boldsymbol{h}) = \underbrace{d_{S_k(\rho)}(\boldsymbol{h}) + {}^{\gamma^{-1}}(|\cdot|^*)(\gamma^{-1} d_{S_k(\rho)}(\boldsymbol{h}))}_{\text{convex}} \underbrace{- \frac{1}{2\gamma} d_{S_k(\rho)}^2(\boldsymbol{h}) + \frac{1}{2\gamma}(\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h})^2}_{g_k(\boldsymbol{h})}. \tag{21}$$

Here, the conjugate function $|\cdot|^*$ of $|\cdot|$ and the distance function $d_{S_k(\rho)}(\boldsymbol{h})$ are convex [21]. It can also be verified that the composition ${}^{\gamma^{-1}}(|\cdot|^*)(\gamma^{-1} d_{S_k(\rho)}(\boldsymbol{h}))$ is also convex. It is therefore sufficient to show convexity of $g_k(\boldsymbol{h})$. Recall that

TABLE I: Computational complexity per iteration

proposed method	GMCC	LHCAF	NLMS
$3n+25$	$2n+4$	$2n+6$	$3n+2$

it is assumed that $\boldsymbol{u}_k \neq \boldsymbol{0}$. The gradient $\nabla g_k(\boldsymbol{h})$ of $g_k(\boldsymbol{h})$ is given by

$$\nabla g_k(\boldsymbol{h}) = \begin{cases} \frac{1}{\gamma}\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h} - \mathrm{sign}(e_k(\boldsymbol{h}))\frac{|e_k(\boldsymbol{h})| - \sqrt{\rho}}{\gamma\|\boldsymbol{u}_k\|_2}\tilde{\boldsymbol{u}}_k, \\ \qquad\qquad \text{if } |e_k(\boldsymbol{h})| > \sqrt{\rho}, \\ \frac{1}{\gamma}\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{u}}_k^\top \boldsymbol{h}, \quad \text{if } |e_k(\boldsymbol{h})| \leq \sqrt{\rho}, \end{cases} \tag{22}$$

and the Hessian $\nabla^2 g_k(\boldsymbol{h})$ of $g_k(\boldsymbol{h})$ is given by

$$\nabla^2 g_k(\boldsymbol{h}) = \begin{cases} \boldsymbol{O}, & \text{if } |e_k(\boldsymbol{h})| > \sqrt{\rho}, \\ \frac{1}{\gamma}\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{u}}_k^\top, & \text{if } |e_k(\boldsymbol{h})| \leq \sqrt{\rho}, \end{cases} \tag{23}$$

where $\boldsymbol{O} \in \mathbb{R}^{n\times n}$ is the zero matrix. From (23), $\nabla^2 g_k(\boldsymbol{h})$ is positive semi-definite as $\boldsymbol{h}^\top(\boldsymbol{u}_k\boldsymbol{u}_k^\top)\boldsymbol{h} = (\boldsymbol{u}_k^\top\boldsymbol{h})^2 \geq 0$ for all $\boldsymbol{h} \in \mathbb{R}^n$, and hence $g_k(\boldsymbol{h})$ is convex. ■

D. Complexity

The computational complexities (the number of multiplications) of the proposed and existing algorithms are given in Table I, where it can be seen that all methods have linear-order complexities. It will be shown in Section IV that the proposed method significantly improves outlier robustness with the moderate computational complexity.

E. Parameter design

Possible choices of ρ are given as follows [27]: $\rho_1 := (1+\sqrt{2})\sigma^2$, $\rho_2 := \sigma^2$, and $\rho_3 := 0$, where n_k is assumed to obey the zero-mean i.i.d. Gaussian $\mathcal{N}(0,\sigma^2)$. An increase of the error bound ρ enhances the "thickness" of the hyperslab $S_k(\rho)$, while the step size λ_k governs the degree of relaxation for the projection onto $S_k(\rho)$. In fact, an increase of ρ may give an impact, similar to a decrease of λ_k, on the performance. We mention here another aspect that an increase of ρ also decreases the frequency of update, thereby reducing the computational loads. Our recommended choice is setting ρ to ρ_1, ρ_2, or ρ_3, and tuning λ_k possibly depending on the noise variance.

The saturation factor γ could be selected based on the the variance of Gaussian noise and the outlier statistics. As depicted in Fig. 2, the Gaussian noise concentrates around the mean value, while outliers reside in those regions apart from the Gaussian noise. The saturation factor γ is desired to be tuned in such a way that the Gaussian noise falls, with high probability, into the central region between the two dotted lines in the figure, and, at the same time, the outliers arise in the saturated regions where the gradient vanishes.

IV. Numerical Examples

We compare the performance of the proposed method to those of the robust adaptive filtering algorithms: LHCAF [12] and GMCC [14]. For reference, we also test the NLMS algorithm. We consider two types of input signal $(u_k)_{k\in\mathbb{N}}$, white and colored. The white input signal is generated from the standard Gaussian distribution. As a colored input, the USASI signal (speech-like wide-sense stationary process) is used, modeled on the autoregressive moving average process characterized by the transfer function [27]

$$H_{\mathrm{USASI}}(z) := \frac{1-z^{-2}}{1-1.70233z^{-1}+0.71902z^{-2}}, \quad z \in \mathbb{C}, \tag{24}$$

where $\mathbb{C}$ denotes the set of complex numbers.

The estimand $\boldsymbol{h}^*$ of length $n = 256$ is generated from the i.i.d. standard Gaussian distribution. The noise n_k is white Gaussian with zero mean and signal-to-noise ratio (SNR) 10 dB and 20 dB, where SNR $:= 10\log_{10}(E[z_k^2]/E[n_k^2])$ with $z_k := \boldsymbol{u}_k^\top\boldsymbol{h}^*$. An impulsive noise is generated by $o_k := b_k v_k$, where b_k obeys the Bernoulli process with success and fail probabilities $P(b_k = 1) = 0.05$ and $P(b_k = 0) = 0.95$, respectively, and v_k is zero-mean Gaussian with variance $\sigma_v^2 = 1000E[z_k^2]$. We adopt the system mismatch defined as

$$\xi(\boldsymbol{h}_k) := 20\log_{10}\frac{\|\boldsymbol{h}^* - \boldsymbol{h}_k\|_2}{\|\boldsymbol{h}^*\|_2}. \tag{25}$$

All result are averaged over 300 trials. The adaptive filters are initialized as $\boldsymbol{h}_0 := \boldsymbol{0}$. The parameters for each input are given in Tables I and II, respectively.

Figs. 3 and 4 depict the results for the white and colored input signals, respectively. It can be seen that the proposed method significantly outperforms the existing robust methods in both cases.

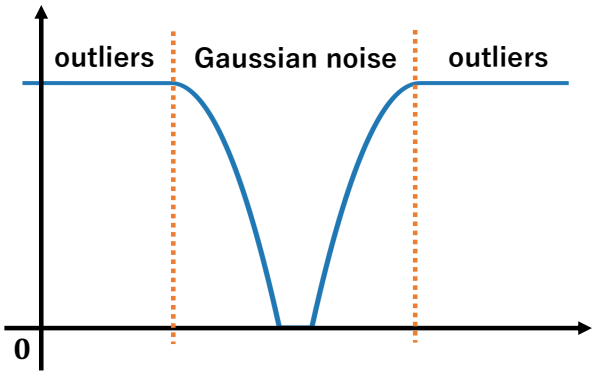


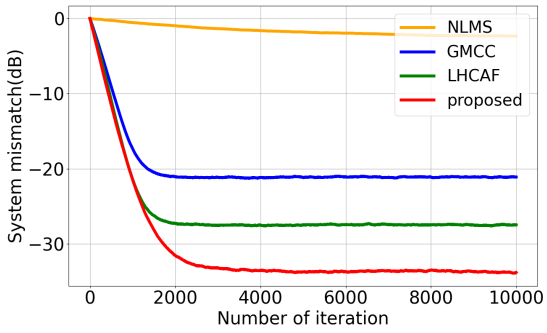
Fig. 2: Graph of $\varphi_\gamma^{\mathrm{MC}}(d_{S_k(\rho)}(\boldsymbol{h}))$ and the saturation factor γ.

TABLE II: Parameter setting for the case of white signals

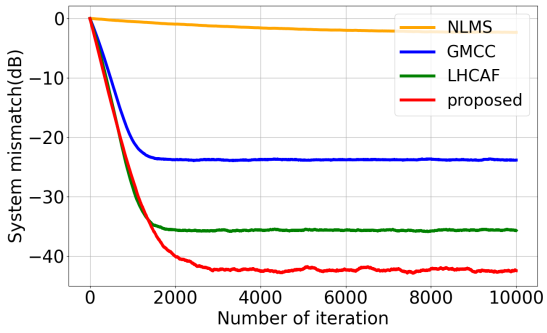
Methods	noise 10 dB	noise 20 dB
proposed	λ_k=0.4, γ=3.5, ρ=0	λ_k=0.5, γ=3.5, ρ=0
GMCC	η=0.025, α=1, λ=0.001	η=0.029, α=1, λ=0.005
LHCAF	λ=0.12, μ= 0.03	λ=0.14, μ=0.035
NLMS	μ=0.03	μ=0.03

V. Conclusion

This paper presented the robust adaptive filtering algorithm based on APSM. The instantaneous cost function consists of three terms: (i) the weakly convex loss given by the Moreau enhancement of the distance function, (ii) the data-dependent

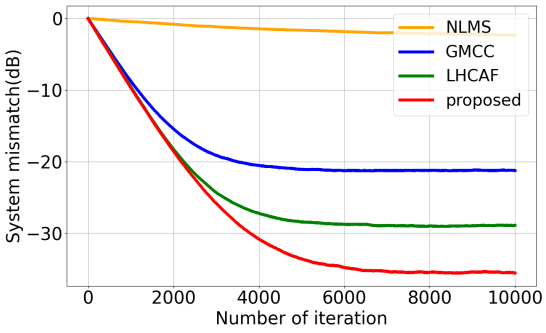


(a) noise 10 dB

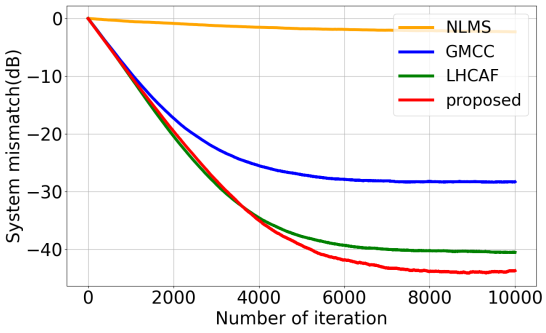


(b) noise 20 dB

Fig. 3: Learning curves for white signals.



(a) noise 10 dB



(b) noise 20 dB

Fig. 4: Learning curves for USASI signals.

TABLE III: Parameter setting for the case of colored signals

Methods	noise 10 dB	noise 20 dB
proposed	λ_k=0.3, γ=5, ρ=0	λ_k=0.3, γ=5, ρ=0
GMCC	η=0.01, α=1, λ=0.005	η=0.01 α=1, λ=0.005
LHCAF	λ=0.12, μ=0.007	λ=0.12, μ=0.007
NLMS	μ=0.03	μ=0.03

penalty function, and (iii) the constant term. The loss function brings significant robustness against impulsive noises, while the penalty function makes the whole cost function be convex without causing serious estimation biases. The constant term makes the minimum of the whole cost be exactly zero, allowing the use of a fixed step size. The numerical examples showed that the proposed method outperformed the existing methods (GMCC, LHCAF) in the presence of impulsive noises for both white and colored input signals.

REFERENCES

[1] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, Robust Statistics for Signal Processing. Cambridge: Cambridge Uni- versity Press, 2018.

[2] T. van Erven, S. Sachs, W. M. Koolen, and W. Kotlowski, "Robust Online Convex Optimization in the Presence of Outliers," Proc. Machine Learning Research, vol. 134, pp. 1–21, 2021.

[3] S. Haykin, "Adaptive Filter Theory", 4th ed., New Jersey: Prentice Hall, 2002.

[4] A. H. Sayed, "Fundamentals of Adaptive Filtering", Wiley-Interscience, New York, 2003.

[5] B. Widrow and M. E. Hoff Jr., "Adaptive switching circuits," in IRE WESCON Conv. Rec., vol. 4, pp. 96–104, 1960.

[6] J. Nagumo and A. Noda, "A learning method for system identification", IEEE Trans. Autom. Control, vol. 12, no. 3, pp. 282–287, 1967.

[7] A. E. Albert and L. S. Gardner, Jr., "Stochastic Approximation and Nonlinear Regression", Cambridge, MIT Press, 1967.

[8] N. Verhoeckx, H. Van Den Elzen, F. Snijders, and P. Van Gerwen, "Digital echo cancellation for baseband data transmission", IEEE Trans. Acoust, vol. 27, no. 6, pp. 768–781, 1979.

[9] M. O. Sayin, N. D. Vanli, and S. S. Kozat, "A novel family of adaptive filtering algorithms based on the logarithmic cost", IEEE Trans. Signal Process., vol. 62, no.17, 2014. 4411–4424.

[10] J. A. Chambers, O. Tanrikulu, and A.G. Constantinides, "Least mean mixed-norm adaptive filtering," Electron Lett, vol. 30, no. 19, pp. 1574–1575, 1994.

[11] J. Chambers, and A. Avlonitis, "A robust mixed-norm adaptive filter algorithm," IEEE Signal Process. Lett., vol. 4, no. 2, pp. 46–48, 1997.

[12] S. Wang, W. Wang, K. Xiong, H. H. C. Iu, and K.T. Chi, "Logarithmic hyperbolic cosine adaptive filter and its performance analysis", IEEE Trans. Systems, Man., and Cybernetics: Systems, vol. 51, no. 4, pp. 2512–2524, Apr. 2021.

[13] F. Huang, J. Zhang, and S. Zhang, "A family of robust adaptive filtering algorithms based on sigmoid cost", Signal Process., vol. 149, pp. 179–192, 2018.

[14] B. Chen, L. Xing, H. Zhao, N. Zheng, and J.C. Principe, "Generalized correntropy for robust adaptive filtering", IEEE Trans. Signal Process., vol. 64, no.13, pp. 3376–3387, 2016.

[15] K. Kumar, M. L. N. S. Karthik, and N. V. George, "Generalized Modified Blake–Zisserman Robust Sparse Adaptive Filters," IEEE Trans. Systems, Man, and Cybernetics: Systems, vol. 53, no. 1, pp. 647-652, Jan. 2023.

[16] M. Yukawa, H. Kaneko, K. Suzuki, and I. Yamada, "Linearly-involved Moreau-enhanced-over-subspace model: Debiased sparse modeling and stable outlier-robust regression," IEEE Trans. Signal Process., vol. 71, pp. 1232–1247, 2023.

[17] K. Suzuki and M. Yukawa, "Robust recovery of jointly-sparse signals

using minimax concave loss function," IEEE Trans. Signal Process., vol. 69, pp. 669–681, 2021.
[18] K. Suzuki and M. Yukawa, "Sparse stable outlier-robust signal recovery under Gaussian noise," IEEE Trans. Signal Process., vol.71, pp.372–387, 2023.
[19] C. H. Zhang et al., "Nearly unbiased variable selection under minimax concave penalty," The Annals of statistics, vol. 38, no. 2, pp. 894–942, 2010.
[20] I. Selesnick, "Sparse regularization via convex analysis," IEEE Trans. Signal Process., vol. 65, no. 17, pp. 4481–4494, 2017.
[21] I. Yamada and N. Ogura, "Adaptive projected subgradient method for asymptotic minimization of sequence of nonnegative convex functions," Numerical Functional Analysis and Optimization, vol. 25, pp. 593–617, 2004.
[22] T. Hinamoto and S. Maekawa, "Extended theory of learning identification," (in Japanese) Trans. Inst. Elect. Eng. Jpn., vol. 95, no. 10, pp. 227–234, 1975.
[23] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," Electron. Commun. Jpn., vol. 67-A, no. 5, pp. 19–27, 1984.
[24] D. Slock, "The block underdetermined covariance (BUC) fast transversal filter (FTF) algorithm for adaptive filtering," in Proc. Asilomar Conf. Signals, Syst., Comput., 1992, pp. 550–554.
[25] D. R. Morgan and S. G. Kratzer, "On a class of computationally efficient, rapidly converging, generalized NLMS algorithms," IEEE Signal Process. Lett., vol. 3, no. 8, pp. 245–247, Aug. 1996.
[26] M. Rupp, "A family of adaptive filter algorithms with decorrelating properties," IEEE Trans. Signal Process., vol. 46, no. 3, pp. 771–775, Mar. 1998.
[27] I. Yamada, K. Slavakis, and K. Yamada, "An efficient robust adaptive filtering algorithm based on parallel subgradient projection techniques," IEEE Trans. Signal Process., vol. 50, no. 5, pp. 1091–1101, 2002.
[28] M. Yukawa and I. Yamada, "Pairwise optimal weight realization — Acceleration technique for set-theoretic adaptive parallel subgradient projection algorithm," IEEE Trans. Signal Process., vol. 54, no. 12, pp. 4557–4571, Dec. 2006.
[29] J. A. Apolinário, Jr., S. Werner, P. S. R. Diniz, and T. I. Laakso, "Constrained normalized adaptive filters for CDMA mobile communications," in Proc. EUSIPCO, 1998, vol. IV, pp. 2053–2056.
[30] M. L. R. de Campos and J. A. Apolinário, Jr., "The constrained affine projection algorithm — Development and convergence issues," in Proc. Balkan Conf. Signal Process., Commun., Circuits Syst., 2000.
[31] M. Yukawa, R. L. G. Cavalcante, and I. Yamada, "Efficient blind MAI suppression in DS/CDMA systems by embedded constraint parallel projection techniques," IEICE Trans. Fundam., vol. E88-A, no. 8, pp. 2062–2071, Aug. 2005.
[32] S. Theodoridis, K. Slavakis, and I. Yamada, "Adaptive learning in a world of projections: A unifying framework for linear and nonlinear classification and regression tasks," IEEE Signal Process. Mag., vol. 28, no. 1, pp. 97–123, Jan. 2011.
[33] M. Yukawa and I. Yamada, "Efficient adaptive stereo echo canceling schemes based on simultaneous use of multiple state data," IEICE Trans. Fundam., vol. E87-A, no. 8, pp. 1949–1957, Aug. 2004.
[34] R. L. G. Cavalcante, M. Yukawa, and I. Yamada, "Set-theoretic DS/CDMA receivers for fading channels by adaptive projected subgradient method," in Proc. IEEE GLOBECOM, SP06-1, 2005.
[35] K. Slavakis, M. Yukawa, and I. Yamada, "Robust Capon beamforming by the adaptive projected subgradient method," in Proc. IEEE ICASSP, 2006, pp. 1005–1008.
[36] M. Yukawa, K. Slavakis, and I. Yamada, "Adaptive parallel quadratic-metric projection algorithms," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 5, pp. 1665–1680, Jul. 2007.
[37] R. L. G. Cavalcante and I. Yamada, "Multiaccess interference suppression in orthogonal space-time block coded MIMO systems by the adaptive projected subgradient method," IEEE Trans. Signal Process., vol. 56, no. 3, pp. 1028-1042, March 2008.
[38] R. L. G. Cavalcante, I. Yamada, and B. Mulgrew, "An adaptive projected subgradient approach to learning in diffusion networks," IEEE Trans. Signal Process., vol. 57, no. 7, pp. 2762-2774, July 2009.
[39] R. L. G. Cavalcante, A. Rogers, N. R. Jennings, and I. Yamada, "Distributed asymptotic minimization of sequences of convex functions by a broadcast adaptive subgradient method," IEEE J. Selected Topics in Signal Process., vol. 5, no. 4, pp. 739-753, Aug. 2011.
[40] N. Agrawal, R.L.G. Cavalcante, M. Yukawa, and S. Stańczak, "Distributed Convex Optimization " Over-the-Air " in Dynamic Environments", IEEE Trans. Signal and Information Processing over Networks, vol. 10, pp. 610-625, 2024.
[41] H. H. Bauschke and P. L. Combettes, "Convex Analysis and Monotone Operator Theory in Hilbert Spaces", 2nd Ed., Springer, 2017.
[42] J. Abe, M. Yamagishi, and I. Yamada, "Linearly involved generalized Moreau enhanced models and their proximal splitting algorithm under overall convexity condition," Inverse Problems, vol. 36, no. 3, pp. 1–36, Feb. 2020.

Non-Target Conversion Based Speech Steganography for Secure Speech Communication System

Mingjun Zhang*, Yan Feng*, YuGao†, Longting Xu*‡
* College of Information Science and Technology, Donghua University, China
E-mail: zmj@mail.dhu.edu.cn, xlt@dhu.edu.cn
† AI Research Center, Midea Group (Shanghai) Co.,Ltd., Shanghai 201702, China
E-mail: gaoyu11@midea.com
‡ Corresponding Author

***Abstract*—The widespread application of speech data increases the risk of speaker identity being compromised during speech communication. To mitigate this risk and protect voice privacy, we propose a system aimed at ensuring the security of speech communication by avoiding the exposure of the actual speaker's identity. Our system comprises three main components. Firstly, the non-target voice conversion system based on generative adversarial network converts the original audio into the audio of a non-existent person while preserving the speaker embedding of the real audio. Secondly, during speech communication, we utilize speech steganography techniques to embed the actual speaker embedding into the converted audio. Finally, at the receiving end, we extract the actual speaker embedding from the transmitted converted audio and use it to reconstruct the original audio. Experimental results validate the effectiveness of our system, showcasing an innovative solution in the field of speech security.**

***Index Terms*—non-target voice conversion, speech steganography, voice privacy, speech communication**

I. Introduction

With the integration of voice user interfaces into devices like AI speakers, speech data has become ubiquitous and widely used in areas such as smart homes and conversational AI, becoming an essential component of future lifestyles. However, speech signals contain a wealth of speaker-specific information, including sensitive attributes such as gender, identity, age, emotions, and sensations. If intercepted during transmission, these sensitive attributes can be extracted and used as biometric features or for malicious purposes like identity theft or voice spoofing [1]. The privacy protection of speech data has gained increasing attention, particularly after the introduction of the General Data Protection Regulation (GDPR) in the European Union [2], emphasizing the importance of restrictions on speech data usage and the implementation of security measures [3].

Voice conversion (VC) refers to the various modifications made to human speech, including the conversion of non-linguistic information such as voice timbre, prosody, or pitch, while keeping the linguistic information unchanged [4]–[6]. The primary purpose of voice conversion is to transform the identity of one speaker into another while retaining the linguistic content of the speech data. Speaker anonymization, on the other hand, is a speech data processing technique aimed at protecting the identity privacy of speakers [7], [8]. This technique involves steps such as content/speaker separation and voice conversion, to remove the speaker's identity information from the original speech while retaining other information such as linguistic content. By using the technique of speaker anonymization to convert the speaker's identity to a randomly generated non-existent identity, non-target voice conversion can be achieved. While the method performs well in removing speaker identity information, it is not sufficient to solely rely on it for protecting voice privacy in the process of speech communication. This is because when the receiving end perceives the received converted speech as mismatched with the known voice of the original speaker, doubts about the authenticity of the speech information arise.

Therefore, in the process of speech communication, the ability of the receiving end to reproduce the original speech is a crucial issue for protecting speaker privacy using non-target voice conversion techniques. To address this problem, speech steganography can play an important role. Speech steganography is a technique that embeds secret information into audio signals. It modifies the speech data to increase the capacity of hidden information, making it difficult for external observers to detect the presence of additional information. Speech steganography can be achieved through various conventional methods, including LSB substitution [9], echo hiding [10], and spread spectrum (SS) [11]. With the popularity of deep neural networks, many recent studies have utilized DNNs to embed and detect information [12]. By utilizing this technique, we embed the speaker embedding of the original speech data into the converted speech transmitted during the communication process, making it possible for the receiving end to reconstruct the original speech data.

In this paper, we propose a novel system, called the non-target Conversion based Speech Steganography System (CSSS), for protecting voice privacy in speech communication. At the sender's end, we employ a Generative Adversarial Network (GAN) [13], [14] based speaker anonymization system to create an artificial speaker embedding space and sample non-existent voices, thereby achieving conversion of the identity of the original speaker to a non-existent one. To reconstruct

This work was supported in part by Donghua University Language and Text Application Research Project (No. Y2024-8).

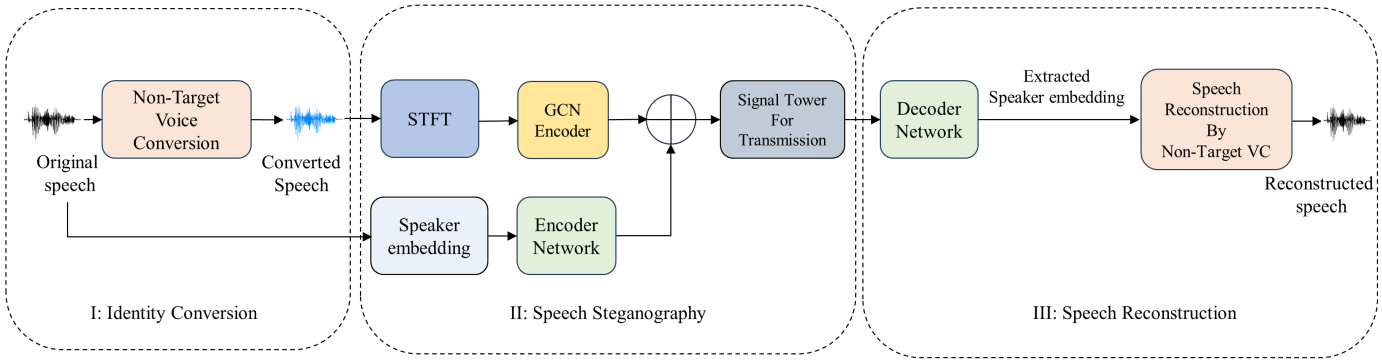


Fig. 1. Architecture of the proposed CSSS. The three sections from left to right correspond respectively to the transmitting end, the transmission process, and the receiving end of the speech signal.

the original speech at the receiver end without leaking identity information, we employ a Convolutional Neural Network (CNN)-based speech steganography technique and introduce a graph encoder [15] based on Graph Convolutional Networks (GCN) [16] to embed the speaker embedding of the original speech into converted speech for transmission. At the receiver's end, the system extracts the hidden speaker embedding and combines it with the prosody information and speech-to-text information of the original speech data to reconstruct the original voice data.

The experimental results indicate that the speech processed by identity conversion and information steganography in CSSS has a high level of privacy, which can effectively prevent privacy attackers from identifying the real speaker. Additionally, the reconstructed speech is close to the original speech in terms of privacy and utility. Furthermore, the reconstructed speech in CSSS demonstrates high intelligibility, and naturalness, and maintains a high perceptual similarity to the original speech.

II. Proposed System

In this section, we will provide a detailed description of the specific architectural design of CSSS. Fig. 1 illustrates the three main components of CSSS, namely the identity conversion, the speech steganography during speech communication, and the speech reconstruction.

A. Identity Conversion

Prosody is closely related to speaker identity, affecting an individual's pitch and speaking speed. Pitch (F_0) and speaking speed not only reveal information about the speaker but also provide essential features for most machine learning models that process speech. Therefore, in privacy-preserving related tasks, maintaining a constant F_0 could potentially lead to identity leakage [17]. Ref. [18] overcame the trade-off between privacy and utility of prosody by implementing prosody cloning, and achieved an anonymization system with high privacy and high pitch relevance. We adopt the speaker anonymization pipeline proposed by [18] to achieve non-target voice conversion, where the converted speech has a high pitch

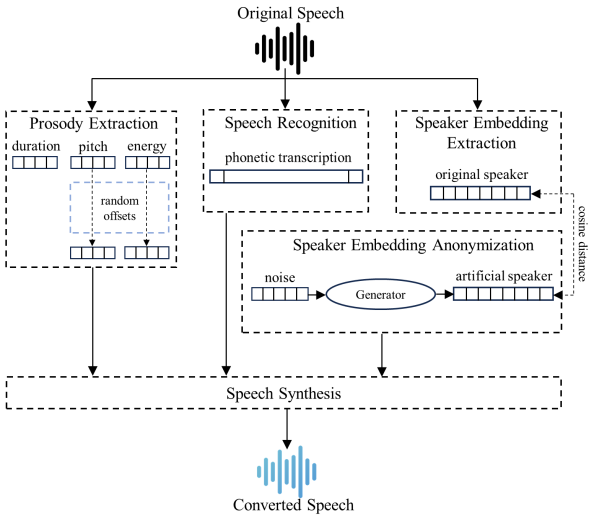


Fig. 2. Illustration of the pipeline for non-target voice conversion. By extracting prosody information, and speech content information, and utilizing cosine similarity to approximate the embedding of the original speaker, randomly generated non-existent speaker identity can be obtained.

similarity to the original speech while having a non-existent speaker identity, as shown in the structure in Fig. 2.

The pipeline extracts three types of information from the input speech: (i) prosodic information as pitch, energy, and duration sequences; (ii) the linguistic content in the form of phonetic sequences by using an Automatic Speech Recognition (ASR) model; and (iii) the speaker information as speaker embedding. Pitch and energy can be adjusted by multiplying them with random offsets to further enhance privacy. Additionally, in the Speaker Embedding Anonymization section depicted in Fig. 2, a GAN is employed to transform random noise into speaker embeddings that are unrelated to the original speaker, with the cosine distance ensuring sufficient similarity between the generated vector and the vector corresponding to the input speaker. Finally, the pipeline utilizes a Text-to-Speech (TTS) model to re-synthesize the audio based on language, prosody, and generated speaker information. The

resulting converted audio maintains the same language content and prosody features as the original audio.

B. Speech Steganography

1) CNN-based Encoder Network: The traditional speech steganography techniques are mainly divided into the time domain and transform domain methods. These methods, however, encounter a couple of challenges. To begin with, the process of manually selecting appropriate redundancy to conceal secret information requires a considerable investment of manpower, resources, and time. Additionally, the manual selection of redundancy for embedding secret information may result in noticeable alterations in the carrier, leaving behind discernible traces.

Steganographic methods based on the Encoder-Decoder structure of neural networks only require training the adopted model. Utilizing a trained DNN for steganography and information extraction in speech carriers can effectively overcome the shortcomings of traditional methods. In the second part of CSSS, we propose a CNN-based steganographic structure that can embed vector data into audio, utilizing it as an Encoder-Decoder network. The specific structure is shown in Fig. 3.

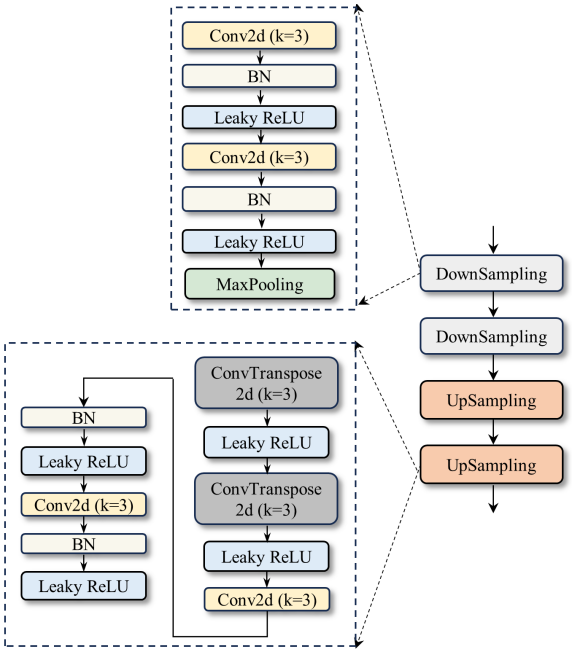


Fig. 3. The structure of the encoder network. BN stands for Batch Normalization.

The encoder network is a two-dimensional fully convolutional neural network based on the U-Net [19] architecture. It consists of a contraction part and an expansion part, which are responsible for the downsampling and upsampling steps. The contraction part comprises two downsampling modules, each containing two 3×3 convolutional layers. Batch normalization and Leaky ReLU [20] activation function are applied after each convolutional layer. The last layer of the downsampling module is a max-pooling layer, used to reduce the spatial dimensions of the speaker embeddings and capture important features. The expansion part consists of two upsampling modules, each primarily composed of two transpose convolutional layers and two 3×3 convolutional layers. Transpose convolutional layers are used to increase the spatial dimensions of the speaker embeddings and recover fine-grained information. In addition, each speaker embedding used for steganography is a matrix of size 128×1.

2) Graph Encoder: To embed richer and more accurate speaker embeddings in converted speech, we introduce the graph encoder proposed in [15]. Specifically, we construct a graph representation for each converted speech and encode its structural details using a dual-layer GCN encoder. The encoded graph features will complement the original feature representation, thereby enhancing the accuracy and richness of the overall feature representation.

The detailed structure of the graph encoder is shown in Fig. 4. After the converted speech undergoes Short-Time Fourier Transform (STFT) operation, a sequence X is obtained. Each element in the sequence X can be regarded as a node in the graph and the connections between nodes can be characterized by graph relationships.

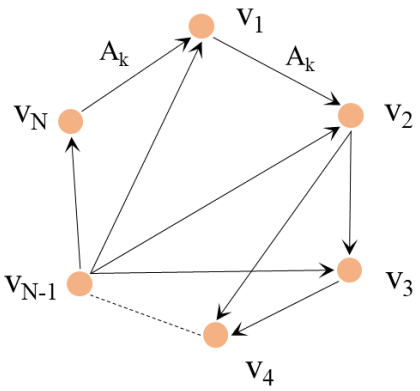


Fig. 4. Illustration of the graph encoder. Each orange dot represents an element in the sequence X, while the value of A_k on each line reflects the presence of structural details between two points.

Specifically, the graph encoder establishes a graph representation $G_{(t)} = (\mathcal{V}_t, \mathbf{A}_k, \mathbf{A}_k)$ for each sequence $X \in \mathbb{R}^{N \times T}$, where $\mathcal{V}_t$ is the set of vertices. Each element in the sequence is treated as a graph vertex, with its value representing the signal residing on the vertex, i.e., the identifier of vertex X at that location. The graph encoder maps the sequence X into the graph domain and examines the structural details among the elements. Assuming there exist structural details between the i_{th} element indexed by vertex v_i and the j_{th} element indexed by vertex v_j, $\mathbf{A}_k(i,j)$ is set to 1; otherwise, $\mathbf{A}_k(i,j) = 0$.

C. Speech Reconstruction

At the receiving end, the speech reconstruction system extracts the hidden speaker embedding from the transmitted converted speech, which is achieved through the decoder module. It is worth noting that the structure of the decoder

module is identical to the encoder module used in speech steganography.

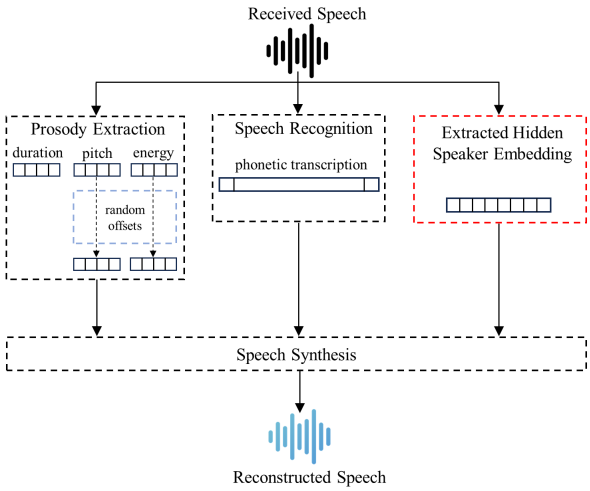


Fig. 5. Illustration of the speech reconstruction module. The part highlighted in red box indicates the difference from Fig. 2.

Besides, the architecture of the reconstruction module is essentially identical to that of the pipeline in the first part of CSSS. The key distinction lies in the reconstruction module's omission of GAN-generated artificial speaker embeddings. Instead, it utilizes the speaker embedding extracted by the decoder module to synthesize the original speech by incorporating the prosody and textual information of the converted speech obtained through transmission. The structure is shown in Fig. 5.

III. Experimentals And Results

A. Setup

The training of the pipeline in the first part of CSSS follows the experimental setup proposed in [18]. The ASR model is trained using the VCTK dataset. We utilize SpeechBrain [21] to train the embedding extractor on approximately 2,800 hours of speaker-validated data from VoxCeleb 1 and 2 [22]–[24]. To synthesize speech, we train the TTS system using the LibriTTS corpus [25]. For the GAN model, we augment the LibriTTS data with the RAVDESS [26] and Emotion Speech Dataset (ESD) [27] to obtain a larger and more diverse corpus for training purposes. For the second part of CSSS, we use VoxCeleb2 to train our speech steganography network. For the third part of CSSS, the speech reconstruction module utilizes the pre-trained model from the non-target voice conversion pipeline.

We validate the performance of CSSS using a subset of LibriSpeech and VCTK as a test set and divided into trial and enrollment data. The privacy of the transmitted speech used for speech communication in CSSS is evaluated using the Automatic Speaker Verification (ASV) attacker of Voice Privacy Challenge 2020 (VPC'20) and measured in terms of Equal Error Rate (EER). The utility of CSSS is measured using the Word Error Rate (WER) of the Voice Privacy Challenge 2022 (VPC'22) ASR model. We compared the CSSS to the original speech and to the main baselines of VPC'20 (BL 1) and VPC'22 (BL 1.b).

B. Evaluation metrics

1) Objective Evaluation: Table 1 lists the EER results for the privacy metric in the first column. For the first four items, the closer the EER score is to 50%, the more difficult it is for attackers to identify the true speaker of the audio. Compared with the original data and the two baseline systems of VPC, the steganographic converted speech used for speech communication in CSSS achieved high privacy standards for both datasets. As for the fifth item, the privacy metric of the reconstructed speech is close to that of the original data, indicating that CSSS can indeed recover the speaker's identity well.

TABLE I
OBJECTIVE EVALUATION RESULTS COMPARED TO BASELINE AND PROPOSED SYSTEMS. IN THE FIRST FOUR ITEMS, THE CLOSER THE EER IS TO 50%, THE BETTER, AND THE LOWER THE WER, THE BETTER. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN BOLD. FOR THE FIFTH ITEM, IT IS COMPARED SEPARATELY TO THE ORIGINAL, AND THE LOWER THE EER AND WER, THE BETTER, INDICATING THE PROXIMITY OF THE RECONSTRUCTED SPEECH BY CSSS TO THE ORIGINAL SPEECH.

model	EER(%)		WER(%)	
	Libri	VCTK	Libri	VCTK
Original	4.39	3.19	4.15	12.82
VPC'20 - BL 1	34.44	31.30	6.73	15.23
VPC'22 - BL 1.b	31.80	24.11	6.08	15.60
CSSS' Steg	**48.92**	**51.74**	**5.93**	**11.07**
CSSS' Recon	5.66	3.70	4.21	11.47

The second column of Table 1 shows that in the first four items, steganographic converted speech still outperforms both baseline systems in terms of the language content (WER) metric. On the VCTK dataset, the WER is even reduced by 13.65% compared to the original data. In addition, the WER metrics of the fifth reconstructed speech are very close to the original data. This indicates that the CSSS-processed transmitted speech and reconstructed speech still have high utility while achieving privacy preservation and restoring the original speaker identity, respectively.

On the other hand, the accuracy of extracting the hidden speaker embedding from the receiving end is very important for the whole system. We use cosine similarity to measure the similarity between the speaker embedding extracted by the decoder module and the original speaker embedding. The results in Table 2 show that the average cosine similarity over the entire test set reaches 0.9999 in the system introducing the graph encoder, with almost no loss of speaker embedding information. However, the system with the introduction of the graph encoder has a reduced SNR metric compared to the system without it, but as the audio quality is still high, and availability of accurate original speaker embedding at the receiver is more critical for reconstructing the original speech.

Therefore, in order to achieve almost lossless transmission of speaker embedding, the introduction of graph encoders is necessary for CSSS.

TABLE II
RESULTS ARE DIVIDED INTO TWO CATEGORIES: CSSS WITH AND WITHOUT GRAPH ENCODER. WE MEASURED THE AVERAGE SNR OF THE STEGANOGRAPHIC SPEECH OBTAINED BY DIFFERENT SYSTEMS ON THE ENTIRE TEST SET, AS WELL AS THE AVERAGE COSINE SIMILARITY BETWEEN THE SPEAKER EMBEDDINGS EXTRACTED BY THE RECEIVING END AND THE ORIGINAL SPEAKER EMBEDDINGS.

System	SNR of Speech	Cos Similarity
with Graph Encoder	38.514	**0.9999**
w/o Graph Encoder	**55.557**	0.9948

2) Subjective Evaluation: In our task, human perception of the original speech and the final reconstructed speech at the receiving end is crucial. This is because if the reconstructed speech at the receiving end diverges significantly from the original speech, users will doubt the authenticity of the transmitted speech information. Therefore, we randomly selected 20 utterances from 10 different speakers from the test set. Each utterance was presented along with the corresponding output from different systems, to evaluate the perceptual similarity, naturalness, and intelligibility of the speech obtained by different systems compared to the original speech. Then, at least 7 participants rated each metric on a 1-5 Likert scale.

TABLE III
MOS MEASURES THE NATURALNESS, INTELLIGIBILITY, AND PERCEPTUAL SIMILARITY BETWEEN THE ORIGINAL UTTERANCE AND THE OUTPUT UTTERANCE OF DIFFERENT SYSTEMS ON A SCALE OF 1-5. THE OUTPUTS OF THE SECOND THROUGH FOURTH SYSTEMS ARE USED FOR PRIVACY PROTECTION, AND THE SIMILARITY METRICS SHOULD BE AS LOW AS POSSIBLE AND NOT COMPARABLE TO THE RECONSTRUCTED SPEECH, SO THE THIRD COLUMN DOES NOT HAVE THE DATA HIGHLIGHTED IN BOLD.

Model	naturalness		intelligibility		similarity	
	MOS	σ	MOS	σ	MOS	σ
original	*4.64*	$\pm$0.75	*4.58*	$\pm$0.43	-	-
VPC'20 - BL 1	4.17	$\pm$0.67	4.29	$\pm$0.79	1.47	$\pm$0.32
VPC'22 - BL 1.b	3.71	$\pm$0.87	4.40	$\pm$0.56	1.55	$\pm$0.44
CSSS' Steg	4.32	$\pm$0.67	**4.67**	$\pm$0.72	1.43	$\pm$0.41
CSSS' Recon	**4.48**	$\pm$0.56	4.65	$\pm$0.54	4.23	$\pm$0.61

Based on the Mean Opinion Score (MOS) and its standard deviation σ shown in Table 3, it can be seen that compared to other systems, the steganographic and reconstructed speech from CSSS both have high intelligibility and naturalness, and the reconstructed speech also achieves high perceptual similarity with the original speech. This further confirms the effectiveness of CSSS in restoring the original speech.

IV. CONCLUSION

In this paper, we present a system called CSSS (non-target Conversion based Speech Steganography System), designed to prevent the leakage of speaker identity information during speech communication. The CSSS system is capable of effectively concealing the speaker's identity and embedding the original speaker's embedding information into converted speech for transmission. Additionally, by extracting the hidden speaker embedding information at the receiving end, CSSS can successfully reconstruct the original speech. With these designs, the CSSS not only avoids the leakage of speaker identity information during speech communication but also prevents suspicions regarding the authenticity of the speech information due to a mismatch between the converted speech and the known sender's voice. Through experimental validation and manual evaluation, it is confirmed that the proposed CSSS has an excellent performance in speaker embedding steganography, and the speech reconstructed by the system has high intelligibility, naturalness, and perceptual similarity with the original speech.

REFERENCES

[1] Z. Wu, J. Yamagishi, T. Kinnunen, *et al.*, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[2] P. Regulation, "Regulation (eu) 2016/679 of the european parliament and of the council," *Regulation (eu)*, vol. 679, p. 2016, 2016.

[3] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The gdpr & speech data: Reflections of legal and technology communities, first steps towards a common understanding," *arXiv preprint arXiv:1907.03458*, 2019.

[4] H. Lu, X. Wu, H. Guo, S. Liu, Z. Wu, and H. Meng, "Unifying one-shot voice conversion and cloning with disentangled speech representations," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 141–11 145. DOI: 10.1109/ICASSP48485.2024.10446296.

[5] J. Lim and K. Kim, "Wav2vec-vc: Voice conversion via hidden representations of wav2vec 2.0," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 326–10 330. DOI: 10.1109/ICASSP48485.2024.10447984.

[6] Z. Łatka, J. Gałka, and B. Ziółko, "Cross-gender voice conversion with constant f0-ratio and average background conversion model," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6825–6829. DOI: 10.1109/ICASSP.2019.8683369.

[7] I.-C. Yoo, K. Lee, S. Leem, H. Oh, B. Ko, and D. Yook, "Speaker anonymization for personal information protection using voice conversion techniques," *IEEE Access*, vol. 8, pp. 198 637–198 645, 2020. DOI: 10.1109/ACCESS.2020.3035416.

[8] M. Matassoni, S. Fong, and A. Brutti, "Speaker anonymization: Disentangling speaker features from pre-trained speech embeddings for voice conversion," *Applied Sciences*, vol. 14, no. 9, 2024, ISSN: 2076-3417.

DOI: 10.3390/app14093876. [Online]. Available: https://www.mdpi.com/2076-3417/14/9/3876.

[9] M. Zamani, A. A. Manaf, R. Ahmad, A. M. Zeki, and S. Abdullah, "A genetic-algorithm-based approach for audio steganography," *World Academy of Science Engineering & Technology*, no. 54, p. 360, 2009.

[10] D. Gruhl, "Echo hiding, proceedings of information hiding," in *International Workshop*, 1996.

[11] R. M. Nugraha, "Implementation of direct sequence spread spectrum steganography on audio data," in *International Conference on Electrical Engineering & Informatics*, 2011.

[12] S. Agarwal, R. Soun, R. Shivani, V. Varanasi, N. Gill, and R. Sawhney, "Hypersteg: Hyperbolic learning for deep steganography," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[14] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*, PMLR, 2019, pp. 7354–7363.

[15] T. Wang, Z. Pan, M. Ge, Z. Yang, and H. Li, "Time-domain speech separation networks with graph encoding auxiliary," *IEEE Signal Processing Letters*, vol. 30, pp. 110–114, 2023.

[16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016.

[17] N. Tomashenko, X. Wang, E. Vincent, *et al.*, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101 362, 2022.

[18] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[20] J. Xu, Z. Li, B. Du, M. Zhang, and J. Liu, "Reluplex made more practical: Leaky relu," in *2020 IEEE Symposium on Computers and communications (ISCC)*, IEEE, 2020, pp. 1–7.

[21] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[22] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101 027, 2020.

[23] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[24] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[25] H. Zen, V. Dang, R. Clark, *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[26] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, e0196391, 2018.

[27] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 920–924.

A Multi-Domain Camera Model Identification Feature Restoration Network to Counter AI Compression Attacks

Jinkai Zhang†, Zijuan Han‡, Yunxia Liu‡ and Yang Yang†*
† School of Information Science and Engineering, Shandong University, Qingdao, China
‡ Center for Optics Research and Engineering (CORE), Shandong University, Qingdao, China
* Corresponding Author
Email:202232729@mail.sdu.edu.cn

***Abstract*—Camera model identification(CMI) has been a well-established passive forensic technique in recent decades. Previous research primarily focused on unaltered images, while real-world scenarios often involve malicious attacks. With the rapid advancement of AI compression technology, it is becoming increasingly popular and prevalent in various applications. However, our experimental study demonstrates that the camera model identification drops dramatically when applied to AI-compressed images without any processing. In order to enhance the robustness of the camera model identification algorithm against AI compression attacks, we propose a multi-domain camera model identification feature restoration Network. Firstly, a spatial and frequency domain dual branch architecture is proposed to restore forensic features from AI compressed image patches. The frequency feature restoration module employs an unbalanced restoration strategy in amplitude and phase for efficient network training and improved identification robustness. Afterwards, a feature fusion and selection module are utilized to fuse and filter restored features, where double pooling cross fusion strategy is adopted in channel attention mechanism for making full use of channel statics. Experimental results show that the proposed method can efficiently restore camera model related forensic features and significantly improve the performance of the CMI model on AI compressed images.**

I. Introduction

The widespread of imaging devices and advancements in image generation algorithms have significantly enhanced the accessibility and quality of images. However, this presents substantial challenges in image authenticity verification [1]. Being a passive forensic technique, Camera model identification(CMI) aims to determine the origin of images by analyzing features such as photo response non-uniformity, sensor noise patterns [2] and so on. It has found wide applications in image forensics, copyright protection, and privacy protection [3].

In recent years, deep learning-based image compression techniques have made significant progress. The Jpeg Association is working on the JPEG AI standard [4]. AI compression is considered as a strong candidate for the next generation of image compression techniques, and is highly likely to receive widespread deployment in industrial applications in the future [5], [6]. However, malicious use of AI compression can seriously disrupt forensic and investigative efforts[4]. To the best of our knowledge, how AI compression affects the accuracy and effectiveness of CMI has not been reported in literatures.

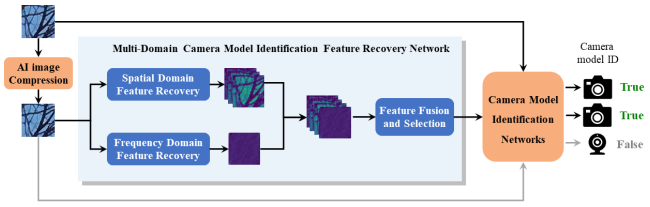


Fig. 1. AI compression significantly decreases the accuracy of camera model identification(CMI). The proposed MDFRNet aims to restore CMI related forensic features from multiple domains to adapt AI compressed patches to prevalent CMI networks for correct camera model prediction.

In this study, we carry out experimental evaluation of the camera model identification performance taken AI compressed image patches as network input to current prevalent CMI networks. Results reveal that most CMI networks fail without any processing. To address this issue, a Multi-Domain camera model identification Feature Restoration Network(MDFRNet) is proposed. As depicted in Fig.1, a dual branch architecture is adopted to restore complementary features from the frequency and spatial domains and successfully improves the accuracy of the CMI network on AI compressed images. The primary contributions of this study are summarized as follows:

- This study pioneers the study of AI compression attack in camera model identification, highlighting that while AI compression achieves high-quality image compression, it often alters critical source camera features, making it susceptible to malicious attacks.
- A multi-domain camera model identification feature restoration network(MDFRNet) is proposed to restore forensic features from both spatial and frequency domains, and successfully improve CMI accuracy on AI-compressed images.
- Experimental validation across multiple CMI methods demonstrate the method's effectiveness in restoring identification features distorted by AI compression. The results affirm the robustness and versatility of proposed network.

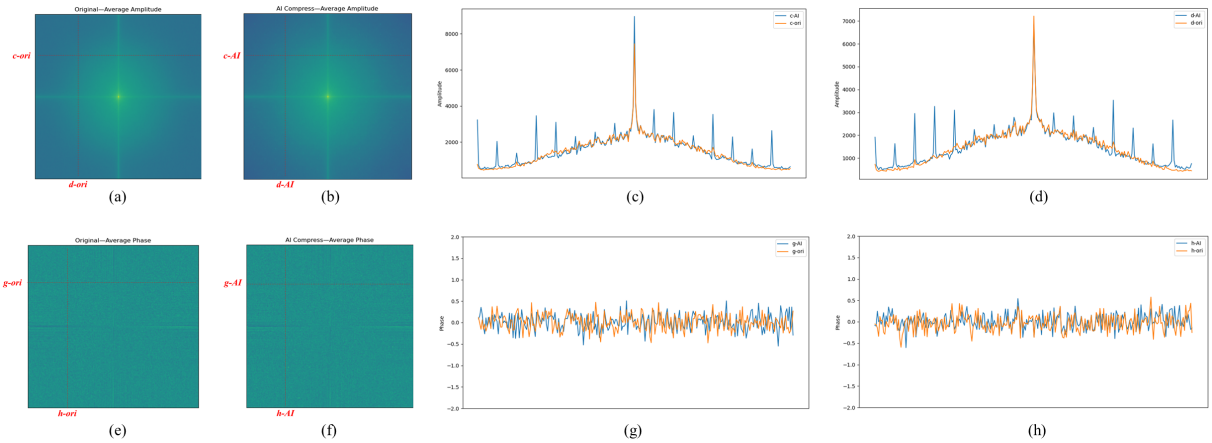


Fig. 2. Frequency domain comparison of images with and without HIFIC compression. (a) Average amplitude of 100 images. (b) Average amplitude of HIFIC compressed images. (c) A typical visualization of lateral average amplitude. (d) A typical visualization of longitudinal average amplitude. (e) Average phase of original images. (f) Average phase of HIFIC compressed images. (g) A typical visualization of lateral average phase. (h) A typical visualization of longitudinal average phase.

II. Related Works

How to extract camera model related features from noise residual to suppress the influence of image content plays an important role in camera model identification. To address this issue, Davide et al. [7] developed a noise residual extraction method using frequency domain constraints. Rafi et al. [8] proposed the Remnet to enhance intrinsic camera fingerprint by suppressing image content, and then combined a dynamic CNN-based preprocessing block with a shallow classifier and verified the applicability of residual blocks for CMI tasks. Bennabhaktula et al. [9] applied MobileNet with pre-trained weights from ImageNet for transfer learning, and study the relationship between deep learning network complexity and recognition accuracy. Our previous work [10] improved VGG with a fine-grained multi-scale residual prediction module to boost recognition accuracy and reduce scene content influence. Huan et al. [11] proposed ConvNeXt with a dual-path attention mechanism and classical residual extraction to capture crucial model features for camera provenance. Although literatures have reported satisfying CMI results, for instance, the state-of-the-art Remnet [8] has reached 94.6% accuracy on Dresden dataset, and DPEC ConvNeXt [11] has achieved 83.1% on VISION [12] dataset. Images inevitably undergo various post-processing during transmission in practice, which will greatly affect its application in real scenarios.

In recent years, with the continuous development of AI compression technology, deep learning-based image compression techniques have made significant progress. Some recent studies [5], [13] have greatly outperformed traditional JPEG in terms of PSNR and SSIM metrics. Toderici et al.[14] introduced a full-resolution image compression method using RNNs and neural networks for entropy coding. Balle et al.[15] proposed an end-to-end variational autoencoder model, incorporating hyperprior and spatial autoregressive models to optimize latent representation encoding. GAN-based compression algorithms now achieve high-quality images at very low bit rates. Mentzer et al. [13] developed HiFiC, a GAN-based technique that achieves perceptual fidelity close to the original image quality at half the original bit rate. AI compression is undergoing rapid development and is expected to be key technology of the next generation of image compression standards in the near future. Consequently, study of potential risks it may bring to related applications is necessary.

Although visually consistent with the original image, AI compression brings certain artifacts as a lossy compression method. Bergmann et al.[6] were the first to study the compression artifacts of HiFiC in the frequency domain, and the detectability of these artifacts was evaluated. Berthet et al. [4] have realized that AI compression is a novel unintended attack method, and used HiFiC to attack the advanced tamper detection network CatNet[16] to verify the aggressiveness of AI compression in the field of tamper detection. To the best of our knowledge, there is no in-depth study on the impact of AI compression in the field of CMI at present. However, existing research on source camera identification mainly focuses on the detection of raw images, ignoring the threats and challenges that AI compression may bring to the source camera identification task.

III. The Proposed Method

A. Frequency domain analysis of AI compressed images

Frequency domain analysis is very important to study the underlying characteristics of artifacts introduced by AI compression. Referring to the work of Bergmannet [6], we experimentally compare the amplitude and phase difference between original and AI compressed images which forms the basis of our later proposed network and loss function design.

We randomly selected 100 images from the VISION dataset [12], which is a widely acknowledged dataset in camera model identification field. Central cropping is applied to both original and HIFIC compressed sets to obtain series of

256×256 patches. Fast Fourier Transform with N = 256 is applied and the average amplitude is calculated for enhanced feature robustness. Average amplitude of the original and AI compressed images are shown in Fig.2(a) and (b), respectively. To further compare their differences intuitively, we depict typical lateral values in Fig.2(c) and longitudinal values in Fig.2(d). Obviously, we can observe that the spectrogram of the AI compressed images regularly shows peaks at 16 bisection positions [6]. Furthermore, the average amplitude maintains stable with respect to image contents when number of images is greater than 50 in our experiment. Consequently, it can serve as a template to regularize frequency domain features in network training.

The average phase of the original and AI compressed images are shown in Fig.2(e) and (f), respectively. Similarly, to compare the difference introduced by HIFIC compression, we also depict a typical lateral and longitudinal slice in Fig.2(g) and (h). As compared with amplitude, phase variations are more disordered and complex. Due its underlying modulus 2π characteristic of phase information, there is no regularity observed in either average phase nor phase of any specific image in our experiments. In contrast to the amplitude which is intuitive and easy to understand, the complex of phase has been reported in literatures [17], [18], and has long been the reason that limits its application in feature representation. However, phase contains rich information of nearby pixels and is of vital importance to restore camera model related subtle features. We will report our effort in exploiting phase information in CMI feature restoration in subsequent sections.

B. Overview of proposed network

Given the fact that camrea model identificaiton (CMI) resutls is significantly affected by artifacts introduced by AI compression, our motivation is to design a CMI feature restoration network to restore camera model identification features. The overall architecture of the proposed multi domain feature restoration network (MDFRNet) is illustrated in Fig.3.

Since the state of the art CMI networks have been able report satisfying performance on patches of size 64×64, we also follow this setting. For the given input patch of 64×64, the spatial domain feature restoration (SFR) module and frequency domain feature restoration (FFR) module are applied to restores camera model related forensic features from the detection features. Then, concatenated features are fused and enhanced by the feature fusion and selection(FFS) module with specifically designed channel attention mechanism. Finally, the fused features are fed into the CMI network to obtain final camera identification results.

C. Multi-domain feature restoration module

To make better use of advantages of both spatial domain and frequency domain methods, a dual branch multi-domain feature restoration module is proposed. As depicted in Fig.3, taking RGB patches as input, the spatial feature restoration(SFR) module aims to restore spatial domain camera model related forensic features with a UNet architecture. SFR outputs are more intuitive as it directly operates on pixel values and is able to provide better local relationship characterization. While the frequency feature restoration (FFR) module has more advantages in dealing with global features and periodic components. As restored features of SFR and FFR modules are finally concatenated for further fusion and selection, special attention is paid to network design so that they have same dimensions of feature maps.

In the encoding stage of the spatial feature restoration (SFR) module, the 3-channel RGB patch is firstly passed through a 3×3 convolutional layer to adjust the number of channels to 64. Each of the next four convolutional layers utilizes a 3×3 convolution kernel to double the number of feature channels to 128 , whose feature map sizes are then halved by Max pooling. In the decoding stage, the feature map is passed through a convolution layer with a kernel size of 3×3 to halve the number of channels. At the same time, the feature map of the encoding layer is spliced with the feature map of the decoding layer, and the number of channels of the feature map is gradually halved. After four convolution and pooling operations, we adjust the number of feature channels to 3 through an output convolution layer with a kernel size of 1×1.

In the frequency feature restoration(FFR) module, different restoration strategies are adopted for amplitude and phase. As for amplitude restoration, we intend to remove the effect of AI compression by subtracting the average amplitude increment (shown as Amplitude Mask in Fig.3). The average amplitude increment is a residual obtained by subtracting the average amplitude of the original patches from those of the AI compressed patches. As for phase restoration, phase of AI compressed patches are fed into the same Unet structure for feature restore, where the original phase is utilized in a supervised manner to guide the network learning of complex phase features. Finally, Inverse Fast Fourier Transform is used to obtain FFR restoration features. Considering the frequency domain stability, FFR also contributes to network training difficulty reduction.

D. Feature fusion and selection module

The dual branch FFR and SFR modules offers complementary benefits. However, it can't be ignored that there is certain redundancy in these two features. A feature fusion and selection module(FFS) is proposed to fully exploit their complementarity and minimize redundancy. As shown in Fig 3, based on spatial correspondence and complementarity between spatial and frequency domain restoration features, the upper branch fuses these features using two 3×3 convolution layers. For the lower branch, we employ an improved channel statistical feature calculation method to reduce feature redundancy. Specifically, we employ two pooling modalities commonly used in channel attention mechanisms: Max pooling and average pooling. Pooling values are fed into two convolution layers with kernel size of 1×1, to obtain the inner features that will be added to their own pooling values and the outer features that will be added to other pooling values. The inner

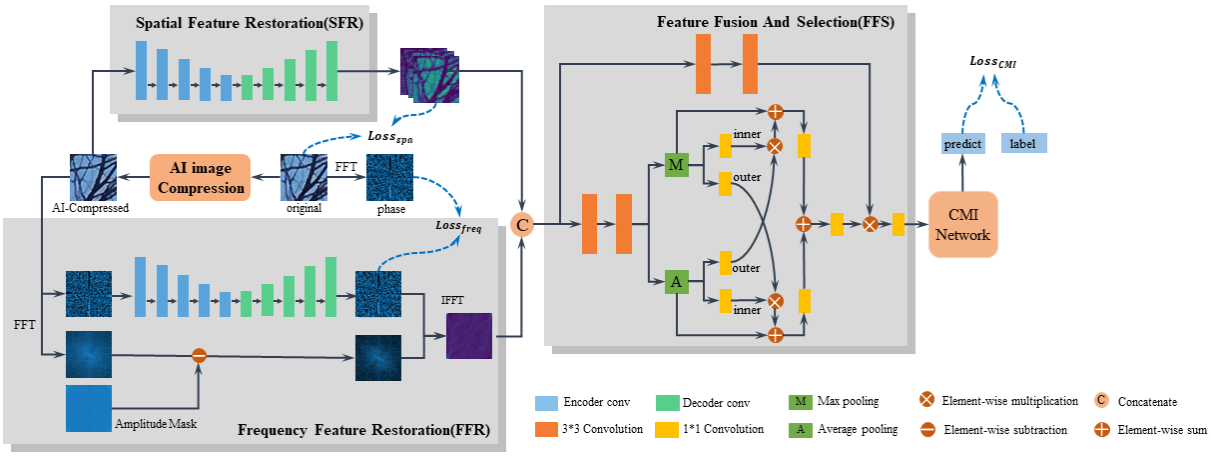


Fig. 3. Overall network structure of the proposed multi domain feature restoration network(MDFRNet).

features are multiplied with the outer features of the other pooling, and then added to the pooling value to output the cross values of each pooling. The maximum cross values and the average cross values are added and convolved to obtain the channel pooling information. The channel pooling information and the features output by the feature restoration network are multiplied channel by channel to obtain the fused and selected restored features.

After a 1×1 convolutional layer to adjust the number of channels, the restored features are sent to the CMI network for detection. This dual-pooling cross-fusion strategy maximizes the utilization of channel statistical information, significantly enhancing feature representation.

E. Loss Function

To guide efficient learning of CMI related features, we utilize three kinds of loss functions in the proposed network, which are the spatial feature restoration loss, the frequency feature restoration loss, and the camera model identification loss.

(1)Spatial feature restoration loss: $Loss_{spa}$ evaluates the spatial feature restoration capability of SFR module by comparing restored spatial features with the original patch:

$$Loss_{spa} = \frac{1}{M}\sum_{j=1}^{M}(y_j - p_j)^2 \tag{1}$$

where y and p denotes the ground truth pixel value in the original patch and restored output of the SFR module, M represents the number of pixels within a patch.

(2)Frequency feature restoration loss: $Loss_{freq}$ evaluates the frequency feature restoration capability by comparing the consistency of restored phase features with phase features of the original image. Inspired by Noiseprint [10], we utilize the geometric mean to arithmetic mean ratio of image phases to constrained training of the FFR. For the original image $I_{ori}(x, y)$, the frequency domain representations $F_{ori}(x, y)$ is obtained by Fourier transform:

$$I(x, y) \overset{FFT}{\rightarrow} F(u, \nu) \tag{2}$$

The phase $P_{ori}(u, \nu)$ is calculated by the phase formula as follows:

$$P(u, v) = \arctan\left(\frac{Im\left(F\left(u, \nu\right)\right)}{Re\left(F\left(u, \nu\right)\right)}\right) \tag{3}$$

where $Im\left(F\left(u, \nu\right)\right)$ denotes the imaginary part of $F\left(u, \nu\right)$ and $Re\left(F\left(u, \nu\right)\right)$ denotes the real part of $F\left(u, \nu\right)$.

Calculate the residual between the $P_{ori}\left(u, \nu\right)$ and phase restoration features $P_{AI}\left(u, \nu\right)$ output to get $R\left(u, \nu\right)$:

$$R\left(u, \nu\right) = P_{ori}(u, v) - P_{AI}(u, v) \tag{4}$$

Assuming that the input batchsize is N, the average phase of each batch can be expressed as $S\left(u, \nu\right)$:

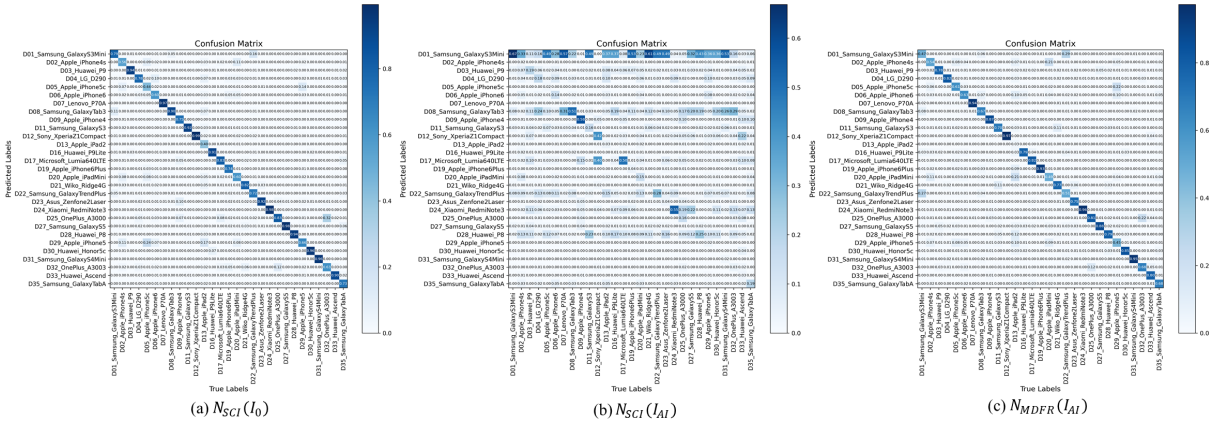
$$S\left(u, \nu\right) = \frac{1}{N}\sum_{k=1}^{N}\left|R_k(u, \nu)\right|^2 \tag{5}$$

The average loss per batch can be obtained by calculating the ratio of the geometric mean to the arithmetic mean of $S\left(u, \nu\right)$ and then taking the logarithm:

$$\begin{aligned} Loss_{freq} &= \log\left[\frac{S_{GM}}{S_{AM}}\right] \\ &= \left[\frac{1}{K^2}\sum_{u,\nu}\log\left(S\left(u, \nu\right)\right)\right] - \log\left[\frac{1}{K^2}\sum_{u,\nu}S\left(u, \nu\right)\right] \end{aligned} \tag{6}$$

Where K represents the input feature size. The ratio of the geometric mean to the arithmetic mean reflects the characteristics of the data distribution: The closer the ratio is to 1, the better the phase feature is restored, and the closer the ratio is to 0, the worse the restored phase feature is.

Source camera identification loss $Loss_{CMI}$ is utilized to provide supervised information by labeled source camera ground truth. It is calculated based on the difference between



(a) $N_{SCI}(I_0)$ (b) $N_{SCI}(I_{AI})$ (c) $N_{MDFR}(I_{AI})$

Fig. 4. Confusion matrix comparison of AI compressed images with/without feature restoration. (a) CMI of the original images (b) CMI of compressed images without feature restoration. (c) CMI of compressed images with feature restored by the proposed MDFRNet.

the camera identification predicted by the network and the actual label of the original patch as:

$$Loss_{CMI} = -\sum_{i=1}^{K} q\left(x_i\right) \log p\left(x_i\right) \tag{7}$$

where q denotes the target distribution and p denotes the predicted matching distribution, and K is total number of camera models involved.

The total training loss L can be expressed as follows:

$$Loss = \lambda_1 Loss_{spa} - \lambda_2 Loss_{freq} + \lambda_3 Loss_{CMI} \tag{8}$$

where λ_1, λ_2 , and λ_3 represent hyper-parameter controlling contributions of different losses that can be empirically determined.

IV. EXPERIMENTAL RESULT

To evaluate the effectiveness of the proposed method, we conduct series of experiments based on the Vision dataset, which is the most well acknowledged dataset in camera model identification field. We followed the experimental settings of our previous work[11] that we selected 29 camera models from them, and each camera model selected 100 images to segment the patch, and the shooting scene was not restricted. Our experiments are based on the deep learning framework Pytorch 1.13.0 and Python 3.8.18. We used the RMSprop optimizer with an initial learning rate of 0.0002 and weight decay of 0.00001. The λ_1 is set to 0.6, λ_2 to 0.5, and λ_3 to 0.8. We performed all the experiments reported in the paper using a Nvidia A4000 graphics card.

In this paper, we verify the effectiveness of our proposed MDFRNet on five CMI models. The results are shown in TABLE 1. We use I_0 to represent the original image, N_{CMI} represents the CMI network, I_{AI} represents the image compressed by AI algorithm, N_{MDFR} means patches are restored by MDFRNet before being sent to CMI network. The right three columns represent the results for different cases. The results show that after AI compression, the accuracy of all CMI models decreases significantly, indicating that AI compression is an effective attack method for the CMI task. After feature restoration via the proposed MDFRNetwork, the detection accuracy of all models is significantly improved. This shows that the proposed MDFRNetwork can effectively resist the AI compression attack for CMI.

TABLE I
IDENTIFICATION ACCURACY COMPARISON OF STATE-OF-THE-ART CAMERA MODEL IDENTIFICATION NETWORKS WITH/WITHOUT FEATURE RESTORATION.

Methods	$\mathbf{N_{CMI}(I_0)}$	$\mathbf{N_{CMI}(I_{AI})}$	$\mathbf{N_{MDFR}(I_{AI})}$
RemNet, Rafi	79.9	10.1	69.7
Res2Net, Liu	70.7	6.8	65.1
ResNet50, Bennabhaktula	78.2	10.1	66.3
MobileNet, Bennabhaktula	76.4	16.3	68.0
DPEC ConvNext, Huan	83.1	16.2	79.0

In Fig. 4, we show the confusion matrices corresponding to $N_{CMI}(I_0)$,$N_{CMI}(I_{AI})$ and $N_{MDFR}(I_{AI})$ using Huan's model. By comparing Fig.4(a) with Fig.4(b), it can be seen the identification accuracy for most camera models drops below 50%. It proves that AI compression causes damage to the camera fingerprint. However, as shown in Fig.4(c), when the identification features are restored by MDFRNet, the identification accuracy of each camera model is significantly improved.

We also compare with the fine-tuning, and do ablation experiments on the network proposed in this paper. The results are shown in Table 2. The fine-tuning uses weight trained using original images, and the corresponding AI compressed patches are used for training.

Setup#2 denotes that the average amplitude increment is excluded, with channel attention relying solely on max pooling. Setup#3 similarly indicates that the average amplitude increment is not utilized. Setup#4 specifies that only the max-pooling strategy is employed. Setup#5 signifies that the feature restoration network is not applied for phase reconstruction. Setup#6 represents our proposed method. The results

TABLE II
HUAN'S MODEL IS USED AS THE CMI NETWORK. EVEN WITH THE REDUCTION OF SOME PROCESSING STEPS, OUR PROPOSED METHOD CAN STILL OUTPERFORM THE COMMON FINE-TUNING(SETUP#1).

Setup	Accuracy
#1 Fine-tune	72.5
#2 backbone	72.9
#3 wo/res	76.8
#4 wo/at	74.3
#5 wo/phase_restore	76.2
#6 proposed	79.4

of these Settings provide further strong evidence to support the effectiveness of our proposed method in restoring the identification features.

V. CONCLUSIONS

In this paper, we experimentally evaluated current camera model identification networks on AI compressed images for the first time and found that AI compression is a powerful attack against the CMI task. To counter this new attack, a multi-domain camera model identification feature restoration network is proposed. Comprehensive experiments are conducted to evaluate the effectiveness of the proposed method. Significant identification performance improvement is observed on AI compressed images. At present, our work is only based on the AI compression algorithm. In the future, we will further investigate the impact of new attacks such as using large models to generate images on the camera model recognition task.

REFERENCES

[1] X. Lin, J. H. Li, S. L. Wang, A. W. C. Liew, F. Cheng, and X. S. Huang, *Recent advances in passive digital image security forensics: A brief review. engineering 4 (1): 29–39*, 2018.

[2] L. Bondi, L. Baroffio, D. Güera, P. Bestagini, E. J. Delp, and S. Tubaro, "First steps toward camera model identification with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 259–263, 2016.

[3] V. U. Sameer and R. Naskar, "Deep siamese network for limited labels classification in source camera identification," *Multimedia Tools and Applications*, vol. 79, no. 37, pp. 28 079–28 104, 2020.

[4] A. Berthet and J.-L. Dugelay, "Ai-based compression: A new unintended counter attack on jpeg-related image forensic detectors?" In *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 3426–3430.

[5] H. E. Egilmez, A. K. Singh, M. Coban, *et al.*, "Transform network architectures for deep learning based end-to-end image/video coding in subsampled color spaces," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 441–452, 2021.

[6] S. Bergmann, D. Moussa, F. Brand, A. Kaup, and C. Riess, "Frequency-domain analysis of traces for the detection of ai-based compression," in *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, IEEE, 2023, pp. 1–6.

[7] D. Cozzolino and L. Verdoliva, "Noiseprint: A cnn-based camera model fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 144–159, 2019.

[8] A. M. Rafi, T. I. Tonmoy, U. Kamal, Q. J. Wu, and M. K. Hasan, "Remnet: Remnant convolutional neural network for camera model identification," *Neural Computing and Applications*, vol. 33, pp. 3655–3670, 2021.

[9] G. S. Bennabhaktula, D. Timmerman, E. Alegre, and G. Azzopardi, "Source camera device identification from videos," *SN Computer Science*, vol. 3, no. 4, p. 316, 2022.

[10] Y. Liu, Z. Zou, Y. Yang, N.-F. B. Law, and A. A. Bharath, "Efficient source camera identification with diversity-enhanced patch selection and deep residual prediction," *Sensors*, vol. 21, no. 14, p. 4701, 2021.

[11] S. Huan, Y. Liu, Y. Yang, and N.-F. B. Law, "Camera model identification based on dual-path enhanced convnext network and patches selected by uniform local binary pattern," *Expert Systems with Applications*, vol. 241, p. 122 501, 2024.

[12] D. Shullani, M. Fontani, M. Iuliani, O. A. Shaya, and A. Piva, "Vision: A video and image dataset for source identification," *EURASIP Journal on Information Security*, vol. 2017, pp. 1–16, 2017.

[13] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 913–11 924, 2020.

[14] G. Toderici, D. Vincent, N. Johnston, *et al.*, "Full resolution image compression with recurrent neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5306–5314.

[15] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[16] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, and C. Kim, "Learning jpeg compression artifacts for image manipulation detection and localization," *International Journal of Computer Vision*, vol. 130, no. 8, pp. 1875–1895, 2022.

[17] A. Leclaire and L. Moisan, "No-reference image quality assessment and blind deblurring with sharpness metrics exploiting fourier phase information," *Journal of Mathematical Imaging and Vision*, vol. 52, pp. 145–172, 2015.

[18] A.-C. Li, S. Vyas, Y.-H. Lin, Y.-Y. Huang, H.-M. Huang, and Y. Luo, "Patch-based u-net model for isotropic quantitative differential phase contrast imaging," *IEEE Transactions on Medical Imaging*, vol. 40, no. 11, pp. 3229–3237, 2021.

Deep Learning-based Intraoperative Video Analysis for Cataract Surgery Instrument Identification

Z. Guo*, Y.H. Chan and N.F. Law
EEE Dept, The Hong Kong Polytechnic University, Hong Kong SAR
E-mail: guozhe9922@gmail.com Tel: +852-65889612

Abstract — **Surgical instrument detection and classification is a critical task for enhancing surgical procedures monitoring, assisting surgical operations, supporting medical education, and enabling the development of intelligent surgical systems. However, there are a few challenges in this domain. The foremost concern is the impact of varying background conditions. Additionally, class imbalance presents another challenge, potentially leading to biased classification results. To solve these challenges, this study proposes a deep learning-based system consisting of two key components: an attention region detection module and a ResNet50 classification model. The attention region detection employs an optical flow-based method to incorporate both temporal and spatial information from the surgical video so that critical attention regions covering surgical instruments are identified. Our experimental results show that the classification accuracy can be improved from 58.7% to 81.9% by using the attention region detection component. To deal with the challenge of class imbalance, we use focal loss and interleaved sampling strategy as solutions. Interleaved sampling uses both the spatial and temporal information of surgical videos to balance the number of samples across different instrument classes, through which some scarce surgical instrument classes are expanded, thus preventing biased learning of the model. And the validation accuracy on the balanced dataset achieves 87.1%. This study demonstrates the effectiveness of deep learning techniques in addressing challenges in cataract surgery video analysis.**

I. INTRODUCTION

Deep learning has made a huge contribution to today's medical field specifically in the fields of image recognition and image processing[1]. Cataract is a disease that is highly prevalent among the elderly. In recent years, deep learning technology has played an irreplaceable role in cataract surgery.

In the field of cataract surgery, the current research status of deep learning technology shows that its application in surgical procedures has potential and indeed has made certain progress[2][3]. These developments are of great significance. On one hand, by analyzing surgical videos and patient health data, deep learning models can not only provide real-time feedback to help doctors adjust treatment strategies and improve surgical success rates[3], but also predict potential complications after surgery[4-6] by analyzing clinical data and imaging, providing doctors with more comprehensive preoperative assessment and patient management strategies.

This study describes a system for identifying and classifying different surgical instruments used during cataract surgery. The core component of our system is a module that can determine the region of attention based on an optical flow method. With attention regions of surgical instruments as the input to the classification model, we can minimize the negative impact of the background portion. This study also proposes an interleaved sampling strategy to address the problem of class imbalance, through which the classification model can fairly learn the features of each class and effectively improve the accuracy.

II. LITERATURE SURVEY

The identification and classification of cataract surgery stages and instruments have extraordinary significance in both medical surgery and artificial intelligence fields[1][7]. For example, Ref. [8] first segments parts of the pupil to extract the attention region and subsequently recognizes surgical instruments with that region using KNN algorithm. However, one of the challenges is that the attention region is not always accurately positioned at the center of the pupil and another is that when some interferences are present in the surgical video because of factors like pupil's outline distortion.

Ref. [9] addresses the challenge of domain shift between different dataset for surgical instrument recognition. Domain shift can be caused by variations in tools, video resolution, and other factors, leading to poor performance of trained models on different dataset. To mitigate this issue, the authors proposed the unsupervised domain adaptation (UDA) method to improve the accuracy in cross-dataset scenarios. However, despite the

improvement of the UDA method, the classification model generalizes poorly when faced with switching dataset. This point is further illustrated in Ref. [10]. The effect of the background portion of the surgical videos on classification accuracy was similarly illustrated in Ref. [8-10]. Therefore, it is important to extract the correct attention region to minimize the negative influences of the background in classification process.

Ref. [11] focuses on phase recognition in cataract surgery videos through deep learning techniques. The authors mention certain limitations, such as the difficulty in accurately predicting categories with infrequent occurrences and scarcity of data. These categories have very low recall, indicating that the model still needs more enhancements to handle less frequent occurrences. Therefore, enhancements are required to increase the size of the dataset, especially for the scarce categories. Zisimopoulos et al. raised the same problem of scarcity of samples of some categories in the dataset in Ref. [2].

III. PROPOSED SURGICAL INSTRUMENT CLASSIFICATION SYSTEM

Since our research is a joint project with the Department of Ophthalmology of the United Hospital in Hong Kong, the 53 cataract surgery video dataset we used in this study were all provided by doctors from the hospital.

A. System Overview

Fig.1 shows the block diagram of the system proposed in this study. The input to the system is intraoperative video clips. The video clip is sampled to obtain video frame pictures of size 1920×1080 each. These images are then used as input to the Attention Region Detection (ARD) Module. The ARD Module detects the area of interest in each frame through an optical flow method and then extracts the attention region in the frame. The size of the attention region is fixed to be 512x512 for all frames. Attention regions are then fed into the ResNet50 model for surgical instrument classification. The ResNet50 model was trained with a dedicated training dataset for classifying one surgical instrument in an image of size 512×512.

To eliminate the negative impact of these factors on the classification process, we introduced the Attention Region Detection Module to confine the condition of the input to the



Fig. 1 Overview block diagram of proposed system.

ResNet50 model. This Attention Region Detection Module uses the optical flow method to extract a 512×512 region that covers only one single surgical instrument within a raw 1920×1080 video frame image. Since the input is now well-conditioned and no down-sampling was performed, the spatial features of the surgical instrument are well preserved during the preparation of the input. Consequently, the system is able to improve the classification accuracy remarkably.

B. Attention Region Detection

In the proposed system, an attention region means the region that the exploited deep-learning model should put its focus on. The model should ignore other irrelevant regions when processing an image. The function of the ARD module is to intercept partial images that contain surgical instruments from the original video clips.

We assume that the surgical instruments appearing in a video frame are the foreground of the frame and all other parts are the background. Fig. 2 provides an example of separating the foreground and the background of a video frame. It is obvious that the background includes the patient's pupils and the parts around the eyes. Factors such as lighting and shooting angles bring a lot of uncertainty to the background. The foreground is actually our concern and it should be the key object in the attention region. Without loss of generality, we can assume that the background does not change too much within a very short time duration during the same surgical phase, while the foreground can have more variability in its position. By taking advantage of the characteristic that the background does not change in a short period, we use the optical flow method to determine the position of the surgical instrument, which is also the position of the foreground. More specifically, the optical flow method compares the grayscale values of the current and the previous video frames to obtain the point of maximum change. This point is considered as the center of the attention region of the current frame. Attention regions generated by the ARD module will be used as input to the classification and recognition model. If the target instrument is successfully captured in the attention region, then we can minimize the impact of the uncertain factors in the background on the classification results and make the classification more robust to the background in the video.



Fig. 2 An example of separating the background and the foreground of a video frame.

To better separate the foreground and the background, it should be noted that an appropriate sampling rate needs to be used. If the sampling interval is too short, the "foreground" may not move significantly, causing it cannot be successfully extracted. If the sampling interval is too long, the "background" may also undergo drastic changes during this period.

C. Methodologies

i. Farneback Method

In this experiment, the signal is often two-dimensional picture information. Optical flow field is a term used to describe the motion vector field of pixels reflected by moving objects in three-dimensional space in a two-dimensional image. To better locate the position of surgical instruments in video frames without being affected by the background, we used the Farneback[12] method. Naturally, for each pair of consecutive images in the dataset, we can compare the grayscale values between adjacent frames to find the point with the maximum difference in grayscale values. Subsequently, this point serves as the center for cropping the original image into a 512×512 pixel-sized picture. In this process, by choosing the appropriate sampling rate, adjusting the parameters of the optical flow method, etc., it is possible to make the cropped image contain a localized image of the surgical instrument.

ii. Loss Function

Due to the significant differences in the duration of different surgical phases, the appearance times of different surgical instruments in the video also vary greatly. Considering the problem of class imbalance, we chose to use focal loss[13] as the loss function. Focal loss allows the network to focus more on difficult-to-classify samples, effectively improving the model's performance on minority samples.

D. Results

i. Optical Flow

Taking a set of adjacent pictures in Fig. 3 as an example, we calculate the optical flow of them. Then we use the Farneback method to determine the maximum change point between two adjacent frames. Fig. 4 shows the energy map. The scale bar on the right shows the energy value represented by each color in the figure. The higher the pixel color value, the greater the energy, indicating a greater change as determined by the optical flow method.

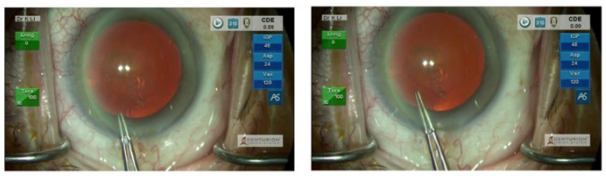


Fig. 3 A set of adjacent pictures in the dataset.

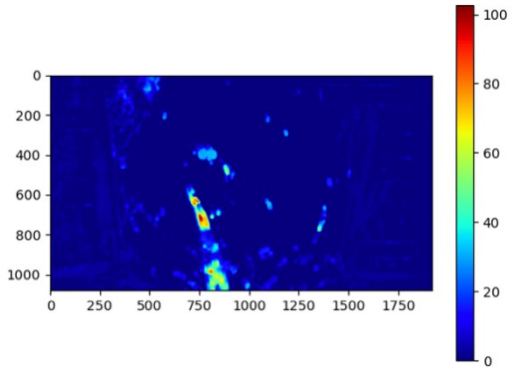


Fig. 4 Energy map obtained based on optical flow.

Similarly, according to the optical flow, the arrow map of the optical flow changes can be obtained as shown in Fig. 5. The direction pointed by the green arrow is the direction of motion of the pixel, and the length of the arrow represents the size of the motion vector of the pixel.

It can be seen from the energy map and arrow map that the Farneback method detects changes in the position of the surgical instruments in the adjacent legends mentioned above, and the changes in the background are much smaller than the changes in the surgical instruments.

ii. Attention Region Detection

After detecting the maximum change point in two adjacent pictures according to the Farneback method, a square of size 512×512 is cut out from the picture with this point as the center. Fig. 6 shows the square area intercepted with the maximum change point as the center.

iii. Model Performance

To mitigate the influence of uncertain factors in the “background”, it is necessary to extract the attention regions that contain surgical instruments separately from video frames.

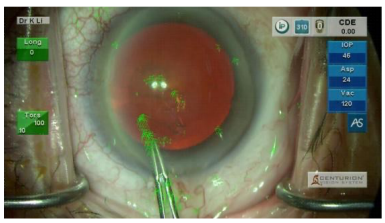


Fig. 5 Arrow map obtained based on optical flow.

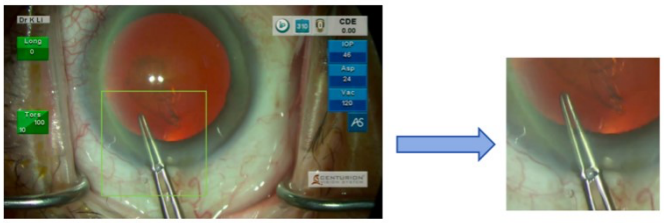


Fig. 6 Attention region obtained based on maximum change point.

Use pictures extracted by ARD module as input to ResNet50. In the ResNet50 model, in addition to using Focal Loss as the loss function to reduce the negative impact of uneven sample distribution, we also use dropout to prevent over-fitting. After the experiment, we got the results in Fig. 7. According to Fig. 7, the accuracy of the best model reached 81.9%, but there is a difference between the accuracy of the validation set and the accuracy of the training set. In the line chart of various evaluation indicators, we can see that except for the early stage of training, the performance of the model in each epoch slowly improves without major fluctuations. This shows that the model maintains good stability while ensuring good training accuracy.

E. Ablation Study

Before verifying the model performance, we performed ablation experiments. First, the performance of the system with only the ResNet50 model is tested, in which Cross Entropy is used as the loss function, and other parameters and settings remain the same.

After sampling the video frames at a sampling rate of 0.2 seconds per frame, they are directly put into the ResNet50. Based on the above steps, the results are shown in Fig.8. The best model achieved an accuracy of 58.1% on the validation set. It can be found that the accuracy on the training set reaches 100%, while the accuracy on the validation set is only 50%-60% on average. We believe there may be three reasons. First, the over-fitting phenomenon is serious during the training process. Second, because the data of the training set and validation set are sampled from different surgical videos, differences between different videos lead to differences in the model performance on the training and validation set. The third is that the imbalance of sample distribution causes the accuracy of the validation set in this experiment to be much lower than the accuracy of the training set. Therefore, we replaced the originally used Cross Entropy loss function with Focal Loss to reduce the impact of uneven sample distribution on classification accuracy. Fig. 9 shows the experiment result.

Compared with Fig. 8, after using Focal Loss, the best accuracy of the validation set increased from 58.1% to 58.7%. Overall, there is no significant improvement in model performance. Therefore, after the ablation experiments, we focused on solving three issues that would arise from speculation. The last chapter has proven that the "background" difference of different videos is too large and the over-fitting phenomenon does have a negative impact on the recognition of the model.

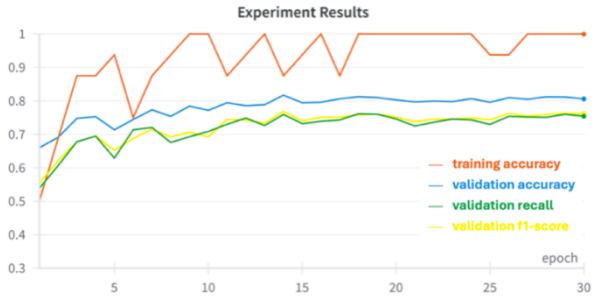


Fig. 7 Training accuracy, validation accuracy, validation recall and f1score.

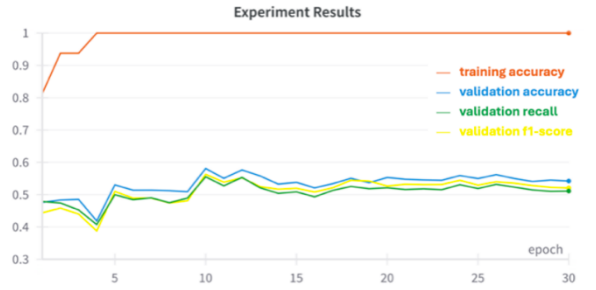


Fig. 8 Training accuracy, validation accuracy, validation recall and f1score.

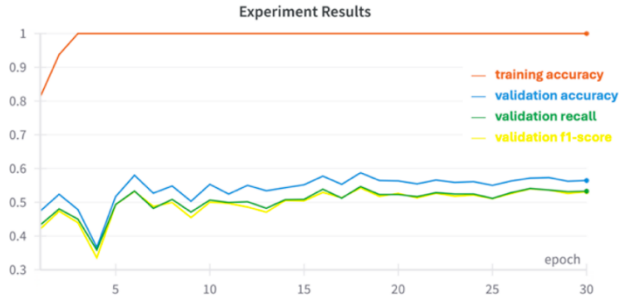


Fig. 9 Training accuracy, validation accuracy, validation recall and f1-score.

IV. PRODUCTION OF BALANCED TRAINING DATASET

A. Dataset

In order to avoid the problem of model "cheating", which occurs when frames from the same video are present in both the training and testing dataset, leading to the model performing exceptionally well, we believe that the images in the training set, validation set, and testing set should be generated from different videos. Therefore, in order to have a fairly unbiased dataset to train the model and test the performance of the model, a preliminary division was made: among these 53 videos, 30 were used for the training set, 15 were used for the validation set, and 8 were used for the testing set, and the videos used between them did not overlap with each other.

B. Challenge

After detailed annotation and sampling of the dataset, the sample number distribution of all classes is shown in Fig. 10. It can be seen that the problem of unbalanced sample

distribution is very serious. This may cause the model to overlearn features for classes with a large number of samples and not be able to fully and thoroughly learn features for scarce samples.

C. *Interleaved Sampling*

In order to verify whether the balance of sample data of each class in the dataset will affect the classification accuracy, we use the interleaved sampling operation to expand the scarce categories of surgical instruments, so that the number of samples of each class is balanced. Fig. 11 shows the principle of interleaved sampling. Suppose there are N consecutive video frames, and we take one frame every 6 frames. Then we can take a series of frame pictures every 6 frames such as Frame 1, Frame 7, Frame 13, Frame 19, Frame 25, etc. as one set. Similarly, we can also use any one of the second to the sixth frames as the starting frames to extract different video frame groups. Each group can be used as input to the Attention Region Detection Module to find the maximum change point between two adjacent frames to provide input data for the classification model. And the video frames in frame groups extracted using the picture frame before the seventh frame as the starting frame do not overlap each other.

The significance of interleaved sampling is obvious. By utilizing both temporal and spatial information, the amount of data can be increased almost exponentially, while the extracted data is not repeated. The temporal information of video frames provides a basis for the arrangement and extraction of video frames, while the spatial information not only allows all video frames to be fully utilized but also allows video frames to be divided into groups and extracted without being reused.

D. *Experiments*

In order to explore whether the problem of unbalanced sample distribution has a negative impact on the model, we performed an interleaved sampling operation on the dataset used in the previous two parts, which can expand the dataset and make the distribution of samples of each class balanced. Fig. 12 is the sample distribution of each class after interleaved sampling operation.

Taking this balanced dataset as the input of the system, and keeping other conditions all the same, we obtained the following result in Fig. 13, in which we can see the accuracy of the best model on the verification set reached 87.1%, and all indicators tended to be stable throughout the training process. This shows that a balanced distribution of samples in each class contributes positively to the accuracy of the classification task. After ablation study and experiments conducted with ARD module and the balanced dataset, we got the results in Table. 1, in which we can see that the effectiveness of ARD module in determining attention region. And it also proves that a balanced dataset has a benign effect on model learning.

Table. 2 shows the performance comparison between our proposed model and the ResNet50 model. The compared indicators include accuracy, precision, recall, and f1-score. Their inputs are identical video clips. It should be noted that they perform different preprocessing on video clips. Our model uses the Farneback method to find the maximum change point and then crops the image to a size of 512×512. ResNet50 model samples video clips at a sampling rate of 5 frames per second, and then perform resize operations to improve the efficiency of training and inference. It resizes the video frame to 225×225 and then center-crop it to 224×224.

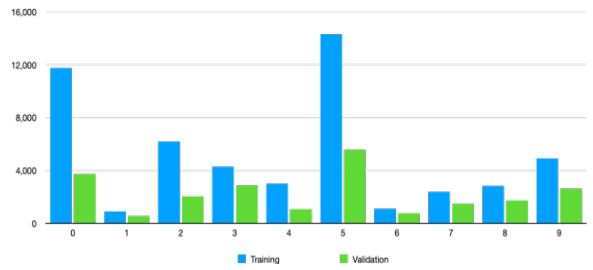


Fig 10. Quantity distributions of different classes.

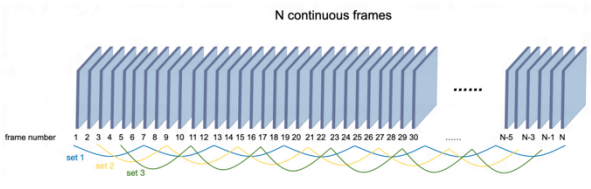


Fig. 11 N continuous frames using interleaved sampling.

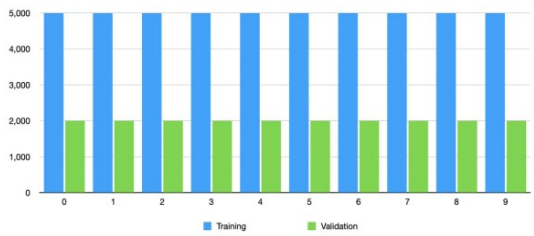


Fig. 12 Quantity distribution of different classes after interleaved sampling.

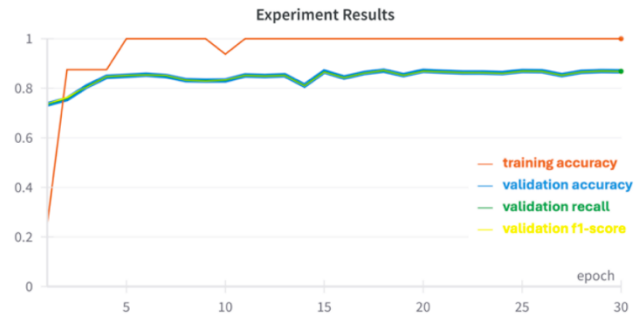


Fig. 13 Training accuracy, validation accuracy, validation recall and f1-score.

Table. 1 The validation accuracy of different methods used.

	Version1	Version2	Version3	Version4
Focal Loss	×	√	√	√
Dropout	×	×	√	√
Attention Region Detection	×	×	√	√
Interleaved Sampling	×	×	×	√
Accuracy	58.1%	58.7%	81.9%	87.1%

Table. 2 Compare with the ResNet50 model. The values used for precision, recall and F1-score are all macro averages.

Method	Accuracy	Precision	Recall	F1-score
ResNet50	0.594	0.642	0.594	0.588
Ours	**0.871**	**0.872**	**0.868**	**0.868**

V. CONCLUSIONS

This study addresses the key challenges associated with surgical instrument detection and classification in cataract surgery videos. The proposed deep learning-based system contains two key components, namely an Attention Region Detection Module and a ResNet50 classification model. The former incorporates both temporal and spatial information to extract attention regions while the latter focuses on these attention regions and performs surgery instrument classification. The Attention Region Detection Module based on the Farneback method plays an indispensable and significant role in promoting this research. It can well extract critical regions of surgical instruments from video clips, which minimizes the negative impact of uncertainties in background portion on the model's ability to accurately identify surgical instruments. As for class imbalance issue, the interleaved sampling strategy effectively utilizes temporal and spatial information to better solve the problem of uneven distribution of sample numbers. Class balancing can significantly improve the generalization ability and fairness of the model, ensuring that the model is not biased toward any specific class.

REFERENCES

[1] Alzubaidi, L. et al. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1), 53.

[2] Zisimopoulos, O. et al. (2018). DeepPhase: Surgical Phase Recognition in CATARACTS Videos. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018.*

[3] Lee, M. et al. (2023). Real-time estimation of the remaining surgery duration for cataract surgery using deep convolutional neural networks and long short-term memory. Translational Vision Science & Technology, 12(2), 24.

[4] Quellec, G. et al. (2014). Real-Time Segmentation and Recognition of Surgical Tasks in Cataract Surgery Videos. IEEE Transactions on Medical Imaging, 33(12), 2352-2360.

[5] Cheng, B. et al. (2020). Deep Learning-Based Keratoconus Screening Using Corneal Topographic Maps. Investigative Ophthalmology & Visual Science, 61(10), 45.

[6] Abbas, A. et al. (2019). Machine Learning for Prediction of Postoperative Complications after HepatoBiliary and Pancreatic Surgery. Annals of Surgery, 270(1), 117125.

[7] Ghamsarian, N. et al. (2021). Relevance Detection in Cataract Surgery Videos by Spatio- Temporal Action Localization. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, (pp. 10720-10727). IEEE.

[8] Palepu, V. et al. (2019). Surgical tools recognition and pupil segmentation for cataract surgical process modeling. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (pp. 1730-1734). IEEE.

[9] Pasa, F. et al. (2020). Cross-dataset adaptation for instrument classification in cataract surgery videos. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (pp. 620-628). Springer.

[10] Sokolova, N. et al. (2020). Evaluating the Generalization Performance of Instrument Classification in Cataract Surgery Videos. In *Proceedings of the 26th International Conference on Multimedia Modeling (MMM2020)* (pp. 626-636). Springer.

[11] Yeh, H.-H. et al. (2021). PhacoTrainer: Deep Learning for Activity Recognition in Cataract Surgical Videos. *Investigative Ophthalmology & Visual Science*, 62(8), 583.

[12] Farneback, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. In *Scandinavian Conference on Image Analysis (SCIA 2003)* (pp. 363-370).

[13] He, K. et al. (2017). Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 29802988).

Disentangling Speaker Representations from Intuitive Prosodic Features for Speaker-Adaptative and Prosody-Controllable Speech Synthesis

Pengyu Cheng*, Zhenhua Ling†, Meng Meng*, Yujun Wang*
* Xiaomi Inc., Beijing, China
E-mail: {chengpengyu, mengmeng, wangyujun}@xiaomi.com
† NERC-SLIP, University of Science and Technology of China, Hefei, China
E-mail: zhling@ustc.edu.cn

***Abstract*—In this paper, we propose a method of speaker-adaptative and prosody-controllable speech synthesis with a disentanglement between intuitive prosodic features and speaker representations. In this method, the intuitive prosodic features include utterance-level pitch, pitch range, speak rate and energy. A residual speaker information encoder with a set of adversarial classifiers is designed to extract the speaker characteristics that can't be described by these intuitive prosodic features. Further, the outputs of the residual speaker information encoder are concatenated with intuitive prosodic features to obtain complete speaker representations for acoustic feature prediction. Experimental results have demonstrated that our proposed method can synthesize speech with better naturalness and higher prosody controllability than its counterpart without the disentanglement.**

***Index Terms*—speech synthesis, speaker adaptation, speaker representation, intuitive prosodic features, disentanglement**

I. Introduction

Currently, state-of-the-art text-to-speech (TTS) systems adopt sequence-to-sequence (seq2seq) acoustic models [1]–[3] to convert input text sequences into Mel-scale spectrograms, and then utilize neural network vocoders [4], [5] to reconstruct audio waveforms from the generated Mel-spectra. On the basis of speaker-dependent model training which usually requires a large amount of training data from the target speaker, many studies have paid their attentions on the speaker adaptation task, which aims to synthesize high-quality voice when the data amount of the target speaker is limited. Most of these studies build a pre-trained TTS model with a multi-speaker dataset at first and then use the recordings from the target speaker to update the whole or part of pre-trained model's parameters. According to the speaker representation method used in acoustic modeling, these speaker adaptation methods can be roughly divided into two main categories, i.e., speaker-encoding-based ones [6]–[11] and speaker-embedding-based ones [12]–[16].

In speaker-encoding-based adaptation methods [6]–[11], a neural speaker encoder is designed to model each sentence in the corpus. Such speaker encoder is usually pre-trained on a large-scale multi-speaker corpus through a speaker verification task and is fixed while training the TTS model. In speaker-embedding-based methods [12]–[16], a speaker embedding vector is designed to represent the global voice characteristics of each speaker. At the training stage, these speaker embedding vectors are usually optimized together with TTS model, and the most widely used strategy is the look-up table. Considering prosody characteristics influence the subjective perception of synthetic speech significantly, our previous work [17] proposed a method of multi-speaker training and speaker adaptation with intuitive prosodic features for seq2seq speech synthesis. The intuitive prosodic features consisted of utterance-level pitch, pitch range, speak rate, and energy. Experimental results demonstrated that this method can effectively improve the subjective similarity of speaker adaptation under both speaker-encoding-based and speaker-embedding-based.

Another advantage of utilizing intuitive prosodic features is that it enables us to control the prosody characteristics of synthetic voice conveniently. This is useful for improving the expressiveness of speech synthesis and creating new voices for avatars. For example, Tuomo et al. [18] proposed a method of prosody modeling and controlling with intuitive prosodic features for single speaker TTS. Morrison et al. [19] presented a deep autoregressive model that supported controllable, context-aware fundamental frequency (F_0) generation. Valle et al. [20] designed a multi-speaker TTS framework combining explicit prosodic variables (F_0 and voicing decision) and latent space modeling. Comparing with the latent prosody representations learned from speech spectra [21], [22], intuitive prosodic features exclude segmental characteristics and have better interpretability at each dimension. One issue with current approaches to integrate intuitive prosodic features into multi-speaker acoustic modeling [17], [20] is the entanglement between speaker representations and intuitive prosodic features. The speaker representations either derived by a speaker encoder or learned as speaker embeddings may contain the prosody characteristics of reference utterances. Thus, they may conflict with the modified intuitive prosodic features at the synthesis stage, which constrains the performance of prosody control.

Therefore, this paper proposes a method of disentangling speaker representations from intuitive prosodic features for

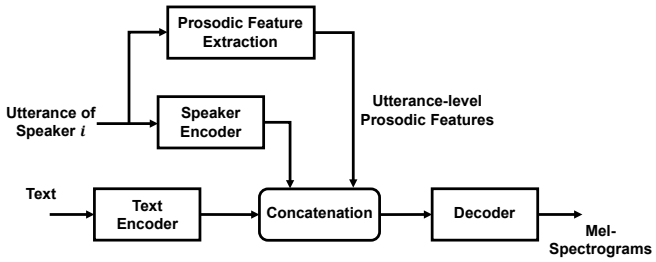


Fig. 1. The architecture of our baseline model [17].

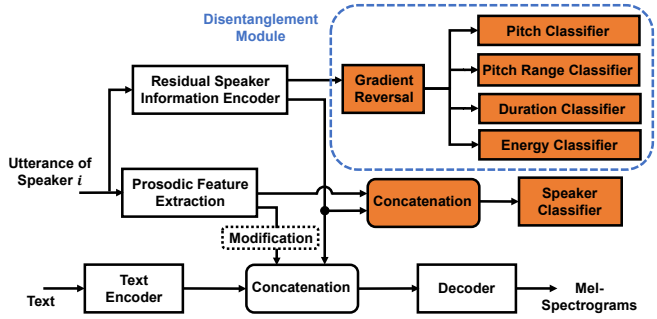


Fig. 2. The architecture of our proposed model, the orange parts are only conducted at the training stage.

speaker-adaptative and prosody-controllable speech synthesis. The four-dimensional utterance-level intuitive prosodic features used in our previous study [17] are adopted here. A residual speaker information encoder is designed to extract the speaker characteristics that are not represented by intuitive prosodic features. A set of prosody classifiers with adversarial losses were applied to achieve the disentanglement. The outputs of the residual speaker information encoder are concatenated with intuitive prosodic features to obtain complete speaker representations, which are constrained by a speaker classifier and are sent into the decoder for acoustic feature prediction. Our experimental results on the AISHELL-3 [23] dataset demonstrated that the proposed method achieved higher naturalness of synthetic speech and better prosody controllability over all prosodic variables than the baseline model [17].

II. Baseline with Intuitive Prosodic Features

The baseline model used in this paper follows our previous work [17]. Its architecture is shown in Fig. 1. It adopts Tacotron2 [1] as its backbone and consists of four parts: a prosodic feature extraction module, a speaker encoder, a text encoder and a decoder. In this model, the extracted prosodic features are all at utterance-level, and the detailed procedure of prosodic feature extraction can be found in [17]. The speaker encoder has a ResNet-based architecture, following the structure proposed in [24]. The parameters of the speaker encoder are pre-trained by a speaker verification (SV) task on a large-scale multi-speaker corpus and are fixed while training the speech synthesis model. The text encoder and the decoder are pre-trained with a multi-speaker dataset and are then adapted to the target speaker. For each training utterance, the speaker representations extracted using the speaker encoder are concatenated with the utterance-level intuitive prosodic features and are then sent into the decoder for predicting the Mel-spectrograms of this utterance. At the synthesis stage, the average speaker representations and intuitive prosodic features of all adaptation utterances from the target speaker are employed. In our previous experiments, this baseline model outperformed its counterpart without intuitive prosodic features (named *Baseline w/o IPF* in this paper) on both similarity and naturalness after speaker adaptation [17].

III. Proposed Methods

A. Overall Structure

The overall structure of our proposed model is illustrated in Fig. 2 which consists of five parts: a prosodic feature extraction module, a disentanglement module, a residual speaker information encoder, a text encoder and a decoder .

The prosodic feature extraction module, the text encoder and the decoder have exactly the same structure as the ones in the baseline model [17]. The residual speaker information encoder receives the 80-dim Mel-spectra of the input utterance and output a disentangled speaker vector which is expected to be irrelevant to the extracted intuitive prosodic features of this utterance. To achieve this goal, a set of prosody classifiers with a gradient reversal layer [25] are designed in the disentanglement module in Fig. 2. To guarantee the completeness of speaker representation, the disentangled speaker vector and ground-truth (GT) intuitive prosodic features are concatenated and then sent into a speaker classifier with a cross-entropy (CE) loss. Like the baseline model, the concatenated speaker representation together with the output of the text encoder are fed into the decoder for predicting corresponding Mel-spectra.

At the inference stage, an utterance randomly selected from the target speaker's adaptation dataset is employed as the input of the residual speaker information encoder, while the average intuitive prosodic features of the speaker's adaptation data is adopted, which can be further modified to control the prosody characteristics of synthetic speech.

B. Residual Speaker Information Encoder

We first tried the ResNet-based [24] structure commonly used in SV tasks to build the residual speaker information encoder, but it did not work well. This may be due to that the structure designed for SV tasks learns to extract prosody-related representations which can help to distinguish between different speakers, resulting in that the output speaker vectors can't be successfully disentangled from intuitive prosodic features by adversarial training.

Therefore, a simpler architecture is adopted in our proposed residual speaker information encoder, which consists of 6 convolutional layers and a bidirectional long short-term memory (BiLSTM) layer. Each convolutional layer is followed by a

batch normalization operation. A post linear layer is attached to the BiLSTM layer as the output layer. Finally a L2-norm activation is used for the output hidden variables to accelerate the convergence of model training.

C. Disentanglement Module

In order to eliminate the information of intuitive prosodic features from speaker representations, a disentanglement module based on domain adversarial training [25] is applied in this paper. Four prosodic classifiers corresponding to the four intuitive prosodic features used in this paper are designed as shown in Fig. 2. A gradient reversal layer is placed between the residual speaker information encoder and the four prosodic classifiers to invert the gradients calculated by the prosody classification loss.

The four prosodic classifiers share the same structure, including a dropout operation with pre-dense and post-dense layers. Since the previously extracted intuitive prosodic features are continuous ones, they are discretized first to get classification labels. For each prosodic feature, its minimum and maximum values in the training set are normalized to $[0, 1]$ respectively. The normalized $[0, 1]$ range is then divided into 256 intervals equally. Thus, each continuous value of prosodic features can be quantized into a categorical label for training the disentanglement module.

D. Model Training Strategy

Three groups of loss functions are used for model training. The first group is the conventional decoder loss, which consists of the original Tacotron2 [1] loss function of Mel-spectrum prediction and a guided attention loss [26] to help model learn alignment faster. The second group contains the adversarial losses given by the prosodic classifiers in the disentanglement module. A cross-entropy (CE) loss is calculated for each classifier since the prosodic features are quantized to discrete labels. The third group is the cross-entropy (CE) loss calculated by the speaker classifier in order to guarantee the completeness of speaker representation.

The model parameters are first pre-trained with a multi-speaker dataset and then adapted to the target speaker. At the pre-training stage, all three groups of losses are applied with equal weights to train all model parameters simultaneously. At the adaptation stage, the text encoder is fixed. The CE loss of the speaker classifier is removed since only the data of the target speaker is used at this stage. The other two groups of losses are utilized to fine-tune other model parameters.

IV. Experiments

A. Datasets

The AISHELL-3 [23] corpus was adopted in our experiments, which contains a total of 88,305 utterances spoken by 218 Mandarin Chinese speakers. Considering that some utterances in this dataset had no phrase boundaries in their corresponding transcripts, these samples were abandoned. The remaining part had a total of 63,263 utterances from 174 speakers, including 31 male and 143 female speakers. 4 male and 4 female speakers were randomly selected as adaptation target speakers. For each target speaker, 80, 10 and 20 utterances were randomly selected as the training set, the development set and the test set for speaker adaptation. The recordings of the remaining 166 speakers were used as the multi-speaker dataset for pre-training.

TABLE I
Objective evaluation results of three systems on the test sets of female and male speakers.

Models	MCD (dB)		F_0 RMSE (Hz)	
	Female	Male	Female	Male
Baseline w/o IPF	3.85	5.36	47.08	27.79
Baseline	3.76	**5.10**	44.07	26.98
Proposed	**3.57**	5.21	**42.23**	**26.14**

B. Experimental Configurations

The original recordings were down-sampled to 16kHz for training the TTS model. The 80-dimensional Mel-spectrograms were computed with 50 ms frame length and 12.5 ms shift.

To verify the effectiveness of our proposed method, two models were constructed for comparison in addition to the proposed model. One was the baseline model introduced in Section II. Another was the *Baseline w/o IPF* model mentioned at the end of Section II. We trained each model on the multi-speaker dataset for 140k steps with a batch size of 40. The pre-trained model was then fine-tuned for fixed 800 steps for each unseen speaker with a batch size of 20. All the three models shared a same Parallel WaveGAN [27] vocoder to generate speech from predicted Mel-spectra. We simply trained the vocoder on the complete AISHELL-3 dataset since this paper focuses on acoustic modeling instead of vocoding techniques.

C. Evaluation on Speaker Adaptation

Objective and subjective evaluations were conducted to compare the performance of different models on speaker adaptation.

We used 25-dimensional Mel-cepstral coefficients (MCCs) and F_0 to calculate Mel-cepstrum distortion (MCD) and root mean square error of F_0 (F_0 RMSE) as objective evaluation metrics. Dynamic time warping (DTW) [28] was conducted based on MCCs to align the sequences of MCCs and F_0 extracted from synthetic speech with the ones from natural recordings for calculating distortions. Objective metrics were calculated on the test sets of 8 target speakers with 20 utterances per speaker.

The evaluation results are reported in Table I. We can see that the overall objective performances of the baseline model and our proposed model were similar. Our proposed method outperformed baseline on female speakers. For male speakers, our proposed method achieved lower F_0 RMSE than baseline, but higher MCD. Both baseline and proposed

TABLE II
RESULTS OF ABX PREFERENCE TESTS (%) BETWEEN TWO SYSTEMS ON THE TEST SETS OF FOUR SPEAKERS, WHERE N/P DENOTES "NO PREFERENCE" AND p MEANS THE p-VALUE OF PAIRED T-TEST.

	Baseline	Proposed	N/P	p
naturalness	24.77	**46.36**	28.87	0.0002
similarity	19.09	22.05	58.86	0.23

TABLE III
MEAN OPINION SCORES ON NATURALNESS (NAT.) AND SIMILARITY (SIM.) WITH 95% CONFIDENCE INTERVALS OF SPEECH SYNTHESIZED BY TWO MODELS WITH PROSODY CONTROL.

Prosody Variables		Baseline	Proposed
Pitch	Nat.	3.115 $\pm$ 0.136	**3.475** $\pm$ **0.114**
	Sim.	3.920 $\pm$ 0.107	3.955 $\pm$ 0.100
Pitch Range	Nat.	3.300 $\pm$ 0.118	**3.645** $\pm$ **0.105**
	Sim.	3.765 $\pm$ 0.104	**3.975** $\pm$ **0.088**
Speak Rate	Nat.	3.005 $\pm$ 0.130	**3.665** $\pm$ **0.100**
	Sim.	3.575 $\pm$ 0.116	**3.995** $\pm$ **0.094**
Energy	Nat.	3.270 $\pm$ 0.125	**3.780** $\pm$ **0.100**
	Sim.	3.825 $\pm$ 0.086	**4.025** $\pm$ **0.087**

models outperformed Baseline w/o IPF, which demonstrated the effectiveness of integrating intuitive prosodic features.

The subjective evaluation was conducted following the experimental configurations in our previous work [17]. 4 target speakers (2 female and 2 male) were randomly selected to control the scale of listening tests. For each target speaker, an ABX preference test was conducted to compare the baseline model with our proposed model. Each test had 10 pairs of utterances synthesized by these two models respectively. At least 11 native listeners participated in each test, and the listeners were asked to give their preference opinion for each pair (which one in the pair was better or there was no preference) on both naturalness and similarity. Paired t-test was also carried out to examine the significance of the preference between two models.

Table II shows the results of ABX preference tests. We can see that the proposed model achieved significantly ($p = 0.0002$) better naturalness of synthetic speech than the baseline. This demonstrates the effectiveness of our proposed disentanglement module, which helped to eliminated the mismatch between speaker representation and intuitive prosodic features at the inference stage. Comparing the similarity of speech synthesized by these two models, the preference difference was insignificant ($p = 0.23$)[1].

D. Evaluation on Prosody Control

To test whether our proposed model can effectively disentangle prosodic information from speaker representations, a prosody control experiment was carried out. Two speakers (one male and one female) were randomly selected from the 8 target speakers and we conducted speaker adaptation for both speakers following the baseline and proposed methods

[1]Samples can be found at https://chengpy22.github.io/DSRIPF_demo/.

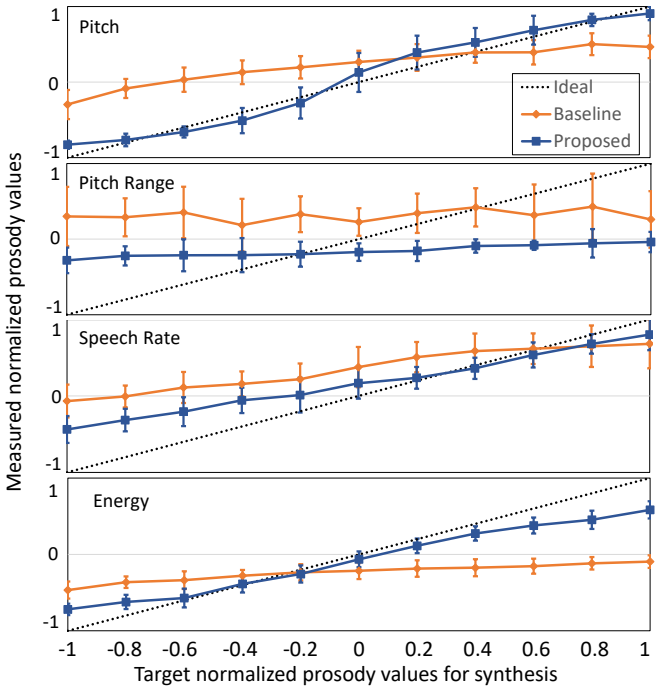


Fig. 3. The mapping curves between measured and target prosody values for different intuitive prosodic features and models on the female target speaker.

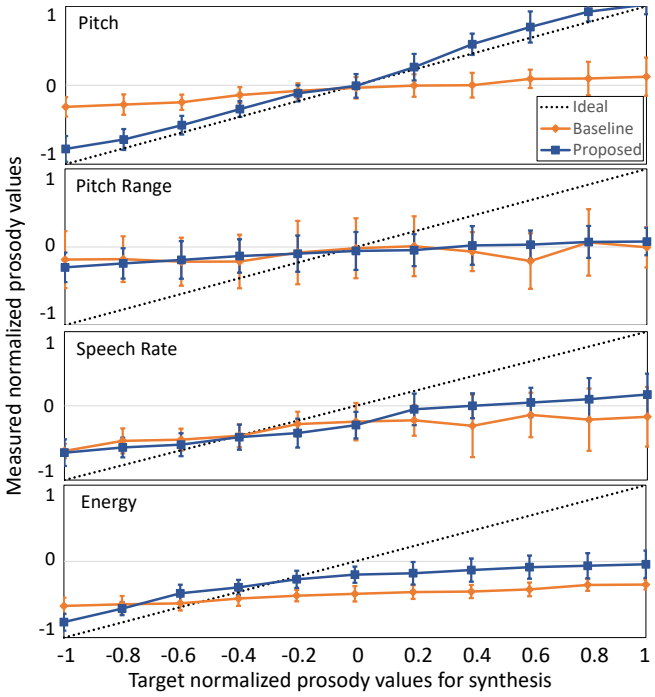


Fig. 4. The mapping curves between measured and target prosody values for different intuitive prosodic features and models on the male target speaker.

respectively. For each intuitive prosodic feature, its minimum and maximum values in the multi-speaker pre-training dataset were calculated after discarding the highest and lowest 10% samples. Then, we mapped the range between the minimum and maximum values to [-1, 1] to get normalized prosody values. In order to control each intuitive prosodic feature at the

synthesis stage, a group of target normalized prosody values were sampled with intervals of 0.2 in [-1, 1] and were then denormalized to the original values of each prosodic feature for decoding the utterances in the test set. For each target normalized prosody value, intuitive prosodic features were extracted from synthetic utterances, and were then normalized to get the measured normalized prosody values. The mean and standard deviation of the measured prosody values were calculated from all test utterances. Finally, the mapping curves between measured and target prosody values were drawn to compare the prosody controllability of different models as shown in Fig. 3 and Fig. 4. The ideal curve should be a diagonal one as shown by the dotted lines in these figures.

The results show that our proposed method can better reflect the target prosody values in synthetic speech than the baseline method in the prosody control experiment. For pitch, the proposed method presented almost ideal correlation between target and measured prosody values on both speakers. The performance of controlling pitch range was the worst among the four prosodic features for both models. The reason may be that the reading style of most speakers in the AISHELL-3 dataset was rather bland, and the models failed to learn the correlations between pitch ranges and Mel-spectra. This issue is worth further investigation in the future.

We also evaluated subjective performance of prosody control for these two models. For each intuitive prosodic feature, we selected ten pairs of synthetic utterances with the same measured prosody values from the samples used to draw Fig. 3 and Fig. 4. The two utterances in each pair had the same contents and were synthesized by the two models respectively. Then, the mean opinion scores (MOS) on naturalness and similarity of these selected samples were evaluated by a listening test. 11 native listeners took part in the test to give a 5-scale opinion score (5:excellent, 4: good, 3: fair, 2: poor, 1: bad) on both naturalness and similarity for each sample.

The subjective evaluation results are shown in Table III. We can see that for each prosodic feature, the proposed model outperformed the baseline model on both naturalness and similarity, except the similarity when controlling pitch. This further confirms the effectiveness of our proposed method on prosody control by eliminating the possible conflict between speaker representations and modified prosodic features.

V. Conclusions

In this paper, we have proposed a multi-speaker TTS and speaker adaptation framework with disentanglement between intuitive prosodic features and speaker representations. Experiments on speaker adaptation demonstrated that our method can achieve better naturalness of synthetic speech than the baseline model with intuitive prosodic features but without disentanglement. The objective and subjective evaluation results of the prosody control experiment further proved that our method outperformed the baseline model on prosody controllability. Our future work includes integrating more intuitive acoustic features, such as voice quality features, into current framework for adaptive and controllable speech synthesis.

References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018.

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[3] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6706–6713, 2019.

[4] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[5] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[6] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain with one-shot speaker adaptation," *arXiv preprint arXiv:1803.10525*, 2018.

[7] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," in *International Conference on Machine Learning*, pp. 3683–3691, PMLR, 2018.

[8] Q. Hu, E. Marchi, D. Winarsky, Y. Stylianou, D. Naik, and S. Kajarekar, "Neural text-to-speech adaptation from low quality public recordings," in *Speech Synthesis Workshop*, vol. 10, 2019.

[9] Z. Kons, S. Shechtman, A. Sorin, C. Rabinovitz, and R. Hoory, "High quality, lightweight and adaptable TTS using LPCNet," in *Interspeech*, pp. 176–180, 2019.

[10] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6184–6188, IEEE, 2020.

[11] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in Neural Information Processing Systems 31*, pp. 4485–4495, 2018.

[12] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," *arXiv preprint arXiv:1802.06006*, 2018.

[13] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie, *et al.*, "Sample efficient adaptive text-to-speech," *arXiv preprint arXiv:1809.10460*, 2018.

[14] Y. Huang, L. He, W. Wei, W. Gale, J. Li, and Y. Gong, "Using Personalized Speech Synthesis and Neural Language Generator for Rapid Speaker Adaptation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7399–7403, 2020.

[15] Y. Deng, L. He, and F. Soong, "Modeling multi-speaker latent space to improve neural tts: Quick enrolling new speaker and enhancing premium voice," *arXiv preprint arXiv:1812.05253*, 2018.

[16] Y. Zheng, X. Li, and L. Lu, "Investigation of fast and efficient methods for multi-speaker modeling and speaker adaptation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6618–6622, IEEE, 2021.

[17] P. Cheng and Z. Ling, "Speaker adaption with intuitive prosodic features for statistical parametric speech synthesis," in *Proceedings of the 6th International Conference on Digital Signal Processing*, pp. 187–193, 2022.

[18] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," in *Interspeech*, 2020.

[19] M. Morrison, Z. Jin, J. Salamon, N. J. Bryan, and G. J. Mysore, "Controllable neural prosody synthesis," *Proc. Interspeech 2020*, pp. 4437–4441, 2020.

[20] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6189–6193, IEEE, 2020.

[21] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, pp. 5180–5189, PMLR, 2018.

[22] C. Lu, X. Wen, R. Liu, and X. Chen, "Multi-speaker emotional speech synthesis with fine-grained prosody modeling," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5729–5733, 2021.

[23] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus," *Proc. Interspeech 2021*, pp. 2756–2760, 2021.

[24] W. Cai, J. Chen, J. Zhang, and M. Li, "On-the-Fly Data Loader and Utterance-Level Aggregation for Speaker and Language Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1038–1051, 2020.

[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[26] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4784–4788, 2018.

[27] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, IEEE, 2020.

[28] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1, pp. 125–128, IEEE, 1993.

A Joint Graph Signal and Laplacian Denoising Network Inspired by Majorization-Minimization

Zepeng Zhang* and Ziping Zhao†
* École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
Email: zepeng.zhang@epfl.ch
† ShanghaiTech University, Shanghai, China
Email: zipingzhao@shanghaitech.edu.cn

***Abstract*—Graph neural network (GNN) models have presented astonishing achievements in various application fields. However, they are shown to be vulnerable to adversarial attacks on graph structure and unnoticeable perturbations on the graph structure can cause significant performance drops in GNN models. Based on recent studies that reveal a class of GNN models is performing graph signal denoising (GSD), in this paper, we design a novel robust GNN model from a joint graph signal and Laplacian denoising problem (GSLD), named GSLDN. Specifically, GSLDN is built based on a block majorization-minimization algorithm for solving the GSLD problem. Designed in such a principled way, GSLDN is endowed with the power to fight against adversarial attacks on graph structure. Experiment results demonstrate the effectiveness of GSLDN.**

I. Introduction

Graph neural networks (GNNs) have shown great power in learning representations for graph-structured data and have achieved impressive performance in various graph-related tasks [1]. Typically, a GNN is a neural network consists of consecutive propagation layers, each of which contains two steps, namely, the feature aggregation step and the feature transformation step [2], [3]. The process of passing through the consecutive propagation layers is also referred to as message passing. Recent studies [4]–[6] have proven that the feature aggregation steps in a class of GNNs can be interpreted as performing graph signal denoising (GSD). Specifically, given a graph $\mathcal{G}$ with N nodes, where each node is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^M$, the GSD problem can be defined as follows:

$$\underset{\mathbf{H}}{\text{minimize}} \quad \|\mathbf{H}-\mathbf{X}\|_F^2 + \lambda \text{tr}\left(\mathbf{H}^T\mathbf{L}\mathbf{H}\right), \tag{1}$$

where $\mathbf{H}$ is the output after denoising (or after the message passing process in a GNN), $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N\times M}$ denotes the input feature matrix, λ represents a weight parameter, and $\mathbf{L}$ is the graph Laplacian matrix. From the underlying GSD problem (1) corresponding to a class of GNNs, e.g., graph convolutional network (GCN) [2] and approximate personalized propagation of neural predictions (APPNP) [7], it can be observed that the graph Laplacian matrix is assumed to be credible enough. However, if the graph is noisy (e.g., contains task-irrelevant edges) or under adversarial attacks, the performance of GNNs may drop significantly and even be worse than the performance of a simple baseline that ignores all the relational information among data features such as multi-layer perceptrons [8], [9].

To improve the robustness of GNN models, some studies suggest pre-processing the graph to refine the graph structure before feeding it to a GNN model. For example, the GNN-Jaccard method [10] proposes to remove edges that connect nodes with feature vectors of low "Jaccard similarity". Under the observation that existing adversarial attack methods tend to increase the rank of the graph adjacency matrix, [11] uses a low-rank approximation version of the given graph adjacency matrix as a substitute. Besides pre-processing the graph, there are also some studies designing "graph learners," which are parameterized models to learn the graph structure matrix (e.g., the graph adjacency matrix) that are co-trained with the GNN models [12], [13]. These previous studies, however, all focus on augmenting extra components to the existing GNN models to improve their robustness instead of designing new GNN architectures that are inherently robust.

Inspired by recent studies that build GNNs from an underlying optimization problems and the corresponding iterative algorithms [14]–[16], in this paper, we present GSLDN, which is designed based on a joint graph signal and Laplacian denoising (GSLD) problem. Specifically, the GSLD problem is first reparameterized using the graph Laplacian operator to reduce the parameter dimension. The reparameterized problem is tackled via the block majorization-minimization (BMM) algorithm, which not only gives iterative steps that are friendly to back-propagation training but also guarantees convergence to stationary points. Then a robust and interpretable message passing scheme is induced from the BMM algorithm, based on which we develop GSLDN. Built on a BMM algorithm, the message passing procedure in a trained GSLDN is naturally a parameter-optimized BMM algorithm for solving the GSLD problem. Thus, GSLDN model is endowed with the power to fight against adversarial attacks on the graph structure. The experiment results on real-world datasets demonstrate that the proposed GSLDN model is resistive to adversarial attacks.

II. Graph Signal and Structure Denoising

We consider a positively weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V}$ represents the vertex set containing N nodes, $\mathcal{E}$ denotes the edge set, and $\mathbf{W} \in \mathbb{R}^{N\times N}$ is the weight matrix with w_{ij} indicating the pairwise relationship

between the i-th node and the j-th node. The degree matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose i-th diagonal element $d_{ii} = \sum_{j=1}^{N} w_{ij}$. The positive semi-definite Laplacian matrix is accordingly defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Assuming each node is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^M$, then the feature matrix is defined as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N\times M}$. Each column of $\mathbf{X}$ can also be interpreted as a graph signal.

A. Problem Formulation

In this section, we will introduce the GSLD problem, which augments the GSD problem (1) with graph Laplacian learning components to help mitigate the influence of adversarial attacks on graph structure.

For the GSD problem (1), we notice that it only concerns with denoising the graph signal while assumes the Laplacian matrix to be accurate enough. However, when the graph structure is maliciously manipulated by an attacker, such assumption will result in unsatisfactory model performance and may lead to dramatic consequences. Specifically, it has been shown that with unnoticeable and carefully designed modifications, the performance of GNN models significantly drops and may even be worse than the performance of a simple baseline that ignores all relational information [8], [9].

In order to handle potentially attacked graph structures, we propose to augment the objective in Problem (1) with a Laplacian learning term, i.e., $\|\mathbf{L} - \mathbf{L}_n\|_F^2$, where $\mathbf{L}$ is the refined Laplacian matrix that needs to be optimized and $\mathbf{L}_n$ is the given attacked Laplacian matrix. Then the objective of the GSD problem becomes

$$f(\mathbf{H}, \mathbf{L}) = \|\mathbf{H} - \mathbf{X}\|_F^2 + \gamma\|\mathbf{L} - \mathbf{L}_n\|_F^2 + \lambda \mathrm{tr}\left(\mathbf{H}^T\mathbf{L}\mathbf{H}\right). \quad (2)$$

Under the assumption that the denoised feature $\mathbf{H}$ is smooth over the refined graph $\mathbf{L}$, by minimizing $f(\mathbf{H}, \mathbf{L})$, the influence of the adversarial attacks can be mitigated.

It is evident that the Laplacian matrix is symmetric with degrees of freedom equal to $\frac{N(N-1)}{2}$. Thus, the Laplacian matrix $\mathbf{L}$ can be determined by a vector $\mathbf{w} \in \mathbb{R}_+^{\frac{N(N-1)}{2}}$ and the $f(\mathbf{H}, \mathbf{L})$ can be reparameterized accordingly. By reparameterizing $\mathbf{L}$ to $\mathbf{w}$ through the graph Laplacian operator, the dimension of variables is reduced and the positive semi-definiteness of $\mathbf{L}$ is satisfied naturally. In the following, we first introduce the notion of graph Laplacian operator [17].

Definition 1. The graph Laplacian operator $\mathcal{L} : \mathbb{R}^{\frac{N(N-1)}{2}} \to \mathbb{R}^{N\times N}$, $\mathbf{w} \to \mathcal{L}\mathbf{w}$, is defined as

$$[\mathcal{L}\mathbf{w}]_{ij} = \begin{cases} -w_{k_{ij}} & i > j, \\ [\mathcal{L}\mathbf{w}]_{ji} & i < j, \\ -\sum_{j\neq i}[\mathcal{L}\mathbf{w}]_{ij} & i = j, \end{cases}$$

where $k_{ij} = i - j + \frac{j-1}{2}(2n - j)$.

Similarly, the inverse graph Laplacian operator is defined as $\mathcal{L}^{-1} : \mathbb{R}^{N\times N} \to \mathbb{R}^{\frac{N(N-1)}{2}}$ such that $\mathbf{w} = \mathcal{L}^{-1}\mathcal{L}\mathbf{w}$. Based on the graph Laplacian operator, we can derive its adjoint operator $\mathcal{L}^*$, which satisfies $\mathrm{tr}\big((\mathcal{L}\mathbf{w})^T\mathbf{Y}\big) = \mathbf{w}^T\mathcal{L}^*\mathbf{Y}$ [17].

Definition 2. The adjoint operator of the graph Laplacian operator $\mathcal{L}$, i.e., $\mathcal{L}^* : \mathbb{R}^{N\times N} \to \mathbb{R}^{\frac{N(N-1)}{2}}$, $\mathbf{Y} \to \mathcal{L}^*\mathbf{Y}$, is defined as

$$[\mathcal{L}^*\mathbf{Y}]_k = Y_{ii} - Y_{ij} - Y_{ji} + Y_{jj}, \quad k = 1, \ldots, \frac{N(N-1)}{2},$$

where $i, j \in \mathbb{Z}^+$ satisfy $k = i - j + \frac{j-1}{2}(2n - j)$ and $i > j$.

To better understand the graph Laplacian operator and its adjoint operator, more discussions and examples can be found in [17]. Using the notion of Laplacian operator, we rewrite $f(\mathbf{H}, \mathbf{L})$ (where f is reused with a little abuse of the notation) as:

$$f(\mathbf{H}, \mathbf{w}) = \|\mathbf{H} - \mathbf{X}\|_F^2 + \gamma\|\mathcal{L}\mathbf{w} - \mathbf{L}_n\|_F^2 + \lambda \mathrm{tr}\left(\mathbf{H}^T\mathcal{L}\mathbf{w}\mathbf{H}\right).$$

Considering that real-world graphs are generally sparse and low-rank [18], we further consider regularization terms, i.e., $||\mathcal{L}\mathbf{w}||_1$ and $||\mathcal{L}\mathbf{w}||_*$, to promote the sparsity and the low-rankness of the learned Laplacian matrix $\mathcal{L}\mathbf{w}$. It can be proved that $||\mathcal{L}\mathbf{w}||_1 = 4||\mathbf{w}||_1$. Besides, since $\mathcal{L}\mathbf{w}$ is positive semi-definite, we have $||\mathcal{L}\mathbf{w}||_* = \mathrm{tr}(\mathcal{L}\mathbf{w}) = 2||\mathbf{w}||_1$. Thus, the sparsity and the low-rankness promoting property can be achieved with a single regularization term $||\mathbf{w}||_1$. To restrict the scale of the Laplacian matrix, we further consider a regularization term $||\mathbf{w}||_2^2$. Finally, the GSLD problem is defined as follows:

$$\underset{\mathbf{H},\mathbf{w}\geq\mathbf{0}}{\text{minimize}} \quad f(\mathbf{H}, \mathbf{w}) + \alpha\|\mathbf{w}\|_1 + \beta\|\mathbf{w}\|_2^2, \quad (3)$$

where α and β are positive weighting parameters.

B. Solving Problem (3) via BMM

In this section, we will develop an efficient and convergent BMM algorithm for solving Problem (3). Denoting $\mathbf{H}^{(k-1)}$ and $\mathbf{w}^{(k-1)}$ as the value of $\mathbf{H}$ and $\mathbf{w}$ after the $(k-1)$-th iteration, with $\mathbf{H}^{(0)}$ and $\mathbf{w}^{(0)}$ being the initial points, we derive the update steps of $\mathbf{H}$ and $\mathbf{w}$ at the k-th iteration in the following. To proceed with the resolution, we introduce a useful lemma.

Lemma 3 ([19])**.** *For a continuously differentiable function $f(\mathbf{X})$ where ∇f is Lipschitz continuous with Lipschitz constant L, the following result holds for all $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N\times M}$ and $\xi \geq L$:*

$$f(\mathbf{X}) \leq f(\mathbf{Y}) + \mathrm{tr}\left((\mathbf{X} - \mathbf{Y})^T\nabla f(\mathbf{Y})\right) + \frac{\xi}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2.$$

1) The Resolution of the $\mathbf{H}$-block Subproblem: With $\mathbf{w}$ being held fixed, the subproblem w.r.t. $\mathbf{H}$ becomes

$$\underset{\mathbf{H}}{\text{minimize}} \quad h(\mathbf{H}\,|\,\mathbf{w}^{(k-1)}) = \|\mathbf{H}-\mathbf{X}\|_F^2 + \lambda\mathrm{tr}\big(\mathbf{H}^T\mathcal{L}\mathbf{w}^{(k-1)}\mathbf{H}\big).$$

Let $\lambda_{\max}^{(k-1)}$ be the largest eigenvalue of $\mathcal{L}\mathbf{w}^{(k-1)}$. Then the Lipschitz constant of $\nabla h^{(k-1)}(\mathbf{H})$ is $2 + 2\lambda\lambda_{\max}^{(k-1)}$. With $\eta^{(k)} \geq 2 + 2\lambda\lambda_{\max}^{(k-1)}$, we can construct a majorization function of $h^{(k-1)}(\mathbf{H})$ based on Lemma 3 as follows:

$$\begin{aligned} &p\big(\mathbf{H} \mid \mathbf{w}^{(k-1)}, \mathbf{H}^{(k-1)}\big) \\ =&\|\mathbf{H}^{(k-1)} - \mathbf{X}\|_F^2 + \mathrm{tr}\Big(\lambda(\mathbf{H}^{(k-1)})^T\mathcal{L}\mathbf{w}^{(k-1)}\mathbf{H}^{(k-1)}\Big) \\ &+ \mathrm{tr}\Big((\mathbf{H} - \mathbf{H}^{(k-1)})^T\mathbf{D}^{(k-1)}\Big) + \frac{\eta^{(k)}}{2}\|\mathbf{H} - \mathbf{H}^{(k-1)}\|_F^2, \end{aligned}$$

where $\mathbf{D}^{(k-1)} = 2\mathbf{H}^{(k-1)} - 2\mathbf{X} + 2\lambda\mathcal{L}\mathbf{w}^{(k-1)}\mathbf{H}^{(k-1)}$. Ignoring the constant terms in $p\big(\mathbf{H} \mid \mathbf{w}^{(k-1)}, \mathbf{H}^{(k-1)}\big)$, we obtain a surrogate optimization problem as follows:

$$\underset{\mathbf{H}}{\text{minimize}} \quad \|\mathbf{H} - \mathbf{H}^{(k-1)} + \frac{1}{\eta^{(k)}}\mathbf{D}^{(k-1)}\|_F^2.$$

This surrogate optimization problem has a closed-form solution, which leads to following update rule:

$$\mathbf{H}^{(k)} = \mathbf{H}^{(k-1)} - \frac{1}{\eta^{(k)}}\mathbf{D}^{(k-1)}. \tag{4}$$

Note that this update step can also be interpreted as a gradient step with a stepsize of $\frac{1}{\eta^{(k)}}$.

2) The Resolution of the $\mathbf{w}$*-block Subproblem:* With $\mathbf{H}$ being held fixed, the subproblem with respect to $\mathbf{w}$ becomes

$$\begin{aligned}\underset{\mathbf{w}\geq\mathbf{0}}{\text{minimize}} \;\; & w(\mathbf{w} \mid \mathbf{H}^{(k-1)}) = \|\mathcal{L}\mathbf{w}\|_F^2 - 2\mathrm{tr}\left(\mathbf{L}_n^T\mathcal{L}\mathbf{w}\right) \\ & + \frac{\lambda}{\gamma}\mathrm{tr}\left((\mathbf{H}^{(k-1)})^T\mathcal{L}\mathbf{w}\mathbf{H}^{(k-1)}\right) + \frac{\alpha}{\gamma}\|\mathbf{w}\|_1 + \frac{\beta}{\gamma}\|\mathbf{w}\|_2^2,\end{aligned}$$

where we omit the constant term $\|\mathbf{L}_n\|_F^2$. Using the adjoint operator $\mathcal{L}^*$, we have

$$w(\mathbf{w} \mid \mathbf{H}^{(k-1)}) = \|\mathcal{L}\mathbf{w}\|_F^2 - \mathbf{w}^T\mathbf{c}^{(k-1)} + \frac{\beta}{\gamma}\|\mathbf{w}\|_2^2,$$

where $\mathbf{c}^{(k-1)} = \mathcal{L}^*\big[2\mathbf{L}_n - \frac{\lambda}{\gamma}\mathbf{H}^{(k-1)}(\mathbf{H}^{(k-1)})^T\big] - \frac{\alpha}{\gamma}\mathbf{1}$ with $\mathbf{1} \in \mathbb{R}^{\frac{N(N-1)}{2}}$ being the all-one column vector. Due to the non-negativity constraint, the $\mathbf{w}$-block subproblem does not have a closed-form solution. To obtain closed-form solution, we employ a majorization technique [20].

According to [17, Lemma 1] and [17, Lemma 3], we conclude that the $\|\mathcal{L}\mathbf{w}\|_F^2$ is L-smooth with $L = 2N$. Thus, we can construct a majorization function of $\|\mathcal{L}\mathbf{w}\|_F^2$ with $\xi^{(k)} \geq 2N$ based on Lemma 3, leading to a majorization function of $w(\mathbf{w} \mid \mathbf{H}^{(k-1)})$ as follows:

$$\begin{aligned}& g\big(\mathbf{w} \mid \mathbf{w}^{(k-1)}, \mathbf{H}^{(k-1)}\big) \\ & = \|\mathcal{L}\mathbf{w}^{(k-1)}\|_F^2 + 2\big(\mathbf{w} - \mathbf{w}^{(k-1)}\big)^T\mathcal{L}^*\mathcal{L}\big(\mathbf{w}^{(k-1)}\big) \\ & \qquad + \frac{\xi^{(k)}}{2}\big\|\mathbf{w} - \mathbf{w}^{(k-1)}\big\|^2 - \mathbf{w}^T\mathbf{c}^{(k-1)} + \frac{\beta}{\gamma}\|\mathbf{w}\|_2^2.\end{aligned}$$

Note that $\xi^{(1)}, \ldots, \xi^{(K)}$ can be set to be the same or varying in different layers. Ignoring the constant terms in $g\big(\mathbf{w} \mid \mathbf{w}^{(k-1)}, \mathbf{H}^{(k-1)}\big)$, we obtain a surrogate optimization problem as follows:

$$\underset{\mathbf{w}\geq\mathbf{0}}{\text{minimize}} \quad \left(\frac{\xi^{(k)}}{2} + \frac{\beta}{\gamma}\right)\|\mathbf{w}\|_2^2 - \mathbf{w}^T\mathbf{b}^{(k-1)}, \tag{5}$$

where $\mathbf{b}^{(k-1)} = \xi^{(k)}\mathbf{w}^{(k-1)} + 2\mathcal{L}^*\mathcal{L}\big(\mathbf{w}^{(k-1)}\big) - \mathbf{c}^{(k-1)}$. Then using the KKT optimality condition, we have the following update rule:

$$\mathbf{w}^{(k)} = \frac{\gamma}{\gamma\xi^{(k)} + 2\beta}\mathrm{ReLU}\left(\mathbf{b}^{(k-1)}\right). \tag{6}$$

In conclusion, in each iteration, the BMM algorithm update $\mathbf{H}$ as in Eq. (4) and update $\mathbf{w}$ as in Eq. (6). The convergence property of the proposed algorithm is stated in the following theorem.

Theorem 4. *With* $\eta^{(k)} \geq 2 + 2\lambda\lambda_{\max}^{(k-1)}$ *and* $\xi^{(k)} \geq 2N$ *for* $k = 1, \ldots, K$*, the BMM algorithm for solving the GSLD problem* (3) *with update rules as in Eq.* (4) *and Eq.* (6) *converges to the stationary points.*

Proof: With $\eta^{(k)} \geq 2 + 2\lambda\lambda_{\max}^{(k-1)}$ and $\xi^{(k)} \geq 2N$ for $k = 1, \ldots, K$, the proposed algorithm is a standard BMM algorithm. Thus, following the convergence results for general BMM algorithms [21], the proposed algorithm ensures convergence to stationary points. ■

III. The Graph Signal and Laplacian Denoising Network

Based on the BMM algorithm developed in Section II-B, we present the GSLDN below.

GSLDN Architecture

$$\begin{aligned}& \mathbf{H}^{(0)} = \mathrm{ReLU}\left(\mathrm{ReLU}\left(\mathbf{X}\mathbf{\Theta}_1\right)\mathbf{\Theta}_2\right), \quad \mathbf{w}^{(0)} = \mathcal{L}^{-1}\mathbf{L}_n, \\ & \text{for } k = 1, \ldots, K, \\ & \begin{cases}\mathbf{H}^{(k)} = \big(1 - 2\eta^{(k)}\big)\mathbf{H}^{(k-1)} \\ \qquad\qquad -2\eta^{(k)}\lambda\mathcal{L}\mathbf{w}^{(k-1)}\mathbf{H}^{(k-1)} + 2\eta^{(k)}\mathbf{X}, \\ \mathbf{w}^{(k)} = \frac{\gamma}{\gamma\xi^{(k)} + 2\beta}\mathrm{ReLU}\left(\mathbf{b}^{(k-1)}\right)\end{cases} \\ & \mathbf{Z} = \mathrm{softmax}\big(\mathbf{H}^{(K)}\big).\end{aligned}$$

In the above GSLDN, $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ are the learnable weight matrices and $\mathbf{Z}$ is the output probability matrix for prediction. Note that we use the decoupled structure as in APPNP [7], i.e., the feature aggregation steps are not intertwined with the feature transformation steps. Specifically, the feature transformation $\mathrm{ReLU}\left(\mathrm{ReLU}\left(\mathbf{X}\mathbf{\Theta}_1\right)\mathbf{\Theta}_2\right)$ is performed before conducting the feature aggregation steps. To ensure the convergence of the BMM algorithm, i.e., the message passing procedure, we can set $\eta^{(k)}$ and $\xi^{(k)}$ as constants satisfying $\eta^{(k)} \geq 2 + 2\lambda\lambda_{\max}^{(k-1)}$ and $\xi^{(k)} \geq 2N$. To improve the model expressiveness and save the computation of $\lambda_{\max}^{(k-1)}$, we set $\eta^{(k)}$ and $\xi^{(k)}$ to be learnable parameters. Since the GSLDN is induced from the BMM algorithm, the message passing procedure in a trained GSLDN is naturally a parameter-optimized BMM algorithm.

Since GSLDN is a GNN architecture, some extra components used in existing graph structure learning methods, i.e., the pre-processing methods [10], [11] and the graph-learner based methods [12], [13] can be used to further improve the model performance, which we leave for future exploration.

IV. Experimental Results

In this section, we conduct experiments on semi-supervised node classification tasks with two real-world citation graphs, i.e., Cora and Citeseer [22] to validate the effectiveness of the proposed GSLDN.

A. Experiment Settings

To evaluate the effectiveness of GSLDN, we compare it with GCN and several benchmarks that are designed from different perspectives to robustify the GNNs, including GCN-Jaccard [10] that pre-processes the graph by eliminating edges with low Jaccard similarity of node feature vectors, GCN-SVD [11] that uses the low-rank approximation of the given graph adjacency matrix, Pro-GNN [12] that jointly learns a graph structure and a GNN model guided by some predefined structural priors, and Elastic GNN [14] that utilizes trend filtering instead of Laplacian smoothing to promote robustness. For GCN-Jaccard, GCN-SVD, and Pro-GNN, we use the implementation provided in DeepRobust [23]. For Elastic GNN, we follow the implementation provided in the original paper [14].

For each graph, we only consider the largest connected component and randomly select 10%/10%/80% of nodes for training, validation, and testing. The Adam optimizer is used in all experiments. The models' hyperparameters are tuned based on the results of the validation set. The search space of hyperparameters is as follows: 1) learning rate: {0.005, 0.01, 0.05}; 2) weight decay: {0, 5e-5, 5e-4}; 3) dropout rate: {0.1, 0.5, 0.8}; 4) model depth: {2, 4, 8, 16}. For GCN-Jaccard, the threshold of Jaccard similarity for removing dissimilar edges is chosen from {0.01, 0.02, 0.03, 0.04, 0.05, 0.1}. For GCN-SVD, the reduced rank of the graph is tuned from {5, 10, 15, 50, 100, 200}. For Elastic GNN, the regularization coefficients are chosen from {3, 6, 9}. For Pro-GNN, we adopt the hyperparameters provided in their paper [12].

B. Performance Under Adversarial Attack

We evaluate the model performance under the training-time adversarial attacks [9], i.e., the graph is first attacked, and then the GNN models are trained on the perturbed graph. Two types of attacks are considered, namely, the global attack that aims to reduce the overall performance of GNNs [9] and the targeted attack that aims to fool GNNs on some specific nodes [8]. For all the experimental results, we give the average performance and standard variance with 10 independent trials. The best model and the runner-up model are highlighted in bold and wavy underlining, respectively.

1) Model Performance Under Global Attack: For the global attack, we use a representative method called meta-attack [9] and the results at a 20% and a 25% perturbation rate are showcased in Table I. For GCN-Jaccard method, we use the results reported in [12]. From the table, we observe that the proposed GSLDN model achieves better or comparative results compared with other methods. Specifically, the GSLDN outperforms other GNN architectures, i.e., GCN and Elastic GNN, indicating that GSLDN can effectively resist the global attack. For instance, GSLDN improves GCN by 11% on the Cora dataset at a 25% perturbation rate. Note that although Pro-GNN outperforms GSLDN in some cases, it requires training an additional graph learner. Moreover, the graph learner used in Pro-GNN can also be used to boost the performance of GSLDN, which we leave for future work.

TABLE I
CLASSIFICATION PERFORMANCE UNDER GLOBAL ATTACK

Ptb. rate	Cora		Citeseer	
	0.2	0.25	0.2	0.25
GCN	61.5 ± 2.2	56.8 ± 1.4	59.7 ± 0.8	60.0 ± 1.0
GCN-Jaccard	65.7 ± 0.9	60.8 ± 1.1	59.3 ± 1.4	59.9 ± 1.5
GCN-SVD	58.8 ± 2.1	59.1 ± 2.7	65.8 ± 0.7	62.3 ± 0.6
Pro-GNN	**70.1 ± 2.5**	67.0 ± 1.6	**70.1 ± 1.1**	**69.7 ± 0.9**
Elastic GNN	68.3 ± 3.5	65.8 ± 2.5	61.6 ± 1.8	64.0 ± 2.2
GSLDN	69.2 ± 2.0	**67.8 ± 1.1**	67.3 ± 0.4	66.5 ± 0.8

TABLE II
CLASSIFICATION PERFORMANCE UNDER TARGETED ATTACK

Ptb. number	Cora		Citeseer	
	4.0	5.0	4.0	5.0
GCN	61.5 ± 2.2	56.8 ± 1.4	62.5 ± 1.6	52.7 ± 2.0
GCN-Jaccard	61.7 ± 1.1	59.5 ± 1.9	76.3 ± 1.5	72.9 ± 1.7
GCN-SVD	58.8 ± 2.1	59.2 ± 2.7	62.2 ± 3.3	60.2 ± 6.7
Pro-GNN	70.1 ± 2.5	67.0 ± 1.6	75.7 ± 4.9	74.0 ± 7.1
Elastic GNN	68.3 ± 3.5	65.8 ± 2.5	72.1 ± 5.6	74.0 ± 3.9
GSLDN	**70.5 ± 1.5**	**68.3 ± 4.5**	**78.2 ± 2.7**	**75.2 ± 3.5**

2) Model Performance Under Targeted Attack: For the targeted attack, we use a representative method called nettack [8]. The results with 4 and 5 perturbations per targeted node are reported in Table II. Specifically, following [12], we choose the nodes in the test set with degrees larger than 10 as targeted nodes and the classification performance is evaluated on target nodes. From the table, we can see that the proposed GS^{2}DNet attains better performance than other baselines in all cases. For instance, on the Citeseer dataset with 5 perturbations per targeted node, GSLDN improves GCN by 22.5% and outperforms other baselines by 1.2%. Such inspiring results demonstrate that GSLDN can better resist targeted attacks than other baseline methods.

V. CONCLUSION

In this paper, we first introduced graph Laplacian learning components into graph signal denoising. Then we developed a block majorization-minimization algorithm for problem resolution, based on which we proposed the GSLDN model. Designed from an optimization perspective, GSLDN is endowed with the power of performing Laplacian learning during the message passing procedure. Experiments validate the robustness of GSLDN under both the global and targeted attack.

REFERENCES

[1] L. Wu, P. Cui, J. Pei, and L. Zhao, *Graph neural networks: Foundations, frontiers, and applications*. Singapore: Springer Singapore, 2022, p. 725.

[2] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[3] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] Y. Ma, X. Liu, T. Zhao, Y. Liu, J. Tang, and N. Shah, "A unified view on graph neural networks as graph signal denoising," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 1202–1211.

[5] M. Zhu, X. Wang, C. Shi, H. Ji, and P. Cui, "Interpreting and unifying graph neural networks with an optimization framework," in *Proceedings of the Web Conference*, 2021, pp. 1215–1226.

[6] Z. Zhang and Z. Zhao, "Towards understanding graph neural networks: An algorithm unrolling perspective," *arXiv preprint arXiv:2206.04471*, 2022.

[7] J. Gasteiger, A. Bojchevski, and S. Günnemann, "Predict then propagate: Graph neural networks meet personalized pagerank," in *International Conference on Learning Representations*, 2018.

[8] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2847–2856.

[9] D. Zügner and S. Günnemann, "Adversarial attacks on graph neural networks via meta learning," in *International Conference on Learning Representations*, 2019.

[10] H. Wu, C. Wang, Y. Tyshetskiy, A. Docherty, K. Lu, and L. Zhu, "Adversarial examples for graph data: Deep insights into attack and defense," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 4816–4823.

[11] N. Entezari, S. A. Al-Sayouri, A. Darvishzadeh, and E. E. Papalexakis, "All you need is low (rank) defending against adversarial attacks on graphs," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 169–177.

[12] W. Jin, Y. Ma, X. Liu, X. Tang, S. Wang, and J. Tang, "Graph structure learning for robust graph neural networks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 66–74.

[13] B. Runwal and S. Kumar, "Robustifying GNN via weighted laplacian," in *IEEE International Conference on Signal Processing and Communications*, 2022, pp. 1–5.

[14] X. Liu, W. Jin, Y. Ma, *et al.*, "Elastic graph neural networks," in *International Conference on Machine Learning*, PMLR, 2021, pp. 6837–6849.

[15] S. Chen, Y. C. Eldar, and L. Zhao, "Graph unrolling networks: Interpretable neural networks for graph signal denoising," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3699–3713, 2021.

[16] Z. Zhang, S. Lu, Z. Huang, and Z. Zhao, "ASGNN: Graph neural networks with adaptive structure," *arXiv preprint arXiv:2210.01002*, 2022.

[17] S. Kumar, J. Ying, J. V. de Miranda Cardoso, and D. P. Palomar, "A unified framework for structured graph learning via spectral constraints.," *J. Mach. Learn. Res.*, vol. 21, no. 22, pp. 1–60, 2020.

[18] K. Zhou, H. Zha, and L. Song, "Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes," in *Artificial Intelligence and Statistics*, PMLR, 2013, pp. 641–649.

[19] D. P. Bertsekas, *Nonlinear programming*, 1999.

[20] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2016.

[21] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[22] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI Magazine*, vol. 29, no. 3, pp. 93–93, 2008.

[23] Y. Li, W. Jin, H. Xu, and J. Tang, "Deeprobust: A pytorch library for adversarial attacks and defenses," *arXiv preprint arXiv:2005.06149*, 2020.

Comparative Analysis of Glottal and Vocal Tract Features in Dysarthria

Geeta Sai Sahasra[1], Kadwasra Swapna[1], Arushi Srivastava[1], Aditya Pusuluri[2], and Hemant A. Patil[3]
[1]Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana
E-mail:2210080009@klh.edu.in, 2210040125@klh.edu.in
[2]Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat
E-mail:202215003@daiict.ac.in, aditya_pss@daiict.ac.in, hemant_patil@daiict.ac.in

***Abstract*—Dysarthria is a neurological disorder associated with the muscles in the vocal tract system that is caused by problems with coordination. It affects the vocal tract system and the glottis with different levels of severity. This research deals with the role of vocal tract (i.e., filter) and glottal excitation source in dysarthric speech, using Glottal Flow Model with Iterative Adaptive Inverse Filtering (GFM-IAIF). By decomposing the vocal tract and glottal source into two separate components, we were able to identify which component (source or filter) is affected the most. Our research showed that the vocal tract system is the most affected part, determining a better classification of features the magnitude spectrum-based Mel frequency cepstral coefficient (MFCC) resulted in 95.75% accuracy for vocal tract components and lower accuracy of 86.5% for glottal source components. In the same way, Modified Group Delay Cepstral Coefficients (MGDCC) correlated with a test accuracy of 94.43% of the vocal tract in comparison with 88.70% of the glottal source. These outcomes reveal that the vocal tract gets most damaged due to dysarthria and thus, emphasizes the use of specific diagnostic and therapy interventions targeted at this area. The research not only explores feature fusion but also recommends further efforts to refine dysarthria diagnosis and treatment.**

Keywords : Dysarthria, glottal flow model using the iterative adaptive inverse filtering algorithm, phase, magnitude-based components.

I. Introduction

An impairment in motor speech, dysarthria, is characterized by muscle weakness and is often the result of neurological damage from either neuromuscular origins (e.g., brain tumor, stroke) or neurodegenerative diseases (e.g., Parkinson's or Huntington's). Impaired speech production mechanism can include poor quality production of speech, such as imprecise articulation, reduced loudness, atypical speech prosody, and varying speech rate. Dysarthria occurs when there is weakness in or paralysis of the muscles involved in speech and can include slurred speech, abnormal speech rhythm, pitch (i.e., fundamental frequency, F_0), loudness control, difficulty controlling voice pitch or volume, drooling, and so forth. Considered as one of the features of disordered speech production, dysarthria is caused by a neurological problem, is still usually the dominant issue [1]. Contrary to this, vowels in dysarthric speech, in general, are produced with the help of insufficient articulation (the failure of the produced vowel to reach the canonical formant frequencies) could suffer this [2]. Quite often, this can be avoided standing for a compression during the critical part of production or a vowel space that is contracted [3].

Speech production mechanism begins with the expulsion of air from the lungs [4]. The vocal folds in the larynx either vibrate to generate voiced sounds or remain open for unvoiced sounds, creating a raw sound source. The tongue, jaw, muscles, and lips then shape the vocal tract system, influencing how the sound resonates [4], [5]. This interaction between the glottal excitation source and the vocal tract system configuration produces the diverse speech sounds.

In this study, we investigate how the glottal source (GS) and the vocal tract (VT) components are affected in dysarthria, focusing on how the impairment in motor control mechanism leads to altered patterns w.r.t. natural speech production mechanism. Our analysis aims to determine which of the two: glottal source *vs.* vocal tract system is affected the most due to dysarthria. We aim to identify whether the primary impact lies in the irregular vibration patterns and airflow generated at the glottal source or in the resonance (i.e., reinforcement of spectrum energy) characteristics shaped by the vocal tract system configuration [6]. We pre-process individually the glottal and vocal source features, with fusion performed at the feature-level. A detailed analysis of the results is presented. This analysis is instrumental in advancing the study and provides insights into which component, GS or VT, is more significantly affected. This understanding aids in better diagnosing and treating the condition of dysarthric patients.

II. Proposed Approach

In this Section, we briefly describe how the excitation source and the vocal tract filter components are separated.

A. Glottal Flow Model using the Iterative Adaptive Inverse Filtering (GFM-IAIF)

The source-filter model of speech production is the combination of four components, namely, an excitation E, VT filter V, lip radiation filter L, and glottis component G, where speech signal is represented in the frequency-domain as S(f) = E(f)G(f)V(f)L(f) (due to linear time invariant assumption and

hence by applying convolution theorem), where lip-radiation is of the form in z domain [7]:

$$L(z) = 1 - bz^{-1}, \quad 0 < b \leq 1.$$

The main role of this model is to separate these components efficiently. Glottal Inverse Filtering (GIF) incorporates both glottal formants and spectral tilt characteristics [1], [8]. It isolates glottal airflow from speech signals by filtering out vocal tract system effects, focusing solely on the glottal airflow by introducing 3rd order filtering in pre-emphasis step. This method is implemented in speech analysis to identify characteristics associated with the glottal excitation source. Iterative Adaptive Inverse Filtering (IAIF) in combination with GIF enhances the estimation of glottal airflow and vocal tract filter parameters [7]. There are mainly four different stages included in it, in order to separate the source and system components, such as gross glottis, gross vocal tract, fine glottis, and fine vocal tract estimation [9]. Applying adaptive filtering ensures that it adapts to the change in characteristics of excitation noise [10]. Vocal tract filter V(z) is estimated using the linear prediction analysis on the highpass filtered signal. Generally in IAIF, LP (Linear Prediction) method estimates the vocal tract filter by finding the coefficients that minimize the residual or prediction error. However, in GFM-IAIF, the glottal air flow is estimated as the residual signal after filtering the speech signal through the LP coefficients derived vocal tract model [11]. Here, the assumption is that the glottal source signal is the broadband excitation that gets shaped by the vocal tract filter [12], [13]. The residual should ideally represent this glottal excitation. Then, the inverse filtering is applied to the original speech signal s(n) with the estimated V(z) to obtain the glottal flow derivative, $\dot{g}(t)$. Here, g(t) represents the glottal flow signal.

$$\dot{g}(t) = s(n) - \sum_{k=1}^{p} a_k s(n-k).$$

As it primarily captures the high frequency components of the glottal source signal sometimes it can also capture the high frequency noise, which are actually not the part of the glottal airflow. To avoid this risk further, smoothing was performed to reduce the noise. Lowpass filter is used to smooth the g(t) [14], [15]. To obtain the LP coefficients for glottal flow filter G(z), the LP analysis is done on this smoothed glottal flow derivative [16]. In particular,

$$\dot{g}(t) \approx \sum_{k=1}^{q} b_k \dot{g}(t-k).$$

The glottal flow filter is as follows [17]:

$$G(z) = \left\{(1 - az^{-1})(1 - a^* z^{-1})(1 - bz^{-1})\right\}^{-1},$$

where a and a* (complex conjugate of a) forms a pole representing the representing resonant frequency and amplitude of oscillations and their characteristics. Glottal flow contribution is represented by b, which shows the affect of the shape and behaviour of the glottal waveform. The glottal flow is estimated using direct steps, where only LP analysis and inverse filtering techniques are included without the need to locate the glottal opening or closure [7], [8]. Each iteration improves the clarity of the signal and adapts to the change in parameters. Here, the window size is taken same as the length of the speech frame, and LP coefficients for the vocal tract system, glottal source, and lip radiation are set with a default order of 48, 3, and 2, respectively [17]. This design ensures the separation of the GS and VT are efficient and simplifies the setup. The higher order formants are more affected and more fluctuations are observed [3]. Fig. 1 illustrates the separated source and filter components using GFM-IAIF for high severity dysarthric speech utterance. Fluctuations are high and formants are highly distorted as the severity of the dysarthria increases.

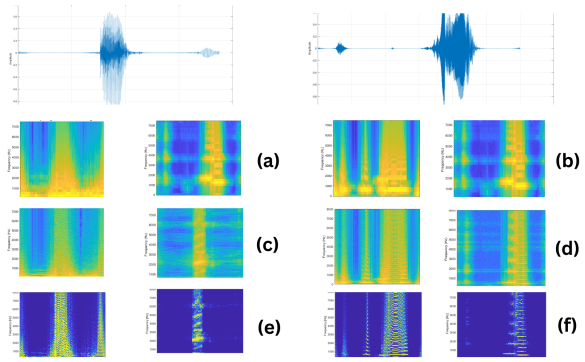


Fig. 1. Spectrogram analysis for normal and high severity, dysarthric female speakers speech signal of vowel 'a' of a female speaker. (a) Mel GS & VT normal speaker, (b) Mel GS & VT dysarthric speaker, (c) linear GS & VT original speaker, (d) linear GS & VT dysarthric speaker,(e) modified group delay GS & VT original speaker, and (f) modified group delay function of GS & VT dysarthric speaker.

B. Group Delay Cepstral Coefficients (GDCC)

To determine which component between the glottal source and vocal tract, is more significantly impacted, it is essential to analyze both magnitude spectrum and phase spectrumcomponents of these sources [2], [18]. To achieve this, our study incorporates the GDCC into classification model. GDCC is computed from the group delay function, which is the negative derivative of unwrapped Fourier transform phase converting them to cepstral of vocal tract domain coefficients. GDCC identifies the resonant frequencies from the phase spectrum. Many researchers have shown the evidences about the phase information in speech using GDCC through various experiments, such as human perception [19], [20]. Using GDCC, we extracted the phase components for both VT and GS. There are certain limitations associated with the application of GDCC. Specifically, GDCC is designed in order to use with minimum phase signals and is not suitable for mixed-phase signals [21]. In dysarthric speech (i.e., a system having all the poles and zeros inside the unit circle in the z-plane)

additional phases characteristics may be introduced due to irregular phonation. Consequently, GDCC demonstrates sub-optimal performance, when applied to both vocal tract and glottal source components of dysarthric speakers. In contrast, MGDCC has shown superior performance in this context [22].

C. Modified Group Delay Cepstral Coefficients (MGDCC)

The MGDCC is derived from the phase spectrum of the speech signal, converting it into cepstral coefficients. The parameters α and γ are optimized experimentally, where α is used for frequency warping, and γ controls the type and degree of cepstral transformation. The modified group delay function is as follows [23], [24]:

$$\tau_m(\omega,t) = \frac{\tau(\omega,t)S(\omega,t)^{\alpha}}{|\tau(\omega,t)|^{\gamma}}.$$

After applying the DCT to the logarithm of the modified group delay function to compute the MGDCC. In particular,

$$c_n = \sum_{k=0}^{K-1} \log(\tau_m(\omega_k,t)) \cos\left(\frac{\pi n(2k+1)}{2K}\right),$$

where c_n are the modified group delay cepstral coefficients, K is the number of frequency bins, and ω_k are the discrete frequencies. In our study, we extract MGDCC features by fixing γ at 0.1 and tuning α. After determining the optimal α, we kept it constant and tuned γ to find the best possible values for both the vocal tract (VT) and glottal source (GS) components. Features were extracted using a Hamming window of 25 ms duration and 10 ms shift for various α and γ combinations across different tasks. The phase-based features are extracted using this concept of MGDCC. Section IV discusses the experimental results obtained from implementing various values of tuning parameters.

D. Baseline Features

1) MFCC & LFCC: These are well known features in speech signal processing, essential for analyzing key features characteristics of signal [25]. MFCC represents the magnitude of different frequency components of a sound signal. MFCC use a non-linear scale called as the Mel scale. These better mimic the human auditory system. As human ear is sensitive to 1 kHz to 4 kHz, MFCC spaces frequency bands more closely at lower frequencies and more widely at higher frequencies similar to how human ear perceives sound. Signals undergo pre-emphasis, are windowed (25 ms with a 10 ms overlap), and transformed using FFT for spectral analysis. MFCC utilizes a Mel filterbank and DCT to extract 20 coefficients with 40 filterbanks, while LFCC employs a linear filterbank, providing superior spectral resolution for similar tasks. DCT is taken for feature vector decomposition, dimensionality reduction and energy compaction.

LFCC uses a linearly-spaced unlike the non-linear filterbank in MFCC. Here, the frequency bands are evenly spaced in the frequency-domain [26], [27]. It also captures the details that may be overlooked by the Mel scale. In this context, we are using both MFCC and LFCC features in order to extract the magnitude-based features that capture the fine spectral details, and the results are discussed in the Section IV.

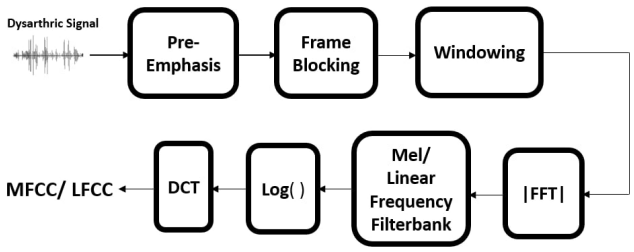


Fig. 2. Functional block diagram of MFCC/ LFCC feature extraction.

E. Proposed Methodology

The process begins with the input of dysarthric speech, which is processed using the Glottal Flow Model - Iterative Adaptive Inverse Filtering (GFM-IAIF) method. This method is crucial for decomposing the speech signal into two primary components, namely, the glottal source and the vocal tract. Once the glottal source and vocal tract components are extracted, they undergo further analysis through two different classification approaches: phase-based classification and magnitude-based classification. Both the glottal source and vocal tract are evaluated separately using these two approaches. Finally, the results from the phase-based and magnitude-based classifications are fed into a Convolutional Neural Network (CNN) for further processing and analysis.

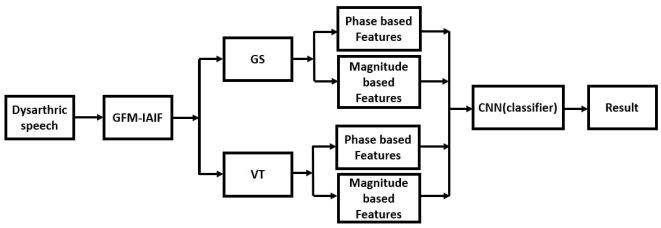


Fig. 3. Functional block diagram of proposed severity-level classification system.

III. EXPERIMENTAL SETUP

A. Dataset Used

The cepstral coefficients features, such as MGDCC, GDCC, MFCC, and LFCC are studied for the assessment of dysarthria. CNNs are employed as pattern classifiers with accuracy, precision, recall, and, F1-score as performance metric, of UA Speech Corpus dataset with various severity-levels high, low, medium, and very low. Finally, to understand the practicality of the proposed feature set, analysis of latency

period is prevented.

In this work, the severity-level classification is performed on Universal Access Speech Corpus (UA-Speech Corpus), and TORGO database [28], [29]. Their statistics are summarized in Table I. These corpora are primarily of the spastic dysarthric type, which is characterized by several factors, such as breathiness, hypernasality, harsh voice, and incorrect articulation that causes unintelligible speech. In this work, the data of 8 speakers (4 males, namely, M01, M05, M07, and M09, and 4 females, namely, F02, F03, F04, and F05) is considered. Each severity-level class, namely, high, low, medium, and very low of dysarthric speaker consists of 751, 926, 930, and 930 samples, respectively. From 735 utterances, 465 utterances per speaker as mentioned in were used. The training data and testing data consists of 80% and 20% of the entire dataset, respectively. The experiments conducted involve a *5*-fold cross-validation (CV).

TABLE I
STATISTICS OF UA-SPEECH AND TORGO CORPUS.

Severity-Level	UA-Speech	Type	Samples	TORGO	Type	Samples
Very low	F05, M09	Spastic	930	F04, M03	Spastic	671
Low	F04, M05	Mixed, Spastic	926	F01, M05	Spastic	627
Medium	F02, M04	Spastic	930	M01, M04	Spastic	684
High	F03, M01	Spastic	751	-	-	-

B. Classifier Used

When conducting the experiments, we applied the Convolutional Neural Network (CNN) classifier because it preserves the translation invariance enabling to grasp the most critical spectrotemporal patterns and variations occurring in the speech signals that it feeds [30], [31], [32]. The model is refined via the stratified *5*-fold cross-validation (CV) strategy with a seed value and a train and validation split of 80% and 20% using the optimizer in this instance of Adam, categorical cross-entropy loss function, and % classification accuracy as the evaluation metric. The algorithm was optimized using grid search with best learning rate and the batch size for 100 epochs. Two types of activation functions, i.e., ReLU and softmax, are employed for each activation neuron. A ReLU is chosen to increase learning speed and reduce the computational requirements, while the softmax activation is used in the final layer for multi-class classification. A normalization layer is added to each convolutional layer in addition to a dropout layer to prevent the issue of overfitting in the CNN model.

IV. EXPERIMENTAL RESULTS

After separating the vocal and glottal sources using the GFM-IAIF algorithm, MFCC and LFCC for each feature (glottal and vocal components) were computed and subsequently analyzed using a CNN classifier. The CNN classifier was trained on both feature sets (GS and VT), and the performance metrics were also evaluated. Initially, we have done the experiments using the UA - Speech dataset. The comparative analysis of VT and GS of this dataset using these features extraction techniques revealed that the accuracy is high in VT, which signifies that the vocal tract system is more getting affected in a dysarthric patients speech. This implies that abnormalities in the vocal tract system contribute significantly to the speech impairments observed in dysarthric patients. Further, research work could be explored for the fusion of MFCC and MGDCC features [33]. Not only feature-level fusion, we can explore score-level fusion and the classifier-level fusion features also in order to deeply analyze the contribution and effects of GS and VT. The combination of these features is hypothesized to capture a broader spectrum of speech characteristics for dysarthric speakers in order to improve the accuracy of detecting either VT or GS is affected much. We have also explored the temporal dynamics of the extracted features and analysed how these features change over the time period. The main motive of this study is that since we were able to predict the vocal tract system much being affected in the speech of the dysarthric patient, findings from this study can be useful to develop targeted therapeutic interventions that would potentially leads to more effective treatment outcomes of dysarthric patient [34].

TABLE II
PERFORMANCE USING LFCC AND MFCC FEATURE FOR GS AND VT COMPONENTS

Feature Set	GS\VT	Fold	Test Accuracy	Precision	Recall	F1-Score
LFCC	GS	86.57	90.25	91.06	91.05	90.99
LFCC	VT	**95.75**	95.76	95.79	96.22	95.97
MFCC	GS	92.82	92.51	93.84	93.05	93.08
MFCC	VT	91.97	90.96	93.41	91.53	91.6

Moving forward, since we have also extracted the phase-based features using MGDCC features, we have got the best α and γ values as 0.3 and 0.1, respectively, for glottal source, where the test accuracy is relatively high when compared to the other values of α and γ. For VT, the best α and γ values are 0.1 and 0.3, respectively, at these values, the accuracies are relatively high. For the GS components, the test accuracy is 88.70 and for the VT, the test accuarcy is 94.43, and here also, we can clearly identify that the vocal tract source is much being effected in the dysarthric speaker when compared with the GS.

TABLE III
ACCURACIES OF GS AND VT OF MFCC AND LFCC FEATURES

Feature Set	GS\VT	Fold	Test Accuracy	Precision	Recall	F1-Score
LFCC	GS	90.28	95.96	95.84	95.69	95.75
LFCC	VT	96.21	96.22	96.36	95.83	96.00
MFCC	GS	93.81	97.98	97.88	97.28	97.88
MFCC	VT	95.89	95.21	95.29	94.81	94.85

We have also extracted the phase-based features using MGDCC feature extraction technique for TORGO dataset where we have found the best values for α and γ for GS is 0.2 and 0.1 respectively and for VT system it is 0.2 and 0.9 respectively. The above mentioned are the accuracies of GS and VT for MFCC and LFCC features as 97.98 and 95.21, respectively, for both GS and VT. Using the CNN classifier these are the results that we have encountered.

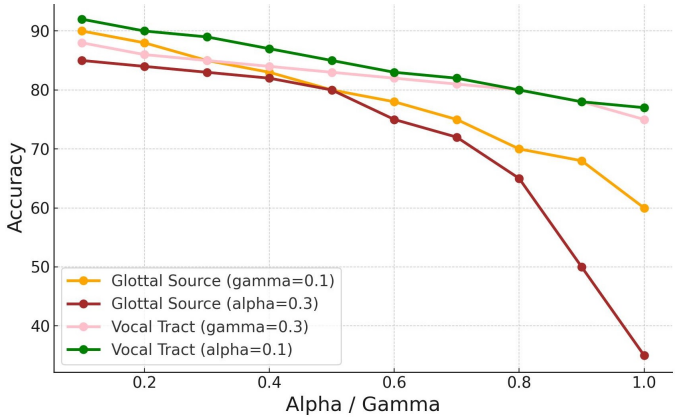


Fig. 4. Accuracies of GS and VT with α = 0.3, and γ = 0.1 of MGDCC features.

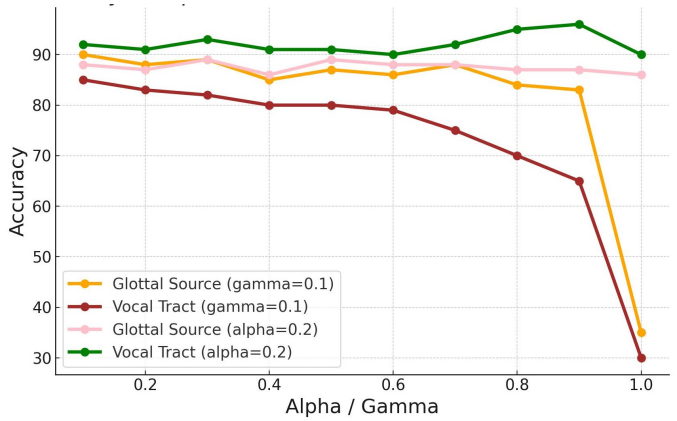


Fig. 5. Accuracies of GS and VT with α = 0.2 and γ = 0.1 of MGDCC features.

V. Summary and Conclusions

This study investigated the impact of glottal and vocal tract dysarthria treatment features of dysarthric speech, using advanced signal processing techniques and machine learning classifiers. The GFM-IAIF is the filter that is used to separate and analyze the contributions of glottal and vocal tract components effectively. From our investigation, analysis results suggest that the vocal tract system is the primary casualty in dysarthric speech, leaving the glottal source relatively less affected. To be precise, the measurement of magnitude-based features (e.g., MFCC and LFCC) exhibited an accuracy of 95.75% and 91.97%, for vocal tract features, while glottal source features presented 86.57% and 92.82% accuracy, respectively. However, using phase-based features, namely, MGDCC, the vocal tract's test accuracy was 94.43%, in contrast to the 88.70% for the glottal source. These findings certainly confirm the fact that vocal tract system is the main source of speech impairments in dysarthric patients. From a diagnostic aspect, the usefulness of vocal tract system in the differentiation of dysarthria severity using MFCC and MGDCC features is quite a performance that is usually seen. However, it deserves more research for effective application of the method. Further research is critical, where the combination of these features can be explored, and the analysis can be extended to different types of dysarthria on larger datasets. The idea is that this will further enhance diagnostic tools and enable precise therapeutic interventions leading to enhanced communication abilities and a better life for dysarthric individuals.

Acknowledgments

This study has been funded by the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, under the sponsorship of project 'BHASHINI', (Grant ID: 11(1) 2022-HCC.(TDIL)). Authors thank DA-IICT Gandhinagar and KL University for their encouragement and support in pursuing this research.

References

[1] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *arXiv preprint arXiv:1912.12604*, 2019, {Last Accessed Date : $22^{nd} July, 2024$}.

[2] M. Kamsali Veera, R. K. Vuddagiri, S. V. Gangashetty, and A. K. Vuppala, "Combining evidences from excitation source and vocal tract system features for indian language identification using deep neural networks," *International Journal of Speech Technology*, vol. 21, pp. 501–508, 2018, Springer.

[3] M. Freixes, M. Arnela, J. C. Socoró, F. Alías, and O. Guasch, "Glottal source contribution to higher order modes in the finite element synthesis of vowels," *Applied Sciences*, vol. 9, no. 21, p. 4535, 2019.

[4] K. Honda, "Physiological processes of speech production," *Springer Handbook of Speech Processing*, pp. 7–26, 2008.

[5] B. E. Lindblom and J. E. Sundberg, "Acoustical consequences of lip, tongue, jaw, and larynx movement," *The Journal of the Acoustical Society of America (JASA)*, vol. 50, no. 4B, pp. 1166–1179, 1971,

[6] P. P. Caffier, A. I. Nasr, M. d. M. R. Rendon, *et al.*, "Common vocal effects and partial glottal vibration in professional nonclassical singers," *Journal of Voice*, vol. 32, no. 3, pp. 340–346, 2018.

[7] P. Mokhtari, B. Story, P. Alku, and H. Ando, "Estimation of the glottal flow from speech pressure signals: Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production," *Speech Communication*, vol. 104, pp. 24–38, 2018.

[8] S. Duvvuru and M. Erickson, "The effect of change in spectral slope and formant frequencies on the perception of loudness," *Journal of Voice*, vol. 27, no. 6, pp. 691–697, 2013.

[9] Degottex, Gilles, "Glottal source and vocal-tract separation," Ph.D. dissertation, Université Pierre et Marie Curie-Paris VI, 2010.

[10] A. H. Sayed, *Fundamentals of Adaptive Filtering*. John Wiley & Sons, 2003.

[11] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26, no. 1, pp. 20–34, 2012.

[12] A. Barney, A. De Stefano, and N. Henrich, "The effect of glottal opening on the acoustic response of the vocal tract," *Acta Acustica united with Acustica*, vol. 93, no. 6, pp. 1046–1056, 2007.

[13] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *the Journal of the Acoustical Society of America (JASA)*, vol. 90, no. 5, pp. 2394–2410, 1991.

[14] J. Hook, "Smoothing non-smooth systems with low-pass filters," *Physica D: Nonlinear Phenomena*, vol. 269, pp. 76–85, 2014.

[15] J. Kaiser and W. Reed, "Data smoothing using low-pass digital filters," *Review of Scientific Instruments*, vol. 48, no. 11, pp. 1447–1457, 1977.

[16] D. Childers and H. Hu, "Speech synthesis by glottal excited linear prediction," *The Journal of the Acoustical Society of America (JASA)*, vol. 96, no. 4, pp. 2026–2036, 1994.

[17] O. Perrotin and I. McLoughlin, "A spectral glottal flow model for source-filter separation of speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 7160–7164.

[18] K. Gurugubelli and A. K. Vuppala, "Analytic phase features for dysarthric speech detection and intelligibility assessment," *Speech Communication*, vol. 121, pp. 1–15, 2020.

[19] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, no. 4, pp. 403–417, 1997.

[20] S. Torresin, G. Pernigotto, F. Cappelletti, and A. Gasparella, "Combined effects of environmental factors on human perception and objective performance: A review of experimental laboratory works," *Indoor Air*, vol. 28, no. 4, pp. 525–538, 2018.

[21] S. Uthiraa, A. Pusuluri, and H. A. Patil, "Modified group delay features for emotion recognition," in *International Conference on Pattern Recognition and Machine Intelligence (PReMI)*, Springer, 2023, pp. 321–330.

[22] M. J. Alam, P. Kenny, and T. Stafylakis, "Combining amplitude and phase-based features for speaker verification with short duration utterances.," in *INTERSPEECH*, 2015, Dresden, Germany, pp. 249–253.

[23] B. Kashyap, P. N. Pathirana, M. Horne, L. Power, and D. Szmulewicz, "Quantitative assessment of speech in cerebellar ataxia using magnitude and phase based cepstrum," *Annals of Biomedical Engineering*, vol. 48, pp. 1322–1336, 2020.

[24] D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, (ICASSP), vol. 1, 2004, Montreal, Canada, pp. I–125.

[25] L. P. Sahu and G. Pradhan, "Significance of filterbank structure for capturing dysarthric information through cepstral coefficients," in *IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2022, IISc Bengaluru, India, pp. 1–5.

[26] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *2011 IEEE workshop on Automatic Speech Recognition & Understanding*, (ICASSP), 2011, Prague, Czech Republic, pp. 559–564.

[27] L. F. Gallardo, M. Wagner, S. Moller, *et al.*, "Advantages of wideband over narrowband channels for speaker verification employing MFCCs and LFCCs," 2014, Singapore.

[28] H. Kim, M. Hasegawa-Johnson, A. Perlman, *et al.*, "Dysarthric speech database for universal access research.," in *INTERSPEECH*, vol. 2008, 2008, Brisbane, Australia, pp. 1741–1744.

[29] G. Schu, P. Janbakhshi, and I. Kodrasi, "On using the ua-speech and torgo databases to validate automatic dysarthric speech classification approaches," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, Rhodes Island, Greece, pp. 1–5.

[30] M. T. Ahmad, G. Pradhan, and J. P. Singh, "Modeling source and system features through multi-channel convolutional neural network for improving intelligibility assessment of dysarthric speech," *Circuits, Systems, and Signal Processing*, pp. 1–19, 2024, Springer.

[31] S. Yang, Z. Wu, B. Shen, and H. Meng, "Detection of glottal closure instants from speech signals: A convolutional neural network based method.," in *INTERSPEECH*, 2018, Hyderabad, India, pp. 317–321.

[32] D.-H. Shih, C.-H. Liao, T.-W. Wu, X.-Y. Xu, and M.-H. Shih, "Dysarthria speech detection using convolutional neural networks with gated recurrent unit," in *Healthcare*, MDPI, vol. 10, 2022, p. 1956.

[33] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech.," in *INTERSPEECH*, 2015, Dresden, Germany, pp. 2092–2096.

[34] R. Palmer and P. Enderby, "Methods of speech therapy treatment for stable dysarthria: A review," *Advances in Speech Language Pathology*, vol. 9, no. 2, pp. 140–153, 2007.

Contrast-Aware DCT for Image Enhancement with JPEG Compatible Coding

Kohei Hayashi*, Soichiro Honda*, Hirokazu Kamei*, Yoshihiro Maeda†, and Norishige Fukushima*
* Nagoya Institute of Technology, Nagoya, Japan
E-mail: fukushima@nitech.ac.jp
† Shibaura Institute of Technology, Tokyo, Japan
E-mail: ymaeda@shibaura-it.ac.jp

***Abstract*—JPEG is a de facto standard image coding format that is still used today. Although JPEG has a lower coding efficiency than new coding methods such as WebP, HIEF, and JPEG-XL, it has a high rate of adoption; thus, there have been several proposals for improved JPEG coding that maintain backward compatibility, such as Guetzli, MozJpeg, and Jpegli. These improved JPEG codecs optimize the quantization table but do not change the discrete cosine transform (DCT). Recent advances in signal processing have led to various extended DCTs. The use of redundant DCTs, which a type of extend DCTs, has made it possible to achieve a wide range of processing. In this paper, we propose a method for extending the redundant DCT to a contrast-aware process and for achieving DCT suitable for image enhancement while maintaining backward compatibility for JPEG. A naïve implementation requires 64 times DCT per block, but this paper proposes a method for calculating it quickly. The experimental results showed that the coding efficiency improved by 2.57 %, and significantly suppressed coding noise when enhancing.**

I. Introduction

The discrete cosine transform (DCT) [1] is a frequency transform widely used for image coding, denoising, and enhancement. The JPEG coding is the de facto standard for image coding, which uses 8×8 block coding with 8-point DCTs [2], [3]. JPEG [4] is a classic, and its coding efficiency is lower than that of the new such as WebP, HIEF, and JPEG-XL, both subjectively and objectively [5]. However, JPEG is still the de facto standard and highly useful considering the hardware penetration rate and backward compatibility. Therefore, there are many tools to improve the JPEG coding rate so that the decoder circuit is fixed. Guetzli [6], MozJpeg, and Jpegli [7] are typical backward compatible JPEG encoders. These methods increase coding efficiency by optimizing quantization tables.

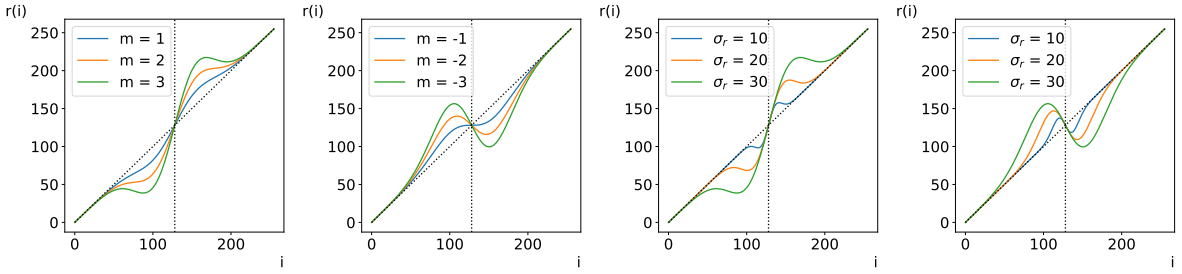
The DCT itself has undergone various expansions, such as the redundant transformation. The redundant transformation, which is spatially or temporally overlapped, is used as modified DCT in speech and is used as redundant frequency denoising [8]–[11] in images, as represented by BM3D denoising [12]. In addition, this redundant transformation is accelerated using a sliding transform [13], [14], and the recursive convolution that it uses is proposed for each DCT definition (-I [15], -III [16], [17], -V [18], [19], -VII [20]), and efficient implementations have been proposed [21]–[23].

This work was supported by JSPS KAKENHI (JP24K02960) and Environment Research and Technology Development Fund (JPMEERF20222M01) of the Environmental Restoration and Conservation Agency of Japan.

Moreover, applying redundant transformations within the range domain enables edge- or contrast-aware signal processing through linear combinations of linear time-invariant processes. The approximate bilateral filters [24]–[27] and the local contrast filters [28], [29] belong to this category. It is established that approximate bilateral filters can be represented as a combination of the original image and local contrast adjustments [30], [31]; therefore, the idea of being aware of contrast and edges is equivalent. In contrast-aware processing, contrast is locally manipulated to emphasize and smoothen the images, and the difference in the intensity of the image is locally changed to perform multiscale processing such as Laplacian pyramid and difference of Gaussian (DoG). The local Laplacian filter (LLF) [28], [30], [32] uses the Laplacian pyramid, and the local DoG filter (LDF) [29] is a DoG version of it. The decomposed multilateral filter (DMF) [33] handles spatial and range domain frequency conversion simultaneously. These processes can suppress extreme responses better than global contrast manipulation. However, when calculated naively, it is necessary to perform redundant multiscale processing, increasing its computational cost. The drawback can be overcome by employing the range domain transformation [28].

In this study, we enhance the DCT to handle contrast awareness by incorporating the concept of manipulating local contrast within the 8×8 DCT framework. Typically, the final step in the camera image processing pipeline is coding, with denoising and enhancement performed beforehand. Nonetheless, applying smoothing as a preprocessing step can lead to losing critical frequency details necessary for coding. Additionally, increasing the frequency through preprocessing enhancement unavoidably generates ringing and mosquito artifacts. If smoothing is applied as a postprocessing step, the filter's effectiveness will decline because the decoder does not have access to an uncorrupted input image, leading to reduced noise suppression ability and an increase in coding noise. If the processing is local-contrast-aware, it is possible to achieve DCT that suppresses ringing, mosquito, and block noises. Moreover, since the inverse DCT (IDCT) is unchanged, it is backward compatible without changing the JPEG decoder. The naïve implementation requires 64 times DCT per block, but it is possible to speed up the process using approximate calculations that convert frequencies to a value domain.

The contributions of this paper are as follows.



(a) enhance (m change) (b) smooth (m change) (c) enhance (σ_r change) (d) smooth (σ_r change)

Fig. 1: Remap functions for enhancement and smoothing with changing m and σ_r ($g = 128$, $m = \pm 3, \sigma_r = 30$, if fixed).

- We introduced contrast-aware DCT (CA-DCT), which allows the forward DCT to consider local contrast.
- CA-DCT has backward compatibility with JPEG.
- CA-DCT can jointly enhance and compress JPEG images.

II. Preliminary

A. DCT

There are eight types of DCT, from I to VIII. Here, we show DCT-II used in JPEG. Let the input signal be $f(x)$ and the output be $F(u)$. $N \in \mathbb{N}$ represents the number of samples.

$$F(u) = \sqrt{\frac{2}{N}} c(u) \sum_{x=0}^{N-1} f(x) \cos \frac{(2x+1)u\pi}{2N}$$
$$c(u) = \begin{cases} \frac{1}{\sqrt{2}} & (u = 0) \\ 1 & (u \neq 0) \end{cases} \quad (1)$$

B. Local Contrast-aware Multiscale Processing

In classical multiscale processing, the multiscale coefficients are increased or decreased to manipulate the detailed information. In contrast, contrast-aware multiscale processing, such as LLF and LDF, directly calculates the coefficients by appropriately remapping to match each coefficient.

In other words, after the contrast around a particular pixel of interest has been enhanced using a remap function, multiscale processing is performed, and only the coefficients for the positions referenced in the remapping are retained. By performing this operation on all layers and all pixels, we obtain multiscale coefficients for the image, with improved contrast at all positions in the image. Therefore, this is a very redundant process because we need to perform multiscale processing of the entire image once to obtain a single coefficient.

The remap function is defined as follows.

$$r(i, g) = i + (i - g) f(i - g), \quad (2)$$

where i and g are the input and reference values, respectively.

The function f is often defined as an even function with decay to suppress halos, and the definition depends on the application. This study uses a Gaussian function with the parameter m that controls the degree of emphasis:

$$f(i) = m \exp\left(\frac{-i^2}{2\sigma_r^2}\right). \quad (3)$$

When $m > 0$, this function emphasizes the image, and when $m < 0$, it smooths the image. In addition, the larger σ, the wider the range of intensities that are emphasized, and the smaller the value, the opposite. Figure 1 shows the remap functions for various values of m and σ_r when $g = 128$.

III. Proposed Method

We introduce a novel forward DCT, named contrast-aware DCT (CA-DCT) and then derive its approximated acceleration. The naïve CA-DCT computes the DCT coefficients after contrast manipulation for 8×8 blocks of the input image for a coefficient of CA-DCT; this processing is redundant. Therefore, we introduce its acceleration using a Fourier series-expanded remap function to reduce the number of DCTs.

A. Naïve Implementation

Let the DCT coefficients of the $W \times H$ image $\boldsymbol{I} \in \mathbb{R}^{W \times H}$ be $C[\boldsymbol{I}] \in \mathbb{R}^{W \times H}$ and the position of the pixel be $\boldsymbol{p} = (x, y)$. The naïve CA-DCT is as follows, and Fig. 2 shows this flow.

(1) $\boldsymbol{I}$ is transformed by the normal DCT per 8×8 block to obtain the coefficients $C[\boldsymbol{I}]$.

(2) Construct an image $\boldsymbol{G} \in \mathbb{R}^{W \times H}$ for the remapping reference value of each position. Here, the input image is used as $\boldsymbol{G}$, but the user can define it as a blurred image, etc.

(3) Repeat (4) for all 8×8 blocks $\mathcal{B}_{\boldsymbol{p}}$ in $\boldsymbol{I}$, which is the set of pixel positions of the block including $\boldsymbol{p}$.

(4) Repeat (a)–(c) for all positions $\boldsymbol{q}$ in the block.

- (a) Referring to the value of $\boldsymbol{G}_{\boldsymbol{q}} \in \mathbb{R}$, apply the remap function to the block where $\boldsymbol{q}$ is located in the input image to create a contrast-transformed block-image $r(\boldsymbol{I}, \boldsymbol{G}_{\boldsymbol{q}}) \in \mathbb{R}^{W \times H}$.
- (b) Transform $r(\boldsymbol{I}, \boldsymbol{G}_{\boldsymbol{q}})$ by DCT and add the result $C[r(\boldsymbol{I}, \boldsymbol{G}_{\boldsymbol{p}})] \in \mathbb{R}^{W \times H}$ to the corresponding block of output coefficients $C[\boldsymbol{O}] \in \mathbb{R}^{W \times H}$.
- (c) The output $C[\boldsymbol{O}]_{\boldsymbol{p}}$ is a division of $C[r(\boldsymbol{I}, \boldsymbol{G}_{\boldsymbol{p}})]$ by $|\mathcal{B}_{\boldsymbol{p}}| = 64$ as an averaged value, but 3 pixels, (0,0), (1,0), (0,1), which are DC and its connected pixels, are copied directly from the DCT coefficients of the input image.

(5) Perform IDCT for $C[\boldsymbol{O}]$ to obtain the output image $\boldsymbol{O} \in \mathbb{R}^{W \times H}$.

The coefficients of the naïve CA-DCT are defined as follows.

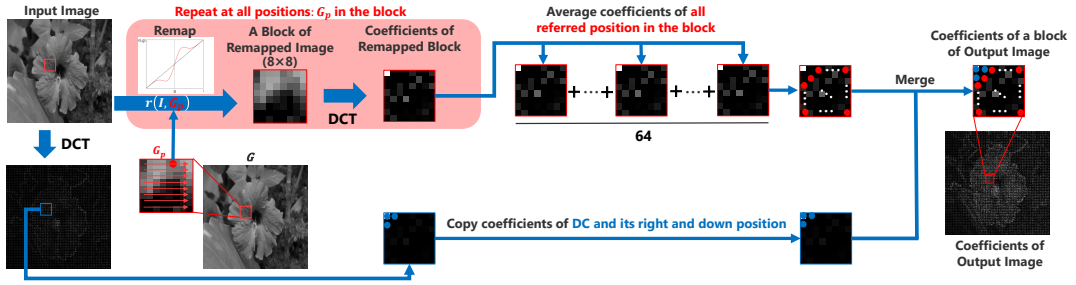


Fig. 2: Overview of naïve implementation of contrast-aware DCT (CA-DCT.)

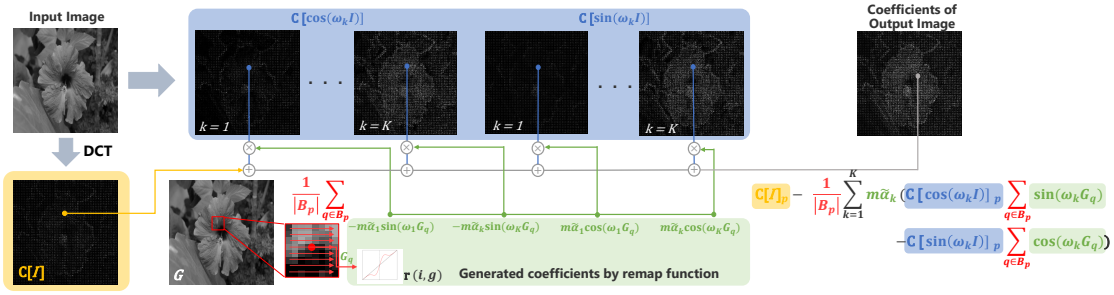


Fig. 3: Overview of approximated implementation of CA-DCT.

$$C[\boldsymbol{O}]_{\boldsymbol{p}} = \begin{cases} C[\boldsymbol{I}]_{\boldsymbol{p}=(x,y)} & (x \bmod 8)+(y \bmod 8) < 2 \\ \frac{1}{|\mathcal{B}_{\boldsymbol{p}}|}\sum_{\boldsymbol{q}\in\mathcal{B}_{\boldsymbol{p}}} C[r(\boldsymbol{I},\boldsymbol{G_q})]_{\boldsymbol{p}} & \text{otherwise} \end{cases} \quad (4)$$

This process is expensive because it performs the 8×8 block DCT for all pixels. However, the calculation is less than that of LLF and LDF, which are processes that involve similar local contrast transformations. While LLF and LDF require remapping and multiscale-processing of the entire image when calculating a single coefficient, the proposed CA-DCT only requires remapping of 8×8 blocks and DCT for these blocks.

B. Approximated Implementation

In this section, we introduce the approximation of the naïve CA-DCT by a Fourier series-expanded remap function to reduce the amount of calculation. The flow of the approximate implementation is shown in Fig. 3.

The weight function f in the remap function is often an even function, and this paper also uses a Gaussian function in Eq. (3); thus, it can be approximated by finite cos functions by Fourier series expansion. The weight function f for remapping an intensity $I_q = [\boldsymbol{I}]_{\boldsymbol{q}} \in \mathbb{R}$ with reference intensity $g_p = [\boldsymbol{G}]_{\boldsymbol{p}} \in \mathbb{R}$ at any position $\boldsymbol{p}, \boldsymbol{q}$ is as follows.

$$f(I_q - g_p) \approx \alpha_0 + 2m\sum_{k=1}^{K} \alpha_k \cos(\omega_k(I_q - g_p)), \quad (5)$$

where $K \in \mathbb{R}$ is the order of approximation, and the larger this value, the more accurate the approximation. $\alpha_k \in \mathbb{R}$ is the k-th order's coefficient and $\omega_k \in \mathbb{R}$ is its angular frequency.

Differentiating this equation, we can approximate the derivative of the weight function as follows:

$$f'(I_q - g_p) \approx 2m\sum_{k=1}^{K} \alpha_k \sin(\omega_k(I_q - g_p))\omega_k. \quad (6)$$

Next, transform the remap function into a form that includes the derivatives of the weight function. Differentiating the weight function and transforming the equation, it matches the two terms of the remap function as follows:

$$\begin{aligned} f'(I_q - g_p) &= -\sigma_r^{-2}(I_q - g_p)f(I_q - g_p), \\ \therefore (I_q - g_p)f(I_q - g_p) &= -\sigma_r^2 f'(I_q - g_p). \end{aligned} \quad (7)$$

Substituting the above equation for the remap function of Eq. (2), the remap function can be expressed as follows:

$$r(I_q, g_p) = I_q - \sigma_r^2 f'(I_q - g_p). \quad (8)$$

Plugging Eq. (6) into (8) shows the approximated function.

$$r(I_q, g_p) \approx I_q - 2m\sigma_r^2 \sum_{k=1}^{K} \alpha_k\omega_k \sin(\omega_k(I_q - g_p)). \quad (9)$$

Moreover, transformed by the additive theorem, the following equation is obtained.

$$\begin{aligned} r(I_q, g_p) \approx I_q - m\sum_{k=1}^{K} \tilde{\alpha}_k \Big(&\cos(\omega_k I_q)\sin(\omega_k g_p) \\ &+ \sin(\omega_k I_q)\cos(\omega_k g_p)\Big) \end{aligned} \quad (10)$$

where $\tilde{\alpha}_k = 2\sigma_r^2\alpha_k\omega_k$.

Substituting this equation into the naïve implementation of Eq. (4), the DCT coefficients of the remapped image can be approximated as follows (the color is related to Fig. 3).

$$\begin{aligned} C[r(\boldsymbol{I},\boldsymbol{G_p})]_p = C[\boldsymbol{I}]_p - \frac{1}{|\mathcal{B}_{\boldsymbol{p}}|}\sum_{k=1}^{K} m\tilde{\alpha}_k \Big(&C[\cos(\omega_k \boldsymbol{I})]_{\boldsymbol{p}} \sum_{\boldsymbol{q}\in\mathcal{B}_{\boldsymbol{p}}} \sin(\omega_k \boldsymbol{G_q}) \\ &- C[\sin(\omega_k \boldsymbol{I})]_{\boldsymbol{p}} \sum_{\boldsymbol{q}\in\mathcal{B}_{\boldsymbol{p}}} \cos(\omega_k \boldsymbol{G_q})\Big) \end{aligned} \quad (11)$$

This equation shows that the DCT to all blocks performed in the approximate implementation is $2K$, i.e., a constant number of times, except for one DCT common to both implementations. Denoting the total number of pixels as N, the DCT to each image block requires $\frac{N}{64}$ times. This means that DCT to one block requires $\frac{NK}{32}$ times for the approximate implementation and N times for the naïve implementation, indicating that the approximate implementation is computationally less expensive when K is less than 32.

IV. Analysis of CA-DCT

First, we demonstrated the signal processing aspect of CA-DCT by image signal enhancing and smoothing and the evaluation of the proposed approximation. We used Intel Core i9-9900K 3.60 GHz (8 cores / 16 threads) and the C++ code compiled by Visual Studio 2022. Trigonometric computations were sped up with table references [34] with subnormal numbers ignoring [35].

A. DCT vs. CA-DCT for Signal Smoothing and Enhancement

We enhanced an image using the proposed CA-DCT and the usual DCT and then compared the outputs. This process for CA-DCT and the usual DCT is represented as follows.

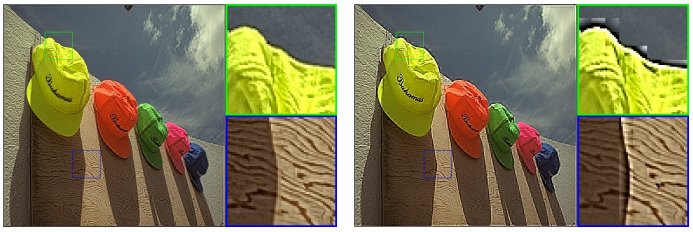
$$\begin{aligned} O^{\text{prop}} &= \text{IDCT}(\text{CA–DCT}(\boldsymbol{I}, m, \sigma_r)), \\ O^{\text{conv}} &= \text{IDCT}(m \cdot \text{DCT}(\boldsymbol{I})), \end{aligned} \tag{12}$$

where O^{prop} and O^{conv} are the outputs for the proposed and usual DCT, respectively. O^{conv}'s DC is not multipied by m.

The proposed enhancement (Fig. 4a) and smoothing (Fig. 5a) and normal DCT enhancement (Fig. 4b) and smoothing (Fig. 5b) are depicted. For normal DCT, the coefficients were multiplied by a constant m, except the DC, that is, the DC is not amplified. CA-DCT suppressed the over-smoothing or enhancement response near steep edges.

B. Approximation Accuracy of CA-DCT

Next, we evaluated the accuracy of the approximate implementation by measuring the peak signal-to-noise ratio (PSNR) between the naïve and approximate output while varying the approximation order K. We tested the case in three different σ_r cases for 10, 20, and 30. The PSNR was averaged over 24 images. Figure 6 shows that the larger the σ_r, the higher PSNR is reached with fewer K. When the PSNR exceeds 59 dB, the two output images are regarded as the same in the 8-bit image case. Focusing on K with PSNR greater than the threshold of 59 dB, K is 14 ($\sigma_r = 10$) , 8 ($\sigma_r = 20$), and 6 ($\sigma_r = 30$). In practical applications, the above K values are expected to provide sufficiently accurate output. This means that the number of DCTs is $\frac{NK}{32}$ to obtain an output equivalent to the naïve; $\frac{7N}{16}$ when $\sigma_r = 10$, $\frac{N}{4}$ when $\sigma_r = 20$, and $\frac{3N}{16}$ when $\sigma_r = 30$, indicating that it requires fewer calculations than the naïve implementation. Although the code is not optimized, the runtime is about 450 msec for the naïve implementation and 150 msec for the approximate implementation.



(a) CA-DCT ($\sigma_r = 30$, $K = 6$, $m = 3.5$) (b) DCT ($m = 3.5$)

Fig. 4: Image enhancement by CA-DCT and DCT



(a) CA-DCT ($\sigma_r = 30, K = 6, m = -0.5$) (b) DCT ($m = 0.5$)

Fig. 5: Image smoothing by CA-DCT and DCT

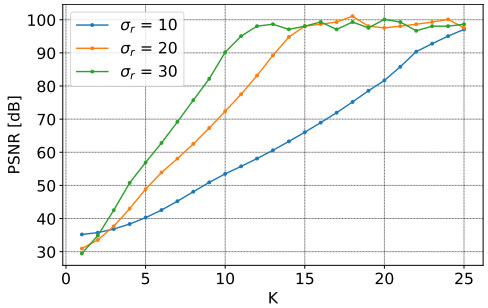


Fig. 6: Accuracy of PSNR between the naïve and approximation for each order K with $\sigma_r \in \{10, 20, 30\}$.

V. Experimental Results of Image Coding

Next, we evaluated the coding performance of CA-DCT for the usual and image enhancement cases.

A. Coding Performance with Quantization

We quantized the coefficients of the proposed CA-DCT and the usual DCT for each quantization parameter $q \in \mathbb{R}$. Also, we used bilateral filtering (with $\sigma_r = 15$ and $\sigma_s = 1$) as prefiltering for the usual DCT as the competitive smoothing. We used the default JPEG quantization table $Q_{u,v} \in \mathbb{N}^{8\times 8}$. Here, u, v are the coordinates in the frequency space, and the input frequency coefficient is $F_{u,v} \in \mathbb{R}^{8\times 8}$. The quantization function $\mathcal{Q}(\cdot) : \mathbb{R}^{8\times 8} \mapsto \mathbb{N}^{8\times 8}$ is defined by the following equation.

$$F^{\text{quant}}_{u,v} = \mathcal{Q}(F_{u,v}) = \text{round}\left(q \cdot \frac{F_{u,v}}{Q_{u,v}}\right) \tag{13}$$

The average PSNR of 24 images was used for evaluation, and bit per pixel (bpp) was computed for each coefficient.

Figure 7 shows the rate distortion curve (PSNR per bpp) of the coefficients of each DCT. CA-DCT always shows 0.1 to 0.3 dB higher PSNR than the usual DCT. CA-DCT slightly smoothed the input image by a local contrast transform to

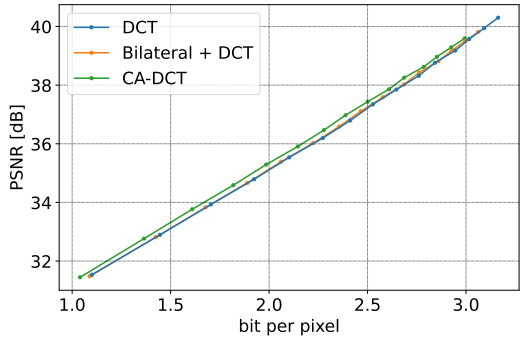


Fig. 7: Coding performance for DCT and CA-DCT ($\sigma_r = 30$, $m = -0.2$, $K = 6$). BD-PSNR = 0.20, BD-rate = -2.57.

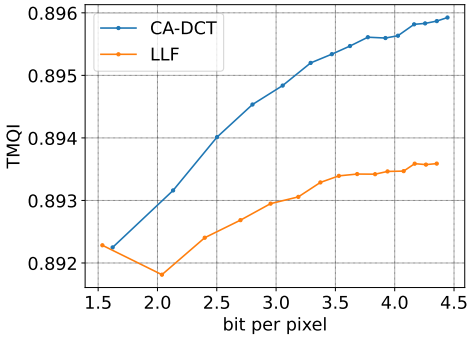


Fig. 8: Coding performance with enhancement evaluated by TMQI ($\sigma_r = 30$, $m = 5$, $K = 6$).

suppress the extreme response at the edges and then quantized it. The proposed PSNR is better than the DCT with bilateral filters, indicating that prior smoothing with the local contrast transform is effective. We also measured Bjøntegaard delta rate (BD-rate) [36] and evaluated the bitrate reduction, and the data was reduced by 2.57% for the same image quality.

Although CA-DCT differs from usual DCT, it can be achieved with multiple DCTs. Since conventional IDCT can be used for the inverse transform, CA-DCT can compress images better than conventional DCT in applications that use DCT, such as image coding, simply by replacing the DCT portion used in the encoding process with CA-DCT.

B. Coding Performance with Enhancement and Quantization

We enhanced and quantized the images using CA-DCT and LLF [32] that simulates the camera image processing pipeline. These processes are defined as follows.

$$\begin{aligned} F_{u,v}^{\mathrm{prop}} &= (\mathcal{Q} \circ \mathrm{CA\text{–}DCT})(\boldsymbol{I}, m, \sigma_r), \\ F_{u,v}^{\mathrm{conv}} &= (\mathcal{Q} \circ \mathrm{DCT} \circ \mathrm{LLF})(\boldsymbol{I}, m, \sigma_r), \end{aligned} \tag{14}$$

where $\mathrm{CA\text{–}DCT}(\cdot)$, $\mathrm{DCT}(\cdot)$, and $\mathrm{LLF}(\cdot)$ are CA-DCT, DCT and LLF functions, respectively, and $\circ$ is the operator for function composition.

We measured the tone-mapped image quality index (TMQI) [37] of the enhanced image and evaluated its naturalness and its structural similarity to the original image. TMQI was averaged from Kodak's 24 images.

Figure 8 shows that the TMQI score of CA-DCT is higher for the same compression ratio. The conventional approach of enhancing with a prefilter before coding is prone to introduce coding noise, and the quality is reduced because the emphasis effect is weakened by quantization. In addition, since CA-DCT can perform both enhancement and quantization at once, it is considered to be computationally less expensive than enhancing an image by some other method and quantizing that image.

VI. Conclusion

In this paper, we propose a contrast-aware DCT (CA-DCT) and a fast approximation method. Since this transformation can be reversed using a normal inverse DCT, it is possible to perform JPEG-compatible coding. In our experiments, we showed a 2.57% improvement in coding performance and also showed that when performed simultaneously with enhancement, it dramatically improves image quality. One limitation of the JPEG coding performance in this paper is that, unlike the conventional backward-compatible JPEG codecs, no processing is performed on the quantization table. However, as it is sufficient as a performance improvement as one of the coding tool blocks and is independent from the optimization of the quantization table, it is necessary to verify it in combination with improvements in other tool groups.

References

[1] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[2] Y. Arai, T. Agui, and M. Nakajima, "A fast dct-sq scheme for images," *IEICE Transactions*, vol. E-71, no. 11, pp. 1095–1097, 1988.

[3] C. Loeffler, A. Ligtenberg, and G. S. Moschytz, "Practical fast 1-d dct algorithms with 11 multiplications," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1989.

[4] G. K. Wallace, "The jpeg still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[5] S. Honda, Y. Maeda, and N. Fukushima, "Dataset of subjective assessment for visually near-lossless image coding based on just noticeable difference," in *Proc. International Conference on Quality of Multimedia Experience (QoMEX)*, 2023, pp. 236–239.

[6] J. Alakuijala, R. Obryk, O. Stoliarchuk, Z. Szabadka, L. Vandevenne, and J. Wassenberg, "Guetzli: Perceptually guided jpeg encoder," *arXiv:1703.04421*, 2017.

[7] M. Bruse, L. Versari, Z. Szabadka, and J. Alakuijala, "Users prefer jpegli over same-sized libjpeg-turbo or mozjpeg," *arXiv:2403.18589*, 2024.

[8] G. Yu and G. Sapiro, "Dct image denoising: A simple and effective image denoising algorithm," *Image Processing On Line*, vol. 1, p. 1, 2011.

[9] S. Fujita, N. Fukushima, M. Kimura, and Y. Ishibashi, "Randomized redundant dct: Efficient denoising by using random subsampling of dct patches," in *Proc. SIGGRAPH Asia, Technical Briefs*, 2015.

[10] Y. Kawasaki, Y. Maeda, and N. Fukushima, "Parallelized and vectorized implementation of dct denoising with fma instructions," in *Proc. International Workshop on Advanced Image Technology (IWAIT)*, 2018.

[11] N. Fukushima, Y. Kawasaki, and Y. Maeda, "Accelerating redundant dct filtering for deblurring and denoising," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2019.

[12] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[13] J. Xi and J. Chicharo, "Computing running dct's and dst's based on their second-order shift properties," *IEEE Transactions on Circuits and Systems I Fundamental Theory and Applications*, vol. 47, no. 5, pp. 779–783, 2000.

[14] V. Kober, "Fast algorithms for the computation of sliding discrete sinusoidal transforms," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1704–1710, 2004.

[15] T. Yano, K. Sugimoto, Y. Kuroki, and S. Kamata, "Acceleration of gaussian filter with short window length using dct-1," in *Proc. APSIPA Annual Summit and Conference*, 2018.

[16] E. Elboher and M. Werman, "Cosine integral images for fast spatial and range filtering," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2011.

[17] D. Charalampidis, "Recursive implementation of the gaussian filter using truncated cosine functions," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3554–3565, 2016.

[18] K. Sugimoto and S. Kamata, "Fast gaussian filter with second-order shift property of dct-5," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2013.

[19] K. Sugimoto and S. Kamata, "Efficient constant-time gaussian filtering with sliding dct/dst-5 and dual-domain error minimization," *ITE Transactions on Media Technology and Applications*, vol. 3, pp. 12–21, 2015.

[20] K. Sugimoto, S. Kyochi, and S. Kamata, "Universal approach for dct-based constant-time gaussian filter with moment preservation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[21] N. Fukushima, Y. Maeda, Y. Kawasaki, *et al.*, "Efficient computational scheduling of box and gaussian fir filtering for cpu microarchitecture," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018.

[22] T. Otsuka, N. Fukushima, Y. Maeda, K. Sugimoto, and S. Kamata, "Optimization of sliding-dct based gaussian filtering for hardware accelerator," in *Proc. International Conference on Visual Communications and Image Processing (VCIP)*, 2020.

[23] Y. Kanetaka, H. Takagi, Y. Maeda, and N. Fukushima, "Slidingconv: Domain-specific description of sliding discrete cosine transform convolution for halide," *IEEE Access*, vol. 12, pp. 7563–7583, 2024.

[24] K. Sugimoto and S. Kamata, "Compressive bilateral filtering," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3357–3369, 2015.

[25] K. Sugimoto, N. Fukushima, and S. Kamata, "Fast bilateral filter for multichannel images via soft-assignment coding," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2016.

[26] G. Deng, "Fast compressive bilateral filter," *Electronics Letters*, vol. 53, pp. 150–152, 2017.

[27] Y. Sumiya, N. Fukushima, K. Sugimoto, and S. Kamata, "Extending compressive bilateral filtering for arbitrary range kernel," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2020.

[28] Y. Sumiya, T. Otsuka, Y. Maeda, and N. Fukushima, "Gaussian fourier pyramid for local laplacian filter," *IEEE Signal Processing Letters*, vol. 29, pp. 11–15, 2022.

[29] K. Hayashi, Y. Maeda, and N. Fukushima, "Local contrast enhancement with multiscale filtering," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 765–770.

[30] M. Aubry, S. Paris, J. Kautz, and F. Durand, "Fast local laplacian filters: Theory and applications," *ACM Trans. on Graphics*, vol. 33, no. 5, 2014.

[31] T. Tsubokawa, H. Tajima, Y. Maeda, and N. Fukushima, "Local look-up table upsampling for accelerating image processing," *Multimedia Tools and Applications*, 2023.

[32] S. Paris, W. Hasinoff, and J. Kautz, "Local laplacian filters: Edge-aware image processing with a laplacian pyramid," *ACM Trans. on Graphics*, vol. 30, no. 4, 2011.

[33] H. Nogami, Y. Kanetaka, Y. Naganawa, Y. Maeda, and N. Fukushima, "Decomposed multilateral filtering for accelerating filtering with multiple guidance images," *Sensors*, vol. 24, no. 2, 2024. DOI: 10.3390/s24020633.

[34] Y. Maeda, N. Fukushima, and H. Matsuo, "Taxonomy of vectorization patterns of programming for fir image filters using kernel subsampling and new one," *Applied Sciences*, vol. 8, no. 8, 2018.

[35] Y. Maeda, N. Fukushima, and H. Matsuo, "Effective implementation of edge-preserving filtering on cpu microarchitectures," *Applied Sciences*, vol. 8, no. 10, 2018.

[36] G. Bjøntegaard, "Calculation of average psnr differences between rd-curves," *ITU-T SG16*, vol. Doc. VCEG-M33, 2001.

[37] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 657–667, 2012.

Dysarthria Severity Classification Using Phase Based Features of LP Residual

Rohini Sri Mannepalli[1] , Aditya Pusuluri[2] and Hemant A.Patil[3]
Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana
E-mail: 2210040095@klh.edu.in
Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat
E-mail:{aditya_pss,hemat_patil}@daiict.ac.in

***Abstract*—Classifying the severity of speech impairment due to dysarthria is crucial for optimizing care and enhancing communication abilities for affected individuals. This study explores the use of the Modified Group Delay Function (MGDF) of LP residual signal in classifying dysarthria severity-levels. Evaluations were conducted using standard UA-Speech and TORGO datasets. A stratified Convolutional Neural Network (CNN) with *5*-fold cross-validation validated the results. Baseline features included Linear Frequency Cepstral Coefficients (LFCC), Mel Frequency Cepstral Coefficients (MFCC), and Whisper module. Key performance evaluation metrics were accuracy, precision, recall, and F1-score. Finally, the latency period was analyzed for practical deployment of the system, system's ability to accurately recognize and process speech from any speaker, without needing to be specifically trained or adapted to individual voice characteristics.**

Keywords: Dysarthria, LP residual, Modified Group Delay Function.

I. Introduction

Dysarthria is a motor speech disorder characterized by impaired movement of the muscles used for speech, resulting in slurred, slow, or difficult-to-understand speech. This condition presents significant challenges in clinical assessment and management due to its heterogeneous nature and varying dysarthric severity-levels. Accurate classification of dysarthria severity is essential for guiding treatment planning and monitoring disease progression. Traditional methods of severity classification, often relying on subjective analysis, are expensive and time-consuming, highlighting the need for more efficient and objective approaches [1].

Early approaches in dysarthria classification utilized acoustic features derived from fundamental or pitch frequency (F_{o}), formant frequencies, and duration measures [2]. However, these traditional features often lack the sensitivity and specificity required for precise severity-level classification, particularly in subtle or nuanced speech impairments. Researchers have found cepstral-based features, such as Mel Frequency Cepstral Coefficients (MFCC), [3] effective for classifying dysarthria severity due to their ability to capture the vocal tract system characteristics. Studies have shown that combining MFCC with auditory features yields better classification results.

This study proposes the use of Modified Group Delay Cepstral Coefficients (MGDCC) for dysarthric severity-level classification [4], leveraging the phase information of linear prediction (LP) residual to capture the characteristics of dysarthric speech. By integrating features, such as MFCC, Linear Frequency Cepstral Coefficients (LFCC), and Linear Frequency Residual Cepstral Coefficients (LFRCC) with MGDCC, [5] the proposed approach aims to enhance the reliability and accuracy of automated dysarthria severity-level classification systems.

The organization of the rest of the paper is as follows as: Section II presents details of proposed phase-based features from LP residual. Section III gives details of experimental setup used for this study. Section IV presents experimental results evaluating the proposed features with various performance evaluation factors. Finally, Section V concludes the paper along with potential future research directions.

II. Proposed Methodology

A. LP Residual

Speech signal is a one-dimensional signal carrying information in the form of dependencies in the sequence of samplesof speech signal and thus, a single speech sample has no perceputal significance. Hence, LP analysis models speech signals as the result of a linear combination of past samples, aiming to predict the current speech sample. The residual represents the difference between the actual signal and the predicted signal. This residual signal contains information about the components of the speech signal that are not well-predicted by the LP model, typically capturing aspects, such as aspiration noise in consonants, glottal pulses in voiced speech, and other high frequency components [6] . In particular,

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k). \tag{1}$$

The parameter p, which denotes the filter order, significantly influences performance in speech recognition tasks, with optimal results achieved using LP order 10. Residual error plays a crucial role in validating and debugging LP models, guiding iterative optimization algorithms, and refining solutions to improve their quality. By analyzing residual errors, one can identify and correct issues within the model, ensuring more accurate and reliable outcomes. It is given by:

$$r(n) = s(n) - \hat{s}(n) \tag{2}$$

LP residual is a versatile tool in speech signal processing, valued for its ability to capture prediction errors and excitation source details that contribute to the accurate representation and analysis of speech signals in various applications.

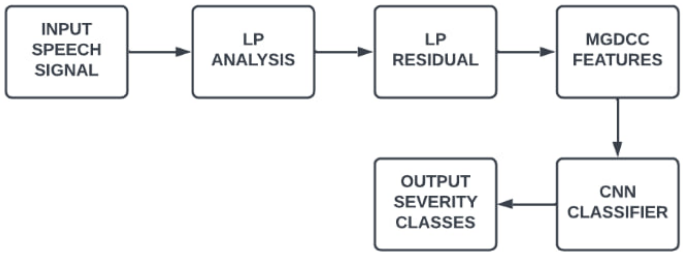


Fig. 1: Functional block diagram of LP residual and MGDCC feature for the dysarthric severity-level classification system.

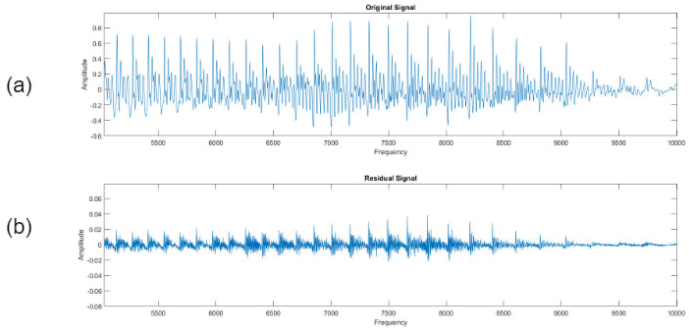


Fig. 2: (a) depicts original signal and (b) depicts MGDCC(lp) of that signal

B. Modified Group Delay Function

Speech signals are complex, characterized by fundamental components, such as frequency, amplitude, and spectral envelope. The vocal tract system, which is generally a maximum phase system, plays a crucial role in shaping these signals by influencing acoustic resonances, known as *formants*, which are essential for distinguishing vowels and certain consonants [7]. The glottal excitation source, with its nonlinear vibration, contrasts with the vocal tract's minimum phase behavior. Analyzing phase provides insights into the temporal coherence and stability of speech components, aiding in tasks, such as speech synthesis, enhancement, and recognition [8]. Identifying resonance frequencies or formant frequencies from the phase spectrum presents a challenge because they become obscured by the phase wrapping phenomenon occurring at multiples of 2π. To overcome this signal processing conflict, the signal must be a minimum phase signal, where the continuous (i.e., unwrapped) phase function is denoted by $\theta(e^{j\omega})$. Minimum phase signals are preferred because their magnitude spectrum and group delay spectrum exhibit similar α characteristics. The group delay function derived as the negative derivative of the unwrapped Fourier transform phase, serves as a measure to quantify this *coherence*. The group delay function is given as,[8]

$$T(e^{j\omega}) = -\frac{d\theta(e^{j\omega})}{d\omega}. \tag{3}$$

The group delay function is specifically applicable to minimum phase signals. However, speech signals are generally mixed-phase systems that contain zeros introduced by noise or nasal sounds. To address these undesirable or spurious spikes, the Modified Group Delay Function (MODGF) is introduced, which effectively mitigates the influence of these zeros near the unit circle in the group delay spectrum of speech signals. The Modified Group Delay Function (MODGF) effectively moves the zeros radially inside the unit circle, thereby reducing the occurrence of spikes in the valleys. Additionally, a cepstrally-smoothed signal is introduced to restore the dynamic range and mitigate the spiky structure of the phase-based features.The modified group delay function is given as:

$$T_m(\omega) = \frac{T(\omega)}{|T(\omega)|}[T(\omega)]^a, \tag{4}$$

where $T(\omega)$ is given by,

$$T(\omega) = \frac{G_R(\omega)H_R(\omega) + G_I(\omega)H_I(\omega)}{|S(\omega)|^{2\gamma}}, \tag{5}$$

where $S(\omega)$ represents the cepstrally-smoothed version of $G(\omega)$, and G_R, G_I, H_R, and H_I indicate the real and imaginary parts, respectively. Two parameters α and γ are introduced , which are used to restore the dynamic range and reduce the amplitude of the unwanted spikes, respectively. The range of α and γ range between $0 < \alpha \leq 1$ and $0 < \gamma \leq 1$.

III. Experimental Setup

A. Dataset Used

This study utilizes two well-known dysarthria speech corpora, namely, the Universal Access Dysarthria Speech Corpus (UA-Speech) and TORGO. Both corpora predominantly exhibit spastic dysarthria, characterized by features, such as breathiness, hypernasality, a harsh voice, and incorrect articulation, which lead to unintelligible speech [9]. The UA-Speech corpus is segmented into four severity-levels: 930 samples of very low severity, 926 samples of low severity, 930 samples of medium severity, and 751 samples of high severity. It includes 8 speakers: 4 males (M01, M05, M07, and M09) and 4 females (F02, F03, F04, and F05). The TORGO corpus comprises a total of 1982 samples distributed across three severity-levels: 671 samples of very low severity, 627 samples of low severity, and 684 samples of medium severity [10]. In both corpora, 80% of the data is allocated for training purposes, while the remaining 20% is reserved for testing. The division ensures that both training and testing sets include

words, non-words, and sentences. The experiments employ a *5* cross-validation (CV) approach, focusing on evaluating model performance exclusively on the training data to assess its robustness and speaker-independence.

B. Classifier Used

The research employed a Convolutional Neural Network (CNN) classifier due to its inherent translation invariance, essential for recognizing features irrespective of their *spatial* location. Moreover, CNNs efficiently capture spectro-temporal patterns and variations in speech signals, making them ideal for robust and accurate classification in such audio processing tasks. The model was trained using a stratified *5* cross-validation approach with a set seed value to ensure consistent data distribution across folds. Each fold involved an 80% training set and a 20% validation set. The adam optimizer was employed with categorical cross-entropy as the loss function and accuracy as the evaluation metric. A grid search was performed to optimize the learning rate and batch size over 100 epochs, ensuring that the model's parameters were fine-tuned for peak performance. Two activation functions were utilized: ReLU and softmax. ReLU enhances learning speed and reduces computational cost, applied throughout except at the final layer, where softmax aids multi-class classification. Each convolutional layer incorporated normalization and dropout layers to curb overfitting. Fine-tuning parameters yielded a learning rate of 0.001, batch size of 128, and 100 epochs.

C. Baseline Features

The baseline for this work includes features, such as MFCC, LFCC, LFRCC, and WSPSR (Web- scale Supervised Pretraining). Additionally, the Whisper Tiny model, is used as a baseline for the task of dysarthria severity-level classification. This Tiny model is the smallest variant, featuring relatively a fewer trainable parameters and layers compared to its counterparts. For both datasets, LFCC features outperform MFCC features, highlighting the effectiveness of the linear frequency scale for classifying dysarthria severity.

IV. Experimental Results

This Section evaluate the proposed LP residual MGDCC feature set for various experimental evaluation factors, such as the parameter tuning, spectrographic analysis, comparision of baseline features and analysis of latency period.

A. Spectrographic Analysis

The spectrograms collectively illustrate varying degrees of spectral complexity via pattern of spectral energy distribution. Exhibit predominantly low-frequency content with a minimal variation, while the others display a richer spectrum with multiple frequency components. In particular, reveals a wide band of frequencies, indicative of a potentially noisy or intricate signal.

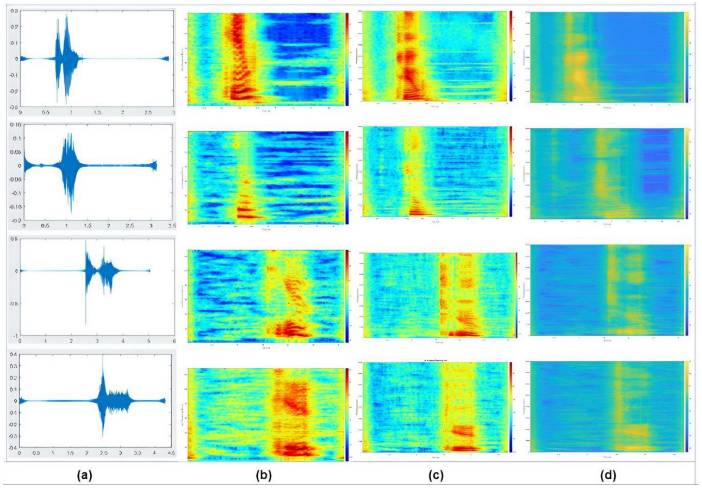


Fig. 3: Plots for dysarthric speech "to" from UA-Speech:Fig. 3(a), Fig. 3(b), Fig. 3(c), Fig. 3(d), of each section depicts the time-domain waveform, Mel Spectrogram, residual spectrogram, and modified group dealy gram, respectively.

B. Parameter Tuning

As discussed in Eq. (4) and Eq. (5), MGDCC consists of two parameters, namely, α and γ that are fine-tuned using a greedy search algorithm and are varied within the range [0.1,1] with a step size of 0.1. The evaluation is performed using a CNN classifier, and a *5*-fold cross-validation accuracy metric. MGDCC features that are tuned for α = 0.1 and γ = 0.1 results in relatively optimum performance for both UA-Speech and TORGO with fold (test) accuracies of 94.63% (93.45%), and 94.63% (93.37%), respectively, indicating α and γ are the generalized parameters for the dysarthria severity-level classification task w.r.t Eq. (4) and Eq. (5).

C. Comparison with Baseline Features

Table 1 shows the relative comparision of LFRCC feature set along with baseline features, such as MFCC, LFCC, and state-of-the-art whisper model-based features using CNN classifier.

For both the datasets, LFCC outperforms MFCC features, indicate effectiveness of linear frequency scale for the task of dysarthria severity-level classification. LFRCC outperforms both MFCC and LFCC baseline features by a fold (test) accuracy margin of 4.72% (4.23%), 0.51% (1.98%) for UA-Speech, and 5.72%(2.81%), 5.58% (2.38%) for TORGO. Furthermore, state-of-the-art web-scale supervised pre-training for speech recognition (WSPSR), also known as Whisper encoder module is used as a baseline for dysarthria severity-level classification task. The proposed LFRCC features perform on par for UA-Speech and outperforms by fold (test) accuracy of 0.46% (0.2%) for TORGO, when compared with

TABLE I: Fold (test) Accuracy, Precision (P), Recall (R) for Various Feature Sets using CNN classifier on UA-Speech and TORGO

Data	Features	Fold Acc.	Test Acc.	Precision	Recall
UA-Speech	MFCC	87.29	90.68	91.52	90.21
	LFCC	91.50	92.93	93.61	92.41
	Whisper	92.01	94.80	93.28	92.44
	LFRCC	92.01	**94.91**	**95.06**	94.85
	Proposed Features MGDCC (LP)	**93.07**	94.63	94.35	**94.65**
TORGO	MFCC	85.71	88.62	89.40	88.64
	LFCC	90.58	91.82	91.51	91.02
	Whisper	90.97	94.10	93.47	94.00
	LFRCC	91.43	**94.20**	93.96	**94.19**
	Proposed Features MGDCC (LP)	**94.95**	93.45	**94.16**	93.51

whisper model, which is an advance machine learning model trained using labelled data of 680,000 hours. Table I indicate the classwise precision, recall, and F1-Score of LFRCC with p = 10 for UA-Speech and TORGO database. We notice the balanced Fl-score across all the severity-levels reflects the model's robustness and consistency in classification performance.

D. Analysis of Latency Period

Analyzing the latency period helps us to identify the minimum number of speech frames, i.e, minimal speech duration required to achieve optimal classification accuracy. By understanding the relationship between the number of frames and the accuracy, we can fine-tune our models to be both efficient and effective, ensuring that we do not process more data than needed while still maintaining high accuracy in our speech classification tasks. Fig.3 the performance metrics

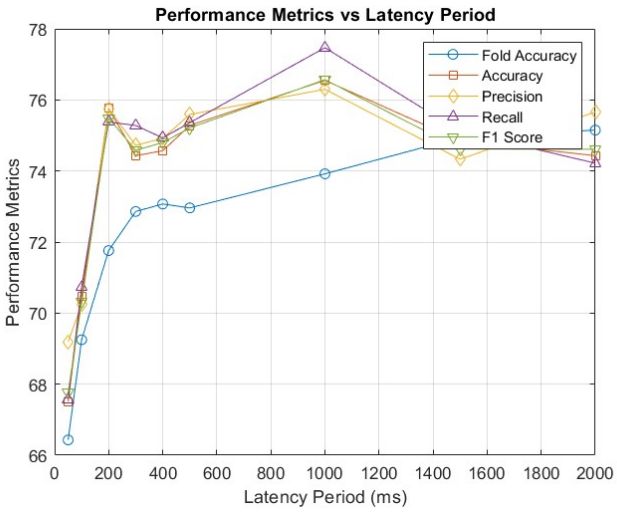


Fig. 4: Latency Period Analysis for different frames of UA-Speech

(fold accuracy, accuracy, precision,recall, and F1 score) against different latency periods (50,100, 200, 300, 400, 500, 1000, 1500, 2000ms). Observing the output plot allows one to see how these performance metrics change as latency increases, reflects how responsive the proposed features are. Typically, it can highlight the trade-offs between quicker decision-making (lower latency) and the performance of the model.

V. Summary And Conclusions

The study suggested utilizing LP residual-based MGDCC features to classify dysarthria severity-levels, using two established dysarthria speech corpora, namely, UA-Speech and TORGO. Through various experiments, we discovered important insights demonstrating the effectiveness of MGDCC features for classifying dysarthric severity. Incorporating phase-based features into a modified group delay function for dysarthric speech involves enhancing the analysis of speech signals by integrating phase information to improve recognition and intelligibility. The group delay function, which typically analyzes phase variations with frequency, is adjusted to better capture the unique spectral and phase distortions present in dysarthric speech.

A comparative analysis with baseline features, such as MFCC and MGDCC, revealed that LFCC features are superior for classifying dysarthria severity-levels. These findings were especially noteworthy because LFCC features exhibited performance at par with the advanced Whisper model features, highlighting their effectiveness in capturing dysarthria-specific characteristics.

Further analysis of the latency period illustrates how varying latency periods (50 ms to 2000 ms) affect performance metrics like accuracy, precision, recall, and F1 score. Shorter latencies may speed up decision-making but could reduce accuracy, while longer latencies often improve accuracy at the cost of increased processing time. The plot helps balance decision speed with model performance to find an optimal latency period. The computational cost typically increases with more complex features and shorter latency periods. Systems optimized for real-time, low-latency operations must often invest in better resources, while those with higher latency tolerance can reduce costs by using more relaxed processing strategies.

One limitation of this work is the lack of direct comparison with current state-of-the-art deep learning techniques. This limitation is acknowledged due to constraints such as resource availability or differences in the scope and application of the study. While the primary focus has been on developing and validating a novel approach, future work should include comprehensive benchmarking against leading deep learning methods to more thoroughly assess the relative performance and generalizability of the proposed solution. Despite this, the results presented offer significant value, particularly in specialized scenarios where the state-of-the-art may not be directly applicable.

VI. Acknowledegements

The authors sincerely thank the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, for sponsoring a consortium project titled 'BHASHINI', (Grant ID : 11(1)2022-HCC (TDIL)). They also thank the

authorities of DA-IICT Gandhinagar for their kind support and cooperation in carrying out this research work.

REFERENCES

[1] J. R. Duffy *et al.*, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Elsevier Health Sciences, 2012.

[2] L. P. Sahu and G. Pradhan, "Significance of filterbank structure for capturing dysarthric information through cepstral coefficients," IISc Banglore, India, 2022, pp. 1–5.

[3] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease," *The journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[4] H. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," vol. 1, May 2003, pp. I–68.

[5] S. Sajiha, K. Radha, D. Venkata Rao, N. Sneha, S. Gunnam, and D. Bavirisetti, "Automatic dysarthria detection and severity level assessment using CWT-layered cnn model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, Jun. 2024.

[6] Srivastava, "Fundamentals of linear prediction," 1999.

[7] B. Yegnanarayana, "Group delay spectrogram of speech signals without phase wrapping," *The Journal of the Acoustical Society of America(JASA)*, vol. 151, no. 3, pp. 2181–2191, 2022.

[8] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *speech communication*, vol. 10, no. 3, pp. 209–221, 1991.

[9] H. Kim, M. Hasegawa-Johnson, A. Perlman, *et al.*, "Dysarthric speech database for universal access research.," in *Interspeech*, vol. 2008, 2008, pp. 1741–1744.

[10] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, pp. 523–541, 2012.

Significance of Entropy Based Features For Dysarthric Severity Level Classification

Meghana Avula[1], Aditya Pusuluri[2]and Hemant A. Patil[2]
[1] Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana
E-mail: 2210040091@klh.edu.in
[2] Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat
E-mail: aditya_pss@daiict.ac.in, hemant_patil@daiict.ac.in

***Abstract*—Dysarthria is a motor speech disorder arising from impairment of muscles that makes it difficult to form or pronounce words while speaking. In this paper, we introduce an approach of multiband entropy based features extracted from dysarthric speech signals for dysarthric severity level classification. Generally, entropy is measured as the number of bits of information contained in each message signal. The information content of these signals measures how much randomness or uncertainty contained in a signal. Extending this, we use a frame-wise processing technique to divide the speech signal into short frames, which allows for detailed analysis of different characteristics of the speech signal. Furthermore, we divide frames into equal sub bands and compute the entropy and zero mean entropy in each sub band using the Gabor filterbank. It is expected that the mean entropy of very low dysarthric severity is lower compared to high dysarthric severity level indicating that randomness is increasing as the severity level of dysarthria is increasing. Experimental analysis were conducted on extensively used datasets namely UA Speech and TORGO. Results were carried out by Convolutional Neural Network (CNN), along with 5-cross validation. The results are compared against standard MFCC, LFCC, and glottal source based LFRCC. It was observed that the addition of entropy information boosted the performance of MFCC by 4.38%, LFCC by 2.54%, and LFRCC by 1.88% compared to their traditional techniques, indicating the crucial information captured by the entropy for dysarthria severity classification.**

Keywords: Dysarthria, Entropy, Gabor Filterbank

I. Introduction

Dysarthria is a speech disorder caused by impairments of neurons that affect the muscles necessary for speech production. These impairments can be caused by many conditions, such as stroke, traumatic brain injury, and Parkinson's disease [1]. Dysarthria is manifested by a variety of speech abnormalities, such as slurred speech, decreased intelligence, changes in speech rate, and changes in speech quality [2]. Given the diverse etiology and presentation of dysarthria, accurately classifying the severity of the disorder is essential for developing tailored treatment plans, monitoring patient progress, and optimizing clinical resource allocation. Traditional methods of assessing dysarthria severity often rely on subjective evaluations by clinicians , which can lead to variability and inconsistency. These assessments depend on the clinician's experience and judgment, which can cause differences in diagnosis and treatment outcomes. This demands the need for automatic speech recognizers (ASR), and dysarthria severity level classification can improve the performance of an ASR.

Earlier approaches focused on employing acoustic characteristics such as features derived from fundamental or pitch frequencies, formant frequencies and cepstral based techniques like Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC) for speech pathology classification tasks. The ability of fundamental frequency and formant frequency for dysarthria severity classification is showcased in [3]. Furthermore, in [4], [5], and [6] the cepstral based features such as MFCC, LFCC are explored, which contain the ability to capture global spectral characteristics.

Entropy measures the randomness or complexity of a signal, making it useful for identifying the detailed variations in speech patterns seen in different severity levels of dysarthria. By looking at the entropy of speech signals, we can better understand the severity of the disorder, as the entropy metric provides information related to the amount of randomness present in the signal, which can help us learn about the controllability of the voice production system for a dysarthric speaker and provide a more accurate classification. Ultimately, this approach aims to provide reliable and effective tools for assessing and managing dysarthria, improving the quality of life for individuals affected by the disorder. By leveraging the power of entropy-based features, we can achieve more accurate and consistent severity classification, leading to better treatment outcomes and enhanced communication abilities for those with dysarthria. Previously, entropy based features were used for various applications such as speech recognition [7], emotion recognition [8], speech/music segmentation [9]. Furthermore, [10] showcased the importance of entropy based features for detection of hypernasal speech, emphasizing the usage of entropy based features for dysarthric speech. To the best of our knowledge, no prior research has explored the ability of multiband entropy for dysarthria severity classification.

The rest of the paper is organized as follows: Section 2 represents a brief technical details about the proposed zero mean entropy using Gabor filter bank, Section 3 gives the details about the database used, the classifiers used, and the baseline features considered for this work. Section 4 consists of the motivation for the application of entropy features for dysarthric severity-level classification tasks. Section 5 contains

the experimental results and the work is concluded with results and conclusion along with potential future research direction.

II. Proposed Approach

A. Shannon Entropy

Entropy is used to separate the voiced and unvoiced regions of the speech signal. The entropy is low for voiced region and high for unvoiced region. Since speech is a non stationary signal, the voiced and invoiced regions vary across each frame. This motivated us to use the multiband entropy for each frame as a feature set instead of calculating the entropy across the entire signal. Entropy measures, such as Shannon entropy, quantify the unpredictability or randomness in the distribution of signal amplitudes over time [11]. In speech analysis, entropy reflects the variability and complexity of vocal characteristics, including pitch, intensity, and timing. Higher entropy values indicate greater disorder or irregularity in speech production, which can be indicative of dysarthric speech patterns characterized by inconsistent articulation and prosody. Entropy H(X), is mathematically expressed as:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

where p(x) denotes the probability of the random variable X taking on the value x In the context of information theory and speech analysis, negative entropy values are not typically encountered, as they would imply a negative amount of uncertainty. Therefore, in practical applications like speech signal analysis, entropy values are non-negative.

Entropy is calculated on speech signals by filtering them through a Gabor filter bank, extracting the linear predictive residual, and then computing the entropy of the filtered signal. *Zero-mean entropy is chosen in order to emphasize the variability of entropy values around the mean value.* It is calculated by subtracting the mean entropy of each frame from its individual entropy values, providing a normalized measure of entropy variation across frames [12]. It can be calculated by,

$$ZeroMeanEntropy = H(X) - \frac{1}{n}\sum_{i=1}^{n} H(x_i)$$

In this process of entropy calculation, Gabor filter banks are used for their ability to capture both frequency and temporal information from speech signals, which is crucial for analyzing dysarthric speech [13]. They decompose the signal into various subbands, allowing the extraction of fine-grained spectral features. *This multi-resolution analysis helps in capturing the small patterns and variations in the speech signal, making it easier to identify and quantify the irregularities associated with dysarthria.* By using a Gabor filter bank, the analysis becomes more robust and comprehensive, as it effectively highlights the relevant spectro-temporal characteristics that are essential for accurate entropy measurement and subsequent classification of dysarthric severity levels [14]. Furthermore, fusion of traditional baseline features and zero mean multi band entropy based features are performed. Fig. 1 contains the block diagram of the proposed feature set.

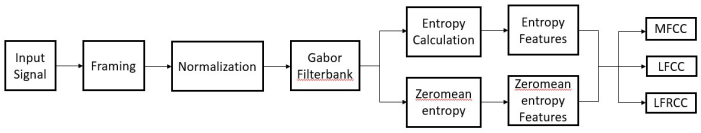


Fig. 1. Block Diagram of Proposed Architecture

The entropy and zero mean entropy features are extracted using 40 Gabor filterbanks while segmenting the speech signal using 20 ms window length with 10 ms hop length. These overlapping windows ensured effective temporal feature capture. Entropy values were computed for each window, normalized to achieve zero-mean entropy, and then arranged in descending order to prioritize the most significant features, enhancing the representation for subsequent processing or classification. Furthermore, the top 20 and bottom 20 features from the ordered sub bands are analyzed separately for entropy and zero mean entropy.

By integrating entropy features with MFCC, LFCC, and LFRCC, we enhance the analysis of dysarthric speech. Entropy measures the unpredictability or variability in the speech signal, which, when combined with these cepstral coefficients, improves the sensitivity to different spectral characteristics. This integration enables a more accurate assessment of speech impairments, making it possible to better differentiate and analyze variations in dysarthric speech.

III. Experimental Setup

A. Dataset Used

In this study, we use two well-known datasets of dysarthric speech: the Universal Access Dysarthria Speech Corpus (UA-Speech) and the TORGO corpus. These datasets mainly feature speech from individuals with spastic dysarthria, characterized by symptoms such as breathiness, hypernasality, a harsh voice, and incorrect articulation, making the speech difficult to understand. The UA-Speech corpus includes recordings from eight speakers (four males: M01, M05, M07, and M09; and four females: F02, F03, F04, and F05). Each speaker contributes 465 utterances, selected from a total of 735 available utterances per speaker. The UA corpus, contains 3537 speech samples categorized into four severity levels: 930 samples of very low severity, 926 samples of low severity, 930 samples of medium severity and 751 samples of high severity. The TORGO corpus, on the other hand, contains 1982 speech samples categorized into three severity levels: 671 samples of very low severity, 627 samples of low severity, and 684 samples of medium severity. For both datasets, we use 80% of the data for training and 20% of the data for testing. The training and testing sets are carefully split to ensure both contain a variety of words,

non-words, and sentences, providing a comprehensive assessment of the model's performance.We employ a 5-fold cross-validation (CV) approach in our experiments. This involves dividing the training data into five subsets, training the model on four of these subsets, and validating it on the remaining subset. This process is repeated five times, with each subset used once for validation.

B. Classifier Used

The experiments used a Convolutional Neural Network (CNN) classifier due to its effectiveness in capturing essential spectro-temporal patterns in speech signals and its ability to maintain translation invariance. The model was trained with a stratified 5-fold cross-validation (CV) approach. This method ensures that each fold has a data distribution similar to the entire dataset, enhancing the robustness and reliability of the training process. The data was split into 80% for training and 20% for validation, and the model was optimized using the Adam optimizer, with categorical cross-entropy as the loss function, and accuracy as the evaluation metric. The CNN used two activation functions: ReLU and softmax. ReLU was selected to improve learning speed and reduce computational cost, while softmax was used in the final layer for multiclass classification. To prevent overfitting, normalization layers and dropout layers were added after each convolutional layer. These layers help the model generalize better by reducing the risk of becoming too specific to the training data, thereby improving its performance on unseen data.

C. Baseline Features

In this study, MFCC, LFCC and LFRCC features are used as baselines. MFCCs are widely used in speech processing and are derived from the short-term power spectrum of a speech signal. They transform the frequency components of the signal to the Mel frequency scale, which is more aligned with human auditory perception. This transformation captures important aspects of speech that are relevant to human hearing, allowing for effective representation of the spectral characteristics of speech.

LFCCs, on the other hand, are computed using a linear frequency scale rather than the Mel scale. This approach is advantageous for capturing detailed frequency variations in the speech signal. LFCCs especially in cases of speech disorders like dysarthria is crucial due to its linear scale.

LFRCCs are derived from the residual signal obtained after applying pre-emphasis and frame-wise processing [15]. The process starts with pre-emphasizing the residual signal to enhance high-frequency components, followed by dividing the signal into frames and applying a windowing function to each frame. The Fourier Transform is then used to compute the magnitude spectrum of each frame. This spectrum is passed through a linear filter bank, and the resulting energies are logarithmically transformed. Finally, the Discrete Cosine Transform (DCT) is applied to these log energies to obtain the cepstral coefficients [16]. This method captures the detailed spectral characteristics of the speech signal on a linear frequency scale, providing a comprehensive representation of the residual speech signal.

To ensure a fair comparison, 20-dimensional feature vectors were extracted from the speech data using the librosa toolkit, with a window length of 25 ms and a hop length of 10 ms for all feature sets. These features are commonly used in speech processing because they effectively represent the speech signal's spectral characteristics, making them suitable for analyzing dysarthric speech patterns [17]. By maintaining consistent extraction parameters across all feature sets, the study ensures that the comparative analysis of different features is reliable and unbiased.

IV. Spectrographic Analysis

A. Effect of entropy

Figure 2 illustrates the distribution of entropy across various subbands and frames for the word "November." Each subplot, labeled from (a) to (d), demonstrates how entropy varies, with subbands on the x-axis and frames on the y-axis. The z-axis represents entropy values, using a color gradient from blue (indicating lower entropy) to red (indicating higher entropy) to make differences easily visible. Subplots (a) and (b) display considerable fluctuations with notable peaks and valleys, which reflect a high degree of entropy variability. In contrast, subplots (c) and (d) have smoother, more uniform surfaces, indicating more consistent entropy values. The spectrogram reveals that as dysarthric severity increases, the overall entropy also increases. This pattern is evident in the transition from low to high severity levels, where the entropy values become more pronounced. This correlation between rising entropy levels and increased dysarthric severity underscores the utility of entropy as a metric for assessing the severity of dysarthria. The visual representation provided by these plots offers valuable insights into how entropy distribution changes with the progression of the condition.

From Fig. 3, the provided bar graphs illustrate the zero-mean entropy of each subband for varying degrees of dysarthric severity. Each subplot, from (a) to (d), represents different severity levels, with the x-axis indicating the subband index and the y-axis showing the zero-mean entropy values. Subplots (a) indicate values are relatively low and stable, indicating quite regular and predictable character of speech. This minimal variation around the mean can suggest controlled phonation. Subplot (b) demonstrates that, when compared to very low severity, the variations around the mean value of entropy increase, indicating the phonatory instability, reflecting high variability in entropy. Subplots (c) and (d) shows as the severity increases, the fluctuations increases and the entropy value increases indicating the irregularity and unpredictability of the speech signal, resulting in source phonation instability. The higher entropy values can indicate severe noise in the speech production system. The overall pattern reveals that as the severity of dysarthria increases, the entropy values also increase, highlighting a direct correlation between higher entropy levels and greater dysarthric severity. This consistent rise

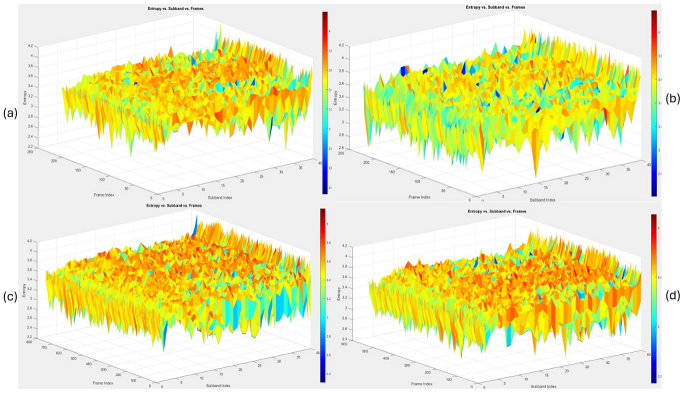


Fig. 2. Spectrogram of multi band entropy for each frame of the word "November" for (a)very low, (b)low, (c)medium, and (d)high severity-level dysarthric signal, respectively. X-axis represents the sub-band index, Y-axis represents frame index, and Z-axis represents the entropy value, respectively.

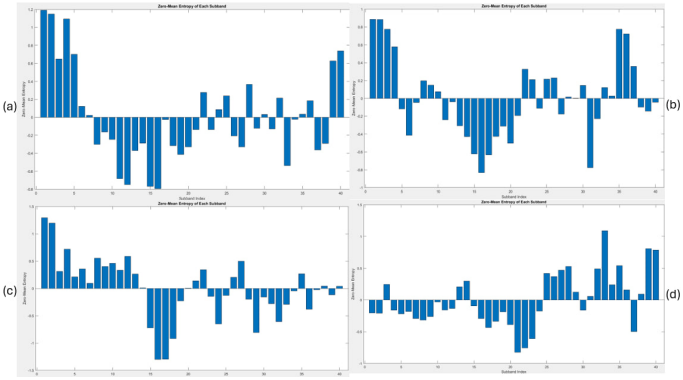


Fig. 3. Effect of entropy for the word "November" for (a)very low, (b)low, (c)medium, and (d)high severity-level dysarthria speaker. X-axis represents the sub-band index and Y-axis represents the zero mean entropy, respectively.

in entropy with increasing severity offers valuable insights into the relationship between entropy distribution and dysarthric speech impairment.

Together, these analyses clearly demonstrate that as dysarthria severity increases, speech becomes more chaotic and disordered, with higher and more variable entropy values.

V. Experimental Results

Table I indicates the classification accuracy of baseline features using cnn classifier.

TABLE I
FOLD, TEST, PRECISION (P), RECALL (R), F1-SCORE FOR BASELINE FEATURES USING CNN CLASSIFIER ON UA-SPEECH AND TORGO.

Data	Features	Fold Acc.	Acc.	P	R	F1
UA-Speech	MFCC	93.09	93.36	93.89	93.38	93.56
	LFCC	94.11	95.45	95.63	95.60	95.57
	LFRCC	92.01	93.78	93.55	94.06	93.71
TORGO	MFCC	85.71	88.62	89.40	88.64	89.01
	LFCC	91.58	91.82	91.51	91.02	91.26
	LFRCC	92.43	94.20	93.96	94.19	94.07

Table 1 contains the fold test accuracies of baseline features. For AU Speech dataset, LFCC outperforms both MFCC and LFRCC baseline features by a fold (test) accuracy margin of 1.02% (2.09%), 2.1% (1.67%) and for TORGO dataset, LFRCC outperforms both MFCC and LFCC baseline features by a fold (test) accuracy margin of 6.72% (5.58%), 0.85% (2.38%). This might be because dysarthric speech is seen to have turbulent noise in a higher frequency range and thus, the linear frequency scale in LFCC and LFRCC offers relatively better frequency resolution than its MFCC counterpart to capture the spectral details at higher frequency region.

TABLE II
FOLD, TEST, PRECISION (P), RECALL (R), F1-SCORE FOR 40 SUB BAND FEATURES ALONG WITH FIRST 20 AND LAST 20 OF DECREASING ARRANGED SUB BAND ENTROPY AND ZERO MEAN ENTROPY FEATURES USING CNN CLASSIFIER ON UA-SPEECH.

Features	Fold Acc.	Acc.	P	R	F1
Entropy	67.77	77.11	76.86	77.13	76.65
Entropy (first 20)	57.71	59.46	59.71	59.46	59.56
Entropy (last 20)	50.95	58.47	58.06	59.26	56.62
Zero mean entropy	**82.75**	**86.15**	**86.65**	**85.87**	**86**
Zero mean entropy (first 20)	79.86	80.36	80.40	80.04	80.17
Zero mean entropy (last 20)	68.86	71.89	71.64	72.57	70.66

TABLE III
FOLD, TEST, PRECISION (P), RECALL (R), F1-SCORE FOR 40 SUB BAND FEATURES ALONG WITH FIRST 20 AND LAST 20 OF DECREASING ARRANGED SUB BAND ENTROPY AND ZERO MEAN ENTROPY FEATURES USING CNN CLASSIFIER ON TORGO.

Features	Fold Acc.	Acc.	P	R	F1
Entropy	71.36	74.81	77	75.53	74.86
Entropy (first 20)	65.55	70.19	70.00	70.60	69.16
Entropy (last 20)	60.28	62.24	64.81	60.10	56.84
Zero mean entropy	**90.93**	**91.93**	**92.15**	**92.17**	**91.93**
Zero mean entropy (first 20)	89.14	84.88	886.25	85.38	84.51
Zero mean entropy (last 20)	68.86	71.89	71.61	72.57	70.66

TABLE IV
FOLD, TEST, PRECISION (P), RECALL (R), F1-SCORE FOR FEATURE LEVEL FUSION USING CNN CLASSIFIER ON UA-SPEECH AND TORGO.

Data	Features	Fold Acc.	Acc.	P	R	F1
UA-Speech	MFCC+Zero mean	95.65	97.74	97.43	97.61	97.51
	LFCC+Zero mean	94.13	97.79	97.33	97.11	97.20
	LFRCC+Zero mean	94.85	95.66	94.46	94.38	94.41
TORGO	MFCC+Zero mean	91.10	96.97	97.06	96.96	96.98
	LFCC+Zero mean	94.51	94.18	92.43	90.87	91.21
	LFRCC+Zero mean	94.24	95.96	96.14	96.33	96.16

Table 2 and table 3 indicate the accuracies for various forms of entropy and zero mean entropy for UA speech and TORGO datasets, respectively. The first 20 and last 20 features indicate the top and bottom 20 features from the descending arranged sub band entropy values. It can be seen that zero mean entropy outperforms the entropy feature by 9.04% for UA Speech and 17.12% for TORGO datasets. Furthermore, it is evident that the information across all the 40 subbands is essential in order to classify the severity level accurately.

Additionally, feature level fusion of baseline features is performed with best performing entropy feature i.e., zero mean entropy and the fusion based features outperform baseline

features by **4.38% (8.35%)** for MFCC, **2.34% (2.36%)** for LFCC, **1.88% (1.76%)** for LFRCC on UA Speech (TORGO), respectively. These experimental results sit right with the spectrographic analysis and emphasize the importance of inclusion of entropy based features for dysarthric severity classification.

These consistent improvements across both datasets highlight the effectiveness of integrating zero mean entropy into feature extraction. This technique not only refines the representation of the audio data but also enhances the overall accuracy of speech recognition models, making them more robust and reliable. The observed gains in accuracy underscore the value of adopting zero mean entropy in the development of advanced speech recognition systems.

VI. Conclusions

This study proposed the use of entropy based features for the application of dysarthric severity classification employing two widely used dysarthric speech corpra: UA-Speech and TORGO. Spectrographic analysis of speech data from individuals with varying degrees of dysarthria reveals distinct entropy patterns associated with each severity level. These patterns help in identifying specific speech alterations that occur as the disorder progresses. For instance, individuals with more severe dysarthria often exhibit higher entropy values due to greater variability in their speech signals. This analysis is supported by several experiments.

Experimental results support the effectiveness of incorporating zero mean entropy into the feature extraction process for classification. These results show that adding entropy information significantly boosts the accuracy of these systems, enabling them to better differentiate between different levels of dysarthria. Furthermore, this work can be extended by inclusion of entropy based information extracted using various filterbanks like linear and mel in the state-of-the-art deep learning models like whisper, wav2vec 2.0 and analyzing the noise robustness ability of entropy which helps us understand the practicality of the proposed methodology.

Acknowledgements

This study has been funded by Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, under the sponsorship of project 'BHASHINI', (Grant ID: 11(1 2022-HCC.(TDIL)). Authors thank authorities of DA-IICT Gandhinagar for their kind support and cooperation to carry out this research work.

References

[1] P. Lieberman, "Primate vocalizations and human linguistic ability," *The Journal of the Acoustical Society of America*, vol. 44, no. 6, pp. 1574–1584, 1968.

[2] R. Palmer and P. Enderby, "Methods of speech therapy treatment for stable dysarthria: A review," *Advances in Speech Language Pathology*, vol. 9, no. 2, pp. 140–153, 2007.

[3] Y. Kim, R. D. Kent, and G. Weismer, "An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria," 2011.

[4] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 5070–5074.

[5] A. Farhadipour, H. Veisi, M. Asgari, and M. A. Keyvanrad, "Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks," *Etri Journal*, vol. 40, no. 5, pp. 643–652, 2018.

[6] A. Kachhi, A. Therattil, A. T. Patil, H. B. Sailor, and H. A. Patil, "Teager energy cepstral coefficients for classification of dysarthric speech severity-level," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 1462–1468.

[7] N. Obin and M. Liuni, "On the generalization of shannon entropy for speech recognition," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2012, pp. 97–102.

[8] S. P. Mishra, P. Warule, and S. Deb, "Speech emotion recognition using mfcc-based entropy feature," *Signal, image and video processing*, vol. 18, no. 1, pp. 153–161, 2024.

[9] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a hmm classification framework," *Speech communication*, vol. 40, no. 3, pp. 351–363, 2003.

[10] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, "Automatic detection of hypernasal speech signals using nonlinear and entropy measurements.," in *INTERSPEECH*, 2012, pp. 2029–2032.

[11] H. Azami, L. Faes, J. Escudero, A. Humeau-Heurtier, and L. E. Silva, "Entropy analysis of univariate biomedical signals: Review and comparison of methods," *Frontiers in Entropy across the Disciplines: Panorama of Entropy: Theory, Computation, and Applications*, pp. 233–286, 2023.

[12] Z. Tüske, "Discriminative feature modeling for statistical speech recognition," Ph.D. dissertation, Dissertation, RWTH Aachen University, 2020, 2020.

[13] A. Lauraitis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Detection of speech impairments using cepstrum, auditory spectrogram and wavelet time scattering domain features," *IEEE Access*, vol. 8, pp. 96 162–96 172, 2020.

[14] J. Schröder, S. Goetze, and J. Anemüller, "Spectro-temporal gabor filterbank features for acoustic event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2198–2208, 2015.

[15] P. Gupta and H. A. Patil, "Linear frequency residual cepstral features for replay spoof detection on asvspoof 2019," in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 349–353.

[16] C. Nadeu, D. Macho, and J. Hernando, "Time and frequency filtering of filter-bank energies for robust hmm speech recognition," *Speech Communication*, vol. 34, no. 1-2, pp. 93–114, 2001.

[17] B. A. Al-Qatab and M. B. Mustafa, "Classification of dysarthric speech according to the severity of impairment: An analysis of acoustic features," *IEEE Access*, vol. 9, pp. 18 183–18 194, 2021.

Feature Extraction for Machine Learning-based Sleep Stage Classification Using PPG-Derived Parameters and Skin Temperature

Suphachok Buaruk*, Sikawat Thanaviratananich†, Peerasit Treesuthacheep‡ and Somrudee Deepaisarn§
* Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
E-mail: d6522300067@g.siit.tu.ac.th, s.buaruk@gmail.com
† Division of Neurology, Department of Medicine, Bumrungrad International Hospital, Bangkok, Thailand
Sleep Disorders Center, Cleveland Clinic Neurological Institute, Cleveland, Ohio, USA
E-mail: thanavir@gmail.com
‡ Chulalongkorn Comprehensive Epilepsy Center of Excellence, King Chulalongkorn Memorial Hospital, Bangkok, Thailand
E-mail: ptpeerasit@gmail.com
§ Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
Research Unit in Sustainable Electrochemical Intelligent, Thammasat University, Pathum Thani, Thailand
E-mail: somrudee@siit.tu.ac.th (Corresponding Author)

***Abstract*—This study explores feature extraction and machine learning methods for automated sleep stage classification using physiological signals, focusing on PPG-derived parameters and skin temperature. We experimented with feature extraction techniques that required different engineering steps, i.e., with or without applying the first derivative and deviation from the mean of raw signals. The statistical values, including mean, standard deviation, maximum, minimum, and median of each 30-second segment of signals, are then extracted and used as input features for our selected classification model, the Extra Trees. The classification performance is evaluated under both subject-dependent and subject-independent scenarios using data from 93 subjects. The data include labeled stages of wake, non-rapid eye movement (NREM) sleep stages N1, N2, and N3, as well as rapid eye movement (REM) sleep. Optimal features were derived from skin temperature measurement along with the normalized timestamp, which is associated with sleep duration. Results show that the subject-dependent experiment yields the highest performance with an accuracy of 0.87 for three-stage classification (wake, NREM:N1/N2/N3, REM) and 0.77 for five-stage classification (wake, N1, N2, N3, REM). This work highlights the potential of artificial intelligence to enable the automatic labeling of sleep stages using non-constrained sensors. The important features related to sleep staging can be investigated, which could contribute to future advances in clinical diagnostics and healthcare applications.**

I. Introduction

Sleep is vital for health and well-being. It helps the body recover, enhances brain function, and maintains emotional stability. Good sleep is crucial for memory, a strong immune system, and proper metabolism. Not getting enough sleep or having poor-quality sleep can lead to health conditions, such as heart disease, diabetes, and cognitive difficulties [1]. Sleep staging is the process of categorizing different phases of sleep. It is crucial for understanding sleep patterns, diagnosing sleep disorders, and developing effective treatment strategies. The gold standard for sleep staging is Polysomnography (PSG), a recording of the bio-physiological changes that occur during sleep [2]. PSG involves the use of electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) signals to monitor brain activity [3]. These signals are essential for distinguishing sleep stages [4]. However, PSG is complex and requires a high level of technical expertise to interpret. According to its complexity and the need for technical expertise to interpret, a study by the American Academy of Sleep Medicine (AASM) found that the sleep staging agreement among 2500 scorers using PSG is around 83% efficient [5]. Another study examined the overall agreement for sleep staging between seven scorers, finding 82% agreement rate [6]. Both studies found the highest agreement on REM sleep, while the lowest was on N1 sleep. The procedure is typically conducted in a clinical setting with many sensors and electrodes, which can be uncomfortable for the patient. There is a growing interest in finding less invasive and more accessible methods for sleep staging [7]. Recent research has shown that skin body temperature holds the potential for sleep staging. Biologically, skin temperature is affected by the thermoregulatory processes, which vary during different sleep stages [8]–[10]. For instance, skin temperature tends to increase during the transition to sleep and remains relatively higher during REM sleep [11]. However, there is a lack of quantitative studies exploring the direct use of skin temperature for sleep staging. In addition, the use of photoplethysmography (PPG) on sleep staging can be used to detect sleep and wake stages. This detection relied on the combined features between heart rate variability (HRV) and ECG-derived respiration (EDR) signals, measured from wearable sensors [12]. To the best of our knowledge, no studies investigate the combination of skin temperature and PPG-derived parameters as analyzing features for sleep staging.

Advancements in machine learning (ML) and deep learning (DL), combined with the biological signals from PSG, have opened new techniques for sleep staging. This approach could be applied to offer more efficiency in the aspect of manpower

compared to the traditional techniques [13]. In the context of sleep staging, the ML technique can process signals from PSG devices to classify different sleep stages. As a subset of ML, DL takes this a step further by using neural networks with multiple layers to model complex and non-linear relationships in the data. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been particularly effective in handling time-series data and recognizing intricate patterns associated with sleep stages. Many studies have adopted CNNs and RNNs for sleep staging with signals retrieved from PSG. Li. F. *et al.* developed an end-to-end automatic sleep staging method by proposing the stacking model architecture with a collection of consecutive convolutional micro-networks (CCNs) and squeeze-excitation (SE) block [14]. The proposed architecture was applied to use different single-channel EEGs. Research has constructed sleep staging models based on the random forest algorithm to distinguish four sleep stages based on core body temperature rhythm, achieving accuracies of 0.70 for males and 0.77 for females [15]. PPG signals, measured by finger pulse oximeters in a high prevalence of obstructive sleep apnea (OSA) patients, were employed with combined CNNs and RNNs to classify sleep stages into three (wake/NREM/REM), four (wake/N1+N2/N3/REM), and five (wake/N1/N2/N3/REM). The performance results showed that the three-stage model achieved 80.1% accuracy, while the four- and five-stage models achieved accuracies of 68.5% and 64.1%, respectively [16].

In this study, skin temperature and a set of parameters derived from PPG sensors are used to classify sleep stages. The statistical features, including the mean and standard deviation of these signals, are utilized as inputs for classical ML approaches. Additionally, the first derivatives of each prepared signal are included to capture the rate of change and further enhance the model's ability to detect complex variations in the data. Various experimental scenarios are explored, including different numbers of sleep stage labels, subject dependency, and groups of features. This comprehensive approach aims to understand the effectiveness and robustness of the proposed methods in classifying sleep stages. By using easy-access tools, it offers a more accessible and less complex alternative to traditional PSG-based techniques, making sleep stage classification more feasible for a broader range of users and applications.

The subsequent sections of this paper are organized as follows. Section II describes the methodologies employed in this study, including an overview of the dataset, data preparation techniques, and classification experiments. In section III, the results and discussion regarding the classification performance of each experimental scenario are presented. Finally, section IV provides a summary of the study and explores potential directions for future research.

II. Method

A. Dataset

This work utilizes a public dataset acquired from 100 unique participants by the Duke University Health System Sleep Disorder Lab [17], [18]. This sleep study aimed to detect and monitor apnea events during sleep using the PSG gold-standard approach to sleep testing. PSG data collection began around 11 PM and continued until the participant naturally awoke around 6 AM, resulting in a 7-hour sleep test on each participant. In addition to PSG components, the Empatica E4 wristband (Empatica Inc., Milano, Italy) was placed on the left wrist of each participant and was deactivated upon awakening. The E4 device collected six raw signals, including blood volume pulse (BVP) from the PPG sensor, accelerometry in three axes, electrodermal activity (EDA), and skin temperature. In the dataset, heart rate (HR) and interbeat interval (IBI) were derived from the BVP. Additionally, technician-annotated sleep stages derived from the PSG data were recorded every 30 seconds. All data were resampled to a frequency of 64 Hz and synchronized using timestamps, which were time-shifted to start from zero. An example of raw signals used in this study, including skin temperature, BVP, HR, and IBI, is presented in Fig. 1.

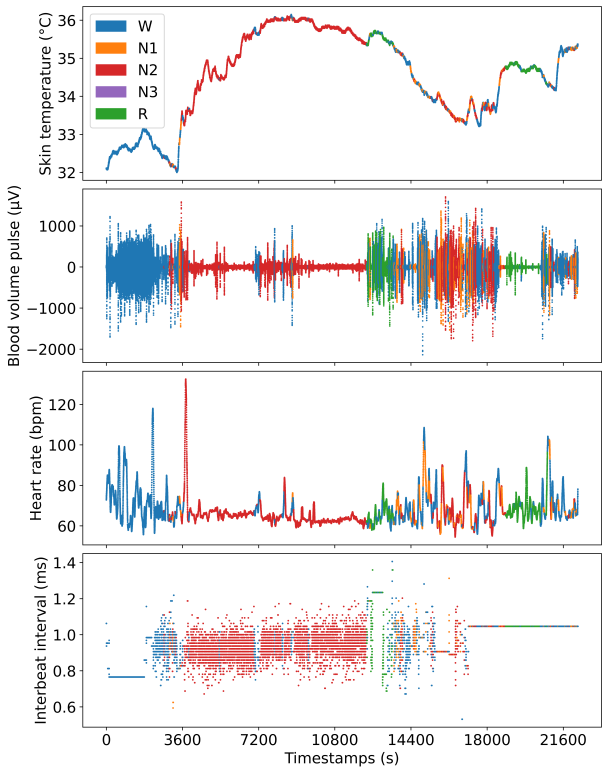


Fig. 1. Signal profiles of skin temperature, blood volume pulse, heart rate, and interbeat interval.

B. Feature Extraction

In this study, a total of 93 subjects were selected from the initial 100 subjects of the dataset. The 7 subjects were excluded from the analysis due to either having more than one sleep stage within certain 30-second intervals or having labels indicating a "preparing stage" (the stage before the PSG recording starts) between the five sleep stages. The preprocessing and

feature extraction steps involved the following:

1) **Feature Signals**: The feature signals used in this work include derived signals from a PPG sensor, specifically heart rate (HR), blood volume pulse (BVP), interbeat interval (IBI), and skin temperature.
2) **Resampling**: All signals were resampled from a 64 Hz sampling frequency to 1 Hz.
3) **Feature Engineering - Derivative Set**: The first derivative was applied to all resampled signals to create the derivative set. For a given signal x, the first derivative Δx_i is calculated as:

$$\Delta x_i = x_i - x_{i-1}, \tag{1}$$

Note that the first sample's derivative is initialized to zero: $\Delta x_1 = 0$. For the last sample, the derivative is computed as the difference between the last two samples to maintain consistency: $\Delta x_n = x_n - x_{n-1}$.

4) **Feature Engineering - Deviation Set**: The resampled signals were subtracted from their mean values to generate the deviation set. For a given signal x with n samples, the mean μ_x is calculated as:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

The deviation of each sample is then calculated as:

$$d_i = x_i - \mu_x \tag{3}$$

5) **Segmentation**: Both the derivative set and deviation set were segmented into 30-second epochs, with each segment labeled according to the corresponding sleep stage.
6) **Final Dataset Creation**: Each segment (or chunk) was processed to compute the mean and standard deviation for both the derivative set and the deviation set. These computed features formed the final dataset used for the classification task in sleep staging.

C. Classification Experiments

The classification experiment investigates various scenarios by varying factors involved in the training and evaluation of sleep staging classification. The parameters include:

1) **Number of stages**: The original dataset includes signals along with labels for five sleep stages: Wake (W), Non-Rapid Eye Movement (N1/N2/N3), and Rapid Eye Movement (R). This study examines three different classification schemes:
   - Three stages: Wake, combined N1/N2/N3, and REM
   - Four stages: Wake, combined N1/N2, N3, and REM
   - Five stages: Wake, N1, N2, N3, and REM
2) **Dependency**:
   - Subject-independence: Ensuring that subjects in the training and test sets are separate, with 74 subjects in the training set and 19 subjects in the test set. All experiments under subject-independent conditions will be evaluated using the same test set.
   - Subject-dependence: Splitting the data by a 70:30 ratio for training and testing, respectively. The splitting is performed by grouping ID and sleep stages, ensuring that all stages from each ID are split with the same ratio in both the training and test sets.
3) **Classification**: The classification training model is set up using the PyCaret framework [19], which provides robust data preprocessing steps. Since the imbalance in the sleep stages labels within the dataset, PyCaret addresses this issue by applying the SMOTE (Synthetic Minority Over-Sampling TEchnique). Additionally, the framework uses stratified K-fold cross-validation to ensure that all stages from each ID are separated in the same ratio in both the training and validation sets. In this work, the Extra Trees (ET) algorithm is employed as the classification algorithm with a 10-fold cross-validation. The performance of the classification model is monitored using several key metrics, including Accuracy, Recall, Precision, and F1-score, applied to the test dataset to compare the performance across different classification experiment scenarios.

III. Results and Discussion

A. Data Exploration

Fig. 1 illustrates an example of measured signals from the Empatica E4 wristband device. The skin temperature measurements obtained from this device exhibit a pattern similar to the distal temperature reported in the study of C. Cajochen *et al.* [20]. This finding suggests that the distal-proximal temperature relationship should be considered when evaluating the performance of the easy-access tool used in this study.

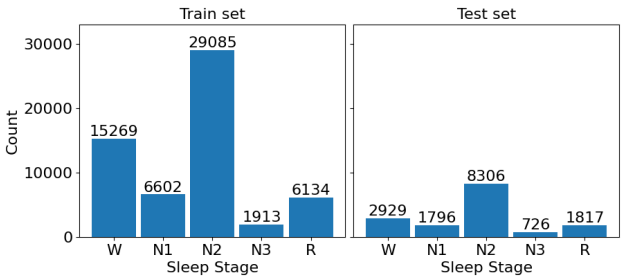


Fig. 2. Number of each sleep stage on the train (left) and test (right) datasets, split by subject-independent scenarios.

Fig. 2 illustrates the significant imbalance in the distribution of labels across the 74 subjects and 19 subjects in training and test sets, respectively, from the subject-independent strategy. The majority of the samples belong to the N2 and Wake stages, while the N3 stage has the fewest samples. This distribution could be influenced by the first night effect (FNE), which increases time in the light sleep stages and decreases time in deep sleep stages [21]. FNE also increases the transition probability from stage N2 to N1 and decreases the transition probability from stage N2 to N3 [22]. Additionally, patients with obstructive sleep apnea (OSA) who have short sleep times

probably experience a decrease in the N3 deep sleep stage and REM stage, along with an increase in N1/N2-stage time due to fragmented sleep architecture [23], [24].

B. Classification Performance

This subsection discusses the performance of each experimental scenario. The parameters include groups of features, including PPG-derived signals, skin temperature, and a combination of both. Additionally, the number of sleep stages used as labels are varied for three-, four-, and five-stages. The details of the subject-dependent and independent scenarios are described as follows:

In the result of subject-independence in Table I, the skin temperature profile shows better performance than using PPG-derived signals alone. However, the use of PPG-derived signals in this study yields lower performance than previous research reported by H. Korkalainen *et al* [16]. This difference might be due to the feature extraction methods in order to represent raw signals.

TABLE I
CLASSIFICATION PERFORMANCE OF SUBJECT-INDEPENDENCE USING STATISTICAL VALUES EXTRACTED FROM RAW SIGNALS, ALONG WITH NORMALIZED TIMESTAMPS.

Features	Stages	Accuracy	Recall	Precision	F1
PPG derived	3	0.37	0.37	0.62	0.39
	4	0.39	0.39	0.53	0.41
	5	0.29	0.29	0.45	0.31
Skin temperature	3	0.58	0.58	0.57	0.57
	4	0.55	0.55	0.51	0.53
	5	0.37	0.37	0.39	0.37
PPG-derived & Skin temperature	3	0.42	0.42	0.61	0.45
	4	0.42	0.42	0.52	0.44
	5	0.29	0.29	0.42	0.30

The significant impact of imbalanced labels also affects the ability to classify less occurred stages accurately. Models may become biased towards the more frequent stages. This imbalance obviously impacts classification performance, especially under the subject-independence. The performance gap between the five-stage classification scenario and the three- and four-stage scenarios is considerable. Separating into five stages requires considering N2 separately from N1 and N3. Since N1 has a significantly lower sample compared to N1 and N2, this results in less support for test classification performance. Consequently, the overall performance in the five-stage classification scenario is lower.

The classification performance improved by 10-16% across all experimental scenarios when using the statistical values of deviation and the first derivative of the raw signal, compared to using the statistical values of the raw signals, as described in Table II.

Table III presents results for the 70:30 train-test split ratios, respectively. The classification performance using skin temperature alone is 87%, 85%, and 77% for three-, four-, and five-stage classification, respectively. Combining both PPG-derived and skin temperature features decreases a bit performance by approximately 1%. across all experiment scenarios. This suggests that including the statistical values of the skin temperature profile as input features can enhance classification performance.

TABLE II
CLASSIFICATION PERFORMANCE OF SUBJECT-INDEPENDENCE USING STATISTICAL VALUES EXTRACTED FROM THE DEVIATION AND FIRST DERIVATIVE OF RAW SIGNALS, ALONG WITH NORMALIZED TIMESTAMPS.

Features	Stages	Accuracy	Recall	Precision	F1
PPG derived	3	0.53	0.53	0.63	0.56
	4	0.52	0.52	0.57	0.54
	5	0.41	0.41	0.46	0.43
Skin temperature	3	0.61	0.61	0.56	0.58
	4	0.56	0.56	0.50	0.52
	5	0.38	0.38	0.37	0.38
PPG-derived & Skin temperature	3	0.55	0.55	0.62	0.57
	4	0.52	0.52	0.55	0.53
	5	0.40	0.40	0.45	0.41

Similar to subject-independent scenarios, the deviation and the first derivative are computed for subject-dependent scenarios. However, as shown in Table IV, the classification performance decreases when using either PPG-derived features or skin temperature features alone. By incorporating a range of known labels for the individuals' sleep stages into the training set, these models can be fine-tuned to enhance their accuracy and reliability. This personalized approach highlights the importance of subject-dependent data in improving sleep stage classification performance.

TABLE III
CLASSIFICATION PERFORMANCE OF SUBJECT-DEPENDENCE USING STATISTICAL VALUES EXTRACTED FROM RAW SIGNALS, ALONG WITH NORMALIZED TIMESTAMPS.

Features	Stages	Accuracy	Recall	Precision	F1
PPG derived	3	0.81	0.81	0.82	0.81
	4	0.79	0.79	0.80	0.79
	5	0.72	0.72	0.71	0.71
Skin temperature	3	0.87	0.87	0.87	0.87
	4	0.85	0.85	0.85	0.85
	5	0.77	0.77	0.77	0.77
PPG-derived & Skin temperature	3	0.86	0.86	0.86	0.86
	4	0.84	0.84	0.85	0.84
	5	0.77	0.77	0.77	0.77

Fig. 3 illustrates the normalized confusion matrix of classification performance using statistical values from the skin temperature profile. This confusion matrix shows the highest performance for REM sleep and the lowest for N1 sleep, aligning with the agreement of sleep staging by scorers [5], [6]. N1 is often characterized as the transitional phase between wakefulness and deeper sleep stages. During this stage, phys-

TABLE IV
CLASSIFICATION PERFORMANCE OF SUBJECT-DEPENDENCE USING STATISTICAL VALUES EXTRACTED FROM THE DEVIATION AND FIRST DERIVATIVE OF RAW SIGNALS, ALONG WITH NORMALIZED TIMESTAMPS.

Features	Stages	Accuracy	Recall	Precision	F1
PPG derived	3	0.76	0.76	0.77	0.76
	4	0.73	0.73	0.75	0.74
	5	0.66	0.66	0.66	0.66
Skin temperature	3	0.85	0.85	0.84	0.84
	4	0.83	0.83	0.83	0.83
	5	0.75	0.75	0.75	0.75
PPG-derived & Skin temperature	3	0.80	0.80	0.80	0.80
	4	0.79	0.79	0.79	0.79
	5	0.71	0.71	0.70	0.70

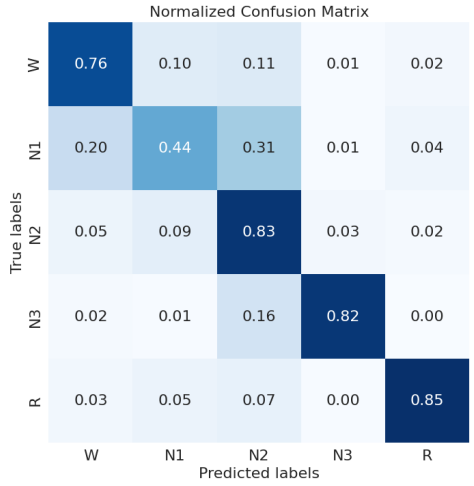


Fig. 3. Confusion matrix for 5 stages classification in subject-dependence, using skin temperature and normalized timestamps.

iological signals, i.e., skin temperature and heart rate, may not differ significantly from neighboring stages, leading the model to misclassify N1 as wake and N2, as presented in Fig. 3. Moreover, it can be seen that the staging performance using skin temperature-related features is better than that of using PPG-derived features. Medical studies have shown that skin temperature, especially near the wrist, increases at the sleep onset and changes predictably during different sleep stages. The distinction between distal and proximal temperature measurements is essential in the context of sleep staging. Distal temperature, measured at locations such as the wrist or ankle over the radial artery [25], can be easily captured using wearable devices.

In this work, the distal skin temperature was also measured using the Empatica E4 wristband device. Understanding the influence of distal temperature on sleep stages is important. During the transition to sleep and throughout NREM sleep, there is an increase in distal skin temperature. This increase is associated with vasodilation, promoting heat loss and helps lower the core body temperature [26]. In fact, distal temperature continues to increase during the N3 deep sleep stage [11]. Additionally, a warm pre-sleep environment before going to bed, known as the "Warm Bath Effect", increases NREM and decreases REM sleep [26]. Furthermore, environmental factors such as ambient temperature should be considered in the development of home-use measurement tools. Higher ambient temperature leads to increased skin temperature, potentially affecting the detection accuracy of sleep stages.

C. Effects of Skin Temperature and Sleep Duration on Sleep Stage Classification

The study on sleep stage classification explores the impact of incorporating normalized timestamps into the feature set. The results, as shown in Section III-B, demonstrate that taking into account the normalized timestamps significantly improves performance in subject-dependent scenarios. This improvement is evident when using either PPG-derived features or skin temperature features alone. However, the improvement is obviously presented with skin temperature features, achieving performance increase in the range of 19-23% depending on the number of sleep stage labels. These findings highlight the importance of temporal information in enhancing the accuracy of machine learning-based sleep staging classification. This suggests the possibility of developing personalized AI models for healthcare, particularly for sleep stage classification tasks.

IV. CONCLUSIONS AND FUTURE WORK

This work demonstrates that simple physiological parameters acquired from the wearable device, including PPG-derived parameters and skin temperature, can be effective for automatic sleep staging classification. The key process is feature extraction and engineering to select the suitable set of feature inputs to train the machine learning model. Our experimental design covers both subject-dependent and independent scenarios for the train-test splitting of the classification model with available sleep stage labels of 93 individual subjects. It was found that the information on skin temperature and sleep duration (i.e., normalized timestamps) led to a highly accurate model. The extra Tree model achieves up to 0.87 in accuracy in the subject-dependent scenario when labeling wake, NREM (N1/N2/N3), and REM. Our proposed model shows agreement with the sleep stage labels by human expert scorers both in terms of the overall classification performance and the bias in labeling stages arising from the ambiguity between staging behavior. The findings can pave the way for artificial intelligent solutions to analyze data from simple, non-constrained sensors to automate sleep stage labeling, which can be useful and practical for the healthcare industry. Environmental factors such as ambient temperature and light intensity can be taken into account in future research. An appropriate experimental design to collect primary data is necessary to improve scoring performance, reduce bias, and gain more impactful insights related to the sleep stage patterns of individuals.

ACKNOWLEDGMENT

The first author (S.B.) acknowledges the Graduate Scholarship Program for Excellent Thai Students, awarded by

Sirindhorn International Institute of Technology, Thammasat University, Thailand, for his doctoral study.

REFERENCES

[1] G. Medic, M. Wille, and M. E. Hemels, "Short-and long-term health consequences of sleep disruption," *Nature and science of sleep*, pp. 151–161, 2017.

[2] S. Keenan, M. Hirshkowitz, and H. Casseres, "Monitoring and staging human sleep," *Principles and practice of sleep medicine*, vol. 5, pp. 1602–1609, 2011.

[3] J. V. Rundo and R. Downey III, "Polysomnography," *Handbook of clinical neurology*, vol. 160, pp. 381–392, 2019.

[4] M. D. Oliver and S. Datta, "Electrophysiological correlates of the sleep–wake cycle," in *The Behavioral, Molecular, Pharmacological, and Clinical Basis of the Sleep-Wake Cycle*, Elsevier, 2019, pp. 17–26.

[5] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: Sleep stage scoring," *Journal of clinical sleep medicine*, vol. 9, no. 1, pp. 81–87, 2013.

[6] H. Danker-Hopfe, P. Anderer, J. Zeitlhofer, *et al.*, "Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard," *Journal of sleep research*, vol. 18, no. 1, pp. 74–84, 2009.

[7] T. Lauteslager, S. Kampakis, A. J. Williams, M. Maslik, and F. Siddiqui, "Performance evaluation of the circadia contactless breathing monitor and sleep analysis algorithm for sleep stage classification," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2020, pp. 5150–5153.

[8] B. H. Te Lindert and E. J. Van Someren, "Skin temperature, sleep, and vigilance," *Handbook of clinical neurology*, vol. 156, pp. 353–365, 2018.

[9] R. Szymusiak, "Body temperature and sleep," *Handbook of clinical neurology*, vol. 156, pp. 341–351, 2018.

[10] X. Xu, J. Zhu, C. Chen, X. Zhang, Z. Lian, and Z. Hou, "Application potential of skin temperature for sleep-wake classification," *Energy and Buildings*, vol. 266, p. 112 137, 2022.

[11] S. Herberger, T. Penzel, I. Fietze, *et al.*, "Enhanced conductive body heat loss during sleep increases slow-wave sleep and calms the heart," *Scientific Reports*, vol. 14, no. 1, p. 4669, 2024.

[12] R. V. Sharan, "Ecg-derived respiration for sleep-wake stage classification," in *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, IEEE, 2021, pp. 1–4.

[13] D. Sarkar, D. Guha, P. Tarafdar, S. Sarkar, A. Ghosh, and D. Dey, "A comprehensive evaluation of contemporary methods used for automatic sleep staging," *Biomedical Signal Processing and Control*, vol. 77, p. 103 819, 2022.

[14] F. Li, R. Yan, R. Mahini, *et al.*, "End-to-end sleep staging using convolutional neural network in raw single-channel eeg," *Biomedical Signal Processing and Control*, vol. 63, p. 102 203, 2021.

[15] X. Xu and Z. Lian, "A sleep staging model based on core body temperature rhythm," *Energy and Buildings*, vol. 310, p. 114 099, 2024.

[16] H. Korkalainen, J. Aakko, B. Duce, *et al.*, "Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea," *Sleep*, vol. 43, no. 11, zsaa098, 2020.

[17] K. Wang, J. Yang, A. Shetty, and J. Dunn, "Dreamt: Dataset for real-time sleep stage estimation using multisensor wearable technology,"

[18] A. L. Goldberger, L. A. Amaral, L. Glass, *et al.*, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, e215–e220, 2000.

[19] M. Ali, *Pycaret: An open source, low-code machine learning library in python*, PyCaret version 1.0, Apr. 2020. [Online]. Available: https://www.pycaret.org.

[20] C. Cajochen, K. Kräuchi, and A. Wirz-Justice, "Role of melatonin in the regulation of human circadian rhythms and sleep," *Journal of neuroendocrinology*, vol. 15, no. 4, pp. 432–437, 2003.

[21] A. Mayeli, S. Janssen, K. Sharma, and F. Ferrarelli, *Examining first night effect on sleep parameters with hd-eeg in healthy individuals. brain sci. 2022, 12, 233*, 2022.

[22] A. Shirota, M. Kamimura, A. Kishi, H. Adachi, M. Taniike, and T. Kato, "Discrepancies in the time course of sleep stage dynamics, electroencephalographic activity and heart rate variability over sleep cycles in the adaptation night in healthy young adults," *Frontiers in Physiology*, vol. 12, p. 623 401, 2021.

[23] S. Nozawa, K. Urushihata, R. Machida, and M. Hanaoka, "Sleep architecture of short sleep time in patients with obstructive sleep apnea: A retrospective single-facility study," *Sleep and Breathing*, pp. 1–8, 2021.

[24] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset.," *Sleep*, vol. 23, no. 7, pp. 901–908, 2000.

[25] F. M. Acosta, B. Martinez-Tellez, D. P. Blondin, *et al.*, "Relationship between the daily rhythm of distal skin temperature and brown adipose tissue 18f-fdg uptake in young sedentary adults," *Journal of Biological Rhythms*, vol. 34, no. 5, pp. 533–550, 2019.

[26] E. C. Harding, N. P. Franks, and W. Wisden, "The temperature dependence of sleep," *Frontiers in neuroscience*, vol. 13, p. 336, 2019.

Enhancing Acoustic Scene Classification with Layer-wise Fine-Tuning on the SSAST Model

Shuting Hao*, Daisuke Saito*, Nobuaki Minematsu*
* Graduate School of Engineering, The University of Tokyo, Tokyo E-mail: mine@gavo.t.u-tokyo.ac.jp

***Abstract*—We introduce a novel approach to Acoustic Scene Classification (ASC) using the Self-Supervised Audio Spectrogram Transformer (SSAST) sophisticated with a focus on layer-wise fine-tuning. Recognizing the challenge in distinguishing similar classes in the TAU Urban Acoustic Scenes 2022 dataset, which sometimes exceeds human perceptual capabilities, this study introduces a novel architecture that categorizes environmental audio streams into predefined semantic labels through integrating multi-layer classifiers and direct fine-tuning. Employing the TAU Urban Acoustic Scenes 2022 Mobile dataset for both fine-tuning and validation, our SSAST model, which was initially pre-trained on the AudioSet and LibriSpeech datasets, was uniquely fine-tuned to enhance ASC-specific feature learning. Here, a combined approach of layer-wise and simultaneous fine-tuning of the backbone was introduced, eliminating the need for reassembling the dataset. This method achieved satisfactory results, our layered SSAST system reached an accuracy of 52.43% and an AUC of 88.51%, marking a notable improvement over the baseline with absolute increases of 1.25% in accuracy and 0.70% in AUC.**

I. Introduction

Acoustic Scene Classification (ASC) is to assign one of the predefined semantic labels, such as bus, office, or home, to an input audio stream recorded in various environments [1]–[3]. Generally, semantic labels stem from environmental sound categorizations that describe the ambiance of audio streams. An overview of a ASC system is shown in Fig. 1. For many years, ASC has been an active area of research. This field has made noteworthy advancements, notably in areas like the Detection and Classification of Acoustic Scenes and Events. Environmental audio signals, especially when compared to speech or music, tend to be more diverse and complex, posing unique challenges. As a result, ASC's progress has been somewhat slower than in other audio-related domains, such as speaker identification and music classification. To tackle the issue of low performance, researchers have introduced a variety of strategies. These include signal processing techniques, data augmentation, advanced feature learning, efficient deep learning modeling, and sophisticated post-processing approaches.

In this paper, we attempt to enhance accuracy without retraining models from scratch. We introduce a novel architecture that incorporates multi-layer classifiers and direct fine-tuning, diverging from the previous models such as Patchout Spectrogram Transformer (PASST) [4] which often require extensive reassembly of training data. As far as we know, the application of the Self-Supervised Audio Spectrogram Transformer (SSAST) model [5] to the TAU Urban Acoustic Scenes 2022 Mobile dataset [6] is not a mainstream approach and has not been extensively studied within the DCASE community. This

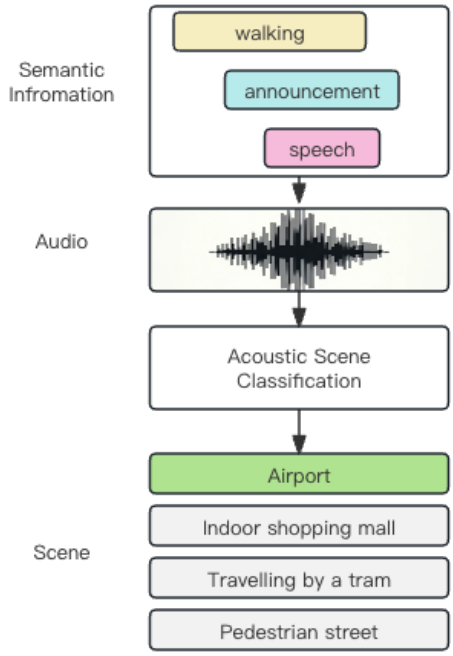


Fig. 1: Overview of an ASC system.

study aims to explore its potential and evaluate its performance on this dataset. Our experiments demonstrate that our system achieves an accuracy of 52.43%, surpassing other methods that rely solely on direct fine-tuning. The TAU dataset is notoriously challenging, containing environmental sounds that even humans struggle to differentiate. This innovative approach not only addresses the challenges of applying pre-trained models to ASC but also advances how these models can be fine-tuned more effectively. One of the main observations in this study is the difficulty in improving accuracies for distinguishing similar categories. To tackle this issue, we have implemented advanced feature extraction techniques designed specifically to enhance differentiation among similar acoustic scenes. Our method avoids concatenating audio data, instead training on original recordings, which preserves the authenticity of the audio contexts.

II. Related and Proposed Models

This chapter describes the methodology employed in the current study, focusing on the innovative use of the Self-

Supervised Audio Spectrogram Transformer (SSAST) as the baseline model for acoustic scene classification (ASC). Traditionally, ASC tasks have predominantly employed models such as the Patchout faSt Spectrogram Transformer (PaSST), which optimizes transformer training on audio spectrograms to achieve state-of-the-art performance with less computational overhead [7]. Unlike PaSST, which still requires pre-training on large labeled datasets and employs patch-based input strategies akin to those used in vision tasks, the SSAST introduces a novel approach by leveraging a self-supervised learning framework to pre-train directly on raw, unlabeled audio data [5].

Another significant strength of SSAST is its ability to perform well without the need to reassemble audio snippets into longer segments for training or evaluation. This capability underscores its robustness and effectiveness in real-world applications, where manipulation of audio data is not always feasible or desirable. The common practice in ASC competitions, such as reassembling audio snippets into longer segments for training [8], exemplified by the best models in the DCASE challenges, poses significant theoretical and practical challenges. This method may create discrepancies between training and operational environments. In real-world applications, audio data often arrive in non-continuous streams without the opportunity for pre-segmentation or reassembly. Training on artificially concatenated segments may lead to models that perform well in a controlled setting but falter in real-world scenarios where such preprocessing is not feasible. This approach may affect the model's generalization to diverse acoustic environments.

A. Baseline: Self-Supervised Audio Spectrogram Transformer (SSAST)

The Self-Supervised Audio Spectrogram Transformer (SSAST) architecture [5], which is illustrated in Fig. 2 as a self-supervised learning framework for Audio Scene Classification (ASC), is built on the Audio Spectrogram Transformer (AST) model [9]. The AST, which is based purely on self-attention mechanisms as in the Vision Transformer (ViT) [10], has demonstrated superior performance over traditional deep learning models that used convolutional neural networks (CNNs) for a variety of tasks. SSAST begins by transforming a τ-sec long audio segment into a series of 128-dimensional log Mel filterbank (fbank) features. These features are derived using a 25ms Hanning window at 10ms intervals, resulting in a $128\times100\tau$ spectrogram. This spectrogram is then divided into 16×16 patches, as shown in A of Fig. 2. And those patches are subsequently flattened into 1D 768-dimensional embeddings through a linear projection, known as the patch embedding layer, producing patch embeddings E_i. To account for the Transformer architecture's lack of understanding the temporal order, a trainable positional embedding of size 768, denoted as P_i, is added to each patch embedding, facilitating the model's understanding of the 2D audio spectrogram's spatial structure, two embeddings are shown in B of Fig. 2. The sequence of embeddings is then processed through the Transformer,

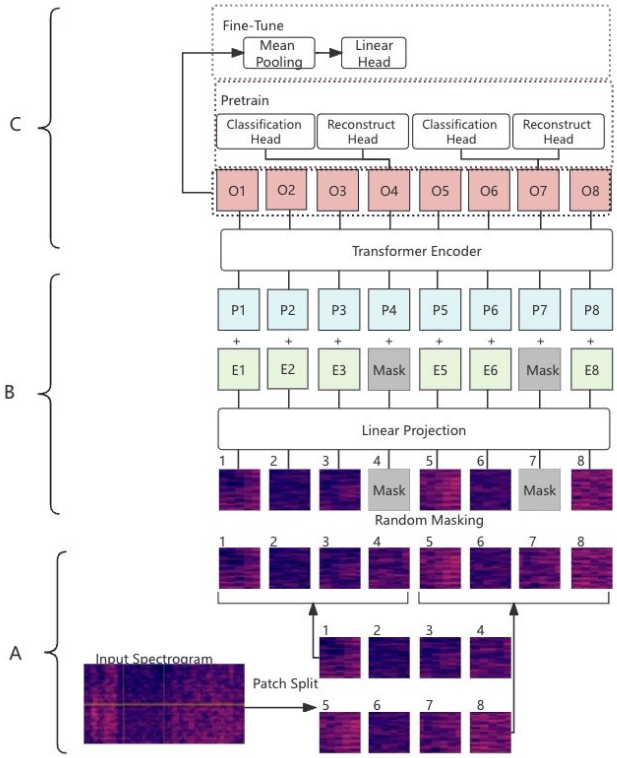


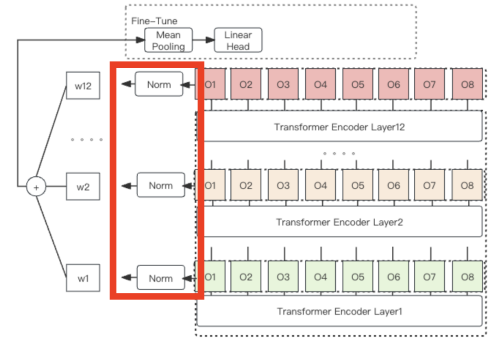
Fig. 2: Block diagram of Self-Supervised Audio Spectrogram Transformer (SSAST)

As shown in the C part of this figure, during self-supervised pretraining, a portion of spectrogram patches are randomly masked. And the task for the model is to 1) find the correct patch at each masked position from all masked patches; and 2) reconstruct the masked patch. The two pretext tasks aim to force the AST model to learn both the temporal and frequency structure of the audio data. During fine-tuning, the purpose apply a mean pooling over all patch representation O_i and use a linear head for classification [5].

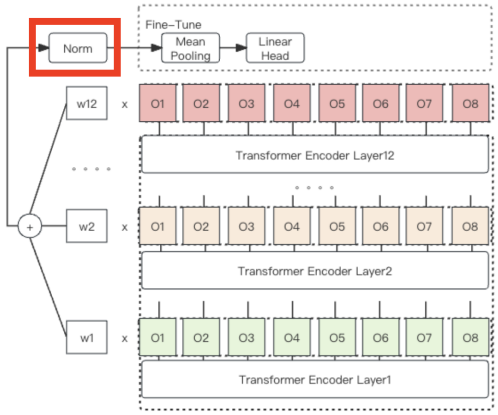
which is composed of multiple encoding layers and decoding layers. In SSAST, however, only the encoding part of the Transformer is utilized, which extracts an embedding feature whose dimension is 768, with 12 layers and 12 heads. The output from the Transformer encoder, referred to as the patch representation O_i, undergoes a mean pooling during fine-tuning and inference to yield an audio clip level representation. Subsequently, a linear head is applied for the classification task.

B. Proposed: Normalization for Layer Wise Classifiers

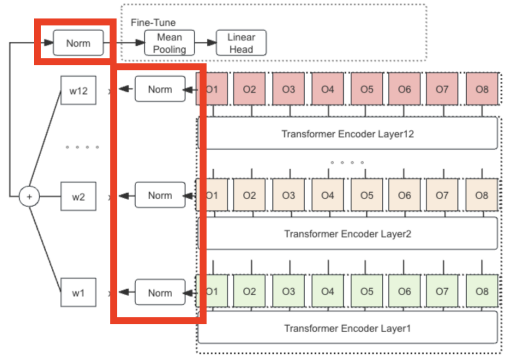
The architecture of our layer-wise classifier is shown in Fig 3 with three specific methods applied. The first method emphasizes the importance of individual layer outputs by normalizing each block's output, denoted as Pre-Norm in Fig 3a. The mean of the resultant vectors, excluding the class token, is computed to derive layer-wise representations. These are weighted by a softmax function with learned weights,



(a) Post-Norm



(b) Pre-Norm



(c) Double-Norm

Fig. 3: Block diagram of layer wise architecture.

allowing for differential emphasis on information from various network depths. This weighted sum is input to a multi-layer perceptron (MLP) head for classification. The equations for this method are:

$$M_i = \text{Mean}(\text{Norm}(O_i)), \qquad i \in \{1, \ldots, L\} \tag{1}$$

$$S = \sum_{i=1}^{L} (W_i \times M_i) \tag{2}$$

$$x = \text{MLPHead}(S), \tag{3}$$

where M_i is the output from the ith layer after normalization and mean calculation, O_i is the output of the ith block, S is the weighted sum of the outputs, W_i are the softmax-normalized weights, and L is the total number of layers.

The second method defers mean pooling until the weighted sum of layers is calculated, denoted as Post-Norm in Fig 3b. The layer outputs are aggregated using softmax-scaled weights and then normalized before passing through the MLP head. This is represented as:

$$M_i = \text{Mean}(O_i), \qquad i \in \{1, \ldots, L\} \tag{4}$$

$$S = \sum_{i=1}^{L} (W_i \times M_i) \tag{5}$$

$$x = \text{MLPHead}(\text{Norm}(S)), \tag{6}$$

To expand on these two methods with a new variation, we introduce a third approach that incorporates normalization both before and after the aggregation of layer outputsdenoted as Post-Norm in Fig 3b.. This method applies normalization initially to each layer output before calculating the weighted sum, and once again after the aggregation, ensuring a more uniform scale and potentially enhancing the stability and performance of the MLP head. Here's how this new method is formulated:

$$M_i = \text{Mean}(\text{Norm}(O_i)), \qquad i \in 1, \ldots, L \tag{7}$$

$$S = \sum_{i=1}^{L} (W_i \times M_i) \tag{8}$$

$$x = \text{MLPHead}(\text{Norm}(S)), \tag{9}$$

In all methods, MLPHead denotes the MLP head, which processes the normalized weighted sum and produces the classification output x. The normalization step ensures that the signal is suitable for classification, making the most of the diverse features extracted from the audio signal.

III. Experimental Setup

In our experimental setup, all evaluations were conducted on the TAU Urban Acoustic Scenes 2022 Mobile dataset [6]. This dataset comprises 64 hours of audio recordings from 12 different European cities, featuring 10 unique acoustic scenes. The audio was captured using four distinct devices, and each audio sample has a length of 1 second, trimmed from the original 10-second recordings. Out of the total audio data, 40 hours were recorded using the main device, while the remaining hours were captured with the three additional, less frequently used devices. The audio format for all devices is single-channel, 44.1kHz, and 24-bit resolution.

In our experimental framework, we leveraged the SSAST model, which features a novel approach integrating both discriminative and generative learning paradigms through Masked Spectrogram Patch Modeling (MSPM). This model was pre-trained using a vast collection of unlabeled audio data sourced from the AudioSet [11] and LibriSpeech datasets [12]. With MSPM, the model masks a strategic subset of spectrogram patches, specifically 400 patches, to challenge the model to predict the missing acoustic information. This technique not only encourages the model to learn a more comprehensive

representation of the audio but also to discern finer details that are crucial for classification. Notably, the effectiveness of this model has been validated by its impressive performance on the ESC-50 dataset [13], a benchmark in environmental sound classification. This is why the SSAST model is regarded as suited foundation on the TAU Urban Acoustic Scenes 2022 Mobile dataset.

For our experiments, we employed a consistent set of hyperparameters across all models to maintain uniformity in the evaluation conditions. These parameters included the learning rate, batch size, and data augmentation techniques, along with the warm-up strategy, number of training epochs, choice of optimizer, and loss function utilized during the training process.

Various data augmentation strategies were explored during experimentation. Mixup augmentation, implemented at the waveform level using a Beta distribution for mixing weights, was attempted but did not lead to significant improvement. Feature-level augmentation was also investigated, specifically by adding random noise to the log mel-filterbank features and applying time-axis shifting, which resulted in a modest improvement. The most effective augmentation technique proved to be the masking strategy inherent to SSAST's self-supervised learning framework, which applies both time and frequency masking to the input spectrogram.

In addition to feature augmentation, a notable deviation from the pre-training phase was the input handling; specifically, we introduced overlapping inputs to enrich the model's exposure to the data. With these consistent parameters in place, we focused on the classifier component of our modified architecture. Starting with randomly initialized weights for the classifier while inheriting the backbone's weights from pre-training, we proceeded with fine-tuning the entire model without freezing any layers. This approach was adopted due to the significant differences between the TAU Urban Acoustic Scenes 2022 Mobile dataset and the AudioSet upon which the model was originally trained. To ascertain the impact of our strategy, we also conducted comparative experiments using models with their backbone layers frozen during fine-tuning.

IV. Results and Discussion

Before delving into the specifics of the experimental outcomes, it's essential to understand the metrics used to measure the performance of the systems. Accuracy (Acc) is a metric that represents the proportion of true results, both true positives and true negatives, in the total number of cases examined. It provides a straightforward measure of a model's overall correctness. On the other hand, the Area Under Curve (AUC) is a performance measurement for classification problems at various threshold settings. AUC represents the degree to which the model is capable of distinguishing between classes. An AUC of 1 indicates perfect prediction, while an AUC of 0.5 suggests no discriminative power, equivalent to random guessing.

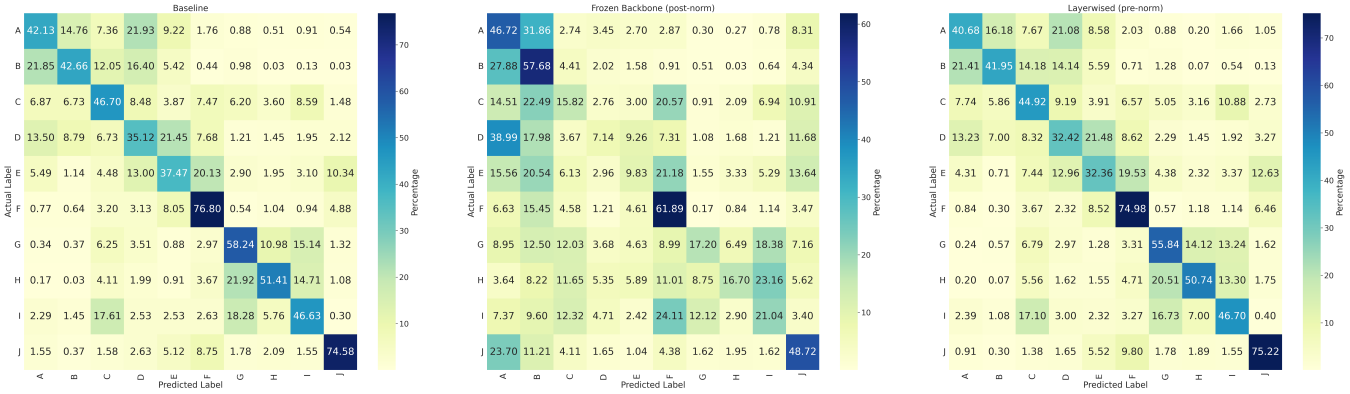
In this study, we only adopt Accuracy and AUC as performance metrics because our TAU Urban Acoustic Scenes 2022 Mobile dataset is a fully balanced dataset, where each class contains an approximately equal number of samples. This makes these two metrics highly appropriate, as they provide an accurate reflection of the model's performance without being skewed by class imbalances.

TABLE I: Performance of our proposed layer-wised SSAST and several comparing systems.

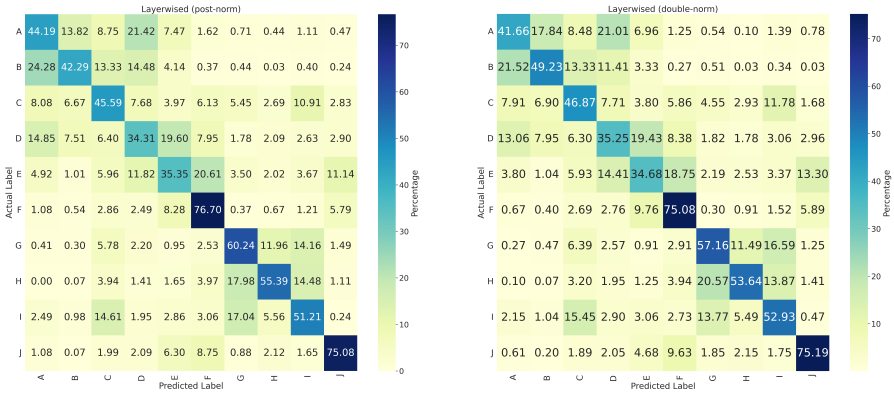
Model	Acc(%)	AUC(%)
Baseline	51.18	87.81
Frozen Backbone (post-norm)	30.27	73.11
Layer-wised (pre-norm)	49.58	87.06
Layer-wised (post-norm)	52.03	88.17
Layer-wised (double-norm)	**52.43**	**88.51**

The experimental results of layer-wised SSAST and several comparing systems are summarized in Table I. The layer-wised model with double-normalization SSAST outperforms the baseline SSAST by absolute 1.25% Acc and absolute 0.70% AUC. Our results support the hypothesis that by extracting information from all layers and implementing comprehensive fine-tuning, we can adapt pre-trained models to extremely challenging and divergent domain tasks effectively even with minimal efforts. Models with frozen backbones show a significant divergence between the SSAST pretraining domain and the DCASE ASC target domain, which training the classifier alone cannot address. Furthermore, the results of the layer-wised model with pre-normalization indicate that the importance assigned to each layer's output varies greatly. Normalizing these outputs before weighting may actually diminish the distinct information of different layer levels, resulting in lower classification performance. Post-normalization, applied after weighting the outputs, preserves the distinctiveness of features from each layer. This method enhances the model's ability to leverage inter-layer discriminative features, potentially improving the robustness and accuracy of the classification across varied and complex scenarios.

We also plot the confusion matrices for 1) Baseline, 2) Frozen Backbone with post-norm, and 3) three versions of layer-wise architecture, depicted in Figure 4. Our findings provide further validation of our initial premise. The confusion matrix from models with a Frozen Backbone reveals that the backbone itself is entirely unsuitable and incapable of facilitating effective classification, as shown in Fig. 4a and Fig. 4b. Moreover, when comparing the baseline with the layer-wised with pre-norm approach, there are no significant differences in classification outcomes for most categories, as shown in Fig. 4c. And, for the post-norm approach, as shown in Fig. 4d, there is a noticeable improvement at class G. However, for the double-norm approach, as shown in Fig. 4e, there is a noticeable improvement across those categories that with low accuracy, such as class B and C. This suggests that the overall architecture enables the model to extract features more effectively and utilize multi-layer features more efficiently for classification. Such an approach could enhance the fine-tuning performance of both the backbone and the classifier components of the model.



(a) Baseline (b) Frozen Backbone with post-normalization (c) Layer-wised with pre-normalization



(d) Layer-wised with post-normalization (e) Layer-wised with double-normalization

Fig. 4: Confusion matrices of the ASC systems examined. The labels from A to J correspond to: A - Airport, B - Indoor shopping mall, C - Metro station, D - Pedestrian street, E - Public square, F - Street with medium level of traffic, G - Travelling by a tram, H - Travelling by a bus, I - Travelling by an underground metro, J - Urban park.

V. Conclusions

In this paper, we proposed a novel ASC system named Layer wised SSAST. Based on our experiments and analysis, it is evident that the layer-wise approach significantly aids pre-trained models to be adapted for tasks where the data distribution undergoes severe changes and presents considerable challenges. This is particularly beneficial for extremely difficult tasks and datasets within the ASC, where the approach shows marked improvements in classifying similar categories. Our system enhances discriminative capability in similar acoustic scenes, improving performance in challenging ASC applications. This method achieved an accuracy of 52.43% and an AUC of 88.51%, representing an absolute increase of 1.25% in accuracy and an absolute increase of 0.70% in AUC.

This study enhances the adaptability of SSAST for ASC tasks, particularly in handling similar class distinctions, providing a new strategy for future ASC systems to improve classification performance in complex scenarios. The method also demonstrates SSAST's potential to better adapt to real-world ASC tasks, paving the way for future research to further expand its applications. Additionally, the research offers a promising solution for deploying complex models on resource-constrained devices, suggesting future work on low-complexity models and knowledge distillation techniques for efficient ASC implementations.

For the future works several promising directions for enhancing ASC systems using the Self-Supervised Audio Spectrogram Transformer (SSAST). Emphasizing low complexity models is crucial for devices with limited computational resources. Future studies should explore using SSAST for knowledge distillation to train efficient student models and leverage techniques like model compression, quantization, and pruning to reduce complexity [14]–[17]. Additionally, evaluating SSAST on datasets beyond urban environments, including rural, wilderness, and underwater soundscapes, could provide insights into its adaptability and generalization. This research is essential to understand the model's performance across different auditory environments [18]. Furthermore, enhancing current classification schemes by addressing overlaps and ambiguities in acoustic scene definitions could improve model precision. A more granular taxonomy of acoustic environments would ensure consistency and reliability across applications [19]–[21].

References

[1] S. Waldekar and G. Saha, "Two-level fusion-based acoustic scene classification," *Applied Acoustics*, vol. 170, p. 107 502, 2020.

[2] T. Virtanen, M. Plumbley, and D. Ellis, Eds., *Computational analysis of sound scenes and events*, English. Springer, 2018, ISBN: 978-3-319-63449-4. DOI: 10.1007/978-3-319-63450-0.

[3] B. Ding, T. Zhang, C. Wang, *et al.*, "Acoustic scene classification: A comprehensive survey," *Expert Systems with Applications*, vol. 238, p. 121 902, Oct. 2023. DOI: 10.1016/j.eswa.2023.121902.

[4] Q. Kong, L. Cances, Y. Wang, T. Iqbal, M. D. Plumbley, and W. Wang, "Passt: Efficient training of audio transformers with patchout," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[5] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 10 699–10 709.

[6] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: Generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop*, 2020, pp. 56–60. [Online]. Available: https://arxiv.org/abs/2005.14623.

[7] K. Koutini, J. Schluter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021, Available at https://arxiv.org/abs/2110.05069.

[8] F. Schmid, S. Masoudian, K. Koutini, and G. Widmer, "CP-JKU submission to dcase22: Distilling knowledge for low-complexity convolutional neural networks from a patchout audio transformer," DCASE2022 Challenge, Tech. Rep., Jun. 2022.

[9] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575. DOI: 10.21437/Interspeech.2021-698.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] J. F. Gemmeke, D. P. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2017, pp. 776–780.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2015, pp. 5206–5210.

[13] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[14] I. Martín-Morató, T. Heittola, A. Mesaros, and T. Virtanen, "Low-complexity acoustic scene classification for multi-device audio: Analysis of dcase 2021 challenge systems," *Applied Sciences*, vol. 11, no. 4, p. 1850, 2021.

[15] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2006, pp. 535–541.

[16] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," *Advances in Neural Information Processing Systems*, 2015.

[17] J. Gou, B. Yu, S. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[18] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ACM, 2014, pp. 1041–1044.

[19] X. Bai, J. Du, J. Pan, H. S. Zhou, Y. H. Tu, and C.-H. Lee, "High-resolution attention network with acoustic segment model for acoustic scene classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 656–660.

[20] S. Chandrakala and S. Jayalakshmi, "Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–34, 2019.

[21] T. Virtanen, M. D. Plumbley, and D. Ellis, "Computational analysis of sound scenes and events," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2018.

Non-blind Deblurring Using Probabilistic Models and Spatial Adaptive Restoration

Chun-Lin Liao [1], Jian-Jiun Ding [2*], and Chun-Jen Shih [3]
[1] National Taiwan University, Taipei, Taiwan
E-mail: r11942060@ntu.edu.tw Tel: +886-2-33663662
[2] National Taiwan University, Taipei, Taiwan
E-mail: jjding@ntu.edu.tw Tel: +886-2-33669652
[3] National Taiwan University, Taipei, Taiwan
E-mail: f10942045@ntu.edu.tw Tel: +886-2-33663662

***Abstract*— The core concept of image deblurring is to establish a prior model and formulate the deblurring process as an optimization problem by using the statistical properties of gradients and noise. In usual, a user-defined parameter is introduced to control the trade-off between the sharpness of the deblurring result and the robustness to noise. However, using a single variable is difficult to make the deblurring algorithm adapt to different types of images. In this work, we aim to overcome this limitation. The criteria to determine different regions can be obtained through noise estimating and construct a space-variant noise model. Utilizing different statistical characteristics of noise, we can approximate their distribution using various hyper-Laplacian models. Thus, we establish two lookup mechanisms. The final result interpolates the result of different regions to obtain a smoother outcome. Additionally, for the significantly degraded image with a low sign-to-noise ratio (SNR), we apply an adaptive denoiser in the frequency domain to stabilize the deblurred image.**

I. INTRODUCTION

Image deblurring has been a thoroughly studied topic for the past several decades. Here, we focus on a specific type of distortion: motion blur [1,2,3] combined with additive white Gaussian noise (AWGN) [4]. This type of distortion is commonly observed in handheld photography, where the camera and the subject experience relative movement during the exposure time. The presence of AWGN, on the other hand, mainly stems from the sensor while reading discrete data, making the restoration process more challenging. In our study, we aim to address the challenges posed by motion blur and AWGN.

To begin with, the process of image blurring can be formulated as follows:

$$\mathbf{y} = \mathbf{x} \otimes \mathbf{k} + \mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(0,\ \sigma^2) \tag{1}$$

where $\mathbf{x}$ denotes sharp image, $\mathbf{k}$ is the point spread function (PSF), also known as the bur kernel, $\mathbf{y}$ is the image after motion blurring, $\mathbf{n}$ is the additive white Gaussian noise (AWGN) with standard deviation σ, and the notation $\otimes$ indicates 2D convolution in spatial domain. All kinds of uniform motion blurring can be simply modeled by (1). The goal is to reconstruct latent image $\mathbf{x}$ based on the blurry input $\mathbf{y}$ and $\mathbf{k}$.

II. RELATED WORKS

Non-blind deblurring algorithms can be roughly categorized into conventional methods and deep learning-based methods. Conventional methods, such as the Wiener filter, are derived based on minimum mean square error (MMSE), while conjugate methods are derived based on least squares error (LSE). Recently, the most popular framework is maximum a posteriori (MAP), which, through Bayes' theorem, maximizes the posterior probability $\mathrm{P}(\mathbf{x}|\mathbf{y},\mathbf{k})$ by maximizing the product of the likelihood $\mathrm{P}(\mathbf{y}|\mathbf{x},\mathbf{k})$ and the image prior $\mathrm{P}(\mathbf{x})$. A classic non-deep learning method utilizes a hyper-Laplacian function as the image prior [5], aiming to solve the following objective function

$$\min_{\mathbf{x}} \sum_{i=1}^{N} \left(\frac{\lambda}{2} (\mathbf{x} \otimes \mathbf{k} - \mathbf{y})_i^2 + \sum_{j=h,w} \left| (\mathbf{x} \otimes f_j)_i \right|^{\alpha} \right) \tag{2}$$

where N denotes the total number of pixels in the image, and f_h and f_w represent the vertical and horizontal gradient filters, respectively. The work suggests that the deblurring problem can be well modeled by the summation of a likelihood term and a hyper-Laplacian distribution prior $e^{-|\mathbf{x} \otimes f|^{\alpha}}$ with two adjustable parameters, α and λ.

To solve (2), half quadratic splitting (HQS) is introduced to separate the objective function into two sub-problems: deconvolution and image regularization. In the deconvolution phase, the latent image is obtained by direct FFT, while the image regularization phase is achieved through an analytic solution. The restored image is obtained by iteratively solving these two sub-problems.



Fig. 1. Deconvolution results from different λ fixing α.



Fig. 2. Deconvolution results from different α fixing λ.

Other non-DL methods follow a similar solving architecture. For example, BM3D-based deblurring [6, 7] utilizes the synthesis features of the BM3D framework, while outlier handling [8] removes saturated pixels during the deblurring process. With the advent of deep learning, methods have shifted. Instead of using the Hyper-Laplacian prior as in [7], the FCN method [12] suggests learning the image prior using a fully convolutional network. The IRCNN method [13], on the other hand, frames the image regularization problem as a denoising problem, solving it with a residual network. These methods are based on single optimization problems. The NBDN [9] trains three separate models for restoration degree, image regularization, and a mask determining the probability of saturation. In INFWIND [10], deblurring is no longer treated as a single optimization problem; instead, it integrates the Wiener filter with image preprocessing and postprocessing.

III. Motivation

For non-DL methods, a single optimization problem restricts the degree of restoration in different regions of an image. As for the DL based methods, we find that for low SNR images, the learned priors are either ineffective or require the noise standard deviation to be known. Hence, blind denoising plays a crucial role in low SNR images deblurring. Our work is highly related to [5]. We first make a few observations about the hyper-Laplacian prior. In (2), the restoration depends on two parameters: λ and α . λ controls the sharpness of the deconvolution result. A greater λ will make the image prior term subtler, resulting in a sharper yet noisier image.

Additionally, α controls the curvature of the image prior and also affects the outcomes. A greater α indicates a more gradual decline in the gradient model, which corresponds to a sharper deconvolution.

We aim to replace the single hyper-Laplacian model with more accurate regularization terms according to local variance. The following section illustrates how we construct a spatially adaptive model, address the ringing effect, and provide the complete algorithm at the end.

IV. Proposed Method

A. Spatial Adaptive Probabilistic Model

In [5], the gradient is simply the difference between two horizontally and vertically adjacent pixels, i.e., $f_h = [1 -1]$, $f_w = [1 -1]^T$. Here, we use the Roberts cross filter [15] to calculate the gradient because the Roberts cross filter maximizes the KL divergence [14] of the gradient distribution between sharp images and blurry ones among conventional edge filters. Therefore, the filter in [5] is replaced by

$$f_h = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \; f_w = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{3}$$

Generally speaking, when choosing λ, we consider two factors: the SNR of the blurry image and the local variance. For low SNR images, we want a smaller λ to increase the proportion of the image prior term. This supports the idea that the image prior term acts as a denoiser. In other words, denoising is more important than deblurring for low SNR images. On the other hand, local variance also determines the degree of restoration. For image patches with low variance, we can assume that the patch is almost constant. In these smooth patches, we apply strong regularization, forcing the gradients to be close to zero. We establish a 2D lookup table that uses noise standard deviation σ_n and local variance v_i to determine the desired λ. We use a rule of thumb to find the relationship between σ_n, v_i, and optimal λ_o. The experimental data consists of 80 undistorted real-world images. Each image is randomly blurred by one of the kernels in Levin [16] and AWGN with σ_n. We then deblur the blurry images using five different λ values: 0, 4, 200, 1000, and 10000. Note that when $\lambda = 0$, it means that only denoising is performed. Next, we calculate the local variance of the original image using a 9×9 block. For each patch $\mathbf{P}^i$ with variance v_i, we compare $\mathbf{P}^i$ with the patch in the same location in the five previously generated results using PSNR. By choosing the patch with the highest PSNR, we can find λ_o^i under σ_n and v_i.

Then, we partition the variance into 16 intervals with the following boundaries: 0, 1, 4, 8, 16, 30, 53, 88, 142, 212, 312, 449, 650, 961, 1467, and infinity. In each interval, there are several pairs of patches $\mathbf{P}_i$s with v_i falling within this range. For the parameters λ_o^is corresponding to these patches, we count the most frequent λ_o^i in this section and set it as the optimal lambda λ_o for the current variance section. To this end, we create a 1-D lookup table from local variance v_i to optimal λ_o under AWGN with noise standard deviation σ_n.

By repeating the above procedure for four different noise standard deviation, where $\sigma_n = 1, 4, 7,$ and 10, we obtain a scatter plot of λ_o versus σ_n and v_i, i.e., the blue dots in Fig. 3. Based on the experiment result, we use the following formula to model the scatter points

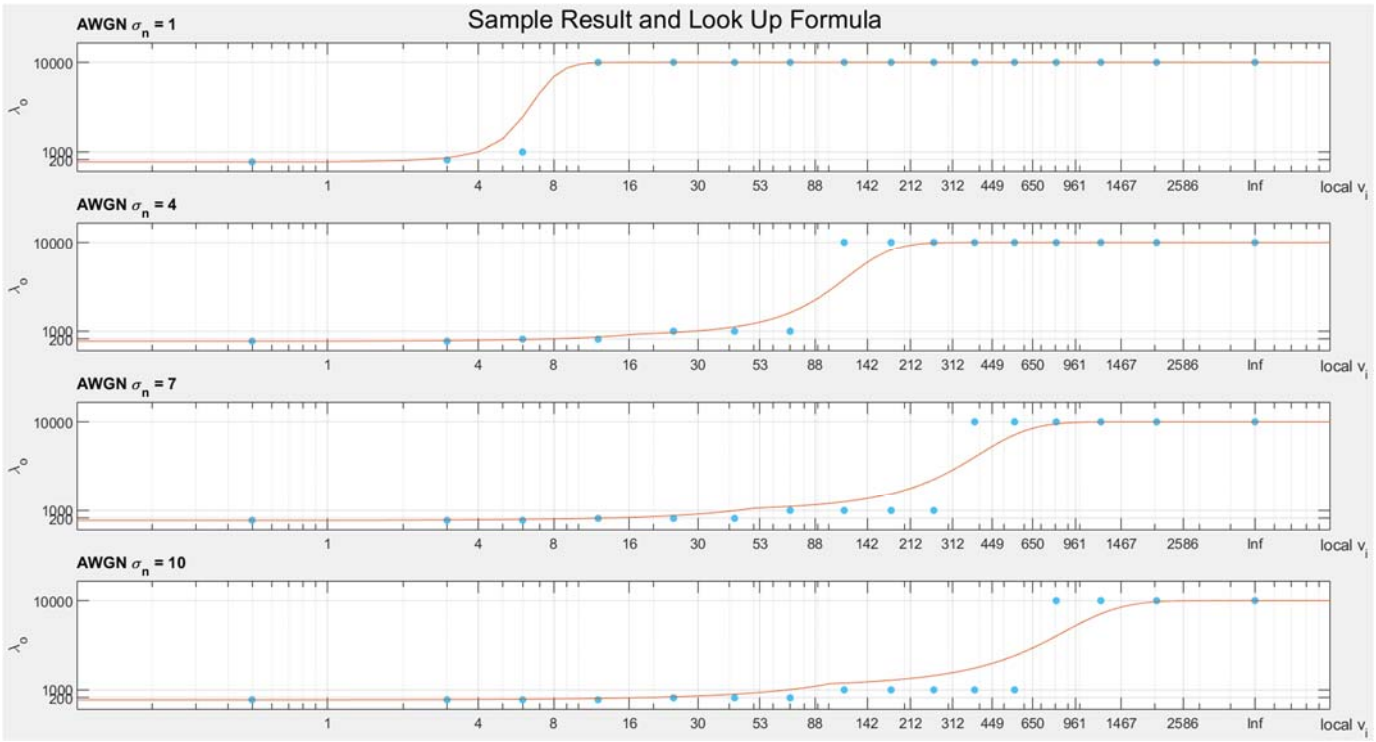


Fig. 3. Proposed 2D lookup table.

$$\lambda_{HighVar} = \frac{10000}{1+e^{\sqrt{\frac{-t}{\sigma_n^5}}}}, \quad t = v_i - (6.32 \cdot \sigma_n^2), \tag{4}$$

$$\lambda_o = \max\left(\lambda_{HighVar}(1-C), 0\right), \; C = \max\left(\frac{\sigma_n^2-(v_i-1)}{\sigma_n^2}, 0\right). \tag{5}$$

We use a sigmoid function for the high-variance intervals. The parameter C regulates the low-variance sections, ensuring that λ_o remains close to zero regardless of σ_n^2. Additionally, when v_i is less than 1, λ_o is set to 0, indicating that no deconvolution is needed. The 2D lookup table generated using (3) and (4) is shown in Fig. 3. by the red line.

So far, we have proposed a complete 2D lookup table for λ_o. To further improve the accuracy of our model, we adjust the hyper-Laplacian model [5] for different variance sections. We reintroduce the parameter b, which was neglected in [5], to achieve a more flexible model. Then, (2) is rewritten as

$$\min_{\mathbf{x}} \sum_{i=1}^{N} \left(\frac{\lambda}{2} (\mathbf{x} \otimes \mathbf{k} - \mathbf{y})_i^2 + \sum_{j=h,w} \frac{\left|(\mathbf{x} \otimes f_j)_i\right|^{\alpha}}{b} \right) =$$

$$\min_{\mathbf{x}} \sum_{i=1}^{N} \left(\frac{\lambda b^{\alpha}}{2} (\mathbf{x} \otimes \mathbf{k} - \mathbf{y})_i^2 + \sum_{j=h,w} \left|(\mathbf{x} \otimes f_j)_i\right|^{\alpha} \right). \tag{6}$$

We sampled a total of 10,120 patches and classified them into the variance intervals mentioned above. For each variance section, we gather all the patches and used a pair of (α, b) to approximate the gradient distribution. Taking the local variance interval [312, 449] as an example, the optimal hyper-Laplacian prior is $\alpha = 0.9$ and $b = 11.9$.

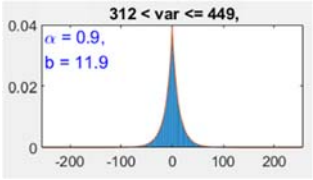


Fig. 4. Histogram of the gradient of loval variance between 312 and 449.

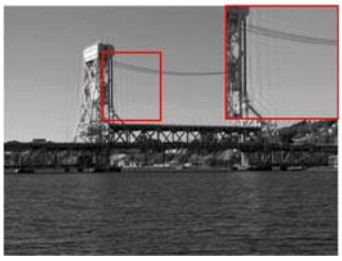


Fig. 5. Resulting image **I** from [5] that exhibits ringing effect.



Fig. 6. (Left) Original image patch; (Right) After ringing effect removal.

B. Ringing Effect Suppression

An effective ringing effect method is to apply local subvoxel-shifts [17]. For every patch in the image, alternative sampling that minimizes the difference between adjacent pixels is applied. Here, we propose an effective way of removing the ringing effect during the deconvolution phase.

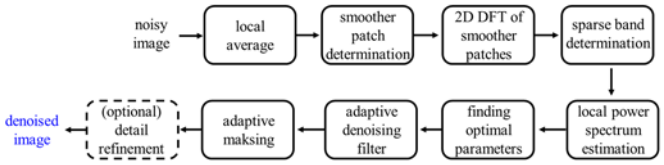


Fig. 7. Flowchart of the adaptive blind denoising algorithm.

For gradient values below a specific threshold (3 in our implementation), we smooth the ringing-distorted image across three scales: 100, 50, and 25. At the first scale of 100, we search for any 100 × 100 area where all gradients are zero. If such an area exists, we apply smoothing to the corresponding area. This procedure is then repeated for scales of 50 and 25. Next, we perform regional smoothing by solving the least-squares approximation using orthogonal projection matrices. Taking the scale of 100 as an example, for the top-left pixel in ringing patch, denoted as p_1, its x-index x_1 and y-index y_1 are set to 1. As we move from left to right, the x-index increases to 100, and as we move from top to bottom, the y-index also increases to 100. This approach provides a total of 10,000 3D data points, i.e., $[p_i, x_i, y_i]$. In other words, we aim to minimize

$$\|\mathbf{p} - (a_2\mathbf{v}_2 + a_1\mathbf{v}_1 + a_0\mathbf{v}_0)\|^2, \tag{7}$$

where $\mathbf{p}$, $\mathbf{v}_2$ and $\mathbf{v}_1$ corresponds to the vectorization of P_r, x-index and y-index, i.e.,

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{10^4} \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10^4} \end{bmatrix}, \mathbf{v}_1 = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{10^4} \end{bmatrix}, \mathbf{v}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{8}$$

Next, we concatenate $\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2$ to form a matrix $\boldsymbol{C}$=$[\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2]$. Following the orthogonal projection theorem, the coefficients of the estimated plane are given by

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\mathbf{p}, \tag{9}$$

$$P_{\text{smooth}} = a_2 \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 100 & \cdots & 100 \end{bmatrix} + a_1 \begin{bmatrix} 1 & \cdots & 100 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 100 \end{bmatrix} + a_0. \tag{10}$$

C. *Blind Denoising*

Since denoising becomes crucial in deblurring low SNR images, we propose a blind denoising method that can restore images degraded by frequency-varying noise. The overall flowchart of the algorithm is shown here without delving into details. The proposed method first estimates the noise spectrum, then performs local denoising using an adaptive Wiener filter. Our method can denoise a 512x512 grayscale image in under 0.5 seconds regardless of noise level, making it robust for deblurring, especially in low SNR images.



Fig. 8. (Left) Deblurred result without denosing; (Right) Apply denoising.

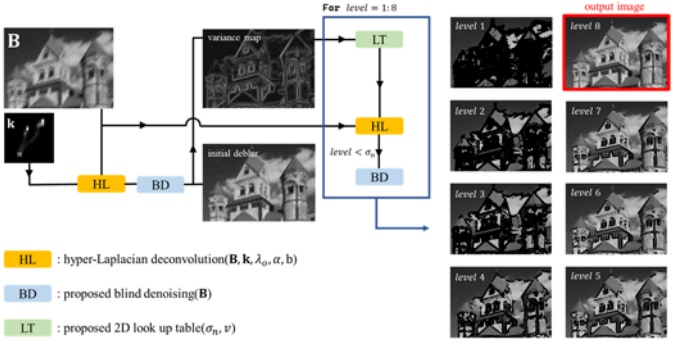


Fig. 9. Overall architecture of proposed non-blind deblurring algorithm.

Algorithm 1 Adaptive Deblurring Using 2D Look Up Table

Input: blurry image $\mathbf{B}$,
blur kernel $\mathbf{k}$,
noise std σ_n
Output: deblurred image S

```
X_pre ← proposed_denoise(B)
X_var ← variance(fast_deconv(B, 6000, 0.5))
for i = 1 : 8 do
    X_map ← v_i ≤ X_var < v_{i+1}
    λ_o ← LookUp(v_i, σ_n)
    X_l ← (fast_deconv(B, λ_o, α_i, b_i))
    if i < σ_n then
        X_l ← proposed_denoise(X_l)
        X_l ← proposed_ringing_removal(X_l)
    end if
    S ← S + X_map ∘ (R ∘ X_l + (1 − R) ∘ X_pre)
    X_pre ← X_l
end for
```

Fig. 10. proposed non-blind deblurring algorithm

D. *Schematic of Implementation*

Finally, we provide a complete framework of the proposed non-blind deblurring algorithm along with a few implementation details. The variance map X_{pre} of the blurry image is calculated by the image with initial denoising and deblurring. Additionally, the output at each variance level is interpolated by a sigmoid-like function to prevent the image from having discontinuous transitions.

$$R = 1/(1 + \exp(-\frac{24(v - \frac{v_{i+1}+v_i}{2})}{v_{i+1}-v_i})) \tag{11}$$

Also, we perform denoising and ringing effect removal only at levels where the variance is less than the noise standard deviation to avoid losing image details. In other words, if the image is not degraded by noise, we will not apply any denoising.

TABLE I. DEBLURRING PERFORMANCE UNDER $\sigma_n = 2$

PSNR	Blackboard (text)	Fox (Animal)	Boat (Night)	Man (Face)	Beefnoodle (Food)	**Average**
BM3D [6, 7]	**31.3842**	23.8940	21.9942	17.2226	20.0036	22.8997
Fast [5]	26.1323	34.0253	33.1821	30.1901	31.7080	31.0476
DMSP [11]	20.9645	28.7640	29.1063	25.4472	25.4122	25.9388
FCN [12]	27.1441	30.9976	30.8542	29.4330	29.6862	29.6230
IRCNN [13]	29.3836	33.0348	32.4839	29.7949	32.1332	31.3661
NBDN [9]	30.6898	34.6576	32.4464	29.7519	24.5466	30.4185
INFWIDE [10]	28.4646	34.0638	25.8255	28.3207	26.3456	28.6040
Proposed	30.8523	**34.9007**	**34.1105**	**31.7858**	**33.5533**	**33.0405**

TABLE II. DEBLURRING PERFORMANCE UNDER $\sigma_n = 5$

PSNR	Blackboard (text)	Fox (Animal)	Boat (Night)	Man (Face)	Beefnoodle (Food)	**Average**
BM3D [6, 7]	26.9748	27.9798	25.4035	20.9018	24.0873	25.0694
Fast [5]	25.3035	29.1439	28.5934	27.2970	28.0058	27.6687
DMSP [11]	20.8497	28.4093	28.7898	25.2838	25.3031	25.7271
FCN [12]	21.1616	26.8437	27.8374	24.5312	24.2157	24.9179
IRCNN [13]	27.0281	26.8644	25.9539	25.2703	25.7853	26.1804
NBDN [9]	**27.5403**	29.0457	27.1494	23.6711	19.5285	25.3870
INFWIDE [10]	27.2039	32.1745	27.4825	27.4527	26.1158	28.0859
Proposed	26.2209	**32.2095**	**32.1896**	**29.2618**	**31.2045**	**30.2173**

TABLE III. DEBLURRING PERFORMANCE UNDER $\sigma_n = 10$

PSNR	Blackboard (text)	Fox (Animal)	Boat (Night)	Man (Face)	Beefnoodle (Food)	**Average**
BM3D [6, 7]	23.8607	30.1757	27.9975	24.1482	26.5962	26.5557
Fast [5]	23.3875	24.1167	23.4251	23.0462	23.3501	23.4651
DMSP [11]	20.9026	**29.1435**	29.5945	26.0530	25.9833	26.3354
FCN [12]	21.2007	26.0540	27.3009	24.0345	23.6572	24.4495
IRCNN [13]	20.3963	16.9625	15.2924	15.9559	16.0320	16.9278
NBDN [9]	23.8940	21.7628	15.4941	18.1882	18.7295	19.6137
INFWIDE [10]	**25.8669**	27.3311	28.3032	25.9292	24.1002	26.3061
Proposed	25.4815	29.0070	**29.7646**	**27.5719**	**29.4297**	**28.2509**

Conversely, if the AWGN has a standard deviation greater than 8, we perform denoising at all levels. The complete process is outlined as in Fig. 10.

V. EXPERIMENTS

We examine our work by comparing the deblurring performance of motion blur with other methods under three different noise standard deviation. The test dataset comprises a total of five distinct image categories: text images, animal images [19, 20], night-time images [21], face images [18], and food images.



Fig. 11. Effect of ringing artifact removal.



Fig. 12. Comparison of the details of the deblurring outputs.

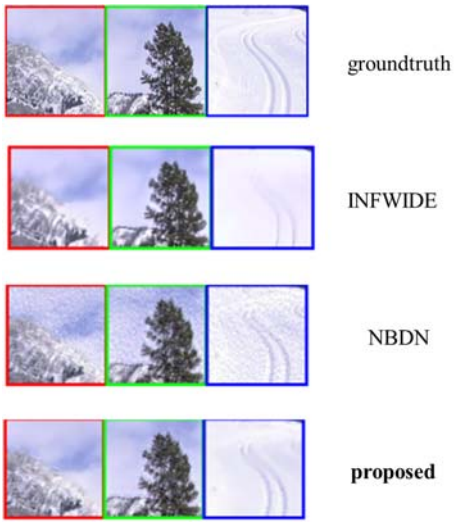


Fig. 13. Comparison of the details of the deblurring outputs.

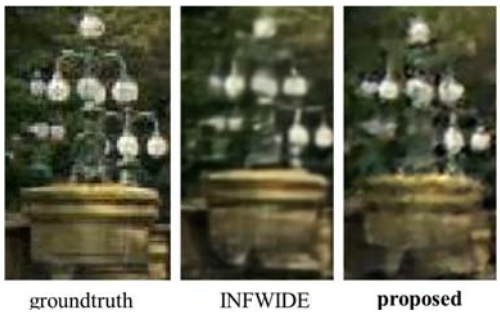


Fig. 14. Comparison of the detail of deblurring outputs.

We use the peak signal-to-noise ratio (PSNR) as our evaluation metric. For a sharp image $\mathbf{y}[m,n]$ and deblurred image $\mathbf{x}[m,n]$, the metrics is evaluated as follows:

$$\text{PSNR} = 10\log_{10}\frac{255^2}{\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}|\mathbf{y}[m,n]-\mathbf{x}[m,n]|^2}. \quad (12)$$

From Tables I, II, and III, our method consistently achieves the best average performance among state-of-the-art techniques across three different scenarios.

At the first noise level ($\sigma_n = 2$), the results from these methods are similar. However, upon closer examination of

image details, our method, which includes ringing effect removal, effectively eliminates the ripple on the roof that the DMSP does not address. Additionally, compared to the INFWIDE, our method better preserves texture details, such as the carvings on the wall.

At the second noise level, where AWGN begins to influence the results, shows that our method preserves tire tracks while also performing denoising. In contrast, the INFWIDE is more likely to sacrifice details for a more stable scene structure, and the NBDN does not specifically address denoising.

In high-noise situations ($\sigma_n = 10$), the results in Table III, show that our method tends to produce sharper edges, while the INFWIDE trades sharpness for overall image stability.

Moreover, in Figs. 11-14, the details of the deblurring outputs of the proposed method and other methods are compared. One can see that, with the proposed methods, the edges and texture of the object can be effectively reconstructed. Meanwhile, the proposed method can well avoid the ringing effect and the noise effect.

VI. Conclusion

An advanced probability-model based deblurring algorithm based on the local variance and the optimal degree estimation was proposed. For low SNR images, we developed an efficient denoising method that handles noise varying in both space and frequency domains. Compared to state-of-the-art non-blind deblurring techniques, the proposed algorithm gives consistent performance across various types of degradation.

References

[1] H. Ji and C. Liu, "Motion blur identification from image gradients," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2008, doi: 10.1109/CVPR.2008.4587537.

[2] G. Boracchi and A. Foi, "Modeling the performance of image restoration from motion blur," *IEEE Trans. Image Processing*, vol. 21, no. 8, pp. 3502-3517, 2012.

[3] Shengyang Dai and Ying Wu, "Motion from blur," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2008, doi: 10.1109/CVPR.2008.4587582.

[4] C. Boncelet, "Image noise models," in *The Essential Guide to Image Processing*. Academic Press, Chap. 7, pp. 143-167, 2009.

[5] D. Krishnan and R. Fergus, "Fast Image Deconvolution using Hyper-Laplacian Priors," *Neural Information Processing Systems*, vol. 22, pp. 1-9, 2009.

[6] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Processing*, vol. 16, no. 8, pp. 2080-2095, 2007.

[7] A. Danielyan, V. Katkovnik, and K. Egiazarian, "BM3D frames and variational image deblurring," *IEEE Trans. Image Processing*, vol. 21, no. 4, pp. 1715-1728, April 2012.

[8] S. Cho, J. Wang, and S. Lee, "Handling outliers in non-blind image deconvolution," in *Int. Conf. Computer Vision*, pp. 495-502, 2011.

[9] L. Chen, J. Zhang, J. Pan, S. Lin, F. Fang, and J. S. Ren, "Learning a non-blind deblurring network for night blurry images," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 10542-10550, 2021.

[10] Z. Zhang, Y. Cheng, J. Suo, L. Bian, and Q. Dai, "INFWIDE: Image and feature space wiener deconvolution network for non-blind image deblurring in low-light conditions," *IEEE Trans. Image Processing*, vol. 32, pp. 1390–1402, 2023.

[11] S. A. Bigdeli, M. Zwicker, P. Favaro, and M. Jin, "Deep mean-shift priors for image restoration," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1-7, 2017.

[12] J. Zhang, J. Pan, W. S. Lai, R. Lau, and M. H. Yang, "Learning fully convolutional networks for iterative non-blind deconvolution," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3817-3825, 2017.

[13] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3929-3938, 2017.

[14] S. Kullback and R. A. Leibler, "On Information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[15] F. Nielsen, *Visual Computing: Geometry, Graphics, and Vision (Graphics Series)*, Charles River Media/Thomson Delmar Learning, 2005.

[16] A. Levin, Y. Weiss, F. Durand and W. T. Freeman, "Understanding and evaluating blind deconvolution algorithms," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1964-1971, 2009.

[17] E. Kellner, B. Dhital,V. G. Kiselev, and M. Reisert, "Gibbs-ringing artifact removal based on local subvoxel-shifts," *Magnetic Resonance in Medicine*, vol. 76, issue 5, pp. 1574-1581, 2016.

[18] T. Karras, S. Laine, and T. Aila, ""A style-based generator architecture for generative adversarial networks," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 4401-4410, 2019.

[19] A. Kumar, "Bird species classification," Kaggle, 2019. [Online] Available: https://www.kaggle.com/datasets/akash2907/bird-species-classification.

[20] S. Banerjee, "Animal image dataset (90 different animals)," Kaggle, 2022. [Online] Available: https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals.

[21] A. Ye, "ExDark dataset," Kaggle, 2022. [Online] Available: https://www.kaggle.com/datasets/washingtongold/exdark-dataset.

A Pilot Study of Applying Sequence-to-Sequence Voice Conversion to Evaluate the Intelligibility of L2 Speech Using a Native Speaker's Shadowings

Haopeng Geng, Daisuke Saito, Nobuaki Minematsu
Graduate School of Engineering, The University of Tokyo
E-mail: {kevingenghaopeng, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

***Abstract*—Utterances by L2 speakers can be unintelligible due to mispronunciation and improper prosody. In computer-aided language learning systems, textual feedback is often provided using a speech recognition engine. However, an ideal form of feedback for L2 speakers should be so fine-grained that it enables them to detect and diagnose unintelligible parts of L2 speakers' utterances. Inspired by language teachers who correct students' pronunciation through a voice-to-voice process, this pilot study utilizes a unique semi-parallel dataset composed of non-native speakers' (L2) reading aloud, shadowing of native speakers (L1) and their script-shadowing utterances. We explore the technical possibility of replicating the process of an L1 speaker's shadowing L2 speech using Voice Conversion techniques, to create a virtual shadower system. Experimental results demonstrate the feasibility of the VC system in simulating L1's shadowing behavior. The output of the virtual shadower system shows a reasonable similarity to the real L1 shadowing utterances in both linguistic and acoustic aspects[1].**

I. Introduction

Feedback from native speakers or language teachers is essential in second language acquisition (SLA), as it can improve the communicative skills of L2 speakers. With the assistance of Large Language Models (LLM) agents, it has become easier for L2 speakers to have their writing proof-read with detailed suggestions. However, L2 speakers still face challenges in receiving word-by-word feedback on their speaking performance, especially when educational resources are limited.

In such cases, Computer-Aided Language Learning (CALL) has proven to be effective. For example, an Automatic Speech Recognition (ASR) engine can generate transcripts of an L2 speaker's utterances, helping them understand how the system processes their speech and identify any unintelligible words. Recent advances in ASR, leveraging LLM and sophisticated feature embedding, have achieved promising performance, with Word Error Rates (WER) below 10% for Global Englishes [1]–[3].

Nevertheless, previous research indicates that relying solely on ASR for speech intelligibility evaluation is not suitable for providing feedback to L2 speakers [4]. Textual feedback alone is often insufficient for L2 speakers to identify unintelligible segments, especially in spontaneous situations, as the system has no access to the ground truth reference which is considered to exist only in the speaker's mind. Moreover, recent ASR systems are specifically designed to predict, or even over-speculate, what speakers actually said, which is contrary to the purpose of CALL systems. [5], [6]. Besides, research indicates that L2 speakers often overestimate their speech intelligibility [7], [8] and replaying their utterances is generally less effective than anticipated in helping them recognize unintelligible parts of their speech. In the worst case, replaying L2 speakers' speech may fossilize their pronunciation.

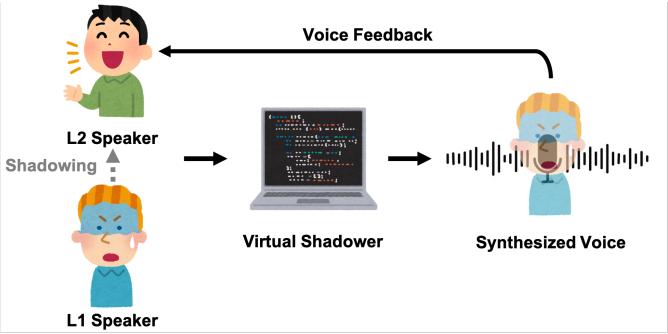


Fig. 1: The concept of the virtual shadower, which simulates the shadowing behaviors of an L1 rater hearing a given L2 speech for the first time. Stuttering or inarticulate production of speech may occur due to listening disfluencies.

How can we expose those unintelligible parts in L2 speech? Native speaker shadowing, wherein an L1 speaker repeats an L2 utterance with as short a delay as possible while listening, is proposed as an advanced method for evaluating L2 speech intelligibility [9]–[11]. In this approach, the L1 speaker immediately repeats what they hear in the given L2 speech using their own accent. Any stuttering or delivery of unrelated words, known as shadowing disfluencies, indicates where listening breakdowns occur. Consequently, shadowing by the L1 speaker highlights the problematic parts of the L2 speech, enabling L2 speakers to know where their speech may be unclear or difficult to understand. However, it is impractical to provide a real shadower for every L2 speaker. In this study, as shown in Figure 1, we aim to develop a virtual shadower system that simulates the process of an L1 speaker shadowing L2 speech

[1]Audio samples are available at: https://secondtonumb.github.io/publication_demo/APSIPA_2024/index.html

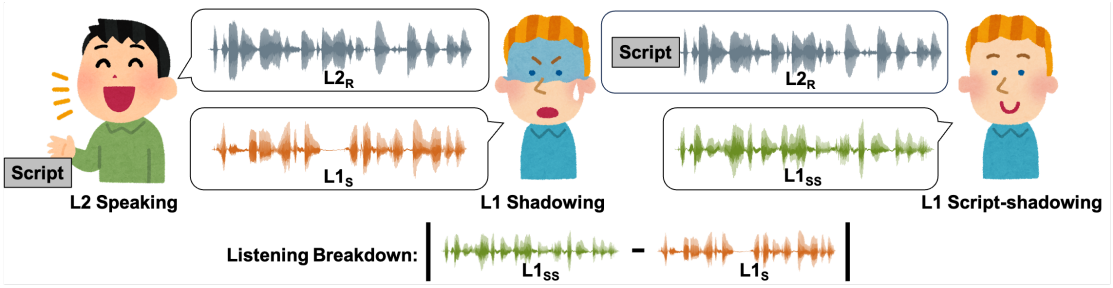


Fig. 2: The proposed shadowing technique aims to identify unintelligible parts in an L2 speaker's reading aloud utterances ($L2_R$). In this approach, $L1_{S1}$ represents a native speaker's initial shadowing, while $L1_{SS}$ denotes the script-shadowing by the native speaker. By calculating the distance between $L1_{S1}$ and $L1_{SS}$, it is possible to pinpoint the native speaker's listening breakdowns, which correspond to the unintelligible parts in the $L2_R$ as well.

(L1-shadowing-L2), thereby offering more comprehensible and constructive feedback for L2 speakers.

Our research contributions are as follows:

- We examine the similarity between an L1 shadower's behavior and Voice Conversion (VC) alignment, utilizing this similarity to construct a virtual L1-shadowing-L2 system, which simulates an L1 speaker's shadowing L2 speech. To the best of our knowledge, this is the first work to apply VC to the L2 intelligibility task.
- We examine the feasibility of using L1 shadowing speech data, a form of semi-parallel data, to develop a VC model. Experimental evaluations on both linguistic and acoustic aspects demonstrate reasonable feasibility in constructing an L1-shadowing-L2 system with Seq2Seq VC.

II. Research Background

A. Native speaker shadowing

Fine-grained annotation of the intelligibility of L2 speech is such a challenging task that it requires specialist knowledge in phonetics for phoneme-level labeling. To address this problem, [10] proposed a two-stage reverse form [2] of shadowing to identify unintelligible parts in L2 utterances. As shown in Figure 2, an L1 speaker first shadows a given L2 utterance alone ($L1_{S1}$). During this process, unintelligible parts result in stuttering or inarticulate production. Following $L1_{S1}$, the L1 speaker performs script-shadowing ($L1_{SS}$), where the L2 script is presented visually along with the L2 utterance. By comparing $L1_{S1}$ and $L1_{SS}$ using Dynamic Time Warping (DTW), we obtain sequential data on broken articulation—shadowing disfluencies that indicate listening disfluencies—based on the distance between the two.

Previous research has shown that alignment based on phonetic posteriorgrams (PPG) performs better than acoustic features such as MFCC or Mel-Cepstrum in assessing pronunciation quality [12]. Additionally, PPG-based DTW between $L1_{S1}$ and $L1_{SS}$ can derive word, syllable, and phoneme-level annotations of intelligibility, as also shown in [11].

[2]In the field of SLA, shadowing typically involves learners repeating L1 speech to enhance their listening skills. However, in our study, learners are shadowed by L1 raters. Thus, we refer to our method as reverse shadowing.

B. Seq2Seq voice conversion

Conventional VC aims to change non-/para-linguistic features while preserving the linguistic content of input speech. However, with the advent of end-to-end architectures and self-supervised speech representations, recent studies on Seq2Seq VC enable a more robust mapping of sophisticated latent features between source and target voices. This includes tasks such as Speaking Style Conversion [13], Voice Emotion Conversion [14], and Foreign Accent Conversion [15], [16].

In [17], the authors proposed the Voice Transformer Network (VTN) that utilizes a Transformer-based Seq2Seq model and employs mel-spectrograms as input features [18]. In [19], the authors modified the input acoustic feature to linguistic representation, resulting in a significant reduction of accentedness. To address the issues of inaccurate duration prediction and repetitive artifacts caused by the auto-regressive nature of transformer models, a non-autoregressive model was proposed using a conformer structure to mitigate these problems [20], [21]. Recent works have further improved upon this by using a Monotonic Alignment Search (MAS) and joint vocoder training, outperforming non-autoregressive models in duration and prosody [22]. Furthermore, a more robust Automatic Alignment Search (AAS) method, derived from a Text-to-Speech (TTS) system, has been developed to address challenges in low-resource settings [23], [24].

In the following sections, these recent advancements in Seq2Seq VC are effectively adapted to implement our purposed L1-shadowing-L2 system.

III. Experimental Settings

A. Representing listening breakdowns with attention alignment failure

Shadowing is a voice-to-voice process in which the listener's instantaneous identification of words is crucial. However, smooth shadowing does not always occur, as listening breakdowns can happen when an L1 listener encounters unintelligible segments while shadowing L2 speech [25]. To explore the potential of VC to build virtual shadowers, Figure 3 shows our pilot experiment on Seq2Seq VC using semi-parallel data, $L2_R$ and $L1_{S1}$. We observe that the decoding path in the $L2_R$-$L1_{S1}$ VC system closely resembles the PPG-DTW editing path. This

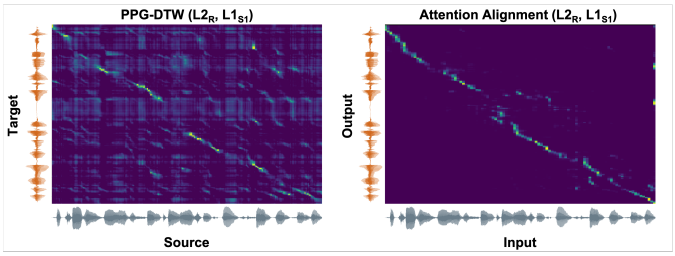


Fig. 3: The similarity of attention alignment to PPG-DTW is illustrated. The left figure shows the attention alignment observed in the inference phase of converting $L2_R$ to $L1_{S1}$, while the right figure shows the PPG-DTW path between $L2_R$ and $L1_{S1}$. Both figures exhibit a prominent diagonal contour.

inspires us to explore whether a VC system can be extended to explicitly represent L2 unintelligibility. If alignment failure (weak or faded-out attention) indicates listening breakdowns, we suggest that such a virtual shadower can be developed to resemble a real human L1 shadower.

B. Experimental setups

1) VC models: To ensure that the VC system's alignment failure is attributed to data mismatch rather than the model's limited alignment capabilities (as discussed in Section III-A), we evaluated two VC models in this study. Alongside the baseline Transformer-based model, VTN [26], we tested a robust alignment method, AAS, proposed in AAS-VC [23].

2) Data preparation: In this study, we utilized a reverse form of shadowing dataset mentioned in [11], [12]. With different degrees of accent, 225 Japanese English speakers' reading-aloud utterances $L2_R$ were collected. We then recruited one male native English speaker who shadowed and script-shadowed all these recordings as $L1_{S1}$ and $L1_{SS}$ respectively. Although each recording session lasted for more than 30 seconds, to avoid potential alignment failure due to lengthy input, we trimmed the original recordings sentence by sentence using forced alignment. This process resulted in 2,695 triplets of $\{L2_R, L1_{S1}, L1_{SS}\}$. Each dataset comprised an average of 3.9 hours of valid phonation, with the average duration of each sentence being 5.0 seconds. 300 utterances were selected as test sets.

3) Source-target selection: To determine the optimal setting for an L1 virtual shadower simulator, we designed three different source-target pairings:

- $L2_R$-$L1_{S1}$: This is the most straightforward approach for building a virtual shadower, as the training target is $L1_{S1}$ itself. However, it is also the most technically challenging setting, as both the linguistic context and speaker identity differ.
- $L2_R$-$L1_{SS}$: This is the typical setting for parallel-data VC. Since the source and target are intended to speak the same content, we expect this setting to reveal the pronunciation distance between L2 and L1.
- $L1_{SS}$-$L1_{S1}$: This unique training setting of the reverse shadowing dataset focuses on the listening breakdowns of the L1 shadower. We expect that this setting will reveal the disfluencies exhibited during the shadowing behavior of L1.

4) Feature embedding and pretrained models for decoding: Speaker-independent features enable stable convergence in our many-to-one VC setting. For feature embedding, we used the original PPG-like bottleneck feature extractor designed by [27]. The framewise dimension of latent feature is 144, and the features were normalized to zero mean and unit variance for embedding.

As for the PPG-to-Spec decoding, we trained a single-speaker PPG-to-Spec decoder following the implementation described in [28], where utterances were collected from the given L1 speaker.

For the vocoder, we employed Parallel-WaveGAN [29], which was trained on joint $L1_{S1}$ and $L1_{SS}$. The input Mel spectrogram has 80 frequency bins with a hop size of 256, and the target sample rate is 16 kHz.

Note that for VTN, both the PPG-to-Spec decoder and vocoder were utilized in the decoding phase. For AAS-VC, we followed the original implementation to map the source PPG to the target Mel-spectrogram, which generates the target's prosodic features better as previous study discussed [23].

IV. Experimental Evaluation

A. Metrics for evaluation

To objectively evaluate our virtual shadower, we focused on two key aspects. First, the linguistic content of the virtual shadower should closely match that of the L1 human shadower. Second, the acoustic similarity of the virtual shadower should resemble that of the L1 shadower. In this pilot study, we assessed the linguistic output of our virtual shadower only holistically. For the acoustic aspect, we calculated the segmental and prosodic similarities between the virtual shadower and the L1 shadower. Before presenting the evaluation results for these two aspects, the following section will show the ASR results for $L1_{S1}$, $L1_{SS}$, and $L2_R$, which are necessary for subsequent discussions.

B. Ablating effects of vocoding module

We first applied ASR to all testing data using a reputable and advanced ASR system[3] [31]. In this study, the ASR results reflect the expressions of the L1 shadower during the shadowing sessions (S1 and SS) and the L2 learner during the recording session (R). Inevitably, the ASR result may include errors, which are quantified using the CER/WER by comparing the results with the reference text used during the recording of $L2_R$. As shown in † of Table I, the differences in CER/WER values between $L1_{S1}$ and $L1_{SS}$ reveal the linguistic variations introduced by shadowing. The $L2_R$ utterances, characterized by a strong accent, consequently display even higher CER/WER compared to the L1 speech.

To eliminate propagation errors potentially caused by VC decoders, we performed a framewise PPG-based VC that only converts the speaker identity based on the implementation in

[3]https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

TABLE I: CER/WER and S1-CER/WER results evaluated on original recordings, PPG-VC, VTN and AAS-VC outputs.

	Model	Training Source	Training Target	Testset	CER↓	WER↓	S1-CER ↓	S1-WER↓
†	-	-	-	$L1_{S1}$	6.83	13.71	-	-
†		-	-	$L1_{SS}$	0.85	3.60	-	-
†		-	-	$L2_R$	10.04	20.91	-	-
$*$	PPG-VC [30]	$L1_{spk}$	$L1_{spk}$	$L1_{S1}$	8.00	15.90	-	-
$*$		$L1_{spk}$	$L1_{spk}$	$L1_{SS}$	1.90	5.70	-	-
$**$		$L2_{spk}$	$L2_{spk}$	$L2_R$	21.67	38.83	-	-
$***$		$L2_{spk}$	$L1_{spk}$	$L2_R$	21.81	39.23	-	-
‡	VTN [17]	$L2_R$	$L1_{S1}$	$L2_R$	22.64	37.29	22.41	36.15
‡		$L2_R$	$L1_{SS}$	$L2_R$	**19.47**	**34.53**	19.28	33.53
‡		$L1_{SS}$	$L1_{S1}$	$L2_R$	32.21	51.24	32.08	50.56
‡	AAS-VC [23]	$L2_R$	$L1_{S1}$	$L2_R$	21.60	38.68	21.53	37.98
‡		$L2_R$	$L1_{SS}$	$L2_R$	**20.44**	**36.74**	19.98	35.97
‡		$L1_{SS}$	$L1_{S1}$	$L2_R$	26.49	46.76	26.49	46.50

TABLE II: Mel Cepstral Distortion (MCD), F0 Root Mean Square Error (F0RMSE), F0 Correlation (F0CORR), and Absolute Duration Difference (DURR) on PPG-VC, VTN and AAS-VC. PPG-VC converts multiple L2 speaker identities into the target L1 speaker to minimize the influence of speaker identity on acoustic features.

Model	Training Source	Training Target	Testset	Reference	MCD↓	F0RMSE↓	F0CORR↑	DURR↓
-	-	-	$L2_R$	$L1_{S1}$	12.84	89.65	0.084	0.337
	-	-	$L1_{SS}$	$L1_{S1}$	6.62	35.24	0.385	0.350
PPG-VC	$L2_R$	$L1_{spk}$	$L2_R$	$L1_{S1}$	8.94	47.27	0.127	0.339
	$L2_R$	$L1_{spk}$	$L2_R$	$L1_{SS}$	8.97	47.03	0.117	0.361
VTN	$L2_R$	$L1_{S1}$	$L2_R$	$L1_{S1}$	7.53	40.35	0.239	0.552
	$L2_R$	$L1_{SS}$	$L2_R$	$L1_{S1}$	7.47	39.76	0.253	0.575
AAS-VC	$L2_R$	$L1_{S1}$	$L2_R$	$L1_{S1}$	7.39	38.76	0.240	0.515
	$L2_R$	$L1_{SS}$	$L2_R$	$L1_{S1}$	7.34	37.78	0.259	0.401

[30] for analysis synthesis (ANA-SYN). The PPG decoder and vocoder described in Section III-B4 were utilized. The CER/WER of the ANA-SYN speech is presented in $*$, $**$ and $***$ of Table I. $*$ and $**$ show the VC results in which the identity of the input speaker was maintained, while $***$ shows the conversion of $L2_{spk}$ to $L1_{spk}$. The minimal variations in CER/WER for $L1_{S1}$ and $L1_{SS}$ suggest that the embedding features do not distort linguistic information. However, a larger degradation in CER/WER for ANA-SYN $L2_R$ is noted in $**$, which is considered to be due to the inevitable noises present in $L2_R$. Note that $L2_R$ recordings were collected separately in various environments using personal devices, making it difficult to ensure standardized acoustic quality. Nevertheless, the very small difference in CER/WER after speaker normalization to $L1_{spk}$, shown in $***$, indicates that these recording disturbances do not significantly affect the decoder's performance.

C. Linguistic similarity of virtual shadower

Since the proposed virtual shadower aims to construct potentially corrupted speech generated by the L1 shadower, the output is different from what the L2 learner actually intends, but expects to be more similar to what the shadower said. While the conventional WER is calculated as follows:

$$\text{WER} = \frac{S + I + D}{N_R}, \tag{1}$$

we introduce S1-WER to evaluate the linguistic similarity between the conversion results and $L1_{S1}$ based on their ASR results.

$$\text{S1-WER} = \frac{S + I + D}{N_{S1}}, \tag{2}$$

where S, I, and D are substitution, insertion, and deletion errors of the converted results, respectively. N_R represents the word count of L2 learners' handcrafted scripts, while N_{S1} is the word count of the recognition hypothesis of $L1_{S1}$ conducted by the mentioned ASR system. Lower S1-WER will represent higher linguistic similarity between the converted speech and the shadowing speech. S1-CER is the character-based error rate, calculated in a similar way to S1-WER.

As shown in ‡ of Table I, compared to $***$, the reduction of WER on $L2_R$-$L1_{S1}$ and $L2_R$-$L1_{SS}$ demonstrates the accent reduction capability of Seq2Seq VC. The best CER/WER are achieved using the $L2_R$-$L1_{SS}$ training pair, as the L1 shadower fluently repeated the content intended by the L2 learner by viewing the learner's script, which is a beneficial condition for VC training. For the $L2_R$-$L1_{S1}$ which is the most straightforward approach to achieve virtual shadowing

by the L1 speaker, the results are reasonable, as the semi-parallel setting brings high barriers for feature mapping from a mispronounced segment to the correct segment. In the $L1_{SS}$-$L1_{S1}$ setting, the relatively high CER/WER indicates that the model struggles to accurately capture the unintelligibility of the L2 speakers. This is likely because $L2_R$'s features are considered out-of-domain.

Regarding S1-CER/WER, although the differences are not significant, all proposed models have lower or equal S1-CER/WER values compared to CER/WER. This indicates that Seq2Seq VC systems have the potential to effectively map semi-parallel linguistic features, such as $L2_R$-$L1_{S1}$ in this study. Particularly, $L1_{SS}$ targeting models outperform all the other training settings, indicating the possibility of building a virtual shadowing system utilizing parallel data only.

D. Acoustic similarity of virtual shadower

Table II illustrates the acoustic similarity between $L1_{S1}$ and the converted speech. Compared to the $L1_{spk}$-normalized $L2_R$ conducted by PPG-VC, the better metrics in MCD and F0 demonstrate the prosody reconstruction ability of the proposed virtual shadower. In particular, AAS-VC, benefiting from its superior alignment ability, shows better prosody similarity, especially in F0 and duration.

Interestingly, although the model was not trained with $L1_{S1}$, the results showed that the output of $L1_{SS}$-targeting model closely resembles the acoustic similarity of the actual $L1_{S1}$. This indicates the potential of using parallel data directly to establish a virtual shadowing system, which supports the conclusion mentioned in Section IV-C.

V. Conclusions and future works

A. Conclusions

In this study, we first proposed the concept of a virtual shadower, which simulates the behavior of an L1 listener who instantly repeats what an L2 speaker utters while listening to it. By reflecting the L1 listener's immediate understanding, the virtual shadower aims to highlight linguistic ambiguities in L2 speech that the speaker may not notice. Experiments conducted on two Seq2Seq VC systems with various training data settings from the $L2_R$, $L1_{S1}$ and $L1_{SS}$ triplet demonstrated the potential of parallel and semi-parallel training from $L2_R$ to $L1_{S1}$/$L1_{SS}$.

B. Future works

1) Generalization ability: Shadowing behavior varies among listeners with different language backgrounds. This study involved only Japanese learners of English and a single rater. Future experiments will involve participants who speak different languages and multiple raters to explore the capability and adaptability of Seq2Seq VC as more flexible and general virtual shadowers.

2) Integrating $L1_{SS}$-$L1_{S1}$ information for virtual shadowing: In this study, we acknowledge and examine the challenge of directly mapping $L2_R$-$L1_{S1}$ as they are both linguistically and acoustically different. However, the promising result with $L2_R$-$L1_{SS}$ encourage us to combine information derived from $L1_{SS}$-$L1_{S1}$. We believe the distance between $L1_{SS}$-$L1_{S1}$ has not been fully utilized yet. For instance, using a loss function derived from PPG-DTW ($L1_{S1}, L1_{SS}$) is considered a more promising and interpretable approach. Additionally, exploring the cascading of $L2_R$-$L1_{SS}$ with $L1_{SS}$-$L1_{S1}$ could yield valuable insights to convert $L2_R$ to $L1_{S1}$.

3) Self-supervised speech representations with prosodic feature considered: Although linguistically-rich embedding like PPG perform well in foreign accent conversion tasks, PPG lacks prosodic features such as intonation and rhythm, which often affect listeners' comprehension. Self-supervised speech representations have shown great potential in reconstructing unintelligible speech, such as electrolaryngeal or dysarthric speech [32]. We will explore the feasibility of using latent features from models like HuBERT [33] and WavLM [34], which are known to include both linguistic and prosodic features.

4) Pedagogical Assessment of virtual shadower: Our virtual shadower offers feedback by pinpointing segments of learners' speech that are unintelligible to listeners. Unlike conventional feedback, which is typically provided as scores or written comments, our feedback is delivered as immediate vocal response. This approach offers more realistic feedback, which we expect will enhance learners' motivation more effectively than traditional methods.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[2] S. Wills, Y. Bai, C. Tejedor-García, C. Cucchiarini, and H. Strik, "Automatic Speech Recognition of Non-Native Child Speech for Language Learning Applications," in *Proc. SLATE*, 2023, 7:1–7:8.

[3] M. P. Y. Chan, J. Choe, A. Li, Y. Chen, X. Gao, and N. Holliday, "Training and typological bias in ASR performance for world Englishes," in *Proc. Interspeech*, 2022, pp. 1273–1277.

[4] A. Neri, C. Cucchiarini, and H. Strik, "Effective feedback on l2 pronunciation in asr-based call," 2001.

[5] Y. Weng, S. S. Miryala, C. Khatri, *et al.*, "Joint contextual modeling for asr correction and language understanding," in *Proc. ICASSP*, 2020, pp. 6349–6353.

[6] R. Ma, M. Qian, P. Manakul, M. Gales, and K. Knill, "Can generative large language models perform asr error correction?" *arXiv preprint arXiv:2307.04172*, 2023.

[7] T. M. Derwing and M. J. Munro, "Accent, intelligibility, and comprehensibility: Evidence from four l1s," *Studies in Second Language Acquisition*, vol. 19, no. 1, pp. 1–16, 1997.

[8] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language learning*, vol. 45, no. 1, pp. 73–97, 1995.

[9] Y. Inoue, S. Kabashima, D. Saito, N. Minematsu, K. Kanamura, and Y. Yamauchi, "A Study of Objective Measurement of Comprehensibility through Native Speakers' Shadowing of Learners' Utterances," in *Proc. Interspeech*, 2018, pp. 1651–1655.

[10] Z. Lin, R. Takashima, D. Saito, N. Minematsu, and N. Nakanishi, "Shadowability annotation with fine granularity on l2 utterances and its improvement with native listeners' script-shadowing.," in *Proc. Interspeech*, 2020, pp. 3865–3869.

[11] C. Zhu, R. Hakoda, D. Saito, N. Minematsu, N. Nakanishi, and T. Nishimura, "Multi-granularity annotation of instantaneous intelligibility of learners' utterances based on shadowing techniques," in *Proc. ASRU*, 2021, pp. 1071–1078.

[12] J. Yue, F. Shiozawa, S. Toyama, *et al.*, "Automatic Scoring of Shadowing Speech Based on DNN Posteriors and Their DTW," in *Proc. Interspeech*, 2017, pp. 1422–1426.

[13] G. Maimon and Y. Adi, "Speaking style conversion in the waveform domain using discrete self-supervised units," in *EMNLP*, H. Bouamor, J. Pino, and K. Bali, Eds., 2023, pp. 8048–8061.

[14] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. ICASSP*, 2021, pp. 920–924.

[15] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM TASLP*, vol. 29, pp. 2367–2381, 2021.

[16] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *Proc. ICASSP*, 2018, pp. 5314–5318.

[17] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining," in *Proc. Interspeech*, 2020, pp. 4676–4680.

[18] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 6000–6010.

[19] W.-C. Huang and T. Toda, "Evaluating Methods for Ground-Truth-Free Foreign Accent Conversion," in *Proc. APSIPA ASC*, 2023.

[20] T. Hayashi, W.-C. Huang, K. Kobayashi, and T. Toda, "Non-autoregressive sequence-to-sequence voice conversion," in *Proc. ICASSP*, 2021, pp. 7068–7072.

[21] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[22] T. Okamoto, T. Toda, and H. Kawai, "E2e-s2s-vc: End-to-end sequence-to-sequence voice conversion," in *Proc. Interspeech*, 2023, pp. 2043–2047.

[23] W.-C. Huang, K. Kobayashi, and T. Toda, "Aas-vc: On the generalization ability of automatic alignment search based non-autoregressive sequence-to-sequence voice conversion," *arXiv preprint arXiv:2309.07598*, 2023.

[24] K. J. Shih, R. Valle, R. Badlani, A. Lancucki, W. Ping, and B. Catanzaro, "Rad-tts: Parallel flow-based tts with robust alignment learning and diverse synthesis," in *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[25] Y. Hamada, "The effectiveness of pre-and post-shadowing in improving listening comprehension skills," *The Language Teacher*, vol. 38, no. 1, pp. 3–10, 2014.

[26] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Pretraining techniques for sequence-to-sequence voice conversion," *IEEE/ACM TASLP*, vol. 29, pp. 745–755, 2021.

[27] S. Liu, Y. Cao, N. Hu, D. Su, and H. Meng, "Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation," in *Proc. ICME*, 2021, pp. 1–6.

[28] W.-C. Huang, S.-W. Yang, T. Hayashi, and T. Toda, "A Comparative Study of Self-Supervised Speech Representation Based Voice Conversion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1308–1318, 2022.

[29] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020, pp. 6199–6203.

[30] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM TASLP*, vol. 29, pp. 1717–1728, 2021.

[31] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[32] L. P. Violeta, D. Ma, W.-C. Huang, and T. Toda, "Pretraining and adaptation techniques for electrolaryngeal speech recognition," *IEEE/ACM TASLP*, vol. 32, pp. 2777–2789, 2024.

[33] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM TASLP*, vol. 29, pp. 3451–3460, 2021.

[34] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

EADSum: Element-Aware Distillation for Enhancing Low-Resource Abstractive Summarization

Jia-Liang Lu, Bi-Cheng Yan, Yi-Cheng Wang, Tien-Hong Lo, Hsin-Wei Wang, Li-Ting Pai, Berlin Chen
Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
E-mail: {jialianglu, bicheng, yichengwang, teinhonglo, hsinweiwang, 61147095s, berlin}@ntnu.edu.tw

***Abstract*—Abstractive summarization aims to generate concise summaries from input documents, with notable advancements driven by the advent of large language models (LLMs). However, the substantial computational demands and the sheer size of LLMs pose significant scalability challenges for the corresponding deployment in real-world applications. Existing methods typically adopt large amounts of training data to train smaller, more compact models to achieve performance on par with LLMs, either through fine-tuning with human-annotated labels or distilling rationales from LLM-generated labels. As an appealing alternative for low-resource abstractive text summarization, we propose EADSum (Element-Aware Distillation for Summarization), a novel training framework which aims to generate fine-grained summaries correlated to the human writing mindset while alleviating the heavy requirement of supervised training data. The proposed EADSum approach first guides LLMs to generate element-aware rationales from the input document, drawing attention to crucial elements such as entities, dates, events, and the results of event. These generated rationales then serve as additional supervision for the subsequent training of compact models within a multi-task learning framework. A series of experiments conducted on the CNN/DailyMail benchmark dataset demonstrate the feasibility and effectiveness of our approach.**[1]

I. Introduction

Text summarization is a crucial task in natural language processing which aims to automatically condense lengthy input documents into shorter, coherent, and informative summaries [1][2]. Research endeavors in text summarization generally fall into two modeling paradigms: extractive and abstractive summarization. The former extracts the most representative sentences from the input documents and forms a summary by aggregating these sentences. The latter, in contrast, concentrates more on generating concise, fluent summaries that capture the core ideas of the input documents. Among these paradigms, abstractive summarization has garnered significant attention due to its ability to produce high-quality summaries that are both verbally innovative and capable of integrating external knowledge [3][4].

Early studies on abstractive summarization focused on both sequence-to-sequence neural modeling [5][6][7] and transfer learning techniques coupled with pre-trained large language models [8][9][10]. These approaches typically train or fine-tune pre-trained language models with large amounts of training data to achieve superior performance in terms of lexical overlap (i.e., ROUGE evaluations [11]) compared to golden summaries [12][13][14]. In recent years, the advent of Large Language Models (LLMs), such as GPT-3 [15] and its variants [16][17], has fueled the advancements in language understanding and generation tasks, leading to remarkable breakthroughs and sparking further research in abstractive text summarization. LLMs demonstrated reasoning abilities through chain-of-thought prompting [18]. This approach effectively captures both the topic structures and the core ideas from source texts, suggesting LLMs' promising potential for generating text summaries. As a follow-up study, for example, Wang et al. leveraged chain-of-thought techniques to guide LLMs in focusing on the news elements of input documents. These derived element-aware rationales then act as additional information to prompt the LLMs to generate summaries in a step-by-step manner [18].

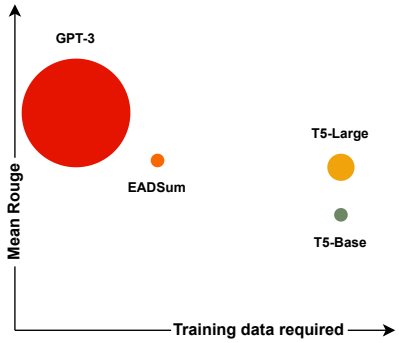


Fig. 1. While GPT-3 offers strong zero-shot performance, it is challenging to implement in practice due to its size and computational requirements. The size of the circles represents the model parameters. Conversely, traditional methods of training smaller task-specific models require large amounts of training data. We propose EADSum, a new paradigm that extracts element-aware rationales from LLMs as additional knowledge to train compact models. This approach reduces both the deployed model size and the amount of data required for training.

Despite the impressive performance of abstractive text summarization achieved by LLMs, deploying these models in real-world applications is confronted with at least two intrinsic challenges. On one hand, the substantial computational demands and large model size of LLMs greatly limit their utility in resource-constrained environments. On the other hand, training a smaller, compact model to achieve performance comparable to LLMs is a data-hungry process that

[1]Source code is available at: https://github.com/VitasLu/EADSum

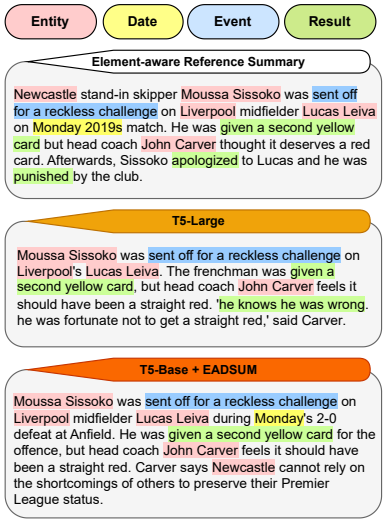


Fig. 2. Comparison of summarization outputs from different models and the Element-aware Reference Summary.

heavily relies on enormous amounts of training data, either by fine-tuning the model with human-labeled data or distilling knowledge from a given LLM. In light of this, we propose EADSum (Element-Aware Distillation for Summarization), a novel training framework for low-resource abstractive text summarization managed to generate high-quality summaries closely aligned with human assessment criteria, such as coherence, consistency, and relevance [19]. Specifically, our method first employs chain-of-thought reasoning [20][21] to guide LLMs in extracting element-aware rationales from the input document, including detailed information of entities, dates, events, and the results of event, as shown in Fig. 2. These extracted summary rationales then act as additional supervision to train compact models through a multi-task training setup with both summary and rationale generation tasks. A series of experiments conducted on the CNN/DailyMail benchmark dataset demonstrate the feasibility and effectiveness of our approach. Experimental results show that our method surpasses the performance of fine-tuning pre-trained language models and is competitive with the zero-shot summarization capabilities of GPT-3 on element-aware reference summary. It is worth noting that the proposed method provides a practical solution tailered for deploying advanced summarization techniques in resource-constrained environments by distilling the capabilities of LLMs into a compact model.

II. METHODOLOGY

We propose a novel method called EADSum (Element-Aware Distillation Summarization) that leverages the capability of Large Language Models (LLMs) to identify the core elements of a document, enabling the training of compact models in a data-efficient manner. Our comprehensive framework, as illustrated in Fig. 3, comprises two main steps designed for low-resource scenarios: (A) rationales extraction, and (B)

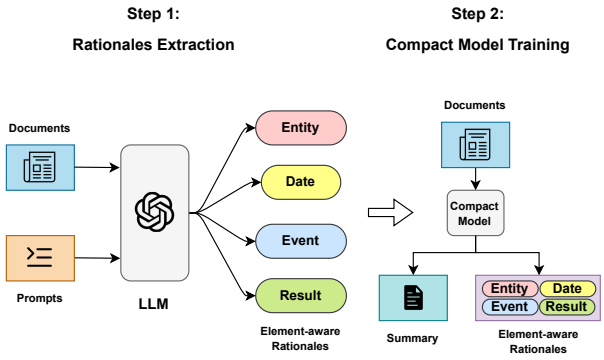


Fig. 3. A conceptual demonstration of our two-step framework, EADSum, which endows compact models with LLM's text summarization capabilities.

compact model training. In the first step, we extract element-aware rationales from LLMs. In the second step, we train a compact model through a multi-task training setup.

A. Rationales Extraction

The extraction of element-aware rationales from Large Language Models (LLMs) draws inspiration from Chain-of-Thought prompting research [20]. This line of inquiry has demonstrated that LLMs are capable of providing step-by-step explanations, or rationales, for their outputs. This capability has been successfully applied to various tasks, including summarization [18]. Our Element-Aware Distillation Summarization (EADSum) method further leverages this approach, guiding LLMs to identify and articulate key elements such as entities, dates, events, and the results of event.

To extract element-aware rationales from Large Language Models (LLMs), we have designed a series of guiding questions. These questions prompt the LLM to identify and extract four core elements—entities, dates, events, and the results of event—from the input document. For each element i, we formulate a guiding question q_i. For instance, the prompt for the entities element $q_{entities}$ can be "What are the important entities in the following article?", and so on.

We then combine these guiding questions for the four key elements into a single prompt set $Q = [q_{entities}, q_{dates}, q_{events}, q_{results}]$. Subsequently, we input both the source document S and the question prompt set Q to the LLM, represented as $[S; Q]$. This process generates a rationale R, which serves as crucial role for the next stages of our method.

B. Compact Model Training

We begin by outlining the existing framework for training task-specific models. Building upon this foundation, we then expand this framework to incorporate extracted rationales from step 1 into the training process. Our dataset is formally represented as $\mathcal{D} = (x_i, y_i)_{i=1}^{n}$, where n denotes the size of the dataset, x_i represents the source document, and y_i represents

its corresponding reference summary. In our experiments, we constrain both x_i and y_i to be natural language.

Standard fine-tuning and task-specific knowledge distillation. The predominant approach for developing task-specific models involves fine-tuning a pre-trained model using supervised data [22]. When human-annotated labels are unavailable, researchers often employ task-specific distillation techniques [23]. In this method, Large Language Models (LLMs) serve as teachers, generating pseudo-noisy training labels [24][25].

In the context of text summarization, we train the compact model f to minimize the summary prediction loss in both fine-tuning and distillation scenarios, defined as:

$$\mathcal{L}_{summary} = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i), \tag{1}$$

where ℓ represents the cross-entropy loss between the generated and reference summaries.

Multi-task learning with element-aware rationales. To create a more explicit connection between the source document x_i and the reference summary y_i, we use rationales r_i extracted from step 1 as additional supervision.

A key innovation in our approach is how we utilize these rationales. Instead of using them as additional input to the model, we frame the learning process as a multi-task problem. This decision is driven by two important factors. First, our compact model has fewer parameters compared to large language models (LLMs), making it challenging to effectively generate rationales using chain-of-thought reasoning [20]. Second, considering deployment challenges, we reduce the need for LLM involvement during inference, making our method more efficient and scalable.

Specifically, we train the model $f(x_i) \rightarrow (\hat{y_i}, \hat{r_i})$ to generate not only the text summaries $\hat{y_i}$ but also the corresponding element-aware rationales $\hat{r_i}$ given the text input:

$$\mathcal{L} = \alpha\mathcal{L}_{summary} + (1-\alpha)\mathcal{L}_{rationale}, \tag{2}$$

where α is a balancing parameter, $\mathcal{L}_{summary}$ is the summary prediction loss in Eq. 1, and $\mathcal{L}_{rationale}$ is the rationale generation loss:

$$\mathcal{L}_{rationale} = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), r_i). \tag{3}$$

This rationale generation loss enables the model to learn to generate the intermediate reasoning steps for the summary, potentially guiding the model to better generate the resultant summary and this is our proposed EADSum distillation framework.

III. EXPERIMENTS

A. Experimental Settings

Evaluation Dataset. Our experiments are conducted on the CNN/DailyMail dataset, a widely-used benchmark in news summarization tasks. This large-scale corpus comprises 287,113 training, 13,368 validation, and 11,490 test examples, each consisting of a news article paired with a multi-sentence summary. The articles are sourced from CNN and Daily Mail websites, while the summaries are created by concatenating the bullet points accompanying each article.

To investigate the efficacy of our approach in low-resource scenarios, we deliberately constrain our experimental setup. Instead of utilizing the full dataset, we employ only 70,835 examples (approximately 24.7% of the full training set) for model training, and 7,871 examples for validation. Furthermore, we evaluate our model on a carefully selected subset of 200 examples from the test set. This resource-constrained configuration allows us to simulate real-world situations where annotated data may be limited, assess our model's performance under data scarcity conditions, and explore the potential of our approach for domains or languages with restricted resources.

Compared Baselines. To evaluate the effectiveness of our proposed approach, we compare it against several strong baseline models:

- PEGASUS-Large[2]: A state-of-the-art abstractive summarization model pre-trained on massive text corpora using gap-sentence generation objectives.
- T5-base[3] and T5-Large[4]: Two variants of the Text-to-Text Transfer Transformer (T5) model, which have demonstrated impressive performance across various natural language processing tasks, including summarization.
- BART-base[5] and BART-Large[6]: Two versions of the BART (Bidirectional and Auto-Regressive Transformers) model, specifically designed for sequence-to-sequence tasks like text summarization.
- GPT-3[7]: The largest language model in our baseline set, known for its remarkable zero-shot learning capabilities across a wide range of language tasks.

For the pre-trained language models (PLMs), we utilize the officially models fine-tuned on CNN/DailyMail dataset (BART-Base[8], BART-Large[9], PEGASUS-Large[10], T5-Base[11], and T5-Large[12]) released on the Huggingface platform for generation tasks. In the case of zero-shot prompting for GPT-3, we use the standard prompt p = "Summarize the above article:" for the CNN/DailyMail dataset by approach of [26].

Implement Detail. We employ GPT-3.5 as our Large Language Model (LLM) to extract element-aware rationales from the input articles. To elicit these rationales, we used the following series of prompts: 'What are the important entities in the following article?', 'What are the important dates?', 'What events are happening?', and 'What is the result of

[2]https://huggingface.co/google/pegasus-large
[3]https://huggingface.co/google-t5/t5-base
[4]https://huggingface.co/google-t5/t5-large
[5]https://huggingface.co/facebook/bart-base
[6]https://huggingface.co/facebook/bart-large
[7]https://openai.com/api/
[8]https://huggingface.co/ainize/bart-base-cnn
[9]https://huggingface.co/facebook/bart-large-cnn
[10]https://huggingface.co/google/pegasus-cnn_dailymail
[11]https://huggingface.co/flax-community/t5-base-cnn-dm
[12]https://huggingface.co/ksstevenT5-large-cnndm

TABLE I
PERFORMANCE COMPARISON OF LARGER MODELS (BART-LARGE, PEGASUS-LARGE, T5-LARGE, GPT-3) AND SMALLER MODELS (BART-BASE, T5-BASE) IN TERMS OF ROUGE-1, ROUGE-2, ROUGE-L, MEAN-ROUGE, AND BERTSCORE ON ORIGINAL REFERENCE SUMMARIES

Model	Training Size	# Params	Original Reference Summary				
			Rouge-1	Rouge-2	Rouge-L	Mean-Rouge	BERTScore
BART-Large	-	406M	39.11	16.56	36.43	30.70	0.8708
BART-Large	100%	406M	43.44	20.51	40.52	34.82	0.8828
PEGASUS-Large	-	568M	36.96	15.10	33.28	28.45	0.8637
PEGASUS-Large	100%	568M	40.34	18.65	38.26	32.42	0.8793
T5-Large	-	770M	41.24	18.66	38.57	32.82	0.8795
T5-Large	100%	770M	41.84	19.28	39.34	33.49	0.8829
GPT-3	Zero-Shot	175B	31.44	10.86	29.25	23.85	0.8655
BART-Base	-	139M	40.80	17.82	38.14	32.25	0.8728
BART-Base	100%	139M	41.86	19.75	39.00	33.53	0.8801
T5-Base	100%	220M	41.06	19.08	38.44	32.86	0.8805
T5-Base + EADSum	24.7%	220M	41.29	18.63	38.39	32.77	0.8780

TABLE II
PERFORMANCE COMPARISON OF LARGER MODELS (BART-LARGE, PEGASUS-LARGE, T5-LARGE, GPT-3) AND SMALLER MODELS (BART-BASE, T5-BASE) IN TERMS OF ROUGE-1, ROUGE-2, ROUGE-L, MEAN-ROUGE, AND BERTSCORE ON ELEMENT-AWARE REFERENCE SUMMARIES

Model	Training Size	# Params	Element-aware Reference Summary				
			Rouge-1	Rouge-2	Rouge-L	Mean-Rouge	BERTScore
BART-Large	-	406M	31.78	12.81	28.47	24.35	0.8737
BART-Large	100%	406M	32.57	12.41	29.21	24.73	0.8804
PEGASUS-Large	-	568M	28.87	9.16	24.34	20.79	0.8645
PEGASUS-Large	100%	568M	31.53	11.82	28.97	24.11	0.8786
T5-Large	-	770M	30.68	10.54	27.47	22.90	0.8762
T5-Large	100%	770M	32.55	12.22	29.21	24.66	0.8813
GPT-3	Zero-Shot	175B	37.52	14.59	33.71	28.61	0.8922
BART-Base	-	139M	32.97	13.52	29.69	25.39	0.8744
BART-Base	100%	139M	33.18	12.44	29.74	25.12	0.8809
T5-Base	100%	220M	31.41	11.84	27.99	23.75	0.8784
T5-Base + EADSum	24.7%	220M	33.73	13.57	30.19	25.83	0.8805

these events?'. This approach helps us obtain comprehensive information about the article's key elements.

For the training stage, we utilize T5-Base, a model with 220 million parameters, as our compact model. We implement a multi-task setup by prepending task prefixes to the input examples. Specifically, we use [summary] to indicate summarization tasks and [rationale] for rationale generation tasks. This approach, inspired by [9], allows us to train the compact model to produce summaries when the [summary] prefix is provided and to generate rationales when the [rationale] prefix is used. We set the balancing parameter α to 0.5 for this multi-task training.

Evaluation Metrics. We employ automatic metrics to evaluate the quality of the generated summaries. Specifically, we utilize the following evaluation methods:

- ROUGE Score [11]: We report the F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L. These metrics measure the overlap between the generated summaries and the reference summaries in terms of unigrams, bigrams, and the longest common subsequence, respectively.
- Mean-ROUGE [11]: This metric is calculated by taking the average of ROUGE-1, ROUGE-2, and ROUGE-L scores. It provides a single, comprehensive measure of the overall quality of the generated summaries, balancing the evaluation across different levels of text granularity.
- BERTScore [27]: This metric evaluates the semantic similarity between the generated summaries and the reference summaries. It operates by computing the cosine similarity between the contextualized embeddings of tokens in the summaries, providing a more nuanced assessment of semantic equivalence.

B. Main Results

In our study, we evaluate the effectiveness of our proposed EADSum training method using two distinct test sets, each designed to assess different aspects of summarization quality. The first test set utilizes the original reference summaries from the CNN/DailyMail dataset, a widely recognized benchmark in text summarization. These summaries provide a general assessment of the models' ability to capture the main points of news articles. The second test set employs element-aware reference summaries [18], meticulously crafted by three news experts following a comprehensive writing protocol. This protocol encompasses micro demands, which focus on incorporating core news elements such as entities, dates, events, and the results of event, and macro demands, which ensure professional quality in terms of fluency, coherence, consistency, and relevance. By utilizing these complementary test sets, we aim to provide a more comprehensive evaluation of our EADSum method, assessing its performance not only in general summarization tasks but also in capturing nuanced, element-specific information that aligns with professional news writing standards.

Based on the evaluation using original reference summaries, our proposed EADSum training method demonstrates perfor-

TABLE III
THE ABLATION STUDY COMPARES THREE SCENARIOS: T5-BASE WITHOUT TRAINING, T5-BASE TRAINED ON 24.7% OF THE DATASET, AND T5-BASE WITH EADSUM TRAINED ON 24.7% OF THE DATASET.

Model	Training Size	Original Reference Summary				Element-aware Reference Summary			
		Rouge-1	Rouge-2	Rouge-L	BERTScore	Rouge-1	Rouge-2	Rouge-L	BERTScore
T5-Base	-	41.55	18.75	38.88	0.8783	31.10	11.00	27.56	0.8760
T5-Base	24.7%	41.29	18.87	38.57	0.8784	33.54	12.78	29.87	0.8798
T5-Base + EADSum	24.7%	41.29	18.63	38.39	0.8780	33.73	13.57	30.19	0.8805

mance comparable to, or even surpassing, some larger models presented in TABLE I. Notably, the T5-Base model with our EADSum method on just 24.7% of the data outperforms the non-finetuned T5-Large model in Mean-ROUGE scores. It also surpasses the PEGASUS-Large model trained on 100% of the data in ROUGE-1 and ROUGE-L scores. Furthermore, the EADSum method exhibits competitive performance when compared to BART-Base and T5-Base models trained on the full dataset.

In the second evaluation using element-aware reference summaries, as presented in TABLE II, our EADSum method demonstrates remarkable performance in Mean-ROUGE scores, ranking second only to GPT-3. While GPT-3, with its 175 billion parameters, can produce more informative summaries, it also demands substantial computational resources for deployment. Notably, our EADSum method achieves a Mean-ROUGE score of 25.83, surpassing most larger models and smaller models trained on the full dataset. This result signifies that our method generates summaries with greater overlap with element-aware reference summaries, showcasing superior performance in identifying and summarizing critical information elements. Moreover, the BERTScore of our EADSum method is competitive with those of BART-Base and T5-Large trained on the full dataset, as well as GPT-3 in a zero-shot setting. This performance is particularly impressive considering that our model uses only a fraction of the training data yet can generate summaries with high semantic similarity to expert-crafted, element-aware reference summaries.

These findings highlight the effectiveness of the EADSum approach in producing high-quality summaries that capture essential information elements while utilizing minimal training data.

C. Ablation Studies

To evaluate the effectiveness of our proposed EADSum method, we conduct ablation studies using the T5-Base model. We compared three configurations: T5-Base without training, T5-Base trained on 24.7% of the dataset, and T5-Base with EADSum trained on 24.7% of the dataset. This setup allows us to highlight the impact of EADSum and assess its performance in low-resource scenarios.

TABLE III presents an intriguing result: our EADSum method performs slightly lower than T5-Base (both without training and when trained on 24.7% of the dataset) in terms of ROUGE scores and BERTScores when evaluated against the original reference summaries. However, EADSum outperforms T5-Base in both scenarios when evaluated against element-aware reference summaries using the same metrics. We attribute this phenomenon to the structural differences between the original and element-aware reference summaries. Specifically, when a generated summary achieves a higher score compared to one type of reference summary, it indicates greater similarity in terms of word occurrence and semantic meaning to that particular reference summary style. Consequently, this may result in less similarity to the other type of reference summary.

These results demonstrate that EADSum not only compensates for the reduced training data but also enhances the model's ability to generate summaries that align more closely with expert-written, element-focused references.

IV. CONCLUSIONS

In this paper, we propose EADSum, a novel distillation framework comprising two steps: extracting element-aware rationales from Large Language Models (LLMs) and training compact models with the extracted rationales. Our method successfully distills the capabilities of LLMs into a compact model. Experimental results on the CNN/DailyMail dataset demonstrate that EADSum, when applied to the T5-Base model, outperforms some fully fine-tuned models on original reference summaries. Moreover, it surpasses most larger and smaller models trained on the full dataset when evaluated on element-aware reference summaries, while requiring significantly fewer computational resources and less training data. This approach not only advances the field of abstractive summarization but also provides a practical solution for real-world applications where computational resources are limited. Future work could explore the applicability of EADSum to other domains and languages, as well as investigate ways to further enhance the element-aware rationale training process.

REFERENCES

[1] M. F. Salchner and A. Jatowt, "A survey of automatic text summarization using graph neural networks," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, *et al.*, Eds., Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6139–6150. [Online]. Available: https://aclanthology.org/2022.coling-1.536.

[2] A. P. Wibawa, F. Kurniawan, *et al.*, "A survey of text summarization: Techniques, evaluation and challenges," *Natural Language Processing Journal*, vol. 7, p. 100 070, 2024.

[3] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural abstractive text summarization with sequence-to-sequence models," *ACM Transactions on Data Science*, vol. 2, no. 1, pp. 1–37, 2021.

[4] Y. Zhang, Y. Liu, Z. Yang, *et al.*, "Macsum: Controllable summarization with mixed attributes," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 787–803, 2023.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[6] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.

[7] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.

[8] M. Lewis, Y. Liu, N. Goyal, *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[9] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[10] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[11] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[12] S. Narayan, Y. Zhao, J. Maynez, G. Simões, V. Nikolaev, and R. McDonald, "Planning with learned entity prompts for abstractive summarization," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1475–1492, 2021.

[13] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[14] M. Zaheer, G. Guruganesh, K. A. Dubey, *et al.*, "Big bird: Transformers for longer sequences," *Advances in neural information processing systems*, vol. 33, pp. 17 283–17 297, 2020.

[15] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[16] P. Kavehzadeh, M. Valipour, M. Tahaei, A. Ghodsi, B. Chen, and M. Rezagholizadeh, "Sorted LLaMA: Unlocking the potential of intermediate layers of large language models for dynamic inference," in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds., St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2129–2145. [Online]. Available: https://aclanthology.org/2024.findings-eacl.141.

[17] J. Achiam, S. Adler, S. Agarwal, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[18] Y. Wang, Z. Zhang, and R. Wang, "Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 8640–8665. DOI: 10.18653/v1/2023.acl-long.482. [Online]. Available: https://aclanthology.org/2023.acl-long.482.

[19] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021.

[20] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[21] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[22] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[23] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[24] F. Iliopoulos, V. Kontonis, C. Baykal, G. Menghani, K. Trinh, and E. Vee, "Weighted distillation with unlabeled examples," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7024–7037, 2022.

[25] S. Arora, A. Narayan, M. F. Chen, *et al.*, "Ask me anything: A simple strategy for prompting language models," in *The Eleventh International Conference on Learning Representations*, 2022.

[26] V. Sanh, A. Webson, C. Raffel, *et al.*, "Multitask prompted training enables zero-shot task generalization," *arXiv preprint arXiv:2110.08207*, 2021.

[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

A Tiny Whisper-SER: Unifying Automatic Speech Recognition and Multi-label Speech Emotion Recognition Tasks

Huang-Cheng Chou
National Tsing Hua University, Taiwan
E-mail: huangchengchou@gmail.com

***Abstract*—Speech Emotion Recognition (SER) is a crucial component in human-computer interaction (HCI). Previous SER research often utilizes Automatic Speech Recognition (ASR) systems to improve SER performance. Some studies have investigated the possibility of modeling multi-task learning for ASR and SER. However, prior studies and our preliminary experiments show that conflicts arise between ASR and SER when standard multi-task learning is used. We adopt a two-stage training strategy with weighted losses to overcome the conflicts between SER and ASR tasks. We utilize the public Whisper model, which supports ASR tasks and other functionalities, and add a small, lightweight adapter on Whisper for the SER task, referred to as Whisper-SER. We first fine-tune the model on the ASR task and then fine-tune it using a weighted loss strategy, balancing the losses between the ASR and multi-label SER tasks during training. The proposed method allows Whisper-SER to recognize emotions and transcribe speech without degrading ASR and SER performance within the same encoder and decoder.**

I. Introduction

Speech is the most common modality for daily human interaction, and emotional signals in speech can improve communication. More and more devices support speech-based requests for users to ask machines to perform tasks. For instance, people can use Amazon Alexa to make restaurant reservations and leave messages for friends. Detecting emotions from human speech can improve HCI and user experiences. For example, the machine can add suitable "emojis" by recognizing emotions from user requests when sending messages on behalf of users. Therefore, machines need the ability to understand both the content and emotions of speech. However, ASR and SER tasks are not easy to integrate into one system, as the two tasks are in conflict, which degrades the performance of both tasks [1].

Additionally, humans use semantics and acoustics to express their emotions and intentions in speech. Therefore, many prior studies [2]–[4] improve the performance of SER systems by integrating linguistic and acoustic information, leading to better recognition of emotions than considering acoustic information alone. A common practice is to deploy a pre-trained ASR system to generate transcripts and use a pre-trained language model, such as BERT, to extract linguistic embeddings as inputs for further improved SER performance [5]–[7]. However, prior studies only focus on SER tasks, and very few works jointly train a *"shared-parameter"* model for SER and ASR tasks due to the well-known catastrophic forgetting issue, which occurs when deep neural networks forget the original tasks when fine-tuning models for new tasks. Therefore, we aim to avoid the catastrophic forgetting issue with the proposed method, allowing the pre-trained ASR model to recognize emotions without ASR performance degradation.

The recent rapid development of large foundation speech processing systems presents an excellent opportunity to unify SER and ASR tasks into one shared-layer model. One of the state-of-the-art (SOTA) ASR systems, Whisper [8], trained with 680,000 hours of weakly-supervised audio data, can handle four tasks: speech recognition, speech translation, voice activity detection, and language identification. However, this model cannot perform SER tasks. While some previous studies have proposed models for multi-task SER and ASR [6], [9], [10], they define the SER task as a single-label task, allowing each utterance to have only one emotion. Nevertheless, emotion perception can involve the co-occurrence of multiple emotions [11]–[13]. Therefore, we follow the emerging perspective of defining SER as a multi-label classification task to reflect the real world. We aim to jointly adopt the Whisper model to validate the proposed method for SER and ASR tasks in one unified architecture. To the best of our knowledge, we are the first to integrate multi-label SER and ASR in a shared-layer model.

In conclusion, we developed a model based on the Whisper model and added a lightweight component for multi-label SER tasks. Unlike recent SER studies that utilize speech Self-Supervised Learning (SSL) models as encoders [1], [14]–[16] and fine-tune them for SER tasks, we investigate whether extracting embeddings from both the encoder and decoder of Whisper can benefit the performance of the SER task. The two main contributions of this work are as follows:

- We comprehensively study the SOTA ASR model's word error rate (WER) on emotion-rich corpora and investigate the relationship between ASR and SER performance, considering inter-rater agreement, speaking rate, duration, and emotion category.
- We are the first to model ASR and multi-label SER tasks with one unified weakly supervised system that enables parameter sharing between these two tasks for both encoders and decoders with the adapted two-stage training strategy, mitigating catastrophic forgetting issues.

II. Background and Related Work

A. SER Systems Using Pre-trained ASR Models

Linguistic information plays an essential role in understanding emotions in speech. Many prior studies have integrated semantics from human-annotated transcripts or predicted ones generated by ASR systems. For instance, Yoon et al. [17] applied a pre-trained ASR system to extract linguistic information, achieving better performance of SER systems compared to using only acoustic information for training SER systems. Additionally, recent papers utilize hidden states from the audio encoder of an ASR system [8] to strengthen the performance of SER systems [7], [18], [19]. However, the aforementioned studies only optimized for a single SER task and did not jointly optimize ASR and SER tasks. Moreover, our preliminary study revealed that

the Whisper Large V3 [1] has a WER of 12.7% on version 1.11 of the MSP-PODCAST, which is higher than most ASR benchmark databases, with a WER of 7.7% on the "Open ASR Leaderboard" [2] [20]. Furthermore, Li et al. [9] reported a similar finding on the well-known emotion corpus, IEMOCAP [21]. Therefore, this work aims to deeply analyze the potential reasons why one of the SOTA ASR systems performs worse on emotional speech than on more neutral utterances by considering the inter-rater agreement, speaking rate, categorical emotions, and duration lengths.

B. Multi-task SER and ASR Systems

For most practical applications, speech-based devices must automatically transcribe speech requests to understand users' intentions, actions, or item information. As a result, many previous studies have combined SER and ASR tasks in their systems. Studies such as [22], [23] apply supervised fine-tuning on pre-trained SSL systems for ASR and SER tasks. However, some research [10], [24] consistently reveals that emotional speech is challenging for ASR systems, leading to higher error rates. Additionally, Gao et al. [10] showed that ASR and SER tasks often conflict with each other when using common multi-task learning. To address this issue, Feng et al. [24] proposed an SER model integrated with an acoustic-to-word ASR model. Gao et al. [10] utilized a two-stage fine-tuning method to separately fine-tune two components of the model for SER and ASR tasks, respectively. Nevertheless, the study [10] still split the SER and ASR tasks using two individual layers, which means the encoder part is not shared. In contrast, our work uses shared encoder and decoder components.

To make the model more efficient, we aim to create a shared-layer model for ASR and SER tasks by adopting a two-stage training approach [10] with weighted loss to jointly optimize ASR and SER tasks without compromising either task. Furthermore, all the aforementioned studies regard the SER task as a single-label task, assigning only one emotion to each utterance. They used majority or plurality rules to obtain a consensus label, but they directly excluded data samples without a consensus label, ignoring minority emotional ratings and making the test easier to predict. This work defines the SER task as a multi-label classification task (section II-C) to reflect a realistic scenario by using the complete test set without discarding ambiguous data samples, as suggested by [13].

C. Disagreement Among Raters and Multi-label SER

Disagreement among raters is common in public emotion databases [25]. However, most previous studies regard disagreement as noise and define the SER task as a single-label task, using majority or plurality aggregation rules [3], [26] to obtain a single consensus label. However, the SER task differs from other speech tasks, like ASR or part-of-speech tagging, as it is genuinely and naturally subjective, making majority/plurality aggregation rules unsuitable for SER tasks [13]. Disagreement does not necessarily need to be removed, and non-consensus samples represent real-world emotion perception. Additionally, conveyed and perceived emotions are not always discrete, as found by psychological studies [11], [12], [27]. In other words, co-occurrence of emotions can occur in a single speech. Therefore, to make SER systems recognize co-occurring emotions, we use the all-inclusive label aggregation rule proposed by [13] to calculate distributional labels as ground truth, and we define SER tasks as multi-label emotion recognition.

III. Methodology

Prior studies have observed conflicts between ASR and SER tasks [10], [24], [28], and we have also found that direct joint training for ASR and SER can degrade the performance of both tasks. Therefore, we are investigating the feasibility of a shared multi-task system for ASR and SER based on the Whisper architecture [8]. We introduce an SER and ASR multitask framework, named Whisper-SER, as shown in Fig. 1, by supervised fine-tuning the SOTA ASR models with a two-stage weighting method to minimize ASR and SER performance degradation. To the best of our knowledge, we are the first to evaluate ASR performance on an emotional dataset with the complete test set, allowing samples to have co-occurring emotions.

A. Distributional-Label Processing for SER

We follow the studies [13] to calculate the distributional labels for training SER systems. For instance, given an emotional audio clip for four-class emotion recognition (neutral (N), angry (A), sad (S), and happy (H)), the votes are N, N, A, S, S. The distributional label can be calculated by dividing the number of votes for each emotion by the total number of votes. In this example, we get the distributional label N, A, S, H= $\{2/5,\ 1/5,\ 2/5,\ 0/5\} = \{0.4,\ 0.2,\ 0.4,\ 0.0\}$. Fig. 2 shows the detailed distribution across different datasets. In conventional SER studies, datasets with mixed emotions are removed for simplicity; however, we kept all the data to reveal the model's actual SER performance, as suggested by [13], [29]. During system evaluation, the threshold $1/C$ is used to convert the distributional predictions into hard decisions, where C represents the number of emotion classes. We use an example in Table I to clearly describe the ad-hoc process. Therefore, in this example, we assign the data sample co-occurring emotions: neutral and sadness.

B. Objective Functions

We employ cross-entropy loss as the objective function for both ASR and SER tasks.

1) Cross-entropy Loss for ASR: In this work, we choose the Whisper model [8] as the backbone, an encoder-decoder-based ASR model using sequence-to-sequence learning. The decoder of Whisper generates the IDs of each token in the input utterances, and Whisper tokens can transform the ground-truth transcripts into IDs. Using the ID vectors of the ground-truth and predicted tokens, we calculate the cross-entropy loss, denoted by $\mathcal{L}_{ASR}$, to update the entire model.

2) Cross-entropy Loss for SER Task: Chou et al. [13] reported that soft-label learning with cross-entropy loss is the best approach for achieving higher recognition in SER systems compared to hard-label learning with cross-entropy loss, soft-label learning with cross-entropy loss, and distributional-label learning with Kullback–Leibler Divergence (KLD) loss. Therefore, we follow this study and use soft-label learning with cross-entropy loss, denoted by $\mathcal{L}_{SER}$, to train the model for the SER task.

3) Standard Multitask Learning for ASR and SER Tasks: The objective function in the standard multitask learning (MTL) becomes

$$\mathcal{L}_{MTL} = \mathcal{L}_{SER} + \mathcal{L}_{ASR}. \tag{1}$$

C. Two Stage Training Method with the Weighting Loss

We aim to enable Whisper to recognize emotions from speech while maintaining the same ASR performance. We use a two-stage training approach to achieve this goal. In the first stage, we fine-tune the Whisper model on the emotional dataset using the training

TABLE I
OVERVIEW OF ONE SAMPLE FOR 4-CLASS EMOTION RECOGNITION. THE **A**, **S**, AND **N**, ARE ANGER, SADNESS, AND NEUTRAL EMOTIONS, RESPECTIVELY. THE NUMBER MEANS THE COUNT OF EMOTIONS. FOR INSTANCE, **A*1, S*2, AND N*2** MEANS A, S, S, N, N

Raw Annotation	A*1, S*2, N*2
Label for **Training** Stage {N, A, S, H}	{0.4, 0.2, 0.4, 0.0}
Label for **Testing** Stage {N, A, S, H}	{1, 0, 1, 0}

[1] https://huggingface.co/openai/whisper-large-v3

[2] https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

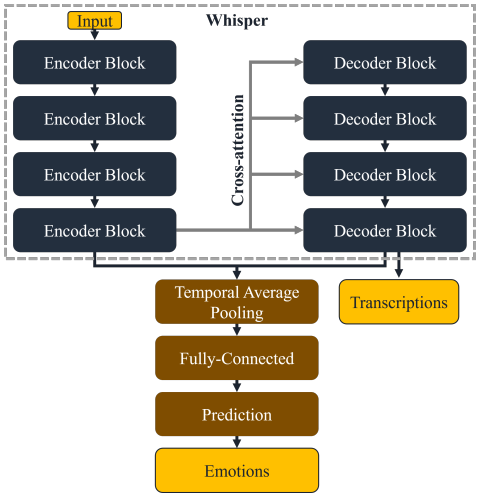


Fig. 1. The figure illustrates the overall structure of Whisper-SER. We added the light models to the Whisper model (brown color) for the SER task, enabling Whisper to do ASR and SER tasks using the proposed training method with weighting loss.

and development sets, focusing only on the ASR task. Then, in the second stage, we apply a weighted loss to fine-tune the model for both SER and ASR tasks. In the weighted loss, we introduce an **alpha** parameter to control the weights between L_{SER} and L_{ASR}, where the alpha value ranges from 0.1 to 0.9, with a step size of 0.1. We also examine the relationship between alpha and ASR/SER performance by introducing the weighting factor. It is important to note that the order of the training tasks matters. In the first stage, we fine-tune the model for the ASR task, and in the second stage, we use the proposed weighted loss for multitasking ASR and SER. Our experiments indicate that if we reverse the order of tasks (i.e., train for the SER task first and then the ASR task), the approach does not work.

$$\mathcal{L}_{MTL}^{\mathcal{W}} = \alpha * \mathcal{L}_{ASR} + (1 - \alpha) * \mathcal{L}_{SER} \tag{2}$$

IV. Experimental Settings

A. The MSP-PODCAST Corpus and Preprocessing

We validate the proposed method on version 1.11 of the MSP-PODCAST emotion dataset [30]. This version has about 237 hours of segmented audio recordings in English, containing 84,030 sentences (134.34 hours) in the train set, 19,815 utterances (31.72 hours) in the development set, 30,647 utterances in the test1 set, and 14,815 utterances in the test2 partition. Test2 set is overall more challenging

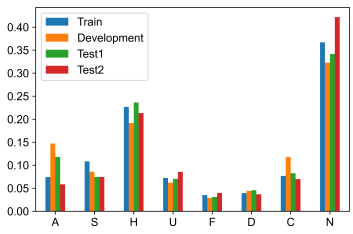


Fig. 2. Distributional-label distribution across different data sets. There are 8 emotions, including anger (A), sadness (S), happiness (H), surprise (U), fear (F), disgust (D), contempt (C), and neutral (N).

than test1 and closer to real-life scenarios described in [30]. To the best of our knowledge, we are the first ones to show the ASR and SER performances on the Test2 set since the prior studies ignored the Test2 set. The work uses the primary emotions where every rater can only select one emotion from the given options, including anger (A), sadness (S), happiness (H), surprise (U), fear (F), disgust (D), contempt (C), neutral (N), and "other". We exclude "other" to define the SER task as 8-class multi-label emotion recognition. The corpus also provides transcripts suitable for developing multitask ASR and SER systems. We use the Whisper tokenizer Fig. 2 illustrates the averaged distributional label across different data sets. to normalize the transcripts and remove special comments, e.g., "["00:30:33" Cross Talk" or "[inaudible 00:08:44]". More details about the corpus are in [30].

B. Whisper-SER and Implementation Details

Fig. 1 illustrates the proposed Whisper-SER framework.We employ the Whisper model proposed by [8] (marked in dark blue in Fig. 1) as the primary backbone model since the Whisper architecture is one of the SOTA ASR models. Whisper models range in size from tiny (39M parameters) to large (1.5B parameters). For computing efficiency, our experiments are based on Whisper Tiny. Whisper Tiny has four transformer-based blocks in both the encoder and decoder, with 384 nodes in each layer. We add a module (marked in brown in Fig.1) for the SER task, consisting of one average pooling layer, one fully connected layer with ReLU activation (with 384 nodes), and one prediction layer. The inputs concatenate the hidden states from the last layers of both the encoder and decoder. Additionally, to investigate whether the encoder encodes more emotion-related information than the decoder, we conduct experiments using hidden states from the last layer of the encoder, the decoder, or both.

There are two training stages for the Whisper-SER framework, as introduced in Section III-C. In the first stage, we fine-tune the pre-trained Whisper architecture on the emotional dataset for ASR tasks by setting $\alpha = 1$ in Equation (2). In the second stage, we introduce a non-zero α and jointly optimize the system using a convex combination of ASR and SER losses (equation (2)). The maximum number of generated tokens is 160. We use the Adam optimizer [31] with an initial learning rate of 1e-7 and a batch size of 32, and we apply a warm-up decay learning rate scheduler with a maximum and minimum learning rate of 1e-5 and 1e-7, respectively, for 5 epochs. We train the models for 100 epochs using DeepSpeed stage 2 [32], and we select the best checkpoint based on the lowest loss on the development set. We use PyTorch and the HuggingFace library. Our experiments are conducted on an NVIDIA Tesla V100 GPU with 16 GB of memory, totaling 100 GPU hours for all results.

C. Evaluation Metric

For the ASR task, we use Word Error Rate (WER) as the evaluation metric, and WER values can be calculated as $WER = (S + I + D)/N$, where S, I, D, and N represent the number of substitution, insertion, and deletion errors, and the number of words in the ground-truth transcripts, respectively. In terms of multi-label emotion recognition, the study by [13] investigated the advantages and disadvantages of using distribution-based assessment (e.g., KLD) versus hard-decision-based assessment (e.g., F1 scores) for the SER task. With distribution-based assessment, it is difficult to observe performance differences between the baseline and proposed model, as the scale is very small, making interpretation challenging. Therefore, we decided to use weighted-F1 (**weF1**), macro-F1 (**maF1**), and micro-F1 (**miF1**) scores to measure multi-label classification accuracy. This involves selecting target classes by applying thresholds to the ground truth data. A prediction is deemed successful if the proportion for a class exceeds $1/8$, as mentioned in Section III-A.

TABLE II

ASR AND SER PERFORMANCE RESULTS OF THE MODELS. THE COLUMN **EXPERIMENTS** EXPLAINS THE EXPERIMENTS WE CONDUCT. THE COLUMN **INPUT** INDICATES THE INPUT SOURCE OF THE SER MODELS MENTIONED IN SECTION IV-B. THE COLUMN **FINE-TUNE** INDICATES IF THE PRE-TRAINED MODELS ARE FINE-TUNED. THE COLUMN **#PARA. (M)** INDICATES THE NUMBER OF MODEL PARAMETERS IN MILLIONS (M). ↑ SIGNALS THAT A METRIC IS BETTER IF THE VALUE IS HIGHER. OTHERWISE, WE USE ↓.

	Test Set				Test1				Test2			
Experiments	Input	Backbone	Fine-Tune	#Para. (M)	WER (%)↓	MaF1↑	MiF1↑	WeF1↑	WER (%)↓	MaF1↑	MiF1↑	WeF1↑
SER SOTA [13]		WavLM Large	V	317	-	0.4850	0.6210	0.6060	-	0.4264	0.6127	0.5842
SER SOTA [13]		WavLM Base+	V	94	-	0.4768	0.6077	0.5978	-	0.4248	0.5993	0.5817
Whisper Large V3 [8]		Large V3		1550	12.57				13.63			
Whisper Tiny [8]		Tiny		39	29.75				54.15			
ASR		Tiny	V	39	20.26				19.98			
SER^{DO}	Decoder	Tiny	V	39	769.2	0.4242	0.6183	0.5679	621.2	0.3591	0.5986	0.5303
SER^{EO}	Encoder	Tiny	V	39	1614	0.3691	0.5911	0.5216	1974	0.3338	0.5924	0.5126
SER^{BO}	Both	Tiny	V	39	348.6	0.4295	0.6191	0.5724	239.8	0.3624	0.5989	0.5332
MTL^{DO}	Decoder	Tiny	V	39	24.86	0.4209	0.6175	0.5646	22.68	0.3509	0.5998	0.5244
MTL^{EO}	Encoder	Tiny	V	39	23.80	0.3186	0.5843	0.4775	21.72	0.2858	0.5908	0.4750
MTL^{BO}	Both	Tiny	V	39	25.30	0.4239	0.6133	0.5642	22.57	0.3568	0.5936	0.5263
Whisper-SER	Both	Tiny	V	39	20.90	0.4111	0.6172	0.5611	19.72	0.3563	0.6009	0.5295

D. Experiments

1) Single-task Baseline: We fine-tune the Whisper Tiny model on the MSP-PODCAST as the ASR baseline using $\mathcal{L}_{ASR}$, denoted by **ASR**. In terms of SER, we add light models, including the temporal average pooling layer, Fully-Connected (FC) layer, and prediction layers. To investigate whether the encoder block has more emotional information than the decoder one, we conduct the experiences to extract the embeddings from the last layer of the encoder block (denoted by **EO**), decoder block (denoted by **DO**), or both blocks (denoted by **BO**). In fact, we fine-tune the Whisper Tiny model on the MSP-PODCAST for the SER task using $\mathcal{L}_{SER}$ as the SER baseline, denoted by **SER**. For instance, the $\textbf{SER}^{BO}$ means the input of the light models for the SER task comes from both of the last layers of encoder and decoder blocks.

2) Multi-task Baseline: To show the conflicts between SER and ASR tasks, we conduct experiments to show the performances using standard loss, $\mathcal{L}_{MTL}$, denoted by **MTL**. We also changed the source of inputs for the light model for the SER task to determine which blocks encode more emotional cues (e.g., $\textbf{MTL}^{BO}$).

3) SER and ASR SOTA Models: To understand the gap between the SOTA and the proposed method, we employ pre-trained Whisper Large V3 model[3] as the ASR SOTA model, denoted by **Whisper Large V3**. In terms of SER SOTA, we use the model[4] proposed by [15], and we fine-tune the model on the MSP-PODCAST using the distributional labels by following the study [13]. We denoted the model by **SER SOTA**. The more details about model structure and training approaches are in [15] and [13].

4) Finding the Best α Value in $\mathcal{L}_{MTL}^{\mathcal{W}}$: To find the best α value in $\mathcal{L}_{MTL}^{\mathcal{W}}$ (Equation 2), we conducted the experiments by assigning the α value as from 0.1 to 0.9, as shown in Fig. 3. The 0.8 is the best α value in the work. Besides, when the α is 0.1 or 0.2, the performances for the SER task is higher than the model $\textbf{SER}^{BO}$, and the finding aligns with the previous studies [1], [9].

5) Whisper-SER: Based on the preliminary experiments, we found that using embeddings from both of encoder and decoder blocks ($\textbf{SER}^{BO}$) has better recognition performance. Therefore, the proposed Whisper-SER uses this same structure. In the first stage, we only fine-tune the Whisper components where in the dotted line in Fig. 1 for the ASR task. Then, in the second stage, we fine-tune the whole Whisper-SER model using the $\mathcal{L}_{MTL}^{\mathcal{W}}$ loss. To find the best α value in Equation 2. We conducted the experiments by assigning the α value as from 0.1 to 0.9, as shown in Fig. 3. The 0.8 is the best α value in the work.

[3]https://huggingface.co/openai/whisper-large-v3

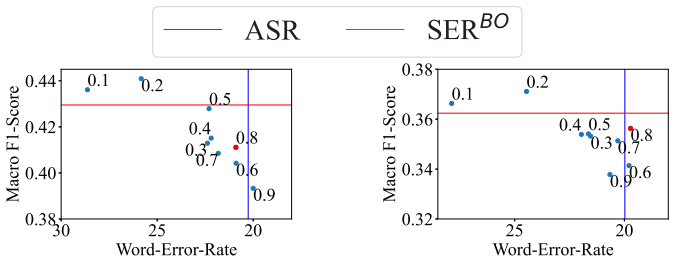
[4]https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

V. RESULTS AND ANALYSIS

Table II summarizes all results on the MSP-PODCAST in various F1 scores across two different test sets. We answer the research questions one by one in detail below.

Does the encoder block of the Whisper have more emotion-related cues? To our surprise, the SER^{DO} and SER^{BO} models have better performance than the SER^{EO}, which might imply that the decoder encodes more emotion-related information than the encoder, which has never been investigated in the previous. We are the first work to report this finding, suggesting we should consider a decoder as well.

Are there conflicts between SER and ASR tasks? We find that the joint training with ASR and SER using vanilla multi-task loss, $\mathcal{L}_{MTL}$, degrades ASR and SER task performance. For instance, the model, **ASR**, has 20.26% WER and $\textbf{SER}^{BO}$ has 42.95% in macro-F1 score on the Test1. However, the $\textbf{MTL}^{BO}$ has 25.30% WER and 42.39% macro-F1 score ASR and SER tasks, respectively. All performances decreased.



(a) Test1 Set (b) Test2 Set

Fig. 3. Relationship between ASR/SER performances and α values in $\mathcal{L}_{MTL}^{\mathcal{W}}$ (Equation 2). We mark the point in red to show the best α considering both ASR and SER performances.

TABLE III

OVERVIEW OF ASR AND SER PERFORMANCES WITH THE THREE FACTORS (INTER-RATER AGREEMENT, SPEAKING RATE, AND DURATION LENGTH). WE USE THE WER (%)↓ AND MAF1↑ TO SHOW PERFORMANCES ON ASR AND SER TASKS, RESPECTIVELY. THE BOLD NUMBERS MEAN THE WORST PERFORMANCES ON ASR OR SER TASKS.

Factor	**Inter-rater Agreement**						**Speaking Rate**						**Duration Length**					
Task	**ASR**			**SER**			**ASR**			**SER**			**ASR**			**SER**		
Experiments	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Slow**	**Medium**	**Quick**	**Slow**	**Medium**	**Quick**	**Short**	**Medium**	**Long**	**Short**	**Medium**	**Long**
Whisper Large V3 [8]	12.64	12.50	**14.15**				**24.74**	12.15	11.47				**16.92**	12.69	11.38			
Whisper Tiny [8]	39.54	38.73	**40.95**				**61.83**	39.69	37.41				**50.59**	40.66	35.43			
ASR	21.16	21.14	**24.05**				**36.68**	21.67	21.60				**28.51**	22.95	18.53			
SER^{BO}				0.411	0.423	**0.296**				0.404	0.410	0.405				0.381	0.407	0.423
MTL^{BO}	22.64	24.77	**31.75**	0.393	0.416	**0.306**	**51.23**	25.94	26.18	0.403	0.404	0.400	**37.52**	27.73	21.84	0.381	0.402	0.413
Whisper-SER^{BO}	23.72	22.87	**27.05**	0.403	0.408	**0.296**	**42.15**	23.88	22.77	0.393	0.397	0.399	**31.42**	24.94	20.35	0.378	0.398	0.398

Why should we use two stages to train the models for SER and ASR tasks? The results of **MTL**^{DO}, **MTL**^{EO}, and **MTL**^{BO} show that the models use the standard multi-task learning without two-stage training. The WER is higher than the results of the **ASR** model, so the standard multi-task learning can not benefit the ASR and SER tasks. However, in the practical application, the ASR task is very important, so our main goal is to maintain the same level of performance on the ASR task and have the model recognize emotions. Based on the results of the proposed framework, Whisper-SER, the two stages are critical to achieving the goal.

Can we unify ASR and multi-label SER multitasks in one shared-layer system without ASR performance degradation? Yes. To overcome the conflicts between SER and ASR, the adapted two-stage training strategy using the weighting loss ($\mathcal{L}_{MTL}^{\mathcal{W}}$) can mitigate this issue by keeping ASR performance the same as **ASR**. Fig. 3 shows the relationship between ASR/SER performances and α values in Equation 2. Using both encoder and decoder features, Whisper-SER, when α is 0.8, achieves the best overall performance for both ASR and SER tasks. The best-performing model, Whisper-SER, has on-par ASR and SER accuracy with the baseline models (**ASR** and **SER**) on Test1 and Test2.

What are the relationships between ASR and SER performances and inter-rater agreement, speaking rate, and duration lengths? We first combine the test1 and test2 as one test set. Table IV summarizes the distributions considering inter-rater agreement, speaking rate, and duration length across Train, Development, and Test1+Test2 sets. We calculate the mean and standard deviation (std.) from the Train set and define two thresholds by (mean - std.) and (mean + std.) to split the data into three groups (high/slow/short, medium, and low/quick/long) based on the values of inter-rater agreement, speaking rate, and duration length, respectively.

TABLE IV

DATA DISTRIBUTIONS CONSIDERING INTER-RATER AGREEMENT, SPEAKING RATE, AND DURATION LENGTH ACROSS TRAIN, DEVELOPMENT, AND TEST1+TEST2 SETS.

Data Set		**Train**	**Development**	**Test1+Test2**
Total Number		**84,030**	**19,815**	**45,462**
Inter-rater Agreement	**Low**	5.33%	4.11%	4.96%
	Medium	79.06%	81.14%	76.79%
	High	15.61%	14.76%	18.25%
Speaking Rate	**Low**	14.97%	3.52%	2.24%
	Medium	72.10%	85.01%	79.48%
	High	12.94%	11.47%	18.28%
Duration Length	**Low**	17.59%	16.40%	21.30%
	Medium	62.28%	64.65%	61.09%
	High	20.13%	18.95%	17.60%

Then, we measure the inter-rater agreement of each sentence by Cohen's Kappa [33], speaking rate, and duration length on the MSP-PODCAST. We calculate the mean and standard deviation (std.) from the train set and define two thresholds by (mean - std.) and (mean + std.) to split the data into three groups (high/slow/short, medium, and low/quick/long) based on the values of inter-rater agreement, speaking rate, and duration length, respectively. The detailed distributions are summarized in Table IV.

Table III summarizes the performances on the ASR and SER tasks considering the three factors. To our surprise, all models in Table III performed worst on the high-agreement samples for ASR and SER tasks. With the finding on the ASR task, we align with prior studies [1], [6], [9], [34] that ASR systems have worse accuracy on emotion-rich utterances than emotion-neutral ones. The potential reason might be because the pronunciation tends to be wrong in strong emotions (e.g., anger) [35]. Also, Chou et al. [13] also reported that adding more ambiguous (lower inter-rater agreement) samples can decrease performance on the samples that have a high inter-rater agreement, but adding the ambiguous samples can have better effectiveness and robustness on the ambiguous samples, which is a real-world scenario that we can not know which sample has a high or low inter-rater agreement. This perspective is very important since the previous studies removed the ambiguous sample from the test set. For instance, the 19.85% samples in the test set were removed by the plurality rule on the MSP-PODCAST.

Besides, for the speaking rate and duration length, we found that the models performed worse on the utterances with a slow speaking rate (less than two words per second) and the ones with a short duration (less than 3.39 seconds) on the ASR task. However, there are no apparent performance differences on the SER task. This is the first work to observe and reveal the relationship between the performance of ASR and SER and the three factors.

VI. CONCLUSIONS AND FUTURE WORK

This paper uses a modified two-stage training strategy with weighted loss to jointly model ASR and **multi-label** SER tasks based on the Whisper architecture, evaluated on the **complete test set** of the MSP-PODCAST corpus without discarding any data samples, including those without a consensus label. Our findings indicate that simultaneous training for ASR and SER with equal weights for each loss can cause catastrophic forgetting of pre-trained ASR tasks. To circumvent this issue, we use the two-stage weighting training method to maintain ASR and SER performance levels comparable to the pre-trained ASR and SER models, respectively. The proposed model, **Whisper-SER**, achieves a 29.75% and 63.58% relative reduction in WER for the ASR task on the Test1 and Test2 sets, respectively, compared to the **Whisper Tiny** baseline model. However, **Whisper-SER** achieves only a 4.28% and 1.68% relative reduction in macro F1

on the Test1 and Test2 sets, respectively, for the SER task compared to $\mathbf{SER}^{BO}$.

In further analysis, we are the first to reveal that modern models perform worse on ASR tasks when utterances have higher inter-rater agreement on emotions, slower speaking rates, and shorter durations. Additionally, this study is the first to demonstrate the possibility of multitasking ASR and multi-label SER models without ASR performance degradation. In future work, we aim to use LoRa [36] to train the larger Whisper model (Whisper-SER) on limited computational GPU resources and address imbalanced emotion distributions using class-balanced loss [37].

REFERENCES

[1] Y. Gao *et al.*, "Enhancing Two-Stage Finetuning for Speech Emotion Recognition Using Adapters," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 316–11 320. DOI: 10.1109/ICASSP48485.2024.10446645.

[2] M. Kang *et al.*, "ZET-Speech: Zero-shot adaptive Emotion-controllable Text-to-Speech Synthesis with Diffusion and Style-based Models," in *Proc. INTERSPEECH*, 2023, pp. 4339–4343. DOI: 10.21437/Interspeech.2023-754.

[3] T. Feng and S. Narayanan, "Foundation Model Assisted Automatic Speech Emotion Recognition: Transcribing, Annotating, and Augmenting," *arXiv preprint arXiv:2309.08108*, 2023.

[4] N. Abhishek and P. Bhattacharyya, "" We care": Improving Code Mixed Speech Emotion Recognition in Customer-Care Conversations," *arXiv preprint arXiv:2308.03150*, 2023.

[5] Y. Li *et al.*, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Interspeech*, Graz, Austria, Sep. 2019, pp. 2803–2807.

[6] Y. Li *et al.*, "Fusing ASR Outputs in Joint Training for Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7362–7366. DOI: 10.1109/ICASSP43922.2022.9746289.

[7] T. Feng and S. Narayanan, "PEFT-SER: On the Use of Parameter Efficient Transfer Learning Approaches For Speech Emotion Recognition Using Pre-trained Speech Models," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, Sep. 2023.

[8] A. Radford *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 23–29 Jul 2023, pp. 28 492–28 518.

[9] Y. Li *et al.*, "ASR and Emotional Speech: A Word-Level Investigation of the Mutual Impact of Speech and Emotion Recognition," in *Proc. INTERSPEECH*, 2023, pp. 1449–1453. DOI: 10.21437/Interspeech.2023-2078.

[10] Y. Gao *et al.*, "Two-stage Finetuning of Wav2vec 2.0 for Speech Emotion Recognition with ASR and Gender Pretraining," in *Proc. INTERSPEECH*, 2023, pp. 3637–3641. DOI: 10.21437/Interspeech.2023-756.

[11] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, E7900–E7909, 2017.

[12] A. S. Cowen and D. Keltner, "Semantic Space Theory: A Computational Approach to Emotion," *Trends in Cognitive Sciences*, vol. 25, no. 2, pp. 124–136, 2021, ISSN: 1364-6613.

[13] H.-C. Chou *et al.*, "Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule," *IEEE Transactions on Affective Computing*, pp. 1–15, 2024. DOI: 10.1109/TAFFC.2024.3411290.

[14] L. Pepino *et al.*, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.

[15] J. Wagner *et al.*, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023. DOI: 10.1109/TPAMI.2023.3263585.

[16] H. Wu *et al.*, *EMO-SUPERB: An In-depth Look at Speech Emotion Recognition*, 2024. arXiv: 2402.13018 [eess.AS].

[17] S. Yoon *et al.*, "Multimodal Speech Emotion Recognition Using Audio and Text," in *IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, Dec. 2018, pp. 112–118. DOI: 10.1109/SLT.2018.8639583.

[18] Y. Gong *et al.*, "Joint Audio and Speech Understanding," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2023)*, Taipei, Taiwan, Dec. 2023.

[19] G. Ioannides *et al.*, "Towards Paralinguistic-Only Speech Representations for End-to-End Speech Emotion Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 1853–1857. DOI: 10.21437/Interspeech.2023-497.

[20] T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, Oct. 2019. DOI: 10.48550/arXiv.1910.03771. arXiv: 1910.03771 [cs.CL].

[21] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[22] W. Wu *et al.*, "Integrating Emotion Recognition with Speech Recognition and Speaker Diarisation for Conversations," in *Proc. INTERSPEECH*, 2023, pp. 3607–3611. DOI: 10.21437/Interspeech.2023-293.

[23] Y. Li *et al.*, "Evaluating Parameter-Efficient Transfer Learning Approaches on SURE Benchmark for Speech Understanding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, Rhodes island, Greece, 2023, pp. 1–5.

[24] H. Feng *et al.*, "End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model," in *Proc. Interspeech 2020*, 2020, pp. 501–505. DOI: 10.21437/Interspeech.2020-1180.

[25] C. Busso *et al.*, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[26] C.-C. Lee *et al.*, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011, Sensing Emotion and Affect - Facing Realism in Speech Processing, ISSN: 0167-6393.

[27] K. Vansteelandt *et al.*, "The co-occurrence of emotions in daily life: A multilevel approach," *Journal of Research in Personality*, vol. 39, no. 3, pp. 325–335, Jun. 2005. DOI: 10.1016/j.jrp.2004.05.006.

[28] M. Tjalve and M. Huckvale, "Pronunciation variation modelling using accent features," in *Proc. Interspeech 2005*, 2005, pp. 1341–1344. DOI: 10.21437/Interspeech.2005-487.

[29] P. Riera *et al.*, "No sample left behind: Towards a comprehensive evaluation of speech emotion recognition system," in *Proc. Workshop on Speech, Music and Mind 2019*, 2019.

[30] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, Oct. 2019. DOI: 10.1109/TAFFC.2017.2736999.

[31] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.

[32] J. Rasley *et al.*, "DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20, Virtual Event, CA, USA: Association for Computing Machinery, 2020, pp. 3505–3506, ISBN: 9781450379984.

[33] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[34] S. Sahu *et al.*, "Multi-Modal Learning for Speech Emotion Recognition: An Analysis and Comparison of ASR Outputs with Ground Truth Transcription," in *Proc. Interspeech*, 2019, pp. 3302–3306. DOI: 10.21437/Interspeech.2019-1149.

[35] Z. Aldeneh and E. M. Provost, "You're Not You When You're Angry: Robust Emotion Features Emerge by Recognizing Speakers," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1351–1362, 2023. DOI: 10.1109/TAFFC.2021.3086050.

[36] E. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR 2022)*, 2022.

[37] Y. Cui *et al.*, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, California, USA, Jun. 2019.

Experimental Evaluation of Speech Enhancement for In-Car Environment Using Blind Source Separation and DNN-based Noise Suppression

Yutsuki Takeuchi*, Taishi Nakashima*, Nobutaka Ono*,
Takashi Takazawa†, Shuhei Shimanoe†, and Yoshinori Tsuchiya†
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: takeuchi-yutsuki@ed.tmu.ac.jp
† MIRISE Technologies, Aichi, Japan

***Abstract*—In this study, we evaluate speech enhancement using multi-channel blind source separation (BSS) and single-channel DNN-based noise suppression in an in-car environment. Specifically, blind source separation is applied to a mixture of target speech and driving noise, and one of the outputs is fed into the DNN. This combination allows for speaker position-independent enhancement through blind source separation, while the single-channel DNN can be trained regardless of the microphone position. To further improve performance, we fine-tune the pre-trained model using the BSS outputs, including driving noise. Our evaluation criteria include signal-to-distortion ratio improvement (SDRi) for noise reduction performance, and accuracy and character error rate (CER) for speech recognition performance. Results demonstrate that SDRi for the proposed method reached 28.37 dB. Additionally, the average speech recognition accuracy improved from 0.785 to 0.826, and CER improved from 15.3% to 11.9%. These findings demonstrate the potential of combining BSS and DNN to enhance speech recognition capabilities in in-car environments.**

I. Introduction

Speech recognition technology has served as a vital interface for various systems, enhancing safety and convenience. In in-car environments, voice-based interfaces are particularly crucial as they allow drivers to operate systems without taking their hands off the steering wheel. Additionally, the development of speech recognition systems for rear seats or while reclined has gained attention owing to the difficulty of physically reaching the controls [1]. However, the in-car environment includes both stationary noises, such as engine and wind noise, and nonstationary noises, such as sounds from the car stereo and conversations among passengers. These noise sources significantly degrade the effectiveness of speech recognition technologies, making reliable noise reduction techniques essential.

When a microphone array is available, beamforming [2] is one of the valuable methods to enhance the speech coming from a target direction. Some previous studies demonstrate this technique [3]–[5]. However, although the driver's position can generally be assumed to remain static, the relative positions of speakers in the rear seats to the microphones may vary depending on the car type and seating arrangement. As a result, reliance solely on prior information is not feasible, and source direction estimation and steering vector estimation become necessary.

In contrast, multichannel blind signal separation (BSS) [6]–[8] works without prior information about the sound sources, making it a flexible framework for noise and speech separation. Thanks to an efficient update rule of parameters [9], this approach requires less computation, and online real-time applications have been developed [10]–[12]. Some previous works show the effectiveness of BSS in car environments [13], [14]. However, because it does not use prior information about sound sources, it may not perform well in complex noisy in-car environments.

Additionally, deep neural networks (DNNs) have been applied to single-channel or multichannel speech enhancement [15]–[19]. Related works show the high separation performance using DNNs in an in-car environment [20], [21]. Although most DNN-based source-separating processing is nonlinear and unsuitable for speech recognition, they can achieve higher noise reduction performance.

Motivated by these considerations, our focus in this study is to improve speech recognition accuracy in car environments by enhancing speech using microphone arrays. To achieve this, we explore a method that combines multi-channel blind source separation (BSS) and single-channel DNN-based speech enhancement. This approach enables enhancement that is independent of the speaker's position through blind source separation, while the single-channel DNN can be trained without regard to the microphone position. Additionally, by using BSS as a frontend, we expect to potentially eliminate the influence of multiple speakers and car stereo sounds, although this aspect is not evaluated in this study. To achieve better performance, we fine-tune the pre-trained model using paired data of BSS outputs and clean speech. We used auxiliary-function-based independent vector analysis (AuxIVA) [9] as the BSS method, as it is a standard multichannel BSS method, and considering future online evaluations [10], [11], [22], [23]. We also used Conv-TasNet [24] as the DNN model. The experimental results show the effectiveness of combining multi-channel BSS and single-channel DNN with fine-tuning.

II. Problem Settings

Let N be the number of microphones. Let $\boldsymbol{s}_t \in \mathbb{C}^N$ and $\boldsymbol{n}_t \in \mathbb{C}^N$ be the multi-channel observation of clean target speech signal and noise in the time domain, where t is the index of the discrete time. The observed signal $\boldsymbol{x}_t \in \mathbb{C}^N$ can be represented as $\boldsymbol{x}_t = \boldsymbol{s}_t + \boldsymbol{n}_t$.

The objective of this study is to recognize speech in an in-car environment. Given that the speech recognition engine used in this study is treated as a black box and no retraining of the engine is considered, our aim is to enhance the speech as much as possible at the frontend through speech enhancement, that is, to estimate s_t from $\boldsymbol{x}_t$.

III. Proposed Approach

A. Combination of Multichannel BSS and Single Channel DNN-based Speech Enhancement

Recent studies demonstrate the effectiveness of DNN for speech enhancement, but in the context of speech recognition, DNN-based speech enhancement is not always effective due to nonlinear filtering, especially when the speech recognition engine is not retrained. To overcome this, we consider the combination of multi-channel BSS and single-channel speech enhancement with DNN. Fig. 1 shows the block diagram of the process. By applying a single-channel speech enhancement with DNN to the output of multichannel BSS, we expect less nonlinear distortion. Previous works show the effectiveness of the combination of DNN and source separation [25].

Let $\boldsymbol{X}_{lk} \in \mathbb{C}^N$ be the multi-channel observation in the time-frequency domain, which is obtained by short-time Fourier transform (STFT) from $\boldsymbol{x}_t$, where l and k are the indices of the time frames and frequency bins, respectively. The source separation by multi-channel BSS can be represented as

$$\boldsymbol{Y}_{lk} = \boldsymbol{W}_k \boldsymbol{X}_{lk} \tag{1}$$

where $\boldsymbol{W}_k \in \mathbb{C}^{N \times N}$ is the demixing matrix estimated by BSS at frequency index k and $\boldsymbol{Y}_{lk} \in \mathbb{C}^N$ denotes the separated signal. Note that $\boldsymbol{W}_k$ works as a time-invariant linear filter.

From $\boldsymbol{Y}_{lk}$, we select the separated signal that contains speech and obtain the time-domain signal y_t by applying inverse STFT (ISTFT). Note that y_t is a single-channel signal, not a multi-channel signal. We then apply DNN-based single-channel speech enhancement to y_t.

$$\hat{s}_t = f(y_t) \tag{2}$$

where $f(\cdot)$ represents the DNN-based single-channel speech enhancement.

B. Fine-Tuning of DNN Using BSS Outputs

Generally, it is essential to train DNN models on highly generalizable datasets and to guarantee performance on untrained data. However, in this study, we use BSS outputs as input data for the DNN, and pre-trained models may not perform adequately on such BSS outputs. Therefore, we propose fine-tuning the pre-trained model with BSS outputs. To verify the effectiveness of this approach, we compare two types of fine-tuning:

Fine-tuning A (FT-A):
As fine-tuning for driving noise in an in-car environment, the DNN is retrained by using mixed signals of speech and driving noise as inputs and clean speech as outputs.

Fine-tuning B (FT-B):
First, multi-channel observations of speech and driving noise are processed by BSS to obtain separated signals. These BSS outputs are then used as inputs, and clean speech as outputs, to retrain the DNN.

For both types of fine-tuning, the dataset is divided into training and validation sets. The DNN is retrained such that the loss function on the training data is minimized. Training is stopped when the loss function value on the validation data starts to increase, indicating overfitting.

IV. Experiments

A. Data Acquisition

In this study, evaluation test data were created from speech sources recorded in an anechoic room, impulse responses, and driving noise recorded in an actual car using a microphone array.

The vehicle Alphard (Toyota, 2022) was used as the actual car. In the car, a microphone array was fixed on the ceiling between the 1st row seats and the 2nd row seats (Fig. 2). The microphone array consisted of eight high-SNR MEMS microphones (TDK InvenSense, SNR=$74\,\mathrm{dBA}$), and the distance between microphones was $6\,\mathrm{mm}$ (Fig. 3). Analog–digital converters (ADCs) were mounted on the microphone printed circuit board (PCB) and controlled by a microcomputer (μC) to synchronize all microphone channel samplings. The time-stretched pulse (TSP) playback system and microphone array were independent regarding sampling timing. On the other hand, eight microphone channels were synchronized with each other (Fig. 4).

To create the target speech signal, 48 utterances of speech sources spoken by two males and two females were recorded in an anechoic room. The utterances likely to be spoken in a car situation were chosen.

Impulse responses were measured using a head and torso simulator (HATS; Type4128, Brüel & Kjær) and the TSP method. A TSP was played from HATS, which was mounted on the 2nd seat of the parked car, and recorded by the microphone array. The recorded TSPs were converted to impulse responses by convolving inverse TSPs. Because the position of HATS changes in accordance with the seat position, the evaluation seat positions were set to three conditions at the 2nd seat (Fig. 5). Pos. 1 is the nearest position between the microphone array and HATS. Pos. 2 is the normal seating position. Pos. 3 is the farthest position between the microphone array and HATS in the 2nd seat. Noise data were recorded in a car driving at $100\,\mathrm{km/h}$ on the Tokai–Kanjo Expressway (Aichi, Japan) and extracted to obtain steady noise with neither road connections nor over-taking cars.

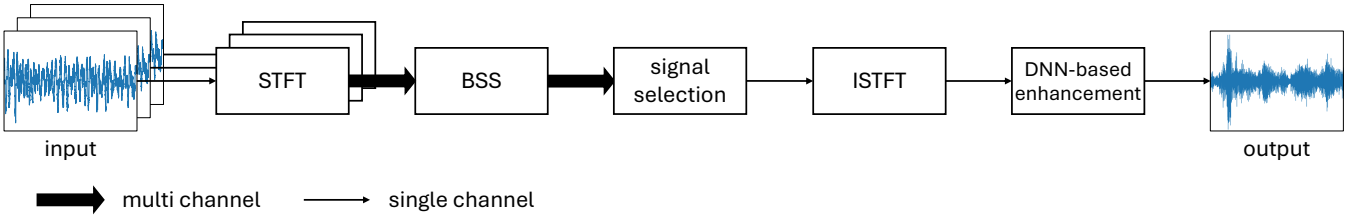


Fig. 1. Block diagram of processing

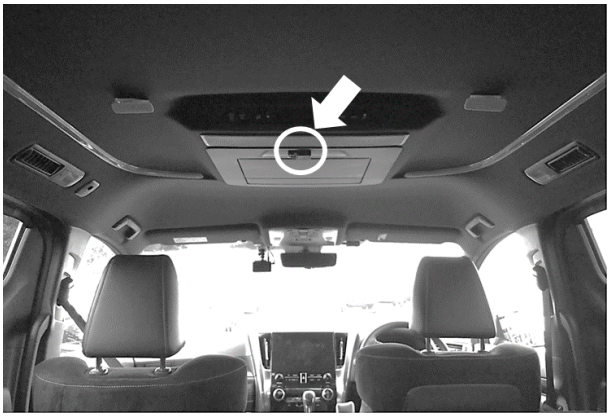


Fig. 2. Position of microphone array

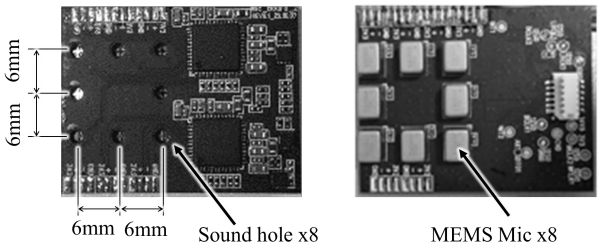


Fig. 3. Details of microphone array

The speech sources, recorded at a 44.1 kHz sampling, were downsampled to a 16 kHz sampling rate. The TSP responses were recorded at the 16 kHz sampling rate, so the impulse responses were also at a 16 kHz sampling rate. The target speech signal was created by convolving the impulse response. The amplitude of the target speech signal was compensated for so that the amplitude was almost the same as that of an actual

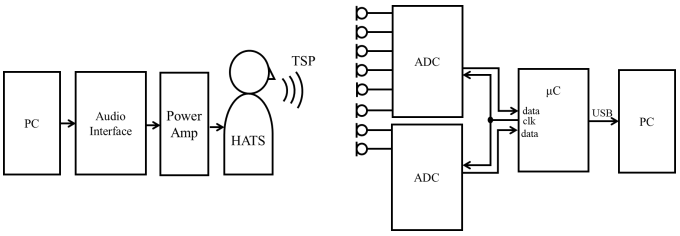


Fig. 4. Block diagram of TSP response recording

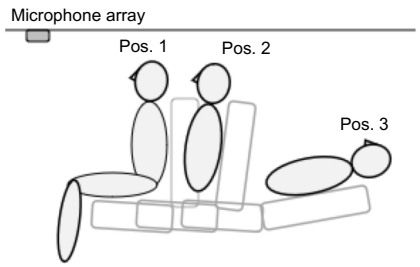


Fig. 5. Seat positions for impulse responses measurement

human voice in the car. The driving noise was also recorded at a 16 kHz sampling rate. The TSP response and driving noise were recorded in the same car and with the same microphone fixture.

B. Methods

The evaluation test data were created by superposing the target speech and driving noise signals with signal amplitude compensation to simulate the actual car noise level. As a result of this postprocess, three conditions with the same driving noise level, the same speech source, and different transfer function of evaluation test data were created. The length of these evaluation test data was 24 s and included 1–3 s of speech.

The training data for fine-tuning described in III-B were also created with these data. We prepared training data by separating the evaluation test data using AuxIVA, and ground truth with clean speech signal. From four speakers, we selected two speakers for training set, one for validation set, and the other for test set.

In this study, we used Conv-TasNet [24] for the training model. This model consists of three blocks: Encoder, Separation, and Decoder. The encoder module transforms the input mixture signal into intermediate representations for each segment. And for them, the separation module applies the mask vector to separate sources. Then, the decoder module reconstructs the signal.

We used the Conv-TasNet model pre-trained with Libri1Mix dataset [26], [27]. We selected this model for its compatibility with the dataset used in this study and its sufficient amount of training. We fine-tuned this model with the training set

TABLE I
EXAMPLES OF SPEECH CONTENT AND MEANING

Speech content	Meaning
Onryo wo ageru.	Turn up the volume.
Annai chushi.	Stop the navigation guidance.

described in IV-A.

For the BSS methods applying DNN as post-processing described in III-A, we employed AuxIVA based on prior experiments. The STFT frame length was set to 4096 samples, and the number of iterations was fixed at 100. In this study, we manually selected the separated signal containing speech from multiple separated signals. Developing an automatic selection method is planned for future work. To reduce low-frequency noise, we first applied a high-pass filter with a cut-off frequency of $100\,\mathrm{Hz}$, and then performed DNN-based post-processing.

Furthermore, to investigate the conditions under which the proposed approach is superior, we evaluated the performance with signal-to-noise ratio (SNR) in $5\,\mathrm{dB}$ increments, ranging from $-10\,\mathrm{dB}$ to $-30\,\mathrm{dB}$.

We used signal-to-distortion ratio improvement (SDRi) [28] to evaluate the performance of noise reduction. The result is obtained by subtracting the SDR before separation from the SDR after separation.

To evaluate speech recognition performance, we defined speech recognition accuracy as the percentage of correct speech recognition results. Since command speech is often used in car navigation systems, and the speech recognition result cannot be said to be correct if the actual action does not match the content of command speech, so we used speech recognition accuracy to evaluate speech recognition performance in this study. As a well-used speech recognition engine with high accuracy, we used Google's software (Cloud Speech-to-Text V1) for speech recognition and manually checked whether the recognition content matched the speech content. Table I shows the examples of speech content.

We also calculated character error rate (CER) to objectively evaluate the speech recognition performance. CER can be computed by dividing the sum of the number of substitutions, deletions, and insertions by the number of characters in the reference. We evaluated the performance after extracting speech segments manually.

C. Results

Fig. 6 shows the SDRi results for the driving noise condition. FT stands for Fine-Tuning. The average SDRi in ideal MaxSNR, which is designed by using true data, is $33.76\,\mathrm{dB}$ in pos. 1, and this is considered the limit value of the linear time-invariant filter. In contrast, the average SDRi value in AuxIVA+DNN with FT-B result in pos. 1 was $29.74\,\mathrm{dB}$, and the average SDRi in all conditions was $28.37\,\mathrm{dB}$. SDRi improved by about $10\,\mathrm{dB}$ compared with the AuxIVA results. This result would be attributed to the fact that Conv-TasNet, the nonlinear model, can better interpret more complex source characteristics and nonstationary noise. The characteristic to

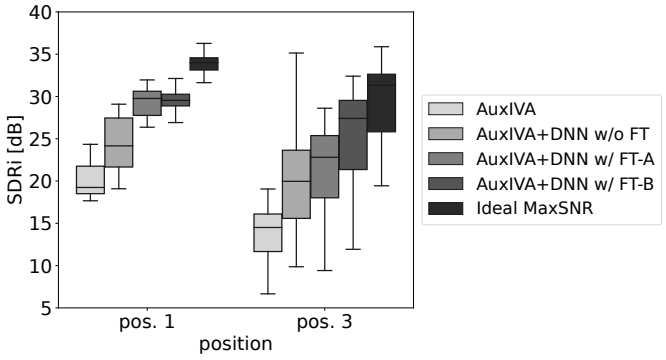


Fig. 6. SDRi for each method (FT-A: fine-tuning with mixture of speech and in-car noise, FT-B: fine-tuning with BSS outputs)

TABLE II
SPEECH RECOGNITION PERFORMANCE FOR EACH METHOD
(*DESINED USING CLEAN SOURCE IMAGES)

Methods		Accuracy		CER [%]	
		Pos. 1	Average	Pos. 1	Average
	Unprocessed	0.938	0.785	2.23	15.3
BSS	AuxIVA	0.917	0.819	4.47	**11.9**
BSS+DNN	AuxIVA+DNN w/o FT	0.875	0.597	15.6	42.1
	AuxIVA+DNN w/ FT-A	0.917	0.757	5.21	17.5
	AuxIVA+DNN w/ FT-B	**0.958**	**0.826**	**1.99**	12.7
	Ideal MaxSNR*	0.958	0.944	1.99	1.99

train significant features from large amounts of data would be one of the most compelling reasons. And comparing with the AuxIVA+DNN without FT result, SDRi improved about $5\,\mathrm{dB}$. From this result, we can confirm the improvement of noise reduction performance by using the BSS outputs as inputs of Conv-TasNet.

Table II shows the results of speech recognition performance under driving noise. The average accuracy in AuxIVA+DNN with FT-B result was 0.041 higher than that of the mixtures. Comparing AuxIVA+DNN with FT-B result with the AuxIVA result, the accuracy was equivalent to or higher score, and it was higher than that of the AuxIVA+DNN without FT result. In addition, the average CER in AuxIVA+DNN with FT-B result was equivalent to ideal MaxSNR result in pos. 1. Generally, the speech recognition performance with Conv-TasNet tends to worsen due to nonlinear filtering, but by using BSS outputs as input, the speech recognition accuracy was almost equivalent to or higher than that of BSS.

Fig. 7 shows the results of speech recognition accuracies for each SNR condition. The results using the proposed approach show the best performance in conditions above $-20\,\mathrm{dB}$. This results demonstrate that proposed approach works under high SNR. In contrast, below $-25\,\mathrm{dB}$, we obtained the best performance with AuxIVA, without Conv-TasNet. When the SNR is low, the noise remaining in the separated signal using AuxIVA is sometimes considered as speech, and the noise is sometimes enhanced as much as target speech. This may be the reason

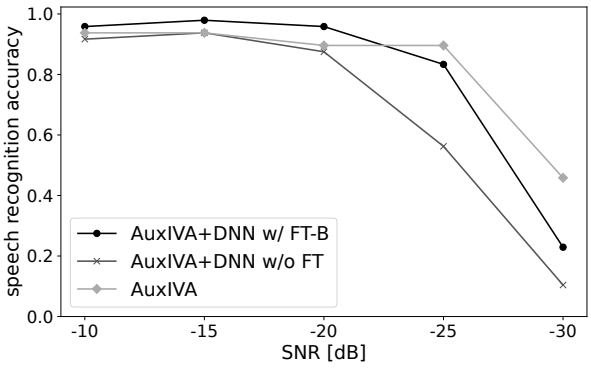


Fig. 7. Speech recognirion accuracy for each SNR (FT-B: fine-tuning with BSS outputs)

for the lower speech recognition rate.

Finally, Fig. 8 shows the spectrograms of the mixtures, the target speech signal, and the separated sound. It can be confirmed that low-frequency noise in the mixture is reduced or almost removed in AuxIVA+DNN with FT-B. However, as seen in before 20 sec of AuxIVA+DNN with FT-B result, the separation filter with Conv-TasNet may interpret the noise as speech. In addition, sparse parts of the actual speech are densely complemented by noise, which could not occur with a linear filter. We consider these results to be the reason for the low speech recognition performance.

V. Conclusion

In this study, we evaluated the performance of a speech enhancement system combining multi-channel BSS and single-channel DNN-based noise suppression in an in-car environment. Our approach demonstrated significant improvements, with an average SDRi of approximately 30 dB. Additionally, the speech recognition accuracy improved in both accuracy and CER criteria. These results indicate that the combination of BSS and DNN-based noise suppression can effectively enhance speech recognition performance in car environments. Future work will focus on leveraging more spatial information by incorporating multi-channel DNNs to further improve speech enhancement performance.

References

[1] M. Fukui, T. Watanabe, and M. Kanazawa, "Sound source separation for plural passenger speech recognition in smart mobility system," *IEEE Transactions on Consumer Electronics*, vol. 64, no. 3, pp. 399–405, 2018.

[2] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2002.

[3] C. Fox, G. Vitte, M. Charbit, J. Prado, R. Badeau, and B. David, "A subband hybrid beamforming for in-car speech enhancement," in *Proceedings of the 20th European Signal Processing Conference*, 2012, pp. 11–15.

[4] A. Chinatto, "In-car speech separation via MVDR," in *Proceedings of XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, Jan. 2017, pp. 122–126.

[5] S. Grimm and J. Freudenberger, "Wind noise reduction for a closely spaced microphone array in a car environment," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, p. 7, Jul. 2018.

[6] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[7] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation*, Springer-Verlag, 2006, pp. 165–172.

[8] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation*, Springer-Verlag, 2006, pp. 601–608.

[9] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 189–192.

[10] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proceedings of Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, 2014, pp. 107–111.

[11] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2017, pp. 216–220.

[12] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online blind source separation based on joint optimization with blind dereverberation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 506–510.

[13] T. Yamada, A. Tawari, and M. M. Trivedi, "In-vehicle speaker recognition using independent vector analysis," in *Proceedings of International IEEE Conference on Intelligent Transportation Systems*, 2012, pp. 1753–1758.

[14] K. Goto, L. Li, R. Takahashi, S. Makino, and T. Yamada, "Study on geometrically constrained IVA with auxiliary function approach and VCD for in-car communication," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2020, pp. 858–862.

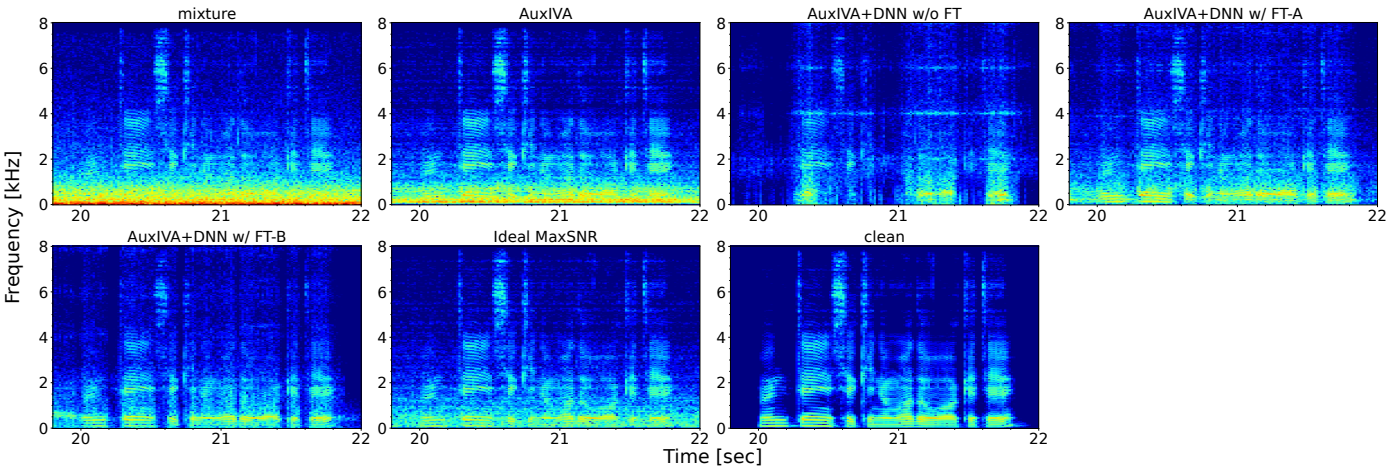


Fig. 8. Spectrograms of mixture, clean, and separated sound for pos. 1 (FT-A: fine-tuning with mixture of speech and in-car noise, FT-B: fine-tuning with BSS outputs)

[15] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3734–3738.

[16] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 116–120.

[17] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 286–290.

[18] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[19] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 196–200.

[20] H. Wang, Z. Ye, and J. Chen, "A speech enhancement system for automotive speech recognition with a hybrid voice activity detection method," in *Proceedings of International Workshop on Acoustic Signal Enhancement*, 2018, pp. 1–9.

[21] V. Kothapally, Y. Xu, M. Yu, S.-X. Zhang, and D. Yu, "Deep neural mel-subband beamformer for in-car speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[22] T. Nakashima and N. Ono, "Inverse-free online independent vector analysis with flexible iterative source steering," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, pp. 749–753.

[23] Y. Kuriki, T. Nakashima, K. Yamaoka, *et al.*, "Efficient low-latency convolution with uniform filter partition and its evaluation on real-time blind source separation," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, pp. 765–769.

[24] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[25] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome, "Unsupervised training for deep speech source separation with kullback-leibler divergence based probabilistic loss function," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 56–60.

[26] M. Hu, *mhu-coder/ConvTasNet_Libri1Mix_enhsingle*, Dec. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4301955.

[27] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, *LibriMix: An open-source dataset for generalizable speech separation*, 2020. arXiv: 2005.11262.

[28] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

PBJDT: Point-Based Joint Detection-and-Tracking

Zhen-Xun Lee[1] and Jian-Jiun Ding [2*]
[1] National Taiwan University, Taiwan
E-mail: r11942118@ntu.edu.tw Tel: +886-937265017
[2*] National Taiwan University, Taiwan
E-mail: jjding@ntu.edu.tw Tel: +886-233669652

***Abstract*—Multiple object tracking is a more and more important technique with the rapid development of autonomous driving. In this work, we introduce a keypoint-based multiple object tracking method aim to address the challenges in complex road environments. It utilizes anchor-free keypoint detection to reduce computational resources while extracting robust features through an improved deep learning model architecture. We implement a background suppression technique to minimize false detections and incorporate the information from adjacent frames to capture object motion characteristics. To address occlusion and overexposure scenarios, we innovatively combine a Kalman filter with the detector and dynamically adjust the tracking strategy based on detection confidence. Experimental results demonstrate that the proposed method can be performed in real time and achieve a MOTA of 92.51% on the KITTI tracking dataset, which outperforms state-of-the-art methods.**

I. INTRODUCTION

Multiple Object Tracking (MOT) is a rapidly evolving and crucial research area within computer vision, with wide-ranging applications in surveillance systems, autonomous driving, human action analysis, and other fields. It aims to simultaneously track the movement trajectories of multiple objects in video sequences. The requirement for efficient and accurate MOT systems has grown exponentially, particularly in autonomous driving, where real-time tracking of numerous vehicles, pedestrians, and other road users is critical for safe navigation. MOT approaches can be broadly categorized based on how they capture object relationships, with tracking-by-detection and tracking-by-attention being prominent paradigms.

There are several well-known tracking-by-detection methods, e.g., CenterTrack [1], FairMOT [2], and SORT [3]. These approaches utilize detection results from adjacent frames to capture object relationships, followed by a refinement stage to manage trajectories. Conversely, prominent tracking-by-attention methods such as TrackFormer [4], ByteTrack [5], and MOTR [6] employ attention mechanisms to capture object relationships across frames.

Furthermore, in the detection step, approaches can be differentiated by their pipeline structure and methodology: some employ pre-defined anchors to detect objects, while others utilize anchor-free methods, directly predicting detection results from the backbone network.

Traditional multiple object tracking methods have predominantly relied on anchor-based mechanisms, wherein multiple predefined anchors are strategically placed within the image for object detection and tracking. While these anchor-based approaches have substantially enhanced detection and tracking accuracy, they are not without limitations. Key challenges include the necessity for precise anchor configuration, high computational demands, and suboptimal performance in tracking small objects and navigating dense scenes.

As the complexity of real-world scenarios escalates, particularly in urban driving environments, these limitations have become increasingly pronounced. Consequently, there is a growing need for novel solutions that can effectively address these challenges, especially in scenarios characterized by occlusion and rapid illumination changes.

To address these challenges, we propose an innovative approach to multiple object tracking. Our method offers more flexible handling of objects under varying illumination conditions and occlusions, reduces computational costs, and demonstrates superior performance in challenging scenarios.

In this paper, we introduce PBJDT (Point-Based Joint Detection and Tracking), a novel framework designed to tackle these MOT challenges. Our main contributions include:

1. An hourglass backbone architecture incorporating multi-scale feature fusion is proposed to enhance the robustness of extracted features.
2. A novel background suppression method is applied to feature heatmaps to mitigate the detection misjudgments caused by background interference.
3. A flexible system that applied Kalman filtering is proposed to enhance the tracking robustness in challenging conditions, particularly in the scenarios characterized by occlusions and rapid illumination changes.

Preliminary results are highly promising, with PBJDT achieving a Multiple Object Tracking Accuracy (MOTA) [17] of 92.51% on the KITTI tracking dataset [16], surpassing existing state-of-the-art methods. This performance underscores the potential of our approach in advancing the field of multiple object tracking.

TABLE I. SUMMARY OF MOT PROCESSES FOR DIFFERENT TRACKING PARADIGMS

Method \ Process	Feature Extraction	Correlation	Refinement
Tracking-by-detection	Uses DNNs to extract features from detected objects	Correlates detected instances with existing tracklets	Applies Re-ID techniques to refine tracking trajectories
Tracking-by-attention	Employs attention modules to extract and focus on relevant features	Correlates tokens (which can represent objects or parts of objects)	Refines input features based on feature correlation

II. RELATED WORKS

Existing literature on multiple object tracking (MOT) can be categorized according to their respective tracking paradigms.

A. Tracking-by-Detection

Tracking-by-detection has emerged as a widely adopted paradigm in multiple object tracking. This approach first detects objects independently in each frame, then links these detections across frames to form trajectories. The tracking-by-detection framework gained prominence with the advent of powerful object detectors such as the Faster R-CNN [7], YOLO [8], the SSD [9], and the CenterNet [10].

Tracking-by-detection methods can be broadly classified into two main categories: one-stage and two-stage methods. Two-stage methods bifurcate the tracking task into object detection and data association, while one-stage methods simultaneously detect object information and learn inter-frame relationships. For instance, CenterTrack [1] exemplified a classic one-stage framework, employing point-based detection for concurrent detection and tracking. For refinement, some methods incorporate additional modules to enhance trajectory management. FairMOT [2] utilized Re-identification (Re-ID) to address id switch issues, while SORT [3] employed the Hungarian algorithm for data association.

B. Tracking-by-Attention

Tracking-by-attention leverages attention mechanisms to enhance the tracking process. This approach enables the model to focus on salient parts of the input data, thereby improving the accuracy and robustness of object tracking, particularly in scenarios involving occlusions and complex object interactions.

Notable examples in this category include:

1. TransTrack [11], which integrated transformers with a traditional detection-based tracking framework.
2. TrackFormer [4], which introduced a transformer-based architecture for end-to-end object tracking, utilizing attention mechanisms to directly model inter-frame object interactions.
3. ByteTrack [5], which proposed a simple yet effective association method that exploits both high and low confidence detections through Kalman filtering, significantly enhancing tracking performance in crowded scenes.

While tracking-by-attention methods offer improved handling of occlusions and complex object interactions, they face several challenges. These include high computational complexity, substantial requirements for training data, and performance limitations in extremely dense scenes.

C. Summary and Other Approaches

We summarize the key processes of tracking-by-detection and tracking-by-attention methods in Table I, highlighting their approaches to feature extraction, correlation, and refinement. Tracking-by-detection and tracking-by-attention are two predominant paradigms in MOT.

Tracking-by-regression: This method directly regresses the object's position in subsequent frames, often using neural networks trained on sequential data.

Tracking-by-segmentation: This approach combines instance segmentation with tracking, providing more detailed object information.

Hybrid approaches: These methods integrate multiple techniques to achieve enhanced performance. For example, some methods combine detection-based tracking with attention mechanisms or incorporate segmentation information into the tracking pipeline.

These diverse approaches reflect the ongoing innovation in the field of multiple object tracking, each offering unique strengths in addressing the challenges of complex tracking scenarios.

III. PROPOSED METHOD

Regarding illumination changes, when vehicles pass through tunnels or shaded areas, images captured by dashcams may be overexposed or underexposed, leading to misjudgments by detectors. The image information in these scenes is compromised due to abrupt lighting changes. On the other hand, object occlusion is commonplace in everyday scenarios, where target objects may be partially obscured by background elements or other objects, or become difficult to detect due to changes in camera perspective.

In this section, we introduce our novel approach, 'Point-based Joint Detection and Tracking' (PBJDT). The overall pipeline is shown in Fig. 3. This method is designed to address the challenges of object detection and tracking in scenarios characterized by rapid illumination changes and frequent object occlusions. PBJDT integrates three key components to enhance detector robustness in these challenging environments:

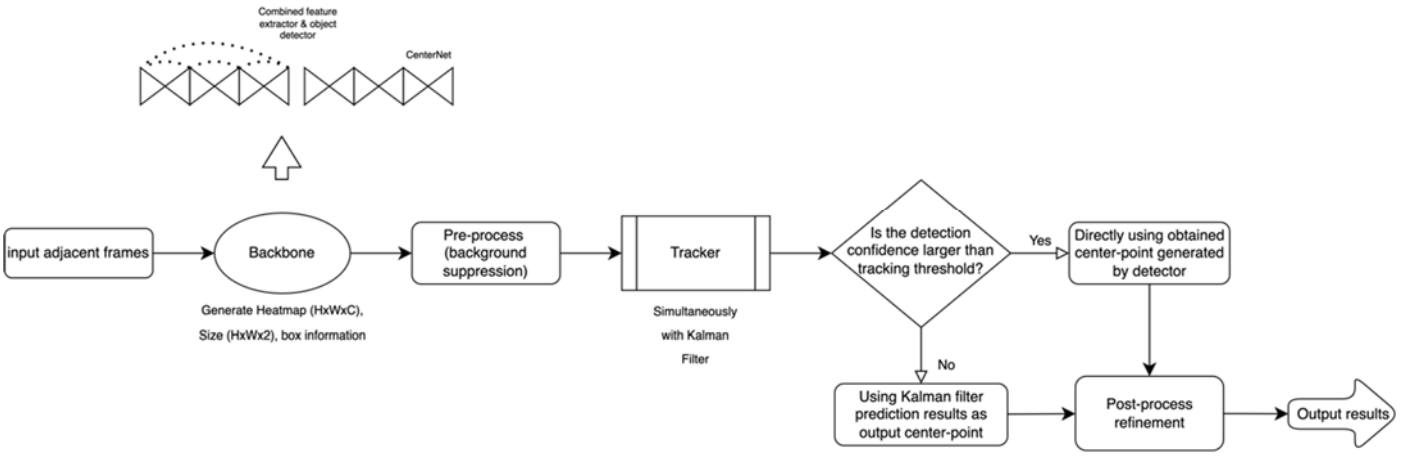


Fig. 1 The overall pipeline of the proposed PBJDT.

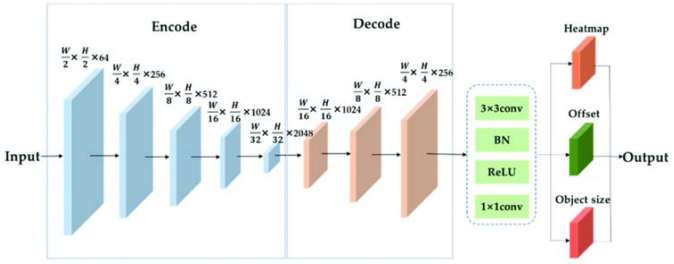


Fig. 2 The proposed hourglass network architecture.

1. Multi-scale feature fusion strategies
2. Background suppression on feature heatmaps
3. Kalman filter-aided tracking

The synergy of these components enables our method to maintain high performance in complex and dynamic scenes, particularly those encountered in urban driving environments.

A. Backbone and Detection Network

Our backbone network employs a sophisticated architecture designed for robust feature extraction and efficient object detection. The network structure utilizes a 6x ResNet-50 stacked hourglass network, incorporating skip-connections and multi-scale feature fusion to enhance the robustness of our extracted features.

The hourglass architecture is specifically designed to capture and consolidate information across multiple scales, which is crucial for accurate object localization and tracking. Each hourglass module consists of a downsampling path, an upsampling path, along with bottleneck and skip connections. This design enables our backbone to identify target keypoints (center-points) under different resolutions through consecutive down-sampling and up-sampling processes. The obtained keypoints are then propagated through skip-layers in each layer, and globally combined to synthesize our final center-point heatmap.

For feature fusion, we employ a Feature Pyramid Network (FPN) structure to fuse features at different scales, analogous to the SIFT [12] concept. The skip-connections play a vital role in combining low-level spatial details with high-level semantic information, further enhancing the network's ability to detect objects across various scales and complexities.

The output of our network generates a center-point heatmap of target objects with 1/4 resolution of the input image. This resolution maintains a balance between computational efficiency and detection precision. Additionally, the network outputs a series of box information including center-point coordinates, object length and width, detection confidence, and object IDs. Thus, we could obtain offset through computing adjacent frames' box information.

The proposed loss function includes:

1. Heatmap Loss: It uses the Focal loss (a modified cross-entropy loss for imbalanced data) for optimizing the center-point prediction, aiming for more precise detection. Our heatmap focal loss formulation:

$$L_{hm} = \begin{cases} -\frac{1}{N}\sum\left(1-p_{ij}\right)^{\alpha}\log\left(p_{ij}\right), & if\ y_{ij}=1 \\ -\frac{1}{N}\sum\left(1-y_{ij}\right)^{\beta} p_{ij}^{\alpha}\log\left(1-p_{ij}\right), & otherwise \end{cases} \quad (1)$$

where p_{ij} is the prediction heatmap response, y_{ij} is the ground truth, and both α, β are the hyper-parameter of the focal loss. The key design consideration is to address the class imbalance in heatmap prediction.

2. Offset Loss: It uses the L_1 loss to optimize the precise location of center-points.

$$L_{off} = 1/N * \sum\left|\hat{O}_k - O_k\right| \quad (2)$$

where $\hat{O}_k$ is the predicted offset and O_k is the ground truth offset.

3. Size Loss: It uses the L_1 loss to optimize the width and height of bounding boxes.

$$L_{size} = 1/N * \sum|\hat{S}_k - S_k| \quad (3)$$

where $\hat{S}_k$ is the predicted size and S_k is the ground truth size.

4. Tracking Offset Loss: It also uses the L_1 loss to optimize object position changes between consecutive adjacent frames.

$$L_{track} = 1/N * \sum|\hat{t}_k - t_k| \quad (4)$$

where $\hat{t}_k$ is the predicted tracking offset and t_k is the ground truth tracking offset.

We all choose the L_1 loss to optimize above three regression tasks due to the property of more robust to outliers. Therefore, our total loss is a weighted sum of these components:

$$L_{total} = \lambda_{hm}L_{hm} + \lambda_{off}L_{off} + \lambda_{size}L_{size} + \lambda_{track}L_{track} \quad (5)$$

where λ_{hm}, λ_{off}, λ_{size}, λ_{track} are weight coefficients to balance the contribution of different loss terms. We combine detection and tracking optimization along with careful tuning of weight coefficients for optimal performance.

B. Background Suppression

To mitigate some misjudgments caused by background interference, we implement a background suppression method on the center-point heatmap. This process involves:

1. Applying a sigmoid activation function to the heatmap to enhance response contrast, pushing background responses towards zero.
2. Implementing a confidence threshold:
   - If the detection confidence score exceeds the threshold, we consider it a target object.
   - Otherwise, we reference adjacent frames' center-point heatmaps to determine if the detection is a false positive or true positive.
3. Multiplying the heatmap with a sigmoid function kernel:
   - This pushes the heatmap response in regions adjacent to target objects towards 1 (with the center-point exactly at 1).
   - Simultaneously, it suppresses background regions towards zero.

It's worth noting that our method's effectiveness may be slightly limited when applied to driving recording data, where the background exhibits relative motion. In such cases, while the background response may not approach zero entirely, our method still maintains a significant suppression effect.

C. Kalman Filter Module

To address challenges posed by rapid illumination changes and occlusion scenarios that are difficult even for human vision to classify, we incorporate a Kalman filter module. This module is specifically designed to track object trajectories in frames where objects are overexposed due to passing trees, entering tunnels, or similar situations.

We apply the Kalman filter [13] to track our center-point coordinates in each frame while our designed detector operates simultaneously, anytime if our detection confidence is lower than tracking threshold, Kalman filter could get involved in our final output center-point. This parallel processing allows for robust tracking even when detection becomes unreliable. When the detection confidence falls below a predetermined tracking threshold, the Kalman filter's predictions are incorporated into our final output center-point. This adaptive approach ensures optimal use of both detection and tracking information.

Moreover, we assume our whole system is linear system given that our input video is processed at 15 frames per second (fps), the center-point trajectory between adjacent frames can be approximated as linear. This assumption simplifies our model and allows for efficient computation. Based on our linear system assumption, we employ a Linear Kalman Filter rather than more complex nonlinear variants like the Extended Kalman Filter (EKF) [14] or Unscented Kalman Filter (UKF) [15], balancing computational efficiency with tracking accuracy with this choice.

Our Kalman filter state vector includes the center-point coordinates and their velocities. The measurement model directly relates to the detected center-point coordinates when available. For update mechanism, in frames where detection confidence is high, we use the detector's output to update the Kalman filter's state. When detection is unreliable, we rely on the Kalman filter's predictions to maintain tracking continuity.

D. Post Process Refinement

To tackle occlusion challenges in complex scenes, we've implemented an enhanced frame referencing strategy. Unlike conventional methods that typically consider only three adjacent frames before and after the current frame, our approach expands this window to five frames in each direction. This extended temporal context allows us to extract more comprehensive reference data for our tracking process.

By inputting box information from a wider range of adjacent frames, we improve our Kalman filter module's accuracy in estimating center-point trajectories. This expanded temporal window enables our system to better capture long-range dependencies and infer object positions even during temporary occlusions. It significantly enhances the performance in challenging environments characterized by frequent occlusions and complex object interactions. It allows for more robust and consistent multiple object tracking in

scenarios where traditional methods might falter due to limited temporal context.

Furthermore, in some scenes where the detection confidence is just slightly above the tracking threshold, there may be some misjudgments. Therefore, we use spatial continuity to check if the objects' motion is reasonable. Typically, objects do not exhibit large movements in adjacent frames. By examining the movement of each object, we can determine whether there may be a false alarm.

IV. Experiments and Results

A. Datasets and Metrics

Datasets. The KITTI tracking dataset [16] is used in our experiments. The KITTI dataset is a comprehensive benchmark for various computer vision tasks, including tracking, and contains annotated images of urban scenes captured from a moving vehicle. This dataset is essential for evaluating object detection and tracking in real-world driving scenarios.

Metrics. Our evaluation metric is Multiple Object Tracking Accuracy (MOTA) [17], which provides a comprehensive measure of tracking performance across each dataset. MOTA encapsulates three key aspects of tracking: detection accuracy, false alarms, and identity preservation. The MOTA [23] metric is defined by the following equation:

$$MOTA = 1 - \frac{\sum_t FN_t + FP_t + IDSW_t}{\sum_t GT_t} \quad (6)$$

where FN_t (False Negatives) represents the number of ground truth objects that were not detected, FP_t (False Positives) represents the number of detected objects that do not correspond to any ground truth object, $IDSW_t$ (ID Switches) represents the number of times that a tracked object changes its identity label, and GT_t represents ground truth, the total number of ground truth objects across all frames.

B. Results

The performance of the proposed and other MOT methods on is shown in Table II, which show that the proposed PBJDT outperformed the state-of-the-art (SOTA) method UCMCTrack [18] with 2.1% improvements in MOTA. This improvement is largely attributed to our background suppression technique, designed hourglass network and adaptive Kalman filter, which together enhance the system's robustness in complex scenarios.

Fig. 3 shows the visualization results that PBJDT can handle illumination changes problem rather than previous similar framework CenterTrack [1].

TABLE II. The Performance on KITTI for Multiple Object Tracking

Methods (Year)	**MOTA**
3DT (2017) [19]	84.52
MOTS (2019) [20]	84.83
JRMOT (2020) [21]	85.70
CenterTrack (2020) [6]	89.44
SRK ODESA (2020) [22]	90.03
QD-3DT (2022) [23]	86.41
EagerMOT (2021) [24]	87.82
DEFT (2021) [25]	88.95
OC-SORT (2023) [26]	90.3
UCMCTrack (2024) [18]	90.4
Proposed PBJDT	**92.51**

V. Conclusions

In this work, we introduce PBJDT, a point-based joint detection-and-tracking MOT framework. It features a specially designed hourglass network backbone, a background suppression technique, and an adaptive Kalman filter, significantly improving tracking performance in scenarios with rapidly changing illumination and occlusion.

References

[1] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *European Conference on Computer Vision*, pp. 474-490, 2020.

[2] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Computer Vision*, vol. 129, pp. 3069-3087, 2021.

[3] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *IEEE Int. Conf. Image Processing*, pp. 3464-3468, 2016.

[4] T. Meinhardt, A. Kirillov, L. Leal-Taixé, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 8844-8854, 2021.

[5] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-object tracking by associating every detection box," in *European Conf. Computer Vision*, pp. 1-21, 2022.

[6] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-end multiple-object tracking with transformer," in *European Conf. Computer Vision*, pp. 659-675, 2022.

(a) The results of CenterTrack



(b) The results of the Proposed PBJDT



Fig. 3 Key frame comparison with CenterTrack [1]. The upper three frames show the results of CenterTrack, while the lower three frames show the results of the proposed PBJDT. It can be seen that PBJDT effectively handles illumination changes and tracks the target object with high precision.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137-1149, June 2016.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conf. Computer Vision*, pp. 21-37, 2016.

[10] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 6569-6578, 2019.

[11] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "Transtrack: Multiple-object tracking with transformer," *arXiv preprint arXiv:2012.15460*, 2020.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, issue 2, pp. 91-110, 2004.

[13] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Engineering*, vol. 82, no. 1, pp. 35-45, 1960.

[14] A. Gelb, *Applied Optimal Estimation*, MIT Press, 1974.

[15] S. J. Julier and J. K. Uhlmann, "New extension of the Kalman filter to nonlinear systems," in *Int. Symp. Aerospace/Defense Sensing, Simulation, and Controls*, vol. 3068, pp. 182-193, 1997.

[16] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 3354-3361, 2012.

[17] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image and Video Processing*, vol. 2008, article 246309, 2008.

[18] K. Yi, K. Luo, X. Luo, J. Huang, H. Wu, R. Hu, and W. Hao, "Ucmctrack: Multi-object tracking with uniform camera motion compensation," in *AAAI Conf. Artificial Intelligence*, vol. 38, no. 7, pp. 6702-6710, 2024.

[19] W. Choi and S. Savarese, "Multiple target tracking in world coordinate with single, minimally calibrated camera," in *European Conf. Computer Vision*, pp. 553-569, 2017.

[20] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "MOTS: Multi-object tracking and segmentation," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 7942-7951, 2019.

[21] A. Shenoi, M. Patel, J. Y. Gwak, P. Goebel, A. Sadeghian, H. Rezatofighi, R. Martın-Martın, and S. Savarese, "JRMOT: A real-time 3d multi-object tracker and a new large-scale dataset," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pp. 10335-10342, 2020.

[22] D. Mykheievskyi, D. Borysenko, and V. Porokhonskyy, "Learning local feature descriptors for multiple object tracking," in *Asian Conf. Computer Vision*, pp. 1-18, 2020.

[23] H. N. Hu, Y. H. Yang, T. Fischer, T. Darrell, F. Yu, and M. Sun, "Monocular quasi-dense 3D object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, issue 2, pp. 1992-2008, 2022.

[24] A. Kim, A. Osep, and L. Leal-Taixe, "EagerMOT: 3D Multi-object tracking via sensor fusion," in *IEEE Int. Conf. Robotics and Automation*, pp. 11315-11321, 2021.

[25] M. Chaabane, P. Zhang, J. R. Beveridge, and S. O'Hara, "DEFT: Detection embeddings for tracking," *arXiv preprint arXiv: 2102.02267*, 2021.

[26] J. Cao, J. Pang, X. Weng, R. Khirodkar and K. Kitani. "Observation-centric sort: Rethinking sort for robust multi-object tracking," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 9686-9696, 2023.

Auxiliary-Function-Based Steering Vector Estimation Method for Spatially Regularized Independent Low-Rank Matrix Analysis

Sota Hirata*, Norihiro Takamune*, Kouei Yamaoka*, Daichi Kitamura†,
Hiroshi Saruwatari*, Yu Takahashi‡, and Kazunobu Kondo‡
* The University of Tokyo, Tokyo, Japan
† National Institute of Technology, Kagawa College, Kagawa, Japan
‡ Yamaha Corporation, Shizuoka, Japan

***Abstract*—Blind source separation (BSS) is a technique to separate each source signal from observed mixtures without prior information, and independent low-rank matrix analysis (ILRMA) is one of the state-of-the-art BSS methods. As not a fully blind method, spatially regularized ILRMA (SR-ILRMA) achieves higher separation performance by using prior information about the demixing matrix or steering vectors (SVs) as a regularizer in ILRMA. In this paper, we propose a new fully blind method that models SVs using the time differences of arrival (TDOAs) and simultaneously estimates TDOAs, the demixing matrix, and other parameters. Then, we derive the update rule of TDOAs on the basis of the majorization-minimization algorithm, which guarantees the monotonic non-increase in the proposed cost function. We also propose the method of initializing TDOAs based on impulse responses for the proposed cost function. Numerical experiments confirmed that the proposed method achieves better separation performance than ILRMA, and initialization based on impulse responses is effective for the proposed method.**

I. Introduction

Blind source separation (BSS) [1] is a technique to separate each source signal from mixtures observed by a microphone array without any information about the transmission system or the characteristics of the source. BSS is an essential preprocessing technique for acoustic signal processing in real-world applications, such as speech recognition and hearing aids. Representative methods of BSS include frequency-domain independent component analysis (FDICA) [2], independent vector analysis (IVA) [3], [4], and independent low-rank matrix analysis (ILRMA) [5]. FDICA estimates the demixing matrix by assuming statistical independence between sources for each frequency bin. However, the outputs in each frequency bin are unordered, and thus, FDICA should align outputs over all frequency bins (permutation problem). IVA and ILRMA simultaneously estimate the demixing matrix and solve the permutation problem using the generative model that has the higher-order correlation between frequency bins in each source. In particular, ILRMA assumes a more sophisticated model than IVA and experimentally achieves a high separation performance [5]. However, it has been reported that there is still room for improvement in the accuracy of solving the permutation problem even in ILRMA [6].

In contrast to fully blind methods, several methods use spatial prior information about the source directions and the microphone array geometry to improve the separation performance [7]–[10]. This spatial prior information often includes some errors due to measurement errors in the directions of the sources and the position of the microphone array, microphone directivity, diffraction around the microphone array, and reverberation, among others. Therefore, in [7]–[10], such prior information is utilized as a regularizer to tolerate some errors. Spatially regularized ILRMA (SR-ILRMA) [9] applies such regularizer to ILRMA and it has been reported that SR-ILRMA achieves higher separation performance than conventional ILRMA.

In terms of applicability, a fully blind method is desirable. In this paper, we propose a new fully blind method that estimates the time differences of arrival (TDOAs), which implicitly retain information about the source directions and the microphone array geometry, simultaneously with the demixing matrix. In [7], the demixing matrix and source directions were simultaneously estimated using different criteria for each. In contrast, we aim to formulate this simultaneous estimation problem as a single minimization problem. Then, for the majorization-minimization (MM) algorithm [11], we design the auxiliary function of the proposed cost function with respect to TDOAs and derive the update rule that guarantees the monotonic non-increase in the cost function. We also investigated the method of initializing TDOAs for the proposed method.

First, we investigated the initial value dependence of TDOAs in the preliminary experiment. Then, we conducted a numerical experiment to confirm the effectiveness of the proposed method.

II. Conventional methods

A. MM algorithm [11]

The MM algorithm is an optimization method used to find $\boldsymbol{s}$ that minimizes the cost function $f(\boldsymbol{s})$. For a given cost function $f(\boldsymbol{s})$, the auxiliary function $f^{+}(\boldsymbol{s}, \tilde{\boldsymbol{s}})$ is designed to satisfy the

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 (for algorithm design) and Tateisi Science and Technology Foundation (for numerical experiment).

following conditions:

$$f(\boldsymbol{s}) \leq f^{+}(\boldsymbol{s}, \tilde{\boldsymbol{s}}), \quad \forall \boldsymbol{s}, \tilde{\boldsymbol{s}}, \tag{1}$$

$$f(\tilde{\boldsymbol{s}}) = f^{+}(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{s}}), \quad \forall \tilde{\boldsymbol{s}}, \tag{2}$$

where $\tilde{\boldsymbol{s}}$ is an auxiliary variable. When the auxiliary function $f^{+}(\boldsymbol{s}, \tilde{\boldsymbol{s}})$ exists, the update rules for the MM algorithm are

$$\hat{\boldsymbol{s}} \leftarrow \boldsymbol{s}, \tag{3}$$

$$\boldsymbol{s} \leftarrow \underset{\boldsymbol{s}}{\operatorname{argmin}}\ f^{+}(\boldsymbol{s}, \hat{\boldsymbol{s}}). \tag{4}$$

The cost function $f(\boldsymbol{s})$ is minimized by the iterative update of (3) and (4) from a given initial value of $\boldsymbol{s}$.

The MM algorithm guarantees the monotonic non-increase in the cost function as

$$f(\boldsymbol{s}^{(l)}) = f^{+}(\boldsymbol{s}^{(l)}, \boldsymbol{s}^{(l)}) \geq f^{+}(\boldsymbol{s}^{(l+1)}, \boldsymbol{s}^{(l)}) \geq f(\boldsymbol{s}^{(l+1)}), \tag{5}$$

where the superscript $^{(l)}$ denotes the parameter updated in the lth iteration. Unlike other gradient-based optimization methods, the MM algorithm does not require step size tuning. The most crucial aspect of the MM algorithm is designing an auxiliary function for each cost function of the optimization problem. Auxiliary functions are not unique for each cost function; thus, it is important to find a better auxiliary function for achieving faster convergence.

B. ILRMA [5]

Let $\boldsymbol{x}_{ij} = (x_{ij1}, ..., x_{ijM})^{\mathsf{T}} \in \mathbb{C}^{M}$, $\boldsymbol{s}_{ij} = (s_{ij1}, ..., s_{ijN})^{\mathsf{T}} \in \mathbb{C}^{N}$, and $\boldsymbol{y}_{ij} = (y_{ij1}, ..., y_{ijN})^{\mathsf{T}} \in \mathbb{C}^{N}$ be the short-time Fourier transforms (STFTs) of the observed, source, and separated signals, respectively. Here, $i \in \{1, ..., I\}$, $j \in \{1, ..., J\}$, $m \in \{1, ..., M\}$, and $n \in \{1, ..., N\}$ are the indices of the frequency bins, time frames, microphones, and sources, respectively, and $\cdot^{\mathsf{T}}$ denotes the transpose. If each source is a point source and the reverberation time is sufficiently shorter than the window length of the STFT, the observed signal is approximately represented as $\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}$. Here, $\boldsymbol{A}_i = (\boldsymbol{a}_{i1}, ...\boldsymbol{a}_{iN}) \in \mathbb{C}^{M \times N}$ is the mixing matrix representing the time-invariant spatial characteristics of the transmission system, and $\boldsymbol{a}_{in}$ is the steering vector (SV) of the nth source. If $M = N$ and $\boldsymbol{A}_i$ is regular, the separated signals can be obtained as

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}, \tag{6}$$

where $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, ..., \boldsymbol{w}_{iN})^{\mathsf{H}} = \boldsymbol{A}_i^{-1} \in \mathbb{C}^{N \times M}$ is the demixing matrix and $\cdot^{\mathsf{H}}$ denotes the Hermitian transpose.

ILRMA assumes that the separated signal y_{ijn} follows a univariate complex Gaussian distribution whose mean and time-variant variance are zero and r_{ijn}, respectively. It is also assumed that the time-variant variance r_{ijn} has low rankness and is modeled by nonnegative matrix factorization (NMF) [12] as $r_{ijn} = \sum_k t_{ikn} v_{kjn}$, where $t_{ikn} \geq 0$ and $v_{kjn} \geq 0$ are the NMF variables, and $k \in \{1, ..., K\}$ is the index of the NMF basis. The cost function of ILRMA $\mathcal{J}_{\text{ILRMA}}$ is given as the following negative log-likelihood of the observed signals:

$$\mathcal{J}_{\text{ILRMA}} = \sum_{i,j,n} \left(\frac{|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{r_{ijn}} + \log r_{ijn} \right) - J \sum_i \log |\det \boldsymbol{W}_i|^2 + \text{const.}, \tag{7}$$

where const. is the term that does not include $\boldsymbol{w}_{in}$, t_{ikn}, or v_{kjn}. The cost function $\mathcal{J}_{\text{ILRMA}}$ is minimized by alternately updating $\boldsymbol{w}_{in}$, t_{ikn}, and v_{kjn}. In [5], the update rules for the NMF variables t_{ikn} and v_{kjn} are derived on the basis of the MM algorithm. For the update of the demixing matrix $\boldsymbol{W}_i$, iterative projection (IP) [13] is applied. IP also guarantees the monotonic non-increase in the cost function $\mathcal{J}_{\text{ILRMA}}$. Therefore, the total update rules for $\boldsymbol{w}_{in}$, t_{ikn}, and v_{kjn} also guarantee the monotonic non-increase in the cost function $\mathcal{J}_{\text{ILRMA}}$.

C. SR-ILRMA [9]

To improve the separation performance of ILRMA, SR-ILRMA uses prior information about the source directions and the microphone array geometry as a regularizer in ILRMA. Note that SR-ILRMA is not a fully blind method.

There are two types of regularizer in conventional SR-ILRMA [9], [14]. In [9], the regularizer $\mathcal{R}_1$ is defined as

$$\mathcal{R}_1 = J \sum_{i,n} \mu_{in}^{(\text{SR1})} \|\boldsymbol{w}_{in} - \hat{\boldsymbol{w}}_{in}\|_2^2, \tag{8}$$

where $\mu_{in}^{(\text{SR1})}$ is the weight of the regularizer and $\hat{\boldsymbol{w}}_{in} \in \mathbb{C}^{M}$ is the prior demixing filter corresponding to $\boldsymbol{w}_{in}$. On the other hand, in [14], the regularizer $\mathcal{R}_2$ is defined as

$$\mathcal{R}_2 = J \sum_{i,n} \mu_{in}^{(\text{SR2})} |\boldsymbol{w}_{in}^{\mathsf{H}} \hat{\boldsymbol{a}}_{in^{(t)}} - \delta_{nn^{(t)}}|^2, \tag{9}$$

where $\mu_{in}^{(\text{SR2})}$ is the weight of the regularizer, and $\hat{\boldsymbol{a}}_{in^{(t)}} \in \mathbb{C}^{M}$ is the prior SV of the target source. Here, $n^{(t)}$ is the index of the target source and $\delta_{nn^{(t)}}$ is the Kronecker delta.

For the convenience of the proposed method, unlike (8) and (9), we define the regularizer $\mathcal{R}$ as

$$\mathcal{R} = J \sum_{i,n,n'} \mu_{inn'} |\boldsymbol{w}_{in}^{\mathsf{H}} \hat{\boldsymbol{a}}_{in'} - \delta_{nn'}|^2, \tag{10}$$

where $\mu_{inn'}$ is the weight of the regularizer, $\hat{\boldsymbol{A}}_i = (\hat{\boldsymbol{a}}_{i1}, ..., \hat{\boldsymbol{a}}_{iN}) \in \mathbb{C}^{M \times N}$ is the prior mixing matrix corresponding to $\boldsymbol{A}_i$, and $\hat{\boldsymbol{a}}_{in}$ is the prior SV of the nth source. Note that the regularizer itself is the same as that proposed in [8], [15]. In this case, the cost function $\mathcal{J}_{\text{SR}}$ is defined as

$$\mathcal{J}_{\text{SR}} = \mathcal{J}_{\text{ILRMA}} + \mathcal{R}. \tag{11}$$

In SR-ILRMA, the cost function $\mathcal{J}_{\text{SR}}$ is minimized by alternately updating $\boldsymbol{w}_{in}$, t_{ikn}, and v_{kjn}. Since the regularizer $\mathcal{R}$ does not include t_{ikn} or v_{kjn}, the update rules for t_{ikn} and v_{kjn} can be derived in the same manner as those in ILRMA. For the update of $\boldsymbol{W}_i$, vectorwise coordinate descent (VCD) [16] can be applied instead of IP because $\mathcal{J}_{\text{SR}}$ consists of a quadratic form, the logarithm of the determinant, and a linear term of

$\boldsymbol{w}_{in}$. Therefore, the update rules for $\boldsymbol{w}_{in}$, t_{ikn}, and v_{kjn} in SR-ILRMA are as follows:

$$t_{ikn} \leftarrow t_{ikn}\sqrt{\frac{\sum_j v_{kjn}|\boldsymbol{w}_{in}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2/(\sum_{k'} t_{ik'n}v_{k'jn})^2}{\sum_j(v_{kjn}/\sum_{k'} t_{ik'n}v_{k'jn})}}, \quad (12)$$

$$v_{kjn} \leftarrow v_{kjn}\sqrt{\frac{\sum_i t_{ikn}|\boldsymbol{w}_{in}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2/(\sum_{k'} t_{ik'n}v_{k'jn})^2}{\sum_i(t_{ikn}/\sum_{k'} t_{ik'n}v_{k'jn})}}, \quad (13)$$

$$\hat{\boldsymbol{D}}_{in} \leftarrow \frac{1}{J}\sum_j \frac{\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathsf{H}}}{\sum_k t_{ikn}v_{kjn}} + \sum_{n'}\mu_{inn'}\hat{\boldsymbol{a}}_{in'}\hat{\boldsymbol{a}}_{in'}^{\mathsf{H}}, \quad (14)$$

$$\boldsymbol{u}_{in} \leftarrow (\boldsymbol{W}_i\hat{\boldsymbol{D}}_{in})^{-1}\hat{\boldsymbol{a}}_{in}, \quad (15)$$

$$\hat{\boldsymbol{u}}_{in} \leftarrow \mu_{inn}\hat{\boldsymbol{D}}_{in}^{-1}\boldsymbol{u}_{in}, \quad (16)$$

$$h_{in} \leftarrow \boldsymbol{u}_{in}^{\mathsf{H}}\hat{\boldsymbol{D}}_{in}\boldsymbol{u}_{in}, \quad (17)$$

$$\hat{h}_{in} \leftarrow \boldsymbol{u}_{in}^{\mathsf{H}}\hat{\boldsymbol{D}}_{in}\hat{\boldsymbol{u}}_{in}, \quad (18)$$

$$\chi_{in} \leftarrow \begin{cases} \dfrac{1}{\sqrt{h_{in}}}, & (\text{if } \hat{h}_{in}=0) \\ \dfrac{\hat{h}_{in}}{2h_{in}}\left[\sqrt{1+\dfrac{4h_{in}}{|\hat{h}_{in}|^2}}-1\right], & (\text{otherwise}) \end{cases} \quad (19)$$

$$\boldsymbol{w}_{in} \leftarrow \chi\boldsymbol{u}_{in} + \hat{\boldsymbol{u}}_{in}. \quad (20)$$

VCD also guarantees the monotonic non-increase in the cost function $\mathcal{J}_{\text{SR}}$. Therefore, the total update rules for $\boldsymbol{w}_{in}$, t_{ikn}, and v_{kjn} also guarantee the monotonic non-increase in the cost function $\mathcal{J}_{\text{SR}}$.

III. Proposed method

A. Motivation

SR-ILRMA requires prior information about the source directions and the microphone array geometry to precalculate $\hat{\boldsymbol{w}}_{in}$ or $\hat{\boldsymbol{a}}_{in}$, i.e., SR-ILRMA is not a fully blind method. However, a fully blind method is still preferable because of its broad range of applications. In this paper, we propose a new fully blind method to estimate SV $\hat{\boldsymbol{a}}_{in}$ simultaneously with the demixing matrix and the NMF variables by minimizing the cost function $\mathcal{J}_{\text{SR}}$.

If we directly optimize (11) with respect to $\hat{\boldsymbol{a}}_{in'}$ considering that the current $\boldsymbol{W}_i$, t_{ikn}, and v_{kjn} are fixed, over-fitting probably occurs. Therefore, we need to introduce a small number of other parameters to model the SV. Then, we use TDOAs, which implicitly have information about the source directions and the microphone array geometry, as the parameters for SV. To guarantee the monotonic non-increase in the proposed cost function during the simultaneous estimation of TDOAs and other parameters, we derive the update rules on the basis of the MM algorithm. We also investigate the method of initializing TDOAs for the proposed method.

B. Modeling

Assuming a plane wave arrival model, $\hat{\boldsymbol{a}}_{in}$ can be represented as

$$\hat{\boldsymbol{a}}_{in}(\boldsymbol{\tau}_n) = (e^{-\mathrm{j}\omega_i\tau_{1n}}, \cdots, e^{-\mathrm{j}\omega_i\tau_{Mn}})^{\mathsf{T}}, \quad (21)$$

where $\boldsymbol{\tau}_n = (\tau_{1n}, \cdots, \tau_{Mn})^{\mathsf{T}} \in \mathbb{R}^M$, j is the imaginary unit, and ω_i is the normalized angular frequency corresponding to the frequency bin i. Here, τ_{mn} denotes the TDOA.

Introducing (21) into (11), we define the proposed cost function $\mathcal{J}_{\text{proposed}}$ as

$$\mathcal{J}_{\text{proposed}} = \mathcal{J}_{\text{ILRMA}} + \sum_{i,n,n'}\mu_{inn'}\left|\boldsymbol{w}_{in}^{\mathsf{H}}\hat{\boldsymbol{a}}_{in'}(\boldsymbol{\tau}_{n'}) - \delta_{nn'}\right|^2. \quad (22)$$

We consider optimizing (22) by alternating iterative updates of $\boldsymbol{w}_{in}$, t_{ikn}, v_{kjn}, and $\boldsymbol{\tau}_{n'}$. Since $\boldsymbol{\tau}_{n'}$ is fixed during updating $\boldsymbol{w}_{in}$, t_{ikn}, and v_{kjn}, the update rules for $\boldsymbol{w}_{in}$, t_{ikn}, and v_{kjn} are the same as those in SR-ILRMA. Therefore, we consider the update rules for $\boldsymbol{\tau}_{n'}$ hereafter.

C. Derivation of update rules for proposed method

$\boldsymbol{\tau}_{n'}$ is only involved in the regularizer of (22); thus, we consider minimizing the following cost function $\mathcal{J}$:

$$\mathcal{J} = \sum_{i,n,n'}\mu_{inn'}\left|\boldsymbol{w}_{in}^{\mathsf{H}}\hat{\boldsymbol{a}}_{in'}(\boldsymbol{\tau}_{n'}) - \delta_{nn'}\right|^2. \quad (23)$$

We substitute (21) into (23) and rearrange the formula as

$$\begin{aligned}\mathcal{J} = \sum_{i,n,n'}\mu_{inn'}\Bigg\{&\sum_{m,m'} w_{inm}^* e^{-\mathrm{j}\omega_i\tau_{mn'}} w_{inm'} e^{\mathrm{j}\omega_i\tau_{m'n'}} \\ &- \delta_{nn'}\sum_m (w_{inm}^* e^{-\mathrm{j}\omega_i\tau_{mn'}} \\ &+ w_{inm}e^{\mathrm{j}\omega_i\tau_{mn'}}) + \delta_{nn'}\Bigg\}, \quad (24)\end{aligned}$$

$$\begin{aligned}&= 2\sum_{i,n,n'}\mu_{inn'}\left(\sum_{m<m'}|w_{inm}||w_{inm'}|\cos\theta_{inn'mm'}\right) \\ &\quad + 2\sum_{i,n'}\mu_{in'n'}\left(\sum_m |w_{in'm}|\cos\phi_{in'm}\right) + \text{const.}, \quad (25)\end{aligned}$$

where $\cdot^*$ denotes the complex conjugation and const. is the term that does not include $\boldsymbol{\tau}_{n'}$. Here, $\theta_{inn'mm'}$ and $\phi_{in'm}$ are respectively defined as

$$\theta_{inn'mm'} = \omega_i\tau_{mn'} - \omega_i\tau_{m'n'} + \angle w_{inm} - \angle w_{inm'}, \quad (26)$$

$$\phi_{in'm} = \omega_i\tau_{mn'} + \angle w_{in'm} + \pi, \quad (27)$$

where $\angle\cdot$ denotes the argument of a complex number.

Since (25) is the sum of cosine functions and the coefficients of the cosine functions are always positive, we can use an auxiliary function for the cosine function to design the auxiliary function for the cost function $\mathcal{J}$. For any ϑ and $\tilde{\vartheta}$, the following inequality [17] is satisfied:

$$\begin{aligned}\cos\vartheta \le \frac{1}{2}\mathrm{sinc}\big(g(\tilde{\vartheta})\big)\big(\vartheta - \tilde{\vartheta} + g(\tilde{\vartheta})\big)^2 \\ + \cos\tilde{\vartheta} + \frac{1}{2}g(\tilde{\vartheta})\sin\tilde{\vartheta}, \quad (28)\end{aligned}$$

where $\mathrm{sinc}(\varphi) = \sin(\varphi)/\varphi$ denotes the unnormalized sinc function and $g(\cdot) = \mathrm{mod}_{2\pi}(\cdot) - \pi$. Here, $\mathrm{mod}_{2\pi}(\cdot)$ denotes the modulo operation with 2π. The equality in (28) holds when

$\vartheta = \tilde{\vartheta}$, and the right-hand side of (28) serves as an auxiliary function for $\cos\vartheta$ with the auxiliary variable $\tilde{\vartheta}$.

Using (28), we can design the auxiliary function $\mathcal{J}^+$ for $\mathcal{J}$ as

$$\begin{aligned}\mathcal{J}^+ = &\sum_{i,n,n'} \mu_{inn'} \left\{ \sum_{m<m'} \alpha_{inn'mm'} \Big(\omega_i \Delta\tau_{mn'} \right. \\ &\left. -\omega_i \Delta\tau_{m'n'} + g(\tilde{\theta}_{inn'mm'}) \Big)^2 \right\} \\ &+ \sum_{i,n'} \mu_{in'n'} \left\{ \sum_m \beta_{in'm} \Big(\omega_i \Delta\tau_{mn'} + g(\tilde{\phi}_{in'm}) \Big)^2 \right\} \\ &+ \text{const.}, \end{aligned} \tag{29}$$

where const. is the term that does not include $\boldsymbol{\tau}_{n'}$, and $\alpha_{inn'mm'}$, $\beta_{in'm}$, and $\Delta\tau_{mn'}$ are respectively defined as

$$\alpha_{inn'mm'} = |w_{inm}||w_{inm'}|\mathrm{sinc}(g(\tilde{\theta}_{inn'mm'})), \tag{30}$$

$$\beta_{in'm} = |w_{in'm}|\mathrm{sinc}(g(\tilde{\phi}_{in'm})), \tag{31}$$

$$\Delta\tau_{mn'} = \tau_{mn'} - \tilde{\tau}_{mn'}, \tag{32}$$

where $\{\tilde{\tau}_{mn'}\}$ is the set of auxiliary variables, and $\tilde{\theta}_{inn'mm'}$ and $\tilde{\phi}_{in'm}$ are defined by substituting $\tau_{mn'} = \tilde{\tau}_{mn'}$ into (26) and (27). The equality in (29) holds when $\tau_{mn'} = \tilde{\tau}_{mn'}$ for all m and n'.

Since $\mathcal{J}^+$ is a quadratic function of $\Delta\tau_{mn'}$, we can deform $\mathcal{J}^+$ by using a quadratic form and linear term of $\Delta\boldsymbol{\tau}_{n'}$ as

$$\mathcal{J}^+ = \sum_{n'} \left(\Delta\boldsymbol{\tau}_{n'}^{\mathsf{T}} \boldsymbol{Q}_{n'} \Delta\boldsymbol{\tau}_{n'} - 2\Delta\boldsymbol{\tau}_{n'}^{\mathsf{T}} \boldsymbol{b}_{n'} \right) + \text{const.}, \tag{33}$$

where const. is the term that does not include $\boldsymbol{\tau}_{n'}$. Here, $\Delta\boldsymbol{\tau}_{n'}$, $\boldsymbol{Q}_{n'} \in \mathbb{R}^{M\times M}$, and $\boldsymbol{b}_{n'} \in \mathbb{R}^M$ are defined as

$$\tilde{\boldsymbol{\tau}}_{n'} = (\tilde{\tau}_{1n'}, \cdots, \tilde{\tau}_{Mn'})^{\mathsf{T}} \in \mathbb{R}^M, \tag{34}$$

$$\Delta\boldsymbol{\tau}_{n'} = \boldsymbol{\tau}_{n'} - \tilde{\boldsymbol{\tau}}_{n'}, \tag{35}$$

$$\begin{aligned}(\boldsymbol{Q}_{n'})_{mm'} &= \begin{cases} \displaystyle\sum_{i,n} \mu_{inn'} \omega_i^2 \left(\sum_{m''\neq m} \alpha_{inn'mm''} \right) & (m = m') \\ \qquad + \displaystyle\sum_i \mu_{in'n'} \omega_i^2 \beta_{in'm}, & \\ -\displaystyle\sum_{i,n} \mu_{inn'} \omega_i^2 \alpha_{inn'mm'}, & (m \neq m') \end{cases} \\ (\boldsymbol{b}_{n'})_m &= -\sum_{i,n} \mu_{inn'} \omega_i \left(\sum_{m'} \alpha_{inn'mm'} g(\tilde{\theta}_{inn'mm'}) \right) \\ &\quad - \sum_i \mu_{in'n'} \omega_i \beta_{in'm} g(\tilde{\phi}_{in'm}), \end{aligned} \tag{36, 37}$$

where $(\cdot)_{mm'}$ denotes the (m,m')th element of a matrix.

Since (33) is separable for each n', the auxiliary function $\mathcal{J}^+$ can be minimized by solving $\partial\mathcal{J}^+/\partial\boldsymbol{\tau}_{n'} = \boldsymbol{0}_M$ for all n'. Here, $\boldsymbol{0}_M \in \mathbb{R}^M$ denotes the M-dimensional zero vector. Therefore, the MM-algorithm-based update rules for $\boldsymbol{\tau}_n'$ are derived as

$$\tilde{\boldsymbol{\tau}}_{n'} \leftarrow \boldsymbol{\tau}_{n'}, \tag{38}$$

$$\boldsymbol{\tau}_{n'} \leftarrow \tilde{\boldsymbol{\tau}}_{n'} + \boldsymbol{Q}_{n'}^{\dagger} \boldsymbol{b}_{n'}, \tag{39}$$

where $\cdot^{\dagger}$ denotes the Moore–Penrose inverse. Note that we use the Moore–Penrose inverse considering if $\boldsymbol{Q}_{n'}$ may not be regular.

Since we can use the same update rules as (12)–(20) for $\boldsymbol{w}_{in}$, t_{ikn}, and v_{kjn}, the total update rules for $\boldsymbol{w}_{in}$, t_{ikn}, v_{kjn}, and $\boldsymbol{\tau}_n'$ in the proposed method are

$$t_{ikn} \leftarrow t_{ikn} \sqrt{\frac{\sum_j v_{kjn} |\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2 / (\sum_{k'} t_{ik'n} v_{k'jn})^2}{\sum_j (v_{kjn} / \sum_{k'} t_{ik'n} v_{k'jn})}}, \tag{40}$$

$$v_{kjn} \leftarrow v_{kjn} \sqrt{\frac{\sum_i t_{ikn} |\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2 / (\sum_{k'} t_{ik'n} v_{k'jn})^2}{\sum_i (t_{ikn} / \sum_{k'} t_{ik'n} v_{k'jn})}}, \tag{41}$$

$$\hat{\boldsymbol{D}}_{in} \leftarrow \frac{1}{J} \sum_j \frac{\boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathsf{H}}}{\sum_k t_{ikn} v_{kjn}} + \sum_{n'} \mu_{inn'} \hat{\boldsymbol{a}}_{in'} \hat{\boldsymbol{a}}_{in'}^{\mathsf{H}}, \tag{42}$$

$$\boldsymbol{u}_{in} \leftarrow (\boldsymbol{W}_i \hat{\boldsymbol{D}}_{in})^{-1} \hat{\boldsymbol{a}}_{in}, \tag{43}$$

$$\hat{\boldsymbol{u}}_{in} \leftarrow \mu_{inn} \hat{\boldsymbol{D}}_{in}^{-1} \boldsymbol{u}_{in}, \tag{44}$$

$$h_{in} \leftarrow \boldsymbol{u}_{in}^{\mathsf{H}} \hat{\boldsymbol{D}}_{in} \boldsymbol{u}_{in}, \tag{45}$$

$$\hat{h}_{in} \leftarrow \boldsymbol{u}_{in}^{\mathsf{H}} \hat{\boldsymbol{D}}_{in} \hat{\boldsymbol{u}}_{in}, \tag{46}$$

$$\chi_{in} \leftarrow \begin{cases} \dfrac{1}{\sqrt{h_{in}}}, & (\text{if } \hat{h}_{in} = 0) \\ \dfrac{\hat{h}_{in}}{2h_{in}} \left[\sqrt{1 + \dfrac{4h_{in}}{|\hat{h}_{in}|^2}} - 1 \right] & (\text{otherwise}) \end{cases} \tag{47}$$

$$\boldsymbol{w}_{in} \leftarrow \chi \boldsymbol{u}_{in} + \hat{\boldsymbol{u}}_{in} \tag{48}$$

$$\tilde{\boldsymbol{\tau}}_{n'} \leftarrow \boldsymbol{\tau}_{n'}, \tag{49}$$

$$\boldsymbol{\tau}_{n'} \leftarrow \tilde{\boldsymbol{\tau}}_{n'} + \boldsymbol{Q}_{n'}^{\dagger} \boldsymbol{b}_{n'}, \tag{50}$$

$$\hat{\boldsymbol{a}}_{in'} \leftarrow \hat{\boldsymbol{a}}_{in'}(\boldsymbol{\tau}_{n'}), \tag{51}$$

where $\boldsymbol{Q}_{n'}$ and $\boldsymbol{b}_{n'}$ are calculated by substituting (49) into (36) and (37), respectively. The update rules for $\boldsymbol{w}_{in}$, t_{ikn}, v_{kjn}, and $\boldsymbol{\tau}_n'$ (40)–(51) guarantee the monotonic non-increase in the cost function $\mathcal{J}_{\text{proposed}}$.

D. Initialization method based on impulse responses

The cost function $\mathcal{J}$ is the sum of cosine functions, and the cosine function has an infinite number of local optimal solutions. Thus, the proposed method seems to be highly affected by the initial values of τ_{mn}. Then, we consider the initialization method for τ_{mn}. Instead of $\mathcal{J}$, we consider $\sum_i \|\hat{\boldsymbol{A}}_i - \boldsymbol{W}_i^{-1}\|_{\mathrm{F}}^2$ for the initialization, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. On the basis of Parseval's theorem, the following equation holds:

$$\begin{aligned}&\sum_i \|\hat{\boldsymbol{A}}_i - \boldsymbol{W}_i^{-1}\|_{\mathrm{F}}^2 \\ &= \gamma \sum_{m,n} \Big\| \mathrm{IDFT}\big[\{e^{-\mathrm{j}\omega_i \tau_{mn}}\}\big] - \mathrm{IDFT}\big[\{(\boldsymbol{W}_i^{-1})_{mn}\}\big] \Big\|_2^2, \end{aligned} \tag{52}$$

where γ is the length of the STFT window and $\mathrm{IDFT}[\{z_i\}] \in \mathbb{R}^{\gamma}$ denotes the inverse discrete Fourier transform of the complex spetrum whose ith frequency bin is z_i

When τ_{mn} is discrete, the $(\tau_{mn}+1)$th element of $\mathrm{IDFT}[\{e^{-\mathrm{j}\omega_i\tau_{mn}}\}]$ becomes one and the other elements become zero. Therefore, to minimize (52), we choose the τ_{mn}

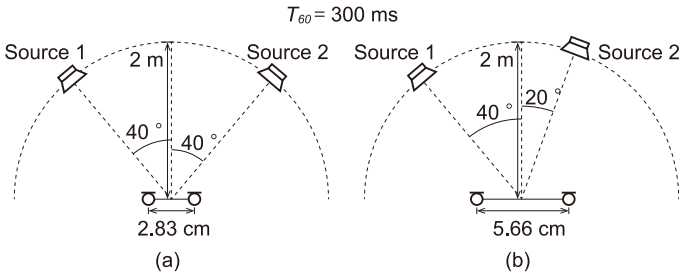


Fig. 1. Recording conditions of impulse responses.

corresponding to the index where $\mathrm{IDFT}\left[\{(\boldsymbol{W}_i^{-1})_{mn}\}\right]$, which means the impulse response (IR) calculated from $\boldsymbol{W}_i$, takes the maximum value. Then, we use this optimal τ_{mn} of the pseudo-problem as the initial value of the proposed cost function.

IV. Experiments

First, we considered a subproblem to estimate TDOAs in the situation where the demixing matrix was fixed and conducted a preliminary experiment to investigate the initial value dependence of the proposed method. Next, we evaluated the separation performance of the proposed method through a numerical experiment. Both experiments were conducted using two sources and two microphones.

A. Experimental conditions

We used three pairs of instruments (bass/drums, drums/vocal, vocal/bass) for each of the six tracks in the DSD100 dataset [18] as dry sources (total 18 pairs). Then, each dry source was convolved with two conditions of impulse responses in the RWCP database [19] (see Fig. 1). The convolved signals were mixed so that the input signal-to-noise ratio became $0\,\mathrm{dB}$. The total number of the observed signals was $18 \times 2 = 36$. The sampling rate was $16\,\mathrm{kHz}$. STFT was performed using a 256-ms-long Hamming window with a shift length of $64\,\mathrm{ms}$.

B. Initial value dependence

To investigate the initial value dependence of the proposed method, we considered the subproblem to estimate the TDOAs in the situation where the demixing matrix $\boldsymbol{W}_i$ was fixed. First, we obtained the demixing matrix by ILRMA as the preprocess. Then, the demixing matrix was fixed, and only the updates related to TDOAs were iteratively performed to observe the behavior of the cost function $\mathcal{J}$ during iteration.

In ILRMA as the preprocess, the demixing matrix was initialized with the identity matrix, and the NMF variables t_{ikn} and v_{kjn} were initialized with random numbers from a uniform distribution on $[0, 1]$. We used 10 random seeds for each observed signal, resulting in 360 demixing matrices. The number of iterations was set to 100. In the estimation of TDOAs, we compared two initialization methods: (i) random sampling and (ii) proposed initialization described in Section III-D (IR-based initialization method). In random sampling, 10 different initial values were uniformly sampled from $[-5, 5]$. The weighting coefficient $\mu_{inn'}$ for the regularizer was set as $\mu_{inn'} = 1$ $(i \neq 1, I)$ and $\mu_{inn'} = 0$ $(i = 1, I)$. The number of iterations was set to 200.

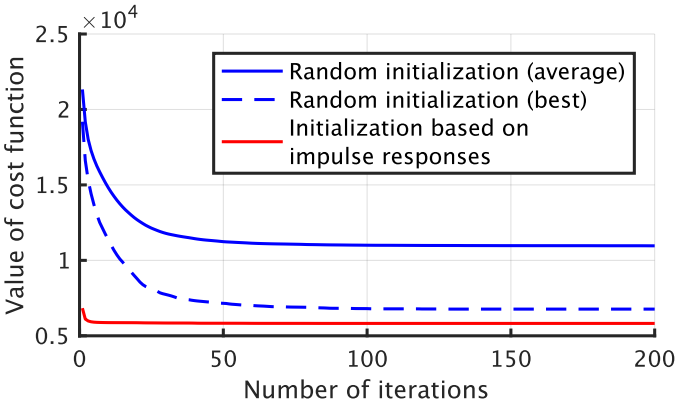


Fig. 2. Behavior of the cost function (23) for each initialization method. The average behavior for random initial values is shown by the solid blue line, the behavior for the random initial values that resulted in the smallest final value of cost function is shown by the dashed blue line, and the behavior for the initial value calculated by the IR-based initialization method is shown by the solid red line.

Fig. 2 shows the behavior of the cost function for the following three cases: the solid blue line shows the average behavior for the 10 random initial values, the dashed blue line shows the behavior for the initial value that resulted in the smallest final value of cost function among the 10 random initial values, and the solid red line shows the behavior for the initial value calculated by the IR-based initialization method.

It can be confirmed that the IR-based initialization method achieves the smallest value of the cost function even at the end and a faster convergence. This suggests that the IR-based initialization method yields the initial value close to the global optimum of the subproblem.

C. Comparison of separation performance

To confirm the efficiency of the proposed method, we compared the following three methods.

- **ILRMA [5]**
- **Proposed method (random init.)**: TDOAs are initialized by a random sampling.
- **Proposed method (IR-based init.)**: TDOAs are initialized by the IR-based initialization method using the demixing matrix precomputed by ILRMA.

In ILRMA, initialization was the same as that described in Section IV-B. The number of iterations was set to 200. In the proposed method (random init.), the demixing matrix and NMF variables were initialized in the same manner as in ILRMA in Section IV-B. TDOAs were initialized with random values from the uniform distribution on $[-5, 5]$. The number of iterations was also set to 200 as in baseline ILRMA. In the proposed method (IR-based init.), to start with better initial values of TDOAs, we first performed 20 iterations of ILRMA to obtain the demixing matrix as the preprocess. TDOAs were then initialized using the IR-based initialization method, followed by 180 iterations of the proposed method.

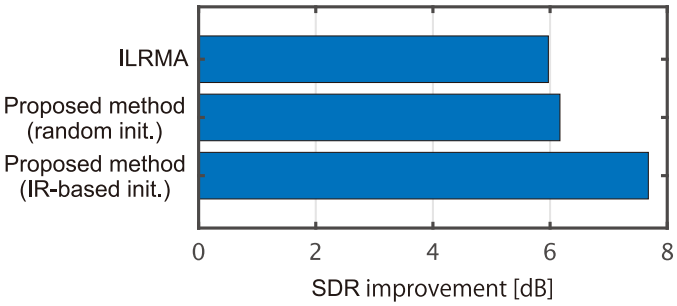


Fig. 3. Average SDR improvement for each method.

The preprocessing ILRMA was initialized in the same manner as in Section IV-B. In the subsequent proposed method, the NMF variables were again initialized as described in Section IV-B. The demixing matrix was initialized with the inverse of the mixing matrix $\hat{\boldsymbol{A}}_i$ computed from the TDOAs obtained by the IR-based initialization method.

Since the plane wave arrival model does not account for reverberation, strong regularization may affect the final separation performance. To mitigate this issue, the weighting coefficient $\mu_{inn'}^{(l)}$ for the regularizer in the lth update was decayed as

$$\mu_{inn'}^{(l)} = \begin{cases} 0, & (i = 1, I) \\ \mu_o \max\left[1/2 - l/L, 0\right], & (i \neq 1, I) \end{cases} \tag{53}$$

where L is the number of iterations for proposed methods, i.e., $L = 200$ in proposed method (random init.) and $L = 180$ in proposed method (IR-based init.). The value of μ_o was determined by a preliminary experiment and set to 10^{-3}.

In each method, we used 10 random seeds for each of the observed signals (total 360 trials) and calculated the average of source-to-distortion ratio (SDR) [20] improvement as the evaluation metric for the separation performance.

Fig. 3 shows the average SDR improvement for each method. It can be confirmed that the proposed method achieves better performance than ILRMA. Additionally, the proposed method (IR-based init.) achieves the largest SDR improvement. Therefore, we can confirm the effectiveness of the proposed fully blind method.

V. Conclusion

In this paper, we proposed a new fully blind method that models SVs using TDOAs and simultaneously estimates TDOAs, the demixing matrix, and NMF variables. To guarantee the monotonic non-increase in the cost function, we derived the MM-algorithm-based update rules. We also investigated the method of initializing TDOAs for the proposed cost function. Numerical experiments confirmed that the proposed method achieves higher separation performance than ILRMA. Additionally, we confirmed the effectiveness of the IR-based initialization method for the proposed method.

References

[1] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. SIP*, vol. 8, no. e12, pp. 1–14, 2019.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[3] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.

[4] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[5] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, pp. 1626–1641, 2016.

[6] D. Kitamura, N. Ono, and H. Saruwatari, "Experimental analysis of optimal window length for independent low-rank matrix analysis," in *Proc. EUSIPCO*, 2017, pp. 1170–1174.

[7] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

[8] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. ICASSP*, 2020, pp. 846–850.

[9] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," in *Proc. ICASSP*, 2018, pp. 746–750.

[10] R. Zhang, T. Ueda, and S. Makino, "Geometrically constrained blind moving source extraction based on constant separation vector and auxiliary function technique," in *Proc. APSIPA ASC*, 2023, pp. 2008–2012.

[11] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *JCGS*, vol. 9, no. 1, pp. 60–77, 2000.

[12] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[13] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[14] Y. Ishikawa, K. Konaka, T. Nakamura, N. Takamune, and H. Saruwatari, "Real-time speech extraction using spatially regularized independent low-rank matrix analysis and rank-constrained spatial covariance matrix estimation," in *Proc. HSCMA*, 2024.

[15] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE TSAP*, vol. 10, no. 6, pp. 352–362, 2002.

[16] N. Makishima, Y. Mitsui, N. Takamune, *et al.*, "Independent deeply learned matrix analysis with automatic selection of stable microphone-wise update and fast sourcewise update of demixing matrix," *Signal Processing*, vol. 178, p. 107 753, 2021.

[17] K. Yamaoka, R. Scheibler, N. Ono, and Y. Wakabayashi, "Sub-sample time delay estimation via auxiliary-function-based iterative updates," in *Proc. WASPAA*, 2019, pp. 130–134.

[18] A. Liutkus, F. R. Stöter, Z. Rafii, *et al.*, "The 2016 signal separation evaluation campaign," in *Proc. LVA/ICA*, 2017, pp. 323–332.

[19] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. LREC*, 2000.

[20] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

Incorporating Auditory Processing into Undergraduate Signal Processing Courses to Enhance Student Learning

Kaibao Nie
Electrical and Computer Engineering, School of Science, Technology, Engineering, and Mathematics
University of Washington, Bothell, USA
E-mail: niek@uw.edu

***Abstract*— This paper describes the educational experience of developing lab components and projects designed to enhance the understanding of key signal processing concepts in undergraduate Signals and Systems courses. Students work with signal-processing algorithms commonly used in auditory models and speech-processing applications for hearing aids and cochlear implants. This approach provides an interactive way to understand digital filters, nonlinear systems, speech analysis and synthesis, Fourier transform, and modulation/demodulation. Assessment results show that students have a highly favorable view of these specially designed labs and projects.**

Index terms — signal processing, student learning outcomes, auditory system, cochlear implant, vocoder.

I. INTRODUCTION

Signal processing is a foundational subject in electrical engineering and other engineering majors, providing students with essential skills for analyzing and manipulating signals, systems, and data. The classic Signals and Systems courses cover the fundamental knowledge of signal operations, system properties, filters, modulation/demodulation, Fourier transform, etc. However, engineering students typically understand the engineering science concepts but lack the skill to apply what they learned to solve real-world engineering problems [1]. Incorporating labs and projects tied to real-world applications into educational curricula offers significant benefits for student learning [2][3]. The hands-on activities can enhance engagement and motivation by connecting theoretical concepts to practical applications, which helps them develop critical thinking and problem-solving skills.

In Signals and Systems courses, many real-world signals can be used to design labs and projects that provide students with hands-on experience. Audio signals, particularly speech, are ideal for teaching time-domain and frequency-domain analysis. Students can visualize waveforms, apply Fourier transform, and analyze spectrograms to grasp concepts like filter, spectrum, and spectrogram. These signal-processing techniques are also widely used in applications such as auditory modeling, hearing aids, and cochlear implants [4-6]. Working with speech signals allows students to immediately observe the effects of their processing techniques, deepening their understanding of fundamental signal-processing concepts. By showcasing how auditory processing algorithms are applied in real-world scenarios—like noise cancellation, speech recognition, and hearing enhancement—students can better understand their studies' relevance and impact on everyday technology.

This paper details a series of laboratory experiments and projects focused on processing audio signals, particularly in the context of auditory modeling and hearing devices. Section II outlines the general design of the labs and projects, emphasizing the connection between signal processing concepts and their real-world applications. It also describes the key components of each lab and project. Section III presents the assessment results from student surveys, while Section IV offers a short summary and conclusions.

II. DESCRIPTION OF THE EXPERIMENTAL LABS AND PROJECTS

The signals and systems course covers a range of essential concepts in signal processing. Figure 1 illustrates the overall design of the lab series, which helps students grasp the primary concepts introduced in the course. One lab begins by exploring linear and nonlinear systems, supported by experiments demonstrating the cochlea's nonlinear processing and the half-wave/full-wave rectification used in auditory processing. The second lab component focuses on filters and Fourier transform, with practical lab procedures on designing low-pass, band-pass, and high-pass filters to filter speech signals. Students also engage in filter design through two projects (Sections II C, and

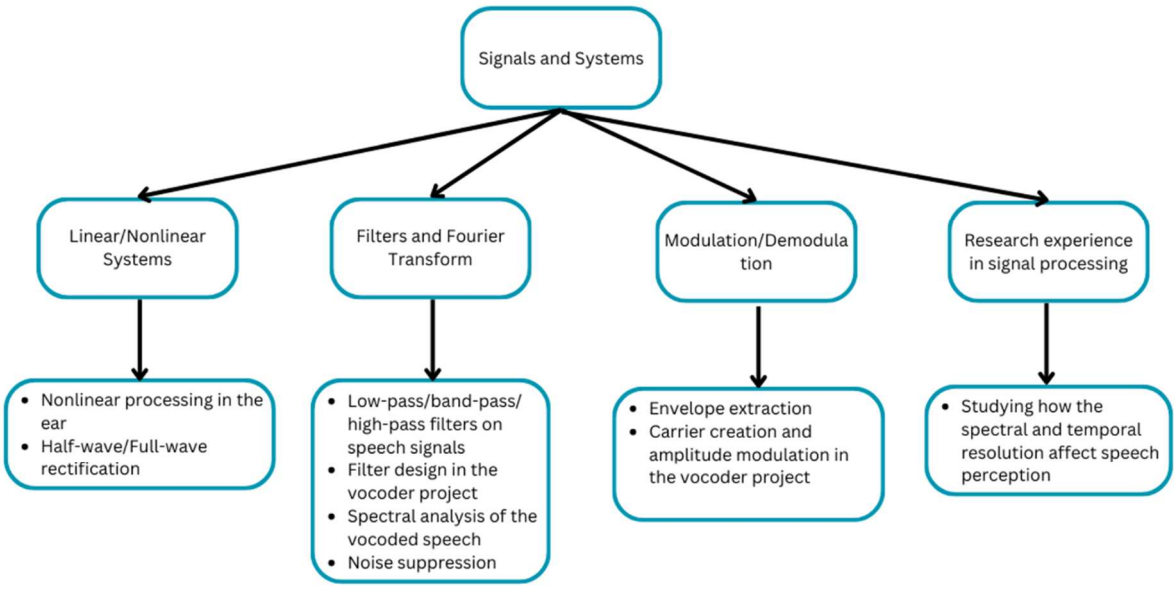


Fig. 1 Overall design of the lab series to connect the key signal processing concepts to real applications

D): developing a speech vocoder and designing noise suppression algorithms. Modulation and demodulation techniques are also introduced, with labs on envelope extraction, carrier creation, and amplitude modulation within the vocoder project. Additionally, students gain research experience by studying how spectral and temporal resolution impact speech perception in the speech vocoder project.

A. Nonlinear processing vs. linear processing

Understanding the differences between linear and nonlinear processing is crucial in signal processing. Nonlinear systems do not follow the principle of superposition. The response from nonlinear processing typically contains distortions and interactions between different frequency components, demonstrated by the presence of new frequency components (Fig. 2). The auditory system exhibits nonlinear behavior at various stages, ranging from the peripheral regions to the cortical level [7]. One great example is the Distortion Product Otoacoustic Emissions (DPOAEs) phenomenon [8], which occurs when the cochlea is simultaneously stimulated with two pure tones at different frequencies, f_1 and f_2. To illustrate how the combination tones ($f_1 + f_2$ and $f_2 - f_1$) are generated

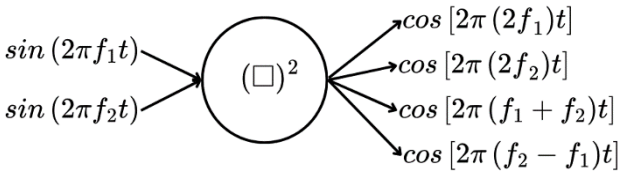


Fig. 2 Nonlinear processing in the auditory system.

due to nonlinear processing; in one lab, students are asked to mathematically derive how the combination tones can be generated by applying the squaring operation to the mixed signal. Next, students work on programming to create two tones at 1,000 Hz and 1,200 Hz and then play each separately. To experience the nonlinear processing in the ear, students write code to mix the two tones and play them simultaneously. Students are instructed to listen to the mixed tone and should hear a noticeable low-pitched tone at 200 Hz and a high-pitched tone at 2,200 Hz, which are not present in the mixed tone. Students not only get the chance to study the fundamental causes of nonlinear distortions but also learn about the practical applications of nonlinearity properties in human hearing screening using DPOAEs.

B. Digital filers and speech signal filtering

Digital filters and their design and implementation are the critical components of signals and systems undergraduate courses. Auditory models often include envelope extraction to simulate how the human ear processes sound. In hearing aids and cochlear implants, signal rectifiers are used to process speech signals to extract features and enhance speech intelligibility. Applying different filters to speech signals and exploring how they affect speech perception can provide an interactive way to understand digital filters' fundamentals.

Filters are commonly used in signal processing for hearing aids, cochlear implants, and other hearing devices. In the lab series, students design and apply three different types of filters (low-pass, band-pass, and high-pass) to process speech signals, allowing them to listen to the results and gain hands-on experience. They can vary the cutoff frequency of each filter to

study how it affects speech perception performance. This practical approach helps students understand the impact of different filter parameters and techniques on real-world audio processing, reinforcing theoretical concepts through direct application.

In addition, students are given the coding task of extracting the envelope signal of a speech sound. First, the speech signal passes through either a half-wave or full-wave rectifier for envelope extraction, another example of nonlinear processing. Then, the filters at cutoff frequencies from 400 Hz down to 4 Hz will be applied to extract a smooth envelope signal. Plots can be created to demonstrate how the cutoff corner frequency can affect the smoothness of the envelope of a speech signal.

C. *Filter design and speech denoising*

One primary ABET(Accreditation Board for Engineering and Technology) student learning outcome is acquiring the ability to identify, formulate, and solve complex engineering problems by applying engineering, science, and mathematics principles. To meet this requirement, a short design project was incorporated into the labs of the undergraduate course. Students are tasked with designing and implementing digital filters to remove noise from the mixture of a clean speech signal and noise signal (Fig. 3). The tasks include:

- *Using FFT to find the dominant frequency range of the noise signal*
- *Determine the filter needed and its parameters*
- *Find the input-output difference equation that represents the filter*
- *Write a FOR loop in Matlab/Python to implement the filter by processing the noisy speech sample by sample*
- *Demonstrate how the cutoff frequency setting and the filter's order affect the speech quality after noise suppression.*

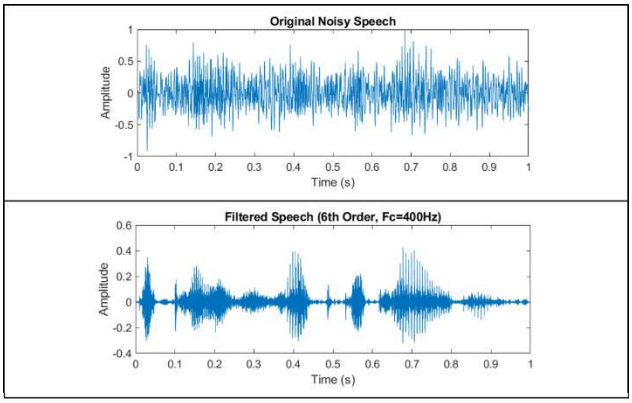


Fig. 3 The waveform of the noisy speech signal and the waveform of the expected output signal after noise removal

D. *The sound vocoder project*

A sound vocoder (Fig. 4) is a signal-processing algorithm used to analyze and synthesize audio signals for applications such as audio compression and cochlear implant simulations [9-11]. It incorporates many of the key signal-processing components typically covered in undergraduate Signals and Systems courses, including band-pass filters, low-pass filters, nonlinear rectifiers, envelope extraction, amplitude modulation, etc. (Fig. 4). Students will also be able to learn speech analysis, speech synthesis, and FFT-based spectrum and spectrogram

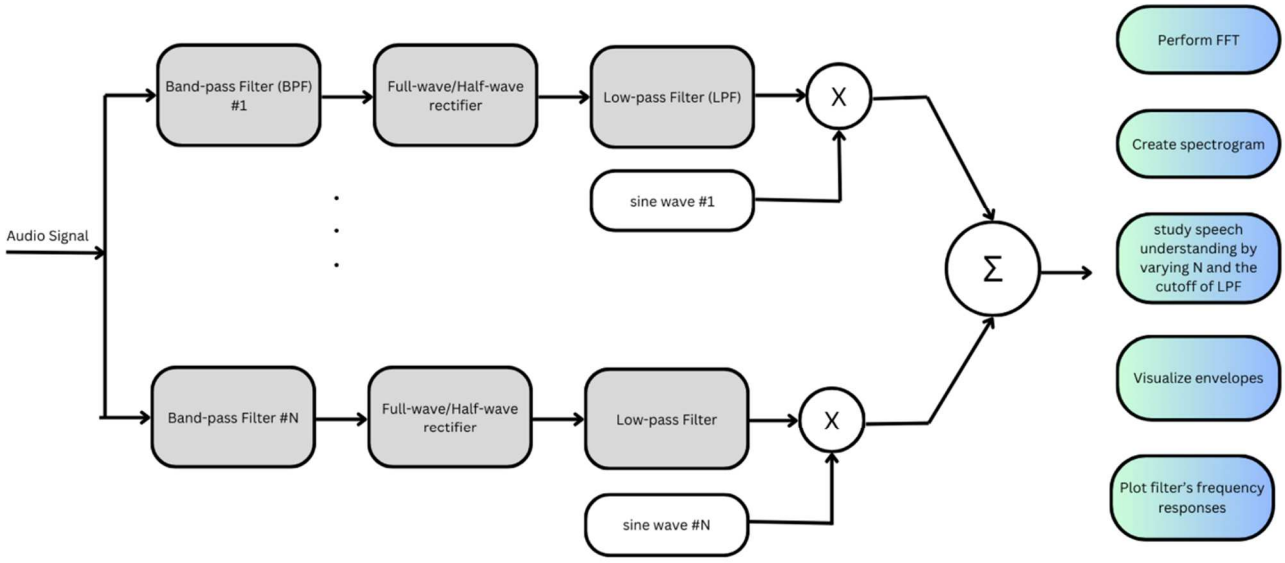


Fig. 4 Block diagram of a speech vocoder and expected leaning outcomes.

generation techniques. It can also help students develop better coding skills. The major requirements are:

- *Design and apply Butterworth band-pass and low-pass filters*
- *Extract envelopes from the filtered signal*
- *Modulate envelopes with sinusoidal carriers*
- *Rethenthesize a sound from N bands*
- *Create spectrograms to demonstrate the spectral contents of vocoded speech (Fig. 5)*
- *Study speech understanding with reduced spectral and temporal cues*

Students can also gain experience conducting research and analyzing data in the vocoder project. One of the assignments is to study how speech understanding performance can be affected by spectral resolution when varying the number of bands, *N*. Fig. 6 presents a sample result by one student, demonstrating that the percentage of words correctly identified increases with the number of bands.

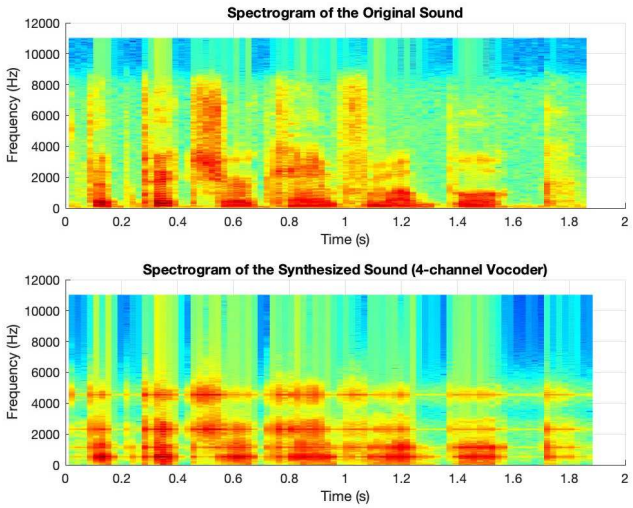


Fig. 5 Spectrograms of a speech signal and the vocoded speech.

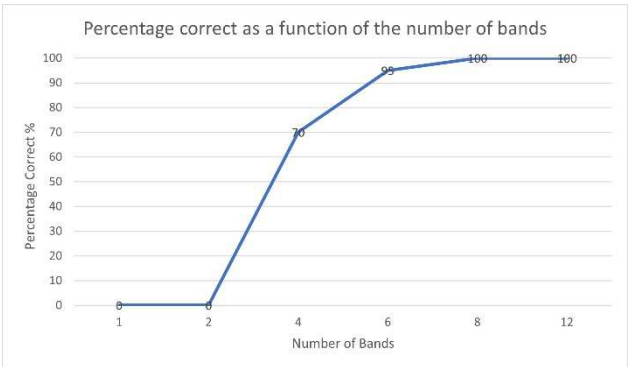


Fig.6 Study results to demonstrate the performance of speech understanding when varying the number of processing bands in the speech vocoder.

III. Assessment Results

To evaluate the effectiveness of the labs and projects involving auditory processing, students were asked to provide feedback at the end of the course. Out of 24 students, 15 responded to the survey questions, as shown in Table I. Over 93% of the students rated the labs and projects as good to excellent, indicating success in stimulating their interest in learning and enhancing their understanding of key signal processing concepts.

Table I Student rating scores on the labs and projects

Survey Questions	Good, very good, and excellent
The lab section as a whole was	100%
The content of the lab section was	93%
The interest level of lab sessions was	100%
The amount you learned in the lab sections was	93%
The relevance and usefulness of lab section content was	93%

IV. Summary and Conclusions

By integrating auditory processing into signal processing courses, students can gain a deeper understanding of theoretical concepts in digital signal processing through engaging examples and hands-on practices. It can stimulate students' interest in learning by illustrating how the auditory system processes sounds and allowing them to hear the effects of signal processing. This approach enhances student learning outcomes and better prepares students for careers in technical fields related to audio and speech signal processing.

V. Acknowledgment

Thanks to the students who participated in the course survey and provided valuable feedback.

References

[1] J.E. Mills and D. F. Treagust. "Engineering education—Is problem-based or project-based learning the answer?" Australasian Journal of Engineering Education 3 pp. 2-16, 2003.

[2] Y. Alqudah, E. Qaralleh, and M. Mace, "Enhancing the Teaching of Digital Signal Processing through Project-Based Learning", International Journal of Online Engineering, Vol. 9, 2013

[3] J. Macias-Guarasa, J. M. Montero, R. San-Segundo, A. Araujo and O. Nieto-Taladriz, "A project-based learning approach to design electronic systems curricula," in *IEEE Transactions on Education*, vol. 49, no. 3, pp. 389-397, Aug. 2006, doi: 10.1109/TE.2006.879784.

[4] Jepsen ML, Ewert SD, Dau T. A computational model of human auditory signal processing and perception. J Acoust Soc Am. 2008 Jul;124(1):422-38. doi: 10.1121/1.2924135. PMID: 18646987.

[5] Loizou PC. Signal-processing techniques for cochlear implants. IEEE Eng Med Biol Mag. 1999 May-Jun;18(3):34-46. doi: 10.1109/51.765187. PMID: 10337562.

[6] Hohmann, V. (2008). Signal Processing in Hearing Aids. In: Havelock, D., Kuwano, S., Vorländer, M. (eds) Handbook of Signal Processing in Acoustics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-30441-0_14

[7] Carney LH, McDonough JM. Nonlinear auditory models yield new insights into representations of vowels. Atten Percept Psychophys. 2019 May;81(4):1034-1046. doi: 10.3758/s13414-018-01644-w. PMID: 30565098; PMCID: PMC6581637

[8] Zelle D, Lorenz L, Thiericke JP, Gummer AW, Dalhoff E. Input-output functions of the nonlinear-distortion component of distortion-product otoacoustic emissions in normal and hearing-impaired human ears. J Acoust Soc Am. 2017 May;141(5):3203. doi: 10.1121/1.4982923. PMID: 28599560; PMCID: PMC5426960.

[9] Cychosz M, Winn MB, Goupell MJ. How to vocode: Using channel vocoders for cochlear-implant research. J Acoust Soc Am. 2024 Apr 1;155(4):2407-2437. doi: 10.1121/10.0025274. PMID: 38568143; PMCID: PMC10994674.

[10] Fu QJ, Shannon RV. Effects of dynamic range and amplitude mapping on phoneme recognition in Nucleus-22 cochlear implant users. Ear Hear. 2000 Jun;21(3):227-35. doi: 10.1097/00003446-200006000-00006. PMID: 10890731.

[11] Friesen LM, Shannon RV, Baskent D, Wang X. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. J Acoust Soc Am. 2001 Aug;110(2):1150-63. doi: 10.1121/1.1381538. PMID: 11519582.

A Comparative Study on the Biases of Age, Gender, Dialects, and L2 speakers of Automatic Speech Recognition for Korean Language

Jonghwan Na*, Yeseul Park† and Bowon Lee‡
* Inha University, Incheon
E-mail: jhna@dsp.inha.ac.kr Tel/Fax: +82 10-4623-0591
† Inha University, Incheon
E-mail: yspark@dsp.inha.ac.kr Tel/Fax: +82 10-2587-8450
‡ Inha University, Incheon
E-mail: bowon.lee@inha.ac.kr Tel/Fax: +82 32-860-7423

***Abstract*—Recent advancements in the field of Automatic Speech Recognition (ASR) have seen the emergence of large-scale models, contributing to a surge in research and development. The performance of recent ASR models has rapidly improved with the utilization of extensive pre-training datasets. However, challenges persist in enhancing the recognition accuracy for non-mainstream groups such as the elderly and speakers of regional dialects. This paper conducts experiments using Korean speech data to compare and analyze the biases related to gender, age, dialects, and second language (L2) Korean speakers using the Conformer, wav2vec 2.0, and Whisper models. The experimental results showed that female results exhibited better performance in ASR models than those of males, and Whisper exhibited lower biases than two other models in most cases. Furthermore, Whisper demonstrated robustness compared to the other two models in the L2 speakers. Additionally, the analysis of characters with high error rates for each group revealed that, in the case of Korean, spacing and particles exhibited high error rates. It was also observed that characters with high error rates were similar within age groups rather than between gender groups. In this study, we conducted the first-ever examination of various biases in Korean ASR. The identified biases through these experiments may serve as a starting point for research aimed at improving the performance of ASR for non-mainstream groups. This study underscores the significance of addressing biases to advance fairness in the field of ASR.**

I. Introduction

Since the introduction of the transformer architecture [1], various models based on transformers have been proposed in the field of Automatic Speech Recognition (ASR), and their performances have significantly improved. Under these circumstances, research efforts in the field of ASR are actively focused on enhancing inclusiveness and accessibility [2]–[8]. As part of such research, studies on biases in ASR performance across different cohorts is being conducted from various perspectives, such as studies analyzing bias across different age groups [7], [9], between genders [7], [9]–[14], and based on racial backgrounds [4], [7], [12], [15]. Studies on such biases are not limited to English [7], [12]–[15] but are also extensively conducted based on various languages, including Dutch [9], [16], [17], French [10], [13], Mandarin [11], and Arabic [18].

According to prior research, gender bias in ASR models generally indicates that these models exhibit higher performance for females than males [9], [11], [13]. However, some studies have reported instances where they demonstrate superior performance for males than females [10], [14]. Regarding age-based bias, Feng et al. reported that, on a Dutch dataset, higher performance is observed in the order of teenagers, older individuals, and young children [9]. Conversely, Liu et al. reported superior performance in the elderly compared to other age groups on the English dataset [7]. Thus, the bias in ASR models appears to vary depending on the language used and the specific model employed in the experiments. To the best of our knowledge, no study comparing and analyzing biases of ASR models on Korean language has been reported in the literature.

In this regard, we conducted experiments in three different cases of ASR training approaches: training a model from scratch, fine-tuning of a model pre-trained through self-supervised learning, and finally, fine-tuning of a model pre-trained through supervised learning using three different ASR models, namely Conformer [19], wav2vec 2.0 [20], Whisper [21], respectively. We evaluated the performance of these ASR models for three age groups – children, adults, and elderly – and two genders, male and female. We examined and compared biases within each group. Additionally, we investigated the regional biases of six dialects of Korea, namely, the dialects of the greater Seoul, the Korea's metropolitan area (MA), Chungcheong (CC), Gyeongsang (GS), Jeolla (JL), Gangwon (GW), and Jeju (JJ). Finally, we compared the speakers whose native languages are English, Chinese, and Japanese, learning Korean as a second language (L2).

The identified biases through these experiments may potentially serve as a starting point for research aimed at reducing the bias of ASR especially for non-mainstream groups including children, elderly individuals, speakers of regional dialects, and L2 speakers. This study underscores the importance of analyzing biases to advance fairness in the field of ASR.

TABLE I: Number of speakers, number of utterances, duration, and the average number of characters in train and test sets

Group	Number of speakers			Number of utterances			Duration (hours)			Avg. number of characters		
	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
train-children	105	135	240	50,000	50,000	100,000	65	62	127	14.23	14.17	14.20
train-adults	465	775	1,240	50,000	50,000	100,000	68	73	141	19.33	19.36	19.35
train-elderly	462	495	957	50,000	50,000	100,000	123	126	249	35.67	35.19	35.43
train-total	1,032	1,405	2,437	150,000	150,000	300,000	256	261	517	23.08	22.91	22.99
test-children	56	72	128	30,000	30,000	60,000	38	38	76	14.31	13.98	14.15
test-adults	273	361	634	30,000	30,000	60,000	39	40	79	19.26	19.52	19.39
test-elderly	245	260	605	30,000	30,000	60,000	74	77	151	36.47	36.01	36.24
test-MA	25	29	54	12,083	13,806	25,889	25	28	53	26.68	26.84	26.77
test-CC	7	8	15	4,588	5,026	9,614	8	9	17	25.92	25.86	25.89
test-GS	9	12	21	4,203	6,449	10,652	8	10	18	28.73	29.02	28.91
test-JL	4	5	9	2,160	2,971	5,131	4	5	9	26.56	26.63	26.60
test-GW	3	3	6	1,637	1,481	3,118	3	2	5	26.35	26.77	26.55
test-JJ	1	1	2	555	512	1,067	1	1	2	28.81	29.14	28.97
test-english L2	37	107	144	673	1,994	2,667	3	7	10	57.16	57.14	57.15
test-chinese L2	32	311	343	1,732	1,828	3,560	6	6	12	57.04	57.20	57.12
test-japanese L2	13	215	228	721	1,847	2,568	2	6	8	57.22	56.29	56.55

II. Experimental Setup

A. Datasets

Table I presents the statistics of the training and testing datasets used in the experiments. The datasets employed in the experiments were sourced from AI-hub[1]. Four distinct types of datasets, namely "Free Conversation Speech (Infants and Children)", "Free Conversation Speech (General Adults)", "Free Conversation Speech (Elderly)", and "Speech Data of Non-Native Korean Speakers for AI Training", were chosen and employed in the experiments. The training data utilized in the experiments were randomly sampled from the remaining three datasets, excluding non-native speakers, with $100,000$ utterances selected from each dataset. In this way, a total of $300,000$ utterances comprised the training data.

Throughout this process, the gender ratio within the training dataset was adjusted to achieve the of ratio of 1:1, aligning with the objectives of the experiments. The training dataset comprises a total duration of 517 hours, and the number of utterances was uniformly distributed across all groups. Examining the average number of characters in transcripts reveals that, in descending order, the elderly, adults, and children exhibit higher counts. The test datasets were uniformly composed of $30,000$ utterances for each age and gender group, and the regional and L2 speaker datasets were obtained as described in Table I. These utterances were randomly extracted for each age and gender category, ensuring the exclusion of shared speakers between the training and test datasets.

B. Character Error Rate

For experiments, the Character Error Rate (CER) is adopted as the evaluation metric to conform with the common practices for Korean ASR due to the unique characteristics of the language. The equation for calculating the CER is

$$CER = \frac{S + D + I}{N}, \tag{1}$$

where it is computed by summing the number of deletions (D), substitutions (S), and insertions (I) of the characters, and then dividing it by the total number of characters (N).

C. Models

We selected three ASR models – Conformer [19], wav2vec 2.0 [20], and Whisper [21] – in our experiments. The reason for selecting the Conformer was to analyze biases in a model trained from scratch. The model based on Conformer is comprised of a Conformer encoder and a Connectionist Temporal Classification (CTC) [22] decoder. The Conformer encoder is configured as Conformer-large with 17 layers and 8 attention heads, amounting to a total of 118.8 million parameters. The initial learning rate was set to 1e-0.6, with a batch size of 32. The experiment utilized a tri-stage learning rate scheduler, and the optimizer employed was Adam.

The rationale behind the selection of wav2vec 2.0 was driven by the need to examine the bias inherent in pre-trained models when utilizing unlabeled data. Like the Conformer, wav2vec 2.0 was a model that demonstrated the SOTA performance upon its release. For the experiment, the encoder was configured as wav2vec 2.0 while the decoder was set as CTC. The wav2vec 2.0 encoder utilized a large model consisting of 24 transformer blocks and attention heads. Additionally, the experiment utilized the pre-trained wav2vec 2.0 XLSR-53 model, trained on a $56,000$-hour dataset spanning 53 languages. The wav2vec 2.0 XLSR model contains 300 million parameters, with an initial learning rate of 4e-4, a batch size of 32, and a linear learning rate scheduler. Adam optimizer was used for the experiments.

Whisper is a pre-trained model like wav2vec 2.0 except it is pre-trained with labeled data. This distinction led to its

[1] https://aihub.or.kr/

inclusion in the experiments to scrutinize biases in models pre-trained through supervised learning on multilingual data. Both its encoder and decoder are structures with transformer-based layers. For the experiments detailed in this paper, the Whisper-small model was utilized, comprising 12 layers and 12 attention heads, totaling 244 million parameters. The initial learning rate was configured at 5e-07, accompanied by a batch size of 32, and a linear learning rate scheduler was employed. The Adam optimizer was used for training. Finally, all models were trained for 25 epochs, incorporating an early stopping mechanism.

III. Results and Discussions

A. Age and Gender Biases

Table II shows the CERs on the test datasets of the different age groups and gender groups. In terms of overall performance, it can be observed that wav2vec 2.0, Whisper, and Conformer exhibit superior performance in the order mentioned, yielding the average CERs of 2.15%, 3.24%, and 5.17% respectively.

When comparing performance based on gender, it can be observed that across all models, the test data for females exhibit slightly better performance than males. This suggests an ASR models have a bias favoring females in Korean, regardless of age. Conformer demonstrated the most pronounced gender-based bias, with a difference in CER values between genders being 1.02. In contrast, wav2vec 2.0 and Whisper demonstrated higher resilience to gender-based bias, with values of 0.43 and 0.71, respectively. These results imply that models pre-trained on various and extensive datasets, such as wav2vec 2.0 and Whisper, are less susceptible to biases compared to the model trained from scratch. It is interesting to observe that including multi-lingual dataset helps the performance of the target language.

Upon comparing performance across different age groups, Conformer demonstrated better performance in the order of adults, children, and elderly testsets. For wav2vec 2.0, the best performance was observed in the order of children, adults and the elderly. Lastly, for Whisper, performance was found to be superior in the order of adults and elderly, with an interesting observation that performance on the children's dataset was the lowest. The reason why Whisper exhibited the lowest performance on the children test dataset, unlike the other models, is likely due to the fact that the Whisper model takes the hyperparameter '*max length*' as input. This is speculated to result in a higher CER for the relatively shorter transcript length, with hallucination issues occurring in some cases for Whisper [23]. In this experiment, as can be seen in Table I, the sentences of children dataset is shorter than those of the other two datasets. The age-based standard deviations of CER are 0.70 for Whisper, 0.78 for Conformer, and 0.98 for wav2vec 2.0, indicating that Whisper shows the smallest bias.

Table III displays the proportions of the sentences for non-zero CERs and Fig. 1 displays the distribution of their CER values by age and gender across three models. As evident from Table III, it is apparent that the model with the highest performance, wav2vec 2.0, exhibits the lowest proportion of non-zero CER. Comparing non-zero CER across genders reveals that females consistently exhibit lower values than males across all cases, consistent with the overall trends observed in Table II. Moreover, the gender disparity in non-zero CER proportions is also least pronounced in wav2vec 2.0. An intriguing observation is that while Table II shows a marginal difference in performance between wav2vec 2.0 and Whisper on the elderly test dataset, Table III indicates that Whisper has a lower proportion of non-zero CER, suggesting that although Whisper may correctly infer more sentences, the average CER of erroneous sentences is higher compared to wav2vec 2.0.

TABLE II: CER (%) on Age and Gender-Specific Test Datasets using Conformer large, wav2vec 2.0, and Whisper-small

Model	Children		Adults		Elderly	
	Male	Female	Male	Female	Male	Female
Conformer	4.98	5.14	4.96	3.93	7.09	4.92
wav2vec 2.0	1.57	1.56	1.88	1.57	3.64	2.65
Whisper-small	4.32	3.57	2.85	2.25	3.63	2.84

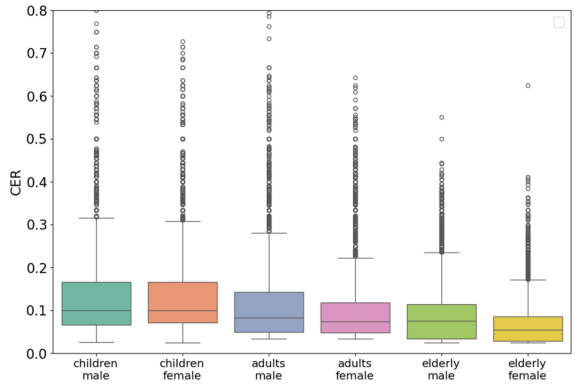
TABLE III: Proportion of non-zero CER sentences in the test set

Model	Children		Adults		Elderly	
	Male	Female	Male	Female	Male	Female
Conformer	0.45	0.45	0.49	0.44	0.82	0.73
wav2vec 2.0	0.15	0.15	0.22	0.19	0.67	0.58
Whisper-small	0.38	0.29	0.29	0.26	0.60	0.52

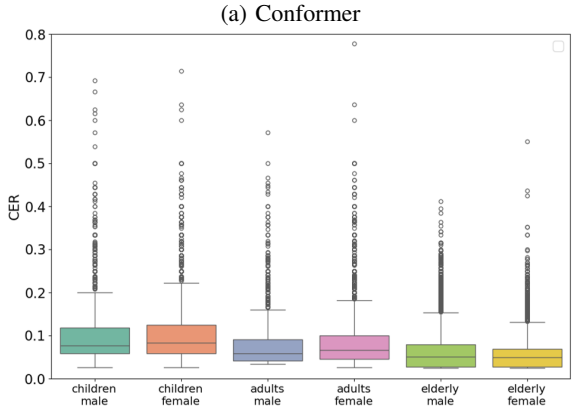
In Fig. 1, for the model with the best overall performance, wav2vec 2.0, the outliers in non-zero CER values are relatively lower compared to the other two models. Additionally, examining the distribution of outliers by gender in Fig. 1 reveals that, paradoxically, in cases where performance was better for females, there are more instances of higher outlier values. Viewing Fig. 1 alongside Table III, it becomes apparent that Whisper exhibits instances where the CER exceeds 0.5 for male and female children, who have the shortest transcripts. Conversely, Table III indicates that Whisper has a lower non-zero CER than the other two models for the elderly, who have the longest transcripts. This is considered as a result arising from the inherent characteristics of the model, as mentioned earlier.

TABLE IV: CER (%) on Korean regional-specific test datasets using Conformer large, wav2vec 2.0, and Whisper-small

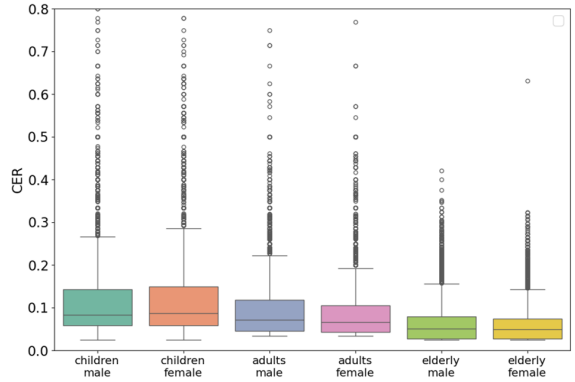
Model	**MA**	**CC**	**GS**	**JL**	**GW**	**JJ**	**Avg**	**SD**
Conformer	5.02	6.25	6.97	4.08	7.96	9.14	6.57	1.87
wav2vec 2.0	2.67	3.10	3.82	2.18	4.14	5.03	3.49	1.04
Whisper-small	2.92	3.52	4.17	2.37	4.23	4.85	3.68	0.92



(a) Conformer



(b) wav2vec 2.0



(c) Whisper-small

Fig. 1: Distributions of non-zero Character Error Rates for Conformer, wav2vec 2.0, and Whisper-small

B. Regional and L2 Biases

The experimental results with six different regions in Korea are summarized in Table IV. It was evident that the performance of the test set from the JL region was the best across all models. Conversely, the performance on the JJ dataset was consistently the lowest across all models. The distinctiveness of the JJ dialect due to its geographical isolation as an island likely contribute to the highest CERs. This is also supported by studies showing that JJ dialect is largely incomprehensible to monolingual speakers of standard Korean, indicating a significant linguistic divide [24]. Examining the

TABLE V: CER (%) on L2 speakers of English, Chinese, and Japanese Test Datasets using Conformer large, wav2vec 2.0, and Whisper-small

Model	English			Chinese			Japanese		
	Male	Female	All	Male	Female	All	Male	Female	All
Conformer	21.72	19.49	20.09	22.74	19.99	21.33	21.65	19.36	20.02
wav2vec 2.0	17.95	16.54	16.89	19.05	17.08	18.03	18.67	17.09	17.52
Whisper-small	10.10	9.96	10.00	13.17	12.00	12.57	12.80	12.59	12.65

average performance for the six regions, wav2vec 2.0 achieved the best performance with a CER of 3.49%, followed by Whisper and Conformer with CERs of 3.68% and 6.57%, respectively. Checking the standard deviation of CERs for the six regions, Whisper exhibited the lowest value at 0.92, indicating that Whisper has the least bias towards regional dialects. This suggests that Whisper, being pre-trained on diverse languages, and datasets, is more robust to biases.

Table V shows the experimental results with three different Korean L2 speakers whose native languages are English, Chinese, and Japanese. The results are further broken down by gender, with separate CER values for male, female, and all speakers combined. The experimental results indicated that English native L2 speakers achieved the highest performance in Korean speech recognition, followed by Japanese and Chinese native L2 speakers. Analyzing gender bias among L2 groups, Chinese native L2 speakers exhibit the most significant gender bias in CER with an average bias of 1.96, while English native L2 speakers show the least gender bias with an average bias of 1.26. Among the three models, Whisper consistently demonstrated superior performance across all language groups and genders, achieving the lowest CERs compared to Conformer and wav2vec 2.0. This suggests that Whisper may be more effective in handling the speech variability of L2 speakers, providing more accurate speech recognition capabilities. The results also highlight a consistent pattern where female speakers tend to have slightly lower CERs compared to male speakers across all models and languages.

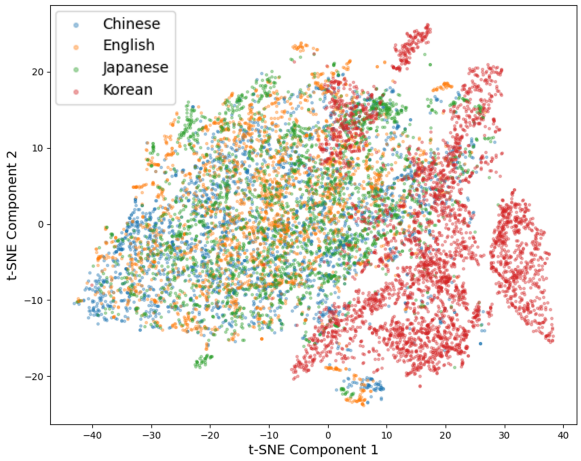


Fig. 2: Distributions of t-SNE of MFCC features for L2 speakers from China, Japan, English-speaking countries, and native Korean speakers

TABLE VI: Top 5 error characters excluding 9 consistently high-CER characters (/ /, /ʌ/, /nɯn/, /ɛ/, /hɛ/, /i/, /ɯn/, /kʌ/, /ɯl/) across models, by gender and age group

		Conformer	wav2vec 2.0	Whisper-small
Children	M	/nɑ/, /ɑ/, /kɑ/, /mjʌn/, /hɑ/	/kɑ/, /ɑ/, /hɑ/, /nɑ/, /il/	/nɑ/, /ɑ/, /kɑ/, /hɑ/, /ko/
	F	/kɑ/, /nɑ/, /ɑ/, /to/, /il/	/kɑ/, /ɑ/, /to/, /it/, /ko/	/kɑ/, /ɑ/, /nɑ/, /ko/, /to/
Adults	M	/hɑ/, /ɑ/, /nɑ/, /ja/, /tɕi/	/ko/, /kɑ/, /ku/, /to/, /hɑ/	/ku/, /ɯi/, /hɑ/, /ko/, /ɑ/
	F	/ɯi/, /nɑ/, /ɑ/, /hɑ/, /tɕi/	/ko/, /ɯi/, /ɑ/, /ku/, /hɑ/	/ku/, /ɯi/, /hɑ/, /ɑ/, /nɛ/
Elderly	M	/hɑ/, /kɑ/, /sʌ/, /nɑ/, /ɯi/	/hɑ/, /kɑ/, /ɯi/, /lɯl/, /ko/	/kɑ/, /hɑ/, /sʌ/, /ɯi/, /ko/
	F	/kʌt/, /kɑ/, /hɑ/, /ɯi/, /sʌ/	/ɯi/, /kʌt/, /kɑ/, /hɑ/, /ɑ/	/kʌt/, /hɑ/, /ɑ/, /kɑ/, /nɛ/

Figure 2 shows the t-SNE plot of Mel-frequency cepstral coefficients (MFCC) features extracted from Korean L2 speakers whose native languages are Chinese, English, and Japanese, as well as native Korean speakers. Notably, there is a significant distinction between native Korean speakers and the other L2 speaker groups. The Chinese L2 speaker cluster exhibits a wider distribution, likely due to the presence of tonal characteristics in the Chinese language. Additionally, the t-SNE distributions of Japanese and English L2 speakers show relatively overlapping regions compared to the other languages. These differences indicate clustering patterns that reflect the underlying acoustic and phonetic influences of the speakers' native languages.

C. Characters with High CER

To analyze the bias results of each group, we examined characters exhibiting high CERs. The investigation revealed that among the top 20 characters with the highest errors for each gender and age group across three models, 9 characters – / /, /ʌ/, /nɯn/, /ɛ/, /hɛ/, /i/, /ɯn/, /kʌ/, /ɯl/ – appeared consistently. Notably, the character '/ /' representing the space between words, was the most error-prone in all cases. Also, the characters corresponding to Korean particles, which are functional elements attached to nouns, verbs, or adjectives to indicate grammatical relations, were frequently misinferred. This is considered to be influenced by the characteristics of the Korean language, attributed to the somewhat intricate rules of spacing and the flexible utilization of particles, which are well known to be two major sources of grammatical errors made even by humans [25], [26]. Table VI presents the top 5 characters, excluding the aforementioned 9 characters, that exert the highest influence on CER for each model across gender and age groups. The top error characters exhibit a high degree of similarity between males and females within the same age group although they differ among the age groups.

IV. Conclusion and Future Works

This paper presented a comparative analysis of various biases for Korean ASR, and the experimental findings provide insights into the performance and biases of the recent ASR models trained with Korean speech datasets. Wav2vec 2.0 and Whisper boast notable robustness to biases, especially regarding the gender difference. The analysis revealed a consistent bias favoring females across all models, with Conformer exhibiting the most pronounced gender-based bias. Furthermore, we observed that age-based bias is least prominent in the adults category across all models, with Whisper exhibiting the smallest age-based bias. Regional biases were also examined, revealing that the Whisper model demonstrates the least amount of regional bias, potentially by training with the vast amount of multi-lingual datasets, and all models exhibit the lowest CER in the JL region. Through experiments, the performance differences of Korean ASR models for L2 speakers were also examined, revealing that the models performed best with English native speakers, followed by Japanese and Chinese.

Additionally, it was observed that the Whisper model outperformed the other two models, which can be attributed to the fact that Whisper was pre-trained on a large amount of labeled data. This pre-training likely contributed to its robustness against various speech patterns and contextual biases in transcriptions.

In line with previous studies of other languages, we have substantiated the presence of biases in Korean ASR models through our experiments. If these biases in ASR models can be addressed properly, it can lead to the development of more inclusive ASR models, which could result in overall performance improvements.

Research on analyzing the identified biases from acoustic, prosodic, and phonetic perspectives through experiments and exploring methods to resolve them is planned as further research. Acquiring additional dialectal datasets will be desirable for more thorough examination of the regional biases. Also, considering the influence of native languages is crucial when evaluating and improving ASR model performance for L2 speakers. Therefore, understanding the clear differences between native languages is deemed necessary. Hence, it is believed that further investigation into the differences among L2 speakers from various perspectives will be required in future studies.

Acknowledgment

This work was supported by the Ministry of Science and ICT of the Republic of Korea and the National Research Foundation of Korea (NRF-2023R1A2C2006725) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) (RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] T. Wang, S. Hu, J. Deng, *et al.*, "Hyper-parameter Adaptation of Conformer ASR Systems for Elderly and Dysarthric Speech Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 1733–1737.

[3] P. DHERAM, M. Ramakrishnan, A. Raju, *et al.*, "Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities," in *Proc. Interspeech 2022*, 2022, pp. 1268–1272.

[4] J. L. Martin and K. Tang, "Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be"," in *Proc. Interspeech 2020*, 2020, pp. 626–630.

[5] R. Gale, L. Chen, J. Dolata, J. van Santen, and M. Asgari, "Improving ASR Systems for Children with Autism and Language Impairment Using Domain-Focused DNN Transfer Techniques," in *Proc. Interspeech 2019*, 2019, pp. 11–15.

[6] M. Moore, H. Venkateswara, and S. Panchanathan, "Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems," in *Proc. Interspeech 2018*, 2018, pp. 466–470.

[7] C. Liu, M. Picheny, L. Sarı, *et al.*, "Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6162–6166.

[8] H. Gao, X. Wang, S. Kang, *et al.*, "Seamless equal accuracy ratio for inclusive ctc speech recognition," *Speech Communication*, vol. 136, pp. 76–83, 2022.

[9] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[10] M. Garnerin, S. Rossato, and L. Besacier, "Gender representation in french broadcast corpora and its impact on asr performance," in *Proceedings of the 1st international workshop on AI for smart TV content production, access and delivery*, 2019, pp. 3–9.

[11] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101 567, 2024.

[12] R. Tatman and C. Kasten, "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions," in *Proc. Interspeech 2017*, 2017, pp. 934–938.

[13] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" In *Ninth European Conference on Speech Communication and Technology*, 2005.

[14] R. Tatman, "Gender and dialect bias in youtube's automatic captions," in *Proceedings of the first ACL workshop on ethics in natural language processing*, 2017, pp. 53–59.

[15] M. P. Y. Chan, J. Choe, A. Li, Y. Chen, X. Gao, and N. Holliday, "Training and typological bias in ASR performance for world Englishes," in *Proc. Interspeech 2022*, 2022, pp. 1273–1277.

[16] M. Fuckner, S. Horsman, P. Wiggers, and I. Janssen, "Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers," in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, IEEE, 2023, pp. 146–151.

[17] Y. Zhang, A. Herygers, T. Patel, Z. Yue, and O. Scharenborg, "Exploring data augmentation in bias mitigation against non-native-accented speech," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–8.

[18] G. Droua-Hamdani, S.-A. Selouani, and M. Boudraa, "Speaker-independent asr for modern standard arabic: Effect of regional accents," *International Journal of Speech Technology*, vol. 15, pp. 487–493, 2012.

[19] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, PMLR, 2023, pp. 28 492–28 518.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[23] A. Koenecke, A. S. G. Choi, K. X. Mei, H. Schellmann, and M. Sloane, "Careless whisper: Speech-to-text hallucination harms," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1672–1681.

[24] C. Yang, W. O'Grady, S. Yang, N. H. Hilton, S.-G. Kang, and S.-Y. Kim, "Revising the language map of korea," *Handbook of the Changing World Language Map*, pp. 215–229, 2020.

[25] D.-J. Kim, "A method for detection and correction of pseudo-semantic errors due to typographical errors," *Journal of the Korea Society of Computer and Information*, vol. 18, no. 10, pp. 173–182, 2013.

[26] S.-H. Lee, M. Dickinson, and R. Israel, "Developing learner corpus annotation for korean particle errors," in *Proceedings of the Sixth Linguistic Annotation Workshop*, 2012, pp. 129–133.

Detecting Spoof Voices in Asian Non-Native Speech: An Indonesian and Thai Case Study

Aulia Adila, Candy Olivia Mawalim, and Masashi Unoki
Japan Advanced Institute of Science and Technology, Japan
E-mail: {adila, candylim, unoki}@jaist.ac.jp

***Abstract*—This study focuses on building effective spoofing countermeasures (CMs) for non-native speech, specifically targeting Indonesian and Thai speakers. We constructed a dataset comprising both native and non-native speech to facilitate our research. Three key features—MFCC, LFCC, and CQCC—were extracted from the speech data, and three classic machine learning-based classifiers—CatBoost, XGBoost, and GMM—were employed to develop robust spoofing detection systems using the native and combined (native and non-native) speech data. This resulted in two types of CMs: Native and Combined. The performance of these CMs was evaluated on both native and non-native speech datasets. Our findings reveal significant challenges faced by Native CM in handling non-native speech, highlighting the necessity for domain-specific solutions. The proposed method shows improved detection capabilities, demonstrating the importance of incorporating non-native speech data into the training process. This work lays the foundation for more effective spoofing detection systems in diverse linguistic contexts.**

I. Introduction

Over the past few years, voice-based authentication systems have become prominent for verifying identities and recognizing spoken utterances, offering convenience in various scenarios. However, they are vulnerable to spoofing attacks, particularly through logical access (LA) scenarios [1] involving Text-to-Speech (TTS) and Voice Conversion (VC). While much research has focused on spoofing detection in English and other languages [2], [3], there is a significant gap in studies addressing non-native accents, particularly within the Asian region. English-speaking accents in these countries present unique challenges, as non-native pronunciations make it difficult for spoofing detection systems to accurately differentiate between genuine and spoofed speech.

This gap becomes critical in applications such as language proficiency tests, where foreign speakers might attempt to fool automated systems using TTS or VC technologies. In high-stakes scenarios like visa applications or educational admissions, spoofing could manipulate speech to sound more proficient, compromising the test's integrity. Research on non-native speech spoofing is essential to develop robust countermeasures, ensuring the fairness and reliability of systems that evaluate language proficiency and identity verification across diverse linguistic backgrounds.

This study focuses on Indonesia and Thailand due to Indonesia's large multicultural population and Thailand's unique dialect variability and tonal representation. Given the lack of publicly available datasets for these linguistic contexts, we developed an audio dataset with speech from Indonesian and Thai non-native English speakers. Our goal is to establish a spoofing countermeasure (CM) to effectively detect TTS and VC attacks for non-native speech accents and patterns.

The contribution of our work is summarized as follows:

- to facilitate the development of CMs to handle non-native speech by creating a dataset with non-native speech, addressing the gap in available speech datasets comprising non-native accents and characteristics,
- to understand the effectiveness on Native CMs in handling Asian non-native speech, and
- to propose CMs that can distinguish bonafide and spoof speech for both natives and non-natives.

In this study, we created a comprehensive dataset consisting of English native and Indonesian-Thai non-native speech to construct the CMs. Furthermore, we investigated how the CMs performed in distinguishing bonafide and spoof speech for both native and non-native speakers. Our findings show that the proposed method significantly improved CM performance in handling non-native speech. Additionally, our CMs utilize common front-end features and back-end classic machine learning (ML)-based classifiers, establishing a foundational baseline for this non-native speech study.

II. Related Works

In response to the growing threat of spoofing, the research community launched the ASVspoof challenges[1], a biennial initiative aimed at developing CMs to detect fake speech, starting in 2015 [4] and continuing through 2024 [5]. As these types of attacks have increased through the challenge editions, TTS and VC attacks have been included since the first edition in 2015.

Several CM systems have been developed to detect audio spoofing, primarily focusing on hand-crafted features that can effectively capture discriminative patterns of artifacts [6]. Early research mainly used conventional classifiers, such as ML-based classifiers.

A study on synthetic speech detection in 2015 found that dynamic Linear Frequency Cepstral Coefficients (LFCC) with a GMM classifier performed best on ASVspoof evaluation sets [7]. In ASVspoof 2017 [8], the best system in the previous challenge was used as the baseline system. It was built using a common GMM back-end classifier with constant Q cepstral coefficients (CQCC). In ASVspoof 2019 [1], the

[1]https://www.asvspoof.org/

top-performing systems employed LFCC features with a light convolutional neural network (LCNN). However, other systems based on standard cepstral features and GMM-based classifiers were only slightly behind in performance. In ASVspoof 2021 [9], most systems operated using short-term spectral features or raw waveforms, utilizing ensemble systems and popular convolutional networks. These efforts primarily addressed spoofing in native language domains, leaving the non-native speech domain largely unexplored. Speech corpora are essential for CM systems, and publicly available datasets with a large number of spoofing attacks have emerged to overcome data bottlenecks. These datasets, which include both bonafide and spoof speech generated by LA algorithms (TTS and/or VC), facilitate the evaluation and benchmarking of different systems. Notable datasets include ASVspoof2015 [4], ASVspoof2019-LA [10], ASVspoof2021-LA [9], and the latest ASVspoof2024 [5]. Other available datasets include ADD2022-LF [3] and Latin-American Voice Anti-spoofing [2]. However, datasets representing non-native English speech are scarce, highlighting the need for research focused on spoofing detection tailored to non-native accents.

While some non-native speech datasets exist [11], [12], such as those designed for automatic speech recognition (ASR) and language learning, they are not specifically tailored for spoofing detection tasks. These datasets typically focus on improving speech recognition for non-native speakers or assessing language proficiency rather than detecting synthetic or manipulated speech. The absence of non-native speech datasets for spoofing detection further underscores the importance of developing and expanding resources to support more effective countermeasures in diverse linguistic environments.

III. Dataset Collection

The dataset used in this study consists of English native and non-native speech, spoken by Indonesian and Thai speakers. The dataset is split into training, validation, and testing sets with the ratio of 70-10-20 to ensure robust evaluation, with no overlap of speakers between the splits to maintain the integrity of the results. Table I summarizes our ENIT Dataset[2].

A. Bonafide Data

Our native English dataset consists of 7,990 utterances sourced from the training sets of ASVspoof 5 [5], which is derived from the English-language subset of the Multilingual Librispeech (MLS) dataset. The MLS dataset is a large multilingual corpus based on LibriVox audiobooks [13], featuring non-studio-quality recordings. This dataset includes recordings from over 4,000 speakers using various devices. As this dataset has a balanced number of speakers, we randomly selected 4,000 utterances from the total of 80 males and 80 females speakers and eliminated 10 utterances randomly to ensure the balance with our non-native English dataset.

The English non-native speech dataset consisted of 7,990 utterances recorded from 21 speakers, including 10 Indonesian speakers (7 males and 3 females) and 11 Thai speakers (7 males and 4 females). These speakers read articles sourced from online English newsletters covering unbiased topics such as health, astronomy, engineering, and technology. We selected similar news topics in Indonesian, English, and Thai, ensuring they did not express hate speech, violence, or harassment to maintain impartiality. Each news article was then segmented into sentences or sub-sentences with 5 to 20 words per utterance to ensure readability.

TABLE I
ENIT Dataset: English Native and Indonesian-Thai Non-Native Speech

Type	Sets	# unique spk.		# utterances		# spoof
		Male	Female	Bonafide	Spoof	attack
Native	Train.	56	56	5,590	33,540	8
	Dev.	8	8	800	4,800	8
	Eval.	16	16	1,600	9,600	8
Non-native	Train.	10 / 12	4 / 5	5,696	33,164	3
	Dev.	1 / 1	1 / 2	542	4,275	3
	Eval.	3 / 3	2 / 3	1,752	10,501	3

The recordings were conducted in a soundproof room using four different recording devices: a condenser microphone, a dynamic microphone, a mobile phone, and a laptop computer, without significant channel or background noise effects. All speech was recorded in a read style, with the speaking pace adjusted to each speaker, while still maintaining the naturalness of spontaneous speech. Audio post-processing included practical trim voice editing using Audacity[3] to separate the recordings by utterances.

B. Spoof Data

We constructed a corpus of 47,940 utterances by selecting a subset from the ASVspoof 5 training data [5]. The selected utterances are from the same 80 males and 80 females speakers used in the bonafide data. The number of spoof utterances is six times larger than the number of bonafide utterances. Table I specifies the number of unique speakers of bonafide and spoof speech on the left and right side of the separator '/', respectively. If the number of speakers is the same for both bonafide and spoof speech, we do not use the separator.

To generate spoof data for non-native English speech, we employed TTS, VC, and synthesis techniques. These methods can be broadly categorized into three distinct approaches.

1) SpeechT5 [14]: SpeechT5 is a model that unifies speech and text processing under a single framework. Inspired by the success of T5 in the text domain [15], SpeechT5 employs an encoder-decoder architecture to learn shared representations for both modalities. It incorporates specialized pre- and post-processing networks to handle the unique characteristics of speech and text, respectively.

Through extensive self-supervised pre-training on large-scale unlabeled speech and text data, SpeechT5 acquires a deep

[2]ENIT Dataset: English Native and Indonesian-Thai Non-Native Speech

[3]Audacity (ver. 3.4.2) is an open source software for recording and editing sounds.

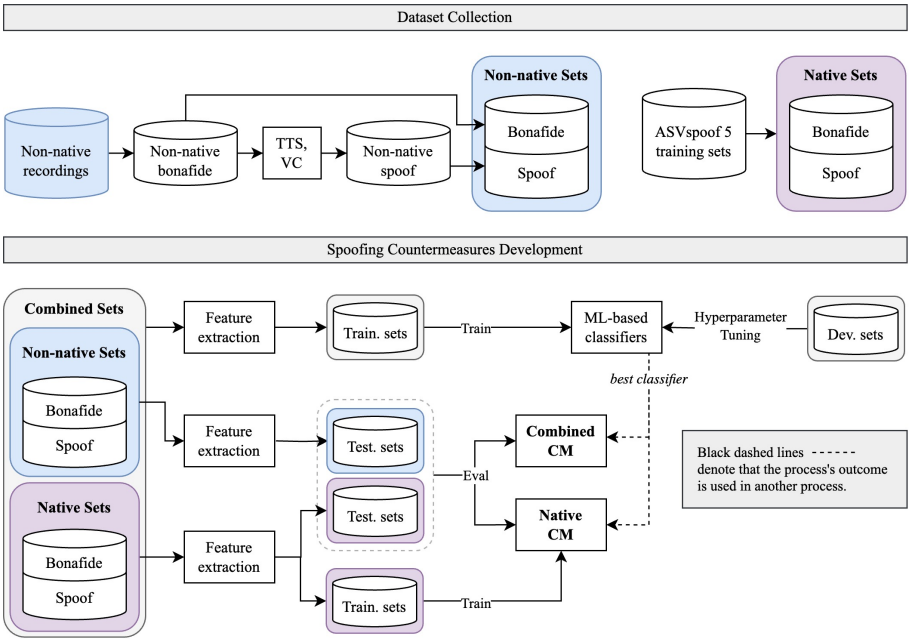


Fig. 1. Block diagrams of our proposed method for building spoofing countermeasures

understanding of both modalities. This versatility empowers it to excel in a wide range of tasks, including automatic speech recognition, speech synthesis, speech translation, VC, speech enhancement, and speaker identification.

Approximately 5,200 synthetic utterances were generated using the SpeechT5 model and x-vector speaker embeddings [16]. To manipulate acoustic and linguistic characteristics, we varied both x-vector and bottleneck features (representing linguistic information). Roughly 2,000 utterances were created using x-vector and bottleneck features extracted from our collected bonafide non-native speaker data. In another set of 1,000 utterances, x-vectors from native speakers in the CMU ARCTIC dataset [17] were combined with bottleneck features from our non-native speakers. The remaining utterances were synthesized by pairing x-vectors from our non-native speakers with bottleneck features from the VCTK dataset [18].

2) FreeVC [19]: We utilized FreeVC, a text-free VC system capable of transforming a speaker's voice with only a single reference audio sample. The model builds upon the ViTS architecture [20] but departs from traditional text-based approaches by learning to disentangle content information directly from raw audio waveforms. FreeVC extracts speaker-independent features that are subsequently processed through a bottleneck layer to isolate content representations using WavLM [21]. To further enhance the model's ability to separate content from speaker identity, spectrogram resizing is employed as a data augmentation technique.

Central to the one-shot conversion capability is a speaker encoder that extracts speaker-specific characteristics. We experiment with two configurations of the speaker encoder: FreeVC-s, which utilizes a non-pretrained encoder, and FreeVC, which benefits from a pre-trained encoder. This comparative analysis allows us to assess the impact of pre-training on the overall performance of the VC system.

3) WORLD [22]: We also employed the WORLD, a vocoder-based speech synthesis system, to generate synthetic spoof data. This system is widely recognized for its exceptional performance in speech analysis, manipulation, and synthesis. It analyzes speech into three parts: fundamental frequency (F0), aperiodicity, and spectral envelope.

To further refine the spectral envelope estimation process, we integrated the CheapTrick algorithm [23] into our pipeline. This method leverages F0-adaptive windowing to enhance spectral resolution. By smoothing the power spectrum and applying spectral recovery techniques in the quefrency domain, CheapTrick better estimates the spectral envelope. This improved accuracy contributes significantly to the overall quality of the synthesized speech. We manipulated speed and F0 with random parameters during the speech synthesis process. This randomization helps to create a more diverse and challenging dataset for subsequent anti-spoofing models.

IV. Proposed Method

A block diagram of our proposed method in building the spoofing CMs using our datasets is shown in Figure 1. To establish a foundation for our newly developed dataset, which represents the characteristics of non-native speech, we employed hand-crafted features and classic ML-based classifiers.

A. Feature extraction

The front-end features are mostly derived from the magnitude or power spectrum, encompassing both short-term and long-term magnitude spectral features [6], to effectively capture the essential characteristics of speech signals, including formant structures and energy distribution across different

frequency bands. These features are typically obtained from MFCC [24], LFCC [25], and CQCC [26].

LFCC is widely used in speaker recognition and has demonstrated strong performance in spoofing detection. MFCC has been extensively explored for accent classification tasks, which closely relate to this study on non-native speech. CQCC has also proven effective in detecting audio spoofing by providing higher frequency resolution at lower frequencies and higher temporal resolution at higher frequencies. This study aims to investigate the effectiveness of LFCC, MFCC, and CQCC in detecting spoofing attacks in non-native English speech, focusing on Indonesian and Thai accented speakers.

B. Classifiers

We use classic ML-based classifiers to build the baseline spoofing CMs. To determine the most effective classifier, we conducted model selection on several widely used ML-based classifiers and evaluated their performance on our development sets. On the basis of our experiments, we identified the two best classifiers for the features used (MFCC, LFCC, and CQCC): CatBoost [27] and XGBoost [28].

CatBoost is a type of binary decision trees-based gradient-boosting predictor that has proven effective in dealing with categorical data [29]. Categorical data consists of distinct values that cannot be compared directly. In this study, we deal with binary categorical data.

Extreme Gradient Boosting, or XGBoost, is a gradient-boosted tree algorithm for supervised learning. Gradient-boosted decision trees excel in learning from noisy data and have achieved cutting-edge results. XGBoost is a gradient-boosting implementation that optimizes computing performance to provide a scalable solution.

Additionally, we included the GMM-based classifier as our baseline, which has been a fundamental method in the ASVspoof challenge series. GMM assumes that data points belong to a mixture of a finite Gaussian distribution [30]. It is commonly used in anti-spoofing systems. There, separate GMMs are learned for each bonafide and spoof dataset. The classification of new input is predicted by calculating the ratio of its log-likelihood of belonging to bonafide and spoof GMMs.

V. Experimental Settings

As the currently available speech corpora primarily consist of native speech, we aim to assess the performance of CMs built upon these datasets in handling non-native speech. We refer to this CM as Native CM. Subsequently, we developed a spoofing CM specifically for non-native speech by combining the current publicly available datasets, which are predominantly native speech, with our non-native datasets. We refer to this CM as Combined CM. Our experiment pipeline is illustrated in Fig. 1.

A. Experiment 1: Assessing the Native CM on Non-Native Speech

First, we built the Native CM by training our selected models (GMM, CatBoost, and XGBoost) using the extracted features (MFCC, LFCC, and CQCC) on the native speech training sets. The features were extracted using the Matlab toolbox and Smileslab implementations [31]. We set the random seed to 42 for all classifiers and adjusted the hyperparameters as detailed in Subsection V-C. Each model was then evaluated on both native and non-native speech evaluation sets, using the specified evaluation metrics. This initial assessment on native speech helps us understand the model's performance in a typical native speech context.

TABLE II
CMs performance evaluation result in Experiments 1 and 2

Exp.	Feat.	Classifier	Non-native		Native	
			minDCF	EER (%)	minDCF	EER (%)
1	MFCC	CatBoost	0.81	41.54	**0.17**	**6.88**
		XGBoost	0.83	42.58	0.18	7.48
		GMM	1.00	100	1.00	100
	LFCC	CatBoost	0.95	40.07	0.24	10.19
		XGBoost	0.98	40.36	0.26	10.56
		GMM	1.00	62.84	0.83	33.89
	CQCC	CatBoost	0.84	38.11	0.29	12.43
		XGBoost	0.79	35.57	0.28	12.56
		GMM	1.00	76.02	0.97	47.69
2	MFCC	CatBoost	0.33	12.66	0.19	7.81
		XGBoost	0.38	13.98	0.21	8.08
		GMM	0.96	46.58	1.00	46.18
	LFCC	CatBoost	0.27	10.90	0.31	13.32
		XGBoost	0.25	10.34	0.33	13.76
		GMM	0.98	53.31	1.00	46.13
	CQCC	CatBoost	**0.19**	**8.56**	0.37	15.06
		XGBoost	0.21	9.63	0.36	14.75
		GMM	0.97	41.55	0.99	91.89

B. Experiment 2: Assessing the Combined CM on Non-Native Speech

With regard to the Combined CM, we trained our selected models (GMM, CatBoost, and XGBoost) using the extracted features (MFCC, LFCC, and CQCC) on the combined native and non-native speech training sets. We set the same hyperparameter tuning as experiment 1 to ensure fair comparison. We also used the same evaluation protocol to facilitate the common assessment and benchmarking of both CMs.

C. Hyperparameter settings

We set different hyperparameters for each classifier on the basis of its architecture as implemented in the scikit-learn library [30]. For the GMM classifier, we used two mixture components, a full covariance type, the k-means method for initializing the weights, and a maximum of 100 iterations. For CatBoost, we set the verbose parameter to 0. For XGBoost, we used logarithmic loss as the evaluation metric. All other hyperparameters were kept at their default configurations.

D. Evaluation Metrics

In this study, we used minimum detection cost function (minDCF) and equal error rate (EER) adopted from the latest ASVspoof 5 challenge, specifically on track 1: stand-alone spoofing and speech deepfake detection [5]. The spoofing detection employed in the challenge is built upon the normalized detection cost function (DCF), defined as follows:

$$\mathrm{DCF}'(\tau_{\mathrm{cm}}) = \beta \cdot P_{\mathrm{cm,\ miss}}(\tau_{\mathrm{cm}}) + P_{\mathrm{cm,\ fa}}(\tau_{\mathrm{cm}}) \qquad (1)$$

where

$$\beta = \frac{C_{\text{miss}} \cdot (1 - \pi_{\text{spf}})}{C_{\text{fa}} \cdot \pi_{\text{spf}}} \quad (2)$$

In Eq. (1), $P_{\text{cm, miss}}(\tau_{\text{cm}})$ and $P_{\text{cm, fa}}(\tau_{\text{cm}})$ are the empirical miss rate for bonafide utterances and false alarms for spoof utterances, respectively. Both are regarded as a function of the detection threshold τ_{cm}. The constant β In Eq. (2) is calculated using miss rate cost C_{miss}, false alarm cost C_{fa}, and detection threshold τ_{cm} with the values set as follows: $C_{\text{miss}} = 1$, $C_{\text{fa}} = 10$, and $\pi_{\text{spf}} = 0.05$. Therefore, the value of β is approximately 1.90.

Furthermore, Eqs. (1) and (2) are used to compute the minimum DCF (minDCF), which measures the CMs performance by using the threshold set on the basis of ground-truth where the lower value indicates a better performance. This metric is calculated as follows:

$$\text{minDCF} = \min_{\tau_{\text{cm}}} \text{DCF}'(\tau_{\text{cm}}) \quad (3)$$

The final evaluation metric used is the EER, one of most widely used metrics for audio spoofing CMs. It represents the CM threshold where the false acceptance rate equals the false rejection rate. A lower EER value signifies better performance.

VI. Results

As seen in Table II, the GMM classifier has comparably high minDCF and EER values on the native speech evaluation sets. It also reached the maximum possible DCF and EER values of 100% on MFCC features, indicating extremely poor performance as it is entirely ineffective at distinguishing between bonafide and spoof speech in both non-native and native contexts. The lowest EER for the GMM classifier was achieved with LFCC features, which is expected as this system is considered the baseline in the ASVspoof challenge.

While the CatBoost and XGBoost classifiers achieved competitive results on native datasets, they still obtained minDCF values greater than 0.79 and EER values higher than 38% on non-native datasets. All the CMs in experiment 1 on average declined around 0.44 in minDCF and 19% in EER relative to their performance on native datasets. These results indicate a significant lack of non-native speech handling capability in Native CMs, demonstrating the inefficacy of the pre-existing speech spoofing CMs in distinguishing between bonafide and spoof speech in non-native contexts.

In experiment 2, all Combined CMs significantly improved when predicting the non-native datasets, with an average relative improvement of around 0.41 in minDCF and around 30% in EER. However, this experimental setting slightly worsened the CMs' ability to detect native speech, as indicated by an average of total increase on minDCF and EER around 0.06 and 1.7%, respectively. This behavior may be due to the increased variability introduced by the non-native speech data within the training sets, making it more challenging for the model to generalize well to the specific characteristics of native speech alone, as it now has to account for a broader range of speech patterns.

Among all the CMs we built, CQCC feature extraction with CatBoost classifier performed the best when evaluated on non-native datasets. These ensemble models tended to perform better due to their boosting algorithm, which iteratively builds an ensemble by training each new model to correct the errors of the previous ones.

Meanwhile, the GMM classifier, which is often used as a baseline model, performed worse than other classifiers. In most cases, the GMM classifier predicted the input speech as a spoof. Despite reducing EER by up to 50% after being trained using the combined native and non-native datasets, its performance as a CM is not satisfactory.

VII. Conclusions

In this study, we constructed spoofing countermeasures (CMs) to handle non-native speech using different acoustic features (MFCC, LFCC, and CQCC) and classic machine learning-based classifiers (CatBoost, XGBoost, and GMM). We also developed the ENIT dataset, which includes speech from Indonesian and Thai non-native English speakers. We conducted two experiments to assess the current CMs (Native CM) in handling non-native speech, as well as the baseline we constructed (Combined CM).

Our findings show that the Combined CMs is better at detecting bonafide and spoof non-native speech than the Native CMs, improving minDCF and EER scores around 0.41 and 30% on average, respectively. Moreover, we also found that CatBoost and XGBoost obtained competitive results using all three features, while GMM performed the worst in all experiments. The future work aims to expand datasets for Asian non-native speech and develop robust CMs using deep learning, focusing on accent and speech characteristics.

Acknowledgments

This work was supported by a Grant-in-Aid for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233), Grant-in-Aid for Challenging Research (Exploratory) (23K18491), Grant-in-Aid for Challenging Research (23K18491), and the SCAT Foundation. This work was a part of the ASEAN IVO project titled 'Spoof Detection for Automatic Speaker Verification'(www.nict.go.jp/en/asean_ivo).

References

[1] A. Nautsch, X. Wang, N. Evans, *et al.*, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," *IEEE Trans. on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, Apr. 2021, ISSN: 2637-6407.

[2] P. A. Tamayo Flórez, R. Manrique, and B. Pereira Nunes, "HABLA: A Dataset of Latin American Spanish Accents for Voice Anti-spoofing," in *Proc. of INTERSPEECH 2023*, 2023, pp. 1963–1967.

[3] J. Yi, R. Fu, J. Tao, *et al.*, "ADD 2022: the first Audio Deep Synthesis Detection Challenge," in *Proc. of ICASSP 2022*, IEEE, 2022, pp. 9216–9220.

[4] Z. Wu, T. Kinnunen, N. Evans, *et al.*, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. of INTERSPEECH2015*, ISCA, 2015, pp. 2037–2041.

[5] X. W. et al., "ASVspoof 5: Crowdsourced data, deepfakes and adversarial attacks at scale," in *ASVspoof 2024 workshop (submitted)*, 2024.

[6] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "Audio Anti-Spoofing Detection: A Survey," *CoRR*, 2024.

[7] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. of INTERSPEECH2015*, ISCA, 2015, pp. 2087–2091.

[8] T. Kinnunen, M. Sahidullah, H. Delgado, *et al.*, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Proc. of INTERSPEECH2017*, ISCA, 2017, pp. 2–6.

[9] J. Yamagishi, X. Wang, M. Todisco, *et al.*, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," *CoRR*, vol. abs/2109.00537, 2021.

[10] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, p. 101 114, 2020.

[11] J. Zhang, Z. Zhang, Y. Wang, *et al.*, "Speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *CoRR*, vol. abs/2104.01378, 2021.

[12] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, and R. Morton, "The ISLE corpus of non-native spoken English," Apr. 2000.

[13] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in *Proc. of INTERSPEECH2020*, ISCA, 2020, pp. 2757–2761.

[14] J. Ao, R. Wang, L. Zhou, *et al.*, "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing," in *Proc. of ACL 2022*, Association for Computational Linguistics, 2022, pp. 5723–5738.

[15] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, 140:1–140:67, 2020.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. of ICASSP 2018*, IEEE, 2018, pp. 5329–5333.

[17] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA ITRW on Speech Synthesis, Pittsburgh, PA, USA, June 14-16, 2004*, A. W. Black and K. A. Lenzo, Eds., ISCA, 2004, pp. 223–224.

[18] J. Yamagishi, C. Veaux, and K. MacDonald, *CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)*, 2019.

[19] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion," in *Proc. of ICASSP 2023*, IEEE, 2023, pp. 1–5.

[20] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. of ICML 2021*, vol. 139, PMLR, 2021, pp. 5530–5540.

[21] S. Chen, C. Wang, Z. Chen, *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[22] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99-D, no. 7, pp. 1877–1884, 2016.

[23] M. Morise, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1–7, 2015.

[24] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 28, no. 4, pp. 357–366, 1980.

[25] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

[26] H. Tak, J. Patino, A. Nautsch, N. W. D. Evans, and M. Todisco, "An Explainability Study of the Constant Q Cepstral Coefficient Spoofing Countermeasure for Automatic Speaker Verification," in *Proc of Odyssey 2020*, K. Lee, T. Koshinaka, and K. Shinoda, Eds., ISCA, 2020, pp. 333–340.

[27] L. O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Proc. of NeurIPS 2018*, 2018, pp. 6639–6649.

[28] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. of ACM SIGKDD 2016*, ser. KDD '16, ACM, Aug. 2016.

[29] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," *CoRR*, vol. abs/1810.11363, 2018.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[31] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Voice spoofing countermeasures: Taxonomy, state-of-the-art, experimental analysis of generalizability, open challenges, and the way forward," *CoRR*, vol. abs/2210.00417, 2022.

Two-stage Framework for Robust Speech Emotion Recognition Using Target Speaker Extraction in Human Speech Noise Conditions

Jinyi Mi, Xiaohan Shi, Ding Ma, Jiajun He, Takuya Fujimura and Tomoki Toda
Nagoya University, Japan
E-mail: {mi.jinyi, xiaohan.shi, ding.ma, jiajun.he, fujimura.takuya}@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

***Abstract*—Developing a robust speech emotion recognition (SER) system in noisy conditions faces challenges posed by different noise properties. Most previous studies have not considered the impact of human speech noise, thus limiting the application scope of SER. In this paper, we propose a novel two-stage framework for the problem by cascading target speaker extraction (TSE) method and SER. We first train a TSE model to extract the speech of target speaker from a mixture. Then, in the second stage, we utilize the extracted speech for SER training. Additionally, we explore a joint training of TSE and SER models in the second stage. Our developed system achieves a 14.33% improvement in unweighted accuracy (UA) compared to a baseline without using TSE method, demonstrating the effectiveness of our framework in mitigating the impact of human speech noise. Moreover, we conduct experiments considering speaker gender, showing that our framework performs particularly well in different-gender mixture.**

I. Introduction

Speech is a significant part of human communication. Besides linguistic information, it contains unique paralinguistic information such as gender, emotion, and age, which is essential to the normal communication. In certain instances, misunderstanding paralinguistic features would distort the correct information conveyed by speech, leading to an ineffective communication. Therefore, it is necessary to develop human-like communication machines that can comprehend paralinguistic data.

Speech emotion recognition (SER), as a branch of affective computing, has garnered growing attention over the past two decades because of its contribution to human-computer interactions [1], [2]. Generally, the mechanism of SER involves extracting and classifying effective emotional features from audio signals so that various emotions of a speaker can be captured, thanks to which SER has been applied in healthcare [3], [4], driver safety [5], [6], call center [7], [8], and online education [9], [10]. At present, research on SER systems on scenarios devoid of background noises, often referred to as clean scenarios, has shown good performance [11]–[13]. However, in real-world environments, the performance of SER significantly degrades, mainly due to the presence of various noises from different sources. These unknown noises severely affect the performance of SER systems, which poses major challenges for the widespread application of SER systems.

Several studies have focused on SER tasks in the environment affected by specific noise sources, including communication systems [14], [15], transportation [15], [16], and industrial activities [15]. In [14], Huang et al. studied SER from speech signals with additive white Gaussian noise, they proposed two speech enhancement methods based on spectral subtraction and masking properties, respectively. In [16], Chenchah et al. used power-normalized cepstral coefficients as acoustic features for improving the robustness of SER systems in noisy environment from cars and trains. In [15], Liu et al. proposed a multi-level knowledge distillation framework, which significantly reduced the affects of noises from channel, car, and factory.

However, the aforementioned studies mainly concentrate on addressing the noise sources associated with non-human activities, leaving a gap in addressing prevalent sources of noise in human-centric environments. While Shi et al. [17] did adopt ASR representations to filter out a specific category of noise related to human activities, which is typically from human physical actions like knocking on doors, a more common category of noise stemming from human speech itself remains underexplored. This type of noise called human speech noise, which is common in activities involving human interactions such as social gatherings, often becomes entangled with target speech data, forming a complex acoustic environment. Therefore, human speech noise becomes more unpredictable and more challenging to address.

On the other hand, humans have an extraordinary ability to selectively concentrate on a single speaker among a complex acoustic environment, commonly called cocktail party effect [18]. To replicate this specialized listening ability in machines, target speaker extraction (TSE) technique has been developed. This technique exploits an auxiliary information of the target speaker and extracts speech of that speaker from the mixture. Švec et al. [19] explored the potential of TSE for extracting target emotional speech. In light of this, we propose a novel two-stage framework by cascading TSE method and SER to mitigate the impact of human speech noise. In the experiments, we utilize the TSE–SER framework to ShiftCNN [12] that is a state-of-the-art SER model, and compare its performance to verify the effectiveness of our framework. Furthermore, we investigate different factors on the performance of SER, includ-

ing training methods and speaker genders. Our contributions to this work are as follows:

- We apply the TSE method on SER tasks to address the practical scenario in which target speech is interfered by human speech noises. To our knowledge, this study is the pioneering effort in exploring the integration of TSE with SER.
- We propose a two-stage framework using different training methods. According to comparative experiments against the baselines without using TSE technique, our framework significantly improves the accuracy of SER in human speech noise conditions.
- We investigate the impact of human speech noise on our framework, especially on same-gender mixture and different-gender mixture. Results indicate that our framework performs better in different-gender mixture.

II. Proposed Method

A. Framework overview

We assume that an observed time-domain mixture signal in human speech noise conditions is defined as

$$\boldsymbol{y} = \boldsymbol{s}_0 + \sum_{i=1}^{I-1} \boldsymbol{s}_i, \tag{1}$$

where I is the number of speakers in the mixture, $\boldsymbol{s}_i$ for $i = 0, ..., I-1$ is the speech signal of the ith speaker. In particular, $i = 0$ indicates the speech signal of the target speaker. Directly using the mixture signal $\boldsymbol{y}$ corrupted by overlapping speakers for SER task, would cause the poor SER performance, as the SER model cannot selectively focus on a single speaker. Therefore, the goal of the proposed framework is to extract the target speaker signal $\boldsymbol{s}_0$ from the mixture signal $\boldsymbol{y}$ for SER training.

As illustrated in Figure 1, the framework contains two stages: First, the TSE model is trained using a large-scale mixed-speech corpus. This ensures a well-trained TSE model that can extract high-quality speech of the target speaker given the input speech and the enrollment utterance of that speaker. Note that we use a neutral speech utterance for enrollment, which is a more convenient and realistic setting than using corresponding emotional speech. Then, we apply the TSE model as the form of data augmentation to extract the exclusive speech information of the target speaker from a mixed emotional speech corpus. This denoised corpus is used for both training and testing in SER tasks. We denote this training method as TSE-SER-base. In addition, we further propose another training method called TSE-SER-ft shown in Figure 2. We start from the same TSE pretraining as the first stage. In the second stage, we introduce mixed emotional speech as input, simultaneously fine-tuning the pretrained TSE model and training the SER model. This joint training process not only refines the TSE system by adjusting its parameters but also benefits the SER training.

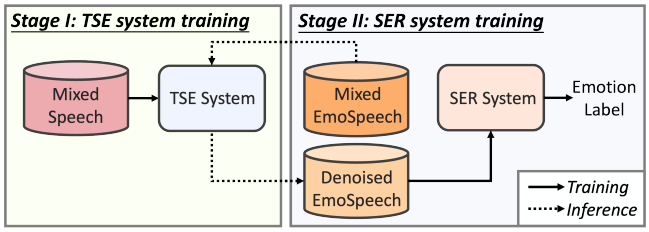


Fig. 1. Two-stage framework with TSE-SER-base.

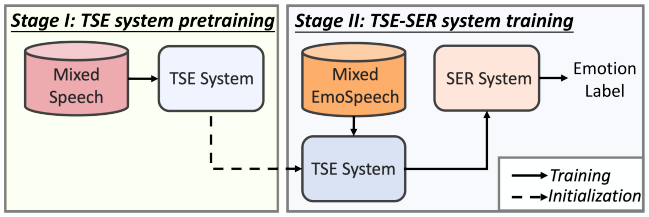


Fig. 2. Two-stage framework with TSE-SER-ft.

B. Target speaker extraction model

TSE refers to the task of reconstructing the speech signal of the target speaker from the mixture given auxiliary information of that speaker. This process can be formulated as

$$\hat{\boldsymbol{s}}_0 = g(\boldsymbol{y}, \boldsymbol{a}_0), \tag{2}$$

where $\hat{\boldsymbol{s}}_0$ is the estimated speech of the target speaker, $\boldsymbol{a}_0$ is the enrollment utterance of the target speaker, g represents the transformation carried out by the TSE system.

In this work, we adopt time-domain SpeakerBeam (TD-SpeakerBeam) [20], [21] as the TSE model. TD-SpeakerBeam consists of an auxiliary network and a speech extraction network, represented by h and f, respectively. The auxiliary network h accepts the enrollment utterance $\boldsymbol{a}_0$ and computes an embedding vector, denoted by $\boldsymbol{E}_0$, to represent the acoustic characteristics of the target speaker, i.e.,

$$\boldsymbol{E}_0 = h(\boldsymbol{a}_0). \tag{3}$$

Then, the speech extraction network f accepts the mixture signal $\boldsymbol{y}$ and the embedding vector $\boldsymbol{E}_0$ of the target speaker as inputs to predict the speech signal of the target speaker, i.e.,

$$\hat{\boldsymbol{s}}_0 = f(\boldsymbol{y}, \boldsymbol{E}_0), \tag{4}$$

where f comprises an encoder $\mathcal{E}$, a mask estimator $\mathcal{B}$, and a decoder $\mathcal{D}$. This process is formulated as

$$\boldsymbol{Y} = \mathcal{E}(\boldsymbol{y}), \tag{5}$$

$$\boldsymbol{M}_0 = \mathcal{B}(\boldsymbol{Y}, \boldsymbol{E}_0), \tag{6}$$

$$\hat{\boldsymbol{s}}_0 = \mathcal{D}(\boldsymbol{Y} \odot \boldsymbol{M}_0), \tag{7}$$

where $\odot$ denotes element-wise multiplication. The mixture signal $\boldsymbol{y}$ is fed into the encoder $\mathcal{E}$ that is represented by a 1D convolution layer. Then, the mask estimator $\mathcal{B}$ maps the output $\boldsymbol{Y}$ of the encoder $\mathcal{E}$ to a mask $\boldsymbol{M}_0$ for the target speaker, utilizing multiple convolution blocks. In particular, a

multiplicative adaptation layer, accepting the embedding vector $\boldsymbol{E}_0$ of the target speaker as auxiliary information, is inserted between the first and second blocks to drive the network towards extracting the target speaker. Finally, the mask $\boldsymbol{M}_0$ and the encoder features $\boldsymbol{Y}$ are fed into a 1D deconvolution layer-based decoder $\mathcal{D}$, to output the time-domain signal of the target speaker.

C. Loss functions

In the first stage, TSE-SER-base and TSE-SER-ft use scale-invariant source-to-noise ratio (SiSNR) [22] as the loss for TSE training. In the second stage, TSE-SER-base uses cross entropy (CE) loss for SER training, whereas TSE-SER-ft jointly trains the pretrained TSE model and SER model, considering both SiSNR and CE losses. The second stage losses of TSE-SER-base (L_{base}) and TSE-SER-ft (L_{ft}) are represented as

$$L_{base} = L_{CE}, \tag{8}$$

$$L_{ft} = L_{SiSNR} + L_{CE}. \tag{9}$$

III. Experimental evaluations

A. Datasets

In this work, we designed two kinds of datasets for comparable experiments: 1) the clean emotional dataset, and 2) the emotional dataset mixed with human speech noise. For the latter, two different human speech datasets were used as noise. All the mentioned datasets were sampled at 16 kHz.

IEMOCAP: The Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus [23], consisting of approximately 12 hours of recordings, includes five dyadic sessions, each with one English male speaker and one English female speaker. For our experiments, we used IEMOCAP as the clean emotional dataset, where we considered only four emotional categories of happy, angry, sad, and neutral. Note that "excited" was merged with "happy" to ensure category balance [13], [24]–[27].

LibriSpeech: The LibriSpeech corpus [28] contains about 1000 hours of read English speech. For our experiments, 105 hours of this corpus were chosen as a source of human speech noise.

ESD: The Emotional Speech Database (ESD) corpus [29] comprises about 29 hours of recordings from 10 English speakers and 10 Chinese speakers. For our experiments, we considered only the English part as another source of human speech noise.

In order to more accurately evaluate the performance of the proposed framework, we adopted leave-one-session-out 5-fold cross-validation to test all the models. Note that we used the following terms to represent different datasets designed: *Clean* means a clean set, *Noisy* means a dataset mixed with human speech noise, and *Denoised* indicates a dataset denoised from *Noisy* by the pretrained TSE model.

B. Experimental procedure

The first experiment investigated the impact of human speech noise on SER (see Section III-D1). The noisy dataset was generated by randomly selecting utterances from different speakers in LibriSpeech and mixing them with clean data of IEMOCAP. The number of speakers in the mixture of speech was limited to two.

The second and third experiments explored the effect of TSE method on SER and the proposed training methods on our framework (see Sections III-D2 and III-D3). We used the same dataset as the first experiment for the second stage of TSE-SER-base and TSE-SER-ft. We adopted LibriMix [30] where 100 hours of LibriSpeech were additionally used to generate mixtures for TSE pretraining. Furthermore, for the target speaker in the mixture, we randomly chose one neutral utterance of this speaker that does not belong to the mixture as the enrollment utterance.

The fourth experiment explored the impact of gender states of mixtures on SER (see Section III-D4). We used the same clean speech from the first experiment and used ESD as noise. We generated two types of mixtures: same-gender mixture and different-gender mixture, where the first type was stipulated that two speakers have the same genders, while the latter type required two speakers of opposite genders.

TABLE I
DEFINITION OF THE MODELS USED IN EXPERIMENTS.

Short Name	Model	Method
SB	TD-SpeakerBeam [21]	-
SC	ShiftCNN [12]	-
SB_SC	TD-SpeakerBeam + ShiftCNN	TSE-SER-base
SB+SC	TD-SpeakerBeam + ShiftCNN	TSE-SER-ft

C. Implementation and metrics

To build TSE model, we followed an open-source SpeakerBeam implementation[1]. For SER model, we used ShiftCNN [12], which has shown advanced performance in clean environments, adopting the same hyperparameters as [12]. All implemented models in experiments are shown in Table I.

For evaluation metrics of SER, we used the unweighted accuracy (UA) and weighted accuracy (WA). UA was the mean of the accuracies for each individual class while WA represented the ratio of correctly predicted samples to the total number of samples. In addition, we used scale-invariant signal-to-distortion ratio (SI-SDR) and scale-invariant signal-to-distortion ratio improvement (SI-SDRi) to evaluate the performance of TSE model.

D. Experimental results

To conduct a comparative study of all the experiments, aside from the proposed TSE-SER-base and TSE-SER-ft, we built two typical SER baselines using ShiftCNN, which were directly trained on *Clean* and *Noisy*, referred to as clean SER model and noisy SER model, respectively.

[1]https://github.com/BUTSpeechFIT/speakerbeam

TABLE II
PERFORMANCE OF SER IN CLEAN AND HUMAN SPEECH NOISE CONDITIONS.

Model	Train Set	Test Set	WA (%)	UA (%)
SC	Clean	Clean	**70.20**	**71.64**
	Clean	Noisy	47.11	46.02
	Noisy	Noisy	53.71	53.94

TABLE III
COMPARISON WITH TSE-SER-BASE AND THE SER BASELINES.

Model	Train Set	Test Set	WA (%)	UA (%)
SC	Noisy	Noisy	53.71	53.94
	Clean	Denoised	51.12	49.50
SB_SC	Denoised	Denoised	**62.20**	**63.42**

1) The impact of human speech noise on SER: To clarify the impact of human speech noise on SER, we compare the clean SER model and the noisy SER model. As shown in Table II, the clean SER model obtains an accuracy of over 70% for both UA and WA in the clean test set. But their performance drops significantly on the noisy test data, with a maximum decrease of up to 23.09% and 25.62% in terms of WA and UA. These results demonstrate that the clean SER model is fragile against human speech noise. Furthermore, when the SER model uses noisy data for training, the relatively better performance can be observed. Nonetheless, the results are still significantly worse than those of the clean SER model on the clean test set, suggesting deficient robustness. We argue that human speech noise severely hinders the SER model from establishing an effective mapping to the target emotional speech with the direct training.

2) The effect of TSE method on SER: To verify the effectiveness of the proposed framework on SER, we compare the performance of the SER model using the TSE method with those of the SER baselines. As presented in Table III, we first observe that the results of the clean SER model on denoised test set are not ideal, demonstrating that the clean model is unable to adapt to the denoised speech with distorted properties. We hence design the corresponding system trained on denoised data using TSE-SER-base, referred to as SB_SC in Table III. Our proposed system SB_SC performs significantly better than the other systems. Especially, the UA reaches 63.42%, which is a 9.48% increase compared to the noisy SER model in noisy conditions. Meanwhile, for the unbalanced training data from IEMOCAP corpus, the WA result is also competitive, closely resembling the UA results. The overall results demonstrate that the proposed framework with TSE-SER-base adapts well to human speech noise, showcasing effectiveness and robustness.

TABLE IV
COMPARISON WITH TSE-SER-BASE AND TSE-SER-FT.

Model	Method	WA (%)	UA (%)
SB_SC	TSE-SER-base	62.20	63.42
SB+SC	TSE-SER-ft	**67.02**	**68.27**

TABLE V
COMPARISON WITH TSE-SER-BASE AND THE SER BASELINES ON SAME- AND DIFFERENT-GENDER MIXTURES.

Model	Train Set	Test Set	Gender State	WA (%)	UA (%)
SC	Clean	Noisy	Same	48.00	46.50
	Noisy	Noisy		54.67	54.84
	Clean	Denoised		45.09	43.43
SB_SC	Denoised	Denoised	Same	**55.09**	**55.95**
SC	Clean	Noisy	Different	46.62	44.60
	Noisy	Noisy		54.29	54.20
	Clean	Denoised		48.87	47.35
SB_SC	Denoised	Denoised	Different	**59.75**	**61.32**

3) The effect of training methods on TSE–SER framework: We compare the performance of TSE-SER-base and TSE-SER-ft. As indicated in Table IV, TSE-SER-ft significantly outperforms TSE-SER-base. Moreover, the UA of TSE-SER-ft reaches a 14.33% improvement compared with the noisy SER model presented in Table III. The SI-SDR of the noisy speech before being processed by the TSE model is 0.09 dB. We calculate the SI-SDRi for the corresponding TSE models of TSE-SER-base and TSE-SER-ft to be 7.68 dB and 12.90 dB, respectively, verifying that TSE-SER-ft can improve TSE performance. Therefore, the TSE fine-tuning, applied to the noisy dataset containing the task-specific emotional data, enables the TSE model to extract purer emotional-related acoustic features to benefit the SER training.

4) The impact of gender states of mixtures on SER: Table V shows the results of TSE-SER-base, the SER baselines on same- and different-gender mixtures. First, the clean SER models unsurprisingly gain the lowest performance. Moreover, the noisy SER model shows a non-obvious trend on the noisy test set across both gender states, whereas the clean SER model in same-gender mixture clearly outperforms that in different-gender mixture, demonstrating better adaptability of SER models to same-gender mixture. In addition, besides the results of TSE-SER-base being all better than those of other systems for both same- and different-gender mixtures, we can see a significant performance gap of TSE-SER-base for dealing with same- and different-gender mixtures, which is opposite to the finding for the clean SER model in noisy conditions. Since our framework uses the TSE model, we conjecture that same-

TABLE VI
ABLATION STUDY FOR TSE-SER-BASE. "SAME" AND "DIFFERENT" INDICATE GENDER STATES.

Model	Train set	Test set	WA (%)	UA (%)
SB_SC	Denoised (Same)	Denoised (Same)	55.09	55.95
		Denoised (Different)	59.12	60.46
	Denoised (Different)	Denoised (Same)	55.32	56.40
		Denoised (Different)	59.75	61.32

gender mixture with the similar acoustic characteristics is more difficult for TSE model to separate. Before being processed by the TSE model, the SI-SDR of same- and different-gender mixtures are 0 dB and 0.02 dB. We calculate the SI-SDRi for TSE model to be 1.09 dB and 5.22 dB for same- and different-gender mixtures, respectively. This also explains why the clean SER model on the denoised test set gives better results in different-gender mixture.

We further conduct an ablation study for TSE-SER-base using two SB_SC systems from Table V on test sets denoised from the same- and different-gender mixtures. We have an interesting finding in Table VI that using training data denoised from different-gender mixture can enhance our model performance on both test sets denoised from same- and different-gender mixtures. We argue that the pretrained TSE model can extract higher-quality denoised data from different-gender mixture, thus finalizing a higher-performance SER model.

IV. CONCLUSION

In this work, we presented a two-stage framework to mitigate the impact of human speech noise on SER. Based on the framework, we designed two training methods, TSE-SER-base and TSE-SER-ft. The effectiveness and robustness of both methods have been verified. Moreover, we investigated the impact of human speech noise on SER, especially on same- and different-gender mixtures. In the future, we plan to explore how our proposed framework performs on human speech noise with different attributes, such as emotion classes and languages. Another possible direction involves multi-interference environments, such as noisy and reverberant environments.

ACKNOWLEDGMENT

This work was partly supported by JST CREST Grant Number JPMJCR19A3, Japan, and JSPS KAKENHI Grant Number 21H05054. In addition, this work was also financially supported by JST SPRING, Grant Number JPMJSP2125. The author would like to take this opportunity to thank the "THERS Make New Standards Program for the Next Generation Researchers."

REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] X. Shi, S. Li, and J. Dang, "Dimensional emotion prediction based on interactive context in conversation.," in *INTERSPEECH*, 2020, pp. 4193–4197.

[3] M. Z. Uddin and E. G. Nilsson, "Emotion recognition using speech and neural structured learning to facilitate edge intelligence," *Engineering Applications of Artificial Intelligence*, vol. 94, p. 103 775, 2020.

[4] M. S. Hossain and G. Muhammad, "Cloud-assisted speech and face recognition framework for health monitoring," *Mobile Networks and Applications*, vol. 20, pp. 391–399, 2015.

[5] M. Grimm, K. Kroschel, H. Harris, *et al.*, "On the necessity and feasibility of detecting a driver's emotional state while driving," in *International Conference on Affective Computing and Intelligent Interaction*, 2007, pp. 126–138.

[6] A. Tawari and M. Trivedi, "Speech based emotion classification framework for driver assistance system," in *2010 IEEE Intelligent Vehicles Symposium*, 2010, pp. 174–178.

[7] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of artificial neural networks in engineering*, vol. 710, 1999, p. 22.

[8] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech communication*, vol. 49, no. 2, pp. 98–112, 2007.

[9] W. Li, Y. Zhang, and Y. Fu, "Speech emotion recognition in e-learning system based on affective computing," in *Third international conference on natural computation (ICNC 2007)*, vol. 5, 2007, pp. 809–813.

[10] A. Tickle, S. Raghu, and M. Elshaw, "Emotional recognition from the speech signal for a virtual education agent," in *Journal of Physics: Conference Series*, vol. 450, 2013, p. 012 053.

[11] Y. Liu, H. Sun, W. Guan, *et al.*, "A discriminative feature representation method based on cascaded attention network with adversarial strategy for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1063–1074, 2023.

[12] S. Shen, F. Liu, and A. Zhou, "Mingling or misalignment? temporal shift for speech emotion recognition with pre-trained representations," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] L. W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] C. Huang, C. Guoming, Y. Hua, B. Yongqiang, and Z. Li, "Speech emotion recognition under white noise," *Archives of Acoustics*, vol. 38, no. 4, pp. 457–463, 2013.

[15] Y. Liu, H. Sun, G. Chen, *et al.*, "Multi-level knowledge distillation for speech emotion recognition in noisy conditions," *arXiv preprint arXiv:2312.13556*, 2023.

[16] F. Chenchah and Z. Lachiri, "A bio-inspired emotion recognition system under real-life conditions," *Applied Acoustics*, vol. 115, pp. 6–14, 2017.

[17] X. Shi, J. He, X. Li, and T. Toda, "On the effectiveness of asr representations in real-world noisy speech emotion recognition," *arXiv preprint arXiv:2311.07093*, 2023.

[18] S. Handel, *Listening: An introduction to the perception of auditory events*. MIT Press, 1993.

[19] J. Švec, K. Žmolíková, M. Kocour, *et al.*, "Analysis of impact of emotions on target speech extraction and speech separation," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2022, pp. 1–5.

[20] K. Žmolíková, M. Delcroix, K. Kinoshita, *et al.*, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[21] M. Delcroix, T. Ochiai, K. Zmolikova, *et al.*, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 691–695.

[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[23] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[24] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7367–7371.

[25] L. Guo, L. Wang, C. Xu, J. Dang, E. S. Chng, and H. Li, "Representation learning with spectro-temporal-channel attention for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6304–6308.

[26] X. Shi, X. Li, and T. Toda, "Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation," in *Proc. Interspeech*, 2023, pp. 765–769.

[27] X. Shi, X. Li, and T. Toda, "Multimodal fusion of music theory-inspired and self-supervised representations for improved emotion recognition," pp. 2024–2350, 2024.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 5206–5210.

[29] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[30] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

Data generation for speaker diarization by speaker transition information

Keigo Ichikawa, Sei Ueno, Akinobu Lee
Nagoya Institute of Technology, Japan
E-mail: k.ichikawa.667@stn.nitech.ac.jp, {sei.ueno, ri}@nitech.ac.jp

***Abstract*—A speaker-turn-aware training data generation method for end-to-end neural diarization (EEND) has been studied. To compensate the lack of real multi-talker training corpus, using artificial multi-speaker conversation data generated from several single speaker's speech has been a promising way to train EEND models. It is still expected that making the training data close to real multi-talker situation can improve EEND performance. In order to generate a data that contributes to the better performance of trained EEND model, we investigated two speaker alternation handling methods. One is a data-dependent model that introduces a probability of speaker transition extracted from multi-talker data. The other one is a data-independent model that uniformly samples the speaker-alternation probability. Experimental results have shown that the data-dependent method improves the DER as compared with baseline simulated conversations (SC) method on 3-speaker and 4-speaker scenarios. Moreover, it was also confirmed that the data-independent method can achieve comparable performance on the best model on 3 and 4-speaker scenarios.**

I. Introduction

Speaker diarization is a task that detects "who spoke when" from multi-talker audio. Speaker diarization outputs are used in the preprocessing step for multi-talker tasks such as source separation and multi-talker speech recognition [1]. In order to perform speaker diarization, a clustering-based model or an end-to-end model can be used. A clustering-based model [2]–[7] is a model that performs a cascade process. In general, it performs speech activity detection, segmentation, speaker feature extraction, and clustering to distinguish the speaker of the input audio.

On the other hand, an end-to-end neural diarization (EEND) model has been investigated [8]–[10]. It directly predicts whether each speaker is talking in a single step, frame by frame. Due to the development of the EEND model, it outperforms the clustering-based model. However, for training the EEND model, a lot of annotated multi-talker data is required. Moreover, it is too costly to prepare the annotated data since we need to label when and how long each person speaks. Therefore, the data generation from the speech data of a single speaker is generally applied for training the EEND model at a lower cost [8], [11]. The simulated data are generated as follows: (1) Preparing source data containing multiple speakers' speeches with each utterance involving only a single speaker's speech. (2) Choosing 2, 3, or 4 speakers and extracting utterances of the chosen speaker from source data. (3) Generating multi-talker speech through concatenating the utterances.

The generated data needs to resemble real data to improve EEND performance. Several approaches have investigated which elements should closely match real data and how to simulate it. Simulated mixtures (SM) [8], [11], used in the original EEND model, considered silence intervals between each speaker's utterances. Simulated conversations (SC) [12], [13] extended SM method and generated multi-talker speech using statistics about distributions of pauses and overlaps estimated on real conversations.

In this work, we also investigate the elements to improve the EEND model and focus on speaker alternation in the generated data. Speaker alternation occurs when one speaker finishes and another speaker begins to speak. In a real conversation, one speaker may speak a lot, while others speak less. To model this conversation, we introduce speaker transition probabilities. Speaker transition probabilities involve not only speaker alternation but also a case where one speaker begins to speak again. The SC method implicitly models the speaker transition probability, but it depends on the source data. We first remove the dependency and then investigate two approaches. One is a data-dependent approach that uses the probabilities extracted from statistics on real conversations. The other is a data-independent approach that uses uniform probabilities.

II. Related works

A. Simulated mixtures (SM)

The SM method [8], [11] is a data generation method for creating multi-talker audio. It involves overlapping the single-speaker audio to create multi-talker audio. Single-speaker audio is created by preparing speech segments from that speaker and concatenating the prepared speech segments with a certain silent interval between them. The silent intervals are determined by random numbers that follow an exponential probability distribution.

By varying the parameters of the exponential distribution, it is possible to adjust the silent intervals, and the overlapping intervals indirectly when multi-talker audio from multiple single-speaker speech is overlapped. It is known that the way overlapping sections are created with SM method may generate data that differs from actual multi-talker audio, the overlapping sections continue for a long time or there are many sections where all speakers overlap, which is not often seen in real conversations.

B. Simulated conversations (SC)

SC method [12], [13] creates multi-talker audio by connecting speech segments from multiple speakers, creating overlapping and silent sections. When connecting speech segments from the same speaker, a certain interval of silence between them is made. When connecting speech segments from different speakers, a certain interval of silence or overlaps between them is made. The length of this overlapping section or silent section is determined by randomly selecting one value from the silent section lengths and overlapping section lengths collected from actual audio data. The choice of whether to create silent intervals or overlapping intervals when different speakers are speaking successively is also determined based on the distribution obtained from actual speech data. The order in which these speakers speak is randomized in simulated conversations.

In simulated mixtures, single-speaker audio is created first, then multi-talker audio is created, but in simulated conversations, multi-talker audio is created from the beginning. This makes it possible to control silent sections and overlapping sections in multi-talker audio more directly. Simulated conversations suppress the generation of speech that is not often seen in actual conversations, such as when there are long overlapping sections or when there are many sections where all speakers overlap. By making the simulated data closer to real data, the performance of EEND is improved.

III. Speaker Transition Probability

The frequency of speaker alternation depends on the domain of a real conversation. For example, speaker alternation in telephone communication is less than in standard face-to-face conversation. We expect that controlling speaker alternation yields an improvement in the EEND model. To model the speaker alternation, we introduce the speaker transition probability. We additionally consider the last speaker and the next speaker, and the probability $P(\text{next speaker}|\text{last speaker})$ of speaker alternation occurring using a first-order Markov model. We also consider P when the next speaker is the same as the last speaker. In total, there are S^2 patterns, where S denotes the number of speakers in a single audio file. Fig. 1 shows an example in a 3-speaker scenario. When the last speaker is speaker A, we need to calculate $P(A|A)$, $P(B|A)$, and $P(C|A)$. We follow the SC method to model other aspects such as overlap and silence ratios.

A. Speaker transition probability on other methods

The speaker transition probability is not determined in the SM method because it involves concatenating speeches from a single speaker and then from multiple speakers. In the SC method, the speaker transition probability is implicitly decided by the source data. It prepares all utterances of the selected speaker and randomly samples from the candidates. Therefore, the speaker transition probability depends on the number of each speaker's candidates of source data. Specifically, in 3-speaker scenario, we consider $P_{SC}(A)$, $P_{SC}(B)$, and $P_{SC}(C)$.

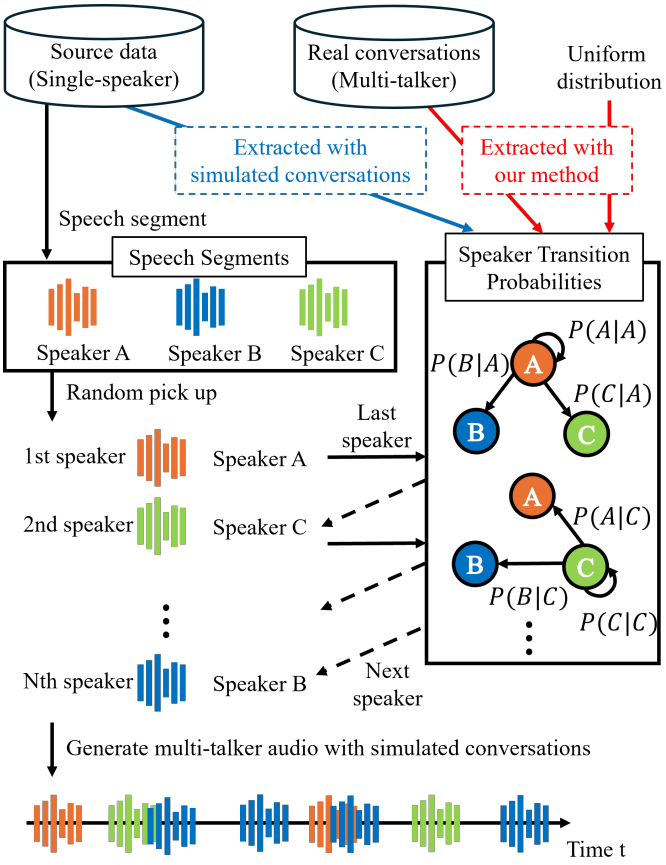


Fig. 1. Overview: Data generation of multi-talker audio using speaker transition probability. Simulated conversations extract speaker transition probabilities from source data. Our method extracts speaker transition probabilities from real conversations or specifies them as equal probabilities.

Each probability is determined as follows:

$$P_{SC}(A) = m_a/M, P_{SC}(B) = m_b/M, P_{SC}(C) = m_c/M$$

where, m_* denotes the number of candidates extracted from speaker-$*$ speeches, and $M = m_a + m_b + m_c$. Because each speech is used only once, $m_* = m_* - 1$ once speaker-$*$ is sampled. The number of candidates is based on the source data. Occasionally, P_{SC} can be biased, which is not expected.

B. Data-dependent approach

As a data-dependent approach, we extract the probability from the real multi-talker data. We calculate the speaker transition probability for each conversation using the multi-talker data. We prepare lists of speaker transition probabilities. In the data generation step, we randomly choose one of the lists and generate a multi-talker speech following the extracted speaker transition probability. To identify speakers A, B, ..., and N, we follow the order in which each speaker speaks in the real conversation.

C. Data-independent approach

As a data-independent approach, we use a uniform probability. All probabilities are $1/N_{spk}$. In the data-dependent

TABLE I
COMPARISON OF THE NUMBER OF AUDIO FILES FROM REAL CONVERSATIONS DATASET.

Dataset	2-speaker	3-speaker	4-speaker
DIHARD 3	157	16	16
CALLHOME 2	148	74	20
AMI	0	4	129
CALLHOME 1 (oracle)	155	61	23

approach, we need to prepare real conversations to extract the speaker transition probability. However, the public dataset provides a limited number of utterances (detailed in TABLE I), which may cause over-fitting. Therefore, we remove the dependency on speaker alternation and analyze how this affects the performance of the EEND model. This data-independent method helps remove the dependency on source data in SC methods.

IV. EXPERIMENTAL EVALUATION

A. Dataset and speaker transition probabilities

For data generation, we used LibriSpeech [14]. LibriSpeech is a speech corpus consisting of 961 hours of single-speaker recordings from 2,338 speakers. Statistical information to control overlap and silence ratio was obtained from the DIHARD 3 dev set [15]. The DIHARD 3 dev set is a speech corpus consisting of 34 hours of multi-talker conversations involving 548 speakers. For the evaluation, we used the 2, 3, and 4-speaker audio from CALLHOME 1, a subset of the 2000 NIST Speaker Recognition Evaluation (CALLHOME) corpus [16]. We divided CALLHOME into CALLHOME 1 and CALLHOME 2 using the Kaldi toolkit [17]. The duration of CALLHOME 1 and CALLHOME 2 are as follows: 2-speaker audio has 3.2 hours and 3.0 hours, 3-speaker audio has 2.1 hours and 3.0 hours, and 4-speaker audio has 2.2 hours and 1.8 hours, respectively. All audio was used at 8kHz.

In the data-dependent approach, we extracted the speaker transition probability from the DIHARD 3 dev set, CALLHOME 1/2, and AMI train set [18]. CALLHOME 1 was used as an oracle. TABLE I shows the number of audio files for each dataset. For the 2-speaker, 3-speaker, and 4-speaker scenarios, DIHARD 3 and CALLHOME 1 and 2 were used. To prevent overfitting, we only used datasets containing more than 30 audio files, and we mixed the datasets when the dataset provided a limited number of audio files. In the data-independent setting, the value of each speaker transition probability was set to 0.25 for the 4-speaker scenario, 0.33 for the 3-speaker scenario, and 0.50 for the 2-speaker scenario as equal probabilities.

B. End-to-end neural diarization

EEND [8], [11] is a model approach that achieves end-to-end speaker diarization by framing the task as a multi-label classification problem, which classifies the presence of multiple speakers within each frame. This model usually uses a transformer encoder, and given acoustic features of length T with D dimensions $(\boldsymbol{x}_t \mid \boldsymbol{x}_t \in \mathbb{R}^D)_{t=1}^T$, the EEND model generates outputs as described in Equation (1).

$$\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T = \mathrm{f}_{\mathrm{EEND}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T) \tag{1}$$

Here, $\boldsymbol{y}_t = [y_{t,1}, \cdots, y_{t,S}]^T \in (0,1)^S$ denotes the probability of speech presence for S speakers in frame t.

C. Experimental settings

In all settings, we generated 2,480.6 hours of audio. We generated audio according to the number of speakers in the evaluation set. Impulse responses were not used, but background noises obtained from the MUSAN corpus [19] were used, and the SNR was set to one of 5, 10, 15, or 20 dB. The input features for the EEND model were 23-dimensional log Mel-filterbanks with a frame length of 25 milliseconds and a frameshift of 10 milliseconds. The EEND model consists of four transformer encoder blocks, each with 4 heads and 256 attention units. The dropout rate was also set to 0.1, and the position-wise feedforward layer had 1024 internal units. The optimizer used was Adam with a noam scheduler, and the warmup period was set to 25,000 steps. The batch size during training was set to 64, and the model was trained for 100 epochs. After 100 epochs of training, averaging the weights was performed on the models from epochs 91 to 100, and this averaged model was used for evaluation.

During inference, the threshold for outputting speech segments was set to 0.5, and a median filter of size 12 frames was applied before determining the speech segments. The diarization error rate (DER) defined by NIST [20] was used as the evaluation metric. DER is broken down into missed speech (Miss), false alarm (False), and speaker confusion (Conf.). During evaluation, errors of less than 250 milliseconds were tolerated.

V. RESULTS AND DISCUSSION

A. Comparison of end-to-end neural diarization performance with 2 and 3-speaker

TABLE II shows the DER for 2-speaker and 3-speaker settings and TABLE III also shows speaker alternation occurrence rates. In the 2-speaker setting, the data-dependent approach using CALLHOME 2 achieved comparable performance compared with the SC method. While all methods improved missed speech, they degraded false alarm and speaker confusion compared with the SC method. All methods caused more speaker alternations than the SC method as shown in TABLE III. We consider that it positively affects missed speech. However, it worsens false alarm and speaker confusion because the 2-speaker scenario is a relatively easy task.

In the 3-speaker setting, we observed that our methods yielded improvement in CALLHOME 2 + DIHARD 3. Different from the 2-speaker scenario, our methods improved speaker confusion. Because the number of speakers increased, detecting speakers in the 3-speaker scenario is a harder task than the 2-speaker scenario. However, the data-dependent approach using CALLHOME 2 worsened the DER. It drastically improved the false alarm, but degraded missed speech.

TABLE II
DER (%) FOR 2-SPEAKER AND 3-SPEAKER SETTING. DER IS DECOMPOSED INTO MISS (MISSED SPEECH), FALSE (FALSE ALARM), AND CONF. (SPEAKER CONFUSION).

#spk	Speaker transition	Miss	False	Conf.	DER
2	SC [baseline]	8.2	3.6	4.4	16.2
	Data-dependent				
	DIHARD 3	7.4	5.4	6.3	19.0
	CALLHOME 2	6.5	5.0	4.6	**16.1**
	CALLHOME 2 + DIHARD 3	7.5	5.3	5.2	17.9
	CALLHOME 1 (oracle)	7.7	3.4	5.8	16.9
	Data-independent (uniform)	7.1	5.0	5.6	17.7
3	SC [baseline]	11.2	6.1	12.5	29.7
	Data-dependent				
	CALLHOME 2	17.4	2.9	12.2	32.6
	CALLHOME 2 + DIHARD 3	10.9	6.5	11.6	**29.1**
	CALLHOME 1 (oracle)	11.8	5.8	11.0	**28.6**
	Data-independent (uniform)	10.8	6.8	10.8	**28.5**

TABLE III
COMPARISON OF SPEAKER ALTERNATION OCCURRENCE RATES (%) IN GENERATED DATA USING DIFFERENT SPEAKER TRANSITION PROBABILITIES FOR EACH NUMBER OF SPEAKERS

Speaker transition \ #spk	2	3	4
SC [baseline]	45.9	63.0	71.7
Data-dependent			
DIHARD 3	58.8	-	-
CALLHOME 2	75.5	81.4	-
CALLHOME 2 + DIHARD 3	66.9	77.3	74.7
AMI	-	-	76.4
AMI + CALLHOME 2	-	-	77.3
AMI + DIHARD 3	-	-	75.5
AMI + CALLHOME 2 + DIHARD 3	-	-	76.3
CALLHOME 1 (oracle)	78.9	81.8	83.1
Data-independent (uniform)	50.1	66.8	75.1
CALLHOME 1 (eval)	81.6	83.1	83.3

The speaker alternation occurrence rate was similar to the evaluation data, but the limited number of samples may have led to overfitting. In the CALLHOME 2 + DIHARD 3 setting, the overfitting is alleviated. The data-independent approach also improved the performance compared with the SC method and achieved comparable performance to the data-dependent approach using CALLHOME 1 (oracle). This result implies that the balance of the number of the speaker's utterances is one of the keys to the EEND model to train the speaker transition.

B. Comparison of end-to-end neural diarization performance with 4-sepaker

TABLE IV shows the DER for 4-speaker setting. We observed that our methods yielded improvement in AMI and AMI + CALLHOME 2 + DIHARD 3. Similarly to the 3-speaker setting, our methods improved speaker confusion. We also observed that our method yielded improvement in CALLHOME 2 + DIHARD 3. Specifically, it did not improve speaker confusion, but it improved false alarm. Using AMI + CALLHOME 2 and AMI + DIHARD 3 improved false alarm compared to using AMI only. Similar to this, we consider that using CALLHOME 2 + DIHARD 3 improved false alarm. But using AMI + CALLHOME 2 and AMI + DIHARD 3 degraded DER. Because the number of AMI is larger, the dataset ratio bias in AMI + CALLHOME 2 and AMI + DIHARD 3 is bigger than in CALLHOME 2 + DIHARD 3 (2-speaker and 3-speaker). We consider that the dataset ratio bias negatively impacts performance.

TABLE IV
DER (%) FOR 4-SPEAKER SETTING. DER IS DECOMPOSED INTO MISS (MISSED SPEECH), FALSE (FALSE ALARM), AND CONF. (SPEAKER CONFUSION).

Speaker transition	Miss	False	Conf.	DER
SC [baseline]	8.2	7.2	17.8	33.2
Data-dependent				
AMI	8.5	7.6	16.4	**32.4**
AMI + CALLHOME 2	8.7	6.4	18.9	34.0
AMI + DIHARD 3	9.2	7.0	17.5	33.7
CALLHOME 2 + DIHARD 3	8.8	6.3	17.5	**32.7**
AMI + CALLHOME 2 + DIHARD 3	9.0	7.4	16.1	**32.4**
CALLHOME 1 (oracle)	7.7	8.2	17.6	33.5
Data-independent (uniform)	8.3	7.0	17.0	**32.3**

The data-dependent approach using CALLHOME1 (oracle) worsened the DER. The speaker alternation occurrence rate was most similar to the evaluation data in the 4-speaker scenario dataset, but the limited number of samples may lead to overfitting as well as using CALLHOME2 in the 3-speaker scenario.

The data-independent approach improved the performance compared SC method and achieved comparable performance to the data-dependent approach using AMI + CALLHOME 2 + DIHARD 3.

C. Discussion of data-dependency in speaker transition

In any speaker scenario, while the best DER of the data-dependent approach was better, the DERs of some data-dependent approaches were worse than the baseline method. This is because the speaker transition highly depends on the dataset. On the other hand, data-independent approaches are more robust than data-dependent methods and achieve the best performance in 3 and 4-speaker settings. The data-dependent approach may improve further, but we need to prepare more data of multi-talker data. We consider that the speaker transition itself can affect the DER performance and can be seen as a hyperparameter of the data generation.

VI. CONCLUSIONS

In this work, we extended the SC method of determining speaker alternation. The SC method was originally based on the speaker transition probabilities from source data. We introduced speaker transition probabilities and two methods to calculate them. One is the data-dependent approach which extracts the probabilities from real conversation data. The other is the data-independent approach that samples speaker alternation uniformly. In the experimental evaluations, we confirmed that our proposed two approaches were effective in 3- and 4-speaker scenarios. It suggests that speaker alternation in the generated multi-talker data is important to improve EEND models.

Future work will involve investigating whether similar trends can be observed using the same method on evaluation corpora

with a larger number of speakers and different tasks. Additionally, we will continue to explore the effects of introducing deeper speaker transition probabilities, such as those modeled by a second-order Markov process.

REFERENCES

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101 317, 2022.

[2] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[3] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2014, pp. 413–417.

[4] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 4930–4934.

[5] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6301–6305.

[6] Q. Li, F. L. Kreyssig, C. Zhang, and P. C. Woodland, "Discriminative neural clustering for speaker diarisation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 574–581.

[7] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: Theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, p. 101 254, 2022.

[8] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[9] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1493–1507, 2022.

[10] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7198–7202.

[11] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 296–303.

[12] F. Landini, A. Lozano-Diez, M. Diez, and L. Burget, "From simulated mixtures to simulated conversations as training data for end-to-end neural diarization," *arXiv preprint arXiv:2204.00890*, 2022.

[13] F. Landini, M. Diez, A. Lozano-Diez, and L. Burget, "Multi-speaker and wide-band simulated conversations as training data for end-to-end neural diarization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[15] N. Ryant, P. Singh, V. Krishnamohan, *et al.*, "The third dihard diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[16] *2000 nist speaker recognition evaluation*. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2001S97.

[17] *Kaldi callhome diarization*, Accessed: 2024-07-28, 2022. [Online]. Available: https://github.com/kaldi-asr/kaldi/tree/master/egs/callhome_diarization/v1.

[18] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The ami meeting corpus," in *Proc. International Conference on Methods and Techniques in Behavioral Research*, 2005, pp. 1–4.

[19] D. Snyder, G. Chen, and D. Povey, *MUSAN: A Music, Speech, and Noise Corpus*, arXiv:1510.08484v1, 2015. eprint: 1510.08484.

[20] NIST, *Rich transcription evaluation*, https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation, version: md-eval-v22.pl.

Capturing Dynamic Identity Features for Speaker-Adaptive Visual Speech Recognition

Sara Kashiwagi*, Keitaro Tanaka*, and Shigeo Morishima†
* Waseda University, Tokyo, Japan
E-mail: sara.kashiwagi@moegi.waseda.jp Tel/Fax: +81-3-5286-3510
† Waseda Research Institute for Science and Engineering, Tokyo, Japan

***Abstract*—This paper describes a multi-task learning method to improve speaker adaptation in visual speech recognition (VSR). VSR models are highly sensitive to variations in lip movements, resulting in degraded accuracy for speakers not encountered during training. A typical solution is to fine-tune pre-trained models with minimal data from the target speaker, but this often faces overfitting to the specific speech content of those samples. Effective speaker adaptation requires VSR models to learn both static and dynamic aspects of individual lip features independently of speech content. However, the dynamic features are time-variant and intertwined with similarly time-variant content information. To address this issue, we introduce an additional task into the fine-tuning process that encourages the model to focus on acquiring disentangled dynamic identity features. Specifically, we apply temporal transformations to the latent visual representations and input them together with the original ones into a dynamic identity discriminator. The discriminator determines whether each entry is original or transformed, where both sets of representations share the same static identity features and speech content, thereby promoting the desired speaker adaptation. Our evaluation demonstrates that the proposed method improves recognition accuracy for all speakers across two different datasets: public online speech videos and our private recordings using a smartphone camera.**

I. Introduction

Visual speech recognition (VSR), also known as lipreading, aims to recognize speech content solely based on lip movements [1]. VSR has attracted attention as an alternative to traditional automatic speech recognition (ASR), particularly in noisy environments or situations where speaking aloud is impractical. Additionally, VSR can be a valuable communication tool for those with hearing or speech impairments. In practical applications, VSR has already been implemented in systems known as silent speech interfaces (SSIs) [2], which enable users to input text without vocalization. To ensure high accuracy in these systems, it is crucial to develop models that are tailored to individual users.

One of the major challenges in VSR is the decreased recognition accuracy for speakers not included in the training data. For example, the word error rate (WER) of a VSR model [3] is 18.0% for speakers seen during training but rises to 30.5% for unseen speakers. This performance drop occurs due to the model's sensitivity to the diverse lip movements of different individuals. The most effective approach to adapt VSR models to unseen speakers is to perform parameter re-optimization (i.e., fine-tuning) on the entire model using a substantial amount of data from these speakers. However, in real-world scenarios, gathering sufficient data for this fine-tuning is challenging, as manual data collection places a significant burden on users.

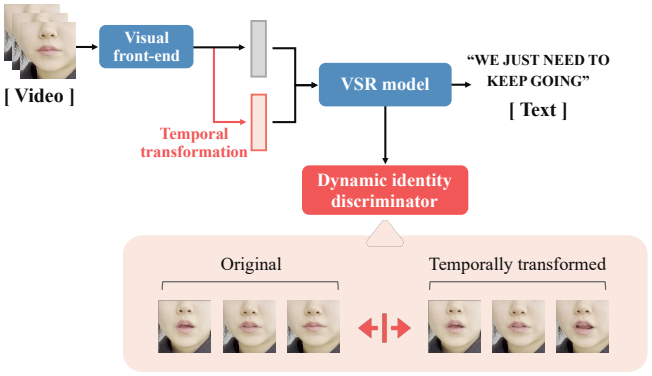


Fig. 1. Our multi-task learning approach employs temporal transformation and a dynamic identity discriminator to adapt VSR models to speaker-specific lip movements independently of speech content.

One remedy to this issue is applying parameter-efficient fine-tuning methods, particularly low-rank adaptation (LoRA) [4]. These methods leverage the extensive knowledge of pre-trained models and require fewer training samples compared to full parameter fine-tuning. While LoRA has proven effective in VSR [5] to some extent, it can lead to undesired learning behaviors when only minimal data from the target speaker is available. In speech videos, the speaker-dependent identity and speaker-independent content information are visually entangled, making it challenging to extract the information required for the target task [6]. Consequently, while LoRA can effectively capture the target speaker's identity with sufficient additional training samples, it tends to overfit the speech content of these samples under minimal data conditions.

In this paper, we present a novel multi-task learning approach that compels the VSR model to focus on the dynamic aspects of lip features (see Fig. 1). Generally, identity information includes two aspects: static (spatial) features related to lip appearance or shape and dynamic (temporal) features related to speaking speed or rhythm [5]. For effective speaker adaptation, the model needs to learn both features independently of speech content. Since the static identity features are time-invariant and unaffected by time-variant content information, the model can naturally acquire them during fine-tuning. Conversely, the dynamic identity features are time-variant and often emerge

alongside content information, such as the duration of pronunciation or the linking of sounds. Therefore, it is necessary to help the model acquire disentangled dynamic identity features in parallel with predicting speech content.

To perform effective speaker adaptation while preserving the robustness of pre-trained models against diverse speech content, we introduce an additional task into the fine-tuning process with generalized low-rank adaptation (GLoRA) [7]. Specifically, our method updates trainable low-rank decomposition matrices such that the model simultaneously recognizes dynamic identity features and produces accurate VSR results. The additional task incorporates temporal transformation and dynamic identity discrimination. The temporal transformation introduces local variations into the original dynamic identity features by randomly adding or removing frames of the latent visual representations with a certain probability. The dynamic identity discriminator receives the embeddings of both the original and transformed latent visual representations and determines whether each embedding entry is derived from the original or transformed representations. Since both sets of representations share the same static identity features and speech content, the model must focus on the target speaker's unique dynamic identity features to accomplish the additional task, thereby promoting the desired speaker adaptation.

Our multi-task learning approach for speaker adaptation prevents VSR models from overfitting speech content without requiring additional data collection. It facilitates efficient learning of speaker-specific lip features and promotes model specialization through targeted adjustments rather than relying on data augmentation or network architecture changes. As a result, our approach can be applied to various existing VSR models. Experimental results show that our method consistently improves accuracy for unseen speakers, both in a publicly available dataset and a custom dataset collected using a smartphone camera to simulate SSI scenarios. Notably, for some speakers, our proposed model trained with 80 samples even outperforms the ordinarily fine-tuned model with 100 samples.

II. Related Work

This section reviews existing research on model adaptation for VSR to handle speakers absent from the training data. We also explore the components of lip feature representation and their effects on speech recognition accuracy.

A. Speaker Adaptation in VSR

Speaker adaptation methods for VSR models have primarily focused on parameter re-optimization. Early research typically adjusted all model parameters using training data that included multiple target speakers. For example, some studies [6], [8] presented a multi-task learning approach that integrated the speaker identification task with VSR. Another approach by Kandala et al. [9] proposed utilizing speaker-specific latent identity vectors as additional inputs to the hidden layers of a VSR model. However, these approaches resulted in significant computational and data collection burdens. To address these issues, researchers have developed parameter-efficient fine-tuning methods that utilize samples from only a single target speaker. Examples include techniques such as input video transformation [10], prompt tuning [11], and low-rank adaptation (LoRA) [5]. These approaches enable speaker adaptation with a small amount of data by leveraging the knowledge embedded in existing pre-trained models.

Despite these advancements, there remains a significant lack of publicly available VSR datasets with adequate speaker labels. As a result, previous research on speaker adaptation has had to rely on datasets that share a limited vocabulary across the training, validation, and test data, such as the LRW dataset [12] and the GRID dataset [13]. Moreover, these studies have focused on classifying input videos into word classes present in the training data. Reflecting current progress in ASR, recent VSR models have shifted towards predicting sentences with unrestricted vocabularies through sub-word classification [14]. In this study, we advance speaker adaptation from word classification to sentence prediction by incorporating sub-word context through connectionist temporal classification (CTC) and attention mechanisms [15]. We also use two datasets with unrestricted vocabularies for our experiments: the publicly available VoxCeleb2 dataset [16] and a custom dataset developed to simulate real-world VSR applications.

B. Lip Feature Representation

Lip feature representation in VSR contains both identity and content information. This nature complicates the process of extracting and applying the information required for specific objectives, especially when the number of training samples is limited. For example, identity information can interfere with text prediction in VSR models, and content information is irrelevant for speaker adaptation. Existing studies on speaker adaptation have primarily addressed this issue by decoupling identity information from content information [6], [8].

Furthermore, He et al. [5] observed that identity information comprises both static (spatial) and dynamic (temporal) features. Static features are related to physiological aspects such as lip appearance and shape, while dynamic features are associated with behavioral aspects such as speaking speed and rhythm. In VSR, variations in lip opening widths caused by different voicing methods have led to performance gaps [17]. Specifically, larger lip openings in unvoiced speech resulted in an 8.5% decrease in accuracy compared to voiced speech, indicating that static identity features substantially affect VSR performance. Additionally, in both ASR and VSR, Pandey and Arif [18] demonstrated the relationship between speaking speed and recognition accuracy. Their results consistently show that native English speakers have lower accuracy than non-native speakers, attributed to their faster speaking speed. This performance decline is caused by shorter pronunciation durations and the linking of adjacent sounds, highlighting the importance of dynamic identity features in recognition tasks [19]. These findings suggest that adapting to both the static and dynamic identity features of the target speaker can contribute to improved recognition accuracy.

III. Proposed Method

In this study, we apply a parameter-efficient fine-tuning method to sub-word-based sentence prediction models. First, we explain the model architecture and the fine-tuning method. Then, an additional task is introduced to mitigate undesired overfitting to the content information.

A. End-to-End VSR Model

As the end-to-end model for predicting sub-word sequences from speech videos, we adopt a CTC/Attention hybrid model based on the conformer architecture [15], as shown in Fig. 2. This model employs an encoder-decoder architecture, with the encoder consisting of front-end and back-end networks. The input video is cropped to a 96×96 bounding box of the lip region using landmarks detected by MediaPipe [20]. The visual front-end comprises a 2D ResNet-18, where the initial layer is replaced with a 3D convolutional layer to extract both spatial and temporal features as latent visual representations from the input. The back-end is a 12-layer conformer encoder that learns temporal dependencies among the latent visual representations and outputs 768-dimensional embeddings per frame.

To leverage both CTC and attention mechanisms, the decoder is divided into two parallel fully connected (FC) layers. One processes the raw embeddings, while the other handles the 256-dimensional outputs from a 6-layer transformer decoder. The ground truth text labels consist of 5,000 sub-word classes generated by SentencePiece [21], and each FC layer produces a probability distribution over these classes. Finally, beam search with a width of 40 is applied to the weighted sum of the output probabilities from the CTC and attention branches, generating the predicted sub-word sequence, i.e., the sentence. The loss function $\mathcal{L}_{\mathrm{VSR}}$, which combines the CTC loss $p_{\mathrm{CTC}}(\mathbf{y}|\mathbf{x})$ and the cross-entropy loss $p_{\mathrm{CE}}(\mathbf{y}|\mathbf{x})$, is employed for training the model. $\mathcal{L}_{\mathrm{VSR}}$ is expressed as follows:

$$\mathcal{L}_{\mathrm{VSR}} = \alpha \log p_{\mathrm{CTC}}(\mathbf{y}|\mathbf{x}) + (1-\alpha) \log p_{\mathrm{CE}}(\mathbf{y}|\mathbf{x}), \quad (1)$$

where $\alpha = 0.1$ is used in the evaluation experiments.

B. Generalized Low-Rank Adaptation (GLoRA)

Given the above-mentioned VSR model pre-trained on large-scale datasets, we consider fine-tuning it with only minimal data from the target speaker. Specifically, we aim to facilitate the model's learning of speaker-specific lip features while preserving the linguistic knowledge of the pre-trained model. For this purpose, we apply generalized low-rank adaptation (GLoRA) [7] to the attention layers of the 12-layer conformer encoder, keeping all other parameters fixed.

GLoRA is an advanced version of LoRA [4], designed for universal parameter-efficient fine-tuning. Let $\boldsymbol{x} \in \mathbb{R}^{H\times 1}$ and $f : \mathbb{R}^{H\times 1} \to \mathbb{R}^{H\times 1}$ be the input and a linear transformation, where H denotes the input dimension. The unified equation representing all trainable parameters is described as follows:

$$f(\boldsymbol{x}) = (\boldsymbol{W}_0 + \boldsymbol{W}_0\boldsymbol{A} + \boldsymbol{B})\boldsymbol{x} + \boldsymbol{W}_0\boldsymbol{C} + \boldsymbol{b}_0 \odot \boldsymbol{D} + \boldsymbol{E} + \boldsymbol{b}_0, \quad (2)$$

where $\odot$ denotes the element-wise product. $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{H\times H}$, $\boldsymbol{C}, \boldsymbol{D}$ and $\boldsymbol{E} \in \mathbb{R}^{H\times 1}$ are the trainable parameters, while the pre-trained weights $\boldsymbol{W}_0 \in \mathbb{R}^{H\times H}$ and bias $\boldsymbol{b}_0 \in \mathbb{R}^{H\times 1}$ remain frozen throughout the fine-tuning process. $\boldsymbol{A}$ scales the pre-trained weight, decomposed into $\boldsymbol{A}_{\mathrm{d}} \in \mathbb{R}^{H\times r}$ and $\boldsymbol{A}_{\mathrm{u}} \in \mathbb{R}^{r\times H}$, where r represents the rank of LoRA. $\boldsymbol{B}$ scales the input and shifts the pre-trained weight, decomposed into $\boldsymbol{B}_{\mathrm{d}} \in \mathbb{R}^{H\times r}$ and $\boldsymbol{B}_{\mathrm{u}} \in \mathbb{R}^{r\times H}$. $\boldsymbol{C}$ acts as the layer-wise manipulation, decomposed into $\boldsymbol{C}_{\mathrm{d}} \in \mathbb{R}^{H\times r}$ and $\boldsymbol{C}_{\mathrm{u}} \in \mathbb{R}^{r\times 1}$. $\boldsymbol{D}$ and $\boldsymbol{E}$ scale and shift the pre-trained bias. We use a random Gaussian initialization for $\boldsymbol{A}_{\mathrm{u}}$, $\boldsymbol{B}_{\mathrm{u}}$, and $\boldsymbol{C}_{\mathrm{u}}$, and zero initialization for $\boldsymbol{A}_{\mathrm{d}}$, $\boldsymbol{B}_{\mathrm{d}}$, and $\boldsymbol{C}_{\mathrm{d}}$, ensuring $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are zeros at the beginning of the training, as well as $\boldsymbol{D}$ and $\boldsymbol{E}$.

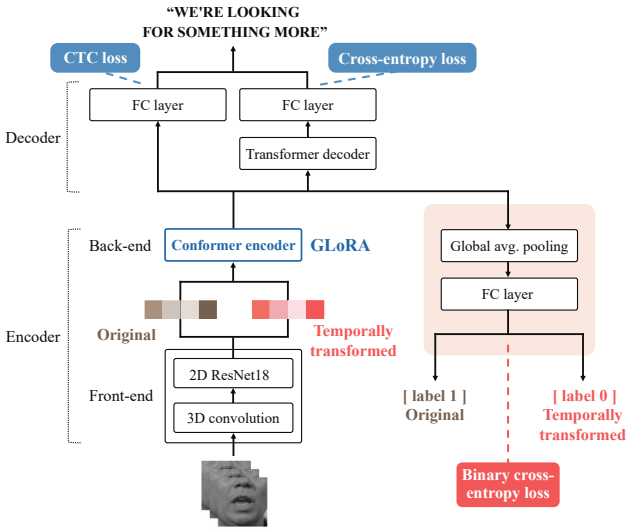


Fig. 2. The overall architecture of the proposed multi-task learning approach. We incorporate a discrimination task that distinguishes between the original and temporally transformed latent visual representations into the output from the conformer encoder, where GLoRA is applied.

C. Multi-Task Learning

Although the model specialized for individual speakers needs to learn both static and dynamic identity features independently of speech content, the dynamic identity features are time-variant and can overlap with similarly time-variant content information. To facilitate the model's learning of disentangled dynamic identity features in parallel with predicting speech content, we introduce an additional task into the fine-tuning process. This task involves temporal transformation and dynamic identity discrimination, as illustrated in Fig. 2.

The temporal transformation introduces local variations into the original dynamic identity features. As illustrated in Fig. 3, this is achieved by randomly adding or removing frames of the latent visual representations produced by the visual front-end, with a probability of 0.2. This process alters the sequence shape from $T \times 512$ to $T' \times 512$, where T and T' represent the number of frames before and after the transformation, respectively. Subsequently, the sequence is realigned to the original shape of $T \times 512$ by either zero-padding or cutting the end of the sequence. By using both the original and transformed versions of each sequence during training, the number of training sequences is doubled from N to $2N$, where N is the number of original data inputs.

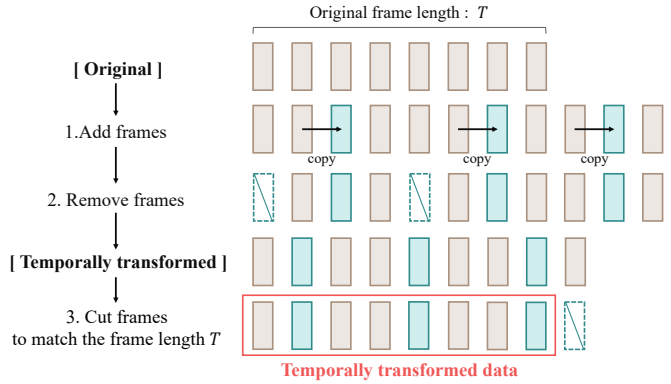


Fig. 3. Example of the temporal transformation process applied to the latent visual representations obtained from the visual front-end.

We then apply a discriminator to the output from the 12th layer of the conformer encoder to enable the model to simultaneously recognize dynamic identity features and generate accurate VSR results by updating GLoRA parameters. The discriminator consists of a global average pooling layer along the temporal axis and an FC layer. The global average pooling layer creates a 768-dimensional embedding for each sample to represent global information about the dynamic identity features. The FC layer takes this embedding and determines whether each embedding entry is derived from the original or transformed representations. Minimizing the binary cross-entropy loss $\mathcal{L}_{\mathrm{BCE}}$ compels the VSR model to focus on the target speaker's unique dynamic identity features, as both sets of representations share common static identity features and speech content. Given the ground truth labels y_i, where $y_i = 1$ for original data and $y_i = 0$ for transformed data, and the predicted probabilities $\hat{y}_i$, $\mathcal{L}_{\mathrm{BCE}}$ can be formulated as

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{2N}\sum_{i=1}^{2N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]. \quad (3)$$

The overall loss in our multi-task learning method is given by:

$$\mathcal{L} = \mathcal{L}_{\mathrm{VSR}} + \mathcal{L}_{\mathrm{BCE}}. \quad (4)$$

IV. Evaluation

This section describes experiments conducted to evaluate the performance of our proposed method for speaker adaptation.

A. Dataset

In our experiments, we used two datasets with unrestricted vocabularies: the VoxCeleb2 dataset [16] and a custom dataset. VoxCeleb2 is a publicly available dataset collected from online videos, originally designed for the speaker recognition task. Therefore, it includes speaker identity labels but lacks ground truth text labels. Here, we used its audio data to automatically generate transcriptions for the text labels, leveraging Whisper models [22]. For the speaker adaptation settings, we selected the five speakers with the most utterances from the test data (S1-S5). The speaker IDs and the size of the training, validation, and test sets per speaker are shown in Table I.

TABLE I
SPEAKER INFORMATION IN THE VOXCELEB2 DATASET

Speaker	ID (`folder_name`)	#Train	#Valid	#Test
S1	id05816 (`SCbhADNyfwg`)	70	30	105
S2	id07874 (`_3D5upGq24U`)	70	30	58
S3	id06692 (`9vs0zAHfI0M`)	56	24	62
S4	id08696 (`Cg9tYEiOCQA`)	56	24	54
S5	id04232 (`tCiPy0q5588`)	42	18	45

To gather the custom dataset, we developed a tailored app using a smartphone camera to simulate real-world VSR conditions. In the recording interface, the text to be read is displayed at the top of the screen, and video is recorded while a button at the bottom is pressed. Five participants (S6-S10), who are native or have equivalent English proficiency, took part in the data collection. We used 300 speech texts from the LRS3 test data [23], with each participant reading each text once. In this setup, S6-S10 all read the same texts, while S1-S5 in VoxCeleb2 each utter different sentences. Out of these 300 samples, 200 were fixed as the test set. To evaluate the impact of adaptation data quantity on performance, we created three different sets from the remaining 100 samples: 100 samples (70 for training and 30 for validation), 80 samples (56 and 24), and 60 samples (42 and 18). The experiments were conducted five times for each set, with a new random split of the training and validation data each time.

B. Training Settings

In this study, to leverage the knowledge from large-scale datasets, we utilized two pre-trained models provided by Ma et al. [24]. For the VoxCeleb2 evaluation, we used the model trained on 438 hours of the LRS3 dataset. For the custom dataset evaluation, we used the model trained on 1,759 hours of the combined LRS3 and VoxCeleb2 datasets. In all experiments, we trained the model for 100 epochs using the AdamW optimizer [25] with a learning rate of 1.0×10^{-5} and a maximum of 1,800 frames per batch. We thoroughly explored parameters and selected the configurations that achieved the best performance. For the rank of GLoRA, we tested $r = 4$ and 8 based on the findings of [26] and set it to 8. In the inference phase, the model was created by averaging the weights of the three best-performing checkpoints based on validation loss.

C. Evaluation Metrics

For evaluation, we used word error rate (WER) for the model outputs. We calculated WER as follows:

$$\mathrm{WER} = \frac{W_S + W_D + W_I}{W_N}, \quad (5)$$

where W_N is the total number of words in the ground truth. W_S, W_I, and W_D represent the number of words substituted for wrong classifications, inserted for those not to be present, and deleted for those to be present, respectively. A lower score indicates higher recognition accuracy.

D. Experimental Results

We conducted experiments with a total of 10 speakers using two different datasets to evaluate our proposed method. The

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR SPEAKERS IN THE VOXCELEB2 DATASET. THE BASELINE REFERS TO THE PRE-TRAINED MODEL, FT REPRESENTS FINE-TUNING WITH GLORA, AND MT DENOTES THE PROPOSED MULTI-TASK LEARNING METHOD

Methods	WER (%)				
	S1	S2	S3	S4	S5
Baseline	63.50	66.69	47.76	76.67	63.59
FT	61.74	61.75	44.54	74.76	64.10
+ MT	**61.17**	**57.55**	**44.19**	**73.95**	**60.78**

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR SPEAKERS IN THE CUSTOM DATASET

#Data	Methods	WER (%)				
		S6	S7	S8	S9	S10
0	Baseline	31.20	16.41	21.26	71.30	58.45
60	FT	30.32	16.90	20.39	68.55	56.78
60	+ MT	**29.09**	16.69	**20.02**	**64.49**	**55.88**
80	FT	28.89	15.72	20.45	60.60	53.18
80	+ MT	**27.31**	**15.51**	**19.64**	**60.11**	**52.45**
100	FT	28.23	15.59	19.81	59.79	52.21
100	+ MT	**26.27**	**15.07**	**18.85**	**58.29**	**52.14**

average WER over five experiments is presented in Table II for the VoxCeleb2 dataset and in Table III for the custom dataset. Table II shows that our proposed method improved accuracy for all speakers. Specifically, the average WER among the five speakers (S1-S5) decreased by 1.83% compared to the regular fine-tuning results and by 4.09% compared to the baseline through our multi-task learning method. For S5, although fine-tuning resulted in lower accuracy than the baseline, our approach effectively mitigated overfitting and achieved an almost 3% decrease in WER.

Table III compares the performance of the fine-tuning and multi-task learning methods across different sample sizes, showing that the multi-task learning method consistently outperformed the regular fine-tuning method. Using 100, 80, and 60 samples for training and validation, the average WER among the five speakers (S6-S10) decreased by 1.00%, 0.76%, and 1.35% compared to the fine-tuning results. The WER reductions compared to the baseline results were 5.60%, 4.72%, and 2.49%, respectively. Furthermore, for S6-S8, the multi-task learning method with 80 samples achieved better accuracy than the fine-tuning method even when using 100 samples. These results demonstrate that our approach promotes more efficient utilization of the target speaker data.

Next, we discuss the variations in recognition accuracy from two perspectives: environmental differences between the VoxCeleb2 and custom datasets and individual differences among speakers within the custom dataset. In the VoxCeleb2 evaluation, both the pre-trained dataset (LRS3) and the fine-tuning dataset (VoxCeleb2) consist of public speech sourced from the Internet. In contrast, the custom dataset evaluation involves significant environmental differences, as the fine-tuning dataset was collected using a smartphone camera in private settings, differing from the pre-trained datasets (LRS3 and VoxCeleb2). Our multi-task learning approach proves effective for both datasets, demonstrating robustness to various environments. Focusing on Table III, despite identical speech content across different speakers, notable performance variations were observed due to individual differences in lip movements. Additionally, speakers with poorer baseline performance exhibited greater accuracy improvements through the multi-task learning method. These results show that our approach can mitigate undesired learning behaviors and adapt the VSR model to the diverse lip features of individual speakers.

TABLE IV
ABLATION STUDY RESULTS FOR SPEAKERS IN THE CUSTOM DATASET. TT DENOTES THE TEMPORAL TRANSFORMATION PROCESS

Methods	WER (%)				
	S6	S7	S8	S9	S10
Baseline	31.20	16.41	21.26	71.30	58.45
FT	28.23	15.59	19.81	59.79	52.21
+ TT	27.12	16.11	19.59	59.29	53.13
+ TT + MT (6th layer)	**26.10**	15.21	18.90	58.69	52.37
+ TT + MT (12th layer)	26.27	**15.07**	**18.85**	**58.29**	**52.14**

Table IV presents the results of the ablation study conducted on the custom dataset using 100 samples for training and validation. We assessed accuracy changes under the following conditions: (1) applying only the temporal transformation process without the discriminator, (2) applying the discriminator to the output from the 6th layer of the conformer encoder, and (3) applying the discriminator to the output from the 12th layer of the conformer encoder (as shown in Table III). The effect of the temporal transformation varied among speakers. It improved accuracy for S6, S8, and S9 compared to the fine-tuning results, but S7 and S10 experienced a decline. For these latter speakers, using temporally transformed data without the discriminator might have introduced noise into the learning process of the target speaker's identity features. Applying the discriminator to the 6th layer consistently achieved higher accuracy than using only the temporal transformation process. There was no significant difference between applying the discriminator to the 6th and 12th layers; still, a slight improvement in accuracy was observed when applying it to the 12th layer, excluding S1.

V. CONCLUSION

We proposed a novel speaker adaptation method to improve accuracy for speakers not included in the training data using only minimal data. Specifically, we presented a multi-task learning approach to suppress overfitting to the specific speech content and promote efficient learning of disentangled dynamic identity features. Through extensive experiments with various speakers, we demonstrated that our proposed method consistently outperformed conventional fine-tuning methods across datasets from two different scenarios, even under challenging conditions. Our future work includes applying the proposed method to various existing VSR models and examining its effectiveness for speakers with more diverse lip features.

ACKNOWLEDGMENT

This work is supported in part by JSPS KAKENHI Nos. 21H05054, 24H00742, and 24H00748.

REFERENCES

[1] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.

[2] Z. Su, S. Fang, and J. Rekimoto, "LipLearner: Customizable silent speech interactions on mobile devices," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–21.

[3] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *International Conference on Learning Representations (ICLR)*, 2022.

[4] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "LoRA: Low-rank adaptation of large language models," *International Conference on Learning Representations (ICLR)*, 2022.

[5] Y. He, L. Yang, H. Wang, Y. Zhu, and S. Wang, "Speaker-adaptive lipreading via spatio-temporal information learning," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024, pp. 10 411–10 415.

[6] Q. Zhang, S. Wang, and G. Chen, "Speaker-independent lipreading by disentangled representation learning," in *IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2493–2497.

[7] A. Chavan, Z. Liu, D. Gupta, E. Xing, and Z. Shen, "One-for-All: Generalized LoRA for parameter-efficient fine-tuning," *arXiv preprint arXiv:2306.07967*, 2023.

[8] M. Wand and J. Schmidhuber, "Improving speaker-independent lipreading with domain-adversarial training," in *Interspeech*, 2017, pp. 3662–3666.

[9] P. A. Kandala, A. Thanda, D. K. Margam, *et al.*, "Speaker adaptation for lip-reading using visual identity vectors," in *Interspeech*, 2019, pp. 2758–2762.

[10] M. Kim, H. Kim, and Y. M. Ro, "Speaker-adaptive lip reading with user-dependent padding," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 576–593.

[11] M. Kim, H.-I. Kim, and Y. M. Ro, "Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition," *arXiv preprint arXiv:2302.08102*, 2023.

[12] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision (ACCV)*, 2016, pp. 6447–6456.

[13] M. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120 5 Pt 1, pp. 2421–4, 2006.

[14] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6447–6456.

[15] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 7613–7617.

[16] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.

[17] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 6219–6223.

[18] L. Pandey and A. S. Arif, "Effects of speaking rate on speech and silent speech recognition," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–8.

[19] L.-F. Lai and N. Holliday, "Exploring sources of racial bias in automatic speech recognition through the lens of rhythmic variation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023, pp. 1284–1288.

[20] C. Lugaresi, J. Tang, H. Nash, *et al.*, "MediaPipe: A framework for building perception pipelines," in *Third Workshop on Computer Vision for AR/VR at IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 66–75.

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*, 2023, pp. 28 492–28 518.

[23] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[24] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-AVSR: Audio-visual speech recognition with automatic labels," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations (ICLR)*, 2019.

[26] A. Baby, G. Joseph, and S. Singh, "Robust speaker personalisation using generalized low-rank adaptation for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024, pp. 11 381–11 385.

Cross Lingual Speech Representation for Infant Cry Classification

Hiya Chaudhari, Arth J. Shah, and Hemant A. Patil
Speech Research Lab, DA-IICT, Gandhinagar, India
E-mail: {202101047, 202101154, hemant_patil}@daiict.ac.in

***Abstract*—Infant Cry Classification is a critical area of research for medical diagnosis, early detection, and timely intervention. Classification of normal *vs*. pathological infant cries can be of potential use to infer the infant's health conditions. In this study, we offer a unique method that employs deep learning models, such as wav2vec, HuBERT, and XLS-R, as well as Whisper feature-based classification. We used Bi-directional Long-Short Term Memory (BiLSTM) as a deep learning classifier in this investigation. We compared XLS-R 0.3B's extracted features to those of other frequently used pre-trained models. Our proposed approach claims 29.02 % increase over wave2vec 2.0 features on standard Baby Chillanto dataset. In addition, a study of latency period is conducted to assess the practical impact of the proposed XLS-R 0.3B characteristics. Finally, the noise robustness of the suggested feature vector has been investigated in this study.**

***Index Terms* - Infant Cry Classification, wav2vec, HuBERT, XLS-R, Multilingual Representation, Bi-LSTM, Deep Learning, Transfer Learning.**

I. Introduction

Infants cry for a variety of reasons, including hunger, pain, discomfort, and the need for attention. Infancy is the period between birth and language acquisition. Infants primarily communicate with us through crying, which makes infant cry classification a crucial task. Crying is generated from a set of complicated and sophisticated physiological activities that involves coordination among the brain, respiratory and motor control, and the vocal system. It is considered that the crying helps in development of infant's physiology by increasing the pulmonary (lung) capacity. Moreover, lungs are power supply for cry/speech production. In [**1**] medical conditions such as asthma, HIE, asphyxia can influence the acoustics of infant cry. Diagnosing ailments in infants can be expensive and time-consuming due to a range of factors such as communication barriers and frequently exhibited non-specific symptoms (such as fever, crying, and irritability) that can indicate a variety of conditions, making diagnosis difficult. Numerous efforts have been undertaken to provide the best possible system for categorizing pathological *vs.* healthy cries. Owing to the latest developments in deep learning and machine learning, these techniques have been used for the classification of infant cries in several prior studies. This interdisciplinary field spans areas such as medicine, signal processing, pattern recognition, linguistics, and machine learning. Collecting infant cry samples is often challenging as some parents / guardians are hesitant to allow recordings. The availability of large-scale data on infant cry classification is limited due to data imbalance and ethical concerns. We are leveraging self-supervised learning and transfer learning techniques to improve model performance, especially for low resource settings, where annotated cry data is scarce. Given the difficulty of finding and storing labeled data, we employ self-supervised speech representations for this problem. Models pretrained on normal *vs.* pathological cries can be fine-tuned to detect specific diseases or conditions. The market demands better approaches to address issues, such as data privacy, which are highly relevant to the problem. Infant cry analysis offers a promising, non-invasive tool for enhancing infant-parent relationship and newborn developments [**2**]. The rest paper is organized as follows: Section II provides the details about related work. Section III presents information about proposed work in this study. Section IV presents information about experimental setup used for this study. Section V presents experimental results and the related discussion. Section VI provides summary, conclusion and the future potential research directions for this study.

II. Related Work

Many attempts have been made to develop an optimal system that classifies pathological *vs.* healthy cries. Due to recent advancement in machine learning and deep learning, many previous studies have employed these processes for infant cry classification. In [1], the authors used magnitude spectrum-based features with Convolutional Neural Network (CNN) as pattern classifiers for infant cry classification task. Many studies reported in recent time, such as in [2] and [3], authors used Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC) for infant cry classification. In [4], the authors used MFCC as features and machine learning approaches, such as Multi-Layer Perceptron (MLP) and CNN as classifier for infant cry classification task. Our previous study [5], employed Modified Group Delay Cepstral Coefficients (MGDCC), for infant cry classification, thereby exploiting phase based characteristics for task.

However, most of the studies fail to explore the characteristics of infant cry signals. According to authors' best knowledge and belief, this is the first study to employ Large Language Models (LLM) features for infant cry classification task. Whisper (pre-trained model by OpenAI) has been used formerly in [6], stands to be lone study on LLM for infant cry classification task. Previously LLM such as wav2vec 2.0 [7], XLS-R [8], Hubert [9] have been used for various classification tasks, such

as vocal intensity category [10], emotion recognization [11], speech recognition [12], speech representation [13], and many more tasks. In this study, we employ various types of pre-trained Large Language Models (LLM's) for this task.

III. Proposed Work

Since infant cry is paralingual in nature, it is non-lexical. Some advanced ASR systems are trained to recognize and annotate non-lexical elements, such as using special tokens to indicate pauses or fillers (e.g., laugh, cry, pause). The processing of non-lexical speech is heavily reliant on training data. If the model is trained on datasets with annotations for non-lexical sounds, it is more likely to recognize and transcribe these elements accurately. Non-lexical speech can occasionally add noise, which reduces the model's accuracy in transcribing the intended lexical content. However, robust methods, such as wav2vec, are built to handle such unpredictability while maintaining high transcription accuracy. Given that XLS-R is trained on broad multilingual data, it may meet a large variety of non-lexical sounds from different languages and cultures. This broad exposure enables the model to better generalize to other types of speech, including non-lexical features. A *cry unit* refers to a specific type of non-lexical vocalization typically found in cry sample data that captures infant crying sounds. Cry units are a specific kind of non-lexical sound. We have trained the models on infant cries. XLS-R model converts raw audio into latent representations that capture both *temporal* and *spectral* features of the sound. This detailed representation helps in distinguishing subtle differences in cries that might indicate normal or pathological conditions. XLS-R model is trained in a self-supervised manner, which allows it to learn rich and diverse features from raw audio without needing large amounts of labeled data. This is especially useful when dealing with specialized tasks such as classifying infant cries, where labeled data may be scarce. Although the primary task here isn't language-based, the cross-lingual training helps the model learn more universal audio patterns, which can be beneficial for distinguishing different types of cries.

A. XLS-R

XLS-R [14], an advanced version of wav2vec 2.0, utilizes cross-lingual self-supervised representation learning to address the challenge of limited resources in speech recognition tasks. XLS-R leverages large amounts of unlabeled cry data, reducing the need for extensive manual labelling. By training on this abundant unlabelled data, XLS-R builds robust models that can effectively classify and understand various cry types without the need for time-consuming data collection and annotation processes.

The training dataset for XLS-R includes diverse speech datasets, such as VoxLingual07, Common Voice, Babel, Multilingual LibriSpeech, and VoxPopuli, which are crucial for learning rich and diverse speech representations. Applying similar principles, infant cry identification can benefit from diverse datasets of infant cries, capturing a wide range of acoustic conditions and cry types. By autonomously extracting meaningful features from raw audio signals, XLS-R generalizes well across different contexts, making it ideal for identifying both normal *vs.* pathological cries in infants. This approach is vital for addressing the challenges posed by limited labeled data in cry analysis, enhancing the accessibility and effectiveness of cry identification systems.

XLS-R offers models of different sizes to balance performance and computational efficiency: XLS-R 300M, XLS-R 1B, and XLS-R 2B. The 300M model, with around 0.3 Billion parameters, is suitable for applications with limited computational resources, offering a good balance between efficiency and performance. The 1B model, containing approximately 1 billion parameters, provides enhanced performance and accuracy, making it ideal for more demanding tasks where higher computational power is available. The largest model, XLS-R 2B, with about 2 billion parameters, delivers the highest accuracy and robustness, suitable for the most demanding cry identification tasks in resource-rich environments. This range of model sizes allows users to choose based on their specific needs and available resources, ensuring optimal performance for various applications in infant cry classification.

The experimental setup for training and evaluating XLS-R models is thorough and involves several key steps. First, large and diverse cry datasets are collected and preprocessed to ensure consistent audio quality. Depending on the available computational resources and desired performance, an appropriate XLS-R model size (300M, 1B, or 2B parameters) is chosen. GPUs are used to train the models using self-supervised learning techniques, where the model learns to predict masked parts of the audio, allowing it to learn from vast amounts of unlabeled cry data. The training process includes optimizing learning rates, batch sizes, and stabilizing training with techniques like gradient clipping. The models are then evaluated on their cry classification performance using metrics like accuracy and their ability to generalize across different cry types. Fine-tuning may be done on specific cry types to improve accuracy further. After training, the models are optimized for deployment by reducing their size and improving inference speed, making them efficient for real-world applications. This comprehensive setup ensures that XLS-R models are robust and high-performing, capable of improving cry identification across diverse acoustic conditions.

XLS-R is pivotal for infant cry identification due to its ability to capture and distinguish various acoustic features effectively. Utilizing a self-supervised learning approach, it learns robust and generalizable representations from large-scale, unlabeled cry data, significantly reducing the reliance on extensive labeled datasets. This makes the model highly efficient and adaptable, allowing for quick fine-tuning on specific cry identification tasks. Consequently, XLS-R enhances the accuracy and performance of cry identification systems, even in resource-constrained environments, by providing rich, pre-trained representations that can be easily adapted to identify multiple types of infant cries.

IV. Experimental Setup

A. Dataset Used

TABLE I
Distribution of Dataset Used

Class	Sub-class	Baby Chillanto
Pathological	Asphyxia	340
	Deaf	879
Healthy	Pain	192
	Hunger	350
	Normal	507

The dataset being used is Baby Chillanto dataset (D1) was created by recordings, which were made by doctors of NIAOE-CONACYT, located in Mexico . It includes 1049 healthy infant cries and 1219 pathological cries. This dataset has 5 classes, which can be divided into two categories: mentioned in Table I.

B. DL Models Used

1) ***wav2vec 2.0:*** It is an unsupervised method for training speech recognition models involving no labeled data, it stands for wave to vector. It uses self-supervised speech representations to segment unlabeled language or speech sounds and then learns a mapping from these representations to phonemes using adversarial training. It masks the speech input in the latent space and solves a contrastive task defined by a quantization of the latent representations that are jointly learned. Whisper is a pre-trained model for speech translation and automated speech recognition (ASR) . Whisper models exhibit a great capacity to generalize to numerous datasets and domains without requiring fine-tuning, having been trained on 680k hours of labelled data.

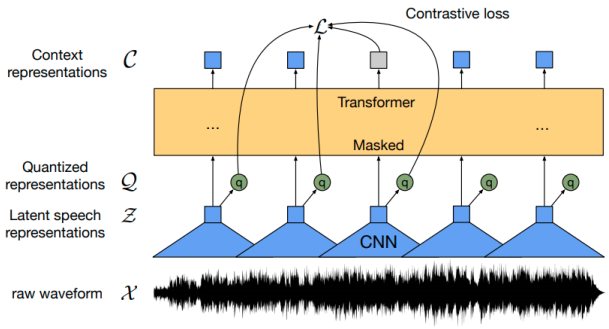


Fig. 1. Architecture of wav2vec 2.0 [7].

2) ***HuBERT:*** It is an acronym for Hidden units Bidirectional Encoder Representations from Transformers. The self-supervised speech representation learning approach known as Hidden-Unit BERT (HuBERT) uses an offline clustering step to generate aligned target labels for a BERT-like prediction loss. It takes in account cross-entropy loss and builds target via a separate clustering process.HuBERT has 3 models Base, Large and XL (Extra Large). We have utilised Large HuBERT model.The architectire of HuBERT is shown in fig. 2.

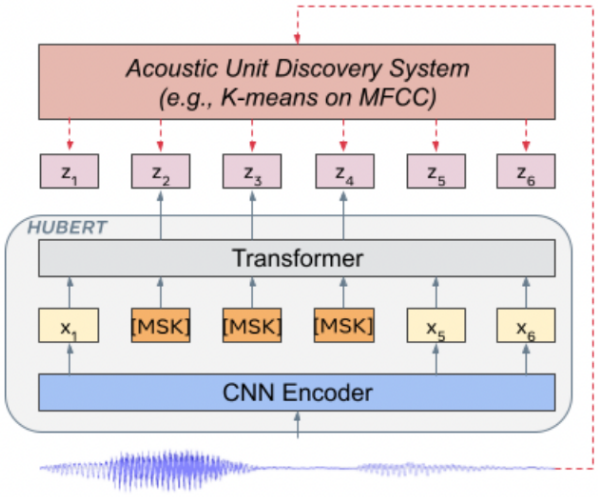


Fig. 2. Architecture of HuBERT [9].

3) ***XLS-R:*** It stands for Cross-Lingual Speech Representation, a multilingual speech recognition model based on wav2vec 2.0. It learns a quantization of the shared latents across languages by solving a contrasting task over masked latent speech representations.Experiments demonstrate that cross-lingual pre-training greatly outperforms monolingual pre-training, and the model is optimized on labeled data. We have used 3 XLS-R models namely XSL-R 2B, which is trained over 2B parameters, XLS-R-1B and XLS-R 300M.

4) ***Whisper:*** Whisper is derived from the acronym WSPSR, which stands for Web-scale Supervised pre-training for Speech Recognition[15] . Whisper has namely five models with increasing number of encoder-decoder blocks and number of trainable parameters, namely, tiny, base, small, medium, and large. We have employed base model for this study.

C. Classifier Used

The bi-directional Long Short-Term Memory classifier is a neural network architecture commonly used for sequence classification tasks. It employs units, which process input patterns in both forward and backward directions [16]. By capturing content from both past and future tokens, BiLSTM enhances its ability to model long-range dependencies. A dropout layer is applied to prevent overfitting, and the final output is obtained by concatenating the hidden states from both directions and passing them through a fully-connected layer. The model is well-suited for tasks, such as sentiment analysis and named entity recognition. Learning rate was chosen to be 0.003, in combination with batch size of 32. Each dropout layer is 25 % dropout, and hidden size is 256. In this paper, the input size was chosen as a constant value of 20.

D. Performance Metrics

In this study accuracy, and F1-score are used as the performance metrics utilized to gauge the effectiveness of the system. These metrics facilitated precise measurement of the

system's performance. Evaluation accuracy, a metric utilized for assessing the performance of classification models, measures the proportion of correctly predicted instances to the total instances in the evaluation set. It offers a clear indication of the model's capability to accurately predict class labels.

F1-Score: This metric effectively balances precision and recall for this problem. The F1-score combines precision and recall to provide a balanced metric. It ranges between 0 and 1. An F1-score close to 1 indicates optimal precision and recall for any problem. F1-score is determined by

$$F1 - Score = \frac{2TP}{2TP + FN + FP}. \quad (1)$$

where FP, and FN denote false positives and false negatives, respectively.

V. Experimental Results

Table II shows the performance of various pre-trained speech models on the Baby Chillanto dataset for classifying infant cries. The models evaluated include wav2vec 2.0, HuBERT, XLS-R 2B, XLS-R 1B, Whisper, and XLS-R 300M. Among these, XLS-R 300M achieved the highest accuracy (99.68 %) and F1-score (99.48 %), indicating its superior ability to correctly classify infant cries. Whisper also performed very well with 95.13 % accuracy and an 95 F1-score. In contrast, wav2vec 2.0 had the lowest accuracy (70.66 %) but still maintained a decent F1-score (86.84 %), suggesting it might be less effective for this specific task compared to the larger and more advanced XLS-R models.

TABLE II
Comparison of Various Feature Vector for Classification of Normal *vs.* Pathological Infant Cry

Baby Chillanto	Accuracy (in %)	F1-score
W2V 2.0	70.66	86.84
HuBERT	80.44	80.23
XLS-R 2B	96.44	94.63
XLS-R 1B	87.11	86.84
Whisper	95.13	95
XLS-R 300M	**99.68**	**99.48**

A. Analysis of Latency Period

Fig. 3 shows the latency period of various pre-trained speech models namely HuBERT, wav2vec 2.0, XLS-R 1B, XLS-R 2B, and XLS-R 300M across different numbers of frames required. As the dimension increases, the latency also increases for all models. XLS-R 1B and XLS-R 2B demonstrate lower latency at lower number of frames (20 and 40), indicating they are more efficient for quick inferences. HuBERT and wav2vec 2.0 show moderate latency at mid-range frames (60 and 80). At the highest frames (100), all models experience increased latency, with XLS-R 300M having the highest latency despite its superior accuracy and F1- score. Thus, while XLS-R 1B and XLS-R 2B provide a good balance of efficiency and performance, XLS-R 300M is the best choice for the highest accuracy, even though with a higher latency trade-off.

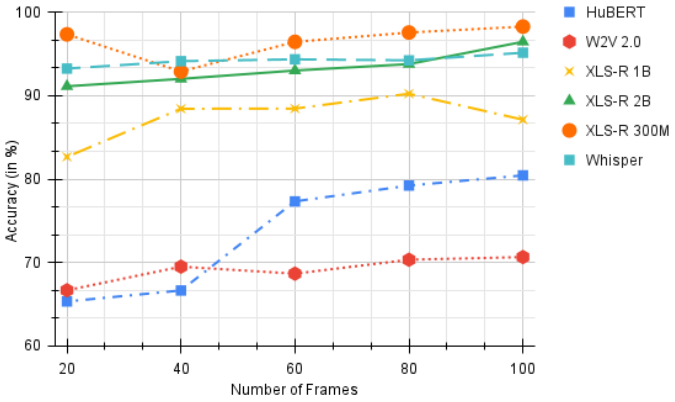


Fig. 3. Latency Analysis of Proposed Method

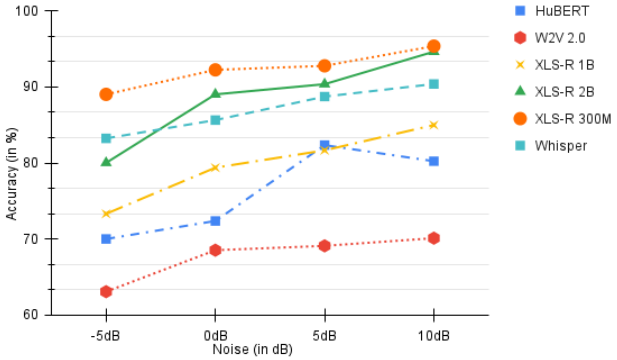


Fig. 4. Noise Robustness of Feature Vector

B. Noise Robustness

Fig. 4 shows analysis of adding babble noise to the LLM models characteristics changes their noise robustness. We also compared the performance of these characteristics to that of HuBERT, wav2vec, XLS-R 2B and XLS-R 1B. A. Furthermore, the effects of noise degradation on the proposed technique are investigated by considering various SNR levels for babble noise, including -5 dB, 0 dB, 5 dB, and 10 dB. Such tests allow us to evaluate the model's capability in real-world conditions. We obtained better results with the XLS-R 0.3B feature vector than with the HuBERT, wav2vec, XLS-R 2B, and XLS-R 1B, demonstrating that the XLS-R 0.3B feature vector outperforms the other features even in noisy conditions. Furthermore, this experiment allows us to estimate the correctness of models in real-life application.

We compared our proposed XLS-R 300M features with existing studies in Table III.

VI. Summary and Conclusions

This work investigated significance of Self-Supervised models for the socially relevant task of infant cry classification. It focused on detecting and identifying pathology in infants and classification of normal *vs.* pathological infant cry. We compared XLS-R 0.3B with other DL models such as wav2vec

TABLE III
COMPARISON WITH DIFFERENT EXISTING WORKS ON BABY CHILLANTO DATASET.

Source	Features	Classifiers	Performance
gupta2023classification	Morse wavelet	GMM	98.19%
[1]	MFCC	GMM	99.16%
[6]	Whisper	Bi-LSTM	97.31%
[6]	Whisper	CNN	97.31%
[1]	LFCC	CNN	99.16%
Proposed	**XLS-R 300M**	**Bi-LSTM**	**99.68%**

2.0, XLS-R 1B, XLS-R 2B, and HuBERT, showed plots corresponding analysis of latency period and noise degradation. We proposed a new optimal system based LLM's for infant cry classification using Baby Chillanto dataset. We studied the effect of babble noise on the proposed XLS-R features and investigated the different dimension of feature vector. Future work involves experiments on variuos types of noise (pink noise, white noise, high frequency noise) and feature fusion.

REFERENCES

[1] S. Uthiraa, A. Kachhi, and H. A. Patil, "Linear frequency residual features for infant cry classification," in *International Conference on Speech and Computer (LNCS)*, Springer, Dharwad, India, 2023, pp. 550–561.

[2] S. P. Dewi, A. L. Prasasti, and B. Irawan, "The study of baby crying analysis using MFCC and LFCC in different classification methods," in *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, 2019, Bandung, Indonesia, pp. 18–23.

[3] A. Pusuluri, A. Kachhi, and H. A. Patil, "Analysis of time-averaged feature extraction techniques on infant cry classification," in *International Conference on Speech and Computer (LNCS)*, Springer, Dharwad, India, 2022, pp. 590–603.

[4] A. Abbaskhah, H. Sedighi, and H. Marvi, "Infant cry classification by MFCC feature extraction with MLP and CNN structures," *Biomedical Signal Processing and Control*, vol. 86, pp. 105–261, 2023.

[5] A. J. Shah, H. Chaudhari, and H. A. Patil, "Modified group delay cepstral coefficients for infant cry classification," in *International Conference on Pattern Recognition (ICPR)*, 2024, Kolkata, India.

[6] M. Charola, A. Kachhi, and H. A. Patil, "Whisper encoder features for infant cry classification," in *Proc. INTERSPEECH*, vol. 2023, 2023, Dublin, Ireland, pp. 1773–1777.

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems (NIPS)*, vol. 33, pp. 12 449–12 460, 2020, Virtual (Covid-19).

[8] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020, {Last Accessed Date: 10th October, 2024}.

[9] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?" In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, Toronto, Canada, pp. 6533–6537.

[10] M. Kodali, S. Kadiri, and P. Alku, "Classification of vocal intensity category from speech using the wav2vec2 and whisper embeddings," in *INTERSPEECH*, Dublin, Ireland, 2023, pp. 4134–4138.

[11] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, Singapore, pp. 6907–6911.

[12] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020, {Last Accessed Date : 2nd July, 2024}.

[13] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, Singapore, pp. 7087–7091.

[14] A. Babu, C. Wang, A. Tjandra, *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021, {Last Accessed Date : 2nd July, 2024}.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. arXiv, {Last Accessed Date: 10th October, 2024}: 2212 . 04356 (eess.AS). [Online]. Available: https://arxiv.org/abs/2212.04356.

[16] M. Charola, S. Rathod, and H. A. Patil, "Robustness of whisper features for infant cry classification," in *International Conference on Speech and Computer (LNCS)*, Springer, Dharwad, India, 2023, pp. 421–433.

Pop Noise Detection Using Group Delay Cepstral Coefficients

Arth Shah, Prathav Kevadiya, and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India

E-mail:{202101154, 202003020, hemant_patil } @daiict.ac.in

Abstract

This proposed study leverages the distinctive pop noise, which is involuntarily generated when pronouncing specific phonemes, such as plosive, fricative, nasal, and affricate sounds. We employ Group Delay Cepstral Coefficients (GDCC) as the key features for the Voice Liveness Detection (VLD) task. To measure the effectiveness of GDCC, we compared the performance using different spectral features, namely, Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), and Gammatone Frequency Cepstral Coefficients (GFCC), employing them with three types of classifiers, namely, Convolutional Neural Networks (CNN), Convolutional Recurrent Neural Networks (CRNN), and Localized Convolutional Neural Network (LCNN). We achieved an accuracy of 79.55% for the GDCC feature vector as its highest, which is comparatively much higher than other feature vectors, thereby proposing a new optimal approach for the VLD task.

1. Introduction

Robust security has become more critical than ever in today's digital world, where remote communication and online transaction systems dominate. Voice biometrics has gained popularity as it's a practical, effective, and natural method for verification of a person's identity. In the meantime, advancements in speech synthesis and transformation technologies have made it possible to produce realistic-sounding artificial (or spoofed) speech with the audio of the target speaker from a supplied text or an inputted audio waveform pronounced by an imposter speaker. Nonetheless, analogous methods might be employed to create counterfeit identities or profiles and initiate spoofing assaults on Automatic Speaker Verification (ASV) systems [1], [2], thereby presenting a substantial risk to the security of these systems. To that effect, researchers of this study focus on developing advanced Voice Liveness Detection (VLD) systems to distinguish between genuine *vs.* spoofed voices, which is more reliable.

Several studies have been explored to increase the reliability of VLD systems in recent years; some countermeasures used the ASVSpoof challenge datasets to develop and evaluate their model. Nevertheless, inspired by [3], we also employ pop noise in our work to protect ASV systems against spoofing assaults. In this context, a new dataset called POCO (POp noise COrpus) [4] has also been developed in the literature to make it easier to design a variety of countermeasures that seek to find pop noise—a marker of human liveness in voice signals. One typical type of speech distortion is called "*pop noise*," which occurs when air from a speaker's lips enters a microphone and is misrepresented by speakers. When assessing an authentication system, it seems sensible to use pop noise as a trustworthy signal of liveness, given its common occurrence. The ability to differentiate between genuine audio and replayed audio (through loudspeakers) may be effectively established by considering a pop noise check.

Phase-based features have become a potential alternative to magnitude spectrum-based features for several speech applications. Phase-based features have noise robustness, perceptual relevance, anti-spoofing ability, and discriminating power. These attributes help to differentiate between authentic human voices and spoof or prerecorded signals because they capture fine-grained *temporal* and *phase*-related properties of the speech signal. In this study, we exploit the Group Delay Function (GDF) for feature extraction to achieve better spectral resolution than the magnitude spectrum for the VLD task. Further, the features obtained from the GDF are combined with Convolutional Neural Networks (CNN), Convolutional Recurrent Neural Networks (CRNN), and Localized Convolutional Neural Network (LCNN) classifiers to detect pop noises, which are found to give promising results.

1.1. Pop Noise

While producing normal speech, a live speaker generates a speech distortion known as pop noise [5]. This phenomenon occurs when the Vocal Tract System (VTS) is stimulated by air from the lungs, causing the vocal folds to vibrate [6]. As the glottal airflow passes through a cascade of organ pipes, it reaches the mouth, where a specific phoneme being uttered determines the location and force of the burst of air, producing a sudden outburst of sound from the mouth. Pop noise may have a different intensity depending on how close the speaker is to the recorded microphone.

By detecting and analyzing pop noise, it becomes possible to protect speech signal information from potential attackers attempting fraud or unauthorized voice recordings. Detecting pop noise provides genuine acoustic cues for VLD, enabling distinguishing between live (natural) speech *vs.* replayed the speech. This distinction is crucial in preventing fraudulent activities, and enhancing the security of voice-based systems [7]. Therefore, investigating and identifying pop noise characteristics and developing techniques to detect and analyze them contribute to the development of effective VLD methods. By leveraging pop noise as an acoustic cue, VLD systems can distinguish between genuine live speech and spoofed or manipulated recordings, enhancing the reliability and security of voice-based authentication and verification systems. This study provides the following novelty:

- Mathematical as well as theoretical analysis of GDCC feature vector.
- Discussion of Results obtained.
- Experimental analysis related to the feature vector's evaluation factor and latency analysis.

- Comparison with different features on different classifiers, and existing approach.

The paper's remaining section is organized as follows: With a thorough theoretical exposition and a comparison of GDF with other features, Section 2 offers computational details on GDCC feature extraction. The dataset, performance measures, classifiers utilized in the experiments, and other spectral properties besides GDCC used for comparison are all described in depth in Section 3. Using the GDCC (proposed) feature set, the results are shown in Section 4 along with a comparison to previous research. Lastly, a summary of the work and recommendations for further research are provided in Section 5.

2. Proposed GDCC Features

The information offered by the magnitude spectrum is supplemented by the group delay spectrum, which provides insightful information about the *temporal* properties of speech signals [8]. The speech signal's envelope and fine structure are determined by the VTS and excitation source, respectively. The objective of extracting features from the magnitude spectrum involves identifying the spectral envelope of the audio signal. It is possible to gain more information about speech signals, and improve the effectiveness of specific speech analysis tasks by integrating *spectral* and *temporal* information.[8]

Because of its lower complexity compared to the phase spectrum, the magnitude spectrum facilitates the relatively straightforward extraction of features containing information about the signal, which requires addressing the intricate task of phase unwrapping in order to reverse the signal processing effects inherent in the arctangent function. [8]. Speech signals are a combination of magnitude and phase spectrum. This study explores phase-based characteristics of signals containing pop noise. In order to achieve this, we use the GDF, which, by performing a negative derivative operation on the unwrapped phase function, enables us to see quick fluctuations in it, i.e., the group delay function [9]. Better resolution of the resonant (peaks) formant is made possible by this GD function characteristic as opposed to the short-time speech spectrum.

Many researchers in the field of speech technology have employed magnitude-based acoustic features, resulting in neglecting of phase spectrum-based characteristics of speech signals. GDF is based on the shift initiated by a *group* of frequencies in a system containing linear phase characteristics [9]. A similar process can be applied to systems with nonlinear phase characteristics by linearly approximating their narrowband input. The input is affected by the system in three ways: it shapes its magnitude, multiplies it using a complex function, and adds a localized linear phase term that represents the delay. The study of GDF is concerned with the delay that a set of frequencies experiences when they are introduced into the system.[10].

For a discrete-time speech signal x(n), its Discrete-Time Fourier Transform (DTFT) represented as X($e^{j\omega_0}$), which is expressed via magnitude-phase representation as [5]:

$$X(e^{j\omega_0}) = \mid X(e^{j\omega_0}) \mid e^{j\phi(e^{j\omega_0})}, \tag{1}$$

where the phase spectrum at frequency ω_0 is represented by $\phi(e^{j\omega_0})$, and the magnitude spectrum is represented by $|X(e^{j\omega_0})|$. The negative derivative of the unwrapped phase, or GDF, has been studied in attempt to extract information from the phase spectrum. In particular,

$$\tau(e^{j\omega_0}) = -\frac{d}{d\omega_0}\phi(e^{j\omega_0}) = -j\,\mathrm{mag}\left[\frac{d}{d\omega_0}\log(X(e^{j\omega_0}))\right].$$

In order to avoid the computationally intensive task of phase unwrapping, an alternative way to compute the GD function is by exploiting the frequency-domain differentiation property of DTFT [11], [12]. In particular,

$$\tau(e^{j\omega_0}) = \frac{X_R(e^{j\omega_0})Y_R(e^{j\omega_0}) + X_I(e^{j\omega_0})Y_I(e^{j\omega_0})}{|X_c(e^{j\omega_0})|^2},$$

where $X(e^{j\omega_0}) \rightarrow$ DTFT of x(n), and $Y(e^{j\omega_0}) \rightarrow$ DTFT of nx(n), R and I are the real and imaginary parts of the Fourier transform, respectively (proof of Eq. 1 is highlighted in Appendix), and $|X_c(e^{j\omega_0})|$ is cepstral smoothing version of $|X(\omega_0)|$.

Algorithm 1 GDCC Features Extraction

1. $x(n)$ = short-time speech signal segment
2. y(n) = n.x(n)
3. $X(e^{j\omega_0})$=DTFT{x(n)}=$X_R(e^{j\omega_0})+jX_I(e^{j\omega_0})$, Y($e^{j\omega_0}$)=DTFT$\{nx(n)\} = Y_R(e^{j\omega_0}) + jY_I(e^{j\omega_0})$
4. $\tau(e^{j\omega_0}) = \frac{X_R(e^{j\omega_0})Y_R(e^{j\omega_0})+X_I(e^{j\omega_0})Y_I(e^{j\omega_0})}{|X_c(e^{j\omega_0})|^2}$
5. GDCC = DCT$\{\tau(e^{j\omega_0})\}$

The Group Delay Cepstral Coefficients (GDCC) features are derived from GDF. The computation of GDCC involves a series of steps (as shown in Algorithm 1) to capture the speech signal's phase information and spectral dynamics. Because of its capacity to capture fine-grained temporal and spectral properties, this feature has demonstrated promising results in a variety of speech processing applications [9]. For the extraction of GDCC features, the Hamming window was 25 *ms* duration, the shift was 10 *ms*, and the *20-D* (to be discussed shortly in Section 4.1) cepstral features were used for experiments. Figure 2 shows the functional block diagram of proposed GDCC feature based methodology.

2.1. Discussion

This Section discusses the significance of the selected feature vector. The concept of GDCC revolves around the concepts of spikes and zeros in the signals. The spikes in the signals are also often known as formants in speech technology. These formants in the z-plane of a real-time speech signal are in the pattern of the mixed-phase system, which means they lie inside and outside the unit circle. In minimum phase signal (like natural speech), if the roots are not near the unit circle, then GDF identifies the signal correctly [13]. In Figure 1 (a), Panel 1 and Panel 2 describe the waveforms of the signal with and without pop noise, respectively. Figure 1 (b) Panel 2 shows the enlarged pop noise region of (0.8 - 1.4 *s*) from (a), which shows the effect of a sudden closure of vocal folds resulting in pop noise formation which is decayed in Panel 1, as it is recorded from a far distance [4]. Figure 1 (c) signifies the difference between the magnitude spectrum of recorded and original speech. Figure 1 (d) represents full-frequency scale GDF, which is further magnified and analyzed in Figure 1 (e) as previous studies found that pop noise lies at a low-frequency region [14]. Figure 3 illustrates both the spectrogram and the group delay spectrogram of the audio signal. It is evident that the frequency bins in the group delay spectrogram exhibit higher resolution in lower frequency regions compared to the spectrogram. However, when

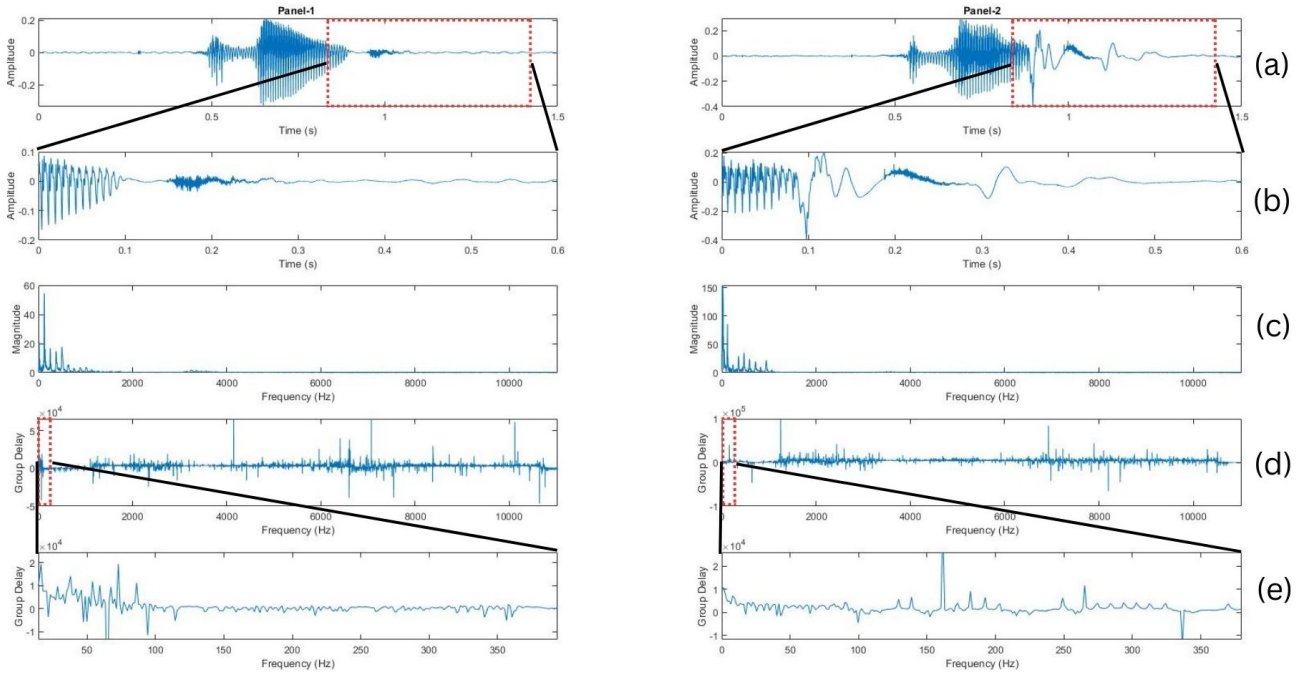


Figure 1: Panel-1 (signal without pop noise), and Panel-2 (signal with pop noise): (a) the original speech signal, (b), (c), (d), (e) shows the short-time signal, corresponding magnitude spectrum, group delay function, and its selected region for 0-400 Hz of the selected signal (as shown via the dotted box in Figure 1(d)), respectively.

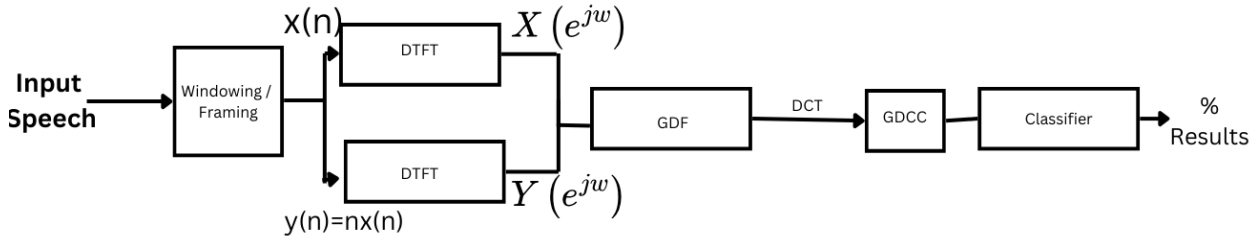


Figure 2: Functional block diagram of proposed GDCC-based VLD system.

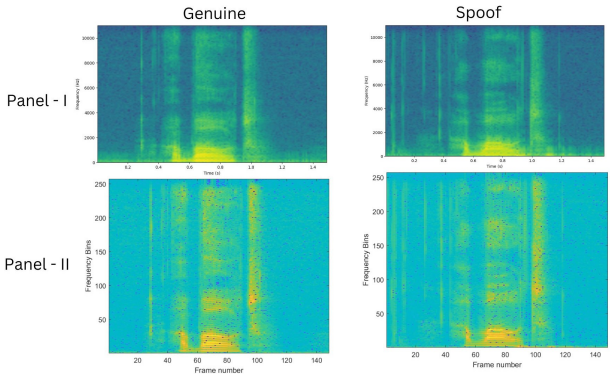


Figure 3: Panel-1 represents simple spectrogram, and Panel-2 represents group delay spectrogram.

examining the spectrogram, there is not a noticeable distinction between genuine and spoof signals. According to the authors' first knowledge, this may be the key reason why the GDCC feature vector outperforms another feature vector *w.r.t.* in its ability to capture unique phase-based characteristics, unlike other magnitude-based features.

3. Experimental Setup

3.1. Dataset Employed

In real-world cases, when a fraudster intends to carry out a Synthetic Spoofing Attack (SSA), they typically need to acquire the target subject's original audio. The most straightforward method to acquire this data involves recording the audio of the specific speaker of interest and subsequently utilizing that recorded audio to integrate a Speaker Specific Adaptation (SSA) into the Automatic Speaker Verification (ASV) system. The pop noise may not be captured since the collection of audio will be carried out from far away [3]. In such cases, it is feasible to identify and distinguish between real *vs.* spoofed speech by taking advantage of the lack of pop noise in replayed samples.

We have been using the POCO dataset for this model, which has been recently released [4]. A total of 66 people, 34 were female, and 32 were male, whose age range is from 18 to 61

years [4]. All the words are in one language (English), and 44 verbs were present in the dataset, sampled at 22050 Hz. The dataset has been released in 3 sub-parts, i.e., RC-A (microphone recordings), RP-A (eavesdropping), and RC-B (microphone array recordings). However, RC-B (a portion of data related to spoofed speech) was not executed during this study as it contains microphone array recordings. The dataset was split up

Table 1: Details of POCO Dataset. After [4].

	No. of Samples	**Total Subjects**	**Male Subjects**	**Female Subjects**
Training	13640	53	26	27
Evaluation	3432	13	6	7

into two categories: training (80%) and evaluation (20%) of all the utterances. Each of these subgroups is split equally between genuine and spoof audio samples. We also made sure that every speaker subset was unique, and that the proportion of male-to-female speakers was kept constant. Table I displays the data distribution statistics in the training and evaluation subgroup. All the detailed information about the dataset is given in [4].

3.2. Other Spectral Features Used

We employed three commonly used features as comparison with baseline features, i.e., Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), Gammatone Frequency Cepstral Coefficients (GFCC) for comparison in this study. Motivated from ability of MFCC for human perspective of speech evaluation on mel scale, and LFCC for human perspective of speech evaluation on linear scale. On the other hand, GFCC is renounced to capture spectral characteristics of human audios, makes it perfect for comparison alongside MFCC, and LFCC.

3.3. Classifiers Used

We have performed the experiments using CNN [15], CRNN [16], and LCNN [17] as classifiers [18]. CNN and CRNN are employed as binary classifiers, where the two classes are speech with pop noise (i.e., genuine human voice) and without pop noise (i.e., spoofed voice) [19]. A learning rate of 0.001 is used in the construction of each model using the Adam optimizer. The loss function for binary classification jobs is binary cross-entropy, with a batch size of 64. The experiment employs *5*-fold cross-validation, with 100 epochs executed during training.

4. Experimental Results

4.1. Effect of Dimension of GDCC feature vector

Before performing comparison experiments, it is important to optimize parameters for a particular feature vector. In contrast to another feature set, GDCC offers the advantage of having only a few parameters that require optimization, with one of them being the dimension of the feature vector. Table 2 represents the effect on accuracy *w.r.t.* change in dimension of the feature vector, which shows state that as we increase the dimension of the feature vector, the accuracy increases to a certain extent, after-by the accuracy starts decreasing due to possible reduction in features as the dimensions are higher than required (*20-D*). This study focused on detecting pop noise through a

Table 2: Effects of Dimension of GDCC Feature Vector using CNN Classifier

Dimension	**Accuracy (in %)**	**F1-Score**	**EER**
10	76.43	75.82	23.77
13	74.07	74.33	26.75
16	78.06	77.68	22.14
20	**79.32**	**79.54**	**20.52**
23	78.53	78.64	21.21
26	78.09	77.96	23.02
30	77.51	78.08	22.14

binary classification task. The dataset was labeled "1" for pop noise files, and "0" for non-pop noise files. In Figure 4, we have compared wordwise accuracies for MFCC, LFCC, GFCC, and GDCC, all using the same CNN classifier. It can be observed from Fig. 4 that MFCC-CNN gives the least accuracy, while the accuracies of GFCC-CNN and LFCC-CNN are almost comparable. In some instances, GFCC-CNN performs better than LFCC-CNN or vice-versa; such behavior could be because some words contain low-intensity pop noises. However, we can see that GDCC-CNN outperforms all other feature sets.

Table 3: Accuracy (in %) for corresponding features and classifiers

Feature Set	CNN	CRNN	LCNN
MFCC	59.12	72.00	74.80
LFCC	76.74	72.73	77.07
GFCC	65.33	68.35	67.92
GDCC	**79.55**	**74.94**	**77.59**

Table 4: F1-score for corresponding features and classifiers

Feature Set	CNN	CRNN	LCNN
MFCC	0.5738	0.7100	0.730
LFCC	0.7357	0.7449	0.7451
GFCC	0.6478	0.6752	0.6861
GDCC	**0.7913**	**0.7432**	**0.7535**

Table 5: EER (in %) for corresponding features and classifiers

Feature Set	CNN	CRNN	LCNN
MFCC	43.32	23.69	26.51
LFCC	23.31	26.04	22.96
GFCC	33.88	32.36	32.79
GDCC	**20.68**	**23.54**	**22.26**

The experimental results in Table 3, Table 4, and Table 5 suggest that by using GDCC features in combination with CNN, the VLD system achieved relatively better performance. The accuracy score obtained was 79.55%, indicating that the system accurately classified the majority of the samples. The F1-score,

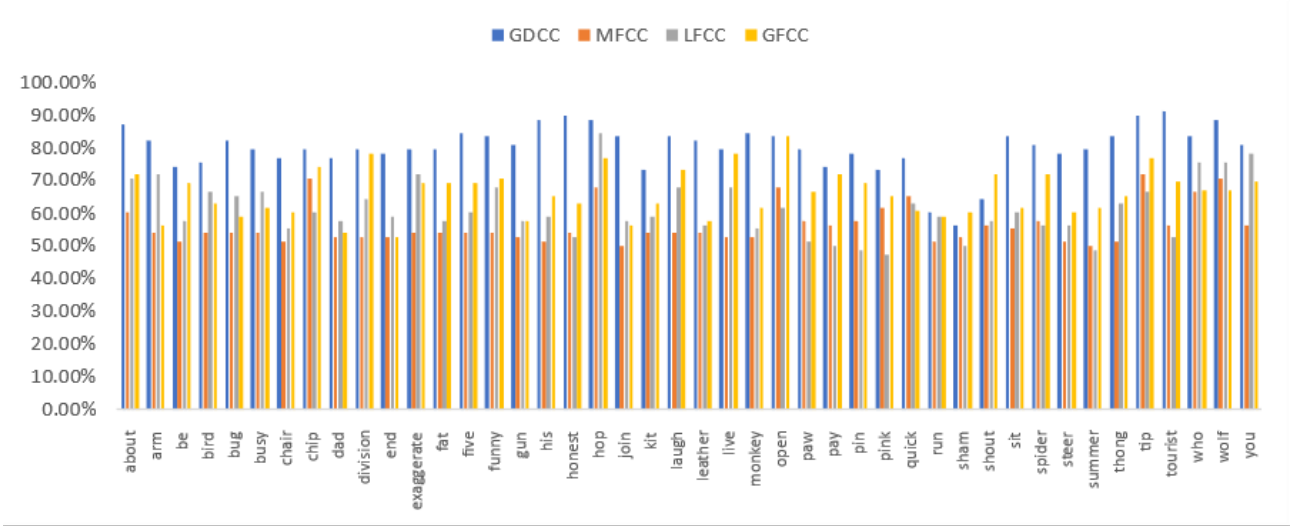


Figure 4: Comparision of wordwise accuracy (in %) for GDCC, MFCC, LFCC, and GFCC features using CNN classifier.

which stands at 0.7931, implies a good equilibrium between precision and recall. This suggests that the system managed to minimize false positives and negatives to a certain extent. The EER score of 20.68% demonstrates the system's ability to achieve a balanced error rate. These findings suggest that phase-based features capture fine-grained temporal and phase-related properties of speech signals, allowing for better discrimination between genuine human voices and spoof or prerecorded signals. Utilizing GDCC features in conjunction with deep learning models can bolster the security and reliability of voice biometric systems.

On the other hand, GDCC features are beating each feature set individually even after changing the classifier. GDCC-LCNN and LFCC-LCNN are close, but GDCC-LCNN still has a minute better result than LFCC-LCNN. Hence, GDCC-CNN is promising for pop noise detection in VLD, offering improved accuracy compared to the other feature extraction techniques.

4.2. Comparison With existing Approaches

In this subsection, we contrast our proposed methodology with established approaches. We achieved higher accuracy compared to recent optimal existing approaches, which utilize acoustic features derived solely from the magnitude spectrum of the speech signal, overlooking the phase spectrum of the speech signal. By comparing this work and showing the superiority of accuracy, we can say that phase-based characteristics carry a greater amount of information as compared to the magnitude spectrum of the pop noise signal. Table 6 shows that we obtained a total of 7.73% increase in accuracy as compared to the baseline work [20].

4.3. Evaluation of Latency Period

The latency period, a critical factor in numerous real-time applications, measures the time delay between the occurrence of an event and its detection [21]. Understanding the performance of different latency calculation techniques is vital for optimizing the efficiency and accuracy of such systems. The minimum amount of speech duration needed to obtain the highest possible classification accuracy is called *latency period*. Figure 5 shows the analysis of the latency period for 3 selective features, i.e., GDCC, MFCC, and LFCC.

Table 6: Comparison with existing works in existing literature.

Study	Feature	Frequency Range (in Hz)	Classifier	Accuracy
[20]	STFT	0-11025	CNN	71.81
[14]	Bump wavelet	0-40	CNN	74.69
(A)	Mel-STFT	0-40	CNN	76.39
Proposed	GDCC	0-11025	CNN	**79.54**

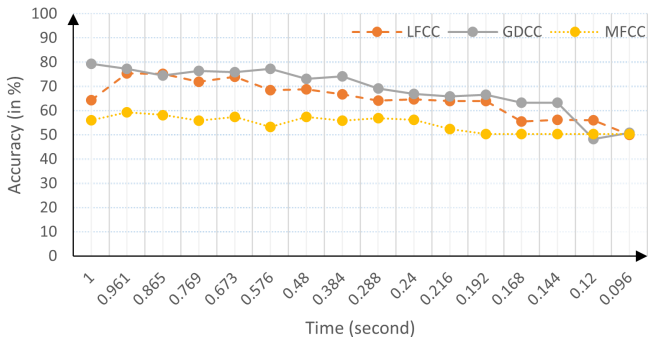


Figure 5: Analysis of latency period using accuracy (in %) as measure using CNN classifier

5. Summary and Conclusions

In this research, different classifiers were employed to assess multiple spectral feature sets, including MFCC, LFCC, GFCC, and GDCC. Despite altering the classifier, the experiments consistently indicated that GDCC exhibited higher accuracy compared to the other three spectral feature sets. These experimental results underscore the efficacy of GDCC features for detecting pop noise in the VLD task. These results are further

supported by a graph comparing wordwise accuracies, which reveals that MFCC-CNN is the least accurate while GFCC-CNN and LFCC-CNN perform similarly. The higher performance of GDCC-CNN raises the possibility of using it to enhance pop noise detection in VLD applications. GDCC-CNN offers improved accuracy compared to other feature extraction approaches, making it a potential option for pop noise identification in VLD. However, GDCC features are computed based on the short time window of speech signals. While they capture information about the spectral shape and group delay properties within each window function, they may lack the ability to capture long-term contextual information, such as prosody or semantic meaning. Also, they use sliding window analysis, resulting in a high-dimensional feature space. This can pose challenges in terms of computational complexity and storage requirements, especially when dealing with large datasets or real-time applications. We aim to focus our future research on exploring the latency period of GDCC features, with the goal of assessing their feasibility for deploying VLD systems in real-world scenarios.

6. References

[1] Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Zhizheng Wu, Federico Alegre, and Phillip De Leon, "Speaker recognition anti-spoofing," *Handbook of Biometric Anti-Spoofing: Trusted Biometrics under Spoofing Attacks*, pp. 125–146, 2014.

[2] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[3] Shihono Mochizuki, Sayaka Shiota, and Hitoshi Kiya, "Voice livness detection based on pop-noise detector with phoneme information for speaker verification," *The Journal of the Acoustical Society of America (JASA)*, vol. 140, no. 4, pp. 3060–3060, 2016.

[4] Kosuke Akimoto, Seng Pei Liew, Sakiko Mishima, Ryo Mizushima, and Kong Aik Lee, "POCO: A voice spoofing and liveness detection corpus based on pop noise," in *INTERSPEECH, Shanghai, China*, 2020, pp. 1081–1085.

[5] Thomas F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Education India, 2002.

[6] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in *INTERSPEECH, Dresden, Germany*, 2015.

[7] Mohit Dua, Rajesh Kumar Aggarwal, and Mantosh Biswas, "GFCC based discriminatively trained noise robust continuous asr system for hindi language," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, pp. 2301–2314, 2019.

[8] J. Tribolet, "A new phase unwrapping algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 2, pp. 170–177, 1977.

[9] Hema A. Murthy and B. Yegnanarayana, "Group delay functions and its applications in speech technology," *Sadhana*, vol. 36, pp. 745–782, 2011.

[10] E.J. Hannan and P.J. Thomson, "Estimating group delay," *Biometrika*, vol. 60, no. 2, pp. 241–253, 1973.

[11] Hema A. Murthy and Venkata Gadde, "The modified group delay function and its application to phoneme recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),Hong Kong*, 2003, vol. 1, pp. I–68.

[12] Alan V Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *The Journal of the Acoustical Society of America (JASA)*, vol. 45, no. 2, pp. 458–465, 1969.

[13] Xiangwei Zhu, Yuanling Li, Shaowei Yong, and Zhaowen Zhuang, "A novel definition and measurement method of group delay and its application," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 1, pp. 229–233, 2008.

[14] Priyanka Gupta, Siddhant Gupta, and Hemant A. Patil, "Voice liveness detection using bump wavelet with CNN," in *9th International Conference on Pattern Recognition and Machine Intelligence*, 2021, Kolkata, India.

[15] Keiron O'Shea and Ryan Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015 {Last Accessed date : 20th February, 2024}.

[16] Courtney J Spoerer, Patrick McClure, and Nikolaus Kriegeskorte, "Recurrent convolutional neural networks: a better model of biological object recognition," *Frontiers in Psychology*, vol. 8, pp. 1551, 2017.

[17] Angjoo Kanazawa, Abhishek Sharma, and David Jacobs, "Locally scale-invariant convolutional neural networks," *arXiv preprint arXiv:1412.5104*, 2014, {Last Accessed : 9th February, 2024}.

[18] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al., "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[19] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[20] Siddhant Gupta, Kuldeep Khoria, Ankur T. Patil, and Hemant A. Patil, "Deep convolutional neural network for voice liveness detection," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, Tokyo, Japan., pp. 775–779.

[21] James William Topliss, Victor Zappi, Andrew McPherson, et al., "Latency performance for real-time audio on beaglebone black," 2014.

NecoBERT: Self-Supervised Learning Model Trained by Masked Language Modeling on Rich Acoustic Features Derived from Neural Audio Codec

Wataru Nakata* Takaaki Saeki* Yuki Saito* Shinnosuke Takamichi*† and Hiroshi Saruwatari*
* The University of Tokyo, Japan
E-mail: nakata-wataru855@g.ecc.u-tokyo.ac.jp
† Keio University, Japan

***Abstract*—We propose *NecoBERT,* a self-supervised learning (SSL) model to extract speech features containing semantic and acoustic information. Unsupervisedly learned representations using massive speech data, such as SSL and neural audio codec (NAC) features, are essential for developing speech foundation models. However, because conventional methods typically capture one specific aspect of speech, i.e., semantic or acoustic information, the learned features are not always suitable for complicated downstream tasks. NecoBERT addresses this issue by masked language modeling similar to HuBERT on NAC-derived rich acoustic features, enabling the acoustic features to represent semantic information as well. We evaluate NecoBERT on recognition tasks taken from the SUPERB benchmark and speech resynthesis tasks. The results show that although NecoBERT does not outperform HuBERT in some recognition tasks, it performs superiorly in speaker verification and speech resynthesis.**

I. Introduction

Speech pretraining using massive data is essential for the developing speech foundation models, i.e., universal spoken language processing (SLP) models [1]. One mainstream approach is self-supervised learning (SSL) [2]–[4], which trains large deep neural networks (DNNs) as a function to extract speech features useful for various downstream tasks [5]. The core technology is to learn mainly *phonetic or semantic* information from speech [6] during pretraining with context modeling, such as masked language modeling (MLM) [7] and contrastive learning [8]. Another approach is speech reconstruction to extract latent variables related to *acoustic* properties of speech. Neural audio codec (NAC) [9] is a typical example of this approach, which aims to compress input audio using residual vector quantization (RVQ) [10] and adversarial training [11] while losing as little of the fidelity of the original audio as possible.

Following the success in many speech *recognition* tasks covered by the SUPERB benchmark [5], SSL techniques have been transferred to speech *generation* tasks [12]–[14]. For instance, SSL features have recently been used as speech representations for speech synthesis, instead of traditional acoustic features such as mel-spectrograms [15], [16]. Performing k-means clustering [17] of speech features derived from a pretrained SSL model, e.g., HuBERT [3], enables us to define discrete speech tokens that can be used as intermediate representations used for speech synthesis [14]. Although high-fidelity speech synthesis involves modeling of both linguistic and non-/para-linguistic information, previous methods often suffer from poor intelligibility or unnatural prosody and thus result in the low fidelity of generated speech [18], [19]. This issue hinders potential applications of learned speech representations in more practical and complicated SLP tasks such as the modeling of spoken dialogue [20].

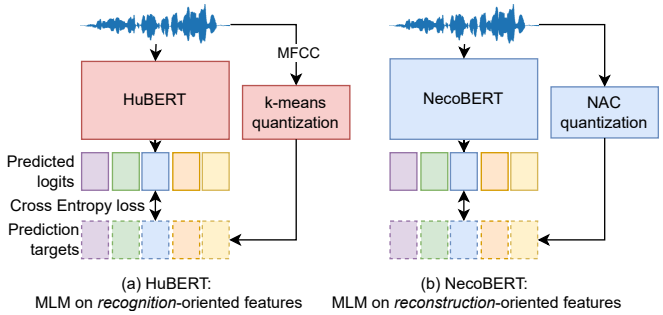


Fig. 1: Differences between (a) conventional HuBERT and (b) proposed NecoBERT.

To this end, we propose *NecoBERT* (**Ne**rual audio **co**dec **BERT**), an SSL model aimed at modeling both semantic and acoustic features of speech. Unlike the conventional HuBERT [3] which relies on the results of k-means clustering on mel-frequency cepstrum coefficients (MFCC), NecoBERT is trained to predict speech tokens quantized by a pretrained NAC model, as shown in Figure 1. This training involves MLM on NAC-derived features to contextualize them and enables the learned features to represent semantic information extracted from rich acoustic features. We evaluate NecoBERT in speech processing tasks including recognition tasks [5] as well as speech resynthesis (i.e., neural vocoding and unit-to-speech) to investigate what kind of information the NecoBERT-derived features represent. Our contributions are as follows.

- We propose *NecoBERT*, a novel speech SSL model that can learn both acoustic and semantic features. The training code and the trained model are available online[1].
- From the evaluation results of recognition tasks, we

[1]https://github.com/Wataru-Nakata/SSL4SpeechSynthesis

show that NecoBERT achieves the highest performance in speaker verification compared to NAC and HuBERT.
- From the evaluation results of speech resynthesis tasks, we demonstrate that NecoBERT improves F0 prediction performance and naturalness of resynthesized speech, regardless of whether we quantize the SSL features.

II. Related work

A. Conventional SSL model: HuBERT

Despite being initially designed for automatic speech recognition (ASR), SSL models have recently demonstrated remarkable performance as feature extractors for various SLP tasks [5]. These models undergo pretraining on abundant speech data initially, which serves as a foundation for subsequent applications across diverse downstream tasks.

HuBERT, for instance, exemplifies this process through two pretraining iterations [3]. Specifically, it first initializes a HuBERT model including Transformer encoders [21] with a convolutional layer by MLM on k-means-clustered MFCCs. Then, it refines the model through another round of MLM on the k-means-clustered 6th layer hidden states of the initial model.

B. NAC and its application in speech synthesis

NAC has recently been established as another approach to discretize audio signals using DNNs trained in an unsupervised manner. The learning process in NAC relies on the use of an information-discretizing bottleneck in an autoencoder structure [9], [22], [23]. By feeding an audio waveform into the encoder part of the trained NAC (i.e., NAC encoder), acoustic tokens descritized by an RVQ layer can be obtained.

As NAC features are expected to contain abundant information related to the acoustic property of original audio, they have been gathering attention as novel representations suitable for speech synthesis [13]. However, NAC faces challenges in capturing temporal dependencies between tokens, and there is no guarantee that the prosody of synthesized speech can be reproduced with high accuracy [19].

III. Proposed SSL Model: NecoBERT

In this work, we propose NecoBERT, which aims to model both semantic and acoustic features from speech. Figure 2 shows the overall model architecture and the training flow of NecoBERT. Our NecoBERT utilizes NAC features suitable for high-fidelity speech reconstruction and learns contextual/semantic information from them by MLM. This strategy is inspired by previous methods that perform MLM on pretrained SSL features, such as vq-wav2vec [2] and w2v-BERT [24].

A. Core architecture

NecoBERT consists of a pretrained NAC encoder and Transformer encoder [21]. The NAC encoder takes raw speech waveform as input and generates d-dimensional NAC latent features $X = \{\boldsymbol{x}_n \in \mathbb{R}^d | n = 1, \cdots, N\}$, where N denotes the number of frames in the feature. The Transformer encoder $f_\theta(\cdot)$ parameterized by θ predicts quantized NAC features (i.e., NAC tokens) $Z = \{\boldsymbol{z}_n \in \mathbb{N} | n = 1, \cdots, N\}$ from X with the positional encoding.

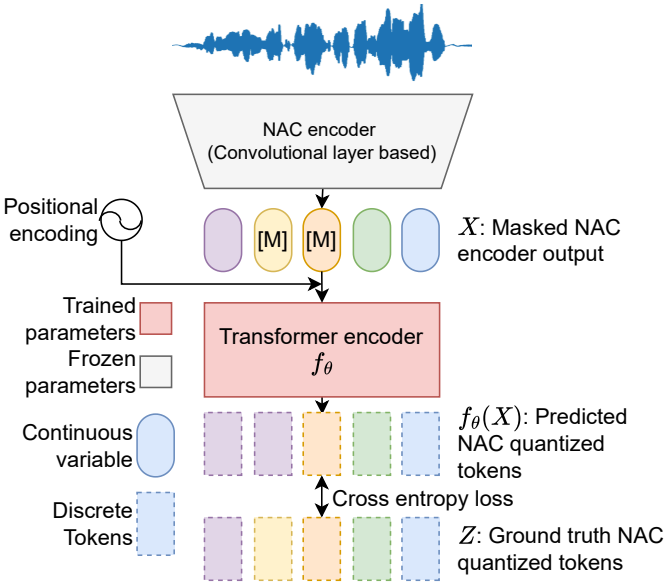


Fig. 2: Model architecture and the training flow of NecoBERT.

Unlike the original framework that uses not only speech but also music and environmental sounds [23], we pretrain the NAC on speech data only. This pretraining enables the NAC-discretized tokens to only capture speech related information, not general audio properties. We empirically observed that this speech-only pretraining significantly improved the speech reconstruction performance of NAC.

B. MLM-based token prediction training

The training of NecoBERT is based on MLM similar to that in existing speech SSL models, which aims to learn contextual information of speech from the NAC-derived latent features. We adopt SpanBERT-based masking [25] following previous work [3]. Specifically, we first select a random segment with a length of l from a target NAC token sequence with a probability of p and mask the selected segment. Then, we replace the masked tokens with a learnable parameter.

C. Training objective and inference

The training loss L is defined as the cross-entropy (CE) between the predicted and target token sequences:

$$L = \sum_{t \in \mathcal{T}} \mathrm{CE}\left(f_\theta(X), Z\right), \tag{1}$$

where T and $\mathcal{T}$ denote the frame length and the set of frame indices to calculate the loss, respectively. In the previous work [3], calculating loss only on the masked regions was effective in ASR tasks. However, its effectiveness in other tasks remains unclear. To see how the model performance differs depending on the loss calculation strategy, we prepared two variants of NecoBERT. The first one is NecoBERT, which is trained with loss calculated in all regions, i.e., $\mathcal{T} = \{1, \ldots, N\}$. The second one is NecoBERT-masked,

which is trained with loss calculated only in masked regions, i.e., $\mathcal{T} = \{m, \ldots, m+l\}$ where m denotes the frame to start the l-length masking.

After the training, we can use the last hidden state of the Transformer encoder in NecoBERT as a speech representation. The representation is expected to represent semantic information extracted from rich acoustic features, such as prosody of input speech.

IV. Experiment

In this work, we trained our NecoBERT models and evaluated their performance on recognition tasks and speech resynthesis from continuous SSL features or their discretized versions.

A. Experimental conditions

For the training, we used LibriSpeech [26] containing 960 hours of English multi-speaker speech samples from audiobooks. For NAC, we used Descript Audio Codec (DAC) [23] and trained it on the LibriSpeech corpus for 150k steps. We resampled all speech data to 24 kHz. For DAC parameters, we set the downsample rate of the encoder to $2, 3, 4, 4, 5$, which resulted in downsampling of $1/480$. The output of the DAC was then set to 50 Hz at an input sampling rate of 24 kHz. This output frequency of 50 Hz was the same as those used in HuBERT [3], WavLM [4], and wav2vec 2.0 [2]. For other parameters, we followed the official implementation of DAC[2].

There were 12 Transformer [21] encoder layers in NecoBERT, and their hidden sizes and number of attention heads were set to 768 and 12, respectively. This was the same configuration for wav2vec 2.0, WavLM-base, and HuBERT-base models. For the speech tokens used for the output of NecoBERT (i.e., Z), we only used the output of the first RVQ layer of DAC. This is because pretraining DAC on speech data only enabled the first layer to capture enough acoustic information in speech. Furthermore, in our preliminary experiments, we empirically confirmed that the use of multiple layer quantization results (i.e., multi-stream NAC tokens) made the pretraining unstable. Therefore, we only used output of the first RVQ layer of DAC in this study.

For optimization, we used AdamW [27], setting its parameters as $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-6}$,, and $\lambda = 1 \times 10^{-2}$. For the learning rate, we linearly increased it from 0 to 2×10^{-4} for first 40k steps and linearly decreased it from 2×10^{-4} to 0 for 460k steps. The total number of training steps is 500k. For MLM, we set the masking probability to $p = 0.08$ and the length of masking to $l = 10$. The training took approximately two days with eight NVIDIA A100 GPUs. NecoBERT consisted of approximately 72M and 91M parameters for the NAC encoder and Transformer encoder, respectively.

B. Evaluation in recognition tasks

We first evaluated our NecoBERT on recognition tasks.

[2]https://github.com/descriptinc/descript-audio-codec

TABLE I: Evaluation results on recognition tasks. Result for HuBERT is quoted from [5]. **Bold** and underlined values are the best and worst in each task, respectively.

Task type	ASR	ASV	ER	SD
Metric	WER↓	EER ↓	ACC ↑	DER ↓
HuBERT	**6.42**	5.11	**64.92**	**5.88**
DAC	29.8	2.60	48.31	12.99
NecoBERT	21.8	**1.80**	53.57	9.96
NecoBERT-masked	21.7	2.79	57.63	8.36

Task description: For recognition tasks, we used automatic speech recognition (ASR), automatic speaker verification (ASV), emotion recognition (ER), speaker diarization (SD), and intent classification (IC) from SUPERB [5].There tasks correspond to the recognition of content (ASR), speaker (ASV), paralinguistics (ER and SD), and semantic information from speech (IC). We selected these tasks to cover all task types introduced in SUPERB and to avoid high computational costs due to large datasets and diverse tasks [28].

Settings and metrics: For the training condition for each task, we followed the conditions presented in the SUPERB paper [5]. For the compared methods, we used the DAC encoder output and the two proposed models: NecoBERT and NecoBERT-masked described in Section III-C. We kept the parameters of these models unchanged and only used them as feature extractors. For NecoBERT and NecoBERT-masked, we took a weighted sum of the hidden states in the Transformer encoder. For the evaluation metrics, we used word error rate (WER) for ASR, equal error rate (EER) for ASV, classification accuracy (ACC) for ER and IC, and diarization error rate (DER) for SD.

Results: Table I shows the evaluation results on recognition tasks. Note that the result of HuBERT is quoted from the original SUPERB paper [5]. From the results, we can see that the NecoBERT improves all evaluation metrics compared to the DAC. One noteworthy result is that NecoBERT achieves the lowest EER among the compared models including HuBERT, the best-performing model regarding ASV-EER in the SUPERB paper. This result shows that through MLM, NecoBERT successfully captured additional semantic information over DAC. NecoBERT-masked performed better than original NecoBERT in all tasks except for the ASV tasks. One reason is that by calculating loss on only masked regions, the MLM training becomes closer to language modeling and thus the learned features change to contain more semantic information rather than acoustic information [3]. Additionally, HuBERT outperforms NecoBERT in most tasks except for the ASV. This result indicates that the HuBERT is actually trained to capture rich semantic information better than NecoBERT.

C. Evaluation in neural vocoding

We then evaluated NecoBERT on neural vocoding, a task of generating speech using *continuous* speech representations.

Task description: To evaluate the performance on the modeling of acoustic information, such as speaker identity and prosody, we performed neural vocoding from the continuous

TABLE II: Evaluation results on speech resynthesis tasks. **Bold** naturalness MOS values are significantly higher than the baseline results ($p < 0.05$). Underlined values denote the worst-performing ones in objective evaluations.

(a) Results for neural vocoding: resynthesis from continuous speech features

Feature extraction method	LogF0 RMSE ↓	MCD [dB]↓	XVector-sim ↑	Naturalness MOS ↑
Ground-truth	-	-	-	**3.82** ±0.187
Mel-spec.	2.82	2.43	0.987	3.54 ±0.222
HuBERT (baseline)	4.43	5.26	0.953	3.37 ±0.206
NecoBERT	2.71	3.24	0.976	**3.67** ±0.206
NecoBERT-masked	2.73	3.41	0.975	**3.79** ±0.201

(b) Results for u2s: resynthesis from discretized speech features

Feature extraction method	LogF0 RMSE ↓	MCD [dB] ↓	XVector-sim ↑	Naturalness MOS ↑
Ground-truth	-	-	-	**4.06** ±0.18
DAC tokens	4.88	5.34	0.888	**2.75** ±0.19
HuBERT (baseline)	7.70	7.61	0.830	1.94 ±0.17
NecoBERT	4.29	5.65	0.874	**2.78** ±0.21
NecoBERT-masked	6.23	6.52	0.855	1.25 ±0.19

features obtained from the trained SSL models. For this task, we used the HiFi-GAN vocoder [29] and trained it on the train-clean-100, clean-360, other-500 subset of LibriTTS [30]. The evaluation data were taken from the test-clean subset of LibriTTS. We compared mel-spectrogram (Mel-spec.) and the last layer hidden states of HuBERT, NecoBERT, and NecoBERT-masked as the input features of HiFi-GAN. The mel-spectrograms were extracted from the speech with a window length of 1024 and a hop length of 256. We added Mel-spec. to the compared methods because it is widely used for speech representation and as the training target in speech synthesis. We implemented HuBERT using the official weight of HuBERT-base distributed on Github[3]. This HuBERT model was trained on LibriSpeech, the same as NecoBERT and NecoBERT-masked.

Settings and metrics: We evaluated the quality of synthesized speech using objective and subjective metrics. For objective metrics, we used square-root mean squared error of log F0 (LogF0 RMSE), mel-cepstrum distortion (MCD) [31], and x-vector [32] cosine similarity (XVector-sim). These three metrics quantify the prediction performances of prosody, vocal timber, and speaker identity of target speech, respectively. We used the WavLM-base speaker verification model available on huggingface[4] to extract the x-vector from speech. We used WORLD vocoder [33] for the F0 analysis. For subjective metrics, we used naturalness mean opinion score (MOS). For this metric, we employed native English speakers via crowdsourcing using Prolific[5] and asked them to rate how natural the presented speech was on a 5-point scale from 1 (very poor) to 5 (very good). There were 60 raters, and each rater evaluated 10 samples.

Results: Table IIa shows the evaluation results on neural vocoding using continuous SSL features. From the objective evaluation results, Mel-spec. achieved the best MCD and XVector-sim values among the compared methods. However, focusing on the LogF0 RMSE values, our NecoBERT and NecoBERT-masked outperformed the others. This result suggests that the NecoBERT-derived features contain richer prosody information by contextualizing DAC-derived acoustic latent variables. From the subjective evaluation results, there were no statistically significant differences between Ground-truth, Mel-spec., NecoBERT and NecoBERT-masked. This result demonstrates that NecoBERT can achieve naturalness comparable to the ground truth. On the other hand, HuBERT performed the worst in this task and had significantly worse naturalness MOS than Ground-truth, NecoBERT, and NecoBERT-masked. This reveals that the HuBERT-derived features lack acoustic information to reconstruct original speech with high naturalness.

D. Evaluation in unit-to-speech (u2s)

Finally, we evaluated NecoBERT on u2s, a task of generating speech using *discrete* speech representations.

Task description: u2s was originally introduced in [1] and aims to resynthesize speech from its discretized representations. Although the descritization of speech enables us to introduce effective methods widely used in the natural language processing field, it may lose the fine structure of original speech that is required for the speech reconstruction. Therefore, we can regard u2s as a more challenging speech resynthesis task than neural vocoding using continuous SSL features. To evaluate the effectiveness of NecoBERT as a speech discretization method, we conducted an evaluation in u2s. In this task, we first extracted continuous features from the speech using NecoBERT in the same manner as in Section IV-C. The extracted features were then discretized using k-means clustering. The number of clusters was set to 1,000. We trained HiFi-GAN [29], which reconstructs the original speech from the discretized representation, using the same training data as described in Section IV-C. HiFi-GAN contained an embedding layer to convert discrete representation to continuous representation. The evaluation data were the test-clean

[3]https://github.com/facebookresearch/fairseq/blob/main/examples/hubert/README.md

[4]https://huggingface.co/microsoft/wavlm-base-plus-sv

[5]https://www.prolific.com/

subset of the LibriTTS corpus. We compared DAC encoder output (DAC tokens), 6th-layer hidden state of HuBERT, 12th-layer hidden state of NecoBERT, and 12th-layer hidden state of NecoBERT-masked. We used the same settings and metrics as described in Section IV-C. For HuBERT, we selected the 6th-layer hidden state following the previous work [1]. For this task, we added DAC tokens as a compared feature extraction method. This is because, in contrast to DAC tokens, mel-spectrogram is unsuitable for the discretization with k-means clustering.

Results: Table IIb shows the evaluation result on u2s. From objective evaluation results, HuBERT performed the worst among the compared methods, indicating that the HuBERT-derived features are inappropriate for speech resynthesis tasks (i.e., both neural vocoding and u2s). In contrast, DAC tokens performed reasonably well in this task with the best LogF0 RMSE and XVector-sim values. This observation is unsurprising because DAC indeed aims to learn quantized representations of speech with high speech reconstruction performance. Our NecoBERT performed comparably to DAC and had the lowest LogF0 RMSE value. This result suggests that the NecoBERT-derived features tend to capture speech prosody better than existing SSL models and NAC by using MLM on DAC-derived features. One noteworthy point is that NecoBERT-masked performed significantly worse than HuBERT in the subjective evaluation. This observation contradicts the result of neural vocoding, suggesting that the latent space of NecoBERT is more suited for discretization by k-means than that of NecoBERT-masked. In summary, the subjective evaluation results reveal a large gap between Ground-truth and u2s-generated speech samples in terms of their naturalness. Further improvement is required for speech resynthesis from the discrete features.

V. Conclusion

In this work, we proposed a novel speech self-supervised learning (SSL) model called *NecoBERT*. It performs masked language modeling (MLM) on neural audio codec (NAC)-derived features and aims to add contextual information on rich acoustic features. We evaluated NecoBERT on recognition tasks and speech resynthesis. The results showed that the while it performed worse in some recognition tasks, it performed better at speech resynthesis than conventional methods leveraging unsupervised learning with massive speech data. Future work includes developing a better discretization method suitable for our NecoBERT and evaluating our model on more challenging tasks such as text-to-speech.

Acknowledgement

This work was supported by JST Moonshot JPMJMS2011.

References

[1] K. Lakhotia, E. Kharitonov, W.-N. Hsu, *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, Vancouver, Canada, Dec. 2020, pp. 12 449–12 460.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[4] S. Chen, C. Wang, Z. Chen, *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[5] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. INTERSPEECH*, Brno, Czech Republic, Sep. 2021, pp. 1194–1198.

[6] T. Ashihara, T. Moriya, K. Matsuura, *et al.*, "SpeechGLUE: How well can self-supervised speech models capture linguistic knowledge?" In *Proc. INTERSPEECH*, Ireland, Dublin, Aug. 2023, pp. 2888–2892.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, U.S.A., Jun. 2019, pp. 4171–4186.

[8] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, vol. abs/1807.03748, 2018.

[9] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, Nov. 2021. DOI: 10.1109/TASLP.2021.3129994.

[10] A. Vasuki and P. Vanathi, "A review of vector quantization techniques," *IEEE Potentials*, vol. 25, no. 4, pp. 39–47, Jul. 2006.

[11] A. Biswas and D. Jia, "Audio codec enhancement with generative adversarial networks," in *Proc. ICASSP*, Barcerona, Spain, May 2020, pp. 356–360.

[12] W.-C. Huang, Y.-C. Wu, T. Hayashi, and T. Toda, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," *ICASSP*, pp. 5944–5948, 2020.

[13] Z. Borsos, R. Marinier, D. Vincent, *et al.*, "AudioLM: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2022.

[14] C. Wang, S. Chen, Y. Wu, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv*, vol. abs/2301.02111, 2023.

[15] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4779–4783.

[16] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI*, Honolulu, U.S.A., Jul. 2019, pp. 6706–6713.

[17] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[18] Z.-C. Liu, Z.-H. Ling, Y.-J. Hu, J. Pan, J.-W. Wang, and Y.-D. Wu, "Speech synthesis with self-supervisedly learnt prosodic representations," in *Proc. INTERSPEECH*, Dublin, Ireland, Aug. 2023, pp. 7–11.

[19] C. Du, Y. Guo, H. Wang, *et al.*, "VALL-T: Decoder-only generative Transducer for robust and decoding-controllable text-to-speech," *arXiv*, vol. abs/2401.14321, 2024.

[20] T. A. Nguyen, E. Kharitonov, J. Copet, *et al.*, "Generative spoken dialogue language modeling," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.

[21] A. Vaswani, N. M. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proc. NIPS*, 2017.

[22] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023, ISSN: 2835-8856.

[23] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Proc. NeurIPS*, 2023.

[24] Y.-A. Chung, Y. Zhang, W. Han, *et al.*, "w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proc. ASRU*, 2021, pp. 244–250.

[25] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2019.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, South Brisbane, Australia, Apr. 2015, pp. 5206–5210.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2017.

[28] Y.-H. Wang, H.-Y. Chen, K.-W. Chang, W. Hsu, and H.-y. Lee, "MiniSUPERB: Lightweight benchmark for self-supervised speech models," in *Proc. ASRU*, Taipei, Taiwan, Dec. 2023.

[29] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 17 022–17 033.

[30] H. Zen, V. Dang, R. Clark, *et al.*, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Proc. Interspeech 2019*, 2019, pp. 1526–1530. DOI: 10.21437/Interspeech.2019-2441.

[31] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. PACRIM*, 1993, pp. 125–128.

[32] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, *X-Vectors: Robust DNN embeddings for speaker recognition*, 2018.

[33] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

Privacy-Preserving Anomaly Detection in Bitstream Video based on Gaussian Mixture Model

Yike Chen, Yuru Song, Peijia Zheng, Yusong Du* and Weiqi Luo
Guangdong Provincial Key Laboratory of Information Security Technology,
MoE Key Laboratory of Information Technology,
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
* Corresponding author, email: duyusong@mail.sysu.edu.cn

***Abstract*—Advanced computer vision technologies enable nearly real-time intelligent monitoring of homes by detecting anomalies such as falls and incidents of domestic violence, thereby enhancing home safety. Leveraging affordable home cameras and cloud computing services, this technology offers significant societal benefits. However, privacy concerns present challenges for its deployment in real-world settings. This paper introduces a method for anomaly detection in encrypted bitstream videos. By analyzing the video compression standard, we incorporate new appearance information that captures variations in object sizes alongside existing motion information, enhancing anomaly detection capabilities. We also refine the feature extraction process to reduce the impact of noise. Finally, we employ Gaussian Mixture Model to model the probability distribution of the data for effective anomaly detection. Experimental results demonstrate that our proposed privacy-preserving anomaly detection method achieves a commendable balance between protecting privacy and maintaining detection performance.**

I. Introduction

As camera prices continue to decrease, video surveillance has become increasingly prevalent, with more individuals using it to ensure their safety, thereby driving the demand and development of private smart monitoring services. However, managing the enormous data volume of surveillance videos and performing intelligent video analysis on resource-constrained local devices is challenging. Cloud servers offer powerful computing capabilities, mature video analysis models, and are willing to provide users with various video analysis and processing tasks [1] for a proper fee.

However, surveillance videos, especially home surveillance videos, contain a significant amount of privacy-sensitive information such as personal identities, family activities, and daily routines. The leakage of this privacy information poses security risks to users of cloud services, making them reluctant to upload home surveillance videos. As a result, cloud servers face the challenging dual responsibility of respecting personal privacy while enhancing the user experience.

Processing videos in the encrypted domain [2] is one of the promising solutions to the aforementioned issues. Users encrypt videos before uploading, in which the cloud servers process the encrypted videos and return the processed results to users. However, directly encrypting raw videos entails significant computational costs. Surveillance videos are similar to continuous movies, generating a vast number of frames daily. Users may not have adequate computational resources to encrypt raw videos. Methods proposed by [3]–[6] advocate encrypting compressed videos, greatly improving the efficiency of video encryption and analysis. These methods compress videos first and then encrypt the compressed bitstream, reducing the amount of data that needs encryption and thereby significantly lowering encryption costs for users.

Existing encrypted domain video analytics, such as moving object detection, have been extensively researched [4]. However, there is limited research on anomaly detection in encrypted bitstream videos. Guo *et al.* [6] used motion information from bitstreams and Adaptive Kernel Density Estimation (AKDE) for anomaly detection, but their focus solely on motion information is insufficient for complex scenes.

In this paper, we build on [6] by introducing new appearance features that partially capture changes in object size, thereby complementing motion information to enhance anomaly detection. We observe there is significant noise in the information extracted from encrypted videos, prompting us to redesign our feature extraction method to mitigate noise impact. To address the challenges of AKDE in handling high-dimensional data within encrypted compressed videos, we employ a Multivariate Gaussian Mixture Model (GMM) [7] for anomaly detection. By inputting high-dimensional features, the model learns intricate relationships between different features, thereby achieving better anomaly detection performance. The main contributions of this paper are as follows:

1. We introduce a new encrypted video information to complement motion information in anomaly detection by partially representing changes in object appearance.
2. We propose new feature extraction methods and use the Multivariate GMM to better capture potential relationships between features.
3. Experimental results show that the proposed privacy-preserving anomaly detection method in bitstream video achieves higher detection accuracy than the existing state-of-the-art method on multiple datasets.

II. Related Works

A. Anomaly Detection Methods

Traditional anomaly detection methods often utilize handcrafted features, such as the hierarchical Bayesian model [8], mixture of dynamic textures [9], and sparse reconstruction

[10]. Recently, Hasan *et al.* [11] use a fully convolutional autoencoder to learn normal patterns. To address the issue that both normal and abnormal events are reconstructed well, Gong *et al.* [12] introduce a memory-augmented deep autoencoder. Yang *et al.* [13] use keyframes to restore video events and explore temporal relationships within the video. Furthermore, to enhance the speed of anomaly detection, some methods operate directly in the compressed domain [14], [15]. However, these methods neglect the need for privacy preservation in anomaly detection.

B. Privacy-Preserving Anomaly Detection Methods

Cheng *et al.* [16] propose a secure video anomaly detection scheme using secret sharing and Bloom filter, but its generalizability is limited. Yan *et al.* [17] mask privacy-sensitive regions in decompressed video data to protect privacy, but their approach is computationally complex and offers limited privacy protection. Allahham *et al.* [18] utilize federated learning techniques, but the method incurs high computational costs, making it unsuitable for real-time processing. Guo *et al.* [6] propose an anomaly detection scheme for encrypted compressed videos, but their scheme lacks robustness and shows a significant performance drop in complex scenarios. Therefore, designing a real-time and robust anomaly detection scheme on encrypted video remains a challenging task.

III. Problem Statement

A. System Model

As illustrated in Figure 1, our encrypted video anomaly detection system involves two main entities: the data owner, $\mathcal{O}$, who possesses the original video data but has limited storage and computational capabilities. The cloud computing service provider, $\mathcal{S}$, acts as a third-party entity, offering $\mathcal{O}$ robust storage and computational resources at a reasonable cost for VAD cloud services. By outsourcing the video data, $\mathcal{O}$ can access the video anytime and anywhere through terminal devices such as smartphones, and benefit from the video anomaly detection cloud services. The cloud server $\mathcal{S}$ performs anomaly detection tasks on the encrypted bitstream video and returns the detection results to $\mathcal{O}$. $\mathcal{O}$ decrypts the results using a key to obtain the anomaly detection outcomes.

B. Threat Model

We adopt the setting established in similar literature [19], assuming $\mathcal{S}$ to be a semi-honest adversary. This means $\mathcal{S}$ will follow the protocol but may attempt additional computations to access private information from the video. We employ the video encryption scheme from [5], [20], which has been extensively validated for its ability to protect video privacy. Only $\mathcal{O}$ possesses the decryption key; $\mathcal{S}$ does not have access to the key. During the execution of anomaly detection, $\mathcal{O}$ does not interact with $\mathcal{S}$, preventing leakage of the video's key information. Therefore, $\mathcal{S}$ cannot obtain sensitive information from the encrypted video. Our encrypted video anomaly detection system ensures security.

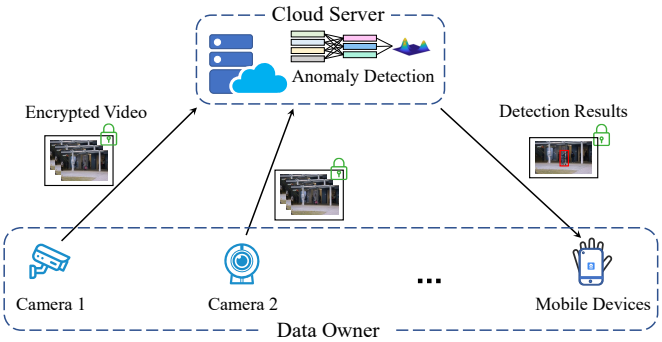


Fig. 1. System model.

IV. Proposed Privacy-Preserving Anomaly Detection Scheme over Encrypted Video

The proposed privacy-preserving video anomaly detection framework is illustrated in Figure 2. To facilitate our discussion, we focus on H.264 videos and use a selective encryption scheme to encrypt the compressed bitstreams, as detailed in [5]. It should be noted that video encryption algorithms compatible with other formats [21] are also applicable to our framework. Our framework directly analyzes encrypted video bitstreams using compression standards, extracting entropy-coded data related to anomalies. From entropy-coded data, we employ statistical analysis methods to extract abnormal behavioral characteristics for each frame in the compressed domain. These extracted compressed domain features serve as representations of each frame. We utilize a Multivariate GMM to construct a probability distribution of abnormal events and employ this distribution to detect abnormal frames. We will now elaborate on each specific part.

A. Abnormal Information Estimation on Encrypted Video

By analyzing video compression standards, we directly extract syntax elements from entropy-coded data to reveal their correlations with abnormal behaviors. Specifically, in addition to the three types of compressed domain abnormal behavior information identified in [6], we introduce a new type: residual density, which effectively captures appearance changes associated with object size variations. Next, we will provide detailed explanations of these four types of compressed domain abnormal behavior information.

1) Consumed Bits: In common video compression standards, the basic coding unit is the macroblock. Video compression aims to represent more content with less data, so repetitive content is predicted to reduce the number of bits per macroblock. In contrast, during abnormal events, affected macroblocks often consume more bits due to unusual object motion. Thus, the bit consumption of macroblocks reveals information about abnormal behaviors.

In the H.264 encoding scheme, its slice data syntax structure is represented as:

$$\left\{\mathbf{SH}, \mathbf{MB}'_1, \mathbf{MB}'_2, \mathbf{SR}_3, \cdots, \mathbf{MB}'_i, \cdots, \mathbf{MB}'_n\right\},$$

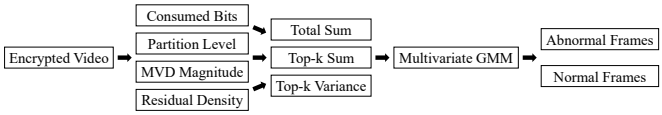


Fig. 2. Overview of the proposed framework.

where **SH** denotes slice header, $\mathbf{MB}'_i$ represents the i-th encrypted macroblock in the slice, and $\mathbf{SR}_3$ indicates that the 3rd macroblock is a skipped macroblock.

The number of bits consumed $b_i(t)$ by the i-th macroblock in frame t is calculated as:

$$b_i(t) = \begin{cases} 0, & \text{skipping } \mathbf{MB}_i, \\ \mathcal{S}(\mathbf{SR}_{i+1}) - \mathcal{S}(\mathbf{MB}'_i), & \text{skipping } \mathbf{MB}_{i+1}, \\ \mathcal{S}(\mathbf{MB}'_{i+1}) - \mathcal{S}(\mathbf{MB}'_i), & \text{otherwise}, \end{cases} \quad (1)$$

where $\mathcal{S}(\cdot)$ denotes the starting address in the bitstream.

2) Partition Level: In compressed videos, b_i in I-frames is generally significantly higher than in P-frames. Additionally, the aggregation or dispersion of normal objects affects b_i. Consequently, relying solely on b_i can result in a high false positive rate in anomaly detection. Typically, 16×16 macroblocks can be subdivided into smaller sub-MBs. To better capture motion during decoding, macroblocks in abnormal areas of P-frames may be divided into smaller sub-MBs. For instance, a regular macroblock might be divided into two 8×16 sub-MBs, whereas a macroblock in an abnormal area may be divided into sixteen 4×4 sub-MBs. The partition level $p_i(t)$ of a macroblock is calculated as:

$$p_i(t) = \mathcal{P}\left(\mathbf{ST}_i, \mathbf{MT}_i\right), \quad (2)$$

where $\mathbf{ST}_i$ denotes the slice type, $\mathbf{MT}_i$ denotes the MB type, and $\mathcal{P}(\cdot,\cdot)$ determines the partition level based on the number of its sub-MBs.

3) MVD Magnitude: Motion vectors (MVs) are widely used features in bitstream-based video anomaly detection [22], [23]. However, MVs in encrypted bitstream videos are unavailable. This paper selects MVD magnitude as a substitute for MV. In bitstream videos, MVs are not directly encoded but are indirectly encoded using predictive motion vectors and actual motion vector differences (MVDs). MVD consists of horizontal $\mathrm{MVD_x}$ and vertical $\mathrm{MVD_y}$ components. We use the magnitude of MVDs from a macroblock as an estimate for MVs to provide richer motion information for anomaly detection. The MVD magnitude $m_i(t)$ of a macroblock is calculated as

$$m_i(t) = \sum_{k \in \mathbb{I}_{mv}(t)} \lfloor \sqrt{L^k_{\mathrm{x}}(t)^2 + L^k_{\mathrm{y}}(t)^2} \rfloor, \quad (3)$$

where $\mathbb{I}_{mv}(t)$ represents the set of addresses for 4×4 blocks within the macroblock, and $L^k_{\mathrm{x}}(t)$ and $L^k_{\mathrm{y}}(t)$ represent the code word lengths of $\mathrm{MVD_x}$ and $\mathrm{MVD_y}$, respectively.

4) Residual Density: Residual density was proposed for detecting moving objects in videos [5]. We introduce residual density as appearance information to complement motion-based anomaly detection. During video compression, residual data undergoes DCT transformation, quantization, and then entropy coding. Given a macroblock, if its DCT block $U^k_{dct}(t)$ contains δ_k 4×4 blocks, the density of non-zero coefficients $d_i(t)$ is calculated as:

$$d_i(t) = \sum_{k \in \mathbb{I}_{dct}(t)} \delta_k \cdot \lfloor \frac{\rho\left(U^k_{dct}(t)\right)}{\sqrt{\delta_k}} \rfloor, \quad (4)$$

where $\rho(\cdot)$ denotes the number of non-zero coefficients in a DCT block, and $\mathbb{I}_{dct}(t)$ denotes the set of DCT block addresses within the macroblock.

B. Feature Extraction from Encrypted Compressed Videos

We extract four types of estimated information from encrypted compressed videos: consumed bits $b_i(t)$, partition level $p_i(t)$, MVD magnitude $m_i(t)$, and residual density $d_i(t)$. We use $\mathbf{f}_i(t)$ to denote any one of these four types of information for the i-th macroblock in frame t, and $\mathbb{I}(t)$ denotes the set of macroblock addresses in frame t. Next, we consider these informational features at the frame level, including the sum of all macroblocks, the sum of representative macroblocks, and the variance of representative macroblocks.

1) Total Sum: For the estimated information $\mathbf{f}(t)$, we compute the total sum of all macroblocks in a frame as a rough feature of that frame, expressed as:

$$\lambda^{\mathbf{f}}_t = \sum_{i \in \mathbb{I}(t)} \mathbf{f}_i(t), \quad (5)$$

2) Top-k Sum: During abnormal events, not all macroblocks in a frame are related to anomalies. In fact, abnormal events only involve a small number of macroblocks with prominent features. Based on this observation, we isolate the top α ranked macroblocks in a frame, termed representative macroblocks. Considering only the information from representative macroblocks can mitigate noise interference from non-representative macroblocks. The sum of information extracted from representative macroblocks in a frame is considered as a refined feature, computed as:

$$\delta^{\mathbf{f}}_t = \sum_{i \in \mathbb{I}_\alpha(t)} \mathbf{f}_i(t), \quad (6)$$

where $\mathbb{I}_\alpha(t)$ denotes the set of addresses of the top α ranked macroblocks in a frame.

3) Top-k Variance: Based on the identified representative macroblocks, in contrast to [6], we consider the variance among representative macroblocks. Since anomalies result in increased variance among representative macroblocks, this variance can be utilized as an additional refined feature for anomaly detection, which is computed by:

$$\phi^{\mathbf{f}}_t = \frac{1}{|\mathbb{I}_\alpha(t)| - 1} \sum_{i \in \mathbb{I}_\alpha(t)} \left(\mathbf{f}_i(t) - \bar{\mathbf{f}}_i(t)\right)^2, \quad (7)$$

where $\bar{\mathbf{f}}_i(t)$ denotes the mean value of $\mathbf{f}_i(t)$ over all representative macroblocks.

Thus, we obtain a total of 12 features:

$$[\lambda_t^c, \lambda_t^p, \lambda_t^m, \lambda_t^d, \delta_t^c, \delta_t^p, \delta_t^m, \delta_t^d, \phi_t^c, \phi_t^p, \phi_t^m, \phi_t^d],$$

which is represented collectively as $[\Lambda_t, \Delta_t, \Phi_t]$. All features together form a high-dimensional feature that represents the frame and is used for subsequent anomaly detection.

C. Anomaly Detection Algorithm

We use the Multivariate GMM for its ability to capture normal patterns in the data, thereby fully exploiting the potential relationships between different features. Specifically, during the training phase, we use the extracted features mentioned above to construct a probability density function using Multivariate GMM based on the training dataset, i.e.,

$$\begin{aligned} \mathbf{x}_t &= [\Lambda_t, \Delta_t, \Phi_t]^T, \\ p_{\mathcal{M}}(\mathbf{x}_t) &= \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}\left(\mathbf{x}_t \mid \boldsymbol{\mu_k}, \Gamma_k\right), \\ \sum_{k=1}^{K} \pi_k &= 1, \quad \pi_k > 0, \end{aligned} \tag{8}$$

where $\mathcal{N}\left(\mathbf{x}_t \mid \boldsymbol{\mu_k}, \Gamma_k\right)$ represents the probability density function of the k-th Gaussian component with mean $\boldsymbol{\mu_k}$ and covariance matrix Γ_k. K denotes the number of Gaussian components, and π_k represents the weight for the k-th Gaussian component.

The obtained GMM provides a representation of the probability of normal behaviors in the video data. During the testing phase, using the high-dimensional features $[\lambda_t, \Delta_t, \Phi_t]$ related to anomalies in each frame of the test video and the obtained GMM, we can compute the probability p_t of an event occurring in that frame. Frames with low probabilities are identified as potential anomalies, indicating deviations from the learned normal patterns. Based on the probability of event occurrence, the confidence s_t that a frame is an anomaly is computed as follows:

$$s_t = 1 - p_{\mathcal{M}}(\mathbf{x}_t). \tag{9}$$

After obtaining the anomaly confidence for each frame, and considering that abnormal events in videos typically develop gradually, we apply median filtering to smooth the confidence values. Finally, by setting an appropriate threshold θ, frames with confidence $s_t \geq \theta$ are classified as anomalous frames, while frames with $s_t < \theta$ are classified as normal frames.

V. Experimental Results

A. Settings

The proposed privacy-preserving anomaly detection scheme is tested on the UCSD Ped1 [9], UCSD Ped2 [9] and CUHK Avenue [24] datasets. For convenience in processing, we resize the video resolution to 1280×720 and set the frame rate to 25 fps. Feature extraction uses $\alpha = 4\%$, and the number of components k in the Multivariate GMM is set to 6, balancing computational complexity and probability fitting effectiveness. All experimental results are obtained using MATLAB on a 64-bit Windows 10 PC with Intel(R) Core(TM) i5-10400 CPU @ 2.90GHz and 16GB memory.

TABLE I
AUC and EER comparison with different methods.

Dataset	*UCSD Ped1*		*UCSD Ped2*		*CUHK Avenue*	
Method	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓
Kmeans [25]	0.59	0.44	0.57	0.53	0.70	0.36
iForest [26]	<u>0.70</u>	<u>0.37</u>	<u>0.76</u>	<u>0.29</u>	0.73	0.32
Guo [6]	0.69	0.38	-	-	<u>0.79</u>	<u>0.29</u>
Proposed	**0.72**	**0.35**	**0.81**	**0.24**	**0.80**	**0.26**

TABLE II
Performance comparison between plaintext and encrypted domains. CB and FS represent inputs of compressed bitstream and frame sequence, respectively.

Plaintext Domain					
Method	Input	*UCSD Ped1*		*CUHK Avenue*	
		AUC↑	EER↓	AUC↑	EER↓
USTN-DSC [13]	FS	-	-	**0.90**	-
Kiryati [14]	CB	-	-	0.72	0.34
Biswas [15]	CB	**0.79**	**0.24**	0.69	0.34
Guo [6]	CB	0.69	0.38	0.79	0.29
Proposed	CB	<u>0.72</u>	<u>0.35</u>	<u>0.80</u>	<u>0.26</u>
Encrypted Domain					
Method	Input	*UCSD Ped1*		*CUHK Avenue*	
		AUC↑	EER↓	AUC↑	EER↓
MVDM [6]	CB	0.62	0.43	0.65	0.40
MVDF [6]	CB	0.58	0.45	0.53	0.47
Guo [6]	CB	<u>0.69</u>	<u>0.38</u>	<u>0.79</u>	<u>0.29</u>
Proposed	CB	**0.72**	**0.35**	**0.80**	**0.26**

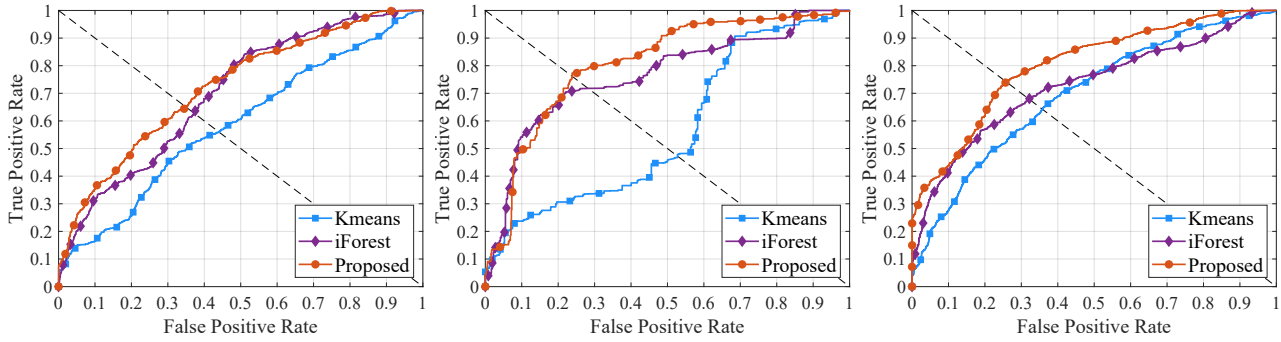
B. Datasets

1) UCSD Peds: The UCSD Peds [9] dataset was captured on a campus and comprises two sub-datasets: UCSD Ped1 and UCSD Ped2. UCSD Ped1 includes 70 videos of size 238×158, captured from a camera with a perpendicular viewpoint to the road, leading to significant variations in the size of moving foreground objects. UCSD Ped2 includes 28 videos of size 360×240, captured from a camera with a parallel viewpoint to the road, resulting in smaller size variations of moving foreground objects. Anomalies in these datasets are related to unexpected objects and pedestrian behaviors.

2) CUHK Avenue: The CUHK Avenue [24] dataset is captured on a campus avenue using a camera with slight jitter. It comprises 28 videos with a resolution of 640×320. Anomalies in this dataset mainly involve abnormal movements such as sudden running and littering.

C. Comparison results

We quantitatively evaluate the video results using widely used metrics in video anomaly detection: Area Under the Curve (AUC) and Equal Error Rate (EER), both obtained from receiver operating characteristic (ROC) curves. A higher AUC indicates better anomaly detection performance, while a lower EER reflects better performance. Quantitative comparisons of the proposed method's detection performance in the encrypted



(a) UCSD Ped1 (b) UCSD Ped2 (c) CUHK Avenue

Fig. 3. Comparison of ROC curves for different methods.

domain, compared with other methods, are shown in Table I. The corresponding ROC curves are depicted in Figure 3.

The proposed method achieves AUC values of 0.72, 0.81, and 0.80, and EER values of 0.35, 0.24, and 0.26 on the Ped1, Ped2, and Avenue datasets, respectively. Comparisons with Guo [6] show AUC improvements of 3% and 1% on the UCSD Ped1 and CUHK Avenue datasets, respectively. Quantitative comparisons show that our method outperforms Guo's, achieving more robust anomaly detection.

In Table I, we also compare our method with other methods, such as Kmeans [25] and iForest [26], using the features extracted in this paper for anomaly detection. On the UCSD Ped2 dataset, our method achieves an AUC 5% higher than the second-best method. Experimental results indicate that the proposed method achieves state-of-the-art performance across all three anomaly detection datasets.

In Figure 3, the proposed method consistently positions above and to the left of the ROC curves, indicating better representation of data probability distributions and thorough exploration of potential relationships between different features, resulting in robust anomaly detection capabilities.

We also compare our method with plaintext domain methods such as Kiryati [14], Bitwas [15], and USTN-DSC [13]. As shown in Table II, experimental results indicate that USTN-DSC [13], a deep learning-based method on frame sequences, achieved the best anomaly detection performance on the CUHK Avenue dataset due to its powerful representation capabilities using neural networks. However, our method performs the best among compressed domain methods and shows comparable performance to deep learning-based methods. On UCSD Ped1, our method also demonstrates performance comparable to similar methods in the compressed domain. For the encrypted video, we compare our results with those reported in [6], demonstrating that the proposed method achieves superior anomaly detection performance compared to other methods.

D. Ablation Study

We conduct ablation experiments to evaluate the effectiveness of residual density for representing appearance information and top-k variance features. As shown in Table III, removing top-k variance features and appearance information results in a reduced AUC of 0.76 and an increased EER of 0.30 on UCSD Ped2, highlighting their importance. Incorporating top-k variance features with the three types of estimated information improves the AUC to 0.77 and reduces the EER to 0.28, demonstrating the efficacy of including variance features.

TABLE III
ABLATION STUDIES ON UCSD PED2 DATASET.

Method		UCSD Ped2	
Top-k Variance	Residual Density	AUC↑	EER↓
✘	✘	0.76	0.30
✔	✘	0.77	0.28
✘	✔	0.78	0.28
✔	✔	**0.81**	**0.24**

TABLE IV
INFLUENCE OF DIFFERENT α VALUES ON AUC AND EER ON UCSD PED2 DATASET.

α	1%	2%	3%	4%	5%	6%	10%
AUC↑	0.74	0.78	0.79	**0.81**	0.80	0.79	0.73
EER↓	0.33	0.26	0.30	**0.24**	**0.24**	0.27	0.36

Incorporating appearance information alone boosts the AUC to 0.78 and reduces the EER to 0.28. With the addition of appearance information and top-k variance features, the model achieves an AUC of 0.81 and an EER of 0.24 on UCSD Ped2, highlighting the benefits of both enhancements.

Furthermore, we examine the impact of different values of α on the performance of representative macroblocks in anomaly detection, as shown in Table IV. We observe that the choice of α significantly affects model performance. Increasing α from a small value of 1% to 4% improves the AUC from 0.74 to 0.81. However, further increasing α results in decreased model performance because macroblocks selected with an α that is too large are influenced more by noise from regions unrelated to anomalies. Therefore, selecting an appropriate value for α is crucial for effective anomaly detection performance.

VI. CONCLUSIONS

We propose an anomaly detection algorithm for encrypted bitstream videos based on the Multivariate GMM. To address the limitation of existing methods, where motion information alone is insufficient to represent anomalies, we introduce

complementary appearance information to enhance detection capabilities. We redesign the feature extraction method to reduce the impact of noise in the encrypted domain on the feature extraction process. Additionally, by integrating the extracted features with the Multivariate GMM, we develop a more robust anomaly detection model for encrypted bitstream videos, particularly in complex scenarios. In future work, we plan to explore end-to-end anomaly detection methods based on deep learning for encrypted bitstream videos, aiming to reduce reliance on manual feature design.

ACKNOWLEDGMENT

This work is supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515030087 and Grant 2022A1515011512, the National Natural Science Foundation of China under Grant 62272498, and the Guangdong Provincial Key Laboratory of Information Security Technology (No. 2023B1212060026). Thanks are due to Dr. Xianhao Tian for assistance with the experiments.

REFERENCES

[1] B. Alouffi, M. Hasnain, A. Alharbi, W. Alosaimi, H. Alyami, and M. Ayaz, "A systematic literature review on cloud computing security: Threats and mitigation strategies," *IEEE Access*, vol. 9, pp. 57 792–57 807, 2021.

[2] C. Lin, K. Muchtar, J. Lin, Y. Sung, and C. Yeh, "Moving object detection in the encrypted domain," *Multim. Tools Appl.*, vol. 76, no. 7, pp. 9759–9783, 2017.

[3] X. Tian, P. Zheng, and J. Huang, "Robust privacy-preserving motion detection and object tracking in encrypted streaming video," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 5381–5396, 2021.

[4] C. Liu, X. Ma, S. Cao, J. Fu, and B. B. Zhu, "Privacy-preserving motion detection for HEVC-compressed surveillance video," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 18, no. 1, pp. 1–27, 2022.

[5] X. Tian, P. Zheng, and J. Huang, "Secure deep learning framework for moving object detection in compressed video," *IEEE Trans. Dependable Secur. Comput.*, vol. 21, no. 4, pp. 2836–2851, 2024.

[6] J. Guo, P. Zheng, and J. Huang, "Efficient privacy-preserving anomaly detection and localization in bitstream video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3268–3281, 2020.

[7] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *ICCV*, 2007.

[8] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception by hierarchical bayesian models," in *CVPR*, 2007.

[9] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR*, 2010.

[10] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR*, 2011.

[11] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *CVPR*, 2016.

[12] D. Gong, L. Liu, V. Le, *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *ICCV*, 2019.

[13] Z. Yang, J. Liu, Z. Wu, P. Wu, and X. Liu, "Video event restoration based on keyframes for video anomaly detection," in *CVPR*, 2023.

[14] N. Kiryati, T. Riklin-Raviv, Y. Ivanchenko, and S. Rochel, "Real-time abnormal motion detection in surveillance video," in *ICPR*, 2008.

[15] S. Biswas and R. V. Babu, "Anomaly detection in compressed H.264/AVC video," *Multim. Tools Appl.*, vol. 74, no. 24, pp. 11 099–11 115, 2015.

[16] H. Cheng, X. Liu, H. Wang, Y. Fang, M. Wang, and X. Zhao, "Securead: A secure video anomaly detection framework on convolutional neural network in edge computing environment," *IEEE Trans. Cloud Comput.*, vol. 10, no. 2, pp. 1413–1427, 2022.

[17] J. Yan, Y. Yang, and S. M. Naqvi, "Object detection oriented privacy-preserving frame-level video anomaly detection," in *ICASSP*, 2024.

[18] A. Al-lahham, M. Z. Zaheer, N. Tastan, and K. Nandakumar, "Collaborative learning of anomalies with privacy (clap) for unsupervised video anomaly detection: A new baseline," in *CVPR*, 2024.

[19] S. Hu, Q. Wang, J. Wang, Z. Qin, and K. Ren, "Securing SIFT: privacy-preserving outsourcing computation of feature extractions over encrypted image data," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3411–3425, 2016.

[20] F. Peng, X. Gong, M. Long, and X. Sun, "A selective encryption scheme for protecting H.264/AVC video in multimedia social network," *Multim. Tools Appl.*, vol. 76, no. 3, pp. 3235–3253, 2017.

[21] Y. Wang, M. O'Neill, and F. Kurugollu, "A tunable encryption scheme and analysis of fast selective encryption for CAVLC and CABAC in H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 9, pp. 1476–1490, 2013.

[22] S. Biswas and R. V. Babu, "Real time anomaly detection in H.264 compressed videos," in *NCVPRIPG*, 2013.

[23] S. Biswas and R. V. Babu, "Anomaly detection in compressed H.264/AVC video," *Multim. Tools Appl.*, vol. 74, no. 24, pp. 11 099–11 115, 2015.

[24] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *ICCV*, 2013.

[25] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *5th Berkeley Symp. Math. Stat. Prob.*, 1967.

[26] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 1, pp. 1–39, 2012.

Targeted Representation with Information Disentanglement Encoding Networks in Tasks

Takumi NAGAWAKI* Keisuke IKEDA* Kohei CHIKE† Hiroyuki NAGANO† Masaki NOSE† Satoshi TAMURA*
* Gifu University, Gifu, Japan
E-mail: {nagawaki,ikeda,tamura}@asr.info.gifu-u.ac.jp
† RICOH, Kanagawa, Japan
E-mail: {kohei.chike,hiroyuki.nagano,masaki.nose}@jp.ricoh.com

***Abstract*—This paper proposes a novel model TRIDENT to obtain better embeddings for speaker, emotion and speech recognition. Conventional features including HuBERT are often chosen to represent speech features for a task, however, such the features still contain the other information irrelated to the task. Our hypothesis is that, if we can disentangle them, better embeddings and higher performance are expected. Our model is designed to obtain speaker-, emotion-, and context-related features exclusively, by enhancing HuBERT features. In addition, these embeddings can be used to reconstruct the original features. We carried out speaker verification, emotion recognition and speech recognition experiments using task-specific embeddings, respectively. The results demonstrated that our scheme could appropriately extract features for various speech processing tasks simultaneously, achieving enough accuracy.**

I. Introduction

Speech has a lot of information such as contents, individuality and emotion. In order to detect theae elements, speech technology has been investigated for decades. Automatic Speech Recognition (ASR) transcribes speech contents. Speaker Recognition (SR) recognizes who spoke given speech waveforms. Speech Emotion Recognition (SER) detects emotion information such as anger, sad and happy, from speech signals. To accomplish these techniques, several kinds of features and models have been proposed. For example, mel-frequency cepstral coefficients and hidden Markov models were used for ASR. Mel-spectral features were used in SR and SER, while machine learning methods such as support vector machine were adopted. Nowadays it is common to employ deep-learning models for the classification and recognition tasks. Deep learning is also utilized to extract more effective features for each purpose.

One of the state-of-the-art deep-learning techniques, Self-Supervised Learning (SSL), has attracted attentions of speech-processing researchers. SSL builds a model on a task, utilizing unlabelled data efficiently. Recent advances in SSL have shown great potential in capturing the diverse attributes from speech [1]. Among such the works, we focus on Hidden Unit BERT (HuBERT) [2]. HuBERT converts speech signals into time-series vectors as useful embeddings. HuBERT embeddings are information-rich, encapsulating phonetic, speaker, and prosodic features, making it a powerful tool for speech processing applications. HuBERT is often employed in many applications for example ASR or SER, however, it is not guaranteed that HuBERT embeddings are designated to have only the information on the target task; sometimes the embeddings encapsulate various attributes together. Therefore, extracting only task-specified features such as speaker identity, emotional, and contextual information is still challenging.

To address this issue, we propose a novel approach TRIDENT. Our model is designed to calculate three kinds of embeddings, each exclusively having speaker-, emotion-, and context-related information. We thus expect that we can obtain better representations and results in ASR, SR and SER respectively, by disentangling HuBERT outputs. In addition, the model has an ability to reconstruct the original HuBERT embeddings after once separating the features, so that we can improve the clarity and consistency of the extracted features much more. We consider that this approach not only enhances speech understanding but also has the potential to significantly improve applications in personalized speech synthesis, emotion recognition, and context-aware systems. In this paper, we introduce our TRIDENT model, and demonstrate its ability to disentangle and re-synthesize speech embeddings. Finally, we highlight its potential contributions to the field of speech processing.

The rest of this paper is organized as follows: Section II briefly introduces related works to our work. The methodology appears in Section III. To evaluate our scheme, experiments were conducted and details are described in Section IV. Finally Section V concludes this paper.

II. Related Work

A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units [2]

Hidden Unit BERT (HuBERT) is a self-supervised learning framework designed to capture diverse speech representations by learning hidden units from unlabeled speech data. HuBERT can learn from unlabeled audio data, which means that a large amount of raw audio data can be used effectively. Therefore, the model expected to understand fundamental structures and patterns of speech, making it applicable to various audio processing tasks. HuBERT clusters audio data to generate

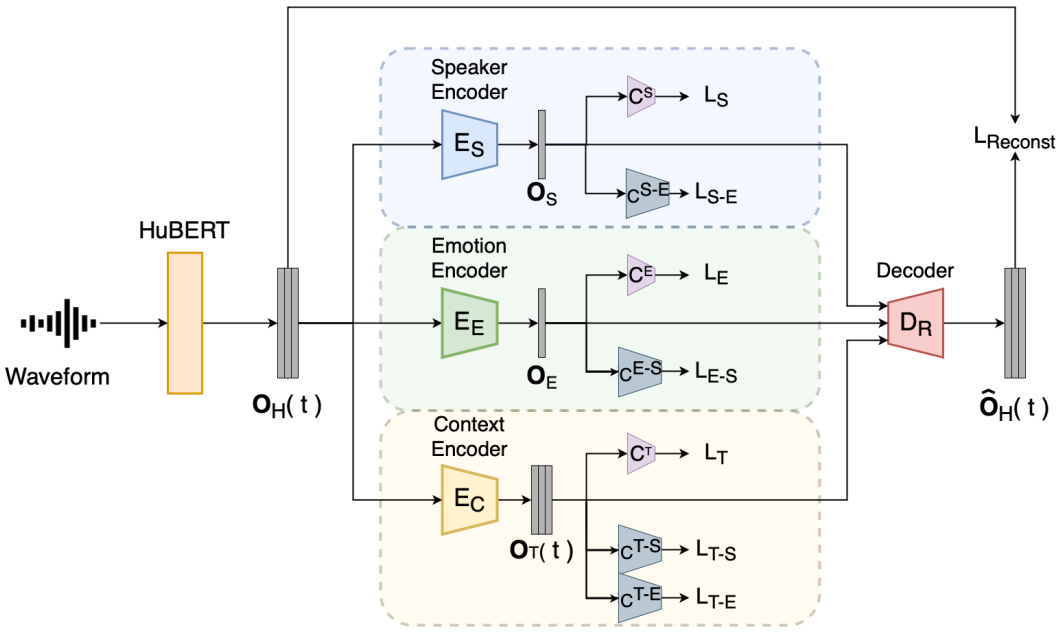


Fig. 1. Our model overview: The proposed method involves extracting embeddings using HuBERT, followed by three encoders: speaker encoder, emotion encoder, and context encoder. Each encoder is connected to classification and suppression models. These models help refine the embeddings by focusing only on the target task and suppressing irrelevant ones. The refined embeddings are then reconstructed by a decoder.

hidden units and sets up a task to predict these units, capturing the temporal features of speech data. Based on BERT, one of the transformer models, HuBERT utilizes bidirectional contextual information to deepen the understanding of audio data, resulting in high performance in many tasks e.g. speech recognition and generation.

B. Non-Parallel Sequence-to-Sequence Voice Conversion with Disentangled Linguistic and Speaker Representations [3]

Non-parallel sequence-to-sequence (seq2seq) voice conversion aims to transform the voice of one speaker into another one without requiring parallel training data. This approach involves disentangling linguistic content from speaker characteristics to achieve high-quality voice conversion. The proposed model consists of a text encoder, a recognition encoder, a speaker encoder, and a seq2seq decoder. The recognition encoder aligns acoustic features with phonetic sequences to produce linguistic representations of audio signals, while the speaker encoder generates embeddings that capture speaker-specific information. These components are integrated by the seq2seq decoder to reconstruct the acoustic features. Our method is inspired by this model structure.

III. Method

This section presents our proposed method: how to extract embeddings from raw speech waveforms leveraging the HuBERT model, how to obtain representations from three different encoders for speaker, emotion and context recognition tasks, and how to incorporate the vectors to reconstruct the original embeddings. Our proposed method, as illustrated in Fig.1, involves two main stages: feature-specific encoding using dedicated encoders, and embedding reconstruction using a decoder.

A. Feature Encoders

Initially, embeddings are extracted from the raw speech waveform using the HuBERT model. Let us denote the embeddings by $\mathbf{O}_H(t)$, where t is a time index. In this paper, the pre-trained HuBERT model is expected. The HuBERT embeddings are then given to three independent encoders to obtain speaker, emotion, and context embeddings, described below.

*1) **Speaker Encoder**:* The speaker encoder E_S is designed to extract a feature vector $\mathbf{O}_S$ related to speaker identity:

$$\mathbf{O}_S = E_S\big(\mathbf{O}_H(t)\big) \tag{1}$$

The encoder is composed as follows: a self-attention layer, two fully connected layers, a Global Average Pooling (GAP) layer, and finally another fully connected layer. (Fig.2) Note that the representation $\mathbf{O}_S$ is obtained for one speech sequence.

Training Objective: To allow the encoded representations to have speaker information exclusively, an SR (identification) classifier C_S is used for model training. In addition, because the encoder is expected to exclude the emotion information, an SER suppression model C_{S-E} is also utilized; the suppression model is trained so that predicted results should be based on uniform distributions of which number corresponds to the number of emotions. This makes the speaker embeddings useless for SER. Consequently, we design a loss function of the speaker encoder according to the above discussion, defined as:

$$L_{\text{Speaker}} = L_S + \lambda \cdot L_{S-E} \tag{2}$$

where both the classifier C_S and the suppressor C_{S-E} consist of a single Linear layer,L_S is a cross-entropy loss for speaker identification, L_{S-E} is a cross-entropy loss for emotion suppression, and λ is a weighting factor.

*2) **Emotion Encoder**:* The emotion encoder E_E captures the emotional state of the speech. Its structure is the same as the speaker encoder. Similar to the former encoder, the emotion encoder also outputs a vector $\mathbf{O}_E$ as follows:

$$\mathbf{O}_E = E_E\big(\mathbf{O}_H(t)\big) \tag{3}$$

Training Objective: The basic idea of training loss function is also the same as the speaker encoder as:

$$L_{\text{Emotion}} = L_E + \lambda \cdot L_{E-S} \tag{4}$$

where both the classifier C_E and the suppressor C_{E-S} consist of a single Linear layer, L_E is a cross-entropy loss for emotion classification, L_{E-S} is a cross-entropy loss for speaker suppression C_{E-S}, and λ is a weighting factor.

*3) **Context Encoder**:* The context encoder E_T extracts contextual information from the given speech. The encoder is composed of a self-attention layer, followed by two fully connected layers. (Fig.2) The representation $\mathbf{O}_E(t)$ is obtained as:

$$\mathbf{O}_T(t) = E_T\big(\mathbf{O}_H(t)\big) \tag{5}$$

In this case, the vectors are generated frame by frame.

Training Objective: An ASR model C_T is adopted to observe a loss function L_T. We also leverage a speaker suppression classifier C_{T-S} and an emotion suppression classifier C_{T-E} to minimize irrelevant information. The training objective for the context encoder is then defined as:

$$L_{\text{Text}} = L_T + \lambda\ (L_{T-S} + L_{T-E}) \tag{6}$$

where both the classifier C_T and the suppressor C_{T-S}, C_{T-E} consist of a single Linear layer, L_T is a CTC loss with text labels, L_{T-S} is a cross-entropy loss for speaker suppression, and L_{T-E} is a cross-entropy loss for emotion suppression, and λ is a scaling factor.

B. Reconstruction Decoder

All the extracted features, which are speaker, emotion, and context embeddings $\mathbf{O}_S$, $\mathbf{O}_E$ and context $\mathbf{O}_T(t)$ respectively, are combined and fed into a reconstruction decoder D_R. The decoder reconstructs the original HuBERT embeddings as:

$$\hat{\mathbf{O}}_H(t) = D_R\big(\mathbf{O}_S, \mathbf{O}_E, \mathbf{O}_T(t)\big) \tag{7}$$

The purpose of this process is to preserve the integrity of the original speech attributes, while enhancing task-specific features. The simple reconstruction loss based on the mean squared error L_{Reconst} is chosen for training.

C. Training Strategy

The training strategy consists of two phases in each iteration. These two phases are learning opposing each other to effectively extract only the desired features.

Phase 1: In the first phase, the encoders (E_S, E_E, and E_T), decoder (D_R) and feature-based classifiers (C_S, C_E and C_T) are trained together, while the suppression classifiers are frozen. The total loss used in this phase includes all the loss values that we have already explained as:

$$L = L_{\text{Speaker}} + L_{\text{Emotion}} + L_{\text{Text}} + L_{\text{Reconst}} \tag{8}$$

Phase 2: Phase 2: In the second phase, only the suppression classifiers (C_{S-E}, C_{E-S}, C_{T-S} and C_{T-E}) are trained to minimize the influence of non-target features, for example, speaker information in emotion features. Specifically, C_{S-E} and C_{E-S} are trained using cross-entropy loss with the correct labels, while C_{T-S}, C_{T-E} employs CTC loss with the correct labels. This means that these models play an adversarial role. This adversarial training helps the encoders to learn more task-specific features by forcing the suppression classifiers to identify unwanted modalities.

IV. Experiment

A. Experimental Conditions

The primary objective of this experiment is to evaluate the performance of the proposed TRIDENT model and disentangled speaker, emotion, and context embeddings from HuBERT representations. The goals include assessing the accuracy of the learned embeddings and verifying the effectiveness of the reconstructed embeddings. Our hypothesis is that the TRIDENT model can effectively separate these attributes, resulting in improved performance in Automatic Speaker Verification (ASV), SER, and ASR tasks.

The model training utilized two datasets. First, the *LibriSpeech train-clean-360* dataset [4], which was a subset of the LibriSpeech dataset, consisting of 360 hours of "clean" English speech, was used to provide a robust training base. Note that, since LibriSpeech is a dataset with no emotion labels assigned, we froze the emotion encoder when training models with LibriSpeech. Second, the *RAVDESS* (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset [5] was employed to enable the model to learn emotional expressions. To further enhance the training data, we randomly added noise to the speech data using the **MUSAN** dataset [6] for data augmentation, ensuring diverse and comprehensive training inputs.

The model in the upstream interface of S3PRL [7], [8] was used as the pre-trained HuBERT model. The model was trained using *LibriSpeech-960* and the model structure follows the HuBERT base model. The output from the final output layer is used as the HuBERT embedding. Regarding the hyperparameter, λ was set to 0.5.

B. Automatic Speaker Verification

ASV was carried out to evaluate the performance of speaker embeddings. ASV tries to verify the identity according to their voices. In this experiment, the *VoxCeleb1* [9] dataset was used for testing, which contains over 100,000 utterances from 1,251 celebrities, extracted from videos uploaded to YouTube. The

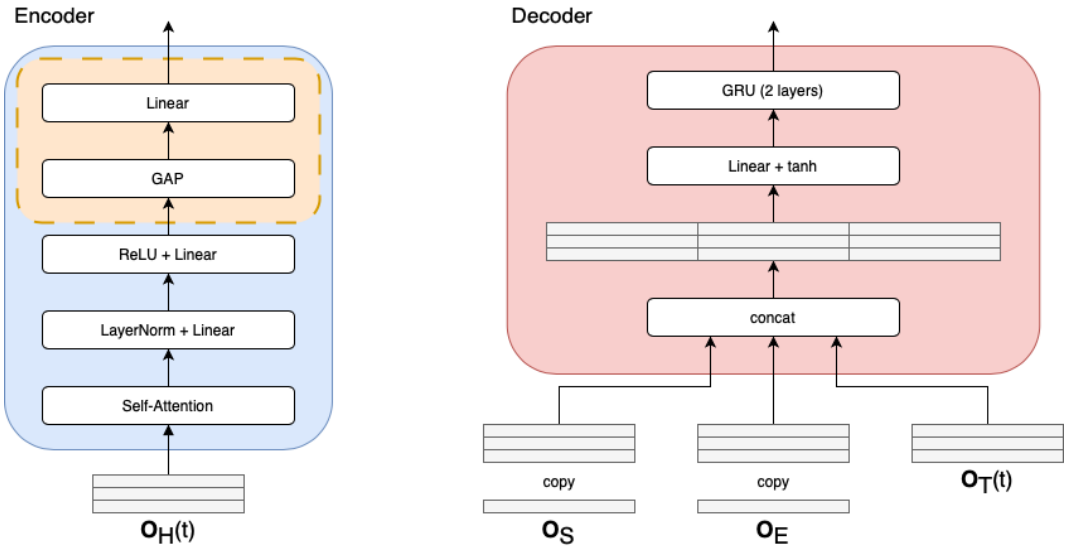


Fig. 2. The architecture of the encoder and decoder used in the proposed model. The encoders extract features from the HuBERT embeddings $\mathbf{O}_T(t)$. For the Speaker Encoder and Emotion Encoder, the components within the dotted region are utilized, including a GAP (Global Average Pooling) layer followed by a Linear layer. These components are omitted when using the Context Encoder. The extracted embeddings $\mathbf{O}_S$, $\mathbf{O}_E$, and $\mathbf{O}_T(t)$ are then passed to the decoder, where they are concatenated and processed through a Linear layer with a tanh activation, followed by a two-layer GRU (Gated Recurrent Unit), to reconstruct the original features

dataset is widely used for speaker verification tasks because of its diversity and real-world conditions.

In our evaluation, speaker embeddings were obtained and compared using the cosine similarity of speaker embeddings from pairs of audio samples. The Equal Error Rate (EER) was calculated to measure the accuracy of these comparisons. A HuBERT model from S3PRL, used as the base model, was compared by learning speaker classification and extracting embeddings with the same data as the proposed model. As a comparison of the accuracy of the reconstructed structure, the accuracy of the model excluding the reconstruction loss of the proposed method was also verified.

The results are summarized in Table I. The EER of our proposed TRIDENT model was 19.7%, slightly higher than the EER of 20.5% for the S3PRL HuBERT used as the base model and 20.1% for the model excluding reconstruction errors.

TABLE I
ASV PERFORMANCE COMPARISON

Model	EER(%)
HuBERT in S3PRL (Baseline)	20.5
Ours (Proposed)	19.7
Ours (w/o reconstruction)	20.1

C. Speech Emotion Recognition

SER was performed to evaluate the performance of emotion embeddings. In this experiment, the *CREMA-D* (Crowd-sourced Emotional Multimodal Actors Dataset) [10] was used, which contains 7,442 original clips from 91 actors. In these clips, actors between the ages of 20 and 74, representing a variety of races and ethnicities, express six different emotions: anger, disgust, fear, happy, neutral, and sad. Each sentence is presented with four different emotion levels: low, medium, high, and unspecified.

In our evaluation, emotion embeddings were obtained and used to train the emotion classifier. Weighted Average Recall (WAR) was a metric commonly used in SER tasks to evaluate the performance. Unlike the accuracy, which can be biased towards the majority class in imbalanced datasets, WAR calculates the average recall (sensitivity) across all classes, weighted by the class distribution. It is defined as the sum of the recall values for each class, weighted by the number of instances in each class, providing an evaluation metric that accounts for class imbalance. It takes values between 0 and 1, with higher values indicating more accurate performance.

The results are summarized in Table II. The WAR of our proposed TRIDENT model was 57.0%, which was lower than the WAR of 69.6% for the base model S3PRL HuBERT. The WAR of 59.7% was obtained in the model excluding reconstruction errors.

TABLE II
SER PERFORMANCE COMPARISON

Model	WAR(%)
HuBERT in S3PRL (Baseline)	69.6
Ours (Proposed)	57.0
Ours (w/o reconstruction)	59.7

D. Automatic Speech Recognition

Automatic Speech Recognition (ASR) was used to evaluate the performance of context representations in the TRIDENT

model. In this experiment, the LibriSpeech datasets were chosen: the *LibriSpeech-train-clean-100* subset for training, and the *LibriSpeech-test-clean* subset for testing. The LibriSpeech corpus is a collection of approximately 1,000 hours of 16kHz read English speech derived from audiobooks, prepared by Vassil Panayotov with the assistance of Daniel Povey.

In the evaluation, the Word Error Rate (WER) was calculated to measure the accuracy of speech recognition. The proposed TRIDENT model was compared to the baseline S3PRL HuBERT model minus reconstruction errors.

The results are summarized in Table III. Our proposed TRIDENT model achieved a WER of 8.61%, exceeding the WER of 11.47% for HuBERT in the baseline S3PRL. Regarding the reconstruction errors. the proposed model with the error slightly outperformed the WER of the model excluding reconstruction errors.

TABLE III
ASR PERFORMANCE COMPARISON

Model	WER (%)
HuBERT in S3PRL (Baseline)	11.47
Ours (Proposed)	8.61
Ours (w/o reconstruction)	8.93

E. Discussion

It is found that significant improvements were observed in ASV and ASR tasks, indicating that the TRIDENT model is particularly adept at separating and utilizing speaker-specific and contextual information. In the SER task, our proposed scheme achieved roughly 60% WAR; since the SER task is still a challenging task, the performance seems to be acceptable. In addition, the experimental results also suggest that the reconstructed structure helps to improve the consistency and accuracy of each embedding.

On the other hand, our method could not achieve the performance of the baseline model. It might be due to distortions in the RAVDESS dataset for model training and the CREMA-D dataset. We believe such the performance degradation sometimes occurs when using several emotion datasets for training, due to inconsistency of emotion labels or unbalanced data as a whole. To confirm the above discussion, visualization of the data which are not used for training in the RAVDESS dataset is conducted. Fig 3 shows the UMAP-based visualization results. The figure shows that the emotions in the RAVDESS dataset are appropriately separated in general.

According to the results, we re-trained the whole model using the CREMA-D dataset when performing SER. The results tell us that the WAR drastically increased to 73.1%. This significant improvement indicates the mismatch between two datasets.

To verify whether each representation was appropriately separated, we conducted supplemental experiments; we evaluated the accuracy of speaker and emotion embedding on another task, respectively. As a result, the WAR using speaker embedding was 48.8%. Regarding ASV, in this case we used

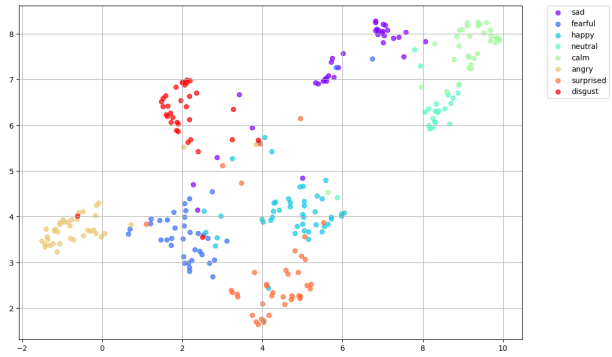


Fig. 3. Visualization of emotional embeddings in the RAVDESS dataset.

classification accuracy: 77.9% for speaker embeddings, and 59.4% for emtion embeddings. It turned out that there was a significant decrease in accuracy if task and embedding were different. This indicates that our TRIDENT model can compute embeddings exclusively for each task, even if not perfectly.

On the other hand, the simplicity of the neural network layers used in the three different encoders raises an important consideration. Although the training process utilized 360 hours of audio data, it remains uncertain whether such a compact model architecture can fully capture the complexity of the data distribution. While the current results are promising, it is essential to explore whether more sophisticated model architectures could better learn the intricate patterns present in the data. Addressing this limitation is a key direction for future work, where we plan to investigate the potential benefits of using more advanced models. This would ensure that the model's complexity is better aligned with the scale and diversity of the training data, potentially leading to further improvements in performance.

Finally, we would like to address the reconstruction. The reconstruction accuracy during testing achieved the MSE of 0.0068. This indicates that speaker-, emotion-, and contextual embeddings can still have significant information to regenerate corresponding HuBERT representations. We then carried out additional experiment. Assuming that a speaker vector is replaced into another one's, we checked the difference between the reconstructed vector and the true vector in the same emotion and context; The results were not good unfortunately, suggesting that our embeddings still need to be improved.

V. CONCLUSION

This paper presents our new model TRIDENT, which is designed to disentangle and re-synthesize speaker, emotion, and context information obtained from HuBERT embeddings. Experimental results demonstrated that TRIDENT improved performance in automatic speaker verification and automatic speech recognition tasks, and achieved acceptable results in emotion recognition. The reconstruction accuracy and classification performance in the tasks without the reconstruction function indicate the effectiveness and enhancement.

Our future works will focus on optimizing the model architecture and training procedures, as well as exploring different speech representations. The ability of TRIDENT to refine and separate speech attributes will offer promising directions for advancements in speech processing applications.

REFERENCES

[1] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, *Superb: Speech processing universal performance benchmark*, 2021. arXiv: 2105.01051 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2105.01051.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[3] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2019.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[5] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, e0196391, 2018.

[6] D. Snyder, G. Chen, and D. Povey, *MUSAN: A Music, Speech, and Noise Corpus*, arXiv:1510.08484v1, 2015. eprint: 1510.08484.

[7] A. T. Liu, S.-W. Li, and H.-y. Lee, *Tera: Self-supervised learning of transformer encoder representation for speech*, 2020. arXiv: 2007.06028 [eess.AS].

[8] S.-w. Yang, H.-J. Chang, Z. Huang, *et al.*, "A large-scale evaluation of speech foundation models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[9] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[10] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

High and Low Frequency Region Separation Method for Adaptive Image Expansion

Shao-Yun Luo [1], Kuei-Chen Chen [2], Jian-Jiun Ding [3*], Cheng-Che Lee [4], and Hsin-Jung Lee [5]
[1, 4, 5] Graduate Institute of Electronics Engineering, National Taiwan University, Taiwan
[2] Graduate School of Advanced Technology, National Taiwan University, Taiwan
[3*] Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
E-mail: r11943162@ntu.edu.tw [1], r10k43002@ntu.edu.tw [2], jjding@ntu.edu.tw [3*], d06943010@ntu.edu.tw [4], d04943010@ntu.edu.tw [5], Tel: +886-233669652

***Abstract*— Signal expansion has played an important role in multimedia signal compression and analysis. There are a variety of signal expansion methods, including DCT, KLT, Legendre polynomial expansion, wavelet transforms, etc. This study proposes a novel approach based on adaptive region segmentation. The main architecture of the proposed method is to divide images into high-frequency and low-frequency components. Then, these components are further divided according to the connectivity and the region size. Then, the adaptive DCT is applied to expand each part. With the proposed method, the truncation of the high-frequency component in the edge part will not affect the expand result in the low-frequency region. It can effectively reduce the ringing effect and decrease the error of image expansion. Experiments demonstrate that, with the proposed frequency-adaptive segmentation method, the accuracy of image expansion can be much improved. The proposed expansion approach is expected to be applied to other two-dimensional signals.**

I. Introduction

Signal analysis has become increasingly important since the last century. Researchers can now tackle much more complex problems thanks to the advent of computers. One of the most classical problems is the four-color theorem, which is the first theorem proved by a computer [1]. Computers also provide faster computing times for previous problems or theories, such as trigonometric functions, logarithms, calculus, and so on. Many methods are used to analyze or compress various one-dimensional signals. For instance, Fourier series [2], Legendre polynomial expansion [3], the wavelet transform [4], the discrete cosine transform (DCT) [5], and the discrete sine transform (DST) [6] are methods regularly applied to all kinds of one-dimensional (1D) signals. Some of these methods can also be extended to two-dimensional (2D) forms, such as 2D-DCT and 2D-DST. In the case of 2D signal analysis, image processing plays an important role. As one of the most popular coding algorithms, the two-dimensional discrete cosine transform (2D-DCT) is a powerful method in the field [7], as its performance approaches that of the Karhunen-Loéve transform (KLT), an optimal method [8]. In recent years, JPEG and JPEG2000 [9] have shown great application in image compression.

However, many former methods have some problems. For instance, the finite form of the Fourier series may exhibit the Gibbs phenomenon, and Legendre polynomial expansion also lacks accuracy in some cases. In this study, we propose a new method based on 1D-DCT for 2D image cases, referred to as adapted-DCT. Through the proposed method, we initially divide the given image into high-frequency and low-frequency parts using edge detection and morphology. Subsequently, we apply adapted-DCT to both the high-frequency and low-frequency parts. In the adapted-DCT stage, we first process the columns of the image, then the rows. Additionally, we observe the relationship between the compression ratio and the cumulative sum ratio and compare the weighted rank simultaneously. The compressed data is obtained through these steps. With the proposed frequency-adaptive segmentation and expansion method, the image approximation error is much less than that of other expansion methods and one can use the same number of basis functions to achieve a much better image expansion result.

II. Previous Work

There are a variety of signal expansion methods. For example, one can use the *Fourier series* to expand a signal using a set of basis functions with different frequencies. Moreover, one can also use *Legendre polynomials* [3], which have the polynomial form and are the solutions of the Legendre differential equation [10], to expand a signal:

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}[(x^2-1)^n]. \quad (1)$$

The *discrete cosine transform (DCT)* expansion method [5] has played an important role in image processing and signal analysis. The DCT of the type 2 (DCT-II) is defined as [11]:

$$F(k) = \sqrt{\frac{2}{N}}\sum_{n=0}^{N-1} \mathrm{b(k)f(n)}cos\left(\frac{\pi(n+0.5)k}{N}\right), \quad (2)$$

$k = 0, 1, 2, \dots, N-1.$

The *discrete wavelet transform (DWT)* is also an important method in signal analysis. The simplest form of the wavelet is the Haar wavelet [12]. It is composed only of -1, 0, and 1. Another well-known wavelet is the Daubechies wavelet transform [13]. This transform is used in various types of signal analysis, signal compression, and denoising. The wavelet transform allows researchers to adjust the vanishing moments, which are half the number of points.

Sinc interpolation is to convolve the samples of signal with a sinc function [14]. Generally speaking, the sinc function is usually defined by the following equation:

$$sinc(x) = \frac{\sin(\pi x)}{\pi x}, x \neq 0 \quad (3)$$

For a signal $y(t)$ with length N, we can decompose the signal in the following form:

$$y(t) = \sum_{n=1}^{N} x(n)\, sinc\left(\frac{(t-n)\cdot T}{T}\right) \quad (4)$$

where T stands for the sparse sampling time. When using sinc interpolation, if the highest frequency of the signal is less than $1/2T$, the signal can be perfectly reconstructed.

In addition to sinc interpolation, B-splines are commonly used in interpolation since they can preserve the trend of the signal. The definition of B-splines is shown as follows:

$$B_{n,0}(t) = 1 \; for \; t_n < t < t_{n+1}, otherwise \; B_{n,0}(t) = 0 \quad (5)$$

$$B_{n,m} = \frac{t - t_n}{t_{n+m} - t_n} B_{n,m-1}(t) + \frac{t_{n+m+1} - t}{t_{n+m+1} - t_{n-1}} B_{n+1,m-1}(t), \; for \; t_n < t < t_{n+m+1} \quad (6)$$

$$x(t) = \sum_{n=1}^{N} x(t_n)\, B_{n,m}(t) \quad (7)$$

In (6) and (7), the number m stands for the type of B-spline. If m = 1, it is the linear B-spline, if m = 2, it is the quadratic B-spline, and m = 3 corresponds to the cubic B-spline [15].

III. Proposed Algorithm

The overall flowchart of the proposed image expansion algorithm is plotted in Fig. 1.

A. Edge Detection

To classify the images into low and high frequency parts, we first perform edge detection. In the step of edge detecting, we use the following edge detector as the separating tool:

$$h[n] = -C * \mathrm{sgn}[n] * \exp(-\sigma|n|) \text{ for } |n| \le L, \quad (8)$$

$$C = \frac{1}{\sum_{n=1}^{L} \exp(-\sigma|n|)}. \quad (9)$$

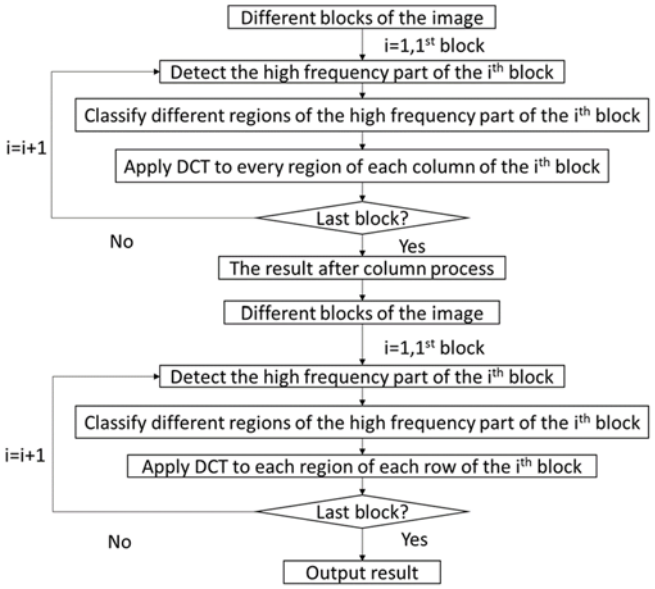


Fig. 1 Flowchart of the proposed adaptive image expansion algorithm.

Note that $h[n]$ is an odd function, and L, σ are adjustable. If we want to detect smaller noise (extracting tiny edges), larger value of σ is essential, and vice versa. Throughout our experiment, we will set L = 10, σ = 1.

After defining $h[n]$, it is time for us to carry out convolution. Convolution will be applied on both the horizontal direction and the vertical direction of the image. For the set of image signal y, we can obtain the horizontal convolution result y_h as well as the vertical convolution result y_v. Afterwards, we calculate $y_g = \sqrt{{y_h}^2 + {y_v}^2}$ as the material for thresholding. Every element's value in y_g larger than the given threshold value will be preserved, otherwise, it will be discarded. The new y_g after thresholding will be sent to the next step of the proposed method.

B. Morphology

After edge detection, we perform morphology to clarify the edges more clearly. Therefore, both erosion and dilation are necessary steps:

$$Erosion: A[m,n] = \min\{A[m,n], A[m-1,n], A[m+1,n], A[m,n-1], A[m,n+1]\},$$

$$Dilation: A[m,n] = \max\{A[m,n], A[m-1,n], A[m+1,n], A[m,n-1], A[m,n+1]\}. \quad (10)$$

If the pixel $[m, n]$ is near the margin or the corner of the matrix, we directly ignore the exceeded parts. To fill holes and eliminate excessive noise, we combine erosion and dilation as closing and opening steps. For the closing operation performed k_1 times, we execute dilation k_1 times, then execute erosion k_1 times. In the case of the opening operation performed k_2 times, erosion is executed k_2 times first, then dilation k_2 times. Now

that morphology is completed, the high-frequency parts and the low-frequency parts of the image are finally separated, and the image is ready for the next step. Some examples of high-frequency part extraction are shown in Fig. 2(b).

C. Adapted-DCT

With the results of morphology, we have already obtained the two parts of the image. Before applying DCT, we will first classify the image into different blocks. After that, we will divide the high-frequency parts into different regions. This is necessary because, after thresholding, some of the high-frequency parts may be divided by low-frequency parts. These steps will also be applied to the low-frequency parts. Next, it is time to apply adapted-DCT to both parts separately. Explaining the algorithm in words is quite complex, so we will illustrate it with a flow chart, as shown in Fig 1.

The steps will also be applied to the low-frequency parts of the image. After employing adapted-DCT on both high-frequency and low-frequency parts, we will combine the results and calculate the cumulative sum ratio. To compress the data, we will set a compression ratio, conserve the data whose cumulative sum ratio is within the given ratio, and set the excluded parts to zero. The compressed data is thus obtained through adapted-DCT.

D. Evaluation

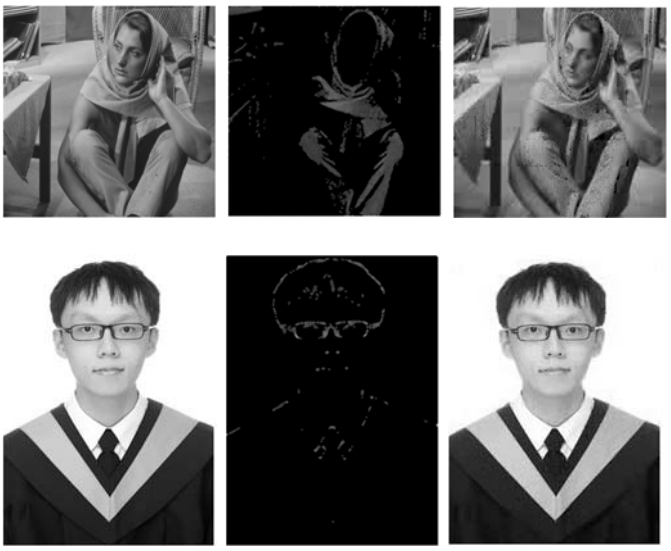
After calculating the coefficients of the proposed methods, we will use the cumulative sum as the metric [16]. To observe the results more clearly, we will calculate the cumulative sum ratio of the given signal. For a signal with length N, the definition of the sth cumulative sum ratio is given as follows:

$$Cumulative\ sum\ ratio = \sum_{n=1}^{s} |C_n|^2 \,/\, \sum_{n=1}^{N} |C_n|^2 \qquad (11)$$

In the above equation, C_n stands for the nth coefficient after calculation. Because the cumulative sum focuses on the preservation of energy, it has a concept similar to Parseval's theorem. In our experiment, we will first sort the coefficients Cn in descending order, then calculate the cumulative sum ratio. Through this sorting method, we can concentrate the energy of the signal at the beginning of the coefficient series. We are also able to compare the efficiency between different methods.

The recovery of a compressed image is quite similar to the inverse of compression. We use the inverse DCT as the main restoration method. It should be noted that because we process the columns first during compression, we have to carry out the row process first during recovery. After applying the recovery process to both high-frequency and low-frequency parts of the compressed data, we can finally recover the original image.

The structural similarity (SSIM) is able to evaluate the perceptual differences between two images [17]. For given images x and y, the equation of the SSIM is shown as follows:



(a) Original images (b) High-frequency parts (c) Reconstructed image

Fig. 2 The high-frequency regions and the reconstructed image (using 5% of the expansion basis functions) of Barbara image (cumulative sum ratio = 0.9833 and SSIM = 0.9472) and Graduation1 image (cumulative sum ratio = 0.9984 and SSIM = 0.9891).

Table II. Summary of the Compared Expansion Methods

Number	Method Name	Number	Method Name
1	Time-domain Expansion	10	8-point Daubechies Wavelet
2	Fourier Series Expansion	11	Sinc Interpolation
3	DCT-II	12	Blockwise Fourier (*Ns* = 32)
4	DCT-I	13	Blockwise Fourier (*Ns* = 64)
5	DCT-III	14	Blockwise DCT-II (*Ns* = 32)
6	DCT-IV	15	Blockwise DCT-II (*Ns* = 64)
7	DST	16	Blockwise Legendre (*Ns*= 32)
8	Haar Wavelet	17	Blockwise Legendre (*Ns*= 64)
9	Multilayer Haar Wavelet	18	Frequency Adaptive Segmentation (Proposed)

$$SSIM(x,y) = \frac{[2\mu_x\mu_y + (c_1L)^2][2\sigma_{xy}(c_2L)^2]}{[\mu_x{}^2 + \mu_y{}^2 + (c_1L)^2][\sigma_x{}^2 + \sigma_y{}^2 + (c_2L)^2]} \qquad (12)$$

In (13), μ_x and μ_y stand for the means of x and y, while σ_x and σ_y mean the variances of x and y. Meanwhile, σ_{xy} is the covariance of x and y. We set L as 255 and set c_1 and c_2 as $\sqrt{255}$ in our further experiments.

IV. Experiments

In this section, we demonstrate the experiments to illustrate the efficiency of the proposed method. We selected seven images, including Baboon, Barbara, Cameraman, Goldhill, Peppers, and two graduation pictures (Graduation1 and Graduation2) of the author.

TABLE III COMPARING THE IMAGE EXPANSION PERFORMANCE IN TERMS OF THE CUMULATIVE SUM RATIO WHEN PRESERVING 5% OF EXPANSION BASIS FUNCTIONS.

Methods	Baboon	Barbara	Cameraman	Goldhill
2	0.9720	0.9269	0.9226	0.9645
3	0.9695	0.9612	0.9217	0.9556
4	0.9702	0.9557	0.9243	0.9646
5	0.9622	0.9544	0.9218	0.9584
6	0.9584	0.9070	0.9070	0.9601
7	0.9628	0.9500	0.9306	0.9474
8	0.2257	0.2837	0.3227	0.2095
9	0.9715	0.9634	0.9259	0.9558
10	0.9814	0.9711	0.9589	0.9637
11	0.0508	0.0508	0.0508	0.0508
12	0.9659	0.9361	0.9101	0.9468
13	0.9627	0.9202	0.9290	0.9595
14	0.9703	0.9558	0.9282	0.9444
15	0.9744	0.9611	0.9471	0.9594
16	0.9710	0.9581	0.9190	0.9432
17	0.9632	0.9439	0.8787	0.9529
Proposed	**0.9480**	**0.9833**	**0.9840**	**0.9958**

Methods	Peppers	Graduation1	Graduation2	**Average**
2	0.9394	0.9714	0.9558	**0.9504**
3	0.9346	0.9510	0.9378	**0.9473**
4	0.9364	0.9590	0.9420	**0.9503**
5	0.9361	0.9555	0.9419	**0.9472**
6	0.9530	0.9485	0.9413	**0.9393**
7	0.9273	0.9576	0.9460	**0.9460**
8	0.3092	0.2978	0.3129	**0.2802**
9	0.9496	0.9654	0.9466	**0.9540**
10	0.9620	0.9799	0.9735	**0.9701**
11	0.0508	0.0522	0.0522	**0.0512**
12	0.9295	0.9479	0.9458	**0.9403**
13	0.9258	0.9498	0.9381	**0.9407**
14	0.9259	0.9604	0.9601	**0.9493**
15	0.9346	0.9541	0.9403	**0.9530**
16	0.9385	0.9664	0.9643	**0.9515**
17	0.9338	0.9558	0.9559	**0.9406**
Proposed	**0.9948**	**0.9984**	**0.9983**	**0.9861**

We compare the proposed methods with 17 other expansion methods, which are summarized in Table II where *Ns* is the size of each block (e.g., *Ns* = 32 means that we partition the images into 32x32 blocks before expansion).

While applying morphology, the condition is set with a threshold of 50, one closing process, and two opening processes. The results after morphology will be illustrated. After compressing, we will first compare the relation between the compression ratio and the cumulative sum ratio with other methods. Second, we will also compare the relation between the compression ratio and the rank. Next, we recover the images with a specified compression ratio.

In Table III, we compare the image expansion performance in terms of the cumulative sum ratio defined in (11) of the proposed method and the methods listed in Table II when using only 5% of the basis functions for expansion. The results of Method 1 are not shown in Table III because its cumulative sum ratios are no more than 0.15, which are much less than those of other methods. The results in Table III show that the proposed algorithm achieves the highest average cumulative sum ratio of 0.9861. That is, with the proposed algorithm, a much more accurate expansion can be achieved.

Moreover, in Figs. 3-6, we show the curves of the cumulative sum ratio versus the compression ratio (i.e., the percentage of the expansion basis functions that are preserved) of Barbara, Goldhill, Peppers, and Graduation2 Images of the proposed algorithm and other expansion methods. Moreover, we also show the rank of the cumulative sum ratio when the compression ratio varies from 0% to 5%. The results show that, with the proposed algorithm, the accuracy of image expansion is much better than that of other methods.

V. CONCLUSION

The adapted-DCT-based method proposed for image compression and recovery demonstrates excellent performance, especially in images where high-frequency and low-frequency components are clearly separated, such as Goldhill, Peppers, and Graduation images. These images achieve not only top efficiency compared to previous methods but also nearly identical recovered images to the originals. For images like Baboon and Barbara, where high-frequency components are heavily mixed with low-frequency components, the recovery results are less satisfactory due to this intermingling, leading to lower performance rankings. Overall, the adapted-DCT-based method shows significant potential in image compression and recovery. Confidence remains high that this method and its underlying principles can be effectively applied to other areas of signal analysis.

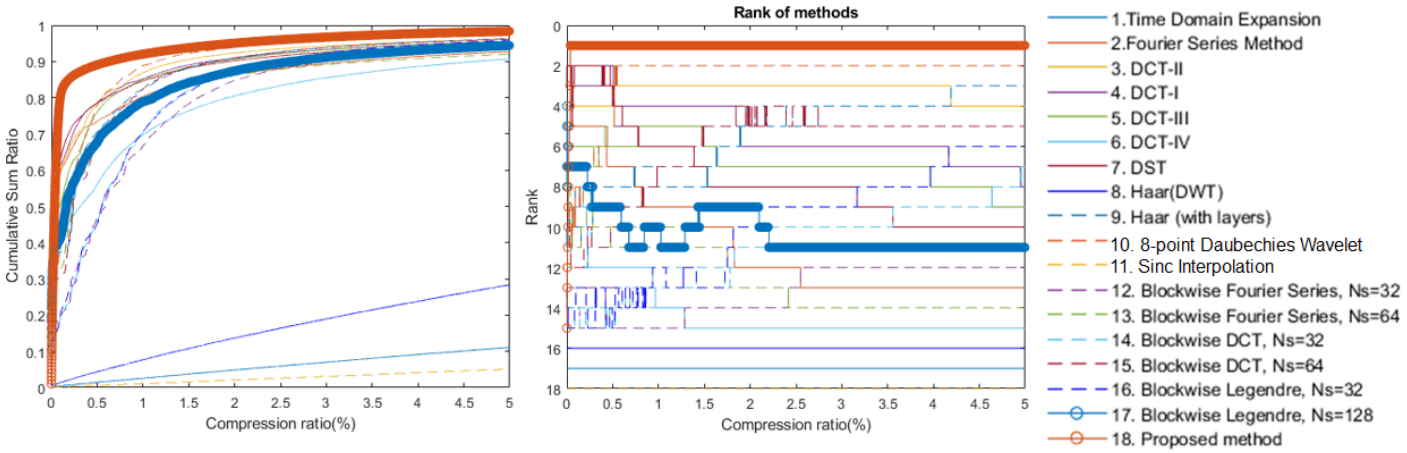


Fig. 3 (Left): The curves of the cumulative sum ratios versus the compression ratios of the proposed algorithm and other expansion methods for Barbara image; (Right): the rank of the cumulative sum ratios for different compression ratios. The bold orange curves are the results of the proposed algorithm.

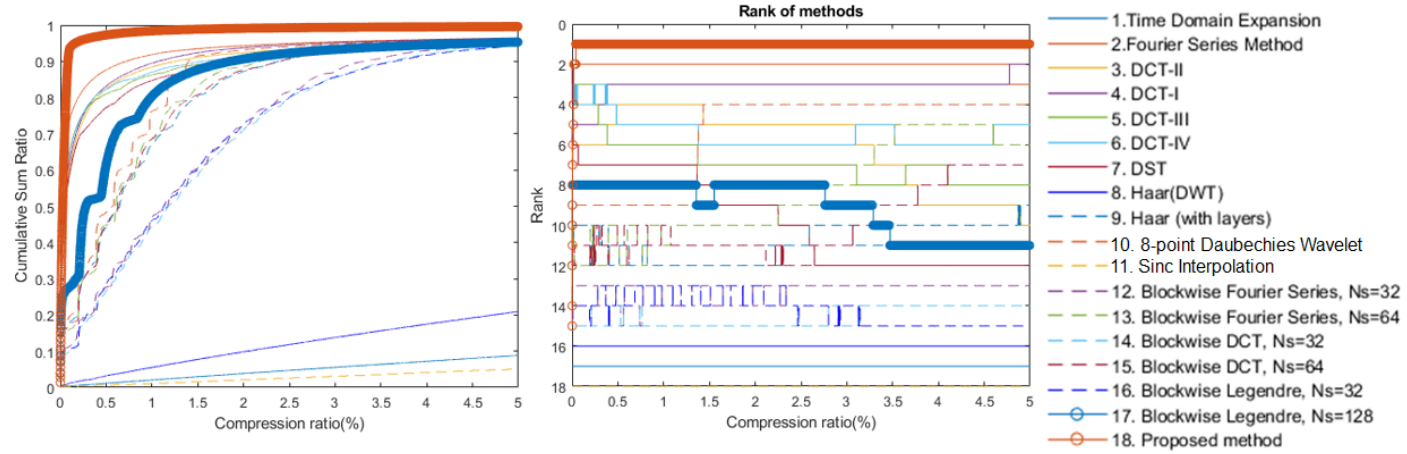


Fig. 4 (Left): The curves of the cumulative sum ratios versus the compression ratios of the proposed algorithm and other expansion methods for Goldhill image; (Right): the rank of the cumulative sum ratios for different compression ratios. The bold orange curves are the results of the proposed algorithm.

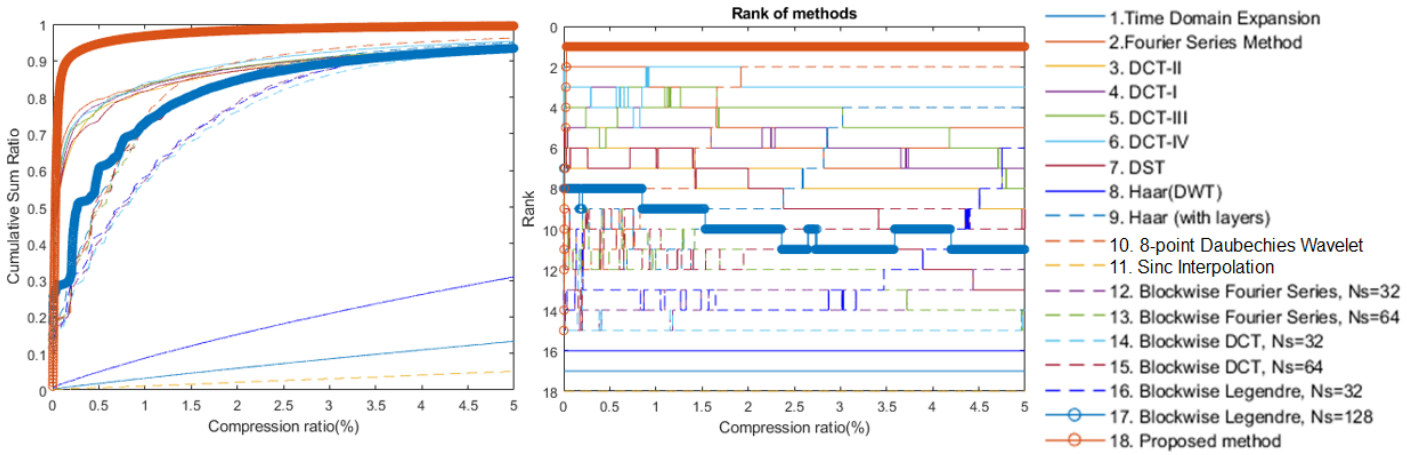


Fig. 5 (Left): The curves of the cumulative sum ratios versus the compression ratios of the proposed algorithm and other expansion methods for Peppers image; (Right): the rank of the cumulative sum ratios for different compression ratios. The bold orange curves are the results of the proposed algorithm.

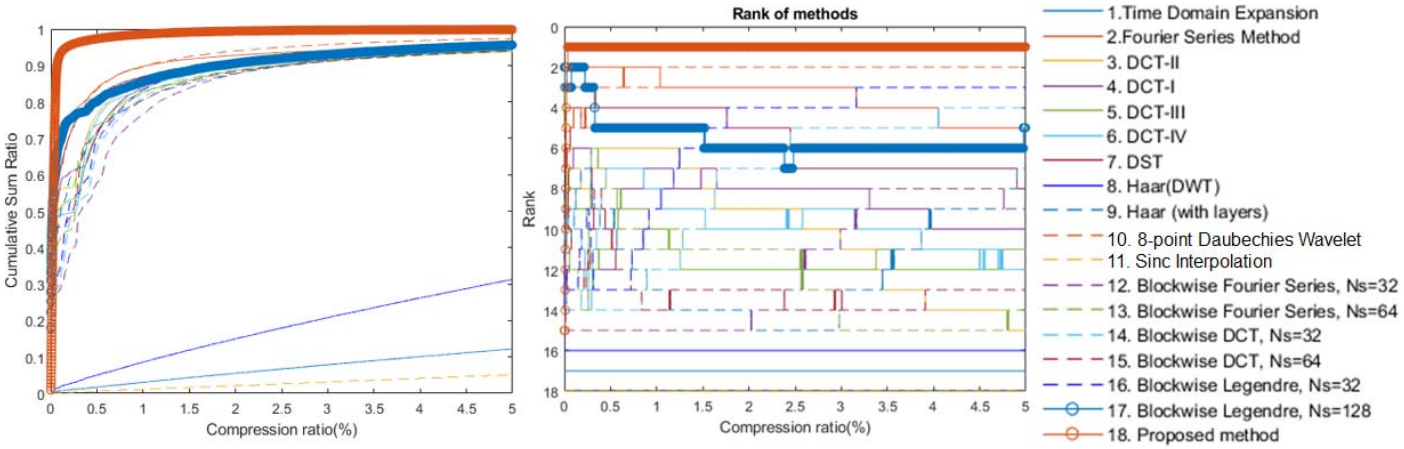


Fig. 6 (Left): The curves of the cumulative sum ratios versus the compression ratios of the proposed algorithm and other expansion methods for Graduation2 image; (Right): the rank of the cumulative sum ratios for different compression ratios. The bold orange curves are the results of the proposed algorithm.

ACKNOWLEDGMENT

We are grateful for the support provided by Dr. Chieh-Hsiung Kuan, who left this world in 2023. His introduction of Precision Calculation has been invaluable to this work. We sincerely thanks for his 30 years of dedicated teaching at National Taiwan University.

REFERENCES

[1] Y. Brun, "The four-color theorem," *Undergraduate J. Mathematics*, pp. 21-28, 2002.

[2] T. Hsu, *Fourier Series, Fourier Transforms, and Function Spaces: A Second Course in Analysis*, vol. 59, American Mathematical Society, 2023.

[3] H. Wang and S. Xiang, "On the convergence rates of Legendre approximation," *Mathematics of Computation*, vol. 81, no. 278, pp. 861-877, 2012.

[4] D. Zhang and D. Zhang. "Wavelet transform," *Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval*, pp. 35-44, 2019.

[5] H. Ochoa-Dominguez and K. R. Rao, *Discrete Cosine Transform*, CRC Press, 2019.

[6] G. Steidl and M. Tasche, "A polynomial approach to fast algorithms for discrete Fourier-cosine and Fourier-sine transforms," *Mathematics of Computation*, vol. 56, no. 193, pp. 281-296, 1991.

[7] Z. Yuan, D. Liu, X. Zhang, and Q. Su, "New image blind watermarking method based on two-dimensional discrete cosine transform," *Optik*, vol. 204, article 164152, 2020.

[8] J. W. Soh, H. S. Lee, and N. I. Cho, "An image compression algorithm based on the Karhunen Loève transform," in *APSIPA ASC*, pp. 1436-1439, 2017.

[9] R. Singh, and V. K. Srivastava, "JPEG2000: A review and its performance comparison with JPEG," in *IEEE Int. Conf. Power, Control and Embedded Systems*, pp. 1-7, 2012.

[10] S. M. Jung, "Legendre's differential equation and its hyers-Ulam stability," *Abstract and Applied Analysis*, vol. 2007, no. 1, article 056419, 2007.

[11] X. Shao and S. G. Johnson, "Type-II/III DCT/DST algorithms with reduced number of arithmetic operations," *Signal Processing*, vol. 88, issue 6, pp. 1553-1564, 2008.

[12] R. S. Stanković and B. J. Falkowski. "The Haar wavelet transform: its status and achievements," *Computers & Electrical Engineering*, vol. 29, issue 1, pp. 25-44, 2003.

[13] D. Popov, A. Gapochkin, and A. Nekrasov, "An algorithm of Daubechies wavelet transform in the final field when processing speech signals," *Electronics*, vol. 7, issue 7, article 120, 2018.

[14] L. Yaroslavsky, "Fast discrete sinc-interpolation: a gold standard for image resampling," *Advances in Signal Transforms: Theory and Applications*, vol. 7, pp. 337-405, 2007.

[15] T. Briand and P. Monasse. "Theory and practice of image B-spline interpolation," *Image Processing on Line*, vol. 8, pp. 99-141, 2018.

[16] E. W. Weisstein, "Cumulative Sum," *MathWorld--A Wolfram Web Resource*, [Online], Available: https://mathworld.wolfram.com/ CumulativeSum.html

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, issue 4, pp. 600−612, 2004.

Unleashing Attributes-content Adaptation with Multi-color Spaces for Food Photo Aesthetic Assessment

Shintami Chusnul Hidayati, Muhammad Valda Rizky Nur Firdaus, Riki Wahyu Nur Dianto, and Sarwosri
Department of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia
E-mail: shintami@its.ac.id

***Abstract*—The rise of social media and digital food culture has elevated the importance of food photography, making aesthetic assessment a key area of interest. This study introduces a novel framework for evaluating the aesthetic quality of food photos by integrating attribute-content adaptation with multi-color space feature representations. Addressing the need for more accurate and nuanced aesthetic assessment models, our method combines high-level attributes with detailed content features. We leverage convolutional neural networks to extract these features, which are then fused with color features and enhanced by a proposed classifier network. Our experimental design includes comprehensive comparisons against baseline models that only use attribute-content adaptations and various classifiers, such as the support vector machine and random forest. Results on public datasets demonstrate that our proposed framework consistently outperforms baseline models, achieving the highest accuracy and F1-score metrics. An ablation study further reveals that the fusion of multiple color spaces significantly enhances performance compared with individual color spaces or no color space integration, underscoring the critical role of color diversity in aesthetic evaluation.**

I. Introduction

The aesthetic assessment of food photos is a crucial aspect of various applications, including social media, online marketing, and culinary arts. Accurate evaluation of food photo aesthetics can enhance user engagement, improve content quality, and support automated systems in curating visually appealing images. Traditional methods for assessing photo aesthetics have relied heavily on handcrafted features and basic image processing techniques, which often fail to capture the nuanced visual details and context-dependent attributes critical for high-quality aesthetic evaluations [1].

Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly enhanced image aesthetics assessment. CNNs excel in computer vision tasks like image classification, object detection, and aesthetic quality evaluation [2], [3]. Studies that leverage CNNs have achieved superior performance by automatically extracting relevant image features. A notable approach is the photo aesthetics ranking network with attributes and content adaptation [4], which dynamically adjusts the importance of aesthetic attributes based on a photo's content. It employs a multi-branch structure, where each branch focuses on a specific attribute, and a content adaptation module fine-tunes the importance of each attribute. This approach has proven effective in capturing context-aware aesthetic characteristics of photos [5]. Hosu *et al.* [6] enhanced model capability by employing multiple CNNs trained on different aesthetic criteria.

Building on these advancements, we propose a novel approach that integrates attribute-content adaptation with multi-color space feature representations to capture a comprehensive feature set that contributes to assessing the visual appeal of food photos. Following [4], we utilize eleven attributes closely related to image aesthetic judgments, spanning traditional photographic principles of color, lighting, focus, and composition. These include *balancing element*, *color harmony*, *interesting content*, *lighting*, *motion blur*, *object emphasis*, *repetition*, *rule of thirds*, *shallow depth of field*, *symmetry*, and *vivid color*. The proposed method leverages color moments from multiple color spaces, thereby allowing richer and more detailed representations of an image's aesthetic quality. By fine-tuning CNNs to extract attribute and content features, we weight these features dynamically based on their relevance to the image content. This adaptive weighting mechanism ensures that the most pertinent features are highlighted in the final aesthetic evaluation. In addition, we use the Extreme Learning Machine (ELM) [7] as the classifier due to its rapid training speed and generalizability.

The performance of the proposed model is evaluated against several baseline models, including those that only used attribute-content adaptations and different well-known classifiers. The results demonstrate that integrating attribute-content adaptation with multi-color space feature representations significantly enhances model performance across all classifiers. In particular, the proposed framework outperformed baseline models.

In an earlier study, Datta *et al.* [8] explored computational approaches to studying aesthetics in photographic images. They developed a framework using both low-level features, such as color and texture, and high-level features, such as the presence of certain objects and compositional rules. Their work laid the groundwork for subsequent research by highlighting the importance of combining various features to achieve a comprehensive aesthetic evaluation. Our approach builds upon this foundation by integrating modern deep learning techniques and multi-color space representations to capture a broader

range of visual and contextual information.

Recent studies further support the importance of incorporating diverse color spaces in aesthetic evaluations. For example, [9]–[11] have shown that multi-color space features can capture nuanced visual details, thereby leading to more accurate assessments of image aesthetics. Our ablation study also highlights that the fusion of multiple color spaces yields superior performance compared to the use of individual color spaces or no color space integration.

The major contributions of this work are as follows:

1) We propose a novel framework that integrates attribute-content adaptation with multi-color space feature representations for food photo aesthetic assessment.
2) We demonstrate the effectiveness of this integrated approach through extensive experiments, demonstrating improvements in performance metrics.
3) We conduct an ablation study to highlight the importance of incorporating multiple color spaces, which further enhances the model's ability to capture nuanced visual details critical for aesthetic evaluation.
4) We provide a comprehensive analysis of the impact of different color spaces and their fusion on model performance, providing insights into the best practices for aesthetic assessment.

The remainder of this paper is organized as follows. Section II details the proposed methodology. Section III describes the experimental setup and results. Finally, Section IV concludes the paper and suggests directions for future research.

II. Methodology

This section details the proposed methodology for classifying a food image into its aesthetic label: positive or negative. Given an input image, we first extract its visual features using the attribute and content representation and color representation modules in parallel. The extracted features are then concatenated. During the training phase, the concatenated features are used to train the classifier. In the recognition phase, they are used for aesthetic evaluation using the trained model. Fig. 1 illustrates these steps.

A. Attribute and Content Representation Module

This step captures the visual characteristics of images by leveraging meaningful photographic attributes and content cues to ensure context-aware aesthetic evaluation. Inspired by the work of Kong *et al.* [4], we dynamically adjust the importance of aesthetic attributes based on the photo content. The network employs a multi-branch structure in which each branch focuses on a specific attribute, and a content adaptation module adjusts their importance. The estimated aesthetic scores from each attribute are concatenated to form the attribute-content adaptation feature representation, denoted by $\mathbf{x}_{AttCont}$. As mentioned earlier, eleven attributes are identified, making the dimentionality of $\mathbf{x}_{AttCont}$ eleven.

Initially, we fine-tuned a CNN backbone using images annotated with aesthetic ratings rescaled to [0, 1] (described later in Section III-A1) with the following Euclidean loss:

$$loss_{reg} = \frac{1}{2N}\sum_{i=1}^{N}\|\hat{y}_i - y_i\|_2^2, \tag{1}$$

where y_i is the average ground-truth rating for image-*i*, $\hat{y}_i$ is the CNN-predicted score, and N is the number of training samples. To further refine the model, we applied a pairwise ranking loss based on the Siamese network to exploit the relative rankings of image pairs:

$$loss_{reg+rank} = loss_{reg} + \omega_r loss_{rank}, \tag{2}$$

where

$$loss_{rank} = \frac{1}{2N}\sum_{i,j}\max(0, \alpha - \delta(y_i \geq y_j)(\hat{y}_i - \hat{y}_j)) \tag{3}$$

and

$$\delta(y_i \geq y_j) = \begin{cases} 1, & \text{if } y_i \geq y_j \\ -1, & \text{if } y_i < y_j \end{cases}. \tag{4}$$

Here, ω_r controls the relative importance of the ranking loss and α is a margin parameter.

The attribute-adaptive process begins by identifying and extracting aesthetic attributes from the input photo. Instead of independently training attribute classifiers, an attribute prediction branch is incorporated on top of the second fully-connected layer in the network. Each attribute is processed through dedicated branches, capturing unique aesthetic information specific to each attribute, which operate in parallel to output individual attribute feature vectors.

To ensure appropriate attribute activations, we employed Euclidean loss ($loss_{att}$) and combined it with rating and ranking losses. The overall loss function is defined as follows:

$$loss = loss_{reg} + \omega_r loss_{rank} + \omega_a loss_{att}, \tag{5}$$

where ω_a controls the importance of attribute fine-tuning. To enhance adaptability, a relevance estimation network processes attribute feature vectors to generate relevance scores, weighting the attribute vectors to ensure that the most significant attributes influence the final aesthetic evaluation.

Simultaneously, the content-adaptive process analyzes the photo to understand its context. This is achieved using the extracted content feature vector, which represents the scene of the captured photo. The content feature vector is then fed into a relevance estimation network similar to that used in the attribute adaptation process. This network outputs a set of relevance scores for the attributes $\mathbf{x}_{\text{AttCont}}$, dynamically adjusting these scores based on the identified content. These scores indicate the relevance of each attribute in the context of the specific content of the photo.

B. Color Representation Module

We employed color moments to extract color features from each color channels. Specifically, color moments capture the color distribution within an image. The first three moments (mean, standard deviation, and skewness) are computed for each channel of these color spaces.

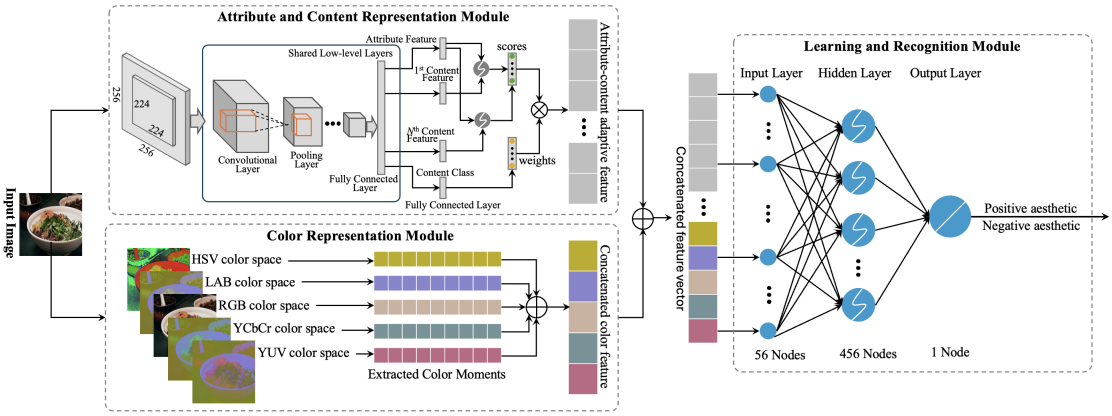


Fig. 1: Illustration of the proposed framework.

The first-order color moment, or mean, captures the average color value in an image and is defined as follows:

$$M_1 = \frac{1}{P}\sum_{i=1}^{P} p_i, \tag{6}$$

where p_i represents the pixel value of the i-th pixel in the color channel, and P is the total number of pixels in the image. The second-order color moment, or variance, measures the spread of color values around the mean:

$$M_2 = \sqrt{\frac{1}{P}\sum_{i=1}^{P}(p_i - M_1)^2}. \tag{7}$$

This equation calculates the standard deviation of the color values, indicating how much the pixel values deviate from the mean. The third-order color moment, or skewness, describes the asymmetry of the color distribution. A positive skew indicates a distribution with a longer tail on the right side, whereas a negative skew indicates a longer tail on the left side. The skewness is given by:

$$M_3 = \sqrt[3]{\frac{1}{P}\sum_{i=1}^{P}(p_i - M_1)^3}. \tag{8}$$

We computed these moments for each channel in the LAB, RGB, HSV, YUV, and YCbCr color spaces. For instance, the RGB color space consists of three channels: red, green, and blue. Thus, for each channel, the mean, standard deviation, and skewness are computed, resulting in nine features for the RGB color space alone. This process is similarly applied to the other color spaces, generating a set of color features. The extracted color moments from all color channels across the different color spaces are concatenated to form a single, unified feature vector, $\mathbf{x}_{\text{MulColor}}$. The feature vector is then used in subsequent stages of the aesthetic assessment process, enabling a detailed analysis of the food photo's color attributes.

C. Learning and Recognition Module

This step uses the Extreme Learning Machine (ELM) [7], a feedforward neural network with a single hidden layer, to classify food photos as "positive" or "negative". ELM is known for its rapid training speed and generalizability. Its simplicity and efficiency provide a robust solution for our aesthetic assessment needs.

The feature representation, $\mathbf{x} = [\mathbf{x}_{\text{AttCont}}, \mathbf{x}_{\text{MulColor}}]$, is formed by concatenating features from the attribute-content and color representation modules. ELM randomly initializes weights between the input and hidden layers and assigns biases to hidden neurons. These weights and biases remain fixed during training to reduce computational complexity.

In the hidden layer, which comprises 456 units, the ReLU activation function is applied to the weighted sum of the input features. Mathematically, the hidden layer output matrix $\mathbf{H}$ is:

$$\mathbf{H} = \text{ReLU}(\mathbf{XW} + \mathbf{B}), \tag{9}$$

where $\mathbf{X}$ is the input feature matrix, $\mathbf{W}$ is the randomly initialized weight matrix, $\mathbf{B}$ is the bias matrix, and ReLU is defined as $\text{ReLU}(\mathbf{x}) = \max(0, \mathbf{x})$.

The output weights $\boldsymbol{\beta}$, which connect the hidden layer to the output layer, are computed using the Moore-Penrose pseudoinverse $\mathbf{H}^{\dagger}$ of the hidden layer output matrix $\mathbf{H}$. This process involves solving a linear system of equations that minimizes the error between predicted and actual outputs:

$$\boldsymbol{\beta} = \mathbf{H}^{\dagger}\mathbf{T}, \tag{10}$$

where $\mathbf{T}$ is the target output matrix.

For classification, a query feature vector undergoes the same hidden layer transformation using the pre-initialized weights and biases. The transformed hidden layer output vector $\mathbf{h}_{\text{query}}$ is then multiplied by the output weights $\boldsymbol{\beta}$ to produce the final output (i.e., the predicted aesthetic class), written as:

$$\mathbf{y}_{\text{query}} = \mathbf{h}_{\text{query}}\boldsymbol{\beta}. \tag{11}$$

III. Experiments

A. Datasets

This study employs two datasets for distinct purposes: the first for learning the attribute-content adaptations and the second for the rest of the models.

TABLE I: Performance comparison of different models. The highest and second-to-best performance is highlighted in boldface and underlined, respectively.

Models	Negative label				Positive label				Overall labels			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
$AttCont_{18}$ + SVM	0.556	0.569	0.170	0.262	0.556	0.564	**0.892**	0.691	0.556	0.567	0.531	0.477
$AttCont_{18}$ + RF	0.558	0.552	0.462	0.503	0.558	0.606	0.689	0.645	0.558	0.579	0.576	0.574
$AttCont_{18}$ + ELM	0.558	0.546	0.434	0.484	0.558	0.598	0.700	0.645	0.558	0.572	0.567	0.565
$AttCont_{50}$ + SVM	0.564	0.569	0.170	0.262	0.564	0.564	**0.892**	0.691	0.564	0.567	0.531	0.477
$AttCont_{50}$ + RF	0.586	0.543	0.441	0.487	0.586	0.598	0.692	0.642	0.586	0.5705	0.567	0.565
$AttCont_{50}$ + ELM	0.584	0.552	0.444	0.492	0.584	0.602	0.700	0.648	0.584	0.577	0.572	0.570
$AttCont_{18}$ + MultiCol + SVM	0.667	0.663	0.587	0.623	0.667	0.687	0.752	0.718	0.667	0.675	0.670	0.671
$AttCont_{18}$ + MultiCol + RF	0.696	0.702	0.653	0.677	0.696	0.656	0.759	0.704	0.696	0.679	0.706	0.691
$AttCont_{18}$ + MultiCol + ELM (ours)	0.717	0.702	0.653	0.677	0.717	0.727	0.770	0.748	0.717	0.715	0.712	0.713
$AttCon_{50}$ + MultiCol + SVM	0.678	0.663	0.588	0.623	0.678	0.687	0.752	0.718	0.678	0.675	0.670	0.671
$AttCon_{50}$ + MultiCol + RF	0.699	0.680	0.635	0.657	0.699	0.712	0.752	0.732	0.699	0.696	0.694	0.695
$AttCon_{50}$ + MultiCol + ELM (ours)	**0.723**	**0.705**	**0.660**	**0.682**	**0.723**	**0.731**	0.770	**0.750**	**0.723**	**0.718**	**0.715**	**0.716**

1) Aesthetics and Attributes Database (AADB): The AADB dataset [4] comprises 10,000 photographic images sourced from Flickr, each licensed under Creative Commons. It deliberately excludes non-photographic materials such as cartoons and advertisements. Five independent evaluators from Amazon's Mechanical Turk (AMT) assessed each image to provide ground-truth aesthetic scores and annotations for 11 crucial attributes, including content appeal, lighting quality, color harmony, and symmetry. The subset containing 8,500 images was used to fine-tune the aesthetic-rating network described in Section II-A.

2) Gourmet Photography Dataset (GPD): The Gourmet Photography Dataset (GPD) [12] includes 24,000 food images sourced from social media, featuring diverse food types, angles, lighting conditions, and layouts. Annotation was conducted using AMT, where online workers evaluated the aesthetic appeal of each image. Images that were skipped three times or had their validity questioned were removed from the dataset. Eight expert photographers reviewed and validated the aesthetic labels to further ensure data quality, retaining only those with consensus among the majority.

B. Experimental Setup

The experiments were conducted using Python with PyTorch as the deep learning framework. Network training was performed on a personal computer with an Intel® Core™ i7-9750G CPU @ 2.60 GHz (12 CPUs) ˜3.5 GHz, 8GB DDR4 RAM, and an Nvidia GEFORCE GTX 1650 GPU. Key hyperparameters, including learning rate, batch size, and the number of epochs, were tuned based on the validation performance for optimal model training.

We selected ResNet-18 and ResNet-50 as backbone networks because of their robust feature extraction capabilities and ability to handle vanishing gradients. ResNet-18 offers a balance of depth and efficiency, whereas ResNet-50 provides enhanced feature extraction for capturing intricate details. The use of both models allowed us to evaluate the performance across different network depths and ensure comprehensive feature representation for attribute-content adaptation.

C. Baselines

To evaluate the performance of our proposed network, we compared it with several baseline models to highlight the added value of integrating attribute-content adaptations (*AttCont*) [4] with multi-color space feature representations (*MulColor*) using ELM. Specifically, we used two types of AttCont features extracted using different backbone networks: $AttCont_{18}$ obtained using ResNet-18 and $AttCont_{50}$ obtained using ResNet-50.

1) Attribute-Content Adaptations Alone: We first assessed a model that used only attribute-content adaptation features, excluding color features, to isolate their impact on overall performance. These features were extracted using the same methods as those used in our proposed network.

2) Comparison with Different Classifiers: To further validate our approach, we explored two different classifiers: Support Vector Machine (SVM) [13] and Random Forest (RF) [14]. Each classifier was applied to classify the different extracted feature representations.

D. Results and Discussion

The performance comparison in Table I demonstrates the effectiveness of integrating attribute-content adaptations [4] with color feature representations and using various classifiers. The models were evaluated using accuracy (Acc.), precision (Prec.), recall (Rec.), and F1-score (F1) for both negative and positive labels, as well as overall performance.

The models that used attribute content adaptations alone ($AttCont_{18}$ and $AttCont_{50}$) generally demonstrated modest performance. The SVM classifier for $AttCont_{18}$ achieved an overall accuracy of 0.556, with a notable precision of 0.569 for the negative label but a very low recall of 0.170, indicating a high rate of false negatives. The RF and ELM classifiers showed slight improvements in recall but still underperformed overall, with both achieving an accuracy of 0.558, although ELM had a slight edge in precision (0.572 vs. 0.579 for RF). For $AttCont_{50}$, similar patterns were observed, with the SVM classifier achieving a modest overall accuracy of 0.564 and precision but poor recall for the negative label. The RF and ELM classifiers performed slightly better than the SVM classifier

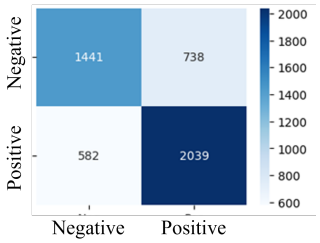


Fig. 2: Confusion matrix illustrating the performance of the proposed network.

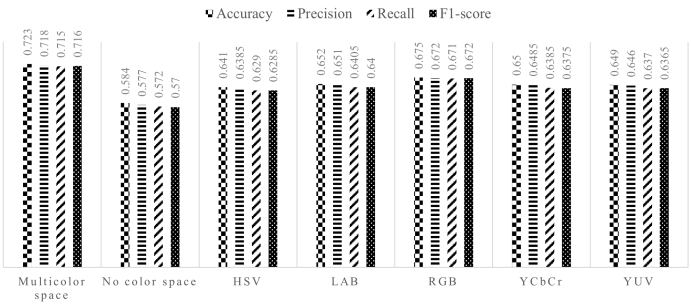


Fig. 3: Results of the ablation study.

with overall accuracy values of 0.586 and 0.584, respectively; however, the improvements remained incremental.

The integration of multi-color space features (MultiCol) with attribute-content adaptations significantly enhanced performance across all classifiers. For the AttCont$_{18}$+MultiCol models, the SVM classifier saw a substantial boost, achieving an overall accuracy of 0.667 with notable improvements in recall (0.670) and F1-score (0.671). RF and ELM further improved performance, with ELM achieving an overall accuracy of 0.717, precision of 0.715, recall of 0.712, and an F1-score of 0.713, underscoring the benefit of combining multiple color spaces with attribute-content adaptations. Similarly, for AttCont$_{50}$+MultiCol models, the SVM classifier continued to improve, with an overall accuracy of 0.678, precision of 0.675, recall of 0.670, and an F1-score of 0.671. RF and ELM performed even better, with RF achieving an overall accuracy of 0.699, and ELM exhibiting the best performance across most metrics, with an overall accuracy of 0.723, precision of 0.718, recall of 0.715, and an F1-score of 0.716.

Key observations from the results include the superior performance of multi-color space integration, which consistently improved the performance of all classifiers, highlighting the value of incorporating color diversity in aesthetic evaluation. The ELM classifier, particularly when combined with AttCont and MultiCol features, outperformed SVM and RF across all metrics, with the highest scores observed in the AttCont$_{50}$+MultiCol+ELM configuration. Additionally, all models demonstrated high recall for positive labels, especially with the ELM and SVM classifiers, indicating a strong ability to correctly identify aesthetically positive images. The combination of attribute-content adaptations with multi-color spaces not only enhanced accuracy but also maintained a balance across precision, recall, and F1-score, thereby contributing to robust and reliable performance in aesthetic assessments.

Fig. 2 illustrates the confusion matrix, where the vertical axis represents the ground truth labels, and the horizontal axis represents the predicted labels. The results demonstrate that integrating attribute-content adaptation and color feature representations significantly enhances the performance of the model. Among the evaluated classifiers, ELM demonstrated superior performance, particularly when combined with both attribute-content and multi-color space features.

E. Ablation Study

To understand the contribution of each color space to model performance, we conducted an ablation study using attribute-content adaptations with the ResNet-50 backbone (AttCont$_{50}$) and ELM as classifier, which provided the highest performance. This study evaluated models using AttCont alone (no color space), AttCont with multi-color space, and AttCont with individual color spaces (HSV, LAB, RGB, YCbCr, and YUV).

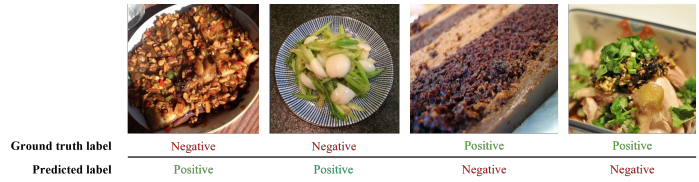
The results, shown in Fig. 3, reveal that models without color space integration show moderate performance, with an accuracy of 0.584, precision of 0.577, recall of 0.572, and F1-score of 0.570. Integrating individual color spaces significantly improves these metrics. For example, using the RGB color space, the model achieves an accuracy of 0.675, precision of 0.672, recall of 0.671, and an F1-score of 0.672. The fusion of multiple color spaces leads to the best performance, with an accuracy of 0.723, precision of 0.718, recall of 0.715, and F1-score of 0.716.

Overall, this study demonstrates that integrating color features significantly enhances the aesthetic assessment network's performance. Among individual color spaces, RGB provides the most substantial improvement, while the fusion of multiple color spaces delivers the highest overall performance, effectively capturing the nuanced aesthetics of food photos.

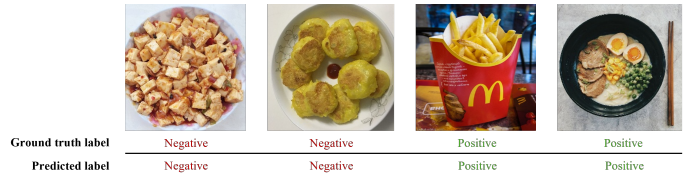
F. Exemplar Results

In Fig. 4, we present examples of the recognition results provided by the proposed method. Images with "negative" actual values tend to have underexposed areas, paler colors, and parts of the food that appear cut off. In contrast, images with "positive" actual values typically exhibit bright colors, proper object placement, adequate lighting, and all objects are fully visible within the camera frame.

Fig. 4a shows examples where the predicted results differ from the actual values. The features extracted using color feature representation, particularly those assessing lighting, heavily influence these predictions. Thus, images with good lighting quality, as shown in the first two figures, are more likely to be rated "positive" by the prediction model. Conversely, incorrect predictions of "positive" examples into the "negative" class, as shown in the last two images, can be attributed to several factors. First, the color quality is relatively dull, with limited color diversity. Second, the images suffer from inappropriate



(a) Incorrect predictions.



(b) Correct predictions.

Fig. 4: Exemplar results demonstrating the proposed network's prediction performance.

photo angles, resulting in significant portions of the food being out of frame. Third, the lighting needs to be adequate; both images cropped and focused on specific areas, compromising overall lighting quality. Fig. 4b, on the other hand, depicts examples of correct predictions. These images typically have vibrant colors, well-composed framing, and sufficient lighting, all of which contribute to their positive aesthetic assessment.

Overall, the analysis indicates that color quality, proper object placement, and adequate lighting are critical factors influencing the aesthetic assessment of food photos. The proposed model effectively captures these nuances, although occasional discrepancies arise due to suboptimal image characteristics.

IV. Conclusions and Future Work

This study presents a novel approach to food photo aesthetic assessment by integrating attribute-content adaptation with multi-color space feature representations. Our experiments demonstrated that this combination significantly enhances the performance of aesthetic assessment models across various classifiers. Specifically, when combined with both attribute-content adaptations and multi-color space features, the proposed classifier network consistently outperformed baseline models that used only one feature type. The results from the ablation study further underscored the importance of incorporating multiple color spaces.

Future work could explore more sophisticated feature extraction techniques, such as transformer-based models or attention mechanisms, to better capture intricate details and contextual information in images. Additionally, expanding the set of aesthetic attributes to include more diverse and comprehensive features could further enhance the model's ability to assess various aesthetic dimensions.

Acknowledgment

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2024.

References

[1] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 80–106, 2017.

[2] S. C. Hidayati, J. Tasyanita, C. Fatichah, and Y. Anistyasari, "Understanding human face shape via inception-resnet neural network architecture," in *Proc. Int. Conf. Advanced Mechatronics, Intell. Manufacture and Industrial Automation*, 2023, pp. 631–636.

[3] D. L. Yonia and S. C. Hidayati, "Assessing food photo aesthetics through dominant color and color harmony with adaptive smoothing regularization," in *Proc. FORTEI-International Conf. Electrical Engrg.*, 2024.

[4] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 662–679.

[5] Y. Wang, Y. Huang, X. Chen, L. Li, and G. Shi, "Modeling content-attribute preference for personalized image esthetics assessment," *Image Vis. Comput.*, vol. 124, p. 104 505, 2022.

[6] V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, 2019, pp. 9375–9383.

[7] S. Ding, H. Zhao, Y. Zhang, X. Xu, and R. Nie, "Extreme learning machine: Algorithm, theory and applications," *Artif. Intell. Rev.*, vol. 44, no. 1, pp. 103–115, 2015, ISSN: 0269-2821.

[8] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. Eur. Conf. Comp. Vis.*, 2006, pp. 288–301.

[9] Q. Dai, L. Tang, and G. Hu, "Neural image aesthetic assessment using adversarial nets," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 2334–2338.

[10] F. Zhang, Z. Lin, X. Shen, R. Mech, and Y. Liu, "Image enhancement and aesthetic assessment with multi-level spatially pooled features," *IEEE Trans. Image Process.*, vol. 29, pp. 1454–1466, 2020.

[11] X. Lu, P. Xu, and W. Zhou, "Deep learning for image aesthetics: A survey," *IEEE Trans. Image Process.*, vol. 29, pp. 9408–9424, 2020.

[12] K. Sheng, W. Dong, H. Huang, *et al.*, "Learning to assess visual aesthetics of food images," *Comput. Vis. Media.*, vol. 7, pp. 139–152, 2021.

[13] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011.

[14] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 3, pp. 5–32, 2001.

PG-MDD: Prompt-Guided Mispronunciation Detection and Diagnosis Leveraging Articulatory Features

Meng-Shin Lin [1], Bi-Cheng Yan[1], Tien-Hong Lo[1], Hsin-Wei Wang[1], Yue-Yang He[1], Wei-Cheng Chao[2], Berlin Chen[1]

[1] National Taiwan Normal University, Taipei, Taiwan
[2] Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan
E-mail: {61147077s, bicheng, teinhonglo, hsinweiwang, yueyanghe, berlin}@ntnu.edu.tw; weicheng@cht.com.tw

***Abstract*—Mispronunciation detection and diagnosis (MDD) manages to pinpoint phonetic errors of L2 (second-language) learners and then provides timely and informative diagnosis on erroneous pronunciation segments. Recently, dictation-based neural methods have emerged as an appealing modeling paradigm for MDD, which simultaneously identifies pronunciation errors and provides diagnostic feedback by aligning the recognized phone sequence to the corresponding canonical phone sequence of a given text prompt. Despite their decent performance in terms of F1-score, dictation-based models still struggle to accurately detect pronunciation errors with balanced precision and recall evaluations, resulting in inferior learning efficiency for L2 learners. In view of this, we propose a novel prompt-guided dictation-based MDD model, dubbed PG-MDD, that can efficiently strike a balance between the precision and recall rates while maintaining a high-performing F1-score. PG-MDD first jointly optimizes the mispronunciation detection and diagnosis processes during the training phase, while aptly guiding the diagnosis process with phone-dependent thresholds in the inference phase. In addition, a novel multi-view audio encoder is introduced to render the fine-grained articulatory cues within learners' speech. A comprehensive set of empirical experiments conducted on the L2-ARCTIC benchmark dataset suggests the practical feasibility of our method in relation to several competitive baselines.**

I. Introduction

Globalization has increased the demand for learning languages like English, Spanish, and Mandarin, driving the development of computer-assisted pronunciation training (CAPT) systems. These systems allow L2 learners to practice pronunciation independently and stress-free [1][2], supplement teachers' instruction, address the shortage of qualified teachers [3], and serve as references for professionals in assessment tasks [4][5].

Mispronunciation detection and diagnosis (MDD) is a crucial part of CAPT systems. MDD aims to identify erroneous pronunciation segments and provide diagnostic feedback to L2 learners. It can assess pronunciation proficiency at both the supra-segmental (e.g., prosody and intonation) [6][7] and segmental levels (e.g., phone and word) [8][9]. Due to the difficulty of achieving high inter-rater agreement for supra-segmental assessment, research has mainly focused on segmental assessment. Efforts in phone-level MDD typically fall into three categories: pronunciation scoring-based, dictation-based, and prompt-based methods. Pronunciation scoring-based approaches relies on a well-trained speech recognition model to derive various types of confidence measurements as indicators of mispronunciation. Commonly used indicators include, but are not limited to, phone durations [10], likelihood ratios [11], and phone posterior probabilities [12]. To better obtain informative diagnostic feedback, dictation-based methods alternatively frame MDD as a phone recognition task by employing a free-phone recognition process to dictate the most likely phone sequence uttered by an L2 learner. The erroneous pronunciation portions are then easily identified by comparing the dictation result with the corresponding canonical phone sequence [13][14]. Among others, prompt-based methods extend the aforementioned methods by simplifying the forced-alignment process, which employs an attention mechanism to calculate the soft alignment between the canonical phone sequence and the L2 learner's input speech while simultaneously generating the diagnostic feedback [15][16]. Comparing these three mainstream paradigms, dictation-based

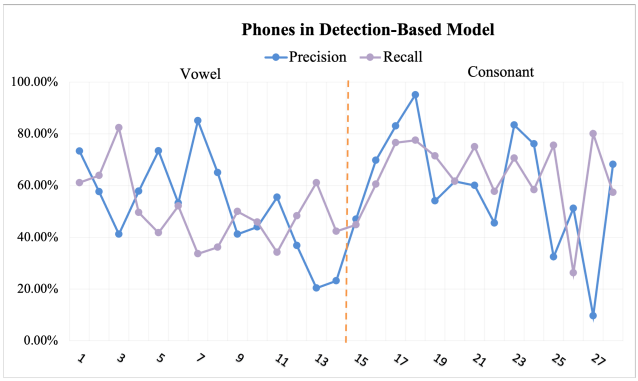
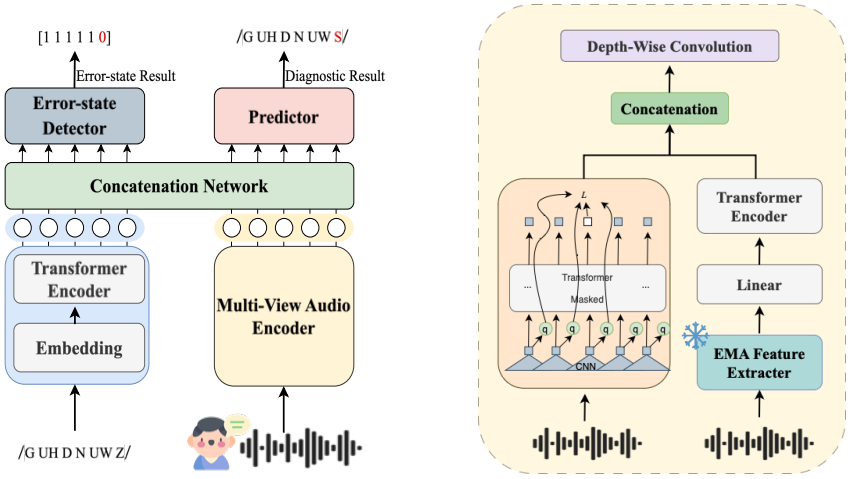


Fig. 1. Precision and recall distributions on different phones for a detection-based model. The orange dashed line indicates the division between vowels (on the left) and consonants (on the right)



(a) Overall Model Architecture

(b) Multi-View Audio Encoder

Fig. 2. A schematic depiction of the model structure of our MDD method.

methods are considered to be the dominant ones, showing considerable promise [17][18].

Despite the effectiveness of existing dictation-based methods in terms of the F1-score metric, they are often faced with challenges in accurately detecting pronunciation errors with balanced precision and recall rates. As an illustration, we demonstrate the performance of mispronunciation detection for disparate phones with a dictation-based MDD model in Fig. 1 where most phones achieve only a precision of 60%, with the corresponding recall rates varying between 40% and 80% for disparate vowels and consonants. This result underscores the adverse impact that correctly pronounced phone segments are frequently misidentified as pronunciation errors by the dictation-based model, which might demotivate students, leading to lower confidence in the MDD system. What is more, failing to detect pronunciation errors potentially slows down the learning process. Building on these observations, this paper introduces a novel prompt-guided dictation-based MDD model, termed PG-MDD, with the goal of achieving a better balance between the precision and recall rates while maintaining a high-performing F1-score for the mispronunciation detection task. To aptly guide the diagnosis process of the proposed model, we first jointly optimize the mispronunciation detection and the diagnosis processes during the training phase. During the inference phase, PG-MDD determines the pronunciation errors in response to the canonical phone sequence of a text prompt using adjustable phone-dependent thresholds. Furthermore, a novel multi-view audio encoder is put forward to capture the fine-grained articulatory cues inherent in learners' speech. Comprehensive experiments conducted on the L2-ARCTIC benchmark dataset reveal that our proposed method achieves significant and consistent improvements over several strong baselines.

In summary, our contributions in this paper are at least three-fold: (1) a novel prompt-guided MDD model is proposed, dubbed PG-MDD, which can flexibly balance performance trade-offs between precision and recall rates using adjustable phone-dependent thresholds; (2) PG-MDD is able to authentically dictate the input speech of an L2 learners, benefiting from cross-modality attention between the input speech and the text prompt, which enables PG-MDD to capture the both inner- and inter-dependencies between the learner's speech and the corresponding canonical phone sequence; and (3) a novel multi-view audio encoder is introduced to exploit the fine-grained articulatory cues within learners' speech.

II. Proposed Method

A. Problem Definition

The proposed PG-MDD tackles the task of MDD through the following processing flow. Given an T-length input audio signal $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ uttered by an L2 learner and an N-length canonical phone sequence $\mathbf{q} = (q_1, q_2, \ldots, q_N)$ converted from the text prompt via a pronunciation dictionary, the proposed model detects an error state sequence $\mathbf{e} = (e_1, e_2, \ldots, e_N)$ in response to q while simultaneously generating a sequence of phonetic diagnostic feedback $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$, where y_n denotes the phone uttered by the L2 learner.

B. Model Components

The overall architecture of our proposed model is illustrated in Fig. 2(a), comprising five primary components that work together seamlessly: (1) a prompt encoder that extracts contextualized phone embeddings from the input canonical phone sequence, denoted as $\mathbf{q}$; (2) a multi-view audio encoder that leverages a pre-trained acoustic model (i.e., wav2vec 2.0) and an articulatory feature extractor to portray the learner's pronunciation from multiple views based on input audio signals, denoted as $\mathbf{x}$;(3) a concatenation network that produces fused representations from the outputs of the prompt encoder and the multi-view audio encoder; (4) an error-state detector that predicts a sequence of error states, denoted as $\mathbf{e}$, based on the fused representations in response to $\mathbf{q}$; and (5) a phone predictor that generates a sequence of phone-level dictation result, denoted as $\boldsymbol{y}$, providing phonetic diagnosis feedback.

C. Prompt Encoder

We adopt Transformer as the backbone model for the prompt encoder. Specifically, for an input canonical phone sequence $\mathbf{q}$, the contextualized phone embeddings, H_{pt}, are derived by mapping $\mathbf{q}$ to the corresponding phone embeddings and then forward propagated through a prompt encoder.

$$\mathrm{E}_{pt} = \mathrm{Embedding}(\mathbf{q}), \tag{1}$$

$$\mathrm{H}_{pt} = \mathrm{PmtEnc}(\mathrm{E}_{pt}), \tag{2}$$

where E_{pt} denotes a sequence of phone embeddings and $\mathrm{PmtEnc}(\cdot)$ denotes the proposed prompt encoder which empirically adopts a single Transformer block.

D. Multi-view Audio Encoder

Our multi-view audio encoder comprises two disparate pre-trained feature encoders, a large-scale pretrained acoustic encoder and an articulatory feature extractor. Specifically, we employ wav2vec 2.0 as the acoustic encoder to extract acoustic features, H_{ac}, from the input audio signals $\mathbf{x}$, which are expected to render the supra-segmental pronunciation cues, such as prosody [26] and intonation [27]. Meanwhile, the articulatory feature extractor takes $\mathbf{x}$ as input and generates electromagnetic articulography (EMA) features, H_{ar}, to characterize the movements of articulatory structures. Here, H_{ar} is a 12-dimensional vector sequence, with each dimension representing the x-y position of an EMA sensor for a specific articulatory structure, such as the lower incisor, upper lip, or lower lip [23].

Afterward, to tightly couple the extracted features, H_{ac} and H_{ar}, we first apply a linear projection to map the articulatory features, H_{ar}, followed by a single transformer block, resulting in $\mathrm{H'}_{ar}$. Next, we concatenate the intermediated acoustic representation, H_{ac}, with $\mathrm{H'}_{ar}$, and then employ a depth-wise convolution layer to refine the concatenated representations, generating a sequence of multi-view feature vectors, H_{mv}. The following equations illustrate the operations of the proposed multi-view audio encoder:

$$\mathrm{H}_{ac}, \mathrm{H}_{ar} = \mathrm{PreEnc}_{ac}(\mathbf{x}), \mathrm{PreEnc}_{ar}(\mathbf{x}) \tag{3}$$

$$\mathrm{H'}_{ar} = \mathrm{TFR}(\mathrm{Proj}(\mathrm{H}_{ar})) \tag{4}$$

$$\mathrm{H}_{mv} = \text{Depth-wise Conv}([\mathrm{H}_{ac}; \mathrm{H'}_{ar}]) \tag{5}$$

where $[;]$ is the concatenation operation, and Depth-wise $\mathrm{Conv}(\cdot)$ is realized as a 1-D depth-wise convolution layer with a kernel size of 3.

E. Concatenation Network

The concatenation network is designed to fuse the textual features, H_{pt}, with multi-view acoustic features, H_{mv}, enabling the model to attend to the full context of the input phone representations both within and between the textual and acoustic information via the self-attention mechanism.

$$[\overline{\mathrm{H}}_{pt}; \overline{\mathrm{H}}_{au}] = \mathrm{ConcatNet}([\mathrm{H}_{pt}; \mathrm{H}_{au}]) \tag{6}$$

where $\mathrm{ConcatNet}(\cdot)$ is a stack of 2 Transformer blocks.

F. Error-state Detector

The error-state detector is a sequential binary labeling model that receives $\overline{\mathrm{H}}_{pt} = (\overline{\mathrm{h}}_1^{pt}, \overline{\mathrm{h}}_2^{pt}, \ldots, \overline{\mathrm{h}}_N^{pt})$ from the concatenation network, and then outputs a sequence of error states $\mathbf{e} = (e_1, e_2, \ldots, e_N)$. Here each element in $\mathbf{e}$ represents the error state of n-th phone in the canonical phone sequence $\mathbf{q}$, with 1 indicating that phone q_n is pronounced incorrectly and 0 indicating correct pronunciation.

For each error state e_n of the canonical phone q_n, the probability of error detection is defined as:

$$\mathcal{P}_{\mathrm{dec}}(e_n = 1|\mathbf{x}, \mathbf{q}) = \sigma\big(\mathrm{proj}_d(\overline{\mathrm{h}}_n^{pt})\big) \tag{7}$$

where $\mathcal{P}_{\mathrm{dec}}(e_n = 1|\mathbf{x}, \mathbf{q})$ is the conditional probability which represents how likely the phone q_n is mispronounced, σ represents the nonlinear function which we used sigmoid function, proj_d is the linear projection layer which transforms the hidden dimension to a single scaler. Furthermore, to estimate the error states in a fine-grained manner, we determine the optimal threshold value for each canonical phone q_n in the development set.

G. Predictor

The phone predictor is responsible for dictating the L2 learner' speech, which takes $\overline{\mathrm{H}}_{au} = (\overline{\mathrm{h}}_1^{au}, \overline{\mathrm{h}}_2^{au}, \ldots, \overline{\mathrm{h}}_T^{au})$ as input and generates a sequence of dictated phones as phonetic diagnosis feedback. The prediction process is optimized with the connectionist temporal classification (CTC) loss which manages the alignment between the input features and the phone sequence $\boldsymbol{y}$. This CTC loss is denoted as follows:

$$\begin{aligned}\mathcal{P}_{diag}(\boldsymbol{y}|\mathbf{x}, \mathbf{q}) &= \sum_{A \in \mathcal{B}_{CTC}^{-1}(\boldsymbol{y})} \mathcal{P}(\boldsymbol{y}|A, \mathbf{x}, \mathbf{q}) \mathcal{P}(A|\mathbf{x}, \mathbf{q}) \\ &= \sum_{A \in \mathcal{B}_{CTC}^{-1}(\boldsymbol{y})} \mathcal{P}(A|\mathbf{x}, \mathbf{q})\end{aligned} \tag{8}$$

The joint probability $\mathcal{P}(A|\mathbf{x}, \mathbf{q})$ is further factorized using the probabilistic chain rule accompanied by a conditional independence assumption.

$$\begin{aligned}\mathcal{P}(A|\mathbf{x}, \mathbf{q}) &= \prod_{t=1}^{T} \mathcal{P}(\mathbf{a}_t|\mathbf{a}_1, \ldots, \mathbf{a}_{t-1}, \mathbf{x}, \mathbf{q}) \\ &\approx \prod_{t=1}^{T} \mathcal{P}(\mathbf{a}_t|\mathbf{x}, \mathbf{q})\end{aligned} \tag{9}$$

The conditional probability $\mathcal{P}(\mathbf{a}_t|\mathbf{x}, \mathbf{q})$ in Eq. (11) is computed as

$$\mathcal{P}(\mathbf{a}_t|\mathbf{x}, \mathbf{q}) = \mathrm{Softmax}(\mathrm{Proj}_{\mathrm{diag}}(\overline{\mathrm{H}}_{au})) \tag{10}$$

where $\mathrm{Proj}_{\mathrm{diag}}(\cdot)$ is a linear projection converting hidden dimension to the vocabulary size.

H. Training objective

The proposed model is trained in an end-to-end manner, with the learning process designed to optimize two objectives: pronunciation error detection and phonetic diagnosis.

Pronunciation Error Detection. The objective function of pronunciation error detection $\mathcal{L}_{det}$ is defined by the negative log likelihood of Eq. (8) over all error states:

$$\mathcal{L}_{det} = -\log \sum_{n=1}^{N} \mathcal{P}_{\text{det}}(e_n|\mathbf{x}, \mathbf{q}) \quad (11)$$

Phone Predictor. The objective function of phone predictor is defined by the negative log-likelihood of Eq. (10) over all possible alignments:

$$\mathcal{L}_{diag} = -\log \sum_{A \in \mathcal{B}_{\text{CTC}}^{-1}(\mathbf{y})} \prod_{t=1}^{T} \mathcal{P}(\mathbf{a}_t|\mathbf{x}, \mathbf{q}) \quad (12)$$

These two modules are linearly combined as the overall objective for model learning:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{det} + \mathcal{L}_{diag} \quad (13)$$

where $\alpha \in [0,1]$ is a controllable parameter: a larger α value means a higher weight on the mispronunciation diagnosis subtask and the value of α is set to be 0.3 based on the development set.

I. Inference

At inference time, a prompt-guided mechanism is put forward for the PG-MDD to strike a balance between recall and precision evaluations. We first identify erroneous pronunciation segments with the outputs of the error-state detector and the phone-dependent thresholds. Meanwhile, the N-best diagnosis results are obtained from the phone predictor, which are then aligned to the canonical phone sequence with Levenshtein distance, individually. Afterward, based on the identified mispronunciation segments, the diagnosis results are obtained by checking where the aligned candidates differ from the canonical phone sequence.

III. Experiments

A. Speech Datasets

TIMIT [21] is a widely used English read speech dataset that was designed for the development and evaluation of automatic speech recognition systems. TIMIT contains recordings of 630 speakers from 8 major dialects of American English.

L2-ARCTIC [20] is a recently released L2-English speech dataset, which was compiled for research on CAPT, accent conversion, and others. L2-ARCTIC contains both correctly pronounced and mispronounced utterances from 24 non-native speakers (12 males and 12 females), whose mother-tongue languages include Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese.

B. Experimental Setup

Experimental Settings. Before the joint optimization of mispronunciation detection and phone recognition task, we fine-tune the acoustic encoder and phone predictor using the TIMIT corpus. We then perform joint optimization of the proposed model using the L2-ARCTIC corpus. Furthermore, to unify the phone dictionaries used in TIMIT and L2-ARCTIC, we map the 61-phone set of the TIMIT corpus to the 39-phone set used in the L2-ARCTIC corpus. In our experimental setup, each attention head in the transformer block is configured as a single-head attention. All model components are optimized using the Adam optimizer [27] for 1000 steps, with an initial learning rate of $7\times10^{-5}$and a batch size of 64.

Implementation Details of the Multi-view Audio Encoder. We adopt wav2vec 2.0-large as the acoustic encoder, which consists of a feature encoder, a context network, and a quantization module. Each component of wav2vec 2.0 is jointly learned through a combination of mask prediction and contrastive learning. Furthermore, wav2vec 2.0 achieves state-of-the-art results in phone recognition, which is highly relevant to CAPT. For the articulatory feature extractor, we adopt an acoustic-to-articulatory inversion (AAI) pre-trained model. This pre-trained model utilizes self-supervised learning techniques to generalize across unseen speaker data, mapping speech signals to an articulatory space that captures the mechanisms of speech production. The model has been trained on the Electromagnetic Articulography (EMA) dataset and demonstrates the interpretability of its feature representations by validating estimated feature representations against actual speech production behaviors.

TABLE I
Experimental results of different methods on mispronunciation detection.

Model	PR (%)↑	RE (%)↑	F1 (%)↑
GOP [9]	69.97	32.54	44.42
CNN-RNN-CTC [13]	**74.78**	36.76	49.29
APL-2 [25]	52.79	54.49	53.62
CoCA-MDD [19]	64.65	51.01	57.03
MDDGCN [16]	51.90	**61.97**	56.49
RNN-T [24]	63.40	55.30	59.10
PG-MDD	60.15	60.06	**60.10**

Evaluation Metrics. We assess the performance of our proposed method using Precision (PR), Recall (RE), and F1-score. For canonical phonemes, true acceptance (TA) and true rejection (TR) denote correctly identified phonemes, while false acceptance (FA) and false rejection (FR) denote misclassified phonemes. These metrics are used to comprehensively evaluate our model's effectiveness.

IV. Experimental Results

A. Comparison with Prior Arts

At the outset, we report on the mispronunciation detection results in Table I to compare the proposed method with several cutting-edge baselines. There are three noteworthy points to these results. First, all the dictation-based methods (CNN-RNN-CTC, APL-2, CoCA-MDD, and RNN-T) significantly outperform the conventional scoring-based method, GOP, in terms of F1-score, demonstrating the promising potential of dictation-based models for MDD tasks. Second, most detection-based methods exhibit high precision but yield inferior results in recall. This result underscores the challenge that dictation-based MDD models face in striking a balance between precision and recall when detecting pronunciation errors. Conversely, the prompt-based model (MDDGCN) tends to aggressively detect pronunciation errors in L2 learners' pronunciations, leading to a higher recall while resulting in a decrease in precision. Finally, the proposed PG-MDD model integrates a multi-view encoder and employs a prompt-guided mechanism, which not only outperforms other state-of-the-art methods in terms of F1-score but also mitigates the limitations in both precision and recall encountered by previous models.

TABLE II
Ablation studies on PG-MDD.

Model	Canonicals		Mispronunciations			RE (%)↑	PR (%)↑	F1 (%)↑	PER(%)↓
	TA (%)↑	FR (%)↓	FA (%)↓	True Rejection					
				Corr Diag. (%)↑	Diag. Error (%)↓				
PG-MDD	93.36	6.63	**39.94**	76.87	23.13	**60.06**	60.15	**60.10**	13.92
-w/o PG	**94.87**	**5.13**	44.33	**80.20**	**19.80**	55.67	**64.45**	59.74	**12.80**
-w/o ES	94.57	5.43	45.19	79.46	20.54	54.81	62.74	58.51	13.20
-w/o ES and PmtEnc	93.69	6.31	43.14	78.20	21.80	56.86	60.07	58.42	13.97
-w/o ES, PmtEnc, and ArEnc	94.05	5.95	45.63	78.35	21.65	54.37	60.38	57.22	13.98

TABLE III
Impact of articulatory features on phone-level substitution error detection and diagnosis.

Error Type	w/o Articulatory Feats		w/ Articulatory Feats	
	Correct Detection	Correct Diag.	Correct Detection	Correct Diag.
z->s	74.62	72.61	**79.64**	**77.63**
v->f	11.86	11.86	**28.81**	**27.11**
ih->iy	28.30	27.67	**30.81**	**30.18**
ow->ao	11.60	10.71	**16.96**	**11.39**
er->ah	**73.46**	**60.20**	66.36	47.95
dh->d	**89.90**	**84.86**	70.64	67.20
d->t	34.24	27.39	**38.35**	**35.61**
p->b	**31.65**	**29.68**	17.18	15.62
th->t	**84.61**	**75.00**	80.76	67.30

B. Ablation Studies

Next, a series of model ablation experiments are conducted, as shown in Table II, to demonstrate the importance of each component in our PG-MDD architecture. This ablation study reports the following settings: 1) inference without the prompt-guided mechanism (-w/o PG) 2) removing the error-state detector (-w/o ES), 3) removing the error-state detector and the prompt encoder (- w/o ES and PmtEnc), and 4) removing error-state detector, the prompt encoder, and the articulatory feature extractor from the multi-view audio encoder (- w/o ES, PmtEnc, and ArEnc). The main motivation of this paper is to balance the performance of recall and precision while preserving the high F1-score. If PG-MDD inference is conducted without the proposed prompt-guided mechanism, the model performance deteriorates in F1-score, leading to an imbalance between recall and precision, even though a more accurate phone recognition rate is achieved. Next, when the error-state detector is removed, performance declines across all evaluation metrics compared to PG-MDD. This result suggests that integrating an error-state detector into the dictation-based MDD model enhances the effectiveness of the mispronunciation detection and diagnosis task. Finally, we assess the impact of articulatory features by comparing the last two rows in Table II. We observe that when the articulatory feature extractor is removed, performance declines in terms of recall while achieving a comparable phone error rate. We attribute this to the fact to articulatory features that are beneficial for the task of mispronunciation detection.

C. Impact of Articulatory Features on MDD

Going one step further, we delve into the impact of articulatory features by reporting the performance changes in the common mispronunciation patterns [20]. From Table III, it is evident that the articulatory features provide more subtle pronunciation cues for the proposed PG-MDD, thereby enhancing the performance of mispronunciation diagnosis for the frequently mispronounced phone pairs, such as (/z/, /s/) and (/ih/, /iy/). Notably, articulatory features show significant improvement for the mispronunciation pattern (/v/, /f/). On the other hand, some mispronunciation patterns result in inferior performance with articulatory features, such as (/p/, /b/).

V. Conclusion

In this paper, we have proposed a novel prompt-guided dictation-based mispronunciation detection and diagnosis (MDD) model, termed PG-MDD. Our model integrates the strengths of a multi-view audio encoder and a prompt-guided mechanism within a unified framework, which jointly optimizes the processes of mispronunciation detection and diagnosis. This approach effectively addresses the common issue of how to properly strike a balance between precision and recall rates, facing dictation-based models. A comprehensive set of empirical results on the L2-ARCTIC benchmark dataset shows that our proposed architecture significantly outperforms several competitive baselines. PG-MDD achieved the highest F1-score, demonstrating robust performance in both precision and recall. This balanced performance is particularly crucial for providing reliable and comprehensive feedback in practical applications, ensuring more effective learning outcomes for L2 learners.

References

[1] Nancy F. Chen, and Haizhou Li. “Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning,” in proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1–7, 2016.

[2] Alistair V. Moere and Ryan Downey, “Technology and artificial intelligence in language assessment,” Handbook of second language assessment, pp. 341–358, 2016.

[3] Keelan Evanini, Xinhao Wang, "Automated speech scoring for non-native middle school students with multiple task types," in Proceedings of Interspeech (INTERSPEECH), pp. 2435–2439, 2013.

[4] Keelan Evanini, Maurice Cogan Hauck, Kenji Hakuta, "Approaches to automated scoring of speaking for K–12 English language proficiency assessments," ETS Research Report Series, pp. 1–11, 2017.

[5] Ramsey Cardwell, Ben Naismith, Geoffrey T. LaFlair, and Steven Nydick. "Duolingo English test: technical manual," Duolingo Research Report, 2022.

[6] Wei Li, Nancy F. Chen, Sabato Marco Siniscalchi, and Chin-Hui Lee, "Improving mispronunciation detection of mandarin tones for non-native learners with soft-target tone labels and BLSTM-based deep tone models," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 12, pp. 2012–2024, 2019.

[7] Kun Li, Shaoguang Mao, Xu Li, Zhiyong Wu, and Helen Meng, "Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks," Speech Communication, vol. 96, pp. 28–36, 2018.

[8] Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, Shira Calamaro, and Bozena Kostek, "Weakly-supervised word-level pronunciation error detection in non-native English speech," in proceedings of Interspeech (INTERSPEECH), pp. 4408–4412, 2021.

[9] Silke M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Communication, vol. 30, pp. 95–108, 2000.

[10] Quy-Thao Truong, Tsuneo Kato, and Seiichi Yamamoto, "Automatic assessment of L2 English word prosody using weighted distances of F0 and intensity contours," in proceedings of Interspeech (INTERSPEECH), pp. 2186-2190, 2018.

[11] Jiatong Shi, Nan Huo, and Qin Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," in proceedings of Interspeech (INTERSPEECH), pp. 3057–3061, 2020.

[12] Shaoguang Mao, Zhiyong Wu, Runnan Li, Xu Li, Helen Meng, and Lianhong Cai, "Applying multitask learning to acoustic-phonemic model for mispronunciation detection and diagnosis in L2 English speech," in proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 6254–6258, 2018.

[13] Wai-Kim Leung, Xunying Liu, and Helen Meng, "CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis," in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8132-8136, 2019.

[14] Bi-Cheng Yan, Meng-Che Wu, Hsiao-Tsung Hung, Berlin Chen, "An end-to-end mispronunciation detection system for L2 English speech leveraging novel anti-phone modeling," in proceedings of Interspeech (INTERSPEECH), pp. 3032–3036, 2020.

[15] Bi-Cheng Yan, Hsin-Wei Wang, and Berlin Chen, "Peppanet: Effective mispronunciation detection and diagnosis leveraging phonetic, phonological, and acoustic cues," in Proceedings of IEEE Spoken Language Technology Workshop (SLT), pp. 1045–1051, 2023.

[16] Bi-Cheng Yan, Hsin-Wei. Wang, Yi-Cheng Wang, and Berlin Chen, "Effective graph- based modeling of articulation traits for mispronunciation detection and diagnosis," in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023.

[17] Linkai Peng, Kaiqi Fu, Binghuai Lin, Deng Feng Ke, and Jinsong Zhan, "A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis," in proceedings of Interspeech (INTERSPEECH), pp. 4448–4452, 2021.

[18] Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng, "Transformer based end-to-end mispronunciation detection and diagnosis," in proceedings of Interspeech (INTERSPEECH), pp. 3954–3958, 2021.

[19] Nianzu Zheng, Liqun Deng, Wenyong Huang, Yu Ting Yeung, Baohua Xu, Yuanyuan Guo, Yasheng Wang, Xiao Chen, Xin Jiang, and Qun Liu, "Cca-mdd: A coupled crossattention based framework for streaming mispronunciation detection and diagnosis," arXiv preprint arXiv:2111.08191, 2021.

[20] Guanlong Zhao, Sinem Sonsaat, A. Silpachai, I. Lucic, E. C.-Hudilainen, J. Levis and R. G.-Osuna, "L2-ARCTIC: A Non-native English Speech Corpus," in proceedings of Interspeech (INTERSPEECH), pp. 2783–2787, 2018.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM," Linguistic Data Consortium, 1993.

[22] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in neural information processing systems 33, pp. 12449-12460, 2022.

[23] Peter Wu, Li-Wei Chen, Cheol Jun Cho, Shinji Watanabe, Louis Goldstein, Alan W Black, Gopala K. Anumanchipalli, "Speaker-Independent Acoustic-to-Articulatory Speech Inversion," in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2023.

[24] Daniel Yue Zhang, Soumya Saha and Sarah Campbell , "Phonetic RNN-Transducer for Mispronunciation Diagnosis," in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2023.

[25] Wenxuan Ye, Shaoguang Mao, Frank Soong, Wenshan Wu, Yan Xia, Jonathan Tien, Zhiyong Wu "An Approach to Mispronunciation Detection and Diagnosis with Acoustic, Phonetic and Linguistic (APL) Embeddings," in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6827-6831, 2022.

[26] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3M: An effective multi-view, multi-granularity, and multi-aspect modeling approach to English pronunciation assessment," in proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 575-582, 2022.

[27] S. Bannò, and M. Matassoni, "Proficiency assessment of L2 spoken English using wav2vec 2.0," in IEEE Spoken Language Technology Workshop (SLT), pp. 1088-1095, 2022.

[28] C. McGhee, K. Knill, and M. Gales, "Towards acoustic-to-articulatory inversion for pronunciation training," In Workshop on Speech and Language Technology in Education (SLaTE), pp. 66-70, 2023.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[30] Konstantinos Kyriakopoulos, Kate M. Knill, and Mark J.F. Gales, "Automatic detection of accent and lexical pronunciation errors in spontaneous nonnative English speech," in proceedings of Interspeech (INTERSPEECH), pp. 3052–3056, 2020.

[31] Yassine El Kheir, Shammur Absar Chowdhury, and Ahmed Ali, "Multi-view multi-task representation learning for mispronunciation detection," in arXiv preprint arXiv:2306.01845, 2023.

[32] Daniel-Yue Zhang, Soumya Saha, and Sarah Campbell, "Phonetic RNN-transducer for mispronunciation diagnosis," in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023

An Explainable Raman Spectral Classification Pipeline via NMF and SHAP: A Case Study of Pen Ink Colors

Pongpon Lapsatid[*,1], Somrudee Deepaisarn[*,†,2] and Pitak Eiamchai[‡,3]
[*] Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
[†] Research Unit in Sustainable Electrochemical Intelligent, Thammasat University, Pathum Thani, Thailand
[‡] National Electronics and Computer Technology Center, Pathum Thani, Thailand
[1] E-mail: m6622040316@g.siit.tu.ac.th, [2] E-mail: somrudee@siit.tu.ac.th (Corresponding Author),
[3] E-mail: pitak.eiamchai@nectec.or.th

***Abstract*—Ink analysis, especially the analysis of coloring components, plays a crucial role in forensic science, as it provides useful information to validate the authenticity of the questioned documents. Raman spectroscopy offers a nondestructive and time- and cost-effective alternative to forensic investigations, providing unique spectral fingerprints of pen ink samples. There are several attempts to analyze Raman spectra using computational techniques. This work proposes an explainable pipeline for Raman spectral classification to distinguish pen ink colors. Preprocessing steps were applied to improve the spectral signal-to-noise ratio. Data augmentation was conducted to increase the training data. Using the weights of the 48 extracted non-negative matrix factorization (NMF) components as inputs for the Random Forest classifier, the test accuracy of 0.8456 was achieved. Moreover, NMF allows underlying linear compositions of spectra associated with ink color classes to be observed. SHapley Additive exPlanations (SHAP) then enables a quantification of the average impact of each NMF component on the classification. Through this pipeline, the scalar multiplication between NMF components, their weights, and their SHAP values enables a human-intuitive explainability of the model prediction results that captures unique Raman spectral patterns as expected for each ink color class.**

I. Introduction

Ink analysis is an essential procedure in forensic science that offers significant insights into the authenticity of the questioned documents by analyzing the dyes and pigments according to their chemical compositions. It plays a significant role in the questioned document examination (QDE) – a fundamental procedure to provide valuable insights in finding the authenticity of a questioned document – which attracts wide interest in both a commercial sense and a legal sense [1], [2], [3]. Physicochemical properties, especially the coloring components, are crucial in such forensic investigation [4]. The traditional chromatographic methods used to analyze physicochemical properties, such as thin-layer chromatography (TLC), high-performance liquid chromatography (HPLC), and gas chromatography-mass spectrometry (GC-MS), yield a high level of effectiveness in distinguishing inks based on their chemical characteristics. Such methods, however, require labor-intensive sample preparation processes and bulky measurement equipment in physical laboratories [5].

Alternatively, Raman spectroscopy (RS) is well-known as a non-destructive approach to collecting the molecular fingerprint representations of substances based on the measurement of the molecular vibrational energy stages following the Raman scattering principles [6], [7]. The result of the RS procedure is the acquisition of the Raman spectra – each spectrum is a one-dimensional signal whose peaks' positions and their relative intensities represent molecular fingerprint information [7]. However, the Raman spectra obtained from conventional RS often suffer from weak signal intensity and fluorescence interference. As a countermeasure, surface-enhanced Raman scattering (SERS) is conducted to intensify the Raman peaks to increase the signal-to-noise ratio [7].

RS and SERS hold significance in real-world applications in pen ink analysis for QDE. The non-destructive nature of RS and SERS is suitable for analyzing highly valuable questioned documents without causing damage to the sample, which is very beneficial to the judicial process. Ref. [5] reported that SERS can offer high discriminative power when analyzing the pen ink Raman spectra. Nevertheless, the traditional chemometric analytical procedures on the Raman spectra remain labor-intensive.

Fortunately, the paradigm of QDE is shifting. Ref. [8], conducted from 2019 to 2022 by Interpol, reviewed several research works and discovered the burgeoning trend of machine learning (ML) applications in forensic document examinations. For document analysis related to writing instruments' ink, RS, and SERS, the ML models often found in the review are principal component analysis (PCA) and the variants of linear discriminant analysis (LDA). Meanwhile,

non-negative matrix factorization (NMF) [9] is capable of extracting nonnegative components, which aligns well with the characteristic of Raman spectra detected from the counts of photons and, thus, nonnegative in nature. Although NMF was used in [10] and yielded poor visual results, the spectral dataset in the study was collected from stamp inks which have different characteristics from writing pen inks, and their spectra were acquired by pixel fusion from hyperspectral images – thus leaving a gap on applying NMF to treat Raman spectra of writing instruments' inks.

Additionally, as [8] concluded, the number of qualified forensic document examiners (FDE) is decreasing in various regions. Considering that training a qualified FDE requires comparatively longer time than other fields in forensics, the FDE field is at risk of FDE shortage in the long run [8]. Therefore, in order to alleviate the current workload of the FDE experts and enable more time for training new generations of FDEs, the field is in need of an automated analytical pipeline that can handle the challenges in spectral analysis such as low signal-to-noise ratio, deliver fast prediction results, and provide a degree of explainability to reassure the FDE domain experts that the pipeline is interpretable and applicable to the real-world pen ink Raman spectral analysis applications.

Our study contributes an explainable ML pipeline for pen ink color classification by using NMF and SHAP for analyzing and explaining the Raman spectral features of the writing instruments' inks. Our dataset was collected from blue, black, and red pen ink samples randomly bought from Thai stationery stores. The proposed pipeline consists of preprocessing steps to tackle low signal-to-noise ratio issues in Raman spectra, a data augmentation step to increase the training data for better ML model training, an NMF model for dimensionality reduction and feature extraction, a machine learning classifier for color classification, and a model explaining step utilizing the NMF components and their SHAP values to construct the spectral feature representations of each class. Our experiment serves as an initial study, paving the way for further complex real-world forensic applications.

II. METHODOLOGY

A. Dataset

Initially, 145 pens were collected from Thailand's markets. There are 53 blue pens, 42 black pens, and 50 red pens where each pen is of a unique model. All pen ink samples underwent SERS measurement using a confocal Raman microscope (inVia, Renishaw®, UK) with optimal parameter setting to acquire Raman spectral dataset. As a result, 9 Raman spectra from each pen are collected via Raman mapping. Each Raman spectrum consists of 1,011 data points ranging from approximately wavenumber 397 to 1,543 cm^{-1}. The raw spectra contain inter-class and intra-class variances, especially in the intensity level and pattern at each wavenumber. Fig. 1 visualizes some samples of raw spectra from each color:

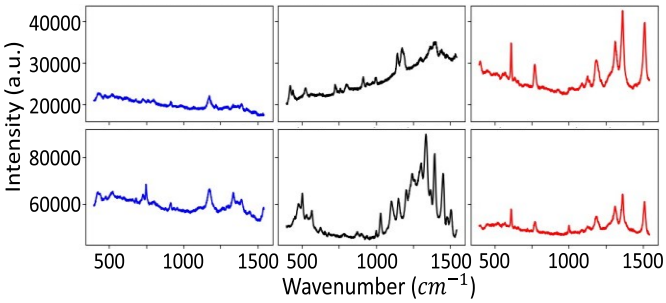


Fig. 1 Sample raw spectra from each pen color.

To train a classification model, the total of 1,305 spectra of pen ink samples are divided into training: validation: test set with the ratio of 40:30:30 for each ink color, resulting in 513: 396:396 raw spectra, where the training spectral data later undergo the data augmentation in Section II-C.

B. Spectral Pre-processing Steps

To handle the low signal-to-noise ratio issue, each raw spectrum needs to be pre-processed prior to the modeling step to account for handling the cosmic ray noise in the spectra, baseline issues, low-intensity issues, and the unwanted noise generated by the baseline correction step. The pre-processing steps include cosmic ray removal using the Whitaker-Hayes despiking algorithm [11] with a window size of 10 and a threshold of 13, smoothing using the Savitzky-Golay filter [12], [13] with a window length of 9 and a polynomial order of 3, baseline correction with IModPoly algorithm [14], [15], [16] with a polynomial order of 7. Then, [11] is applied again to despike the first 50 and the last 50 wavenumbers by setting the window size to 10 and the threshold to 1. Finally, each spectrum is normalized with respect to its own maximum value to handle low-intensity issues.

C. Data Augmentation using Weighted Linear Combination

The raw 513 original spectra in the training set, consisting of 189 blue, 144 black, and 180 red pen ink spectra, undergo data augmentation using the weighted linear combination technique [6] to increase the number of training spectra and enhance the machine learning performance, as the technique synthesizes Raman spectra while preserving the spectral characteristic of the original spectra effectively [6]. Each synthetic spectrum is generated from two original Raman spectra acquired from the same pen following (1):

$$\hat{x} = x_i + (w \cdot x_j) \quad (1)$$

where $\hat{x}$ is the augmented synthetic spectrum, x_i, x_j are two original spectra, and w is a scaler weight ranging from (0, 1). In this study, the weights are set at w = 0.01, 0.02, … , 0.09, 0.10. As a result, 36 synthetic spectra can be generated from

one pen in the training set, creating 20,520 synthetic training data. The synthetic training set is sampled from those 20,520 augmented spectra so that each class consists of 5,700 augmented spectra to ensure class balance. Each augmented spectrum is normalized with respect to its maximum intensity before proceeding to the next step.

D. Non-negative Matrix Factorization (NMF)

Nonnegative matrix factorization (NMF) [9] is a matrix factorization technique that can be used for dimensionality reduction and feature extraction. It can factorize any two-dimensional nonnegative matrix into a low-dimensional feature representation matrix and a component matrix [17]. Equation (2) describes the NMF's factorization model [18], [19] and the notations used in our study:

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}} \qquad \boldsymbol{X} \in \mathbb{R}^{M\times N}, \quad \boldsymbol{W} \in \mathbb{R}^{M\times R}, \quad \boldsymbol{H} \in \mathbb{R}^{N\times R} \tag{2}$$

where $\boldsymbol{X}$ is the $M \times N$ data matrix which stores all spectra in a given dataset, M is the number of spectra in the given dataset, m is the index of a spectrum in the given dataset where $m \in \{0, 1, \dots, M-1\}$, N is the number of data points in each spectrum (i.e., the number of dimensions of a spectrum in $\boldsymbol{X}$), R is the number of NMF components in the NMF model where $R \leq N$, r is an NMF component index where $r \in \{0, 1, \dots, \mathrm{R}-1\}$, $\boldsymbol{W}$ is the low-dimensional feature representation matrix where each row of $\boldsymbol{W}$, $\boldsymbol{w_m}$, is the R-dimensional representation of the mth spectrum in the given dataset, and $\boldsymbol{H}$ is the component matrix where each rth column of $\boldsymbol{H}$, $\boldsymbol{h_r}$, is the rth NMF component.

In this study, the number of NMF components, R, was varied. Eventually, 48 components were found appropriate to capture the real variations of the spectra for color classification. Note that too many components extracted may lead to random variations being captured and contaminating the classification model. The selected 48-component NMF model is fitted to all original spectra. The trained NMF model is then utilized to transform each pre-processed spectrum into a lower-dimensional representation with 48 features for training the classification model.

E. Machine Learning Modeling

"PyCaret" [20] AutoML library provides the capability to explore various classical machine learning classifiers including Random Forest (RF) [21], Decision Tree (DT) [22], and K-Nearest Neighbors (KNN) [23], [24]. The 10-fold cross-validation is applied, and the classical machine learning classifiers are trained with our synthetic training set after being transformed by NMF into 48-dimensional spectra. The evaluation metrics used in our study to select the best color classification model are accuracy and area under the receiver operating characteristic curve (AUC), whose definitions are well-summarized in [25], [26].

F. Explaining Model using SHAP

Shapley Additive exPlanations (SHAP) is a technique that quantifies the importance of each feature in predicting the result of a classification model following a game theoretic approach [27]. One can interpret the visualization of SHAP values and understand the importance of each feature intuitively. In our proposed pipeline, the mean absolute SHAP value for each NMF component is obtained through the model training process, which informs the average impact of each NMF component on the classification of each ink color spectrum. The calculation of SHAP values is implemented through SHAP libraries in Python [27], [28] with their default hyperparameters.

III. Results and Discussions

A. Spectral Pre-processing Results

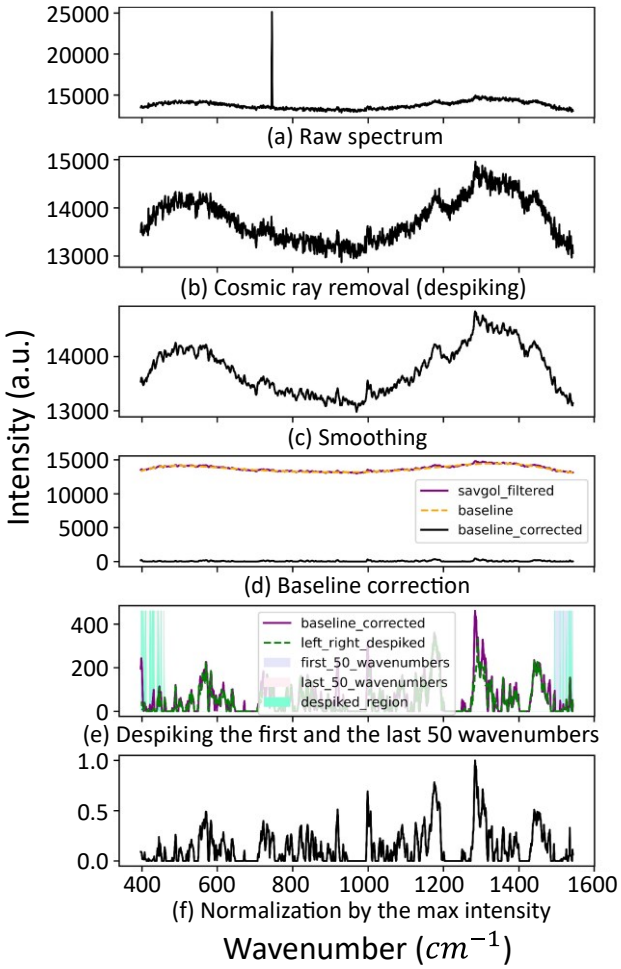


Fig. 2 A noisy raw spectra before pre-processing (a) and after pre-processing at each step in (b), (c), (d), (e), and (f)

Fig. 2 demonstrates each data pre-processing step and the results. Our pre-processing pipeline is capable of handling the cosmic ray noise in the spectra (although some minor portions of feature peaks might be slightly distorted, which can be one of the improvement points to be considered in future works), baseline issues, low-intensity issues, and the unwanted noise

generated by the baseline correction step, which can improve the signal-to-noise ratio of the spectra.

B. Data Augmentation for the Training Set

Section II-C provides the synthetic training set which would later undergo the NMF feature extraction process and machine learning modeling. The final number of the pre-processed spectra in the training set, the validation set, and the test set used for the NMF feature extraction process and machine learning modeling are summarized in Table 1.

TABLE 1
THE FINAL NUMBER OF PRE-PROCESSED SPECTRA IN THE TRAINING SET, VALIDATION SET, AND TEST SET FOR NMF FEATURE EXTRACTION AND MACHINE LEARNING MODELING

Dataset	Number of Blue Spectra	Number of Black Spectra	Number of Red Spectra	Total Number of Spectra
Train (Synthetic)	5,700	5,700	5,700	17,100
Validation	144	117	135	396
Test	143	117	135	395

C. Model Selection and Evaluation

The top 3 best models suggested by PyCaret are K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF) classifiers – each with 1.00 training accuracy and AUC. To select the suitable model among these three, their performances on the validation set and the test set are observed, as shown in Table 2:

TABLE 2
TOP 3 MODELS EVALUATED BY ACCURACY AND AUC ON THE VALIDATION SET AND THE TEST SET

Model	Validation Accuracy	Test Accuracy	Validation AUC	Test AUC
KNN	**0.7525**	0.7595	0.8139	0.8226
DT	0.6591	0.7063	0.7483	0.7847
RF	0.7298	**0.8456**	**0.9056**	**0.9630**

The RF model performed the best in terms of validation AUC, test accuracy, and test AUC. Thus, it is selected as the best model in our study. Fig. 3 visualizes the confusion matrix showing the performance of the RF model on the test set.

D. Model Explainability

Our proposed pipeline utilizes three types of key information – the NMF components, the weight of each NMF component of a spectrum, and the importance of each NMF component for classifying each pen ink color to explain the model

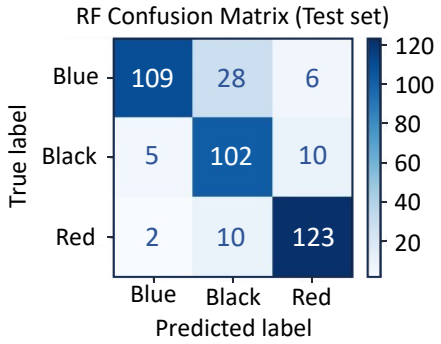


Fig. 3 Confusion matrix evaluating our RF model on the test set.

performance. Referring to the definitions in Section II-D, each of these three key information is described as follows.

Firstly, NMF is applied to extract 48 spectral components, hence reducing the dimensionality of data and capturing the underlying spectral characteristics. Each extracted component is denoted by $\boldsymbol{h_r}$ where $r \in \{0, 1, \ldots, 47\}$.

Secondly, each spectrum in the training set, the validation set, and the test set is transformed into $\boldsymbol{w_m}$ as a result of NMF matrix factorization. Let us denote the rth element of $\boldsymbol{w_m}$ as $w_{m,r}$. In our study, $w_{m,r}$ is referred to as "the weight of the rth NMF component of the mth spectrum".

Thirdly, the importance of each NMF component toward predicting each color is calculated from the mean of the absolute SHAP value. By ranking the NMF components based on their mean absolute values from the greatest to the smallest, the most important NMF components for classifying each pen ink color can be identified. Let us denote $importance_{r,k} = mean(|SHAP\ Value|)_{r,k}$ which is the importance of rth NMF component toward the prediction of class k where k can be "Blue", "Black", or "Red".

To demonstrate the model explainability methods of our pipeline, a spectrum from the validation set was picked. Let us denote that spectrum as "sig_50", and the ground truth class of sig_50 was class "Blue". Fig. 4 visualizes the top 10 most important NMF components for classifying the blue color which are ranked based on $importance_{r,Blue}$ and Fig. 5

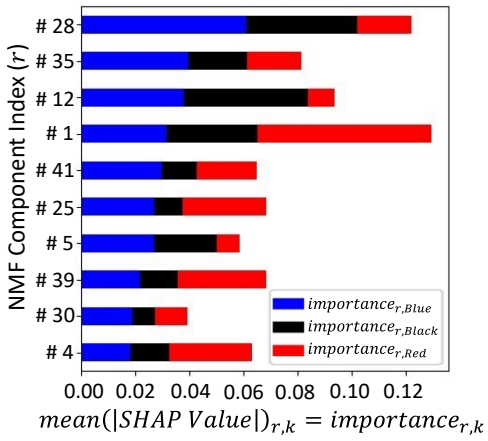


Fig. 4 Top 10 most important NMF components for class "Blue", sorted by $importance_{r,Blue}$.

demonstrates how to utilize $\boldsymbol{h_r}$, $w_{m,r}$, and $importance_{r,k}$ to explain the model prediction on any given spectrum.

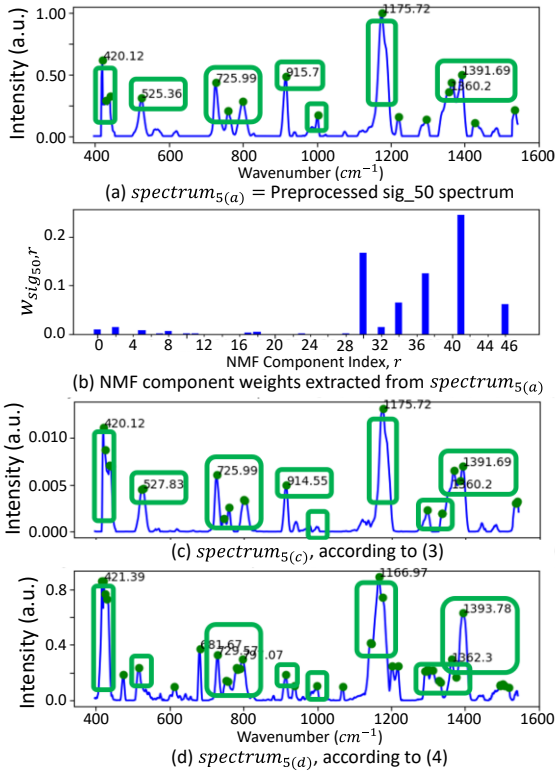


Fig. 5 Model Explainability via NMF and SHAP

Fig. 5(a) displays the sig_50 spectrum after passing through the pre-processing steps in Section II-B. Fig. 5(b) is the $spectrum_{5(a)}$ after being transformed by the NMF model into a 48-data-point spectrum – each data point is $w_{sig_50,r}$ where $r = 0,1,\ldots,47$ (zero-indexed), and visualized as a bar graph.

Fig. 5(c) is a feature representation of the pre-processed sig_50 as perceived by the classification model, which can be constructed by (3):

$$spectrum_{5(c)} = \sum_r importance_{r,k} * w_{sig_50,r} * \boldsymbol{h_r} \quad (3)$$

where $r \in \{r \mid Top\ 10\ Highest\ w_{sig_50,r}\}$ and $k =$ the label of class "Blue" in this example. The term $w_{sig_50,r} * \boldsymbol{h_r}$ approximates the original value of sig_50 according to the NMF factorization principle, and the term $importance_{r,k}$ helps to suppress some Raman features that are not perceived to be important to class k by the trained classification model. Thus, the $spectrum_{5(c)}$ offers insights into the feature representation of sig_50 as perceived by the RF model.

Fig. 5(d) is a feature representation of class "Blue" as learned by the trained classification model, which is constructed by (4):

$$spectrum_{5(d)} = \sum_r importance_{r,k} * \boldsymbol{h_r} \quad (4)$$

where $r \in \{r \mid Top\ 10\ Highest\ importance_{r,k}\}$ and $k =$ the label of class "Blue" in this example. The $spectrum_{5(d)}$ is the summation of the NMF components – each weighted by its overall importance toward classifying a given data as being class k. Thus, it offers insights into the feature representation of class "Blue" as learned by the RF model in this example.

A significant amount of shared Raman peak features (as highlighted by the green rectangular shapes) can be found between the feature representation of sig_50 in Fig. 5(c) and the feature representation of the color "Blue" in Fig. 5(d). In other words, the classification model can detect the important features of class "Blue" from sig_50. The prediction score of class "Blue" for sig_50 is 0.86, while the prediction scores for class "Black" and "Red" are 0.11 and 0.03, respectively. Therefore, the high prediction score of class "Blue" can be explained by the high degree of similarity among the Raman peak features between $spectrum_{5(c)}$ and $spectrum_{5(d)}$.

IV. Conclusions And Future Works

Our study contributes an explainable ML pipeline for pen ink color classification on Raman spectral datasets, which offers a fast and accurate prediction and a higher degree of explainability to forensic investigators and Raman spectroscopists. The proposed method opens new possibilities to interpret pen ink Raman spectra in a human-intuitive approach by reconstructing the learned feature representations of each class by the NMF components, their corresponding weights, and their SHAP values. Future works include the utilization of the learned feature representation of each pen color for more complex pen ink Raman analysis tasks, such as the identification or discrimination of the potential pen models from the ink samples in forensic document examination.

V. Acknowledgment

This research is financially supported by the National Research Council of Thailand, Thailand Advanced Institute of Science and Technology, National Science and Technology Development Agency, Institute of Science Tokyo (formerly Tokyo Institute of Technology), Sirindhorn International Institute of Technology, Thammasat University under the TAIST-Tokyo Tech Program. We gratefully acknowledge fruitful research discussion and advice from Professor Konstantinos Slavakis, Institute of Science Tokyo, Japan.

References

[1] A. V. Kravchenko, D. V. Makeeva, I. Tumkin, and A. V. Kalinichev, "Rapid analysis and classification of blue ballpoint pen inks using high-performance thin-layer chromatography and digital color analysis," Talanta, vol. 275, p. 126117, Aug. 2024, doi: 10.1016/j.talanta.2024.126117.

[2] A. Braz, M. López-López, and C. García-Ruiz, "Raman spectroscopy for forensic analysis of inks in questioned documents," Forensic Sci Int, vol. 232, no. 1–3, pp. 206–212, Oct. 2013, doi: 10.1016/j.forsciint.2013.07.017.

[3] L. A. Chesson, B. J. Tipple, J. E. Barnette, T. E. Cerling, and J. R. Ehleringer, "The potential for application of ink stable isotope

analysis in questioned document examination," Science & Justice, vol. 55, no. 1, pp. 27–33, Jan. 2015, doi: 10.1016/j.scijus.2014.05.010.

[4] N. Kapoor, P. Sulke, R. K. Shukla, R. Kakad, P. Pardeshi, and A. Badiye, "Forensic analytical approaches to the dating of documents: An overview," Microchemical Journal, vol. 170, p. 106722, Nov. 2021, doi: 10.1016/j.microc.2021.106722.

[5] N. Nuntawong et al., "Improved discrimination of pen inks on document by surface-enhanced Raman substrate fabricated by magnetron sputtering," Optik (Stuttg), vol. 201, p. 163499, Jan. 2020, doi: 10.1016/j.ijleo.2019.163499.

[6] S. Laitrakun et al., "Toward Practical Augmentation of Raman Spectra for Deep Learning Classification of Contamination in HDD," Journal of information and communication convergence engineering, vol. 21, no. 3, pp. 208–215, Sep. 2023, doi: 10.56977/jicce.2023.21.3.208.

[7] A. Phetsahai, P. Eiamchai, K. Thamaphat, and P. Limsuwan, "The Morphological Evolution of Self-Assembled Silver Nanoparticles under Photoirradiation and Their SERS Performance," Processes, vol. 11, no. 7, p. 2207, Jul. 2023, doi: 10.3390/pr11072207.

[8] M. Deviterne-Lapeyre and S. Ibrahim, "Interpol questioned documents review 2019–2022," Forensic Sci Int, vol. 6, p. 100300, 2023, doi: 10.1016/j.fsisyn.2022.100300.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: 10.1038/44565.

[10] S. Wang, H. He, R. Lv, W. He, C. Li, and N. Cai, "Classification modeling method for hyperspectral stamp-pad ink data based on one-dimensional convolutional neural network," J Forensic Sci, vol. 67, no. 2, pp. 550–561, Mar. 2022, doi: 10.1111/1556-4029.14909.

[11] D. A. Whitaker and K. Hayes, "A simple algorithm for despiking Raman spectra," Chemometrics and Intelligent Laboratory Systems, vol. 179, pp. 82–84, Aug. 2018, doi: 10.1016/j.chemolab.2018.06.009.

[12] Abraham. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.," Anal Chem, vol. 36, no. 8, pp. 1627–1639, Jul. 1964, doi: 10.1021/ac60214a047.

[13] J. Luo, K. Ying, and J. Bai, "Savitzky–Golay smoothing and differentiation filter for even number data," Signal Processing, vol. 85, no. 7, pp. 1429–1434, Jul. 2005, doi: 10.1016/j.sigpro.2005.02.002.

[14] F. Gan, G. Ruan, and J. Mo, "Baseline correction by improved iterative polynomial fitting with automatic threshold," Chemometrics and Intelligent Laboratory Systems, vol. 82, no. 1–2, pp. 59–65, May 2006, doi: 10.1016/j.chemolab.2005.08.009.

[15] C. A. Lieber and A. Mahadevan-Jansen, "Automated Method for Subtraction of Fluorescence from Biological Raman Spectra," Appl Spectrosc, vol. 57, no. 11, pp. 1363–1367, Nov. 2003, doi: 10.1366/000370203322554518.

[16] J. Zhao, H. Lui, D. I. McLean, and H. Zeng, "Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy," Appl Spectrosc, vol. 61, no. 11, pp. 1225–1232, Nov. 2007, doi: 10.1366/000370207782597003.

[17] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative Matrix Factorization for Signal and Data Analytics: Identifiability, Algorithms, and Applications," IEEE Signal Process Mag, vol. 36, no. 2, pp. 59–80, Mar. 2019, doi: 10.1109/MSP.2018.2877582.

[18] A. CICHOCKI and A.-H. PHAN, "Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E92-A, no. 3, pp. 708–721, 2009, doi: 10.1587/transfun.E92.A.708.

[19] C. Févotte and J. Idier, "Algorithms for Nonnegative Matrix Factorization with the β-Divergence," Neural Comput, vol. 23, no. 9, pp. 2421–2456, Sep. 2011, doi: 10.1162/NECO_a_00168.

[20] M. Ali, "PyCaret: An open source, low-code machine learning library in Python," PyCaret version 1.0, Apr. 2020. [Online]. Available: https://www.pycaret.org.

[21] L. Breiman, "Random Forest," Mach Learn, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[22] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Belmont, CA: Wadsworth International Group, 1984.

[23] E. Fix and J. L. Hodges, Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties. USAF School of Aviation Medicine, 1951.

[24] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans Inf Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[25] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," Sci Rep, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[26] O. Rainio, J. Teuho, and R. Klén, "Author Correction: Evaluation metrics and statistical tests for machine learning," Sci Rep, vol. 14, no. 1, p. 15724, Jul. 2024, doi: 10.1038/s41598-024-66611-y.

[27] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[28] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat Mach Intell, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

EMO-Codec: An In-Depth Look at Emotion Preservation Capacity of Legacy and Neural Codec Models with Subjective and Objective Evaluations

Wenze Ren*, Yi-Cheng Lin*, Huang-Cheng Chou†, Haibin Wu*, Yi-Chiao Wu
Chi-Chun Lee†, Hung-yi Lee*, Hsin-Min Wang‡, Yu Tsao‡
* National Taiwan University, Taiwan † National Tsinghua University, Taiwan ‡ Academia Sinica, Taiwan
E-mail: {r11942166, r12942075}@ntu.edu.tw

***Abstract*—Neural codecs reduce speech data transmission latency and serve as the underlying tokenizer for speech language models (speech LMs). Preserving emotional information in these codes is essential for effective communication and contextual understanding. However, there is a lack of research on emotional loss in existing codecs. This study evaluates both neural and legacy codecs using subjective and objective methods on emotion datasets such as IEMOCAP. Our study identifies which codecs best preserve emotional information at various bitrates. We found that training a codec with both English and Chinese data had limited success in retaining emotional information in Chinese. Additionally, resynthesizing speech using these codecs degrades speech emotion recognition (SER) performance, especially for emotions such as sadness, depression, fear, and disgust. Human listening tests confirmed these findings. This study will guide the development of future speech technologies to ensure that new codecs maintain the emotional integrity of speech.**

I. Introduction

Audio codecs are initially introduced to compress speech signals into fewer bits to achieve low-latency communication. They generally include an encoder and a decoder, which respectively compress a speech signal into codes and reconstruct the codes back into the speech signal. Given their success, researchers are exploring how speech perception tasks can benefit from the advanced capabilities of large language models (LLMs), which have demonstrated remarkable performance in text modeling and surpassed human capabilities in a variety of tasks [1]. Recent advances in speech language models (speech LMs) use the codes of speech codecs as discrete tokens [2], [3]. Speech LMs are distinct in their ability to extract rich information, such as emotions, from spoken languages [4]. In addition to capturing content information, they delve into the nuances of the speaker's identity and emotion that cannot be fully captured by text alone.

Since codec models are widely used to reduce communication latency and serve as tokenizers for speech LMs, their codes should preserve signal integrity, including emotional information. For instance, when an individual communicates with a virtual assistant through voice commands, the emotional information embedded in the speech signal can provide valuable context for the assistant to deliver a more empathetic and tailored response [5]. However, speech codecs used in communication pipelines inadvertently distort or discard critical emotional cues during the compression process. As a result, the virtual assistant may struggle to accurately perceive the user's emotional state, thereby affecting the interaction. Therefore, preserving the emotional information in the speech signal is crucial for the effectiveness of speech LMs.

Many advanced neural codec models have been developed using different techniques [6]. However, existing evaluations of these codecs have mainly focused on signal-level metrics, overlooking critical paralinguistic elements such as emotion [7]. While Codec-SUPERB [8] endeavors to compare emotion preservation in speech reconstructed by neural codecs, its evaluation was limited to a single dataset in a single language. Similarly, Siegert et al. [9] studied the intelligibility of codec-compressed emotional speech but neglected to evaluate the emotional content of the speech. Previous work compared speech distorted by legacy codec compression algorithms and evaluated speech emotion recognition (SER) performance through human perception [10] and automatic methods such as Gaussian mixture model [11] or Support Vector Machine (SVM) [12]. However, these studies are limited to legacy codecs, which have poor performance compared to neural codecs. The SER models used are also less accurate than today's advanced models.

There is an urgent need for a comprehensive comparative analysis of the ability of codec models of different languages on different downstream SER systems under staged or real-world, multilingual, and multi-speaker dataset conditions. We consider various factors of the codec, such as bitrate, pretraining dataset language, and architecture. Different evaluations are conducted to comprehensively assess these factors and their impact on the accuracy of pre-trained SER systems and human emotion perception. Our goal is to provide the research community with valuable insights to guide the design of new codecs. Our study comprehensively evaluates the efficacy of 14 neural and 3 legacy codecs in preserving emotional information across 15 different SER models on 6 datasets, revealing their potential for enhancing affective computing in real-world applications. The main flow of Emo-Codec is shown in Fig. 1.

Our main contributions are as follows:

- Emo-Codec provides comprehensive performance benchmarks for 14 codec models and 3 legacy codecs in 6 emotion datasets, highlighting their ability to preserve emotional information.
- Descript Audio Codec (DAC) series [13] consistently outperforms other codecs in SER at the same bitrate. Additionally, AcademiCodec [14] and SpeechTokenizer [15] show considerable performance in low-bitrate scenarios.
- Training a codec using Chinese and English data resulted in limited improvements in preserving Chinese emotional information compared to training a codec using only English data.
- Of all the emotions investigated, some negatively valenced emotions, such as sadness, depression, fear, and disgust, showed higher performance declines than others.

II. Methodology

This section first provides an overview of our rationale for conducting a large-scale evaluation and then explains each part in detail.

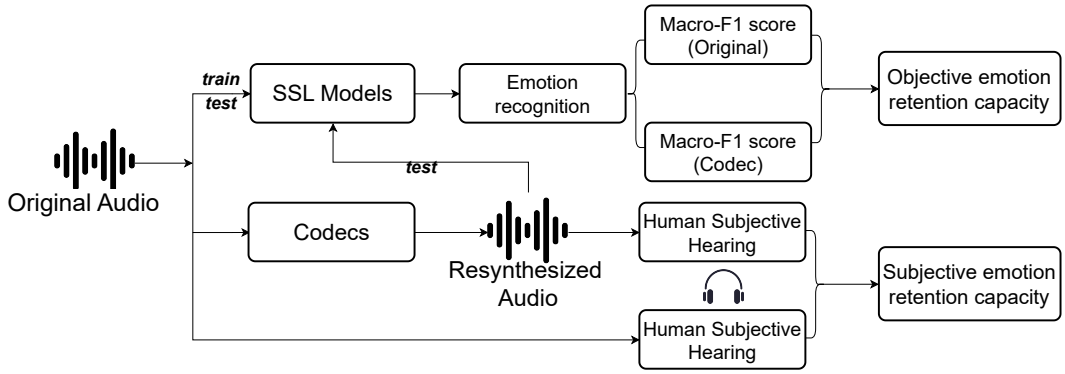


Fig. 1: The pipeline of Emo-Codec. We only use the original training audio from emotion datasets to train the emotion recognition model and fine-tune the SSL model at the same time. Then, we evaluate the testing set's original audio and audio resynthesized using various codecs on SER models. We calculate the F_1 score difference of SER to obtain the objective emotion retention capacity. We also conduct human subjective listening tests using original audio and resynthesized audio to obtain the subjective emotion retention capacity.

A. Rationale for the evaluation

We evaluate both neural and legacy codecs to ensure a comprehensive analysis. The legacy codecs serve as established benchmarks, enabling us to measure the progress achieved with neural codecs. For the neural codecs, we select models based on their functionality and design principles. These codecs incorporate a range of innovative architectural designs and sophisticated methods to ensure optimal performance and versatility in processing various types of speech data. Specifically, we target models explicitly tailored for speech LM tokenization. Additionally, we consider models trained on mixed Chinese and English data (for example Funcodec) to ensure comprehensive language coverage and take into account the nuances and emotional variances inherent in both languages. We evaluate all codec models at similar bitrates to ensure a fair comparison.

For the SER models, we use different representations of self-supervised learning (SSL) models (shown in Table I) to train the SER model because the SSL paradigm has achieved state-of-the-art performance in SER tasks [16], [17]. We employ a variety of emotion datasets to increase diversity in languages, dataset collection methods (real world, improvised act, scripted act), and speakers.

B. Codec models

We carefully selected seven cutting-edge, high-fidelity neural codec models for comparative analysis, as shown in Table II. **Encodec** [18] serves as the baseline. We used the 2, 4, 8, 16, and 32 layer settings to compare with other codecs at similar bitrates. Building on Encodec, **AudioDec** [19] introduces a novel approach that employs group convolution to accelerate and streamline operations. **AcademiCodec** [14] uses group-residual vector quantization to reduce codebook usage while maintaining comparable performance. We used a universal version of this model. **FunCodec** [20] proposes a frequency domain codec that achieves comparable performance with lower computational and parameter complexity. Additionally,

TABLE I: SSL models used in this study and their number of parameters (Parm.

Model	Parm. (M)	Model	Parm. (M)
Wav2Vec 2.0 XLS-R-1B	965	VQ-Wav2Vec	34
WavLM large	317	Wav2vec-large-960h	33
Wav2Vec 2.0 Large Robust	317	TERA	21
HuBERT Large	317	NPC	19
Wav2Vec2 large 960h	317	VQ-APC	5
Data2vec large 960h	313	APC	4
DeCoAR 2.0	90	Modified CPC	2
Mockingjay	85		

TABLE II: Codecs evaluated in this study and their transmission bitrates. The column **ID** shows the identification number of all codecs, **kbps** represents the bitrate in kilobits Per Second, and **sr** shows the sampling rate in kHz.

ID	Codec	Codec Configuration	kbps	sr
U	AudioDec	symAD_libritts_24000_hop300	6.4	24
C	AcademiCodec	large universal	2	16
S	SpeechTokenizer	hubert_avg	4	16
D1	DAC	DAC_16k	6	16
D2		DAC_24k	24	24
E1	Encodec	Encodec_24k	1.5	24
E2			3	24
E3			6	24
E4			12	24
E5			24	24
F1	Funcodec	en_libritts_16k_nq32ds320	16	16
F2		en_libritts_16k_nq32ds640	8	16
F3		zh_en_16k_nq32ds320	16	16
F4		zh_en_16k_nq32ds640	8	16
N	Soundstream	Soundstream	6	16
M1	MP3	-	6	-
M2		-	24	-
M3		-	192	-
O1	Opus	-	6	-
O2		-	24	-
A1	AAC	-	6	-
A2		-	24	-
A3		-	192	-

SpeechTokenizer [15] introduces a unified speech tokenizer tailored for speech LMs, integrating HuBERT units as semantic teachers in the first layer of RVQ. **Descript Audio Codec (DAC)** [13] leverages advanced Snake activation from BigVGAN [21] and utilizes a novel complex STFT discriminator at multiple time scales to further enhance audio fidelity. We used the 16k and 24k sample rate models. Lastly, **SoundStream**[22] uses RVQ in its encoder to represent audio signals more efficiently and compactly, resulting in higher quality reconstruction at lower bitrates[1].

To more comprehensively compare the ability of codecs to retain emotional traits, we also selected three legacy codecs for comparison: **MP3** [23], which is widely used for its efficient compression and good sound quality at different bitrates; **Opus** [24], which is known for its adaptability to a wide range of audio and low latency; and **AAC** [25], which provides high fidelity and is commonly used for streaming and

[1]We used the implementation from: https://github.com/kaiidams/soundstream-pytorch

TABLE III: Datasets used in this study and their setting. **Anno** represents the annotation process used; **P** and **S** represent primary and secondary labeling scenarios, respectively. **Speaker** represents the number of speakers in the dataset.

Dataset	Language	Setting	Anno	Speaker
CREAM-D	English	Acted	P	91
IEMOCAP	English	Acted	S	10
IMPROV	English	Acted	S	12
MSP-PODCAST	English	Real-world	P	2172+Unknown
BIIC-PODCAST	Chinese	Real-world	P	Unknown
NNIME	Chinese	Acted	S	43

broadcasting. These legacy codecs help us benchmark neural codecs against established standards.

C. Speech Emotion Recognition Datasets

To evaluate the codecs, we used six public datasets partitioned by EMO-SUPERB [26]. These datasets are classified according to their source (acted or real-world) and language (Chinese or English), as shown in Table III. The datasets have two annotation scenarios: *Primary* (P) requires each annotator to select only one emotion during annotation, while *Secondary* (S) allows annotators to select multiple emotions for a clip.

III. Experimental Setup

A. Speech Emotion Recognition Models

In contrast to previous SER methods that perform single-target prediction , we adopt a distribution-like representation to model the multi-dimensional complexity of emotions, as suggested by [27]. Furthermore, to enhance the performance of the SER model, we incorporate label smoothing [28] into the emotion distribution, effectively regularizing the classifier layer by setting the smoothing parameter to 0.05.

We adopted the SER model architecture from the S3PRL [29] toolkit, which is based on a CNN-Self Attention network consisting of three Conv1d layers, a self-attention pooling layer, and two linear layers. We used a fixed learning rate of 10^{-4} and the AdamW optimizer [30] to train the SER model until the loss on the development set stopped decreasing for 5 epochs. We used class-balanced cross-entropy loss [31] to mitigate the impact of imbalanced labels in emotion.

B. Evaluation Metrics

We used the macro-F_1 score [32] as the evaluation metric for the SER task. Since the output of the SER model is a probability distribution, the emotion prediction is successful if the corresponding probability of the ground-truth label exceeds the threshold $\frac{1}{n}$ for the n-class SER model, following [27], [33].

C. Human Subjective Evaluation

To determine whether humans perceive emotions differently in synthesized versus original speech, we conducted human subjective listening tests. We randomly selected 45 audio samples from the well-known emotion dataset IEMOCAP. We used three codecs, namely Encodec, DAC, and the legacy codec Opus, to resynthesize the audio at bitrates of 6kbps and 24kbps.

We hired evaluators from the Prolific platform to evaluate the resynthesized audio. Each audio sample was evaluated by 5 male and 5 female evaluators. We required evaluators to be from the US and must have a past assignment acceptance rate above 90%. Each evaluator was asked to answer three questions for each audio sample: (1) select one or more emotions of the speaker from the pre-defined emotions, the same process as in IEMOCAP; (2) rate speech quality based on everyday speech communication on a scale of 1 to 5; (3) rate speech quality according to the degree of distortion on a scale of 1 to 5.

Annotators of the IEMOCAP dataset originally used both video and audio to label multiple emotions. However, our study only focused on audio. This difference may affect emotion perception. Therefore, we also asked these evaluators to re-annotate the emotions of the original audio to make a fair comparison.

After collecting annotations and ratings, we report average values for speech quality and distortion factors. In terms of emotion, we calculate the distributional label for each audio sample, and binarize the label using a threshold in the same way as introduced in Sec. III-B to measure the macro-F_1 score. We take the new-collected labels elicited by audios as the ground truth and calculate the accuracy, and define this as *subjective emotion capacity*.

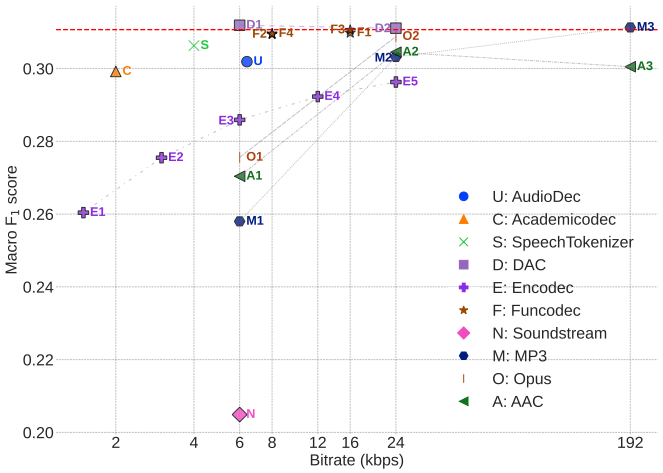


Fig. 2: SER performance (macro-F_1) for different codecs on the IEMOCAP dataset. The red dashed line represents the SER performance of the original audio (the topline).

IV. Results and Discussion

A. Impact of Bitrate on Machine Emotion Recognition

Fig. 2 shows the emotion preservation capabilities of individual codecs at different bitrates on the IEMOCAP dataset. Across most codecs, whether neural or legacy, there is a consistent trend: emotion preservation increases as the codec bitrate increases. It highlights the direct impact of bitrate on the amount of voice information that can be transmitted and retained. A higher bitrate is beneficial to retaining more detailed emotional information. While lower bitrates often compromise the retention of emotional information, some codecs exhibit excellent emotion retention even at extremely low bitrates, such as **SpeechTokenizer** and **Academicodec**. This indicates that these codecs are particularly effective at preserving the integrity of emotional information despite the limitations of lower bitrates. Of all codecs, Soundstream is the worst at retaining emotional information, showing the worst emotion recognition performance.

Legacy codecs (**MP3**, **Opus**, and **AAC**) generally demonstrate effective retention of emotional information at higher bitrates, e.g., **Opus**_24kbps and **MP3**_192kbps perform well, approaching the topline performance. However, all three legacy codecs perform poorly at low bitrates. Of the three legacy codecs, **AAC** is the worst. In comparison, neural codec **DAC** consistently outperforms all neural and legacy codecs at the same bitrate of 24kbps, highlighting DAC's advanced design and effectiveness in retaining emotional information. The superior performance of DAC may be attributed to two tricks: the snake activation function and balanced data sampling during training. DAC also performs well at 6kbps. In addition, **Funcodec**'s

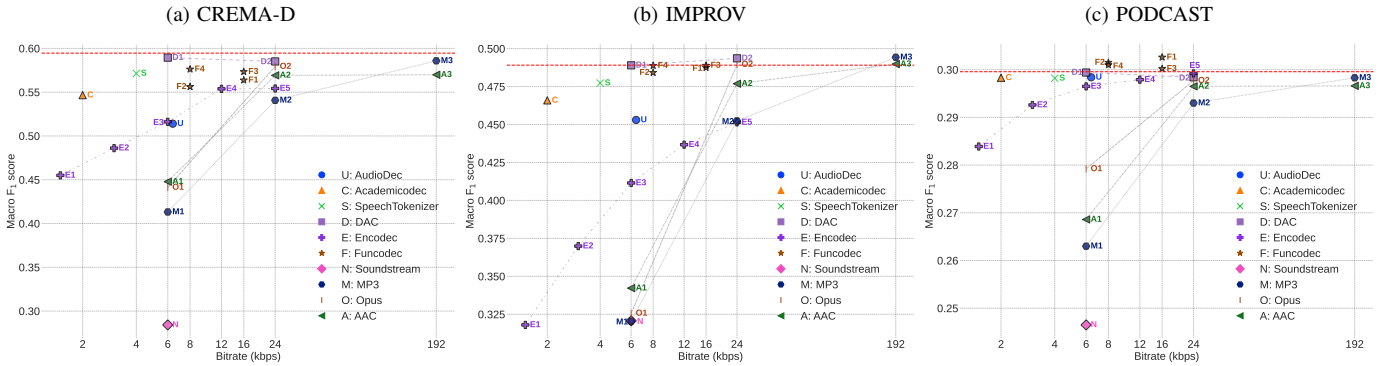


Fig. 3: SER performance (macro-F_1) for different codecs on the (a) CREMA-D, (b) IMPROV, and (c) PODCAST datasets. The red dashed line represents the SER performance of the original audio (the topline).

performance under different bitrates is also good and stable. Our results demonstrate that neural codecs are more suitable for retaining the integrity of emotional information in bitrate-constrained environments.

B. Trend Across English Datasets

Next, we evaluate the emotional information retention of codecs on three additional English datasets (CREMA-D, IMPROV, and PODCAST). Obviously, the same trends as in Fig. 2 can also be observed in Fig. 3. This analysis extends the analysis on the IEMOCAP dataset in Sec. IV-A, highlighting consistent trends in the emotional information retention capabilities of codecs.

C. Variability Across Chinese Datasets

Fig. 4 and Fig. **??** in the Appendix show the emotion preservation capabilities of various codecs at different bitrates on the Chinese datasets BIIC-PODCAST and NNIME. Consistent with the trend in the English datasets, higher bitrate codecs generally retain emotional information better in the Chinese datasets; the **DAC** still provides superior performance compared to other codecs of the same bitrate; neural codecs **SpeechTokenizer** and **Academicodec** also maintain good performance at lower bitrates; and legacy codecs are always weaker than neural codecs.

While the overall trend is consistent, there are some differences when comparing the performance of certain codecs on the English and Chinese datasets in detail. In the case of **Encodec**, a significant increase in emotional information retention with increasing bitrate can be observed in the English datasets. However, this trend slows down in the Chinese dataset NNIME and even decreases in the BIIC-PODCAST dataset. Compared with English, emotional information in Chinese has different phonetic and tonal characteristics, and the neural codec may not have been specifically optimized for these differences, causing the retention of emotional information to slowly increase or even decrease as the bitrate increases.

Furthermore, specific codecs are trained on English-only data, such as **funcodec_en_libritts** (F1 and F2), while conversely **funcodec_zh_en** (F3 and F4) is trained on mixed English and Chinese data. From Fig. **??**, we can see that on the Chinese dataset BIIC-PODCAST, the codecs F3 and F4 trained on mixed English and Chinese data are better at preserving Chinese emotional information than their counterpart codecs F1 and F2 (F4 vs F2, F3 vs F1) trained on English-only data. However, F3 did not beat F1 on the Chinese dataset NNIME. This observation shows the importance of research on codec training data settings in order to accommodate universal representations of different emotional information across languages.

D. Specific Emotion Losses

As can be seen from Table IV, speech resynthesized through a codec may lose some important emotional information, especially for emotions that are challenging for SER models, such as disgust or

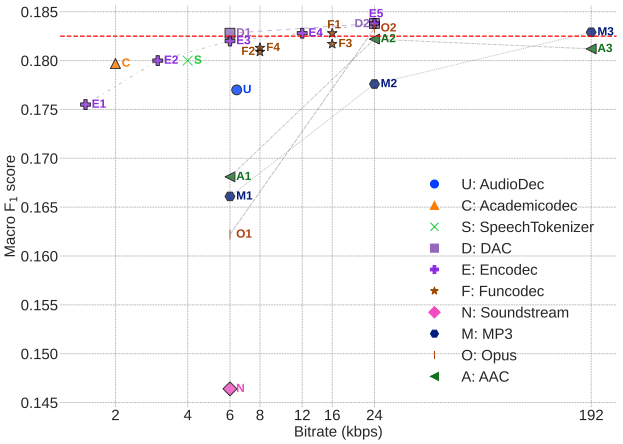


Fig. 4: SER performance (macro-F_1) for different codecs on the NNIME Chinese datasets. The red dashed line represents the SER performance of the original audio (the topline).

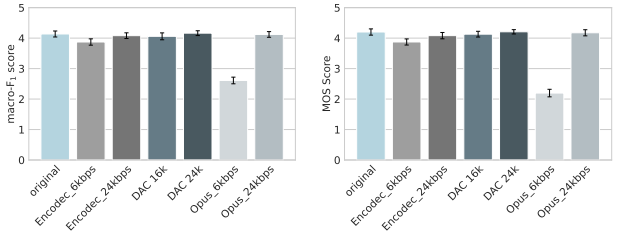
fear. For instance, the F_1 scores for the fear emotion category on the IMPROV and IEMOCAP datasets are only 1.86% and 9.58%, respectively. The F_1 scores for fear in the resynthesized speech dropped by 57.22% and 18.45%, respectively, Significantly higher than most other emotions. This shows that when the original audio has low emotion recognition performance, any additional distortion from the codec will have a pronounced impact. For the SER model, fear is inherently a difficult emotion to identify. The substantial performance drop highlights the inability of the codec to retain the emotional information of fear, which poses a greater challenge to the SER model. In addition to fear, other emotions such as depression and sadness also experience SER performance degradation when speech is resynthesized through these codecs. This highlights the codecs' shortcomings in retaining complex and subtle emotional information. While codecs are proficient in compressing and resynthesizing speech, they struggle to preserve the nuanced emotional cues needed for accurate emotion recognition, especially for more complex emotions. This demonstrates the need for further optimization and enhancement of codec models to better preserve the emotional integrity of resynthesized speech.

E. Human Subjective Listening Tests

This section provides the subjective evaluation results. The original audio samples are randomly selected from the IEMOCAP dataset. Fig. 5 shows the Mean Opinion Score (MOS) of the original audio and the codec-resynthesized audio. The original audio has the highest MOS. The 24k bitrate codecs (Encodec_24kbps and DAC 24k) result in a slight drop in MOS, but still maintain a score close to that of

TABLE IV: SER performance (macro-F_1) degradation of resynthesized speech relative to the original speech. Positive values mean improvement in SER performance. Emotions considered include depression (P), frustration (T), anger (A) sadness (S), disgust (D), excitement (E), fear (F), neutral (N), surprise (U), and happiness (H). The **Ori** row represents the performance of the original audio. The bold text indicates the emotion with the highest, the second-highest, and the third-highest level of average degradation.

	IMPROVS										IEMOCAP								
ID	P	T	A	S	D	E	F	N	U	H	T	A	S	D	E	F	N	S	H
Ori.	38.16	53.34	47.91	57.33	14.35	57.07	1.86	84.04	29.12	69.21	69.27	59.70	61.36	00.89	55.10	9.58	67.54	25.19	49.28
U	-50.49	-11.88	-8.68	-19.15	-35.13	2.05	-70.65	-0.51	-6.90	-3.37	-1.93	-3.06	-5.62	-30.80	-4.20	-7.76	-2.67	-12.10	-6.99
C	-42.27	-8.22	-5.96	-13.03	-25.01	1.37	-73.60	-0.48	4.85	-2.45	-1.81	-4.79	-8.90	-25.63	-5.03	-20.83	-6.77	-8.13	-7.44
S	-23.12	-5.39	-5.61	-5.93	-24.68	0.36	-48.77	-0.31	-3.41	-1.88	-1.37	-1.67	-4.07	-3.81	-2.62	-11.04	-1.78	-7.79	-4.36
D1	-7.58	-1.75	-1.37	-2.06	-3.44	0.64	-23.91	-0.16	-1.76	-0.47	-0.31	-0.19	-0.72	-8.84	-0.47	-2.97	-0.40	0.22	-0.72
D2	-1.84	-0.80	-1.21	-0.37	-9.82	0.65	-21.38	-0.08	1.28	-0.14	-0.02	-0.40	-0.40	-27.97	-0.31	-1.81	-0.13	0.16	-0.30
E1	-88.21	-42.40	-22.08	-68.06	-70.39	-7.00	-94.65	-16.25	-28.31	-6.26	-4.91	-10.47	-37.24	-19.09	-12.82	-41.13	-28.06	-34.14	-18.71
E5	-43.75	-6.24	-2.94	-19.94	-39.15	1.84	-57.34	-1.22	-3.94	-3.39	-1.43	-3.90	-12.28	-10.62	-5.65	-18.05	-6.68	-10.58	-5.62
F1	-12.57	-0.54	-1.63	-3.10	-10.31	1.15	-26.03	-0.07	-0.07	-0.82	-0.72	-0.47	-1.12	-5.97	-1.73	-4.23	-0.55	-5.07	-1.48
F3	-11.86	-1.29	-2.03	-3.25	-11.97	1.08	-30.95	-0.10	3.63	-0.61	-1.02	-0.98	-1.34	-24.16	-1.78	-9.88	-1.08	-1.00	-1.70
M1	-76.03	-62.86	-48.19	-57.12	-88.15	-14.26	-94.21	-15.82	-4.42	-8.45	-16.58	-20.97	-23.86	1.29	-17.02	-54.93	-14.26	-33.99	-28.78
M3	-8.07	-5.33	-7.00	-2.82	-1.36	-11.41	-40.98	0.49	-9.42	-2.08	-0.21	-0.10	-3.99	-28.14	-1.54	0.28	-0.63	-5.75	-5.41
O1	-87.75	-44.84	-27.26	-68.92	-81.77	-12.65	-92.32	-15.93	-35.08	-7.74	-10.67	-14.65	-19.91	-19.66	-14.21	-37.22	-7.63	-33.98	-21.71
O2	-23.29	-5.32	-6.04	-6.27	-5.44	-6.98	-49.64	0.51	-11.87	-2.24	-0.28	-1.18	-4.34	-18.19	-1.82	-6.20	-1.46	-7.46	-7.00
A1	-57.40	-61.04	-52.67	-38.45	-94.45	-11.11	-95.81	-14.95	-15.95	-6.56	-10.71	-15.80	-19.79	-12.93	-17.42	-49.14	-11.81	-30.28	-25.50
A3	-15.53	-6.24	-8.17	-5.52	-8.68	-10.30	-38.00	0.21	-19.70	-1.83	-1.29	-3.92	-10.34	-46.57	-3.02	-11.84	-2.71	-17.44	-9.60
Avg	**-36.65**	-17.61	-13.39	-20.93	**-33.98**	-4.30	**-57.22**	-4.31	-8.74	-3.22	-3.55	-5.50	-10.26	**-18.74**	-5.98	**-18.45**	-5.77	**-13.82**	-9.69



(a) Communication (b) Distortion

Fig. 5: Subjective quality assessment of speech (MOS) based on (a) everyday speech communication and (b) distortion, following ITUT P835 [34]. Higher scores indicate higher quality.

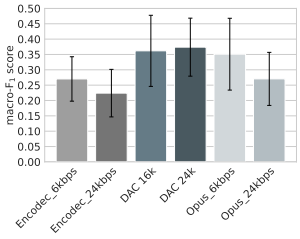


Fig. 6: Subjective emotion retention capacity (macro-F_1) for six codecs.

the original audio. In contrast, the MOS for the 6k bitrate codecs drops significantly, except for DAC 16k (as shown in Table II, its bitrate is 6kbps), which maintains a high MOS. Opus_6kbps has the lowest MOS in this evaluation. These results align with the objective evaluation results in the previous sections.

Fig. 6 shows the human SER performance on the codec-resynthesized audio. Again, we can see that DAC 24k and DAC 16k perform well, and **DAC 24k** achieves the best performance. Encodec_24kbps is worse than its lower bitrate counterpart Encodec_6kbps, and Opus_24kbps is also worse than its counterpart Opus_6kbps. Surprisingly, Opus_6kbps performs well in this human SER evaluation. The reason still needs further study. The discrepancy between subjective human listening tests and the objective assessments presented in Fig. 2 is emphasized, and this inconsistency suggests that objective metrics may not fully capture elements of human perception of emotional content in language. Further research into the perceptual aspects of audio quality and emotion recognition is recommended to address this issue.

V. Discussion and Limitation

Existing neural codecs are mainly trained on English and Chinese datasets. However, there are thousands of spoken languages around the world. Whether existing codecs can be generalized to other languages and how to train codecs that preserve multilingual emotion information are unsolved issues.

While subjective evaluations provide valuable insights, current evaluators are limited to a small number of people. Future research should aim to include a more diverse group of evaluators to account for different emotional perceptions across age groups, cultures, and backgrounds.

Due to limited scope, this work did not consider how the interactivity and context of a conversation might affect the effectiveness of emotion preservation. Future research should explore how codecs perform in interactive dialogue systems, where context and conversational history play crucial roles in emotion recognition and response generation. We will also implement more comprehensive subjective testing protocols and refine objective evaluation metrics to better align with the framework for assessing human perceptual differences.

VI. Conclusion

This work provides insights into the emotional information preservation capabilities of neural codecs. We provide different perspectives to evaluate codec performance. We confirm that codecs with higher bitrate preserve more emotional information. DAC performs the best among all compared neural codecs, while AcademiCodec and SpeechTokenizer can preserve a considerable amount of emotional information at limited bitrates. Legacy codecs perform worse than neural codecs at low bitrates. Furthermore, a codec trained with Chinese and English data may have a limited improvement in Chinese emotion information preservation capacity compared to a codec trained with only English data. We find that the resynthesis of speech by neural codecs reduces emotional information such as sadness, depression, fear, and disgust. Our future work will train neural codec models to preserve emotional information across diverse language usage contexts.

ACKNOWLEDGMENT

NSTC supported this research under Grants 112-2634-F-002-005. We thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] H. Touvron *et al.*, *Llama 2: Open Foundation and Fine-Tuned Chat Models*, 2023. arXiv: 2307.09288 [cs.CL].

[2] C. Wang *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[3] D. Yang *et al.*, "Uniaudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023.

[4] C.-y. Huang *et al.*, *Dynamic-SUPERB: Towards A Dynamic, Collaborative, and Comprehensive Instruction-Tuning Benchmark for Speech*, 2023. arXiv: 2309.09510 [eess.AS].

[5] S. Guha and R. Iqbal, "DESCo: Detecting Emotions from Smart Commands," in *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2022.

[6] H. Wu *et al.*, *Towards audio language modeling – an overview*, 2024. arXiv: 2402.13236.

[7] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "Warp-Q: Quality Prediction for Generative Neural Speech Codecs," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[8] H. Wu *et al.*, *Codec-SUPERB: An In-Depth Analysis of Sound Codec Models*, 2024. arXiv: 2402.13071 [eess.AS].

[9] I. Siegert *et al.*, "Emotion Intelligibility within Codec-Compressed and Reduced Bandwidth Speech," in *Speech Communication; 12. ITG Symposium*, 2016.

[10] O. Niebuhr and I. Siegert, ""High on Emotion "? How Audio Codecs Interfere With the Perceived Charisma and Emotional States of Men and Women," in *33. Konferenz Elektronische Sprachsignalverarbeitung, ESSV 2022*, 2022.

[11] A. Albahri and M. Lech, "Effects of band reduction and coding on speech emotion recognition," in *2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2016.

[12] A. F. Lotz *et al.*, "Audio Compression and its Impact on Emotion Recognition in Affective Computing," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, 2017, pp. 1–8.

[13] R. Kumar *et al.*, "High-fidelity audio compression with improved RVQGAN," *Advances in Neural Information Processing Systems*, 2024.

[14] D. Yang *et al.*, "Hifi-codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.

[15] X. Zhang *et al.*, "Speechtokenizer: Unified speech tokenizer for speech large language models," *arXiv preprint arXiv:2308.16692*, 2023.

[16] J. Wagner *et al.*, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2023.

[17] E. Morais *et al.*, "Speech Emotion Recognition Using Self-Supervised Features," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[18] A. Défossez *et al.*, "High Fidelity Neural Audio Compression," *arXiv preprint arXiv:2210.13438*, 2022.

[19] Y.-C. Wu *et al.*, "Audiodec: An Open-Source Streaming High-Fidelity Neural Audio Codec," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[20] Z. Du *et al.*, "FunCodec: A Fundamental, Reproducible and Integrable Open-source Toolkit for Neural Speech Codec," *arXiv preprint arXiv:2309.07405*, 2023.

[21] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.

[22] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, *Soundstream: An end-to-end neural audio codec*, 2021. arXiv: 2107.03312 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2107.03312.

[23] K. Brandenburg and G. Stoll, "Iso-mpeg-1 audio: A generic standard for coding of high-: Quality digital audio," 1994. [Online]. Available: https://api.semanticscholar.org/CorpusID:58871544.

[24] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, *High-quality, low-delay music coding in the opus codec*, 2016. arXiv: 1602.04845 [cs.MM]. [Online]. Available: https://arxiv.org/abs/1602.04845.

[25] M. Bosi, K. Brandenburg, S. R. Quackenbush, *et al.*, "Iso/iec mpeg-2 advanced audio coding," *Journal of The Audio Engineering Society*, vol. 45, pp. 789–814, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:60516290.

[26] H. Wu *et al.*, "EMO-SUPERB: An In-depth Look at Speech Emotion Recognition," *arXiv preprint arXiv:2402.13018*, 2024.

[27] H.-C. Chou *et al.*, "Minority Views Matter: Evaluating Speech Emotion Classifiers with Human Subjective Annotations by an All-Inclusive Aggregation Rule," *IEEE Transactions on Affective Computing*, pp. 1–15, 2024. DOI: 10.1109/TAFFC.2024.3411290.

[28] C. Szegedy *et al.*, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] S.-w. Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021.

[30] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[31] Y. Cui *et al.*, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] J. Opitz and S. Burst, "Macro f1 and macro f1," *arXiv preprint arXiv:1911.03347*, 2019.

[33] P. Riera *et al.*, "No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems," in *Proc. SMM19, Workshop on Speech, Music and Mind 2019*, 2019.

[34] ITU, "Itu-t p.835, subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *International Telecommunication Union*, 2003.

Pressure Matching Using Data-Driven Estimation for Sound Fields and Transfer Functions

Koki Horikoshi*, Gen Sato*, Izumi Tsunokuni* and Yusuke Ikeda*
* Tokyo Denki University, JAPAN
E-mail: {24fmi25, 24udc03, 21udc02}@ms.dendai.ac.jp, yusuke.ikeda@mail.dendai.ac.jp

***Abstract*—Sound field reproduction is a technology that uses multiple loudspeakers to reproduce an immersive sound field. The pressure matching method, a well-known sound field reproduction method, is challenging to implement because it requires a large number of measurements at the control points to obtain the transfer functions of loudspeakers and the desired sound pressures. In this study, we propose a method to derive the driving functions of loudspeakers based on pressure matching using data-driven estimation of the sound field. The proposed method estimates the desired sound pressures and the transfer functions from the limited number of microphones. Simulation experiments demonstrate that the proposed method significantly improves reproduction accuracy and reduces the number of microphones by approximately 70 %.**

I. Introduction

The sound field reproduction using a loudspeaker array physically reproduces the desired sound field within the array. This technology is expected to be applied in various fields, including immersive virtual reality for movies and games, as well as simulation for acoustic design, such as classrooms and concert halls.

There are two main methods for sound field reproduction [1]: analytical and numerical. Well-known analytical methods for sound field reproduction include WFS [2], [3] and HOA [4]. The numerical method is based on optimization and aims to minimize the error between the reproduced sound field and the desired sound field. The typical approaches are the Pressure Matching (PM) method and the Mode Matching (MM) method [5].

The PM method determines the driving functions of the loudspeakers by minimizing the error between the reproduced sound pressures and the desired sound pressures at control points, where the reproduction region is discretized [6]–[9]. The PM method generally assumes that the transfer functions from the loudspeakers to the control points are known. However, obtaining these transfer functions requires many measurements, as the control points must be placed at a sufficiently high density relative to the wavelength. This makes the PM method difficult to implement. To address this, a method that spatially interpolates and estimates the transfer functions (or room impulse responses: RIRs) from a limited amount of measured data is effective.

In recent years, many methods for estimating RIRs from a limited number of microphones have been proposed based on physical models and deep learning. For example, methods for estimating the early-part RIRs based on physical models and compressed sensing have been studied [10]–[14]. Based on deep learning with the wave equation, many methods for sound field reconstruction have also been proposed [15]–[18].

Furthermore, in [19], [20], a PM method based on deep learning was proposed, which derives the driving functions directly from recorded signals while performing spatial interpolation of the desired sound field using a convolutional neural network. However, most sound field reproduction methods that use deep learning for spatial interpolation assume that the transfer functions or RIRs from the loudspeakers are known. Moreover, since the loudspeakers are fixed and regularly arranged, the model must be retrained with new data when the loudspeakers are freely placed or rearranged.

In this study, we propose a PM method that use data-driven estimation of sound field and transfer functions from the loudspeakers. The proposed method can obtain the transfer functions of the loudspeakers and the desired sound field at finer control points from a smaller number of microphone signals using the same trained model. By controlling the sound pressures based on PM at these finer control points, a more accurate reproduction of the sound field can be achieved. Moreover, the proposed method can be adapted to any configuration of loudspeaker arrays, including irregularly arranged loudspeakers.

The rest of the paper is organized as follows. Section II details the proposed method, the formulation of the PM based on deep learning, and the network architecture based on the U–Net. Section III shows the simulation experiments to evaluate reproduction accuracy. Finally, we concluded in Sect. IV.

II. Proposed Method

A. PM with sound field interpolation based on deep learning

The PM method controls the sound pressures on the control points that discretize the reproduction region. The desired sound pressures at N control points can be reproduced by the loudspeakers as follows,

$$\mathbf{p}_{\mathrm{cp}} = \mathbf{G}_{\mathrm{cp}}\mathbf{d}, \tag{1}$$

where $\mathbf{p}_{\mathrm{cp}}(\in \mathbb{C}^{N\times 1})$ is the desired sound pressures, $\mathbf{G}_{\mathrm{cp}}(\in \mathbb{C}^{N\times L})$ is the transfer function matrix from L loudspeakers to all control points. $\mathbf{d}(\in \mathbb{C}^{L\times 1})$ is the driving functions. In the PM method, because of spatial Nyquist frequency, the density of the control points depends on the wavelength. Thus, when reproducing a sound field at higher frequencies, it is required to place the control points with higher densities.

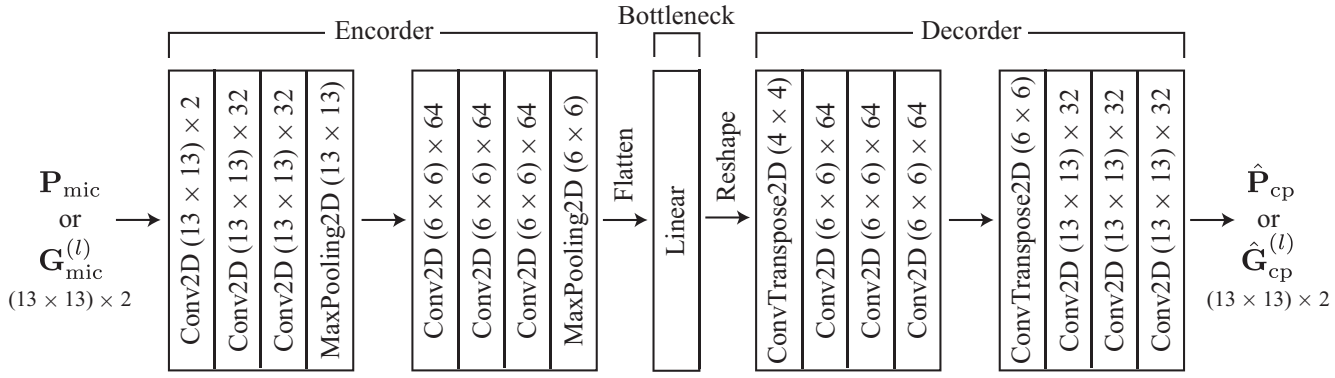


Fig. 1. Network Architecture for estimating the sound pressures or transfer functions at the control points. $\mathbf{P}_{\mathrm{mic}}$ and $\mathbf{G}_{\mathrm{mic}}$ are the sound pressures and the transfer functions at microphones, respectively. $\hat{\mathbf{P}}_{\mathrm{mic}}$ and $\hat{\mathbf{G}}_{\mathrm{mic}}$ are the estimated sound pressures and the estimated transfer functions at control points, respectively.

To reduce the number of measurement points, the proposed method estimates the desired sound field and the loudspeakers' sound fields at N control points from $K(< N)$ measurement points, allowing for higher resolution control of the sound field. The input $\mathbf{P}_{\mathrm{mic}}(\in \mathbb{C}^{M\times M})$ to the deep learning model $\mathcal{U}$, which will be detailed in the following section, is a matrix representing the desired sound field at the grid control points ($N = M \times M$). However, sound pressure values are assigned to the matrix only at the control points where the microphones are located, with zeros at all other locations. Subsequently, the desired sound pressures $\hat{\mathbf{P}}_{\mathrm{cp}}(\in \mathbb{C}^{M\times M})$ at control points are estimated as follows.

$$\hat{\mathbf{P}}_{\mathrm{cp}} = \mathcal{U}(\mathbf{P}_{\mathrm{mic}}). \tag{2}$$

Similarly, consider the transfer function matrix $\mathbf{G}_{\mathrm{mic}}^{(l)}(\in \mathbb{C}^{M\times M})$ from the l-th loudspeaker to the control points, and assign transfer function values only to the positions of the K microphones and zero to all others. The matrix $\mathbf{G}_{\mathrm{mic}}^{(l)}$ is inputted to the same model $\mathcal{U}$ to estimate the transfer functions $\hat{\mathbf{G}}_{\mathrm{cp}}^{(l)}(\in \mathbb{C}^{M\times M})$ as follows.

$$\hat{\mathbf{G}}_{\mathrm{cp}}^{(l)} = \mathcal{U}(\mathbf{G}_{\mathrm{mic}}^{(l)}). \tag{3}$$

That is, in this study, a single model is used to estimate both the desired sound field and the transfer functions from each loudspeaker.

For the deep learning model, the loss function is defined using the mean squared error (MSE) as follows.

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{N}||\mathbf{P}_{\mathrm{cp}} - \hat{\mathbf{P}}_{\mathrm{cp}}||_2^2, \tag{4}$$

where $\mathbf{P}_{\mathrm{cp}}$ is ground truth. Similarly, the loss function for the transfer functions is obtained.

From the estimated sound pressures $\hat{\mathbf{p}}_{\mathrm{cp}}(\in \mathbb{C}^{N\times 1})$, $\hat{\mathbf{G}}_{\mathrm{cp}}(\in \mathbb{C}^{N\times L})$, the driving functions $\hat{\mathbf{d}}_{\mathrm{dlpm}}(\in \mathbb{C}^{L\times 1})$ can be obtained using the regularized least squares method (RLS) as follows.

$$\hat{\mathbf{d}}_{\mathrm{dlpm}} = (\hat{\mathbf{G}}_{\mathrm{cp}}^{\mathsf{H}}\hat{\mathbf{G}}_{\mathrm{cp}} + \lambda\mathbf{I})^{-1}\hat{\mathbf{G}}_{\mathrm{cp}}^{\mathsf{H}}\hat{\mathbf{p}}_{\mathrm{cp}}, \tag{5}$$

where λ denotes the regularization parameter, $[\cdot]^{\mathsf{H}}$ denotes the Hermitian transpose, and $\mathbf{I}$ denotes the identity matrix.

Finally, by using the driving functions $\hat{\mathbf{d}}_{\mathrm{dlpm}}$, the reproduced sound pressures $\hat{\mathbf{p}}_{\mathrm{cp}}^{\mathrm{syn}}(\in \mathbb{C}^{N\times 1})$ for all control points can be represented as follows.

$$\hat{\mathbf{p}}_{\mathrm{cp}}^{\mathrm{syn}} = \mathbf{G}_{\mathrm{cp}}\hat{\mathbf{d}}_{\mathrm{dlpm}}, \tag{6}$$

where $\mathbf{G}_{\mathrm{cp}}$ is the transfer function matrix between the loudspeakers and the control points.

B. Network Architecture

Fig. 1 shows the network architecture in the proposed method. An encoder-decoder type of network architecture was adopted [21]. The inputs to the network is a two-channel real number vectors that are the real and imaginary parts of the complex sound pressure matrix $\mathbf{P}_{\mathrm{mic}}$ or $\mathbf{G}_{\mathrm{mic}}^{(l)}$.

The encoder extracts features using the convolutional and maxpool layers. The decoder performs upsampling using the convolutional and transposed convolution layers. The bottleneck layer uses the flatten and linear layers to learn important latent features of the data by compressing and reconstructing the encoder data. The encoder consists of six convolution layers with 32 filters in the first three-layers and 64 filters in the second three-layers. In the subsequent flattening layer, the encoder output is transformed into a one-dimensional vector, which is then passed to the linear layer. The decoder consists of six convolutional layers with 64 filters in the first three-layers and 32 filters in the second three-layers.

In all layers, kernel size was 3×3, a stride was one $\times$ one and padding was set to one. However, only the first maxpooling layer and the second ConvTranspose layer have a stride of 2×2. In addition, zero-padding was applied for the maxpool and ConvTranspose layers. Furthermore, PReLU [22] is used as an activation function in layers excluding the final Conv2D layer, the maxpooling layer, and the ConvTranspose layer.

Fig. 2 shows the arrangement of microphones and control points. As shown in the figure, the number of control points was doubled in the x-axis and y-axis directions, respectively, relative to the arrangement of microphones on the grid. Thus, the number of microphones was $K = 49$ (7×7), while the number of control points was $N = 169$ (13×13).

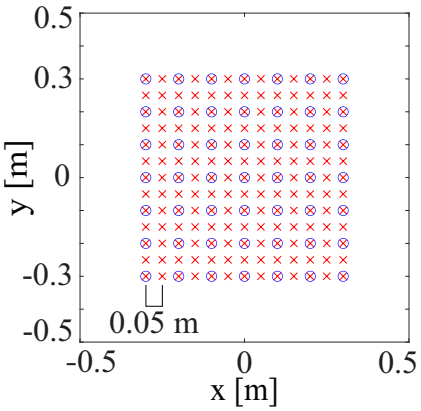


Fig. 2. Arrangement of microphones and control points (× : Control point, ○ : Microphone)

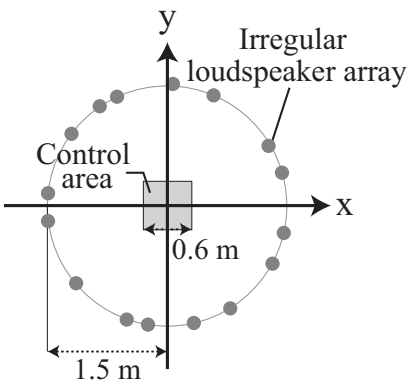


Fig. 3. Arrangement of irregularly-arranged loudspeaker array and control region

TABLE I
SIMULATION CONDITION

No. of training data	3072
No. of validation data	768
No. of test data	100
No. of microphones	7×7 (0.10 m interval)
No. of control points	13×13 (0.05 m interval)
Batch size	256
Learning rate	0.0001
Epoch	30000
Optimization method	Adam

The dataset for training consists of 7 × 7 microphones in the normalized desired sound field analytically calculated with the SFS toolbox [23]. Note that the input was zero-padded to match the 13 × 13, which is an arrangement of the output.

III. Simulation Experiment

A. Simulation condition

We conducted the simulation experiments to compare the reproduction accuracy of the reproduced sound field between the proposed deep-learning-based PM (DL–PM) method and the conventional PM methods. In the DL–PM method, training data were calculated from a single point source randomly placed between 1.5 m and 4.0 m in radius. The positions of point sources for evaluation were not included in the training data. Table I shows the conditions of the dataset and deep learning. The microphone spacing was set at 0.1 m, compared to a wavelength of approximately 0.17 m. Gaussian noise with an SNR of 30 dB was added to the microphone signals as input. For deriving the driving functions by RLS, the regularization parameter λ was 0.01.

Two types of loudspeaker arrays were used in the evaluation: a regularly-arranged array and an irregularly-arranged array. The loudspeaker array consists of 64 loudspeakers and is circular with a radius of 1.5 m. Fig. 3 shows an example of an irregularly arranged loudspeaker array. The center of the loudspeaker array is the origin of the coordinate system. For a regularly arranged loudspeaker array, the loudspeakers were spaced at 0.15 m intervals.

The normalized reproduction error (NRE) was used for the evaluation of the reproduction accuracy at the point $\mathbf{r}$.

$$\mathrm{NRE}(\mathbf{r}) = 10\log_{10}\frac{|\hat{p}^{\mathrm{syn}}(\mathbf{r}) - p(\mathbf{r})|^2}{|p(\mathbf{r})|^2}, \tag{7}$$

where p and $\hat{p}^{\mathrm{syn}}$ denote the desired and reproduced complex sound pressures, respectively. The evaluation points were 60 × 60 with a spacing of 0.01 m in the control region.

B. Results

The proposed DL–PM method was compared with three conventional PM methods: PM, LI–PM, and Full–PM. In the PM, the driving functions were determined so that the sound pressures were matched only at microphone positions. In the LI–PM, PM method was conducted by using the linear interpolated signals at the control points from amplitudes and unwrapped phases [24], [25] of sound pressures at the microphones. In the Full-PM, the desired sound pressures and transfer functions at all control points were known, and the PM method was conducted using all of this information.

Fig. 4 shows the reproduced sound fields at 2 kHz and NRE distributions for a point source at $(x, y) = (2.8, 0.5)$ m, which was not included in the training data. From Figs. 4(a) and (e), the reproduced sound field by PM was distorted. In PM, the sound field was controlled only by the microphone positions. Since the microphone spacing was 0.1 m and the wavelength was approximately 0.17 m, the spatial Nyquist frequency criterion was not satisfied at 2 kHz, which may have contributed to the degradation of reproduction accuracy. In addition, as shown in Figs. 4(b) and (f), LI–PM controlled the sound field at control points that satisfied the spatial Nyquist frequency by using linear interpolation. Consequently, LI–PM slightly improved reproduction accuracy compared to PM.

In contrast, from Figs. 4(d) and (h), DL–PM significantly improved the reproduction accuracy compared to PM and LI–PM. Furthermore, DL–PM achieved a similar performance to Full–PM, as shown in Figs. 4(c) and (g).

Next, the mean of NREs for 100 desired sound fields are shown in Table II. Note that in the case of irregular loudspeaker arrays, the mean NRE was calculated using 10

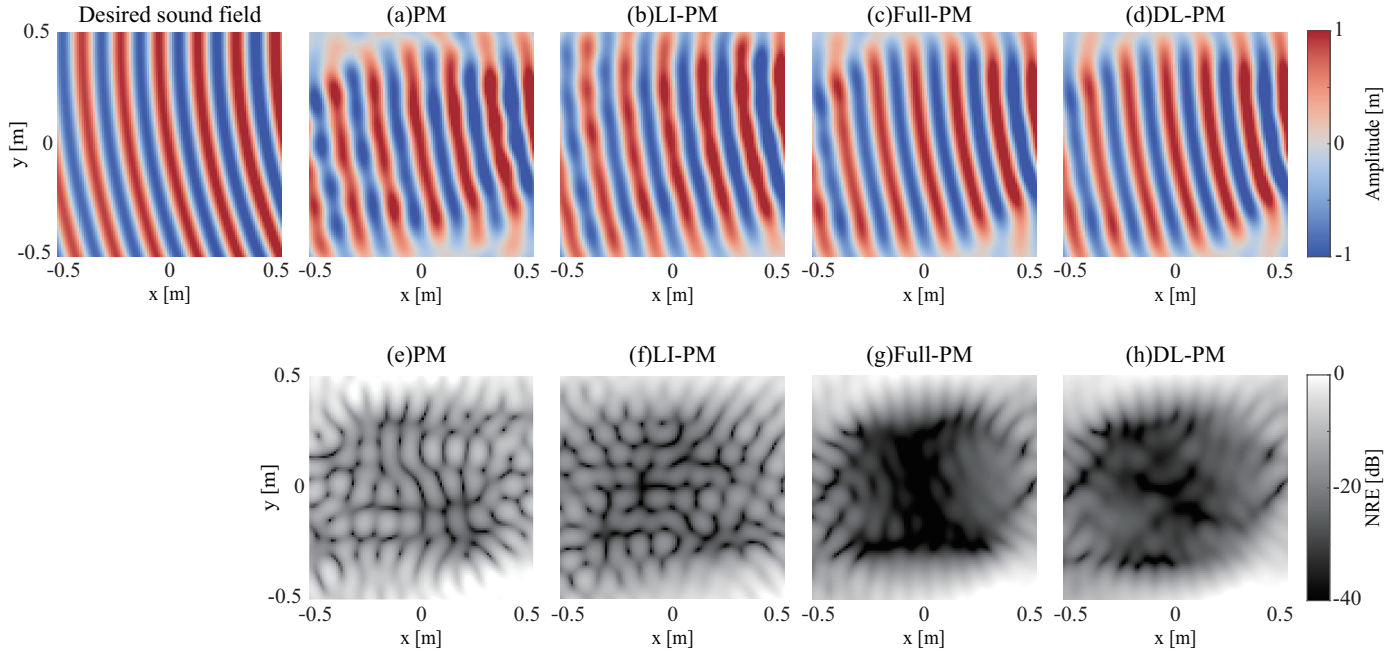


Fig. 4. Reproduced sound fields and NRE distributions for a source at $(x, y) = (2.8, 0.5)$ m at 2 kHz. The loudspeaker array was circular with a radius of 1.5 m and was arranged in a regular configuration. The number of loudspeakers L = 64.

irregular arrangements and 100 desired sound fields; that is, the mean NRE was calculated from 100 $\times$ 10 results. First, we compared the results for a regular arrangement. While the LI–PM improved reproduction accuracy by 1.5 dB compared to the PM, its mean NRE was still 16.2 dB below that of the Full–PM. By contrast, the mean NRE of DL–PM was $\mathrm{NRE} = -26.5$ dB, significantly improving PM and LI–PM performance. In addition, the difference between DL–PM and Full–PM was only approximately 3.7 dB. Consequently, the DL–PM achieved reproduction accuracy close to the Full–PM.

Furthermore, when the loudspeakers are irregularly arranged, the spacing between them is uneven, resulting in configurations that are difficult to reproduce accurately depending on the sound source location. This led to a decrease in the average NRE for all methods. Thus, even at Full–PM, the reproduction accuracy was 5.3 dB lower than the results for regular loudspeaker arrangement. The irregularly arranged loudspeakers reduced the mean NRE of the DL–PM by approximately 3.5 dB. However, the difference in mean NRE between Full–PM and DL–PM was reduced to 2.9 dB in the irregularly placed loudspeaker array because of the small decrease in DL–PM compared to the decrease in mean NRE of Full–PM. Therefore, the results show that DL–PM can reproduce the sound field in irregularly arranged loudspeaker arrays with reproduction accuracy close to that of Full–PM. In addition, the proposed method reduces the number of microphones by approximately 70% compared to the conventional method.

TABLE II
COMPARISON OF MEAN NREs WITH REGULAR- AND IRREGULAR-ARRANGED LOUDSPEAKER ARRAYS

Arrangement	PM	LI–PM	Full–PM	DL–PM
Regular [dB]	-12.5	-14.0	-30.2	-26.5
Irregular [dB]	-9.5	-11.2	-24.9	-22.0

IV. CONCLUSION

In this study, we proposed a PM method using data-driven estimation for the transfer functions of loudspeakers and desired sound fields. Simulation experiments demonstrated that the proposed method successfully interpolated the sound field at the control points with high resolution and achieved more accurate sound field reproduction compared to conventional methods. Additionally, the results showed that the sound field was reproduced with high accuracy even when using an irregular loudspeaker array. In future work, the application of the proposed method to the desired sound fields and loudspeaker transfer functions involving reflected sounds and multiple sources will be investigated.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number 22K12099 and Research Institute for Science and Technology of Tokyo Denki University Grant Number Q24J-04 / Japan.

References

[1] J. Ahrens, *Analytic Methods of Sound Field Synthesis*. Springer Berlin, Heidelberg, 2012.

[2] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[3] J. Ahrens, R. Rabenstein, and S. Spors, "The theory of wave field synthesis revisited," *Journal of The Audio Engineering Society*, pp. 17–20, 2008.

[4] D. Ward and T. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 697–707, 2001.

[5] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of The Audio Engineering Society*, vol. 53, no. 11, pp. 1004–1025, 2005.

[6] O. Kirkeby and P. A. Nelson, "Reproduction of plane wave sound fields," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2992–3000, 1993.

[7] P. Nelson, "Active control of acoustic fields and the reproduction of sound," *Journal of Sound and Vibration*, vol. 177, no. 4, pp. 447–477, 1994.

[8] P.-A. Gauthier, A. Berry, and W. Woszczyk, "Sound-field reproduction in-room using optimal control techniques: Simulations in the frequency domain," *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 662–678, 2005.

[9] T. Lee, J. K. Nielsen, J. R. Jensen, and M. G. Christensen, "A unified approach to generating sound zones using variable span linear filters," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 491–495.

[10] R. G. Baraniuk, "Compressive sensing [lecture notes]," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, 2007.

[11] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[12] M. Pezzoli, M. Cobos, F. Antonacci, and A. Sarti, "Sparsity-based sound field separation in the spherical harmonics domain," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1051–1055.

[13] O. Das, P. Calamia, and S. V. Amengual Gari, "Room impulse response interpolation from a sparse set of measurements using a modal architecture," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 960–964.

[14] S. A. Verburg and E. Fernandez-Grande, "Reconstruction of the sound field in a room using compressive sensing," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3770–3779, 2018.

[15] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2020.

[16] M. Pezzoli and a. A. S. F. Antonacci, "Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, 2023, pp. 2177–2184.

[17] X. Karakonstantis and E. Fernandez-Grand, "Room impulse response reconstuction using physics-informed neural networks," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, vol. 378, 2023, pp. 3181–3188.

[18] I. Tsunokuni, G. Sato, Y. Ikeda, and Y. Oikawa, "Spatial extrapolation of early room impulse responses with noise-robust physics-informed neural network," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E107-A, no. 9, 2024.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, 2012.

[20] L. Comanducci, F. Antonacci, and A. Sarti, "A deep learning-based pressure matching approach to sound-field synthesis," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, 2015, pp. 234–241.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, The Royal Society London, 2015, pp. 1026–1034.

[23] H. Wierstorf and S. Spors, "Sound field synthesis toolbox," *Journal of The Audio Engineering Society*, 2012.

[24] M. A. Herráez, D. R. Burton, M. J. Lalor, and M. A. Gdeisat, "Fast two-dimensional phase-unwrapping algorithm based on sorting by reliability following a non-continuous path," *Appl. Opt.*, vol. 41, no. 35, pp. 7437–7444, 2002.

[25] M. F. Kasim, *Fast 2d phase unwrapping implementation in matlab*, 2017. [Online]. Available: https://github.com/mfkasim91/unwrap_phase/.

Analytic Study of Text-Free Speech Synthesis for Raw Audio using a Self-Supervised Learning Model

Joonyong Park*, Daisuke Saito*, Nobuaki Minematsu*
* The University of Tokyo, Japan
E-mail: {jpark, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

***Abstract*—We examine the text-free speech representations of raw audio obtained from a self-supervised learning (SSL) model by analyzing the synthesized speech using the SSL representations instead of conventional text representations. Since raw audio does not have paired speech representations as transcribed texts do, obtaining speech representations from unpaired speech is crucial for augmenting available datasets for speech synthesis. Specifically, the proposed speech synthesis is conducted using discrete symbol representations from the SSL model in comparison with text representations, and analytical examinations of the synthesized speech have been carried out. The results empirically show that using text representations is advantageous for preserving semantic information, while using discrete symbol representations is superior for preserving acoustic content, including prosodic and intonational information.**

I. Introduction

Current speech synthesis has significantly advanced through deep learning models, greatly surpassing the performance of traditional speech synthesis models [1]. These models generally obtain speech feature vectors from the input text through an encoder, then output a Mel-spectrogram using methods such as Attention or Variational Inference, and finally convert it into speech through a Vocoder [2]–[4]. At this stage, speech synthesis models essentially learn the correspondence between the training speech source and the 'input representation,' which describes the speech source and is traditionally represented as a transcribed text script [5].

However, a constraint of such models is that enhancing performance necessitates the additional task of pairing input representations as labels to the training speech source, which involves human effort and thus incurs significant costs. Furthermore, such text-based input representations vary by language, requiring even more resources when creating multilingual speech synthesis. To overcome these constraints, the use of Self-Supervised Learning (SSL) models, which can extract usable information from the raw speech source, is being considered.

This study evaluates the performance of speech synthesis through several different input expressions, which analyzes the factors to consider when achieving such as "zero-resource synthetic speech", and evaluates the intelligibility, naturalness, and quality of the synthesized speech to see how each factor affects those speech.

II. Prior Studies

First, traditional speech synthesis models have focused on converting transcribed natural language text into a format more suitable for speech synthesis to obtain potential input representations. This includes text normalization, tokenization, the assignment of morphemes and parts of speech, phonemizing, and assigning accents to natural language texts through pre-prepared dictionaries and libraries. However, such methods have limitations in generalizing beyond single-language or high-resource language domains.

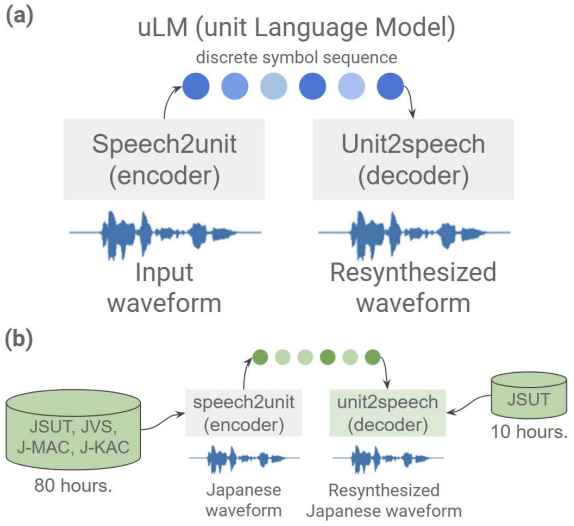


Fig. 1. (a) Architecture of GSLM and (b) Application to Japanese Language

Therefore, methods that can obtain effective intermediate representations for speech synthesis tasks without text transcription, using audio data as a modality, have recently been researched. Self-supervised learning models are trained on a large amount of unlabeled audio data to predict hidden units or symbols not directly observed in the input audio data. Consequently, obtaining corresponding input representations for unlabeled speech has become more accurate, and it has also become possible to incorporate paralinguistic and non-verbal information into these input representations.

In a previous study on this approach, the Generative Spoken Language Modeling (GSLM) [6] provided a methodology for speech synthesis from raw audio. Figure 1(a) shows a schematic illustration of the GSLM model architecture. GSLM processes and resynthesizes speech not through commonly used natural language texts but through discrete symbols, which are feature representations extracted from speech using SSL models. The model consists of an encoder called

speech2unit, a decoder called unit2speech, and a unit language model (uLM). The speech2unit module converts audio waveforms into discrete symbols in text form through feature representations via SSL models like CPC [7], wav2vec 2.0 [8], and HuBERT [9]. It then quantizes the features with a predetermined codebook to obtain the discrete symbol sequence. The codebook is obtained by applying k-means clustering to the framewise features of the training data.

Conversely, the unit2speech module generates audio waveforms from sequences of discrete symbols using traditional text-to-speech models like Tacotron 2 [5] and neural vocoder models. In this process, the speech synthesis model is trained with pairs of training speech and discrete symbols outputted by the encoder. In the case of speech resynthesis, there is a pipeline where the speech encoded through the speech2unit module is transformed back into synthesized speech through the unit2speech module. However, since unit2speech is basically trained in a single language, language dependency needs to be resolved to use it for multiple languages. Figure 1(b) shows a Japanese application of the GSLM model architecture, where language dependence was resolved to some extent by training the speech2unit's k-means clustering model and unit2speech module with a Japanese dataset. The uLM module, positioned between the two modules mentioned above, treats discrete symbols like character symbols and operates as a language model using the Transformer [10] network.

Additionally, as a prior study on the evaluation of synthesized speech without labels, the Zero Resource Speech Challenge (ZRC) can be mentioned. The research sets several tasks with the goal of constructing language processing models using only audio data by removing text labels from a ground truth speech corpus. The objective is to surpass the topline models, which are trained from transcribed text, in metrics for each task. Among the tasks, in the 'Discrete Resynthesis' task, the input representations obtained through SSL models are resynthesized, and clarity and naturalness are evaluated through Character Error Rate (CER) and Mean Opinion Score (MOS). The experimental results have shown that models adopting certain SSL methods achieved better MOS results than the topline, and it was also confirmed that the more bit-information the input representations contained, the higher the metrics of synthesized speech were. [11]–[13]

In addition, through structures such as GSLM, research on solving audio tasks as end-to-end using untranscribed audio data has continued to show superior results compared to previous baselines [14]–[16]. From such prior studies, it is suggested that by obtaining input representations through self-supervised learning methods for audio data without pre-transcribed labels, it is possible that SSL representations may show superior performance compared to text transcriptions for speech synthesis in general.

III. Comparative Study on Methods Based on Input Representations

In the case of the Zero Resource Speech Challenge, only the evaluation of intelligibility and naturalness after synthesis was present, while the quality of the acoustics was not evaluated. Additionally, prior studies have not conducted complete analysis of multiple languages or multiple encoder-decoder pairs, and this has not been analyzed with several conditions such as symbol size and output layer of Transformer.

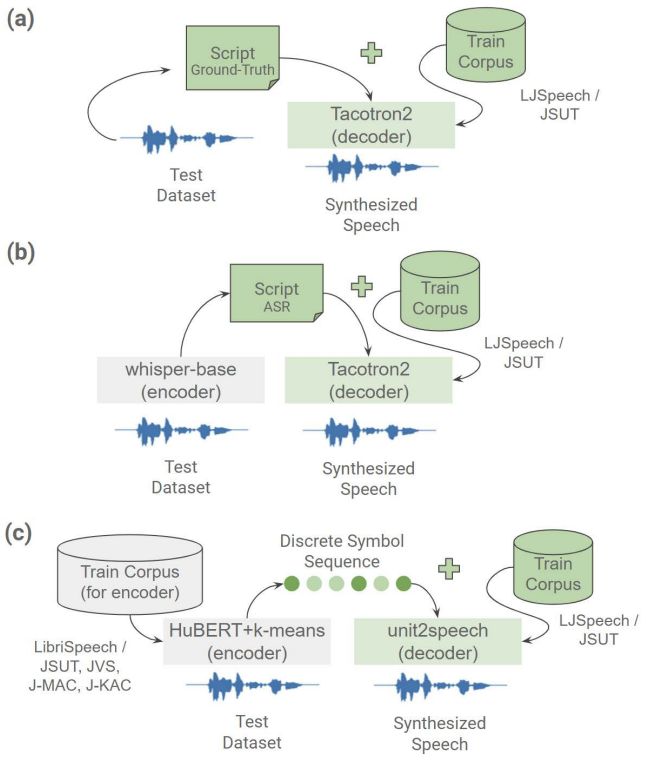


Fig. 2. Building a speech synthesis system using (a) ground-truth script labels, (b) speech recognition (ASR) model, and (c) self-supervised learning (SSL) model

Therefore, this study evaluates by adding metrics that can measure acoustic quality, in addition to the experimental results on clarity and naturalness, along with changes in the models used. Furthermore, SSL models are often trained in a specific language, which could show degraded performance in speech synthesis for multiple languages due to language dependency. Hence, this research proceeds with experiments not only in English but also in Japanese, investigates the presence or absence of language dependency in the representations outputted through SSL models, and ultimately analyzes the impact of such factors on the nature of synthesized speech.

A. Experiment Setting

In this experiment, models for three different input representations were created for two languages: English and Japanese. The structure is shown in Figure 2.

The first model (2-(a)) is used as a reference, synthesizing speech using text scripts as input representations, which serve as the correct (gold) labels.

The second model (2-(b)) serves as a baseline, synthesizing speech after recognizing the training speech dataset through an ASR model. The ASR model used is Whisper [24], with

TABLE I
CORPORA USED FOR TRAINING HYPOTHESIS MODEL

Language	S2u(multiple speakers)	u2S(single speaker)
English	LibriSpeech[17]	LJSpeech[18]
Japanese	Reazonspeech [19] (HuBERT) JSUT[20], JVS[21], JKAC[22], JMAC[23] (k-means)	JSUT

the base model used for both English and Japanese. For these above two models, preprocessing includes the use of the English cleaner provided by Tacotron2 for English, and morphological analysis through Mecab for Japanese.

The third model (2-(c)) is used as a hypothesis, synthesizing speech by obtaining input representations in discrete symbol form via the speech2unit module from the self-supervised learning model presented by GSLM. [1] For this, pipelines for English and Japanese are presented. The datasets for each language used in speech2unit and unit2speech are shown in Table 1; the sampling rate of the audio was unified to 16k for all pre-trained, training, and test datasets. We use HuBERT-base for outputting the input representation in the form of a discrete symbol through a k-means model.

In addition, several differential elements were introduced in the SSL model and the model was analyzed using the ablation method. Firstly, in order to analyze the language dependency of speech2unit, the difference between the two synthetic speech is analyzed as shown in Figure 3, assuming that speech2unit is trained in the same language as unit2speech and in a different language. This creates four combinations of S2u-u2S pair in total. Secondly, in order to evaluate the effect of changing the code length of the discrete symbol, synthetic speech that is synthesized from three different discrete symbols are compared, by training the k-means model's clustering number to 50, 200, and 1000, respectively. For all cases, processing was performed to remove repeated symbols from the discrete symbol sequences if length n or more is continuous for each symbol. For this experiment, n was fixed to 1. Thirdly, to identify differences in output between Transformer layers within speech2unit, synthetic speech from the representation by speech2unit's different layers is been compared: in this experiment, 6th and 12th layer. Attempts have been conducted to find effective features within SSL models, and it has also been found that the differences between these layers has made difference in 'linguistic' and 'acoustic' factors [25], [26]. These conditions also verify that the prior research is consistent with the variables of language and code length.

In common with all three models, the Tacotron 2 model was used for generating synthesized speech through input representations. The model is designed to train Mel-Spectrograms from the text of input representations and to output speech using a Vocoder under the same conditions for each language, except for the differences in input representations.

For the test datasets, 100 utterances selected from

[1]As this experiment focuses only on speech resynthesis, which does not require the uLM, we only analyze the speech2unit and unit2speech equivalent modules.

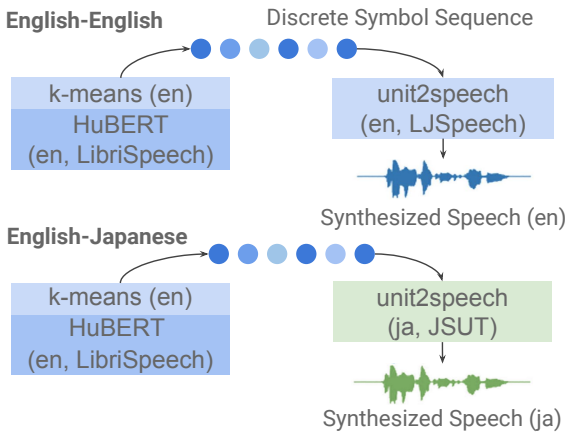


Fig. 3. An example of a SSL system that speech2unit-unit2speech pairs are language-matched and unmatched

LibriSpeech-dev were used for English, and for Japanese, 100 utterances selected from utterances not used as training data in the JSSS corpus [27] were used. The test dataset is also converted into the form of same representations that are used in the training phase, such as transcribed text (GT), ASR script (ASR), and discrete symbol sequence (SSL) respectively. Those input representations are used for inference.

IV. EXPERIMENT RESULTS

The experiment is broadly divided into evaluations of speech language and acoustic quality. The speech language evaluation aims to assess changes in the linguistic semantic elements of synthesized speech based on input representations, while the acoustic quality evaluation focuses on elements other than the language contained in the synthesized speech audio.

The data in Table 2 are the result of SSL speech synthesis comparing with the reference GT and baseline ASR. The characters in the item represents 'S2u-u2S language matching', 'symbol size', and 'Transformer output layer', respectively. For example, in the case of *match-200-L12*, it refers a synthesized speech that used same S2u with u2S, 200 symbol size and discrete symbols obtained from the 12th layer.

A. Speech Intelligibility

For the language evaluation, the impact of each input representation on the intelligibility of speech language was examined by investigating the error rates for the target languages. The output from Whisper-base and the correct scripts of each test dataset were compared, investigating Word Error Rate (WER) for English and Character Error Rate (CER) for Japanese. As an initial step in evaluating the performance of the speech recognition model used in the experiment, the error rate was investigated against the correct scripts for the GT test dataset processed by speech recognition. The results showed a word error rate of 2.41 % for English and a character error rate of 19.35 % for Japanese.

Also, to perform a more detailed analysis using a language-independent metric, phonemes were obtained from each piece of speech information, and their errors were also investigated.

TABLE II

METRICS COMPARING SYNTHESIZED SPEECH FROM THE INPUT REPRESENTATION OF GROUND TRUTH, ASR MODEL, AND VARIOUS SSL MODELS FOR ENGLISH AND JAPANESE RESPECTIVELY. ITEM IS EMBOLDENED WHEN IT RECORDED THE MOST SUPERIOR VALUE AMONG SSL MODELS, AND BASELINE IS UNDERLINED IF THE BASELINE VALUE IS SUPERIOR THAN THE ITEM.

	English					Japanese				
	WER(%)↓	PER(%)↓	UTMOS↑	WARP-Q↑	SDR(dB)↑	CER(%)↓	PER(%)↓	UTMOS↑	WARP-Q↑	SDR(dB)↑
GT	4.22	8.14	3.93 ± 0.067	-	-	20.68	41.59	2.68 ± 0.045	-	-
ASR	5.41	9.92	3.41 ± 0.084	2.71	-18.55	24.62	44.91	2.57 ± 0.049	2.23	-23.53
match-50-L6	8.37	19.09	3.31 ± 0.096	2.62	-19.91	42.88	82.19	2.10 ± 0.064	2.20	-23.95
match-50-L12	7.82	16.06	3.49 ± 0.076	2.78	-19.32	35.73	69.02	2.21 ± 0.055	2.29	-23.81
match-200-L6	6.95	15.24	3.45 ± 0.068	2.86	-17.14	32.95	59.08	2.44 ± 0.041	2.37	-23.56
match-200-L12	6.63	14.53	3.60 ± 0.057	2.89	-17.05	27.04	50.07	2.62 ± 0.053	2.35	-23.41
match-1000-L6	6.25	13.09	3.63 ± 0.060	**2.93**	-17.74	29.54	53.35	2.45 ± 0.051	2.36	-23.45
match-1000-L12	**5.45**	**10.37**	**3.72 ± 0.063**	2.92	-17.37	**26.58**	**49.40**	**2.64 ± 0.043**	2.35	-23.39
unmatch-50-L6	10.25	25.42	3.29 ± 0.093	2.74	-19.44	43.67	83.75	2.12 ± 0.067	2.18	-23.92
unmatch-50-L12	8.90	20.51	3.41 ± 0.079	2.69	-18.94	38.06	77.29	2.18 ± 0.059	2.21	-23.61
unmatch-200-L6	8.84	20.13	3.37 ± 0.069	2.78	-17.46	33.58	62.76	2.43 ± 0.053	2.33	-23.80
unmatch-200-L12	7.37	15.42	3.47 ± 0.058	2.84	**-16.73**	28.22	52.32	2.59 ± 0.047	2.37	-23.44
unmatch-1000-L6	6.31	13.96	3.61 ± 0.057	2.80	-17.09	31.83	59.28	2.45 ± 0.054	2.36	**-23.07**
unmatch-1000-L12	6.29	13.59	3.64 ± 0.065	2.87	-16.69	27.45	50.81	2.57 ± 0.048	**2.40**	-23.47

The Phoneme Error Rate (PER) was investigated using the speech of the GT test dataset as a reference, with each speech is analyzed as hypothesis. We used the phoneme recognizer, *allosaurus* [28], to obtain phoneme directly from the speech.

The average error rates obtained for speech synthesized from ground-truth scripts, ASR model, and SSL model input representations are shown in Table 2. For both languages, speech synthesis using the GT script showed the lowest error rate. On the other hand, for label-less cases, synthesis through the ASR model showed lower error rates than synthesis through the SSL model. Also, it was confirmed that the relationship between PER and WER or CER showed a proportional relationship with all data. This shows that the intelligibility consistently affects the errors from low-level elements like phonemes to high-level elements like words. These results suggest that using input representations obtained through ASR, which is natural text, leads to better conveyance of linguistic elements like intelligibility. This aligns with the tendency in the ZRC's prior study, where the topline using text transcription for both English and other languages showed superior ABX values in the ABX task used for phoneme discrimination [13].

Moreover, regarding the intelligibility of synthesized speech through language matching, the language-matched model's results showed slightly lower error rates than the unmatched model's, showing that language dependence of SSL model has non-negligible consequences. For the codec length, performance improved as the codec length of k-means clustering increased, reaching a level comparable to the ASR result in the case of English WER. In addition, in all cases, the synthesized speech through the 12th layer was able to obtain a lower error rate than the synthesized speech through the 6th layer. This is consistent with the results of previous studies that the layer close to the final output has more 'semantic' information than the intermediate layer. [25]

B. Speech Naturalness

In the case of ZRC, the subjective metric Mean Opinion Score (MOS) was used to evaluate naturalness, which can be considered an indicator that comprehensively evaluates both linguistic and extra-linguistic information. On the other hand, this study employed the pre-trained UTMOS model. UTMOS, trained on multiple languages including English, predicts automated MOS. UTMOS is evaluated independently without a comparative subject, allowing for the analysis of absolute linguistic and paralinguistic information, but making it impossible to compare voice changes between GT and synthesized speech.

The average UTMOS values obtained for speech synthesized from correct labels, ASR model, and SSL model input representations are shown in Table 2. [2] Common across all languages, speech synthesis using correct scripts showed the highest values. Meanwhile, under label-less synthesis conditions, the MOS for synthesized speech from SSL models was slightly higher across all language conditions than for those from ASR models, with this difference being greater in English synthesized speech than in Japanese. Additionally, speech2unit-unit2speech pairs of the same language showed predominant MOS scores for both languages.[3]

Moreover, the performance improved as codec length increased; however, while there was an big increase without exception when the token increasing from 50 to 200, there was a mild increase or even decrease when the token increasing from 200 to 1000. Thus, it can be said that naturalness is correlated to intelligibility, but this correlation decreases as the number of tokens increases. Also, the output obtained in the 12th layer obtained better results in terms of naturalness compared to those obtained in the 6th layer, following the trend has shown in the intelligibility: while some errors are within confidence levels.

These results propose a hypothesis that SSL discrete rep-

[2]Although 95 % confidence level was indicated in the table, as UTMOS is not a subjective value but an objective value from the model, the value was calculated from the standard deviation of the sample.

[3]The difference in absolute UTMOS values between Japanese and English can be attributed to the UTMOS evaluation model being primarily trained on English speech data.

resentations containing paralinguistic information like accents and intonations can enhance the naturalness of synthesized speech more than input representations from ASR models, suggesting that this can change depending on the language dependency of the SSL model and the amount of target language data it was trained on. It also suggests that increasing the language dependency of SSL models could incorporate more language-appropriate paralinguistic information within discrete representations. This hypothesis aligns with the results in ZRC, where the synthesized speech results for Indonesian, a language not trained in ZRC, more frequently failed to surpass the MOS of top-line models utilizing text transcription for English, a trained language [13].

C. Audio Quality and Noisiness

Following the evaluation of linguistic and naturalness aspects, which are related to speech language, the overall acoustic elements of the synthesized speech were assessed through quality evaluation.

Initially, the quality of digital audio was evaluated based on codecs, considering input representations as compressed audio representations, which can be viewed as neural speech codecs. Therefore, assuming the speech synthesis system as a single system, the output audio was analyzed as degraded audio compared to a reference. For such evaluations, PESQ is a representative metric, but in this case, speaker information and speech type can influence the results. Thus, to compensate for temporal mismatches between GT and resynthesized signals internally and to be resilient against errors commonly occurring in audio codecs, the WARP-Q metric was added to the evaluation. To address this, synthesized speech using correct scripts was used as a reference, and synthesized speech using ASR and SSL was compared as hypotheses.

The average values of WARP-Q [29] obtained for speech synthesized from ASR model and SSL model input representations are shown in Table 2. Generally, synthesized speech generated through the SSL model showed higher metrics compared to that synthesized through the ASR model. However, although there was a slight tendency due to Transformer layer and token length, differences due to the SSL model's encoder did not show a consistent trend across languages or tasks, with only minor increases or decreases observed.

The quality of audio in terms of noise inclusion was also evaluated. Signal Distortion Rate (SDR), although an indicator used in source separation, can be considered for evaluating the degree of noise inclusion in the output audio relative to the input audio, hence indicative of acoustic quality. The average SDR values obtained for speech synthesized from ASR model and SSL model input representations are shown in the table. Overall, although the values were low, synthesized speech through the SSL model showed slightly better metrics. Moreover, regarding differences due to the SSL model's encoder, systems crossing languages for English synthesized speech showed better results, while for Japanese synthesized speech, systems not crossing languages did better, indicating no clear trend. There is little trend for Transformer layer and token length as well.

Thus, speech synthesis models generated using speech recognition showed slightly more codec distortion and noise compared to those using SSL models, and synthesized speech's dependency on token language did not bring significant differences in speech quality.

V. Conclusion

This study investigated input representations in speech synthesis systems and created synthesized speech through systems constructed using each type of input representation. These were then compared and analyzed across languages and representations through newly considered metrics and results from prior research.

The findings revealed that no input representation demonstrated higher metrics than correct labels. Natural language input representations through speech recognition models showed dominance in linguistic vocal aspects, while discrete symbol input representations through self-supervised learning models were superior in aspects of naturalness and acoustic quality. Furthermore, the study highlighted differences in self-supervised learning model performance based on language dependency initiated by the language of the data. Also, the metrics shown overall improvement as the codec length of k-means clustering increased, more dramatically in the speech language metrics than the acoustic metrics. Furthermore, regardless of the language using, it is able to be found that Transformer layers close to the final output in linguistic elements had more semantic information than those that did not, and it was possible to demonstrate the tendency of previous studies.

Future efforts will focus on devising methods to minimize such language dependency. The goal is to advance the evaluation of tasks in more multilingual contexts than currently possible by transforming text into language-informed discrete tokens through tokenizers without preprocessing the text according to language-specific rules.

References

[1] L. Borgholt, J. D. Havtorn, J. Edin, L. Maaløe, and C. Igel, "A brief overview of unsupervised neural speech representation learning," *CoRR*, vol. abs/2203.01829, 2022. arXiv: 2203.01829.

[2] J. Shen, R. Pang, R. J. Weiss, *et al.*, *Natural tts synthesis by conditioning wavenet on mel spectrogram predictions*, 2018. arXiv: 1712.05884.

[3] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *ArXiv*, vol. abs/2006.04558, 2020.

[4] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 5530–5540.

[5] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural TTS synthesis by conditioning Wavenet on mel-spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4779–4783.

[6] K. Lakhotia, E. Kharitonov, W.-N. Hsu, *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[7] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. arXiv: 1807.03748.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[11] E. Dunbar, J. Karadayi, M. Bernard, *et al.*, "The Zero Resource Speech Challenge 2020: Discovering Discrete Subword and Word Units," in *Proc. Interspeech 2020*, 2020, pp. 4831–4835.

[12] E. Dunbar, M. Bernard, N. Hamilakis, *et al.*, "The zero resource speech challenge 2021: Spoken language modelling," in *Proc. Interspeech 2021*, 2021, pp. 1574–1578.

[13] E. Dunbar, N. Hamilakis, and E. Dupoux, "Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1211–1226, 2022.

[14] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6562–6566, 2021.

[15] X. Li, Y. Jia, and C.-C. Chiu, "Textless direct speech-to-speech translation with discrete speech representation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096797.

[16] H. Kim, S. Kim, J. Yeom, and S. Yoon, "UnitSpeech: Speaker-adaptive Speech Synthesis with Untranscribed Data," in *Proc. INTERSPEECH 2023*, 2023, pp. 3038–3042. DOI: 10.21437/Interspeech.2023-2326.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

[18] K. Ito and L. Johnson, *The lj speech dataset*, https://keithito.com/LJ-Speech-Dataset/, 2017.

[19] Y. Yue, M. Daijiro, and F. Seiji, *Reazonspeech: A free and massive corpus for Japanese ASR*, https://research.reazon.jp/_static/reazonspeech_nlp2023.pdf, 2023.

[20] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis," *CoRR*, vol. abs/1711.00354, 2017. arXiv: 1711.00354.

[21] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: Free Japanese multi-speaker voice corpus," *CoRR*, vol. abs/1908.06248, 2019. arXiv: 1908.06248.

[22] W. Nakata, T. Koriyama, S. Takamichi, *et al.*, "Audiobook speech synthesis conditioned by cross-sentence context-aware word embeddings," in *Proc. The 11th ISCA SSW*, 2021.

[23] S. Takamichi, N. Wataru, T. Naoko, and S. Hiroshi, "J-MAC: Japanese multi-speaker audiobook corpus for speech synthesis," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association*, ISCA, 2022, pp. 2358–2362.

[24] A. Radford, K. Jong Wook, X. Tao, B. Greg, M. Christine, and S. Ilya, *Robust speech recognition via large-scale weak supervision.* https://cdn.openai.com/papers/whisper.pdf, 2022.

[25] P. Kumar, V. N. Sukhadia, and S. Umesh, "Investigation of robustness of hubert features from different layers to domain, accent and language variations," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6887–6891.

[26] K. Qian, Y. Zhang, H. Gao, *et al.*, "ContentVec: An improved self-supervised speech representation by disentangling speakers," in *Proceedings of the 39th International Conference on Machine Learning*, 2022, pp. 18 003–18 017.

[27] S. Takamichi, M. Komachi, N. Tanji, and H. Saruwatari, "JSSS: free japanese speech corpus for summarization and simplification," *CoRR*, vol. abs/2010.01793, 2020. arXiv: 2010.01793.

[28] X. Li, S. Dalmia, J. Li, *et al.*, "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8249–8253.

[29] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "Speech quality assessment with WARP-Q: From similarity to subsequence dynamic time warp cost," *IET Signal Processing*, vol. 16, no. 9, pp. 1050–1070, 2022.

Acoustic model adaptation in noisy and reverberated scenarios using multi-task learned embeddings

Aditya Raikar* and Meet Soni† and Ashish Panda ‡ and Sunil Kumar Kopparapu ‡
* Cerence Services (India) LLP
E-mail: aditya.raikar@cerence.com
† Rakuten Institute of Technology, Bangalore, India
E-mail: meetkumar.soni@rakuten.com
‡ TCS Research, Tata Consultancy Services, Mumbai, India
E-mail: {ashish.panda, sunilkumar.kopparapu}@tcs.com

***Abstract*—Distortions, such as revebererations and background noise in an acoustic environment pose a significant challenge for large-scale Automatic Speech Recognition (ASR) systems. In this paper we explore acoustic model adaption, to be used in conjuction with multi-conditioning, to address this challenge. Recently, it has been standard to use i-vectors for Acoustic Model (AM) adaptation. With a varying degree of success, Single Task Learned (STL) neural embeddings such as x-vector and r-vector, have also been used for AM adaptation. This paper proposes the use of Multi Task Learned (MTL) embeddings for large vocabulary hybrid acoustic model adaptation in noisy and reverberant environments. Our experiments on noise added Librispeech database as well as real reverberated HRRE database show that this approach provides significant performance gain when used along with the traditional i-vectors.**

I. Introduction

ASR systems have undergone tremendous transformations over the last decade due to the emergence of neural network based models, availibility of large data sets and computationally advanced machines [1], [2]. However, modern ASR's robustness is still affected by adverse acoustic conditions such as channel distortions, background noise and room reverbations [3]. Researchers have tried to improve the robustness of such systems by employing enhancement front-ends [4]–[9], by data augmentation algorithms [10], [11] or by using acoustic model (AM) adaptation techniques [12]–[14].

In this paper, we focus on the AM adaptation aspect of robust ASR. The end goal of such technique is to improve the robustness by effectively handling mismatched conditions between the training and test data distributions, achieved by adapting on utterance specific auxiliary information such as channel, ambiance, speaker, language and accent, etc. Traditionally, i-vectors [12]–[14] have been used for acoustic model adaptation and have been shown to capture the speaker and channel information quite efficiently. Similarly, deep neural network embedding based x-vectors, have also been shown to capture speaker, channel, and phrase related characteristics. Although x-vectors outperform i-vectors in the speaker recognition/verfication task [15], their use in AM adaptation for improving ASR robustness is still in initial stages and is receiving a lot of attention recently. A detailed study on the correlation between speaker embedding's performance in speaker recognition task and ASR was studied in [16]. A comparision study between i-vectors, x-vectors and deep convolutional neural network (CNN) based embeddings is done in [17], where x-vector have been shown to provide no improvement in AM adaptation. Accent embeddings, similar to x-vector, were used in [18] to eliminate the mismatch between native and non-native speech utterances. An overview of different adaptation techniques was presented with a focus on domain, speaker and accent adaptation in [19]. Contrary to [17], x-vector adaptation was found to be better than i-vector adaptation in low-resource scenarios in [20]. Another STL embedding called r-vector has been successfully shown to be effective for reverberant conditions in [21].

Multi Task Leaning (MTL) [22] in not new in the parlance of ASR. MTL networks have been employed for various tasks in ASR such as: predicting speaker identity [23], noise [24] and accent [25]. However MTL embeddings stacked with feature vectors have not been explored for AM adaptations. MTL embeddings have certain advantages in the sense that they can be trained on a database, different from the ASR training corpus, which have been annotated with speaker label, noise label, Room Impulse Response (RIR) labels etc. A multi-task learning neural network can be trained to exploit these annotations which will make it possible to extract an embedding from it which will encode various parameters related to an acoustic environment. Such an embedding will be highly desirable for AM adaptation. Indeed, in [26], we have shown that such an MTL embedding can encode speaker and room size information. In this paper, we study AM adaptation using MTL embeddings for large vocabulary ASR systems, where the multi-task learning network is trained for speaker, room size and noise classification. The noise classification is important, since noises can be classified by type or level (Signal-to-Noise-Ratio or SNR). The main contributions of this paper are:

- This work explores MTL embeddings for AM adaptation for speaker, room size and noise level.
- We show that the proposed MTL embeddings are better

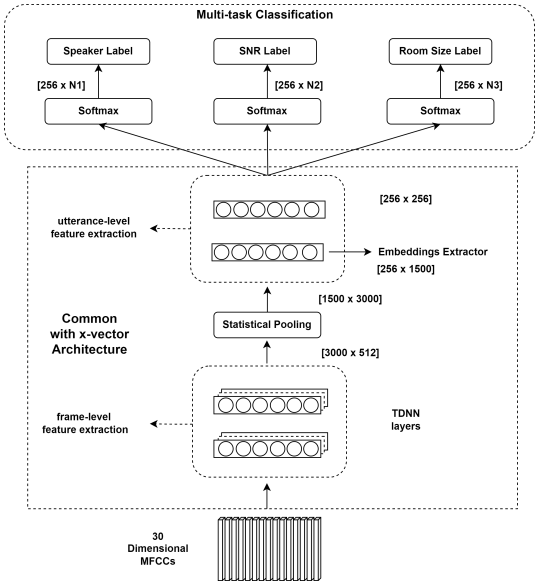


Fig. 1. Proposed extraction of MTL embeddings with three task classifier for speaker (N1 number of speaker labels), SNR (N2 number of SNR labels) and room size (N3 number of room size labels).

than STL embeddings such as x-vectors and r-vectors for the ASR for noisy and reverberant speech when used along with i-vectors.

II. Multi Task Learning based Embeddings

The neural embeddings, such as x-vector and r-vector, are trained over room acoustics and speaker information respectively, with an objective to discriminate between corresponding single task labels. Such embeddings are learned using the TDNN layers [27] [16], followed by a pooling layer, which collects the temporal statistics of the utterance. Finally, such embeddings are usually extracted from the subsequent affine layer in the form of a fixed dimensional vector.

In order to learn MTL embeddings, the single task learning architecture is modified to learn multiple tasks based classifications as shown in Figure 1. With this modification, the embeddings will encode all the information related to different auxillay tasks, that may be important from the ASR perspectives, in a fixed dimensional common embedding space. The proposed architechture is trained to classify speaker identity, room-size and Signal-to-Noise Ratio (SNR). It is important to note that our approach to labeling of the classes are unique for reverberation as well as noise. While r-vector implementation [21] takes individual RIR as a classification label, we use room size as classification label. Doing this makes learning task simpler, since all the RIRs with similar characterstics are put in one group. Additionally, it also enables the use of large number of RIRs without increasing the output size of the network. We have similarly used noise level (SNRs) as classification label. We found this to be more suitable than noise type, as it is very difficult to define the noise types for unseen noises. Using noise types (such as wind noise, dog bark etc.) as classification labels resulted poor classification accuracy. However, using SNRs as the labels the classification accuracy was very high.

To compute the loss incorporating *n* different auxillary tasks, loss function is defined as:

$$\mathcal{L}_{\text{multi-task}} = \lambda_{\text{Task}_1}\mathcal{L}_{\text{Task}_1} + \lambda_{\text{Task}_2}\mathcal{L}_{\text{Task}_2} + ... + \lambda_{\text{Task}_n}\mathcal{L}_{\text{Task}_n} \quad (1)$$

Each task loss $\{\ \mathcal{L}_{\text{Task}_1}, \mathcal{L}_{\text{Task}_2}, ..., \mathcal{L}_{\text{Task}_n}\}$ is weighted by loss weighting $\{\lambda_{Task_1}, \lambda_{Task_2}, ..., \lambda_{Task_n}\}$ respectively. With this formulation, the weights are choosen to adjust the level of influence a task will have in the AM adaptation process. Losses with equal weights, have equal amount of influence in AM adaptation. This way we can tune the weights based on which task is more important from the ASR perspective. We have used and modified the implementation employed in [28], for this work.[1]. The architecture of embedding extractor is same as that of the x-vector, until the embedding layer, which has got 256 hidden units. It is followed by 2 hidden layers, with 256 dimension each. This is further projected on to the number of classes related to different auxillary tasks. The loss function used in the classification task is cross entropy loss.

III. Experimental Setup

A. MTL Embedding extractor

The MTL extractor is trained on the reveberated and noisy version of the Libripseech data [29]. The dataset spans over 960 hours and consists of around 2338 number of speakers. To augment with reverberation, 80% of the total simulated Room Impulse Responses (RIRs) are taken from [30] [31] and are pre-labeled with small, medium and large, based on the room size. An additional anechoic class is added, thus making the total room size labels as 4. Reverb label is choosen with equal probability and augmented to each utterance of the dataset. To augment with noise, generative noise is added with different parameters at various SNRs with labels distributed in 3 class labels, defined as, Class 1: 0-10dB, Class 2: 11-20dB, Class 3: 21-30dB. We used the generative noise alrogithm described in [10] to create these artificial noise. This ensures that the noise used for MTL embedding extractor is different from the ones used for acoustic model testing. In order to evaluate the MTL extractor, a testset is constructed using reverberated noisy version of test-clean set of Librispeech. All the networks were trained on 30 dimensional with 350 frames of MFCC and total of 100,000 iterations. The batch size is kept at 500, with a small held out set of training utterances kept for validation. Stochastic gradient descent with learning rate 0.2 was used along with 0.5 momentum.

B. ASR Test Set

We test the proposed MTL embeddings based AM adaptation in three scenarios: speech with only additive noise ("noise only"), speech with only reverberation ("reverb only") and speech with both additive noise and reverberation ("reverb and noise"). For this we have used four distinct test sets. To evaluate the only additive noise scenario, we added babble

[1]https://github.com/cvqluu/MTL-Speaker-Embeddings

noise to the test-clean set of Librispeech at various SNRs such as 0dB, 5dB, 10dB and 20dB. The babble noise was created from MUSAN speech data so that it is different from the noises used in acoustic model training and MTL embedding extractor. To evaluate the reverberation only scenario, we created reverberated versions of the test-clean set of Librispeech. RIRs used to corrupt the test-clean set were different from the ones used for training either the MTL embedding extractor or the acoustic model. For this purpose, we use have used real RIRs: 271 from MIT impulse response survey [32], 188 from voiceHome [33], 2325 from [34] and 327 from [30] [31]. The naming convention followed for reverberated Librispeech test-clean set is $\{$test-c-rev-set-#$\}\{\# = 1,2,3\}$. To test both additive noise and reverberation, we used noise added version of the Highly Reverberant Real Environment (HRRE) dataset [35]. The dataset consists of around 13.4 hours clean speech utterances taken from Aurora-4 which is recorded in 20 different kinds of highly reverberant real environments, with varying reflection coefficients and speaker-microphone pair distances. We added the babble noise and noises from ESC-50 dataset [36] to create two different versions of noisy HRRE. The additive noises were added to the utterances from ESC-50 datasets at various SNRs, namely 0dB, 5dB, 10dB and 20dB, randomly and with uniform probability. Thus, the noisy HRRE test sets have both unseen additive noise as well as unseen reverberations.

C. Acoustic Model Training

All ASR models evaluated in this work are trained on Librispeech [29] dataset. We have constructed three different augmented versions of Librispeech 100 hours training dataset, i.e, "reverb only", "noise only" and "reverb and noise". To simuate reverb, we corrupted the the training speech with RIRs that are same as used to train the MTL embedding extractors. This ensured that the RIRs used for training and testing the acoustic models were different. Additive noises used in augmenting the dataset are taken from Musan noises at different SNRs. This again ensured that the additive noises are different for training and testing.

All ASR systems trained using default kaldi recipes [30] with Librispeech train-clean-100 set, we use kaldi Librispeech recipe[2] and modify it as follows. We train GMM-HMM system till $tri4$ using 100h subset and use it to obtain alignments of training data. Then we use default chain recipe to extract high-resolution MFCC features, train i-vector extractor and i-vector extraction, and perform AM training using Lattice-Free Maximum Mutual Information (LF-MMI) objective. During decoding, we use standard Language Models (LM) available for each task. For Librispeech evaluation subsets, we use small 3-gram LM and rescore the lattices generated using large 4-gram LM. For HRRE evaluation set, we use small 3-gram wallstreet LM.

[2]https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/run.sh

IV. Results and Discussion

In this section we will report and analyze the performance of our proposed MTL-embedding based AM adaptation and other STL-embedding based AM adaptation.

A. Experiments on "Noise Only" Scenario

The experimental results related to "Noise only" scenarios is reported in Table I. We have tested the performance of only embeddings and embeddings in addition to i-vectors. It can be observed that x-vector on its own provides slightly better performance only in high noise cases of 0dB. The proposed MTL-embeddings, do not provide improvements over i-vectors on their own. However, when used in conjuction with the i-vectors, the MTL-embeddings improve the performance significantly. This shows that the MTL-embeddings have information that complimentary to the i-vectors. MTL-embeddings provide performance gain for both the test combination of weights in the loss function. However, MTL(1,1) provides the best performace gain resulting in 1.38% absolute gain for 0dB case. It shows that the noise and speaker information have equal effect on the ASR performance.

Embeddings	**Babble Noise**			
	test_clean			
	0dB	**5dB**	**10dB**	**15dB**
i-vector	48.15	20.33	9.59	6.41
x-vector	47.94	20.79	9.92	6.75
MTL(1,0.1)	49.41	21.04	9.83	6.79
MTL(1,1)	50.57	21.45	10.04	6.76
i-vector+x-vector	47.64	19.78	9.45	6.41
i-vector+MTL(1,0.1)	47.14	19.51	9.33	**6.33**
i-vector+MTL(1,1)	**46.76**	**19.14**	**9.27**	6.37

TABLE I
RESULTS (% WER) OF ONLINE I-VECTOR, X-VECTOR AND THE PROPOSED MTL EMBEDDING FOR "NOISE-ONLY" SCENARIO. MTL(W1,W2) SIGNIFIES ASSIGNMENT OF WEIGHT W1 FOR SPEAKER INFORMATION LOSS AND W2 FOR NOISE INFORMATION LOSS.

B. Experiments on "Reverb Only" Scenario

For the sake of completeness of this paper, we are reproducing some of the results from "reverb only" scenario from [26] here. The experimental results reported in Table II provide an insight into the performance of various embeddings in "reverb only" scenario. Again, we test the embeddings on their own and in conjuction with i-vector. It can observed that for reverberated speech, x-vector performs better than i-vector and r-vector performs better than both. However, the proposed MTL-embeddings perform the best irrespective of the the weight assigned to speaker information loss and room size loss. Out of the three sets of weights tested by us, assigning weights 0.1 to speaker information loss and 1 to room size loss provided the best performance resulting in 1% absolute performance gain over the r-vectors. The weights assigned in the loss function suggest that room size information is more important for performance of ASR systems in the case of

reverberated speech. The next set of experiments explored possibility of fusion with online i-vectors. Such fusion techniques invariably improved the performance over online i-vectors. It seems to suggest that i-vectors alone may not be sufficient for AM adaptations in reverberated conditions. While r-vectors provided better performance than x-vectors on their own, they were worse performers in fusion. Stacking x-vectors with i-vectors resulted in an average WER of 28.73% in reverberated conditions while stacking r-vectors with i-vectors resulted in 29.15%. This shows that x-vectors compliment i-vectors better than r-vectors. The proposed MTL embeddings, however, comfortably outperform both other fusion schemes. Proposed MTL embeddings with equal weights to speaker information and room size loss was the best performer when fused with i-vectors. This fusion resulted in a relative performance gain of 13.27% over i-vector alone and 3.5% relative improvement over i-vector and x-vector fusion. Though the weights assigned to individual task loss is important, it is noteworthy that all three selected weights combinations outperform x-vectors and r-vectors. Another observation is, although the training of MTL embedding extractor and AM adapataion is done using simulated RIRs, it is able to perform quite good on disjoint real RIR test set, and outperform i-vector by a significant margin.

Embeddings	**Test Data WER (in %)**				
	test	**test-c-rev-set**			
	clean	**1**	**2**	**3**	**Avg.**
i-vector	5.43	31.12	32.69	32.01	32.94
x-vector	5.46	28.76	30.81	30.08	29.88
r-vector	5.49	28.16	30.09	29.57	29.54
MTL (1,0.1)	5.42	27.87	29.78	29.36	29.00
MTL (1,1)	5.46	27.47	29.63	29.41	28.83
MTL (0.1,1)	5.42	27.26	29.32	29.00	28.52
ivector+xvector	5.23	27.85	29.39	28.95	28.73
ivector+rvector	5.37	28.04	30.02	29.41	29.15
ivector+MTL(1,0.1)	**5.18**	26.75	28.95	**28.12**	27.94
ivector+MTL(1,1)	5.29	**26.62**	**28.51**	28.15	**27.72**
ivector+MTL(0.1,1)	5.3	27.03	29.11	28.83	28.32

TABLE II
RESULTS (% WER) OF ONLINE I-VECTOR, X-VECTOR, R-VECTOR AND THE PROPOSED MTL EMBEDDING FOR VARIOUS REVERBERATED CONDITIONS. MTL(W1,W2) SIGNIFIES ASSIGNMENT OF WEIGHT W1 FOR SPEAKER INFORMATION LOSS AND W2 FOR ROOM SIZE LOSS.

C. Experiments on "Reverb and Noise" Scenario

The results for the "reverb and noise" scenario are presented in Table 3(a) and Table 3(b), with ESC-50 and Babble noise respectively . We have tested only MTL(1,1,1) (equal weights to speaker, room and noise information) in this section as we observed from "noise only" and "reverb only" experiments that all these informations are equally important from ASR perspective. Firstly, it can be observed from Table 3(b) that, for all methods, worse result is obtained from babble noise. This is probably because, this type of noise is completely unseen from the acoustic model training perspective. Further, it can be observed from Table 3(a) that the ESC noises do not fare so badly probably because the MUSAN noises used for acoustic model training contain similar types of noises as ESC. In both cases, however, it can be seen that the proposed MTL-embeddings when used along with i-vector outperform other types of embeddings. Out of the 20 individual test sets of the HRRE database, the proposed embeddings outperform i-vector and r-vector in all, while they outperform the x-vector in a majority of the 20 test sets. On an average too, the MTL-embeddings perform significantly better than i-vector, x-vector and r-vector.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced MTL-embeddings for AM adaptation for improved ASR performance in three challenging conditions, i.e, "reverb only", "noise only" and "reverb and noise". The MTL-embedding extractor was trained to classify multiple tasks, namely: speaker identity, room size and SNR levels. We repoted results on experiments conducted on noise added Librispeech test data and noise added real reverberated HRRE test data. The experimental results suggest that the proposed MTL-embeddings have information that are complimentary to i-vector and when used along with i-vector for acoustic model adatation, they provide significant improvements. It should be noted that these improvements are realized on top of multi-conditioning, thereby improving the multi-conditioning performance by some margin. It would be interesting to see whether these MTL-embeddings have complimentary informations to x-vector and r-vector and how they can affect the peformance when used along with these STL-embeddings. Although, we have selected speaker ideantiy, room-size and SNR as the classification labels for the MTL-embedding extractor to good effect, it would be interesting to see whether the classification labels can be futher optimized for better performance.

REFERENCES

[1] G. Saon *et al.*, "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[2] C.-C. Chiu *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 4774–4778.

[3] V. Mitra, H. Franco, R. M. Stern, *et al.*, "Robust features in deep-learning-based speech recognition," in *New era for robust speech recognition*, Springer, 2017, pp. 187–217.

[4] K. Janod, M. Morchid, R. Dufour, G. Linares, and R. De Mori, "Denoised bottleneck features from deep autoencoders for telephone conversation analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1809–1820, 2017.

[5] V. Mitra, H. Franco, C. Bartels, J. van Hout, M. Graciarena, and D. Vergyri, "Speech recognition in unseen and noisy channel conditions," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 5215–5219.

Table 3(a)

Testset (Reverberated Noisy)	Embeddings			
	i-vector			
		+MTL(1,1,1)	+x-vector	+r-vector
test_HRRE_rev_ESC_set_1	9.12	7.36	7.41	9.12
test_HRRE_rev_ESC_set_2	9.25	9.02	8.57	9.06
test_HRRE_rev_ESC_set_3	14.29	13.11	12.93	12.82
test_HRRE_rev_ESC_set_4	32.64	28.06	31.46	30.11
test_HRRE_rev_ESC_set_5	42.03	35.71	37.98	36.48
test_HRRE_rev_ESC_set_6	4.58	4.65	4.93	4.73
test_HRRE_rev_ESC_set_7	7.32	7.29	7.27	7.31
test_HRRE_rev_ESC_set_8	10.54	9.32	9.94	9.91
test_HRRE_rev_ESC_set_9	15.62	14.91	14.89	14.37
test_HRRE_rev_ESC_set_10	28.94	25.87	29.18	26.99
test_HRRE_rev_ESC_set_11	3.92	4.07	4.17	4.18
test_HRRE_rev_ESC_set_12	6.03	5.27	5.81	5.74
test_HRRE_rev_ESC_set_13	7.14	6.63	6.78	6.74
test_HRRE_rev_ESC_set_14	13.15	11.91	12.31	12.42
test_HRRE_rev_ESC_set_15	16.94	14.81	16.79	15.84
test_HRRE_rev_ESC_set_16	3.79	3.61	3.89	3.81
test_HRRE_rev_ESC_set_17	4.84	4.51	4.75	4.61
test_HRRE_rev_ESC_set_18	4.67	4.28	4.26	4.39
test_HRRE_rev_ESC_set_19	5.32	5.12	5.51	5.45
test_HRRE_rev_ESC_set_20	7.12	6.97	7.23	7.55
Average	12.36	**11.12**	11.80	11.58

Results (% WER) of online i-vector, x-vector, r-vector and the proposed MTL embedding for "reverb and noise" scenario on HRRE dataset with additive **ESC-50** noise. MTL(w1,w2,w3) signifies assignment of weight w1 for speaker information loss, w2 for noise information loss and w3 for room size loss.

Table 3(b)

Testset (Reverberated Noisy)	Embeddings			
	i-vector			
		+MTL(1,1,1)	+x-vector	+r-vector
test_HRRE_rev_babble_set_1	33.98	27.89	30.41	**33.21**
test_HRRE_rev_babble_set_2	29.65	28.49	28.41	29.37
test_HRRE_rev_babble_set_3	35.14	34.02	35.71	36.58
test_HRRE_rev_babble_set_4	64.84	62.69	62.81	65.96
test_HRRE_rev_babble_set_5	71.01	70.11	70.65	72.93
test_HRRE_rev_babble_set_6	17.99	16.21	16.85	17.82
test_HRRE_rev_babble_set_7	26.17	25.18	25.35	26.34
test_HRRE_rev_babble_set_8	29.82	28.43	29.29	29.93
test_HRRE_rev_babble_set_9	40.44	37.89	38.74	39.83
test_HRRE_rev_babble_set_10	63.29	60.94	60.83	63.93
test_HRRE_rev_babble_set_11	15.09	14.35	14.87	15.37
test_HRRE_rev_babble_set_12	21.02	20.44	21.05	20.66
test_HRRE_rev_babble_set_13	22.98	21.71	22.51	23.54
test_HRRE_rev_babble_set_14	38.45	34.52	36.09	37.04
test_HRRE_rev_babble_set_15	45.11	40.84	43.04	46.87
test_HRRE_rev_babble_set_16	17.07	15.84	16.63	16.96
test_HRRE_rev_babble_set_17	19.45	17.54	18.57	19.39
test_HRRE_rev_babble_set_18	16.21	14.68	15.52	16.05
test_HRRE_rev_babble_set_19	18.01	17.32	17.17	17.65
test_HRRE_rev_babble_set_20	23.69	21.07	22.04	22.6
Average	32.47	**30.51**	31.33	32.60

Results (% WER) of online i-vector, x-vector, r-vector and the proposed MTL embedding for "reverb and noise" scenario on HRRE dataset with additive **babble** noise. MTL(w1,w2,w3) signifies assignment of weight w1 for speaker information loss, w2 for noise information loss and w3 for room size loss.

[6] A. Maas, Q. V. Le, T. M. O'neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr," 2012.

[7] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 1996–2000.

[8] C.-T. Do and Y. Stylianou, "Improved automatic speech recognition using subband temporal envelope features and time-delay neural network denoising autoencoder.," in *INTERSPEECH*, 2017, pp. 3832–3836.

[9] B. Das and A. Panda, "Robust front-end processing for speech recognition in noisy conditions," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 5235–5239.

[10] M. H. Soni, S. Joshi, and A. Panda, "Generative noise modeling and channel simulation for robust speech recognition in unseen conditions.," in *INTERSPEECH*, 2019, pp. 441–445.

[11] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[12] N. Dehak *et al.*, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[13] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, IEEE, 2013, pp. 55–59.

[14] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. Černockỳ, "Ivector-based discriminative adaptation for automatic speech recognition," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, IEEE, 2011, pp. 152–157.

[15] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors.," in *Odyssey*, 2018, pp. 105–111.

[16] J. Rownicka, P. Bell, and S. Renals, "Embeddings for dnn speaker adaptive training," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 479–486.

[17] J. Rownicka, P. Bell, and S. Renals, "Analyzing deep cnn-based utterance embeddings for acoustic model adaptation," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 235–241.

[18] M. T. Turan, E. Vincent, and D. Jouvet, "Achieving multi-accent asr via unsupervised acoustic model adaptation," in *INTERSPEECH 2020*, 2020.

[19] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation algorithms for speech recognition: An overview," *arXiv preprint arXiv:2008.06580*, 2020.

[20] M. Karafiát *et al.*, "Analysis of x-vectors for low-resource speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6998–7002.

[21] Y. Y. Khokhlov *et al.*, "R-vectors: New technique for adaptation to room acoustics," in *INTERSPEECH*, 2019, pp. 1243–1247.

[22] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[23] G. Pironkov, S. Dupont, and T. Dutoit, "Speaker-aware multi-task learning for automatic speech recognition," in *23rd International Conference on Pattern Recognition (ICPR)*, 2016.

[24] S. Kim, B. Raj, and I. R. Lane, "Environmental noise embeddings for robust speech recognition," *CoRR*, vol. abs/1601.02553, 2016. arXiv: 1601.02553. [Online]. Available: http://arxiv.org/abs/1601.02553.

[25] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and

multi-task learning," in *Proc. Interspeech 2018*, 2018, pp. 2454–2458. DOI: 10.21437/Interspeech.2018-1864. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1864.

[26] A. Raikar, M. Soni, A. Panda, and S. K. Kopparapu, "Acoustic model adaptation in reverberant conditions using multi-task learned embeddings," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 145–149. DOI: 10.23919/EUSIPCO55093.2022.9909579.

[27] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5329–5333.

[28] C. Luu, P. Bell, and S. Renals, "Leveraging speaker attribute information using multi task learning for speaker verification and diarization," *arXiv preprint arXiv:2010.14269*, 2020.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[30] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[31] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 5220–5224.

[32] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, E7856–E7865, 2016.

[33] E. Camberlein and R. Lebarbenchon, "Voicehome-2 corpus-localization and speech enhancement baseline-code," 2017.

[34] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černockỳ, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[35] J. P. Escudero, V. Poblete, J. Novoa, *et al.*, "Highly-reverberant real environment database: Hrre," *arXiv preprint arXiv:1801.09651*, 2018.

[36] K. J. Piczak, *ESC: Dataset for Environmental Sound Classification*. Harvard Dataverse, 2015. DOI: 10.7910/DVN/YDEPUT. [Online]. Available: https://doi.org/10.7910/DVN/YDEPUT.

Generating Room Impulse Responses Using Neural Networks Trained with Weighted Combinations of Acoustic Parameter Loss Functions

Hualin Ren[1], Christian Ritz[1, *], Jiahong Zhao[2], Xiguang Zheng[1], Daeyoung Jang[3]
[1]School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, Australia
E-mail: hualin@uow.edu.au, E-mail: critz@uow.edu.au, E-mail: xiguang@uow.edu.au
[2]Institute of Sound and Vibration Research (ISVR), University of Southampton, Hampshire, UK
E-mail: Jiahong.Zhao@soton.ac.uk
[3]Media Coding Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea
E-mail: dyjang@etri.re.kr

***Abstract*— This paper investigates loss functions based on room acoustic parameters for training conditional generative adversarial networks (CGANs) used for generating room impulse responses (RIRs) at specific locations within a real room. The CGANs are trained on RIRs recorded at multiple positions in the room and then used to generate RIRs at new positions. The study evaluates the effectiveness of adaptive and fixed weightings, applied to various combinations of acoustic parameter-based loss functions, including those based on reverberation time (RT) and early decay time (EDT), along with the time-domain mean squared error (MSE) and frequency-domain multi-resolution short-time Fourier transform (MRSTFT). Results from reconstructing RIRs in two real rooms show that adaptive weightings are more effective than fixed weightings. MRSTFT surpasses MSE in accurately reconstructing RIRs across all frequency bands. Additionally, the loss functions using combinations of MRSTFT, RT30 and EDT with adaptive weightings significantly enhance the accuracy of RIR generation.**

I. INTRODUCTION

The room impulse response (RIR), which represents the acoustic characteristics of a room, is crucial in various audio signal processing tasks. For a realistic auditory experience in virtual reality (VR), the anechoic sound source must be continuously convolved with different RIRs as the listener's position changes within the virtual space. In streaming VR applications, providing an immersive experience in a real room with a complex shape by recording and transmitting RIRs for every possible listening position is impractical. Instead, RIRs can be modeled using the room's geometry, the reflection and absorption coefficients of its walls. However, this approach typically only produces satisfactory results for rooms with clear, simple geometries. Additionally, these methods often struggle to accurately model low frequencies and diffraction effects [1]. To address these limitations, artificial neural networks have been employed, to predict the desired data, offering a potential solution for generating RIRs in more complex environments.

Recently, generative adversarial networks (GAN) [2] based neural network approaches have been proposed for generating RIRs. GANs consist of two competing networks, a generator and a discriminator, which are trained in a back-and-forth manner to challenge each other. WaveGAN was the first attempt to generate raw-waveform audio using a GAN [3]. As an extension, a conditional generative adversarial net (CGAN) ensures that generated data match specific criteria [4]. Several CGAN-based models have been developed for RIR generation. Image2Reverb synthesizes RIRs using a CGAN with mean absolute error (MAE) loss and reverberant time 60 (RT60) losses [5]. FAST-RIR is a fast diffuse RIR generator employing mean squared error (MSE) and RT60 losses [6]. MESH2IR generates RIRs using MSE and energy decay relief losses [7].

However, the exploration of CGAN has only been partially explored in acoustics [8]. Existing RIR generators mainly focus on basic loss functions such as MAE and MSE [5], [6], with limited use of room acoustic parameters as loss functions. This study investigates the effects of some key room acoustic parameters, including RT and early decay time (EDT), on improving CGAN-based RIR generation, alongside the impact of fixed and adaptive weightings. These parameters are crucial for accurately reflecting acoustical properties of a room and enhancing the realism of the generated sound field [9], [10]. The hypothesis is that the acoustic characteristics of the RIR in a particular room can be learned, allowing for the generation of RIRs at arbitrary locations without needing detailed knowledge of the room geometry and surface reflection parameters.

Main Contributions: (1) Inclusion of acoustic parameters in loss functions. The study examines the use of key room acoustic parameters, such as RT and EDT, along with the time-domain MSE and frequency-domain multi-resolution short-time Fourier transform (MRSTFT) in loss functions for a CGAN-based network. (2) Investigation of the performance of

the fixed and adaptive weightings in the context of RIR generation. Fixed weightings remain constant throughout the training, but they might lead to imbalance weightings of loss functions. Adaptive weightings adjust with each iteration to balance the loss functions. (3) Uniquely integrating the MRSTFT with acoustic parameters in the loss functions, rather than using MRSTFT alone [11], [12], [13]. (4) Identification of the best-performing combination of all the mentioned acoustic characteristics in the loss functions for RIR generation. Evaluations compare recorded and generated RIRs using various metrics, including the MRSTFT, RT, EDT, direct-to-reverberant ratio (DRR), normalized root mean squared error (NRMSE), and DROQM for listening quality (LQ) [14].

II. CGAN-Based RIR Generation

In this study, CGAN is trained to learn how to generate RIRs based on receiver and source positions in 3-dimensional Cartesian coordinates, as well as room dimensions, represented as a 9-dimensional embedding vector ε_p. The architecture of the CGAN model builds on the FAST-RIR model [6] and the Stage-I structure of StackGAN [15], with modifications including an additional upsample layer in the generator to accommodate the size of the input signal. The generator was organized by multiple transposed convolutional layers, while the discriminator receives a pair of real and generated RIRs as the input. The performance of the network is controlled by loss functions including MSE, MRSTFT, RT and EDT.

A. Generator Adversarial Loss

Conditioned on ε_p, the generator G is trained using a modified CGAN loss to minimize the overall error between the generated RIR $\tilde{\mathbf{p}}$ and the real-world recorded RIR $\mathbf{p}$. The discriminator, denoted by D, is used to distinguish between the generated RIRs and the real RIRs.

$$\mathcal{L}_{ADV}(G) = \mathbb{E}[(D(\tilde{\mathbf{p}}) - 1)^2] \tag{1}$$

B. MSE Loss

The MSE loss computes between each sample of the real-world recorded and generated RIRs. The $\|\cdot\|_{MSE}$ is the MSE loss between real and generated RIRs.

$$\mathcal{L}_{MSE}(\mathbf{p}, \tilde{\mathbf{p}}) = \| \mathbf{p} - \tilde{\mathbf{p}} \|_{MSE} \tag{2}$$

C. MRSTFT Loss

MRSTFT is extensively used in audio processing [11], [12], [13] as it effectively captures the time-frequency distribution of realistic waveforms [12]. It also prevents the generator from overfitting to a short-time Fourier transform (STFT) representation, which could lead to suboptimal performance in the waveform domain [11]. In this study, MRSTFT is calculated using four distinct STFT resolutions: 1024, 2048, 512, and 128. The hop size is set to half of the STFT length, to allow overlapping segments during transformation. A Hann window is applied in STFT, where $\|\cdot\|_F$ denotes the F-norm, and $|\mathfrak{F}\{\cdot\}|$ is the magnitude of a STFT operator that converts time-domain sound pressure vectors $\mathbf{p}(t)$ and $\tilde{\mathbf{p}}(t)$, into spectrograms of dimension $L \times W$, with L being the number of frequency bins and W the number of frames. The MRSTFT is computed as the sum of two terms, the spectral convergence $\mathcal{L}_{\mathrm{SC}}$ and the spectral log-magnitude $\mathcal{L}_{\mathrm{SM}}$ across these R different STFT resolutions.

$$\mathcal{L}_{SC}(\mathbf{p}, \tilde{\mathbf{p}}) = \frac{\| \, |\mathfrak{F}\{\mathbf{p}\}| - |\mathfrak{F}\{\tilde{\mathbf{p}}\}| \, \|_F}{\| \, |\mathfrak{F}\{\mathbf{p}\}| \, \|_F} \tag{3}$$

$$\mathcal{L}_{SM}(\mathbf{p}, \tilde{\mathbf{p}}) = \frac{1}{N} \| \, log|\mathfrak{F}\{\mathbf{p}\}| - log|\mathfrak{F}\{\tilde{\mathbf{p}}\}| \, \|_F \tag{4}$$

$$\mathcal{L}_{MRSTFT}(\mathbf{p}, \tilde{\mathbf{p}}) = \sum_{r=1}^{R} \mathcal{L}_{SC}(\mathbf{p}, \tilde{\mathbf{p}}) + \mathcal{L}_{SM}(\mathbf{p}, \tilde{\mathbf{p}}) \tag{5}$$

D. RT Loss

The RT loss is calculated in accordance with the standard ISO 3382-1:2009. In this paper, RT30 is chosen for the RT loss since the calculated RT60 sometimes exceeded the duration of the recorded RIRs in the selected databases.

$$\mathcal{L}_{RT}(\mathbf{p}, \tilde{\mathbf{p}}) = |\mathbf{p}_{RT30} - \tilde{\mathbf{p}}_{RT30}| \tag{6}$$

E. EDT Loss

EDT is determined by fitting a straight line to the initial 10 decibels (dB) of sound decay, reflecting changes in perceived reverberance [9]. To predict reverberance ratings accurately, an average of EDT across frequency bands from 125 Hz to 2 kHz is calculated [16], following ISO 3382-1:2009 standard.

$$\mathcal{L}_{EDT}(\mathbf{p}, \tilde{\mathbf{p}}) = |\mathbf{p}_{EDT} - \tilde{\mathbf{p}}_{EDT}| \tag{7}$$

F. Total Loss Function

An additional weighted loss term is used in the generator to ensure data consistency, represented by λ_{ADV}, λ_{MSE}, λ_{MRSTFT}, λ_{RT} and λ_{EDT}, respectively. A weight of 0 means this loss is not used to find the total loss while the values of non-zero weights determine the impact of this loss on the total loss.

The accuracy of RIR generation can be enhanced by incorporating additional auxiliary losses, such as acoustic parameters, alongside primary losses MRSTFT and MSE. This paper explores fixed and adaptive weightings. Fixed weightings are set as hyperparameters, remaining unchanged throughout training. However, they are often unsuitable for auxiliary tasks, as finding the appropriate weightings may require multiple training sessions. Without these adjustments, optimization imbalances can occur, where auxiliary losses either overshadow the main loss or are too weak to be effective.

To address this issue, adaptive weightings are applied based on the MetaBalance method [17]. This approach adjusts the gradient magnitude of auxiliary losses to better align with the target loss. By flexibly modifying the gradient of an auxiliary loss, it ensures a magnitude comparable to the target loss, promoting balanced optimization across multiple losses.

$$\mathcal{L}_{\mathrm{G}} = \lambda_{ADV}\mathcal{L}_{ADV} + \lambda_{MSE}\mathcal{L}_{MSE} + \lambda_{MRSTFT}\mathcal{L}_{MRSTFT} + \lambda_{RT}\mathcal{L}_{RT} + \lambda_{EDT}\mathcal{L}_{EDT} \tag{8}$$

$$\mathcal{L}_D = \mathbb{E}\left[(D(\mathbf{p}) - 1)^2 + \left(D(\tilde{\mathbf{p}})\right)^2\right] \tag{9}$$

Table 1. The evaluation of generated RIR by different loss functions in CUBE dataset.

Task	Loss function implementation	MRSTFT↓ (dB)	RT30↓ (s)	EDT↓ (s)	DRR↓ (dB)	NRMSE↓ (/)	DROQM↑ (LQ, DS, /)
1	MSE (main loss, fixed weightings)	4.38	0.38	0.56	3.21	2.74	0.45
2	MRSTFT (main loss, fixed weightings)	3.51	**0.01**	**0.04**	2.56	2.99	0.41
3	MSE (main loss, adaptive weightings)	4.38	0.24	0.48	3.60	2.88	0.49
4	MRSTFT (main loss, adaptive weightings)	**2.48**	0.29	0.20	**2.41**	**1.72**	**0.53**
5	Baseline (fixed weightings)	3.46	0.18	0.16	2.09	2.37	**0.47**
6	MRSTFT (main loss), RT30 (adaptive weightings)	2.78	**0.02**	0.07	2.19	2.41	0.44
7	MRSTFT (main loss), EDT (adaptive weightings)	2.95	0.04	0.09	2.28	2.58	0.43
8	MRSTFT (main loss), RT30, EDT (adaptive weightings)	**2.48**	**0.02**	**0.06**	**1.90**	**2.19**	0.45

Table 2. The evaluation of generated RIR by different loss functions in MeshRIR dataset.

Task	Loss function implementation	MRSTFT↓ (dB)	RT30↓ (s)	EDT↓ (s)	DRR↓ (dB)	NRMSE↓ (/)	DROQM↑ (LQ, DS, /)
1	MSE (main loss, fixed weightings)	2.71	0.03	0.08	0.86	1.68	0.51
2	MRSTFT (main loss, fixed weightings)	1.81	**0.01**	**0.02**	1.75	2.06	0.38
3	MSE (main loss, adaptive weightings)	5.62	0.25	0.30	3.32	2.45	0.45
4	MRSTFT (main loss, adaptive weightings)	**1.56**	0.06	0.05	**0.63**	**1.61**	**0.61**
5	Baseline (fixed weightings)	2.92	0.06	0.04	**0.68**	1.69	0.56
6	MRSTFT (main loss), RT30 (adaptive weightings)	1.47	0.07	0.12	2.00	1.58	**0.68**
7	MRSTFT (main loss), EDT (adaptive weightings)	1.47	**0.04**	0.04	0.75	1.71	0.55
8	MRSTFT (main loss), RT30, EDT (adaptive weightings)	**1.39**	**0.04**	**0.03**	**0.68**	**1.57**	0.63

* The bold and underlined text indicates the better performance for Table 1 and 2. The "/" means no unit.

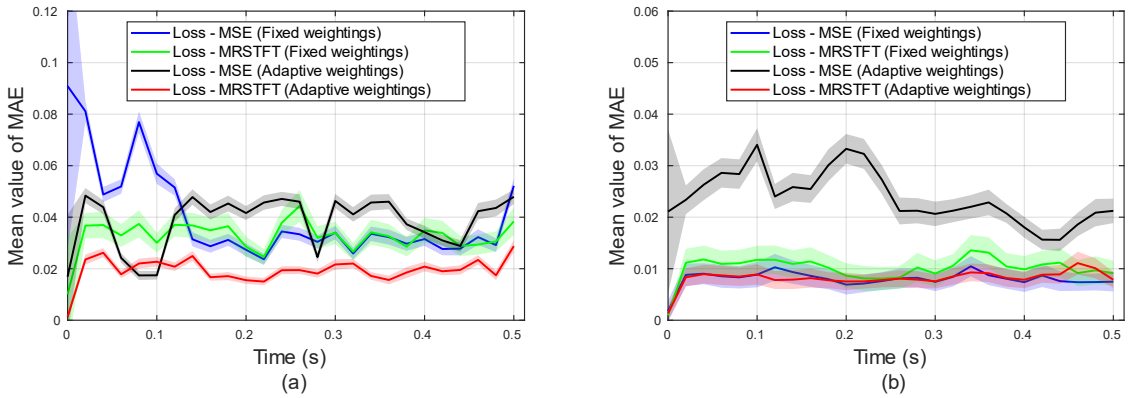


Fig. 1. Mean value of MAE with 95% confidence interval (shaded area) for fixed and adaptive weightings for MSE and MRSTFT as the main loss functions across time. (a) CUBE dataset. (b) MeshRIR dataset.

III. TRAINING

A. *Experimental Datasets*

The realistic RIR dataset CUBE was recorded in a studio measuring 10.5 m × 12 m × 5 m, with RT60 of 0.65 seconds (s) [18]. The dataset features 24 sources and 30 microphone positions, resulting in a total of 720 RIRs.

Another real recorded RIR dataset, MeshRIR, features a room with dimensions of 7 m × 6.4 m × 2.7 m and RT60 of 0.38 s [19]. For training and testing, the first, fifth, and last layers of the 3D cuboidal region (which has 9 layers) from the single source sub-dataset are selected, resulting in a total of 1323 RIRs.

Similar to [6], [13], all real-world recorded RIRs are resampled to a 16 kHz sampling rate and truncated to the first 8192 samples (0.512 s) to manage the resource limitations of the training server. Both datasets are divided into 80% for training and 20% for testing.

B. *Experimental Setup*

The CGAN model is trained on a GeForce RTX 3080Ti GPU for 800 epochs on both the MeshRIR and CUBE datasets. The optimization is performed with the RMSprop algorithm, a batch size of 128, and a learning rate of 8×10^{-5}. To minimize the impact of varying hyperparameters, all hyperparameters and network architecture are kept constant. The target output RIR

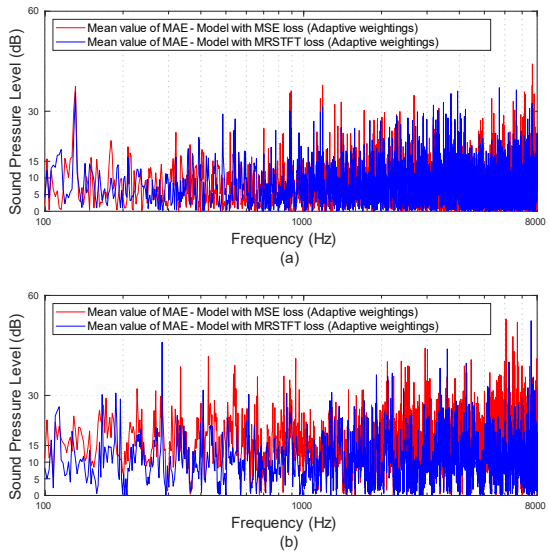


Fig. 2. Mean value of MAE of sound pressure level averaged over testing dataset for loss function MSE and MRSTFT. (a) CUBE dataset. (b) MeshRIR dataset.

is a 1D array of length 8192 samples at 16 kHz sampling rate, focusing on the direct sound (DS) and early reflection segments. The baseline model uses the loss functions from FAST-RIR [6], which include MSE and RT60 losses with fixed weightings. Eight experimental tasks are conducted to evaluate the performance of various combinations of loss functions. The tasks from 1 to 4 in both datasets (Table 1 and 2) investigate the weighting problems. The fixed weightings for the losses MSE, MRSTFT, RT30 and EDT are selected from [6] (λ_{ADV}= 1, λ_{MSE} and λ_{MRSTFT}= 0 or 20480, λ_{RT} and λ_{EDT}= 0 or 40), while adaptive weightings treat MSE and MRSTFT as main losses separately [17]. Tasks from 5 to 8 in Table 1 and 2 compare the performance of baseline, with models trained by the main loss MRSTFT and the acoustic parameters RT30 and EDT. Note that DRR was also considered as another potential loss function. However, during training, it was found that the convergence of this loss was unstable sometimes. Future work will investigate this further.

C. *Objective Evaluation*

The performance of the model is assessed using MRSTFT, RT30, EDT, DRR, NRMSE, and DROQM. DROQM is employed to evaluate the LQ (ranging from 0 to 1, where higher is better, with over 0.4 acceptable) of DS part (the most important part) of generated RIR with reference signal at the same position [14]. To analyze the effectiveness of MSE and MRSTFT in the frequency domain, the mean value of MAE of sound pressure level across the primary frequency range is calculated, as illustrated in Fig. 2, Table 3 and 4. In the time domain, the mean value of MAE across the entire signal is computed over short time windows of 20 milliseconds (ms), as

Table 3. Mean value of MAE of different frequency ranges averaged over testing datasets for loss function MSE and MRSTFT in CUBE dataset.

Frequency bands (Hz)	Loss function implementation	Mean value of MAE (dB)
100-500	MSE	**7.05**
	MRSTFT	7.16
501-1000	MSE	8.26
	MRSTFT	**7.57**
1001-4000	MSE	7.69
	MRSTFT	**7.00**
4000-8000	MSE	7.66
	MRSTFT	**7.06**

Table 4. Mean value of MAE of different frequency ranges averaged over testing datasets for loss function MSE and MRSTFT in MeshRIR dataset.

Frequency bands (Hz)	Loss function implementation	Mean value of MAE (dB)
100-500	MSE	15.84
	MRSTFT	**9.74**
501-1000	MSE	12.58
	MRSTFT	**10.16**
1001-4000	MSE	15.83
	MRSTFT	**10.63**
4000-8000	MSE	15.49
	MRSTFT	**10.59**

* The bold and underlined text indicates the better performance for Table 3 and 4.

shown in Fig. 1 and 3. Equation 10 details the MAE calculation, where $\mathbf{p}_m(k)$ denotes the measured sound pressure from M reference RIRs, and $\tilde{\mathbf{p}}_m(k)$ represents the estimated sound pressure of generated RIRs at various frequency points k.

$$Error_{MAE}(k) = \frac{1}{M}\sum_{m=1}^{M} |\mathbf{p}_m(k) - \tilde{\mathbf{p}}_m(k)| \quad (10)$$

IV. Results

A. *Fixed and Adaptive Weightings*

The findings demonstrate that adaptive weightings outperform fixed weightings. When using MRSTFT as the main loss function, adaptive weightings consistently achieve the lowest errors in MRSTFT, DRR, and NRMSE, as well as the highest DROQM scores across tasks from 1 to 4 in both datasets, as shown in Table 1 and 2. For the task focused on minimizing RT30 and EDT errors, the model with MRSTFT as the main loss function and fixed weightings proves more effective. These patterns are also evident in Fig. 1, where the model with MRSTFT loss and adaptive weightings (red line) generally outperforms others, achieving the lowest mean value of MAE. Although the MSE model with fixed weightings (blue line) shows the highest mean value of MAE from 0 to 0.1 s in Fig. 1(a), it performs comparably to the MRSTFT model with adaptive weightings in Fig. 1(b). This means that fixed weightings with MSE loss is not suitable for more reverberant room.

B. *MSE and MRSTFT Performance Comparison*

In tasks from 1 to 4, as presented in Table 1 and 2, models utilizing MRSTFT as the main loss function, whether with

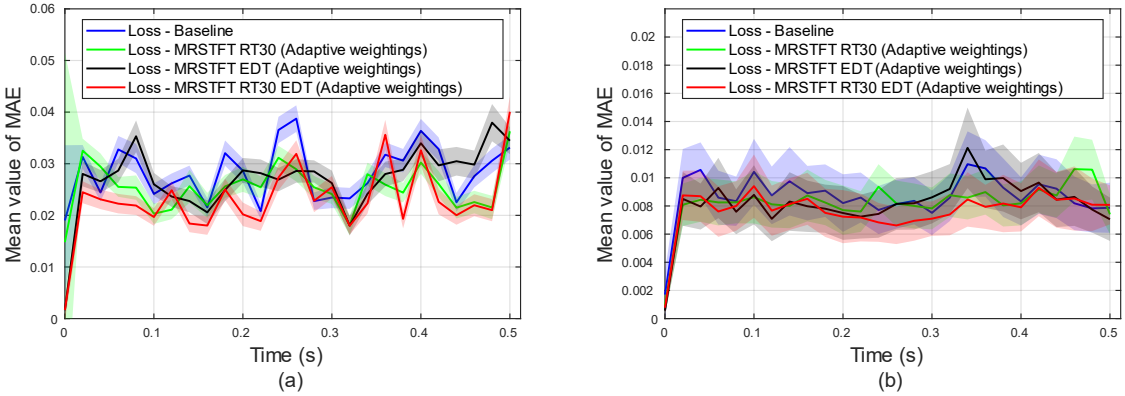


Fig. 3. Mean value of MAE with 95% confidence interval (shaded area) for different combinations of loss functions across time. (a) CUBE dataset. (b) MeshRIR dataset.

adaptive or fixed weightings, consistently achieve lower errors in most objective measurements compared to models using MSE as the main loss function. Additionally, the model using MSE as the main loss function with adaptive weightings (black line) performs the worst after 0.1 s in Fig. 1(a) and consistently underperforms in Fig. 1(b). As shown in Fig. 2(a) and 2(b), Table 3 and 4, the MRSTFT loss-based model demonstrates lower and more consistent mean value of MAE across all frequency bands. In contrast, the MSE loss-based model shows a higher and a wider range of mean value of MAE, as indicated in Table 3.

C. Baseline and Other Models Comparison

In Table 1 and 2, the model with MRSTFT, RT30, and EDT loss functions achieves the best performance across multiple objective measurements, including MRSTFT, RT30, EDT, DRR, and NRMSE. This combination outperforms both the baseline model and models using other loss functions. Specifically, the model with MRSTFT and RT30 loss functions results in the lowest RT30 error from Task 5 to 8 in Table 1, while the model with MRSTFT and EDT loss functions achieves the lowest RT30 error from the same tasks in Table 2. Considering the DS part simulation, the DROQM score is similar in more reverberant room in Table 1(CUBE dataset). However, in less reverberant room (Table 2, MeshRIR dataset), the models combining MRSTFT and RT30 losses, as well as those combining MRSTFT, RT30, and EDT losses, show better performance (top 2 scores in Table 1 and 2) in the DROQM metric. Fig. 3 supports these findings. The model with MRSTFT, RT30, and EDT losses (represented by the red line) consistently reaches the lowest error points, whereas the baseline model displays higher errors in most cases. Additionally, RIRs generated by models in the less reverberant room generally achieve lower MRSTFT, EDT, DRR, and NRMSE errors, along with higher DROQM scores, as shown from Task 5 to 8 in Table 1 and 2, and display more stability in Fig. 3(a) and 3(b), indicating more accurate generation in the less reverberant environment.

V. Discussion

The MRSTFT loss function consistently outperforms the MSE loss function, demonstrating superior performance in both frequency domain (as indicated by MRSTFT metric) and time domain (as shown by NRMSE metric). Specifically, the MRSTFT enhances the recovery of RIR energy across all frequency bands, particularly at higher frequencies. Additionally, incorporating various combinations of acoustic parameters in the loss functions significantly improves the performance compared to that of the baseline model. In experiments using both datasets, the loss function combining MRSTFT, RT30, and EDT losses with adaptive weightings demonstrates better performance, achieving the lowest errors in most cases. The effectiveness of the room acoustics parameters varies with different room conditions. In the more reverberant room, the differences in errors are more observable, with variations sometimes reaching a factor of ten, making the impact of different combinations of the loss functions more detectable. Conversely, in the less reverberant room, the errors are closer to each other. Overall, RIR generation performs more accurately in the less reverberant room.

VI. Conclusions

This study evaluates the effectiveness of fixed and adaptive weightings, compares the performance of MSE and MRSTFT, and explores various combinations of loss functions with key acoustic parameters in CGAN-based RIR generators, focusing on their ability to accurately synthesize RIRs. The findings demonstrate that adaptive weightings outperform fixed weightings. The MRSTFT shows significant advantages over the MSE, enhancing the performance of RIR generation in both the frequency and time domains. Notably, highly accurate results are obtained by incorporating adaptive weightings with the MRSTFT, acoustic parameters RT30 and EDT as loss functions. This innovative integration offers a new perspective on improving RIR generation. Further explorations of loss functions could lead to even more precise synthesized RIRs and then their applications such as higher-order Ambisonics.

VII. ACKNOWLEDGMENT

This research was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government, grant number 23ZH1200 (The research of the basic media contents technologies).

REFERENCES

[1] M. Tamulionis, T. Sledevič, and A. Serackis, "Investigation of Machine Learning Model Flexibility for Automatic Application of Reverberation Effect on Audio Signal," *Applied Sciences*, vol. 13, no. 9, Art. no. 9, Jan. 2023, doi: 10.3390/app13095604.

[2] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, in NIPS'14. Cambridge, MA, USA: MIT Press, Dec. 2014, pp. 2672–2680.

[3] C. Donahue, J. McAuley, and M. Puckette, "Adversarial Audio Synthesis," presented at the International Conference on Learning Representations, Feb. 2018.

[4] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv.org. Accessed: Sep. 02, 2023. [Online]. Available: https://arxiv.org/abs/1411.1784v1

[5] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 286–295. doi: 10.1109/ICCV48922.2021.00035.

[6] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "Fast-Rir: Fast Neural Diffuse Room Impulse Response Generator," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 571–575. doi: 10.1109/ICASSP43922.2022.9747846.

[7] A. Ratnarajah, Z. Tang, R. Aralikatti, and D. Manocha, "MESH2IR: Neural Acoustic Impulse Response Generator for Complex 3D Scenes," in *Proceedings of the 30th ACM International Conference on Multimedia*, Lisboa Portugal: ACM, Oct. 2022, pp. 924–933. doi: 10.1145/3503161.3548253.

[8] P. Gerstoft, H. Groll, and C. F. Mecklenbräuker, "Parametric Bootstrapping of Array Data with A Generative Adversarial Network," in *2020 IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Jun. 2020, pp. 1–5. doi: 10.1109/SAM48682.2020.9104371.

[9] J. S. Bradley, "Review of objective room acoustics measures and future needs," *Applied Acoustics*, vol. 72, no. 10, pp. 713–720, Oct. 2011, doi: 10.1016/j.apacoust.2011.04.004.

[10] S. Saini, I. Engel, and J. Peissig, "An end-to-end approach for blindly rendering a virtual sound source in an audio augmented reality environment," *J AUDIO SPEECH MUSIC PROC.*, vol.2024, no. 1, p. 16, Mar. 2024, doi: 10.1186/s13636-024-00338-6

[11] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6199–6203. doi: 10.1109/ICASSP40776.2020.9053795.

[12] S. Joshi *et al.*, "Defense against Adversarial Attacks on Hybrid Speech Recognition System using Adversarial Fine-tuning with Denoiser," Sep. 2022, pp. 5035–5039. doi: 10.21437/Interspeech.2022-10977.

[13] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft, "Generative models for sound field reconstruction," *The Journal of the Acoustical Society of America*, vol. 153, no. 2, pp. 1179–1190, Feb. 2023, doi: 10.1121/10.0016896.

[14] H. Ren, C. Ritz, J. Zhao, and D. Jang, "Towards an Objective Quality Metric for Interpolated Directional Room Impulse Responses," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 8205–8209. doi: 10.1109/ICASSP48485.2024.10446507.

[15] H. Zhang *et al.*, "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5908–5916. doi: 10.1109/ICCV.2017.629.

[16] M. Barron, "Subjective Study of British Symphony Concert Halls," *Acta Acustica united with Acustica*, vol. 66, no. 1, pp. 1–14, Jun. 1988.

[17] Y. He, X. Feng, C. Cheng, G. Ji, Y. Guo, and J. Caverlee, "MetaBalance: Improving Multi-Task Recommendations via Adapting Gradient Magnitudes of Auxiliary Tasks," in *Proceedings of the ACM Web Conference 2022*, in WWW '22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 2205–2215. doi: 10.1145/3485447.3512093.

[18] K. Müller, "CUBE B-format RIR dataset (Soundfield ST450 MKII)". Available: https://phaidra.kug.ac.at/o:104435

[19] S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnström, "MESHRIR: A Dataset of Room Impulse Responses on Meshed Grid Points for Evaluating Sound Field Analysis and Synthesis Methods," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2021, pp. 1–5. doi: 10.1109/WASPAA52581.2021.9632672.

Investigating the Language Independence of Voice Activity Projection Models through Standardization of Speech Segmentation Labels

Yuki Sato* Yuya Chiba† Ryuichiro Higashinaka*
* Graduate School of Informatics, Nagoya University, Japan
sato.yuki.y1@s.mail.nagoya-u.ac.jp, higashinaka@i.nagoya-u.ac.jp
† NTT Communication Science Laboratories, NTT Corporation, Japan
yuya.chiba@ntt.com

***Abstract*—Voice Activity Projection (VAP) is crucial for ensuring natural turn-taking in dialogue systems. One known limitation in VAP is that, according to an experiment in a previous study, a model trained in one language cannot be applied to other languages. However, in that study, specific languages were associated with specific datasets, making it unclear whether the performance of the VAP model was influenced by the language or by the dataset. In this study, utilizing the Transformer-based VAP model, we trained and evaluated the model using multiple datasets across several languages in order to determine what is affecting the performance. The results indicated that the performance of the VAP model is more dependent on the dataset than on the language. Specifically, differences in how speech segmentation labels are provided significantly impacted the performance. The results also showed that a model trained in one language can be applied to other languages, suggesting that the VAP model captures common features independent of language.**

I. Introduction

In recent years, the naturalness of responses generated by dialogue systems has significantly improved owning to large language models (LLMs) [1], [2]. However, when it comes to multimodal interaction, it has also become evident that the quality of dialogues is still insufficient, leading to a growing demand for systems that can provide more natural conversations [3].

In spoken dialogue, turn-taking [4] is crucial. Extensive research has investigated on how to facilitate appropriate turn-taking in spoken dialogue systems [5]–[8]. Among others, the Voice Activity Projection (VAP) model [9], which can predict voice activity based on acoustic information and forecast future turn-taking events, has been attracting attention for its potentially wide applicability and generalizability. Further research has been exploring how to extend the VAP model and apply it to actual dialogue systems [10]–[13].

Most work on the VAP model has been focused on English, but for the VAP model to be more widely utilized, it needs to be able to cope with other languages. To this end, several studies have applied the VAP model to languages other than English and to multiple datasets [11], [12], [14]. One such study pointed out that a VAP model trained on a single language cannot be directly applied to another language [15], which severely limits the model's applicability. However, in that study, specific languages were associated with specific datasets, making it unclear whether the performance of the VAP model was influenced by the language or by the dataset. In particular, criteria for speech segmentation labeling were different among the datasets used, which could potentially be affecting the VAP performance more than the languages.

Therefore, in this study, we prepare at least two different datasets each for multiple languages, that is, English, Japanese, and Chinese, and then train and evaluate the VAP models. We also use Voice Activity Detection (VAD) to standardize the criteria for speech segmentation labeling across these datasets and investigate how this standardization affects the behavior of the VAP model. The contributions of this research are as follows:

1) We evaluated the performance of the VAP model using multiple languages and types of datasets, showing that the VAP performance depends more on the dataset than on the language.
2) We demonstrated that standardizing the criteria for speech segmentation labels could enable proper evaluation of the VAP model.
3) Experimental results indicated that the VAP model trained on a specific language can be applied to other languages, suggesting that the VAP model captures common features independent of language.

II. Voice Activity Projection Model

As a preliminary, we describe the VAP model proposed by Ekstedt and Skantze [9], on which our experiment will be based. The VAP model predicts the future voice activity of two speakers based on audio input. Voice Activity (VA) refers to the presence or absence of speech in a given interval, regardless of the content.

A. Model architecture

The VAP model is based on Cross-attention Transformer, as shown in Fig. 1. This model takes the past 20 seconds of speech from both speakers as input and predicts the voice

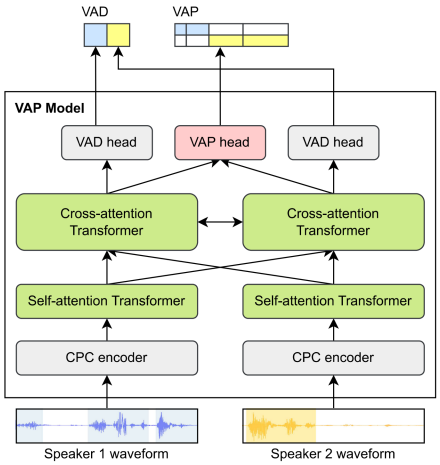


Fig. 1. Voice activity projection model based on Cross-attention Transformer for two-speaker voice input.

activity of the two speakers for the next two seconds. Following previous studies [9], [15], the Transformer has a hidden layer size of 256, one Self-attention Transformer layer, three Cross-attention Transformer layers, and four attention heads. The VAP model used in this study is publicly available at https://github.com/ErikEkstedt/VAP.

The VAP model ultimately generates two outputs on the basis of multitask learning. The first task is the primary objective, VAP, which outputs probabilities corresponding to 256 classes (described in the next subsection). The second task is VAD, which detects the current VA of the two speakers. Since VAP depends on the current VA, setting VAD as a subtask stabilizes VAP learning.

B. Problem setting

The model outputs a single prediction for the voice activity pattern of both speakers over two seconds. Specifically, voice activity is divided into four bins for each speaker, totaling eight bins. These bins correspond to future intervals of 0–200 ms, 200–600 ms, 600–1200 ms, and 1200–2000 ms for each speaker's speech and are discretized into "speech" or "no speech." Consequently, the number of possible output combinations is 256 (2^8), making the VA of the two speakers a 256-class classification problem.

C. Loss function

The VAP model outputs $p_{vap}(y)$, the probability corresponding to the 256 VAP classes indexed by $y \in (1, ..., 256)$, and $p_{vad}(s)$, the probability that speaker s's voice activity is present at the current time. The cross-entropy loss for the true labels is calculated as

$$L_{vap} = -\log p_{vap}(y),$$

$$L_{vad} = -\sum_{s=1}^{2}\{v_s \log p_{vad}(s) + (1 - v_s)\log(1 - p_{vad}(s))\},$$

TABLE I
BREAKDOWN OF DATASETS (IN HOURS)

Dataset	Language	Train	Val.	Test
SWBD	ENG	65.2	38.4	13.0
HKUST	ZHO	65.2	14.4	14.6
Travel	JPN	65.2	8.1	8.4
Call ENG	ENG	32.6	7.3	2.3
Call ZHO	ZHO	32.6	4.4	1.3
Call JPN	JPN	32.6	4.2	1.8
MDT	ZHO	–	–	10.2
CEJC-phone	JPN	–	–	2.6

where $v_s \in (0, 1)$ represents the voice activity of speaker s (1 for speech, 0 for no speech). Finally, the combined loss of VAP and VAD is summed to form the final loss function, as

$$L = L_{vap} + L_{vad}. \quad (1)$$

D. Evaluation metric for VAP

A previous study [9] proposed various metrics, such as hold-shift and shift-prediction, to evaluate the prediction accuracy of specific turn-taking events. However, as these metrics assess the prediction accuracy of specific turn-taking events, they cannot be used to evaluate the overall performance of the VAP model. Since L_{vap} can represent the overall accuracy of VAP and is not limited to specific turn-taking events, in this study, we use L_{vap} as the evaluation metric for the model.

III. DATASETS AND SPEECH SEGMENTATION LABELS

A. Datasets

For our experiments, we utilize various spoken dialogue datasets for training and evaluation, including those used in [15]. Specifically, we use the Switchboard Corpus (SWBD), HKUST Mandarin Telephone Speech (HKUST), and Travel Agency Task Dialogues (Travel), as well as CALLHOME (Call), ASR-MultiDeviCCSC (MDT), and the Corpus of Everyday Japanese Conversation (CEJC).

In previous studies [9], [15], all audio sources were resampled to 16 kHz. However, the datasets we use in the current study include recordings at 8 kHz, 16 kHz, and 32 kHz, so to standardize conditions across datasets, we resampled all audio sources to 8 kHz.

The amount of data varies between datasets. To address this, for training data, the three primary datasets (SWBD, HKUST, and Travel) are divided into training, validation, and test subsets with an 8:1:1 ratio. Then, the amount of training data is adjusted to the size of the smallest one (i.e., Travel). In contrast, for validation and test data, differences in data quantity are not expected to significantly impact the results, so the size of the original data is maintained.

Below, we describe the six different datasets used in this study. The total hours of data and their breakdowns are summarized in Table I.

1) Switchboard Corpus (SWBD) [16]: This dataset contains English telephone conversations. It consists of 2,438 dialogues, with a total duration of approximately 259 hours. After dividing into subsets and adjusting the amount of training data, the

training set was approximately 65.2 hours, the validation set 38.4 hours, and the test set 13.0 hours.

2) HKUST Mandarin Telephone Speech (HKUST) [17]: This dataset contains Chinese telephone conversations. It consists of 873 dialogues, totaling 144.7 hours of speech. After dividing into subsets and adjusting the amount of training data, the training set was approximately 65.2 hours, the validation set 14.4 hours, and the test set 14.6 hours.

3) Travel Agency Task Dialogues (Travel) [18]: This dataset contains simulated customer service dialogues conducted in Japanese through Zoom. Specifically, it involves role-played interactions where the clerk, experienced in customer service, engages in tourism consultation at a travel agency. A notable feature of this dataset is that the customers range from children to the elderly, representing a broad age spectrum. It comprises 330 dialogues with a total duration of 115.5 hours. Certain dialogues were excluded because of different audio lengths for the interlocutors, resulting in 233 dialogues and approximately 82.1 hours of data used. After dividing into subsets, the training set amounted to about 65.2 hours, the validation set to 8.1 hours, and the test set to 8.4 hours.

4) CALLHOME (Call) [19]: This multilingual dialogue dataset comprises conversations in English, Chinese, Japanese, German, and Spanish. Each language subset consists of casual conversations initiated by native speakers residing in North America calling acquaintances abroad, typically in their home countries. For this study, we utilize the English (ENG), Chinese (ZHO), and Japanese (JPN) subsets.

In this dataset, transcriptions and speech segmentation labels are provided for certain intervals of each recording (approximately 10 minutes), and some recordings lack speech segmentation labels altogether. The limited transcribed portions are insufficient for training VAP models, so we use this dataset only for test, although for the experiments using automatic speech segmentation labels (described in the next subsection), we also use it for training. In this study, 10% of the recordings with speech segmentation labels were randomly selected as the test set. In the test set, only the intervals with speech segmentation labels were used for both the experiments using the original labels and the experiments using the automatic segmentation labels.

For the experiments using automatic speech segmentation labels, the remaining recordings (which include those with and without speech segmentation labels) were randomly divided into training and validation sets in an 8:1 ratio. In the training and validation sets, all intervals of each recording were utilized. The amount of training data was adjusted to match the smallest subset, which was Chinese. Consequently, each language (ENG, ZHO, JPN) had approximately 32.6 hours of training set. The validation and test sets were approximately 7.3 hours and 2.3 hours for ENG, 4.4 hours and 1.3 hours for ZHO, and 4.2 hours and 1.8 hours for JPN, respectively.

5) ASR-MultiDeviCCSC (MDT): This dataset[1] contains Chinese telephone conversations. It consists of 59 dialogues with a total duration of 10.2 hours. Due to its small size, this dataset is insufficient for training and is used only for test.

6) Corpus of Everyday Japanese Conversation (CEJC-phone)[20]: The CEJC contains naturally occurring Japanese conversations in everyday situations. In this study, for comparison with other datasets collected via telephone, we used the portion of the data of telephone conversations. The data includes eight dialogues totaling approximately 2.6 hours of telephone conversations. We use this data only for test.

B. Speech segmentation labels

The datasets above have their original speech segmentation labels; however, for the purpose of investigating their effects on the VAP model, we also labeled them with automatic (standardized) segmentation labels using WebRTC VAD[2]. Therefore, we have two types of the above datasets, with the Original label and the WebRTC VAD label, as follows:

Original label We use the speech segmentation labels provided with each dataset.

WebRTC VAD label We perform speech segmentation detection uniformly across all datasets using WebRTC VAD, ensuring consistent labeling criteria. The procedure for annotating speech segmentation labels is as follows:

1) VAD is conducted every 30 ms for audio signals.
2) Sequential processing is conducted frame by frame. If more than 90% of the past ten frames are classified as speech, a speech segment starts from the beginning of that identified range.
3) Within a speech segment, if more than 90% of the past ten frames are classified as non-speech, the segment ends at the beginning of that identified range.
4) The obtained segments are labeled as speech segments, with their start and end times annotated as speech segmentation labels.

IV. Experiments

To investigate whether the performance of the VAP model is dependent on language or dataset, we first evaluate the model trained using a method similar to that in [15] with datasets different from the training data. Next, to examine the impact of differences in speech segmentation labeling criteria, we train and evaluate the VAP model using both Original labels and WebRTC VAD labels, and compare the evaluation results. The details of the experiments are described below.

A. Training conditions for the VAP model

Training was conducted with a dropout rate of 0.1, early stopping criteria of 10 epochs, AdamW optimization, a learning rate of 3.63×10^{-4}, and a batch size of 4. During training, as stated in Eq. (1), the combined loss of VAP and VAD was used as the loss, and the model with the smallest loss on the validation data was considered the best model. These experimental settings are identical to those in the previous study [9]. Each model was trained with seeds ranging from

[1] https://magichub.com

[2] https://webrtc.org

TABLE II
EVALUATION OF THE ORIGINAL LABEL-TRAINED MODEL (L_{vap})

Training data	Test data							
	SWBD	HKUST	Travel	Call ENG	Call ZHO	Call JPN	MDT	CEJC-phone
SWBD	**2.57**	3.69	3.00	3.04	3.83	3.91	**2.66**	**4.01**
HKUST	4.95	**2.16**	3.45	3.16	3.26	4.22	3.16	5.63
Travel	4.31	3.22	**2.26**	**2.86**	**3.11**	**3.60**	2.95	4.62
Avg.	3.94	3.02	2.90	3.02	3.40	3.91	2.92	4.75

TABLE III
EVALUATION OF THE WEBRTC VAD LABEL-TRAINED MODEL USING ORIGINAL LABEL (L_{vap})

Training data	Test data							
	SWBD	HKUST	Travel	Call ENG	Call ZHO	Call JPN	MDT	CEJC-phone
SWBD	3.33	3.49	3.08	2.94	**3.58**	**3.57**	2.70	4.55
HKUST	**3.13**	3.46	2.92	**3.00**	3.58	3.79	**2.65**	**4.11**
Travel	3.76	**3.44**	**2.63**	3.17	3.60	3.64	2.93	4.34
Avg.	3.41	3.46	2.88	3.03	3.59	3.67	2.76	4.33

0 to 4, and the final evaluation value was the average of the test VAP loss of five seed conditions.

B. Comparison models

To investigate the dependence of the VAP model on language or dataset, as well as the impact of the criteria for speech segmentation labeling, we prepared two kinds of models: (a) Original label-trained model, for which training is conducted using the speech segmentation labels provided with each dataset, and (b) WebRTC VAD label-trained model, for which training is conducted using the speech segmentation labels assigned by WebRTC VAD.

Note that the Original label-trained model is trained on each of the three datasets (SWBD, HKUST, and Travel), while the WebRTC VAD label-trained model is trained on these three datasets plus those of three languages in the Call dataset.

V. RESULTS

A. Evaluation of the Original label-trained model

To investigate whether the VAP model is dependent on language or dataset, we evaluated three models trained using the same method as in [15] with datasets different from the training data. Specifically, models were prepared by training on the SWBD, HKUST, and Travel datasets using the Original label. As test data, we utilized the datasets used in training as well as Call (ENG, ZHO, JPN), MDT, and CEJC-phone using the Original label. Assuming that the VAP model learns the differences of language, we expect that the loss for a certain test dataset will be lower for the model trained in the same language compared to those trained in other languages.

The results are listed in Table II. As also demonstrated in [15], the loss was particularly low when the training data and test data were from the same dataset. In contrast, for other test data, there was no noticeable tendency for the loss to be lower when the training data and test data were in the same language. In addition, Call JPN and CEJC-phone are both Japanese telephone datasets and are thus expected to have similar characteristics. However, while the model trained on Travel performed best on Call JPN, the model trained on SWBD performed best on CEJC-phone. From these results, we can conclude that the VAP model is probably learning differences in datasets rather than differences in language.

B. Evaluation of the WebRTC VAD label-trained model using Original label

One of the important factors in determining the characteristics of spoken language datasets is the criteria for speech segmentation labeling, the difference of which could significantly impact the performance of the VAP model. Therefore, to investigate the impact of speech segmentation labels in the training data on the output of the VAP model, we evaluated the WebRTC VAD label-trained model. Specifically, models were trained on the SWBD, HKUST, and Travel datasets using the WebRTC VAD label, and then we evaluated the models with the datasets having the Original label.

The results are listed in Table III. When comparing the increase and decrease in loss between Tables II and III for each combination of training data and test data, we observe cases where the loss increases or decreases depending on the combination. For example, in the evaluation using SWBD, the loss increases with the model trained on SWBD, while it decreases with the models trained on HKUST and Travel. This indicates that the criteria of the WebRTC VAD label have become further from those of the Original label for SWBD, and that the criteria of the WebRTC VAD label have become closer to those of the Original label for SWBD than to the criteria of the Original label for HKUST or Travel. Additionally, focusing on the evaluation with Call datasets, the loss of the model trained on SWBD decreases across all three Call languages, while the loss of the model trained on Travel increases across all three languages. This suggests that the criteria for the Original labels in the Call datasets differ from the Original labels in SWBD (i.e., WebRTC labels are more similar) but are similar to those in the Original labels in Travel (i.e., WebRTC labels are not as similar). These results demonstrate that the criteria of the Original labels vary across datasets, and that the output of the VAP model is influenced by these differences.

TABLE IV
EVALUATION OF THE WEBRTC VAD LABEL-TRAINED MODEL USING WEBRTC VAD LABEL (L_{vap})

Training data	Test data							
	SWBD	HKUST	Travel	Call ENG	Call ZHO	Call JPN	MDT	CEJC-phone
SWBD	**3.03**	3.05	3.01	**2.89**	3.16	**3.35**	2.70	**3.47**
HKUST	3.60	**2.71**	2.84	3.00	**3.12**	3.42	**2.66**	3.56
Travel	3.98	3.52	**2.39**	3.18	3.35	3.41	2.92	3.84
Avg.	3.54	3.10	2.75	3.03	3.21	3.39	2.76	3.63

TABLE V
EVALUATION OF THE WEBRTC VAD LABEL-TRAINED MODEL USING WEBRTC VAD LABEL IN CALL (L_{vap})

Training data	Test data							
	SWBD	HKUST	Travel	Call ENG	Call ZHO	Call JPN	MDT	CEJC-phone
Call ENG	3.84	3.21	3.00	**2.79**	3.08	3.23	3.09	3.49
Call ZHO	4.04	3.55	3.11	2.84	**2.94**	3.20	3.05	**3.47**
Call JPN	**3.68**	**3.06**	**2.89**	2.88	3.09	**3.06**	**2.85**	3.68
Avg.	3.85	3.27	3.00	2.84	3.03	3.16	3.00	3.55

C. Evaluation of the WebRTC VAD label-trained model using WebRTC VAD labels

To investigate the impact of standardized speech segmentation labels, we trained models with the WebRTC VAD label for SWBD, HKUST, and Travel, and evaluated the models using the WebRTC VAD label.

The results are listed in Table IV. Compared to Table III, the loss has decreased in many cases. While a simple numerical comparison cannot be made due to the change in test labels, it is noteworthy that the difference in loss across the test data has become smaller. For example, the loss for MDT, which had the smallest loss in Table III, has remained almost unchanged, while the loss for CEJC-phone, which had the largest loss, has significantly decreased. These changes occurred when changing the labels of the test data from the Original label to the WebRTC VAD label, which suggests that the Original labels for MDT are similar to those of the WebRTC VAD labels, whereas the Original labels for CEJC-phone are not similar to those of the WebRTC VAD labels.

Focusing on the models trained on SWBD and HKUST in Table IV, we observe that for English and Chinese test data, the loss tends to be smaller when the language of the training data matches the language of the test data. For example, in the evaluation using Call ENG, the loss of the model trained on SWBD is 2.89, while the loss of the model trained on HKUST is 3.00 and that of the model trained on Travel is 3.18. The same tendency is observed with other test data, except when Travel is used for training, that is, when the model trained with Travel is evaluated on Call JPN and CEJC-phone. As mentioned in Section V-A, Travel is a dataset collected through Zoom, and its turn-taking method is considered different from the other datasets. The difference in turn-taking methods is presumably what influenced the results over language.

As described above, by standardizing the criteria for speech segmentation labels, we obtained results that align with our intuitive expectations: when the conditions of the datasets are sufficiently similar, the learning of the VAP model is dependent on the language. Therefore, by standardizing the criteria for speech segmentation labels, it may be possible to properly evaluate the VAP model.

D. Evaluation of the WebRTC VAD label-trained model using WebRTC VAD labels in Call datasets

To further investigate the impact of the VAP model when the criteria for speech segmentation labels are standardized, we evaluated the WebRTC VAD label-trained model for each of the three languages in Call using the WebRTC VAD label.

The results are listed in Table V. In the evaluation using Call, the loss is smallest when the training data and test data are in the same language. However, even when the training and test data are in different languages, the difference in loss compared to when they are in the same language is sufficiently small. For example, in the evaluation using Call ENG, the model trained on Call ENG shows a loss of 2.79, while the model trained on Call JPN, which has the highest loss, shows a loss of 2.88. The difference is 0.09 at most. These results indicate that if the features such as labels and recording methods are sufficiently similar, a model trained in one language can be applied to other languages as well. This leads us to conclude that the VAP can captures features that are common across languages.

Furthermore, when comparing the evaluation of Call with the models in Table IV, we can see that the models trained on Call data have lower losses across all languages compared to the models trained on other datasets. Specifically, SWBD and HKUST are larger datasets involving telephone conversations, and one might expect models trained on them to perform better than those trained on Call. However, the models trained on Call in different languages show better results. This also suggests that there are similarities based on the dataset beyond the criteria for speech segmentation labels.

VI. CONCLUSION

In this study, we conducted experiments using multiple languages and datasets to clarify whether the Voice Activity Projection (VAP) model is dependent on language or dataset. Additionally, we investigated the impact of differences in the

criteria for speech segmentation labels on the VAP model by using standardized speech segmentation labels.

Our findings showed that the evaluation values depend on the dataset rather than on the language, arising from the differences in speech segmentation labels associated with the datasets. This suggests that standardizing the criteria for speech segmentation labels could enable proper evaluation of the VAP model. In addition, the results indicated that when the conditions of the datasets are sufficiently similar, the learning of the VAP model is dependent on the language. Our experiments also indicated that a model trained on a specific language can be applied to other languages, suggesting that the VAP model captures common features independent of language.

Our experiments clarified that there are similarities based on the dataset beyond the criteria for speech segmentation labels that influence the evaluation value. These similarities are presumably related to factors such as data collection and recording conditions, but the specific factors remain unidentified. To create a more general-purpose model in the future, it is necessary to establish methods to reduce the influence of dataset-dependent conditions. Additionally, the datasets used in this experiment were limited to three languages: English, Chinese, and Japanese. To demonstrate that there are universal features of turn-taking common to all languages, further experiments using datasets from a wider variety of languages are required. Moreover, since this study utilized speech segmentation labels based on specific criteria from WebRTC VAD, it is unclear if the training and evaluation with these labels align with evaluations in real-world conversations. Thus, it is also necessary to further pursue optimal criteria for speech segmentation labels.

VII. Acknowledgment

This work was supported by JST Moonshot R&D Grant number JPMJMS2011. We used the computational resources of the "Flow" supercomputer at the Information Technology Center, Nagoya University.

References

[1] K. Shuster, J. Xu, M. Komeili, *et al.*, "BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage," *arXiv preprint arXiv:2208.03188*, 2022.

[2] V. Hudeček and O. Dusek, "Are large language models all you need for task-oriented dialogue?" In *Proc. SIGDIAL*, 2023, pp. 216–228.

[3] R. Higashinaka, T. Takahashi, M. Inaba, *et al.*, "Dialogue system live competition goes multimodal: Analyzing the effects of multimodal information in situated dialogue systems," in *Proc. IWSDS*, 2024.

[4] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: A review," *Computer Speech & Language*, vol. 67, pp. 1–26, 2021.

[5] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa, "Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems," *Information and Media Technologies*, vol. 1, no. 1, pp. 296–304, 2006.

[6] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks," in *Proc. SIGDIAL*, 2017, pp. 220–230.

[7] M. Roddy and N. Harte, "Neural generation of dialogue response timings," in *Proc. ACL*, 2020, pp. 2442–2452.

[8] R. Yahagi, Y. Chiba, T. Nose, and A. Ito, "Multimodal dialogue response timing estimation using dialogue context encoder," in *Proc. IWSDS*, 2022, pp. 133–141.

[9] E. Ekstedt and G. Skantze, "Voice activity projection: Self-supervised learning of turn-taking events," in *Proc. Interspeech*, 2022, pp. 5190–5194.

[10] K. Onishi, H. tanaka, and S. Nakamura, "Turn-taking prediction model using single speaker features," *HCG Symposium*, pp. 1–5, 2023, (in Japanese).

[11] K. Onishi, H. Tanaka, and S. Nakamura, "Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation," in *Proc. HAI*, 2023, pp. 13–21.

[12] K. Inoue, B. Jiang, E. Ekstedt, T. Kawahara, and G. Skantze, "Real-time and continuous turn-taking prediction using voice activity projection," in *Proc. IWSDS*, 2024.

[13] Y. Chiba, K. Mitsuda, A. Lee, and R. Higashinaka, "The Remdis toolkit: Building advanced real-time multimodal dialogue systems with incremental processing and large language models," in *Proc. IWSDS*, 2024.

[14] Y. Sato, Y. Chiba, and R. Higashinaka, "Training and evaluating voice activity projection models using multiple japanese datasets," in *Proc. O-COCOSDA*, 2024.

[15] K. Inoue, B. Jiang, E. Ekstedt, T. Kawahara, and G. Skantze, "Multilingual turn-taking prediction using voice activity projection," in *Proc. LREC-COLING*, 2024, pp. 11 873–11 883.

[16] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP*, vol. 1, 1992, pp. 517–520.

[17] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, "HKUST/MTS: A very large scale mandarin telephone speech corpus," in *Proc. ISCSLP*, Springer-Verlag, 2006, pp. 724–735.

[18] M. Inaba, Y. Chiba, Z. Qi, *et al.*, "Travel agency task dialogue corpus: A multimodal dataset with age-diverse speakers," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2024.

[19] B. Wheatley, M. Kaneko, and M. Kobayashi, *CALLHOME Japanese Speech, LDC96S37, Linguistic Data Consortium*, 1996.

[20] H. Koiso, H. Amatani, Y. Den, *et al.*, "Design and evaluation of the corpus of everyday Japanese conversation," in *Proc. LREC*, 2022, pp. 5587–5594.

Generalized SpecAugment: Robust Online Augmentation Technique for End-to-End Automatic Speech Recognition

Meet Soni* and Ashish Panda† and Sunil Kumar Kopparapu†
* Rakuten Institute of Technology, Bangalore, India
E-mail: meetkumar.soni@rakuten.com
† TCS Research, Tata Consultancy Services, Mumbai, India
E-mail: {ashish.panda, sunilkumar.kopparapu}@tcs.com

***Abstract*—Since its introduction, SpecAugment has become a default augmentation technique in many End-to-End Automatic Speech Recognition systems. It is computationaly efficient and provides significant performance boost without increasing training time due to its online nature. Time-masking and Frequency-masking, the operations that contribute the most to the performance gain in SpecAugment, replace the time-stamp and certain frequency bands with either zero or mean value of the input features. In this paper, we propose a framework called Generalized SpecAugment (Gen-SA), where masked values can be replaced with any valid magnitude value. In our implementation of the Gen-SA, we replace the time and frequency mask values in the input Mel-Spectrum with scaled Mel-Spectrum of a white noise signal. Gen-SA has similar computational complexity as the SpecAugment while providing significant gain in robustness and uses just one additional signal for augmentation. Experiments on Librispeech, Aurora-4 and TED-LIUM datasets show that Gen-SA consistently outperforms baseline SpecAugment with similar parameters, provides better cross-dataset performance and improves robustness against the additive noise.**

I. Introduction

Data augmentation has been one of the key contributors in improving robustness of Automatic Speech Recognition (ASR) systems and making them more robust to various degradation conditions. As ASR models become large with more layers and more number of parameters, they tend to overfit even with large amount of training data. Data augmentation helps in diversifying the effective training data by applying perturbations to training utterances. Many data augmentation techniques have been proposed and studied for ASR task. We survey some of these techniques here.

One of the earliest data augmentation techniques in ASR was introduced in [1] that applies Vocal Tract Length Perturbations on training utterances. Speed Perturbation was used in [2] and has become an integral part of modern ASR systems. Synthesized noisy speech was used in [3]–[5] by adding noise signals to training utterances to improve noise robustness of ASR system. Similarly, [2], [6] uses room impulse responses to simulate conditions suitable for far-field and reverberant ASR. TTS system are used to synthetically generate speech from text in [7], [8] to train the ASR system. All these techniques are offline techniques in the sense that some processing is applied on the audio signals/features and the processed audio signals/features are stored for training. These techniques are resource and time consuming. This renders them difficult to use in large-scale tasks or in resource constrained environments.

To overcome the limitations of the above approaches, many online augmentation approaches have been proposed where augmentation is applied on-the-fly on training samples. Feature dropout was used in [9] and examined extensively in [10] in the context of Multi-stream ASR system training. [11], [12] presents an approach where channels in CNNs are systematically dropped while training. SpecAugment proposed in [13] has become an effective augmentation technique that provides significant performance improvement in End-to-End ASR systems. SpecAugment consists of Time-warping, Time-Masking and Frequency-Masking operations. It is an attractive augmentation technique due to its simplicity, low computational complexity, online nature and the performance improvement it provides. The extension of SpecAugment on large-scale dataset [14] shows that SpecAugment scales well with large datasets. It has also been applied in hybrid ASR systems, e.g., in BLSTM hybrid HMM-based ASR in [4], [15], [16], and also in frame-level hybrid ASR in [17] with significant success.

In SpecAugment, the most contributing factors are the Time-Masking and the Frequency-Masking operations applied on the input Mel-spectrum. The masks are selected based on a random process as per decided policy and masked values are replaced with either 0 or mean value of the Mel-spectrum. In [18] SpecSwap was proposed, where Time and Frequency values are randomly swapped in the input Mel-spectrum. Here, mask values are swapped with different portion of the input spectrum for both Time and Frequency values. The performance of SpecSwap was comparable to, though not better than SpecAugment on medium scale dataset. However, SpecSwap demonstrates that the mask values in Mel-spectrum can be replaced by values other than 0 or mean value.

In this paper, we propose Generalized SpecAugment where Time and Frequency mask values can be replaced by any real values. In practice, the mask values can be replaced with Mel-spectrum of any signal. The signal can be a noise signal, speech

of another speaker, different utterance of the same speaker, different part of the same speaker, etc. Hence, Generalized SpecAugment may require additional signal(s) for augmentation. Such formulation makes Generalized SpecAugment a framework using which augmentation policies like SpecAugment and SpecSwap can be obtained as a special case. This paper implements a form of Generalized SpecAugment where we use white noise signal to mask certain portions of the input Mel-spectrum. We replace the masked portion of Mel-spectrum with scaled Mel-spectrum channels of a white noise signal. This makes the augmentation implementation simpler and only one additional signal is required for augmentation. By scaling the channels of white noise Mel-spectrum, more variability can be introduced in augmented Mel-spectrum with computational complexity comparable to SpecAugment. We show that this realization of Generalized SpecAugment results in better performance gain than SpecAugment using Librispeech, Aurora-4 and TED-LIUM datasets. Our experiments also show that the proposed approach is more robust against additive noise compared to SpecAugment and it performs better on cross dataset evaluations.

II. Generalized SpecAugment

A SpecAugment policy includes composition of three augmentation operations: Time-Warping, Time-Masking, and Frequency masking. We focus on Time-Masking and Frequency-Masking for our formulation. These operations can be summarized as follows:

- Time-Masking involves masking t consecutive time stamps $[t_0, t_0 + t\}$. Here, t is chosen from a uniform distribution $[0, T]$ and T is Time-masking parameter. t_0 is chosen from $[0, \tau - t]$, where τ is the length of the signal.
- Frequency-Masking involves masking f consecutive frequency channels $[f_0, f_0 + f\}$. Here, f is chosen from a uniform distribution $[0, F]$ and F is Frequency-Masking parameter. f_0 is chosen from $[0, M - f\}$, where M is the number of Mel channels.

Both these operations can be applied multiple times on Mel-spectrum based on augmentation policy. The number of Time and Frequency masks is controlled using parameter n_mask. For both Time-Masking and Frequency-Masking, the masked portion is replaced by either 0 or mean value of the Mel-spectrum. Hence, the Mel-spectrum of a signal x ($MFBE_x(t, f)$) after SpecAugment masking operations can be written as follows.

$$MFBE_x(t_0 : t_0 + t, f) = mean(MFBE_x), \quad (1)$$
$$MFBE_x(t, f_0 : f_0 + f) = mean(MFBE_x). \quad (2)$$

In practice, when the log-Mel-spectrum features are normalized the mean value becomes zero. It makes SpecAugment computationally inexpensive, easier to implement, and it does not require any additional signals. The mask values can be applied online while training and for each training epoch, the network will be presented with different versions of one input Mel-spectrum. This property makes SpecAugment very attractive to use, given the advantage of computational complexity and performance improvement.

However, in principal, the masked portions can be replaces with any valid values. Mathematically, valid values for mask can be any real number. For practical purposes the mask value can be chosen as Mel-spectrum of any signal that shares similar properties of speech signal, such as sampling rate, bit rate etc. With this modification, we come up with Generalized SpecAugment as follows:

$$MFBE_x(t_0 : t_0 + t, f) = MFBE_y(t_0 : t_0 + t, f), \quad (3)$$
$$MFBE_x(t, f_0 : f_0 + f) = MFBE_y(t, f_0 : f_0 + f), \quad (4)$$

where $MFBE_y(t, f)$ is the Mel-spectrum of an external signal $y(t)$ that is used to replace the masked values. $MFBE_y(t, f)$ can be Mel-spectrum of any signal including, but not limited to, noise signal, speech of a different speaker, different utterance of the same speaker, etc. By this modification, SpecAugment now requires additional signals for masking. At training time, $MFBE_y(t, f)$ can be selected for each example or each batch from a pre-loaded set of signals in an online manner. This increases the computational complexity slightly. However, choosing mask values other than 0 or mean values provides more variations of a training example while training the network. With $MFBE_y(t, f) = 0$, we obtain the original SpecAugment formulation.

In this work, we select $y(t)$ as a white noise signal. We scale the Mel-spectrum of white noise signal $MFBE_{wn}(t, f)$ by multiplying values of each Mel channel with a value in $0 - 1$. This scaling makes certain Mel channels have different weights randomly for each sample as follows:

$$MFBE_x(t_0 : t_0 + t, f) = MFBE_{wn}(t_0 : t_0 + t, f) * S, \quad (5)$$
$$MFBE_x(t, f_0 : f_0 + f) = MFBE_{wn}(t, f_0 : f_0 + f) * S, \quad (6)$$

where $S^{1\times M} \in [0, 1]$ and $*$ denotes element-wise multiplication of S with each frame of $MFBE_{wn}$.

Figure 1 shows the comparison of SpecAugment and Generalized SpecAugment with scaled white noise Mel-spectrum. As it can be observed, the Generalized SpecAugment approach provides more variations in input Mel-spectrum. The scaling of white noise Mel-spectrum channels ensures that for each iteration, different time stamps and frequency stamps are masked with different values. Moreover, by using white noise for augmentation, Mel-spectrum of only one additional signal is required, that can be pre-loaded for computation. This makes the computational complexity of the proposed method similar to SpecAugment. Generalized SpecAugment has all the desirable properties of SpecAugment such as ease of implementation, online nature, and low computational complexity.

Using Generalized SpecAugment in training of an Acoustic Model (AM) can provide better generalization for unseen test utterances. SpecAugment forces AM to underfit the training

examples by introducing degradation to the input features. Generalized SpecAugment increases the degradation amount by many-fold due to its large span of values. Moreover, using white noise signal can provide some level of noise robustness as well. Here, masking the values is identical to adding band-limited white noise as done in [5] with $-\infty$ dB Signal-to-Noise Ratio (SNR) in masked portion and ∞ dB SNR elsewhere. We study generalization as well as noise robustness of Generalized SpecAugment with scaled white noise features in following Section.

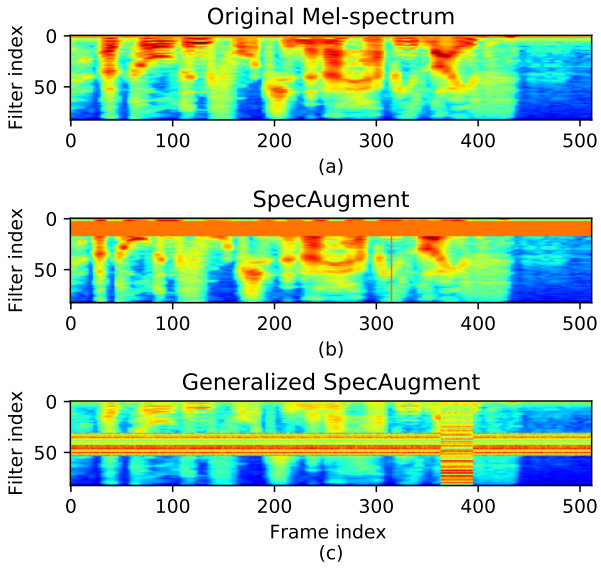


Fig. 1. (a) Mel-spectrum of an utterance, Mel-Spectrum after applying (b) SpecAugment, and (c) Generalized SpecAugment with scaled white noise.

III. Experiments

A. Dataset and ASR system

We perform all experiments on Librispeech [19] dataset using ESPNet [20] toolkit to develop ASR systems. We use Librispeech 960h train set to train Transformer Acoustic Models (AM) in ESPNet using single NVidia RTX 2080 Ti GPU. The AM is trained using ESPNet Librispeech recipe with following modifications to use single GPU:

- Transformer AM [21], [22] with 12 encoder and 6 decoder layers is used. 4 attention heads are used in each layer with 256 attention dimension.
- The model is trained using 80 dimensional Mel-filterbank energies and pitch features extracted using Kaldi [23]. Output of the network is 5000 sub-word units extracted using Sentencepiece [24] library. We *do not* use 3-way speed perturbation, but we expect that using speed perturbation will further improve performance of the system.
- We use default configuration parameters in ESPNet transformer model[1] available with librispeech recipe and modify it for single GPU. We increase batch size to 64, and set accum-grad to 2. transformer-lr was set to 5.0 and we train the network for 70 epochs. All other training parameters were kept as per default values.
- We use pre-trained sub-word token-based large LSTM Language Model (LM) available from ESPNet model repository[2].
- Decoding is done using beam size 10, ctc-weight is set to 0.4, and lm-weight is set to 0.6.
- For Gen-SA, we take a white noise signal and compute Mel-filterbank energies and pitch in the same way as train utterances. We then normalize the features using CMVN stats computed from the training data. We modify the SA implementation in ESPNet by pre-loading the features of white noise signal and replace masked values with scaled white-noise features.
- Augmentation parameters are as follows for both SA and Gen-SA:
  - Time warping: max_time_warp=5
  - Frequency masking: F=30, n_mask=2
  - Time masking: T=40, n_mask=2

The model was tested on librispeech dev (dev-clean, dev-other) and test (test-clean, test-other) sets. To demonstrate the robustness of our approach we test the models on corrupted versions of librispeech dev and test sets. We add babble noise with 15, 10, and 5dB SNR using kaldi noise addition functionality and test the performance of various augmentation techniques. Moreover, we train AM on Aurora-4 dataset [25] to further analyze the noise robustness. For Aurora-4 dataset, we use pre-trained 65,000 vocabulary word LM with available with ESPNet for WSJ task[3]. We also perform cross-dataset evaluation on TED-LIUM [26] dev and test set using AM trained on Librispeech. We report CER and WER for various experiments.

B. Results

1) Baseline results: Table I shows the baseline results in terms of CER and WER with various augmentation strategies. It can be observed from Table I that incorporating SA in training pipeline significantly improves performance of ASR system over not using any augmentation. Gen-SA improves over SA across all test-sets, with the most improvement seen in the dev-other (0.4% absolute) set. This shows that replacing mask values with scaled white noise filterbank values improves the ASR performance over replacing mask values with 0. Figure 2 shows the validation accuracy for SA and Gen-SA starting from epoch 10 to 70. It can be observed that Gen-SA provides better validation accuracy as training progresses and indicates that Gen-SA provides better generalization in performance.

[1] espnet/egs/librispeech/asr1/conf/tuning/train_pytorch_transformer.yaml
[2] LSTM LM for Librispeech
[3] WSJ LM for Aurora-4 task

TABLE I
BASELINE RESULTS WITH VARIOUS AUGMENTATION TECHNIQUES.

Aug.	dev-clean		dev-other		test-clean		test-other	
	CER	WER	CER	WER	CER	WER	CER	WER
no-SA	4.1	3.1	10.9	8.6	4.3	3.5	10.8	8.6
SA	4	2.9	9.5	7.1	**4.2**	**3.2**	9.4	7.4
Gen-SA	**3.9**	**2.8**	**9.1**	**6.8**	**4.2**	**3.2**	**9.3**	**7.3**

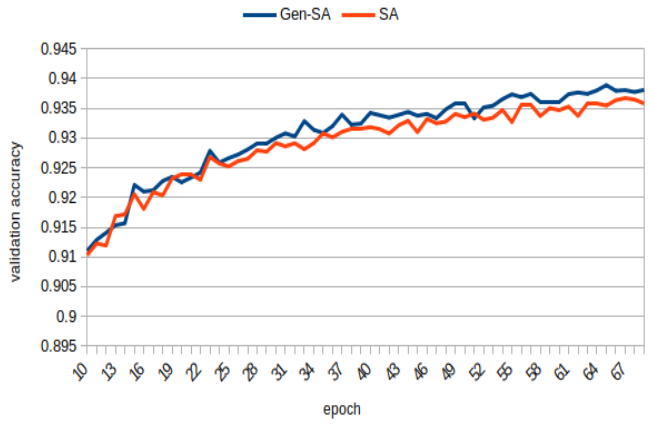


Fig. 2. Validation accuracy on Librispeech dev sets from epoch 10-70 for SA and Gen-SA.

TABLE II
RESULTS OF ABLATION STUDY FOR SPECAUGMENT AND GENERALIZED SPECAUGMENT

	dev-clean	dev-other	test-clean	test-other
SA				
- Time masking	3.1	7.6	3.4	7.5
- Freq. masking	3	7.8	3.2	7.8
Gen-SA				
- Time masking	3.3	7.5	3.4	7.5
- Freq. masking	3.1	8.1	3.4	8.1

We perform ablation study as per [13], where we remove one component at a time from augmentation pipeline to observe contribution of each operation. Results of the ablation study are tabulated in Table II. As observed in [13], we find that for SA and Gen-SA both, Frequency masking contributes the most in performance improvement. In the case of Gen-SA, the contribution of frequency masking was found to be even more significant than SA. Removing Time masking had little impact on the overall performance of both SA and Gen-SA.

2) Cross-dataset performance: To observe the generalization capabilities of the augmentation techniques, we test the performance of ASR systems trained on Librispeech dataset on other out-of-domain dataset, namely, TED-LIUM. We use dev and test set of TED-LIUM for cross-dataset evaluation. We perform decoding using trained model and same decoding parameters without any LM. Table III shows the results of this cross-dataset evaluation. It can be observed that SA greatly improves cross-dataset performance over no augmentation. This suggests that SA has good generalization capabilities in terms of cross-dataset performance. Gen-SA improves over SA significantly, showing superior generalization capabilities compared to SA. This observation suggests that the choice of masking value in SA not only improves same dataset performance, but also improves cross-dataset performance and provides better generalization in ASR system.

TABLE III
CROSS-DATASET RESULTS FOR VARIOUS AUGMENTATION TECHNIQUES. THE AM IS TRAINED ON LIBRISPEECH TRAIN SET AND PERFORMANCE IS EVALUATED ON TED-LIUM DEV AND TEST SET IN TERMS OF CER AND WER WITHOUT ANY LM.

	tedlium-dev		tedlium-test	
	CER	WER	CER	WER
No-SA	21	19.4	19.8	18.9
SA	20	18.3	18.7	17.4
Gen-SA	**19.5**	**17.9**	**18.3**	**17**

TABLE IV
RESULTS OF VARIOUS AUGMENTATION TECHNIQUES FOR NOISY TEST SET. BABBLE NOISE WAS ADDED WITH 15, 10, AND 5DB SNRS IN LIBRISPEECH TEST-CLEAN AND TEST-OTHER UTTERANCES TO CREATE NOISY TEST SET.

		clean	15 dB	10 dB	5 dB
No-SA	test-clean	3.5	5.8	14.1	40.4
	test-other	8.6	18.1	34.8	66.2
SA	test-clean	**3.2**	4.6	8.8	26.9
	test-other	7.4	12.7	24	51.8
Gen-SA	test-clean	**3.2**	**4.4**	**7.9**	**23.2**
	test-other	**7.3**	**11.8**	**21.2**	**46.2**

3) Robustness to noise: We now evaluate noise robustness of various augmentation techniques. Both SA and Gen-SA introduces corruptions in training MFBEs by masking certain portion. Equivalently, masking can be represented as adding noise with $-\infty$ SNR in the masked portion, since no signal magnitude is present in the masked portion. Such corruption may introduce noise robustness in the trained model. Table IV shows the results on librispeech dev and test set corrupted with babble noise with various SNRs. As hypothesized, SA indeed introduces significant noise robustness in the trained model over no augmentation. In this case also, Gen-SA provides the best performance over both the models. The difference in performance of SA and Gen-SA increased with the increased amount of noise in test utterances. Table V shows results on Aurora-4 dataset. Gen-SA provides significant performance improvement over SA on Aurora-4 as well. Especially, on test set B (additive noise) and D (channel degradation + additive noise) the performance difference is observably large. This further bolsters the superior generalization capability of the proposed approach for noisy conditions.

TABLE V
RESULTS OF VARIOUS AUGMENTATION TECHNIQUES ON AUROA-4 MULTICONDITION DATASET.

	A	B	C	D	Average
no-SA	9.5	16.6	15.6	29.3	21.5
SA	**4.7**	7.5	6.2	15	10.4
Gen-SA	**4.7**	**7.1**	**6.1**	**14.7**	**10.1**

IV. CONCLUSIONS AND FUTURE WORK

We propose a generalized version of SpecAugment (SA) where Time and Frequency mask values can be any value, not just 0 or mean of the features. For practical purposes the mask values can be replaced with features of any other signal. In this paper, we present one realization of Generalized SpecAugment (Gen-SA), where we replace the masked region with scaled features values of white noise signal. The results of our experiments show that Gen-SA performs better on Librispeech dev and test datasets compared to SA for similar augmentation parameters. Moreover, results on noisy version of Librispeech dev and test set, and results on AURORA-4 and TED-LIUM dataset suggest that Gen-SA provides better noise robustness and cross-dataset performance compared to SA. This performance improvement requires only one external signal and it has similar computational complexity as SA. We show that replacing mask values in SA with values other than 0 or mean value can provide significant performance boost and robustness to the acoustic model. In future, we plan to use other signals, such as speech of different speaker, more noise signals, music signals, etc. to replace mask values and analyze the robustness of the ASR system to other perturbations and degradation conditions.

REFERENCES

[1] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.

[2] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[3] A. Hannun, C. Case, J. Casper, *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[4] C. Liu, Q. Zhang, X. Zhang, K. Singh, Y. Saraf, and G. Zweig, "Multilingual graphemic hybrid asr with massive data augmentation," in *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, 2020, pp. 46–52.

[5] M. H. Soni, S. Joshi, and A. Panda, "Generative noise modeling and channel simulation for robust speech recognition in unseen conditions.," 2019.

[6] C. Kim, A. Misra, K. Chin, *et al.*, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," *Proc. Interspeech 2017*, pp. 379–383, 2017.

[7] R. Gokay and H. Yalcin, "Improving low resource turkish speech recognition with data augmentation and tts," in *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*, IEEE, 2019, pp. 357–360.

[8] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Multi-speaker sequence-to-sequence speech synthesis for data augmentation in acoustic-to-word speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6161–6165.

[9] S. H. Mallidi and H. Hermansky, "Novel neural network based fusion for multistream asr," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 5680–5684.

[10] G. Kovács, L. Tóth, D. V. Compernolle, and M. Liwicki, "Examining the combination of multi-band processing and channel dropout for robust speech recognition," 2019.

[11] G. Kovács, L. Tóth, D. Van Compernolle, and S. Ganapathy, "Increasing the robustness of cnn acoustic models using autoregressive moving average spectrogram features and channel dropout," *Pattern Recognition Letters*, vol. 100, pp. 44–50, 2017.

[12] L. Tóth, G. Kovács, and D. Van Compernolle, "A perceptually inspired data augmentation method for noise robust cnn acoustic models," in *International Conference on Speech and Computer*, Springer, 2018, pp. 697–706.

[13] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[14] D. S. Park, Y. Zhang, C.-C. Chiu, *et al.*, "Specaugment on large scale datasets," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6879–6883.

[15] D. Le, X. Zhang, W. Zheng, C. Fügen, G. Zweig, and M. L. Seltzer, "From senones to chenones: Tied context-dependent graphemes for hybrid speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 457–464.

[16] W. Zhou, W. Michel, K. Irie, M. Kitza, R. Schlüter, and H. Ney, "The rwth asr system for ted-lium release 2: Improving hybrid hmm with specaugment," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7839–7843.

[17] X. Li, Y. Zhang, X. Zhuang, and D. Liu, "Frame-level specaugment for deep convolutional neural networks in

hybrid asr systems," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 209–214.

[18] X. Song, Z. Wu, Y. Huang, D. Su, and H. Meng, "Specswap: A simple data augmentation method for end-to-end speech recognition," *Proc. Interspeech 2020*, pp. 581–585, 2020.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, IEEE, 2015, pp. 5206–5210.

[20] S. Watanabe, T. Hori, S. Karita, *et al.*, "Espnet: End-to-end speech processing toolkit," in *Proc. Interspeech 2018*, 2018, pp. 2207–2211.

[21] S. Karita, N. Chen, T. Hayashi, *et al.*, "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 449–456.

[22] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration," in *Proc. Interspeech 2019*, 2019, pp. 1408–1412. DOI: 10.21437/Interspeech.2019-1938.

[23] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[24] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71.

[25] N. Parihar and J. Picone, "Analysis of the aurora large vocabulary evaluations," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[26] A. Rousseau, P. Deléglise, and Y. Estève, "TED-LIUM: An automatic speech recognition dedicated corpus," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA), May 2012.

A Byte-based GPT-2 Model for Bit-flip JPEG Bitstream Restoration

Hao Qin, Haoran Sun, and Yi Wang
Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR
E-mail: {fuyuyu.Qin, haoran-eee.sun}@connect.polyu.hk, yi-eie.wang@polyu.edu.hk

***Abstract*—In this paper, we investigate the application of large language models (LLMs) for the recovery of corrupted bitstreams, specifically focusing on JPEG image data. We propose a byte-based GPT-2 model that directly processes byte sequences and predicts the subsequent byte, enabling its application to JPEG bitstream recovery. This architecture allows the model to capture the relationships between consecutive byte data within the bitstream of a JPEG image, such that the model can restore the bit-flip errors due to the damaged storage and malicious attack. We evaluate the model's performance on bit-flip JPEG datasets with varying bit error rates (BERs). The experimental results demonstrate the model's ability to implicitly learn patterns in the bitstream and correct erroneous bytes, showcasing the potential of LLMs in binary processing tasks. Our findings highlight the promise of byte-based LLMs in addressing data corruption issues and open up new avenues for research in this domain.**

I. Introduction

In the face of increasing transmission bandwidth demand for streaming media, audio and video are transmitted under various lossy compression techniques [1], such as MPEG-4 [2] and JPEG [3]. These techniques provide the codec to compress the data into a stream of bytes at a specified compression ratio and decompress it back into images and audio. In scenarios where transmission and storage may be interfered with or intentionally attacked [4]–[6], bit flips or missing will occur in the stream which corrupts blocks or the entire bitstream. Recovering data from a corrupted bitstream is a challenging task. Typically, based on the optimization principle, the data segments in the bitstream of a lossy compression method do not contain redundant data for error correction and use irregular lengths and code words for different data[7].

On the one hand, researchers have focused their problem-solving efforts on coding techniques. Modern error correcting codes (ECC) such as LDPC [8] are used in wireless transmission and digital video broadcasting controlling errors in data transmission. Message-digest algorithms such as MD5 [9]are used for verifying data integrity and creating digital signatures. Other researchers have proposed models to extract pattern and structural features from the correctly decoded portion of the bitstream, complementing the image or interpolating it [10]. Nevertheless, corruption caused by bit flips or missing bits often results in relatively long segments of data that cannot be decoded. Repair of decoded content relies on the temporal and spatial continuity of the data, which is not always satisfactory.

In this work, we modified the GPT-2 [11] model to accept a sequence of bytes as input and predict the next byte. We intend to use the GPT-2 model for JPEG bitstream restoration. The idea is inspired by the fact that image bitstreams in JPEG format have a fixed metadata structure [3], and if the byte sequences in data segments are considered as statements, the codebook in the metadata defines all the semantics recurring in the bitstream. We modeled bit-flip corruption on JPEG image bitstream with existing image datasets such as ImageNet [12] and CIFAR-10 [13]. Trained on these datasets, our proposed binary model successfully recovers a portion of corrupted files. To further evaluate the performance, we restricted the bitstream corruption to occur in different regions and evaluated the model at different bit error rates (BERs). The results show that LLM can capture the relationship between stream metadata and data segments.

Overall, this paper makes the following contributions:

- We adopt Large Language Models (LLMs) to process byte sequences directly and predict the subsequent byte. The proposed byte-based GPT-2 model can be applied directly to the task of JPEG bitstream recovery.
- We evaluate our method on bit-flip JPEG datasets. Experimental results show the potential of the GPT-2 model in this field. To the best of our knowledge, this marks the first instance of utilizing an LLM to restore corrupted bitstreams.

II. Related Work

JPEG structure. The file bitstream of JPEG format is partitioned into two parts: the metadata segment and the image data segment. The image data is compressed by Discrete Cosine Transform (DCT) and then converted into a variable-length coding structure after quantization, run-length encoding (RLE), and Huffman coding as shown in Fig. 1. The metadata stores the important parts such as the Huffman coding table, quantization table, etc., which accounts for a small percentage of the whole bitstream. The data segments are sequence-dependent, so bit-flip and missing errors will seriously interfere with decoding, resulting in image corruption or even the inability to read the entire bitstream. Image corruption can happen in a variety of patterns.

Bitstream restoration. Much of the previous work has focused on using techniques for specific codecs for concealment or fixing. Error concealment is a common decoder-side post-processing technique used to repair erroneous regions in decoded videos [14], [15]. It can be divided into several

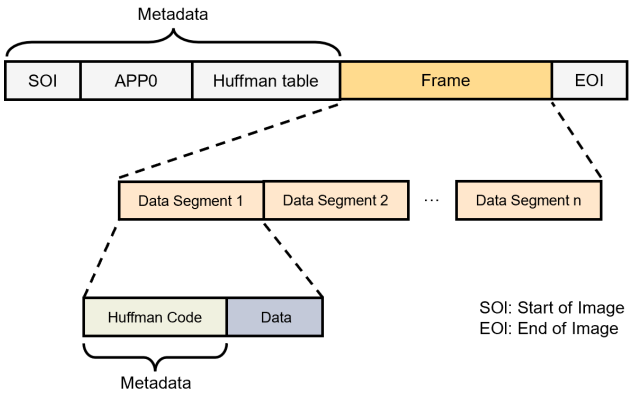


Fig. 1. The structure of JPEG bitstream.

types, including spatial, temporal, and hybrid spatiotemporal methods, covering both bitstream and pixel levels [15], [16]. In recent years, deep learning-based methods usually assume traditional damage patterns and use experimental masks to simulate stripes or block losses [14], [17]–[19]. However, this approach is not suitable for recovering videos with damaged bitstreams because the damage caused by actual packet loss is often difficult to predict and has no regularity. The approach in [20] constructed a robust decoder that attempted to skip over erroneous blocks and parse the remainder of the bitstream. The parsed corrupted images were restored using a two-stage compensation and alignment framework.

Byte models. From early LSTM to today's language models based on the Transformer architecture [21], they play a vital role in understanding, generating human language, and simulating intelligence. Text tokenization breaks down text into smaller units (such as words or subwords) as model input [21]. The Generative Pre-trained Transformer (GPT) model marks a major advancement, utilizing self-supervised learning and next-token prediction pre-training to capture the structure and semantics of language. Researchers are also exploring byte-level encoding methods to improve the performance of existing models. For example, compressing byte sequences using language models [22] offers new insights into leveraging large pre-trained models [23]. Byte-level byte pair encoding (BBPE)[24] has shown promise in enhancing multilingual model pre-training and machine translation [25]. ByT5 [26] builds upon this by processing byte sequences with a standard Transformer model, generalizing an unlabeled encoding approach for improved noise robustness and spelling sensitivity in multilingual scenarios.

III. Our Method

In this section, we describe the byte-based GPT-2 modeling. Processing data from the byte level leads to a model with high granularity and extremely long sequences. The computational effort of self-attentive scaling in Transformer-based models grows quadratically with the length of the sequence [27]. To avoid excessively long sequence lengths, bytes are partitioned

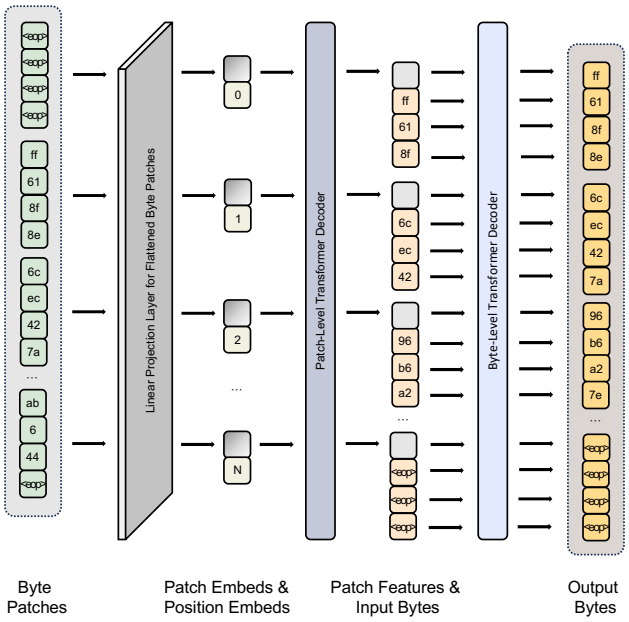


Fig. 2. The structure of byte-based two-layer GPT-2 model.

into patches in a multi-layer transformer, referring to the work of Wu et al. [27].

As shown in Fig. 2, the byte-based GPT-2 model contains three parts, the linear projection layer followed by positional coding to the sequence, the patch-level decoder, and the byte-level decoder.

Bitstream Preprocessing. In a bitstream, data is not always organized by bytes. There are many semi-byte blocks and flag bits in different bitstreams (e.g., communication protocols, standard file types, CPU states, etc.). In variable-length bitstream, at the end of the file, bits often cannot be rounded up to a byte, and padding is required to fill up the data. For a sequence of bytes (bitstream) $B = [b_0, b_1, b_2, b_3, ..., b_{T-1}], b_i \in \mathbb{Z}_{256} = \{0, 1, 2, ..., 255\}$ of length T, we pad $8 - |b_{T-1}|$ zeros at Byte b_{T-1}.

The direct use of B as input leads to long sequence lengths, and to solve this problem, inspired by previous studies[20], [28], we split the B into patches with a certain size S, forming a sequence of patches $P = [p_1, p_2, ..., p_K]$ with the length of K, where $K = \lceil \frac{T}{S} \rceil$. To locate the endings of the bitstreams, noting that the common data file storage formats usually have a start-of-file marker and an end-of-file marker, we add start-of-patches (SOP) and end-of-patches (EOP) marker to the beginning and end of the byte sequence. The SOP and EOP markers share one single symbol, each occupying a byte position. A special patch is formed by adding S markers to the beginning to indicate that this is the start patch, and at the end we need to add a sufficient number of markers to indicate that this is the end patch, and the number of patches to add is $S + Mod(-K, S)$.

Linear Projection Layer. Firstly, there are only 256 hexadecimal digits that can be represented by a byte, so a byte can

be represented by a one-hot vector of length 256, and adding a symbol to stand for the EOP marker gives a length of 257. A patch contains S bytes, which can be represented as a matrix of $S \times 257$. This matrix is flattened to a vector of length $S \times 257$ and linearly weighted to perform token embedding on patches. For error correction type applications, the byte vector will be in Gray code order. For each input patch, the linear projection layer maps the sparse input vector to a dense vector D of size (S, H), with H being the size of the hidden layer, and this process can be formulated as follows:

$$D_i = V_i \cdot W_{linear}, 1 \leq i \leq K \tag{1}$$

where W_{linear} denotes the weight matrix of the linear projection layer.

Byte Decoder. We used two layers of transformer decoders to make predictions for positionally encoded patches and bytes, respectively. First, the patch-level transformer makes predictions based on the sequence of embedded patches to get predicted features for the predicted patches. Then byte-level decoder predicts the next byte based on the previous patches' prediction features and the sequence of bytes within the current patch. Considering the existence of byte-level semantics between codewords in the byte stream, the computations performed at the byte decoder are not independent for each patch, unlike the work of Wu et al. [27].

Loss Function. Our goal is to predict the next byte of the sequence and try to keep it consistent with the uncorrupted sequence, given the existing corrupted sequence, which requires us to adopt a generative modeling framework for the process. For a sequence of bytes $B = \{b_1, b_2, b_3, ..., b_T\}$ of length T, the log-likelihood of the next byte needs to be maximized, so the loss function can be defined as:

$$\mathcal{L}(\Theta) = -\sum_{i=1}^{T-1} \log p(b_{i+1}|b_1, b_2, b_3, ..., b_i; \Theta) \tag{2}$$

where Θ denotes the model parameter $p(\cdot)$ denotes the conditional probability distribution of the next byte prediction. bitstream restoration is a special case of generative modeling, where the model predicts bytes consistent with the original content in most cases. In this case, not all bytes in the sequence are meaningful, and the sequence may have potential semantic sparsity, which remains to be investigated.

Inference. For inference, the patches predicted by the model are unwrapped and the padded zero bits at the end of the bitstream are removed to obtain the predicted sequence.

IV. Experiments

A. Experimental Settings

1) Datasets: For bitstream corruption, there is no publicly available dataset other than a benchmark for video bitstream corruption recovery [29]. To evaluate the model in this paper, we use the ImageNet and CIFAR-10 datasets, which are image datasets used for image recognition and classification, to which we artificially add single-bit-flip errors to validate the viability of the model for JPEG bitstream recovery.

Bit errors happen randomly throughout the whole bitstream and the errors occurring in the Huffman coding table cause the entire file to be unreadable, while errors occurring in the data segments lead to multiple forms of error patterns. Among these, metadata such as the Huffman coding table accounts for less than 0.1% of the total bitstream size. With a low bit error rate (BER), i.e., no larger than 10^{-2}, the bit flips that occur in the metadata are extremely rare. Metadata has a fixed structure and logical order, as well as a limited number of possible values, and we can use its structural information for error correction. Hence, the bit flips in metadata do not have a significant problem for file reading. The bit-flip errors in metadata can be efficiently corrected using the robust decoder proposed in the study of [20].

To simulate corruption, we added noise to the JPEG image bitstream at several certain BER levels and augmented some of the files to suggest where different errors occurred for the same bitstream. Considering the training cost, we use the context length of 8k, which means that the size of the JPEG image file should not be more than 7.8 KB or 128×128. We filtered the images that do not exceed 7.8 KB in the ImageNet dataset to be used as the training data. The image resolution of the CIFAR-10 dataset is 32×32, which meets the demand.

2) Evaluation Metrics: In the field of communication, the quality of a digital signal after transmission can be evaluated in terms of peak-signal-to-noise ratio (PSNR). For the bitstream restoration problem in this paper, the upper bound of the signal-to-noise ratio ratio is the BER if the network does not produce incorrect predictions. It is one-sided to use only one single metric, the peak-signal-to-noise ratio, to evaluate the effectiveness of the model. For the bitstream recovery problem, we propose new definitions of accuracy and recall to replace PSNR, as follows.

$$Precision = 1 - \frac{\sum_i min\{L_i, L_i^{gt}\}}{\sum_i L_i^{gt}}, \tag{3}$$

$$Recall = 1 - \frac{\sum_i N_i^D}{\sum_i L_i^{gt}}, \tag{4}$$

where L_i^{gt} represents the Levenshtein distance [30] between the ground truth and the corrupted bitstream, L_i represents the Levenshtein distance between the predicted bitstream and the correct bitstream, and N_i^D denotes the number of successful fixes of the prediction result at the wrong bits. In addition, we use a metric of human sensing of the image to evaluate the quality of the recovered image. which is based on the PSNR of the pixels of the image to the original image after reading the bitstream as an image. This metric is specifically calculated as follows: the peak-signal-to-noise ratio of pixels to the ground truth image after the recovered stream is read as an RGB image.

3) Parameter Settings: The embedding and hidden state dimensions of Byte-Level are set to 3 and 768 respectively. The embedding and hidden state dimensions of Patch-Level were set to 12 and 768 respectively. The Adam optimizer [31]

TABLE I
OVERALL PERFORMANCE OF THE MODEL ON THE IMAGENET DATASET.

Bit error rates	ImageNet	
	Precision(%)	Recall(%)
10^{-2}	0.380	1.424
10^{-3}	8.950	13.719
10^{-4}	17.265	21.101
10^{-5}	22.430	22.544

TABLE II
OVERALL PERFORMANCE OF THE MODEL ON THE CIFAR-10 DATASET.

Bit error rates	CIFAR-10	
	Precision(%)	Recall(%)
10^{-2}	0.632	3.215
10^{-3}	8.950	13.719
10^{-4}	21.123	28.545
10^{-5}	19.989	22.468

TABLE III
OVERALL PERCENTAGE AND AVERAGE PSNR OF IMAGES RECOVERED.

Bit error rates	Recovery Rate(%)		PSNR(dB)	
	ImageNet	CIFAR-10	ImageNet	CIFAR-10
10^{-2}	4.082	1.460	3.523	3.389
10^{-3}	16.226	13.510	12.063	11.221
10^{-4}	56.122	36.180	19.994	18.190

TABLE IV
COMPARISON OF DIFFERENT METHODS IN OVERALL PERCENTAGE AND AVERAGE PSNR OF IMAGES RECOVERED A BER LEVEL OF 10^{-4}.

Model	Recovery Rate(%)		PSNR(dB)	
	ImageNet	CIFAR-10	ImageNet	CIFAR-10
Robust decoder [20]	44.358	26.115	17.733	16.889
Our Model	56.122	36.180	19.994	18.190

was used. The initial learning rate was 0.0001 and the batch size was 32. A drop rate of 0.1% was used during training.

B. Results

In our experiments, we evaluated the performance of the model at four BER levels on two datasets. In particular, a total of 2,402 images from the ImageNet dataset were used, which were allocated in a 10:1 ratio of training/test set. The training set was augmented by adding noise and the model was trained with both the corrupted and uncorrupted images, totaling 11020 images. For the CIFAR-10 dataset, a total of 50,000 images from 10 categories were used as the training set and 10,000 as the test set. Noised images were aligned with the uncorrupted images and a total of 100000 images were provided for training. For comparison with approaches related to bitstream restoration, we choose the stage-one robust decoder in [20] as the baseline, since this process is for bitstreams and not for image data.

1) Overall Performance: **Precision and recall.** The overall precision and recall of the prediction results for the ImageNet test set are shown in Table I and the overall results for the CIFAR dataset are shown in Table II. As can be seen in Tables I and II, the prediction accuracy of the model basically improves gradually as the BER decreases. Optimal performance is achieved on the CIFAR-10 dataset at a BER of 10^{-4} and on the ImageNet dataset at a BER of 10^{-5}. Theoretically, the lower the BER, the better the performance of the model. The reason that CIFAR-10 achieves optimal performance on 10^{-4} BER is that 10^{-4} and 10^{-5} will only produce bit errors below 1 bit on average for the image size 32×32 of the CIFAR-10 dataset. We only tested the files that produced bit errors, so it is conditionally the same for CIFAR-10. Precision and recall cannot be reasonably calculated for images recovered using the robust decoder in [20]. It adopts a zero-padding strategy to skip blocks that cannot be parsed. As a result, the precision and recall of the robust decoder have significantly reduced, and the comparison is not meaningful.

Pixel's level performance. Since the metadata (e.g., Huffman code) segment is relatively sparse in variable-length bitstream, the JPEG stream is mostly filled with data segments. When the bits in the data segment are changed, it only affects the color, texture, and other data of the specific image block and does not affect the decoding of the bitstream, so errors occurring in data segments have less effect on the image. It is impossible to evaluate the quality of the image recovered from the metrics of bitstream restoration alone, so we use the PSNR as an evaluation metric. The image recovered with higher PSNR is considered as successfully recovered image, and the results are shown in Tables III. Compared to the results in Tables I and II, the performance of the model is better, this is because the model is unable to identify whether the data segments are noisy or not and thus cannot make an accurate prediction of the segment, which reduces the accuracy of the model. At an error rate of 10^{-5}, the model recovered 56.122% of the corrupted images on ImageNet and 36.180% of the corrupted images on CIFAR-10.

Table IV demonstrates the results of the performance comparison between our model and the robust decoder in [20] at a BER level of 10^{-4}. Compared to the robust decoder, our model achieves a performance improvement of 2.26dB PSNR and 11.76% recovery rate on ImageNet. These results show that our method is more advanced in bitstream restoration.

Visual Results. Fig. 3 and Fig. 4 show a comparison of the visual quality of the images before and after the restoration of the robust decoder [20] and our model on the two datasets, where the recovery gets progressively better from left to right. It is worth noting that the robust decoder is unable to produce effective parsing results because the image size of CIFAR-10 is too small. The samples selected in Fig. 3 and Fig. 4 are selected from corrupted bitstreams that can still be decoded, albeit with severe discoloration and block artifacts. It is important to note that these samples represent a subset of the corrupted files, as some corrupted bitstreams cannot be decoded. From the figures, it can be seen that after the recovery by our model, some images are completely restored while others are

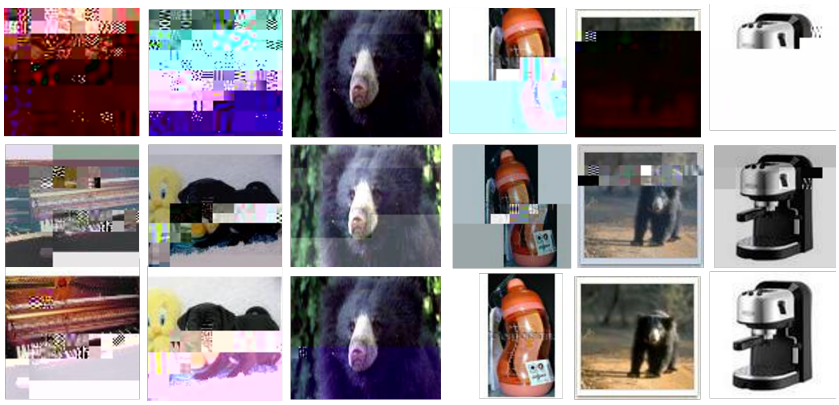


Fig. 3. Visual comparison of the images before (at the top) and after restoration of the robust decoder [20] (at the middle) and our model (at the bottom) on ImageNet at a BER level of 10^{-4}.

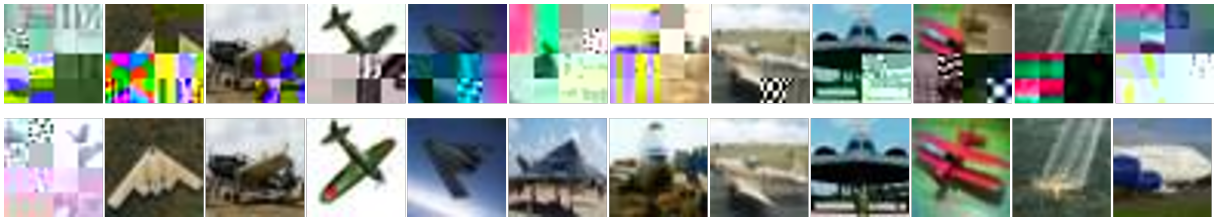


Fig. 4. Visual comparison of the images before (at the top) and after (at the bottom) restoration on CIFAR-10 at a BER level of 10^{-4}.

TABLE V
RECOVERY RATE ON THE BENCHMARK OF NOISY METADATA SEGMENTS.

Bit error rates	Dataset	
	ImageNet	Cifar
10^{-2}	0.000%	0.000%
10^{-3}	0.090%	0.810%
10^{-4}	17.440%	20.010%
10^{-5}	62.190%	81.780%

compensated for the color distortion and block shifts, thus providing better visual effects.

2) Recovery in Metadata of Bitstream: In order to verify that the model achieves recovery of the bitstream and not only a pseudo-positive due to bit errors in data segments, we constructed a benchmark using the test set of both datasets. We made twenty copies of all images and added errors at different locations in the metadata segment. From Table V, it can be seen that for samples corrupted only in the metadata segment, the model still recovers the data, which indicates that the model learned the inter-byte relationship of metadata from the bitstream.

V. CONCLUSIONS

In this paper, we propose a novel JPEG bitstream recovery method based on GPT-2, which explores the potential of large language models (LLMs) in the restoration of corrupted bitstreams. Our model employs a large language model to process binary information, achieving an end-to-end processing flow. Experimental results demonstrate that the proposed method achieves a 56% recovery rate using an ImageNet dataset at low BER levels. In future work, we aim to explore more LLMs and variable content-length embedding for generalized bitstreams to enhance the model's prediction performance and improve the quality of recovered JPEG images.

REFERENCES

[1] A. J. Hussain, A. Al-Fayadh, and N. Radi, "Image compression techniques: A survey in lossless and lossy algorithms," *Neurocomputing*, vol. 300, pp. 44–69, 2018.

[2] D. Le Gall, "Mpeg: A video compression standard for multimedia applications," *Communications of the ACM*, vol. 34, no. 4, pp. 46–58, 1991.

[3] G. K. Wallace, "The jpeg still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[4] A. S. Rakin, Z. He, and D. Fan, "Bit-flip attack: Crushing neural network with progressive bit search," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1211–1220.

[5] F. Carpi, C. Häger, M. Martalò, R. Raheli, and H. D. Pfister, "Reinforcement learning for channel coding: Learned bit-flipping decoding," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2019, pp. 922–929.

[6] J. Dong, H. Qiu, Y. Li, *et al.*, "One-bit flip is all you need: When bit-flip attack meets model training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4688–4698.

[7] I. E. Richardson, *Video codec design: developing image and video compression systems*. John Wiley & Sons, 2002.

[8] R. Gallager, "Low-density parity-check codes," *IRE Transactions on information theory*, vol. 8, no. 1, pp. 21–28, 1962.

[9] R. Rivest, *Rfc1321: The md5 message-digest algorithm*, 1992.

[10] A. S. Parihar, D. Varshney, K. Pandya, and A. Aggarwal, "A comprehensive survey on video frame interpolation techniques," *The Visual Computer*, vol. 38, no. 1, pp. 295–319, 2022.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[13] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[14] S. Ye, M. Ouaret, F. Dufaux, and T. Ebrahimi, "Hybrid spatial and temporal error concealment for distributed video coding," in *2008 IEEE International Conference on Multimedia and Expo*, IEEE, 2008, pp. 633–636.

[15] J. Koloda, J. Østergaard, S. H. Jensen, V. Sánchez, and A. M. Peinado, "Sequential error concealment for video/images by sparse linear prediction," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 957–969, 2013.

[16] M. Kazemi, M. Ghanbari, and S. Shirmohammadi, "A review of temporal video error concealment techniques and their suitability for hevc and vvc," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 12 685–12 730, 2021.

[17] A. Sankisa, A. Punjabi, and A. K. Katsaggelos, "Video error concealment using deep neural networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 380–384.

[18] C. Xiang, J. Xu, C. Yan, Q. Peng, and X. Wu, "Generative adversarial networks based error concealment for low resolution video," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 1827–1831.

[19] Y. Wang, K. Wu, W. Liu, K.-H. Yap, and L.-P. Chau, "Image representation and deep inception-attention for file-type and malware classification," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2023, pp. 1–5.

[20] W. Liu, Y. Wang, K.-H. Yap, and L.-P. Chau, "Bitstream-corrupted jpeg images are restorable: Two-stage compensation and alignment framework for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9979–9988.

[21] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] G. Delétang, A. Ruoss, P.-A. Duquenne, *et al.*, "Language modeling is compression," *arXiv preprint arXiv:2309.10668*, 2023.

[23] J. Hoffmann, S. Borgeaud, A. Mensch, *et al.*, "Training compute-optimal large language models. arxiv 2022," *arXiv preprint arXiv:2203.15556*, vol. 10, 2022.

[24] J. Wei, Q. Liu, Y. Guo, and X. Jiang, "Training multilingual pre-trained language model with byte-level subwords," *arXiv preprint arXiv:2101.09469*, 2021.

[25] C. Wang, K. Cho, and J. Gu, "Neural machine translation with byte-level subwords," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 9154–9160.

[26] L. Xue, A. Barua, N. Constant, *et al.*, "Byt5: Towards a token-free future with pre-trained byte-to-byte models," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 291–306, 2022.

[27] S. Wu, X. Tan, Z. Wang, R. Wang, X. Li, and M. Sun, "Beyond language models: Byte models are digital world simulators," *arXiv preprint arXiv:2402.19155*, 2024.

[28] Y. Wang, W. Liu, K. Wu, K.-H. Yap, and L.-P. Chau, "Intra-and inter-sector contextual information fusion with joint self-attention for file fragment classification," *Knowledge-Based Systems*, vol. 291, p. 111 565, 2024.

[29] T. Liu, K. Wu, Y. Wang, W. Liu, K.-H. Yap, and L.-P. Chau, "Bitstream-corrupted video recovery: A novel benchmark dataset and method," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[30] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[31] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, Ieee, 2018, pp. 1–2.

A Preliminary Study on Analysing Mandarin Tone Values of Romance L2 Mandarin Learners

Wu-Hao Li*, Te-hsin Liu† and Chen-Yu Chiang‡
* College of Electrical Engineering and Computer Science, National Taipei University, Taiwan
E-mail: hank12451@gmail.com Tel/Fax: +886-2671-0893
† Graduate Program of Teaching Chinese as the Second Language, National Taiwan University
E-mail: tehsinliu@ntu.edu.tw Tel/Fax: +886-3366-1549
‡ Dept. of Communication Engineering, National Taipei University, New Taipei City, Taiwan
E-mail: cychiang@mail.ntpu.edu.tw Tel/Fax: +886-2671-0893

***Abstract*—The 5-level tone value labeling system helps characterize pitch contours of syllables for L2 Mandarin learners and teachers to facilitate tone acquisition and tone error analysis, respectively. In the past, linguistic experts did the tone value labeling and initially used it to represent or analyze the pitch contours of monosyllables. Some L2 Mandarin teachers also used the tone value labeling system to analyze pitch contours of continuous speech and illustrate pitch realizations for L2 Mandarin learners. However, the pitch contours of the labeled tone values of continuous speech were rarely objectively measured or analyzed. This paper presents a preliminary analysis of tone values of a continuous read speech corpus uttered by six Romance L2 Mandarin learners. The pitch contours of the tone value classes were explored, and some experiments related to the automatic recognition of tone and tone value classes were reported. We expect the results reported by this study to shed light on the development of an automatic tone value classification mechanism, which will be used in computer-assisted pronunciation training for L2 learners in the future.**

***Index Terms*—Mandarin, tone recognition, tone value, computer-assisted pronunciation training**

I. Introduction

L2 Mandarin learners would utter syllables with incorrect tones. These incorrect tones would be affected by learners' mother tongues and sound differently from native Mandarin speakers. The most intuitive method to access the tone uttered by L2 Mandarin learners is to utilize the results of a machine tone recognizer. The tone recognition results can be converted to some metrics as measures for tone assessments or error detection [1]–[3]. However, the tone recognition results, e.g., the posterior probabilities of tones, do not easily illustrate how the pitch contours of speech uttered by L2 Mandarin learners differ from native Mandarin speakers' tones. For example, if the tone recognizer produces the score for a target Mandarin syllable with scores of 0.1, 0.2, 0.3, 0.3, and 0.1 for tones 1 to 5, this score information is vague for an L2 Mandarin learner to change his/her produced pitch contour into correct one.

For linguistic background researchers or L2 Mandarin teachers, using the 5-level tone values to characterize pitch contours of speech uttered by L2 Mandarin learners is more intuitive or more comprehensive than using tone classes. The 5-level tone values were first introduced by Chao [4], [5] to initially classify the four lexical tones in Mandarin as typical tone value sequences of 55, 35, 214, and 51, for T1, T2, T3, and T4, respectively. In most of the previous studies or teaching materials [6], [7], the tone value transcriptions were based on human ear annotations, supplemented by pitch contours. For example, L2 Mandarin learners may utter a T1 syllable with not a high-enough 44 tone value in the sequence, which has a lower level than the standard T1 with a tone value of 55.

Since the labeling involves human perception in audio and vision, the tone values labeled may suffer from inconsistencies between labelers. Though each utterance can be labeled by many human labelers with consistency checked by discussion between the labelers, the discussion process would be time-consuming and still involve human subjective biases. Besides labeling tone value by perception, Shi [8] proposed mapping from logF0 to Chao's 5-level tone values with maximum and minimum logF0 of speakers. Edmonson [9] proposed to first map from measured logF0 to semitone and then to Chao's 5-level tone values. The successive studies [10]–[12] generally follow the formula proposed by Shi and Edmonson and modify the formula regarding measuring methods to remove or normalize inter-speaker pitch differences.

The formulas proposed by Shi [8] and Edmonson et al. [9] were mainly used to represent tone values of lexical tones in monosyllables. Therefore, the above-mentioned 5-level tone representation and their variations for normalization would no longer be suitable to label the tone values of syllables in a continuous speech due to prosodic variation made by coarticulation and prosodic phrasing though many studies still use this 5-level tone value system to annotate syllable tones of continuous Thai [13], [14] speech.

This paper presents a preliminary analysis of tone values of continuous read speech uttered by six romance L2 Mandarin learners. The tone values of syllables are labeled by linguistic experts with the 5-scale values. To objectively visualize the tone values labeled, we first stylize the pitch contours of syllables by polynomial basis functions, which describe the mean, slope, acceleration, and curvature of syllable logF0 contour. We, then analyze the typical patterns of each tone value with the average observed logF0 contours represented

by the coefficients of the polynomial functions. Last, we conducted automatic machine tone and tone value classification experiments to obtain baselines for tone and tone value classification tasks.

Contributions made by this paper are summarized in two folds. First, typical logF0 patterns for tone values in continuous speech uttered by L2 Mandarin learners are objectively explored. Most previous literature related to tone values considered primarily illustrating pitch contours of tones, especially for pronouncing monosyllables. Second, this paper reports the improvement of the tone recognition accuracy with respect to the result reported in 2023 [15]. Based on this improvement, this paper constructs the baseline for the performance of the tone value class recognition. So far, to our knowledge, this is the first paper to report the performance of tone value recognition.

This paper is organized as follows: Section II introduces the experimental databases used in this study. Section III analyzes the pitch characteristics of tone classes and tone value classes. Section IV conducts tone class and tone value class recognition experiments. The last section reports some conclusions and future works.

II. Experimental Databases

A. Romance L2 Mandarin Speech Corpus

1) Recording and Labeling: The Romance L2 Mandarin speech corpus is used as the material for this preliminary study. The dataset contains six (3 males and 3 females) French learners of Mandarin who participated in creating audio recordings. The participants were second-year (intermediate level) students majoring in Chinese at Institut National des Langues et Civilisations Orientales in Paris. None of the students had previous experience with other tonal languages. The participants were asked to read two short texts (completable within 5.5 min on average) at a natural speed. The first was a brief introduction to the Lunar New Year; the second was an anecdote from a high school student written in the first person.

There are a total of 4,381 syllables in this corpus, and each syllable is labeled with the 5-level tone values by three native Mandarin-speaking linguistic experts in linguistics. After the discussions by the three linguistic experts, the inconsistent or conflict tone value labeling was reduced as much as possible. There are 41 types of tone values labeled by the three experts. For labeling tone classes, each syllable of the corpus was first labeled with a ***lexical tone*** by the linguistic processor of the Speech Labeling and Modeling Toolkit (SLMTK) Version 1.0 [16]. Then, three other native Mandarin speakers labeled the corpus's syllables with perceived tones using the following procedure: the lexical tone labels for syllables were first provided to the three labelers as references, and the labelers relabeled the syllables with the tone classes that they heard if their perceived tone classes were different from the lexical tones.

2) Consistencies of Tone and Tone Value Labelings: Table I shows the percentages for the numbers that denote how many different labels a syllable received. It is found that the consistency in tone class labeling is higher than the consistency in tone value class labeling, maybe because the number of classes for the tone value labeling, i.e., 41, is larger than the number of tone classes, i.e., 5. Table II displays the correlation matrix for the lexical-perceived tone pairs. It is found that syllables with lexical tone 3 would be easily perceived as tone 2 due to tone three Sandhi. Except for tone three Sandhi, the percentages not on the diagonal may indicate the erroneous tones produced by the six L2 Mandarin learners.

TABLE I: percentages for the numbers 1 to 3 that denote how many different labels a syllable received. For example, 1 means all three labelers assigned the same tone tag to a syllable.

	1	2	3
tone classes	80.8%	12.0%	7.2%
tone value classes	60.0%	30.4%	9.6%

TABLE II: Correlation matrix for lexical and perceived tones: The data annotated by three labelers were combined, resulting in a dataset three times larger than the original.

		perceived tones (%)					
		1	2	3	4	5	Total#
lexical tones	1	85.6	7.1	1.5	5.3	0.5	2565
	2	6.4	84.8	1.4	6.7	0.7	2961
	3	3.7	10.3	79.3	6.5	0.2	2007
	4	3.4	4.8	1.9	89.3	0.7	4242
	5	1.8	1.5	0.4	1.4	95.0	1368

3) Relationship between Tone Value Classes and Tone Classes: By analyzing the co-occurrence between tone value classes and tone classes, the top-2 tone value classes for each of tones 1 to 5 are the same as summarized in Table III. The top-2 tone values match well with the common-agreed levels and contours of the five Mandarin tones. The T1 has the lowest entropy over the corresponding tone value labels, inferring that the speakers could pronounce T1 well. On the contrary, T3 has the highest entropy and both low-dipping tone value (21) and rising tone value (24), indicating that the speakers probably may not know how to process tone three Sandhi. T1 (with a standard tone value label of 55), the high-level tone, had an accuracy rate of 35.58%. Nevertheless, it was also produced as 44 with 33.75%. The rising tone T2 (35) had an accuracy rate of 19.76%, and was realized as 24 with 38.04%. This indicates that when L2 learners produce a rising tone, the ascent is not large enough, leading to a tonal deviation.

TABLE III: The top-2 tone value classes and entropies regarding 41 tone classes for each of lexical and perceived tones.

syllable tone	top-2 tone values	entropy
lexical T1	44 (34%), 55 (36%), others (30%)	1.874
lexical T2	24 (38%), 35 (20%), others (42%)	2.196
lexical T3	21 (38%), 24 (11%), others (51%)	2.374
lexical T4	53 (32%), 42 (28%), others (40%)	2.136
lexical T5	3 (26%), 2 (16%), others (58%)	2.339
perceived T1	44 (37%), 55 (39%), others (24%)	1.654
perceived T2	24 (40%), 35 (21%), others (39%)	2.084
perceived T3	21 (44%), 24 (10%), others (46%)	2.193
perceived T4	53 (37%), 42 (29%), others (34%)	1.937
perceived T5	3 (27%), 2 (17%), others (56%)	2.291

B. Speech Corpora For Pre-trained Model

Since the size of the target corpus (the Romance L2 Mandarin speech corpus) is small, we use the following three speech corpora to obtain a pre-trained tone recognition model for the following fine-tuning of the target tasks, i.e., tone class and tone value class recognitions for the target corpus:

1) AISHELL-1: AISHELL-1 [17] is an open-source Mandarin speech corpus for constructing Mandarin automatic speech recognition (ASR) systems. The corpus features recordings from 400 diverse Chinese speakers, high-quality audio at 16kHz, and over 95% transcription accuracy, ensured by expert annotation and thorough quality checks.

2) AISHELL-3: The AISHELL-3 [18] corpus is a collection of speech data with a total length of 85 hours with 997,998 syllables. It is mainly intended for constructing a text-to-speech system for multiple speakers. The database is labeled with high-accuracy phonetic and tone transcription, which makes it suitable for training and testing tone recognition models. In this study, we use the AISHELL-3 corpus to evaluate the performances of the pre-trained tone recognizers.

3) NER: The speech data of the NER corpus [19] originated from the National Education Radio Station in Taiwan. In total, the corpus contains about 5,129 hours and 6,658 files, of which 1,218 hours of data was transcribed by professional dictators, and the rest of the data was transcribed by ASR systems. We remove the utterances that can not be force-aligned successfully by the force-aligner used in the SLMTK [16] because we found that most utterances without successful speech alignments usually contain erroneous transcriptions. As a result, we used a total of 2,318 hours of the NER corpus to construct a pre-trained tone recognition model.

III. Pitch Analyses for Tone Values

A. Parameterized Raw Pitch Contour (RPC) Representation

This study employs the discrete orthogonal polynomials [20] widely used in modeling syllabic pitch contours of Mandarin [21]–[24] and Chinese dialects [25] to parameterize variable-length frame-based logF0 values (extracted by the RAPT [26]) with fixed-dimensional representation and small fitting errors. We can easily model the logF0 contours for some specific tone values and visualize them for illustration with this representation by

$$a_n(j) = \frac{1}{M_n+1}\sum_{m=0}^{M_n} F(n,m)\phi_j(m/M_n) \text{for } j = 0 \sim 3 \quad (1)$$

where $F(n,m)$ is the observed logF0 value at m-th voiced frame of the n-th syllable; $\phi_j(m/M_n)$ is the j-th orthogonal polynomial basis (j-th order polynomial) with a length of M_n+1 voiced frames; $F(n,m)$ for $m = 0, 1, ...M_n$ is the so-called Raw Pitch Contour (RPC); $a_n(j)$ is called the orthogonal expansion coefficient (OEC) for j-th orthogonal polynomial basis, $\phi_j(m/M_n)$. The frame logF0 values can be reconstructed with a negligible fitting error by the four OECs, i.e., $a_n(j)$ for $j = 0, 1, 2, 3$ with:

$$\hat{F}(n,m) = \Sigma_{j=0}^{3} a_n(j)\phi_j(m/M_n) \text{ for } m = 0 \sim M_n \quad (2)$$

The four elements, i.e., $a_n(j)$ for $j = 0, 1, 2, 3$ in order represent respectively the mean, slope, acceleration and curvature of the logF0 contour.

B. Typical logF0 Patterns for Tones and Tone Values

Fig. 1a and Fig. 1b show average logF0 contours $\Sigma_{j=0}^{3}\beta_t(j)\phi_j(m/M)$, i.e., average RPCs, for the five lexical tones and perceived tones, respectively; where $\beta_t(j)$ is the average OEC for tone t calculated from all the syllable samples of the Romance corpus:

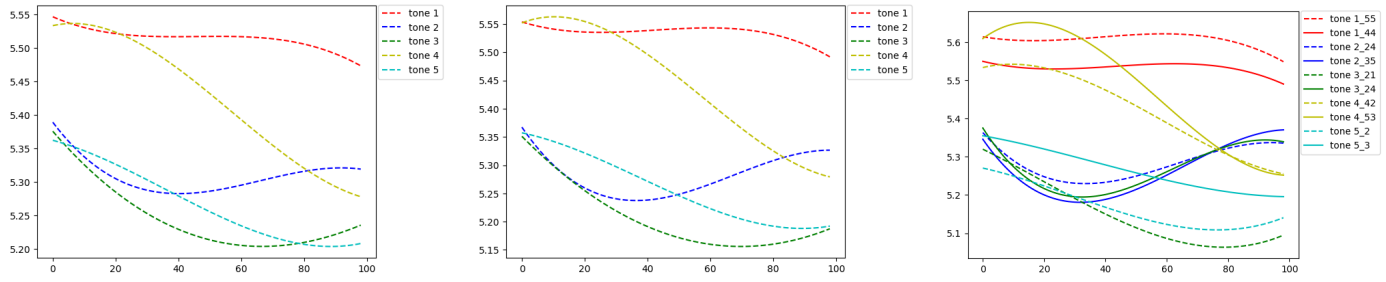
$$\beta_t(j) = (\Sigma_{n=1}^{N} a_n(j)\delta(t_n = t))/(\Sigma_{n=1}^{N}\delta(t_n = t)) \quad (3)$$

where t_n is the n-th syllable's lexical or perceived tone class.

It was found that the average logF0 for perceived T1 is more flat than the one for lexical T1. In addition, perceived T2 and T4 have more significant curvatures than lexical T2 and T4. We then calculate the mean square errors (MSEs) of the fitting for logF0 contour represented in OEC, a_n by the typical tone patterns, β_{t_n} , i.e., $\mathbb{E}[\|a_n - \beta_{t_n}\|_2^2]$. It is found that the MSE for the lexical tones (0.0761 ($\log \text{Hz}^2$)) is larger than the one for perceived tones (0.0704 ($\log \text{Hz}^2$)). This infers that the speakers may utter different or erroneous tones from lexical tones. The perceived tones labeled by humans could better represent tone classes of syllables.

Fig. 1c shows average RPCs for the ten significant tone values as shown in Table III paired with five lexical tones. For T1, tone value 55 has a higher logF0 level than tone value 44, matching the prior knowledge about tone values' pitch level. For T2, tone value 24 is not shown paralleled with tone value 35, manifesting a different logF0 from our prior knowledge about tone values 24 and 35. The two pitch contours of T2 are not significantly different at the beginning and end and do not conform to the expert's markings. The apparent difference in slope between the two contours may indicate that the acceleration of the pitch contour affects the perception of pitch by humans. T3 has two major tone values of 21 for low dipping and 24 for tone three Sandhi. For T4, tone value 53 shows a higher pitch beginning than tone value 42. The neutral tone (T5) with tone value 3 is higher in pitch than tone value 2. The level-four pitch in set "24" is obviously not at the same height as level-four in set "44", and the same situation also occurs in level-five in sets "35" and "55", which is probably due to coarticulation. The above findings indicate that the average logF0 contours (RPCs) of tone values in continuous speech do not generally match well with conventional 5-level tone values suggested by linguistics. Further investigation is required for future studies.

We also calculate the MSE of the fitting for logF0 contour by the typical tone value patterns as shown in Fig. 1c. The MSE for the tone value classes is 0.0565 $\log \text{Hz}^2$, which is smaller than the MSEs for the lexical and perceived tones. The reduction of the MSE infers that tone value could give a



(a) lexical tones (b) perceived tones (c) tone value classes

Fig. 1: Typical logF0 contours for the lexical tones (a), perceived tones (b), and tone values (c) calculated from the Romance L2 Mandarin speech corpus. Note that the legends in (c), i.e., tone_x_yy denotes lexical tone x and tone value yy

more sophisticated description of logF0 contours for describing non-native L2 Mandarin prosody.

IV. Automatic Tone Value Recognition

The Romance L2 Mandarin speech corpus is used to conduct tone value recognition experiments with a transformer-based recognition model, which is a stack of three modified transformer encoder layers. The architecture of the recognition model is the same as the one used in the previous tone labeling [15] and tone representation experiment [24], achieving a tone recognition accuracy of 87.48%, comparable to the related studied reported [27]. The input feature vector of the recognizer for each n-th syllable consisted of 81 10ms-interval Mel-spectrogram around the target syllable and information about syllable boundary and frame index. Since the Romance L2 Mandarin corpus is too small to construct a robust recognition model, we adopted tone recognition models trained with AISHELL-1, AISHELL-3, and NER corpora as pre-trained models. We then fine-tuned these pre-trained models with the Romanace L2 Mandarin corpus to obtain more robust models than the ones without the pre-training.

A. Pre-trained Model

The pre-trained models are trained with the corpora considered in Section II, validated, and evaluated with the validation and test sets of AISHELL-3, which are the same sets used in the previous study [15]. As shown in Table IV, increasing training data indeed enhances the model's accuracy in tone recognition. Although Model 3's overall accuracy is higher than Model 2's, Model 2 has a higher accuracy for T2. This is mainly because the AISHELL-1 and NER datasets were not manually corrected; the tone labels were taken directly from the lexicon without considering tone Sandhi in consecutive T3 syllables. To address this problem, we re-labeled the training data for Model 3 by the re-labeling strategy proposed in [15]. The overall accuracy of the re-labeled Model 3 is 90.7%, slightly higher than the original Model 3. Additionally, the accuracy of each tone is higher than that of Model 2. This demonstrates that the tone re-labeling method can effectively address the tone Sandhi issue in tone labeling.

TABLE IV: The accuracies of tone recognition experiments on the test set of AISHELL-3 with various training datasets.

	training data	syllable#	overall accuracy (%) acc. of T1/T2/../T5
Model 1 (baseline)	AISHELL-3	774,714	87.4 90.3/87.2/80.0/91.3/75.2
Model 2	AISHELL-1&3	2,728,112	89.4 91.0/88.7/83.9/93.1/79.4
Model 3	AISHELL-1&3+ NER	37,399,546	90.4 92.9/87.9/85.5/94.2/83.4
relabeled Model 3	AISHELL-1&3+ NER	37,399,546	90.7 92.6/89.6/85.2/94.2/83.8

B. Tone and Tone Value Recognitions for the L2 Dataset

We construct recognition models for multi-speaker tone and tone value recognition tasks on the Romance L2 Mandarin corpus. In both tasks, we use the utterances corresponding to the first short text in the corpus as training data (2084 syllables) and the rest of the utterances as testing data (2179 syllables). Table V shows the tone recognition results for the Romance L2 Mandarin dataset. Note that the L2 dataset is labeled with perceived tones by the three native Mandarin-speaking annotators. The tone labels made by each of the annotators are regarded as one set, resulting in three training and test sets corresponding to the three annotators. We, therefore, construct the three tone recognition models with the three training and test sets to evaluate performance.

The strict-sense accuracies were calculated by averaging the accuracies of three sets labeled by the three annotators. On the other hand, the non-strict sense accuracies considered a predicted tone correct if it is one of the tones in the three sets labeled by the three annotators. We found that the larger the dataset used to train the pre-trained model, the better the performance after fine-tuning. Contrary to expectations, the relabeled Model 3 did not achieve better results. We found the accuracy drop from Model 3 to relabeled Model 3 is due to the degradation of tone recognition accuracies for T4 and T3 syllables.

For the tone value recognizer, we changed the output layer dimension from 5 (five tones) to 41 (tone value classes labeled by the annotators). As shown in Table VI, similar to the tone recognition task, the relabeled model achieved lower performance both before and after fine-tuning. Note that the

TABLE V: The strict and non-strict sense tone recognition accuracies with various model settings: a) W/O represents that the models are trained with the Romance L2 dataset only, and b) Model 1/2/3 and relabeled Model 3 are fine-tuned from the pre-trained models reported in Subsection IV.A. Note that the numbers left to and right to the arrows are accuracies made by the models before and after fine-tuning to the L2 dataset, respectively.

pretrained model	strict sense acc.(%)	non-strict sense acc.(%)
W/O	47.2	53.6
Model 1	58.8→65.7	66.5→73.2
Model 2	59.9→67.3	67.7→74.5
Model 3	71.2→75.0	79.6→83.1
relabeled Model 3	70.7→74.7	78.9→82.5

calculation of the strict and non-strict sense accuracies follows the definition for the tone recognition results reported in Table V. In addition to using the proposed pre-trained models, we also experimented with the upstream self-supervised learning (SSL) model, XLS-R [28], for the task of tone value recognition. We changed the input parameters from Mel-spectrograms to representations of the output of the ninth layer of XLS-R (the best-performing layer tested) and implemented it using a simple DNN model with 3 layers (1024x512x256). Although the result obtained is better than that using no pre-trained model, it is worse than those using the proposed pre-trained model.

Table VII shows the confusion matrix for tone value recognition. We found that although the overall accuracy for tone value recognition was only 47.4%, the accuracy rates for 6 of the 9 major tone value classes exceeded this number. This result is highly consistent with the distribution of training samples. For example, the tone value '53', which has the largest number of samples, achieved an accuracy of 84.62%, while the tone value '2', which has the smallest number of samples, had the worst performance, with an accuracy of only 14.47%. The tone value classes regarding level pitch contours, i.e., 44 and 55, are easily mutually confused. The tone value classes regarding falling pitch contours, i.e., 42 and 53, are also easily mutually confused. The rising-pitch tone value classes, i.e., 24 and 35, would be jointly confused with the standard T3 lower dipping tone value, i.e., 21, maybe due to tone Sandhi. The tone values corresponding to neutral tones, i.e., 2 and 3, are easily confused jointly with the tone value 21. These confusions between the tone value classes seem related to the similarities between the typical logF0 patterns of tone value classes obtained in Fig. 1c, suggesting that the tone value recognition result is reasonable.

V. Conclusions and Future Works

This study explored the relationship between tone and tone values, visualized the typical pitch contours of tone values, and attempted to train tone and tone value recognition models for the L2 Mandarin speakers. For tone recognition, fine-tuning can indeed improve accuracy, but the accuracy for the L2 speakers is much lower than that for the L1 speakers (in the AISHELL-3 test set). For tone value recognition,

TABLE VI: The strict and non-strict sense tone value recognition accuracies with various model settings: a) W/O represents that the models are trained with the Romance L2 dataset only, b) Model 1/2/3 and relabeled Model 3 are fine-tuned from the pre-trained models reported in Subsection IV.A, and c) XLS-R is a cascade of the pre-trained SSL model, XSL-R [28], with a downstream trainable DNN.

pretrained model	strict sense acc. (%)	non-strict sense acc. (%)
W/O	36.0	47.2
Model 1	46.6	62.4
Model 2	47.5	65.0
Model 3	47.4	66.1
relabeled Model 3	46.0	64.8
XLS-R	42.8	56.0

TABLE VII: The stripped-down confusion matrix of the nine major tone value classes by the fine-tuned Model 3 in the strict sense.

tone value	top-3 recognized tone value classes and percentages (%)
55	55 (71.1%), 44 (13.1%), 53 (8.1%), others (7.8%)
44	44 (42.4%), 55 (30.6%), 42 (10.2%), others (16.8%)
24	24 (69.3%), 21 (11.1%), 35 (5.7%), others (13.8%)
35	35 (57.4%), 24 (22.9%), 21 (6.8%), others (12.9%)
21	21 (63.6%), 24 (11.3%), 32 (8.9%), others (16.2%)
42	42 (56.2%), 53 (27.2%), 44 (4.4%), others (12.2%)
53	53 (84.6%), 42 (10.7%), 55 (1.8%), others (2.9%)
2	21 (42.1%), 2 (14.5%), 32 (10.1%), others (33.3%)
3	3 (31.1%), 21 (13.7%), 32 (13.7%), others (41.5%)

although the model trained using our proposed pre-trained model performs better than the model trained using the SSL upstream representations (XLS-R), the accuracy for most of the tone values is still not high. Based on the experimental results, we propose the following future research directions:

1) Analyze the samples of tone recognition errors in L1 and L2 corpora to understand the characteristics of recognition error samples for each tone class and the specific differences between L1 and L2 speakers in the pronunciation of tones.
2) Use the proposed tone value recognition model to label the unlabeled L2 corpus, conduct linguists analyze whether this method can accelerate their manual tone value annotation works, and improve the quality of Mandarin language teaching.
3) Add corpora from various languages to train the pre-trained model to improve the model's ability to analyze pitch information from the spectrogram, thereby enhancing the performance of the tone and tone value recognition model.

References

[1] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving Mispronunciation Detection of Mandarin Tones for Non-Native Learners With Soft-Target Tone Labels and BLSTM-Based Deep Tone Models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2012–2024, Dec. 2019, ISSN: 2329-9290, 2329-9304.

[2] Y.-S. Cheng, T.-H. Lo, and B. Chen, "Exploiting text prompts for the development of an end-to-end computer-assisted pronunciation training system," in *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, 2020, pp. 290–303.

[3] Z. Gan, T. Zhao, X. Yu, and H. Yang, "Deep feedforward sequential memory networks based mispronunciation detection for tibetan students' mandarin," in *2021 2nd International Conference on Information Science and Education (ICISE-IE)*, IEEE, 2021, pp. 509–512.

[4] Y. Chao, *Tone, Intonation, Singsong, Chanting, Recitative, Tonal Composition, and Atonal Composition in Chinese*. 1956.

[5] Y. Chao, *A Grammar of Spoken Chinese*. University of California Press, 1968, ISBN: 978-0-520-00219-7.

[6] C. Zhu, *Pronounce Chinese as Chinese*. Beijing: Yuwen Publishing, 1997.

[7] C. Zhu, *Pronounce Chinese as Chinese (new edition)*. Taiwan: New Xue Lin Publishing Co., Ltd., 2013.

[8] C.Shih, "The phonetics of the chinese tonal system," AT & T Bell Laboratories, Tech. Rep., 1986.

[9] J. A. Edmondson, J.-L. Chan, G. B. Seibert, and E. D. Ross, "The effect of right-brain damage on acoustical measures of affective prosody in Taiwanese patients," *Journal of Phonetics*, vol. 15, no. 3, pp. 219–233, 1987, ISSN: 0095-4470.

[10] J. Fon and W.-Y. Chiang, "What does chao have to say about tones?-a case study of taiwan mandarin," *Journal of Chinese Linguistics*, vol. 27, no. 1, pp. 13–37, 1999.

[11] P. Rose, "Considerations in the normalisation of the fundamental frequency of linguistic tone," *Speech communication*, vol. 6, no. 4, pp. 343–352, 1987.

[12] X. Zhu *et al.*, "Shanghai tonetics," 1995.

[13] J.-S. Jhang, *Let's study Thai language*, Xiu ding yi ban. Xinbei Shi: Uni-President Publishing House Co., Ltd., 2022, OCLC: 1373231644, ISBN: 978-986-99369-8-9.

[14] J.-M. Liang, *Learn basic Thai in 7 days(in chinese)*, first edition. Taipei: Wo-Shih publishing house, 2014, OCLC: 900451375, ISBN: 978-986-5785-42-0.

[15] W.-H. Li, C.-Y. Chiang, and T.-H. Liu, "Tone labeling by deep learning-based tone recognizer for mandarin speech," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 873–880.

[16] C.-Y. Chiang, W.-H. Li, Y.-T. Lin, *et al.*, "The speech labeling and modeling toolkit (slmtk) version 1.0," in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2022, pp. 1–5.

[17] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[18] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," in *Proc. Interspeech 2021*, 2021, pp. 2756–2760.

[19] Y.-F. Liao, Y.-H. S. Chang, Y.-C. Lin, W.-H. Hsu, M. Pleva, and J. Juhar, "Formosa Speech in the Wild Corpus for Improving Taiwanese Mandarin Speech-Enabled Human-Computer Interaction," en, *Journal of Signal Processing Systems*, vol. 92, no. 8, pp. 853–873, Aug. 2020, ISSN: 1939-8018, 1939-8115.

[20] S.-H. Chen and Y.-R. Wang, "Vector quantization of pitch information in Mandarin speech," en, *IEEE Transactions on Communications*, vol. 38, no. 9, pp. 1317–1320, Sep. 1990, 24, ISSN: 00906778.

[21] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, "Unsupervised joint prosody labeling and modeling for Mandarin speech," en, *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1164–1183, Feb. 2009, ISSN: 0001-4966.

[22] S.-H. Chen, C.-H. Hsieh, C.-Y. Chiang, *et al.*, "Modeling of Speaking Rate Influences on Mandarin Speech Prosody and Its Application to Speaking Rate-controlled TTS," en, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1158–1171, Jul. 2014, 09, ISSN: 2329-9290, 2329-9304.

[23] C.-H. Lin, C.-L. You, C.-Y. Chiang, Y.-R. Wang, and S.-H. Chen, "Hierarchical prosody modeling for Mandarin spontaneous speech," en, *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2576–2596, Apr. 2019, 41, ISSN: 0001-4966.

[24] W.-H. Li, T.-H. Liu, and C.-Y. Chiang, "Tone Value Representation for Computer-Assisted Pronunciation Training," in *Proc. Speech Prosody 2024*, 2024, pp. 712–716.

[25] C. Y. Chiang, "Cross-Dialect Adaptation Framework for Constructing Prosodic Models for Chinese Dialect Text-to-Speech Systems," en, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 108–121, Jan. 2018, 16, ISSN: 2329-9290, 2329-9304.

[26] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995, 22.

[27] J. Tang and M. Li, "End-to-end mandarin tone classification with short term context information," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 878–883.

[28] A. Babu, C. Wang, A. Tjandra, *et al.*, *Xls-r: Self-supervised cross-lingual speech representation learning at scale*, 2021. arXiv: 2111.09296 [cs.CL].

GMNER-LF: Generative Multi-modal Named Entity Recognition Based on LLM with Information Fusion

Huiyun Hu†, Junda Kong†, Bo Xiao†*, Fei Wang‡, Hongzhi Sun‡ and Yang Ge‡
†Beijing University of Posts and Telecommunications, Beijing, China
‡State Grid Shandong Electric Power Company Dezhou Power Supply Company, Dezhou, China
E-mail: hhy150@bupt.edu.cn kongjd@bupt.edu.cn xiaobo@bupt.edu.cn
Wf6102@163.com 328401563@qq.com 15005340985@163.com
*Corresponding Author

***Abstract*— Multi-modal Named Entity Recognition (MNER) leverages visual information to enhance the effectiveness of text-only Named Entity Recognition (NER). Currently, many methods are based on sequence labels, but the model architecture is relatively complex. Large Language Model (LLM) has recently demonstrated powerful generative and comprehension abilities, so we propose GMNER-LF, a new paradigm for MNER with generative method. Firstly, we retrieve relevant image-text pairs to provide prior knowledge for recognition. Secondly, we construct our task into MRC task so that the LLM can better understand the problem. In addition, we designed a multi-modal fusion module and added a gating mechanism to help filter noise information in the image to obtain high-quality fusion representations. The multi-modal fusion module is injected into the LLM block to achieve deep fusion of LLM and multi-modal representations and fully explore the internal knowledge of LLM. The proposed method not only fully explores the internal knowledge of LLM, but also filters important modal information through the gating mechanism. Experimental results show that compared with other generative methods, the proposed method has improved performance on both Twitter-2015 and Twitter-2017 datasets.**

I. INTRODUCTION

Named Entity Recognition (NER) is a crucial component of information extraction, focusing on the identification and classification of name entities in text, including times, locations, and personal names [1]. This task plays a crucial role in extracting key information from the text, enabling models to better comprehend the semantic content and perform downstream tasks more efficiently. Examples of these tasks include information extraction [2], machine translation [3], and text generation [4].

Social media platforms such as Twitter and Instagram have provided a significant amount of user-generated content in the form of text and images. The phenomenon of text information needing to be understood through related images is becoming increasingly common. However, traditional NER methods focus on mining semantic information from limited text, resulting in a lack of effective information and limited model generalization ability. To solve this problem, some researchers propose to introduce multi-modal data information into NER to supplement the effective semantic information missing from uni-modal text data.

Traditional NER task primarily relies on statistical learning or deep learning methods to identify and extract entities such as personal, locations, and organizations names from textual data. Early MNER research concentrated on simple combinations of images and text to enhance entity recognition performance through joint training models. For instance, some studies utilized Convolutional Neural Networks (CNNs) for processing image information and combined Recurrent Neural Networks (RNNs) or Transformer architectures for processing text information in joint learning on multi-modal data [5,6].

With the advancements in deep learning and Graph Neural Networks (GNN), MNER research has made remarkable progress. Recent research methods include the use of Multi-head Attention mechanisms and multi-channel Graph Neural Networks (MGNN) to capture complex relationships within graphic data. These approaches significantly improve entity recognition accuracy by establishing more fine-grained associations between images and text [7,8].

In recent years, transformer architecture is ideal for handling multi-modal data due to its ability to capture long-distance dependencies and perform parallel computations. The combination of pre-trained language models, such as BERT and Roberta, with visual models, such as Vision Transformers (ViT), significantly enhances the effectiveness of multi-modal information processing. Additionally, the emergence of large language models, such as GPT-3 and T5, provides robust support for MNER through their powerful generative and contextual understanding capabilities [9]. These models excel at understanding and generating cross-modal information, enabling the deep integration and accurate identification of multi-modal data.

Generally, NER methods based on CNN or RNN still have shortcomings due to the limitations of their models in the deep fusion of multi-modal information. Additionally, GNN computation usually involves more complex graph structure

operations and may experience over-smoothing (i.e., feature homogenization, where node features become increasingly similar) during multi-layer information propagation [10]. Moreover, traditional text LLM faces challenges in learning information from image features [11]. Therefore, this paper proposes a multi-modal named entity recognition method that utilizes LLM and deep modal fusion to accomplish the generative MNER task. The main contributions of this paper are as follows:

1. We propose a new paradigm for generative multi-modal named entity recognition based LLM with information fusion (GMNER-LF). We treat multi-modal named entity recognition tasks as generative tasks, utilizing machine reading comprehension (MRC) methods to fully exploit LLM generation capabilities, and introducing prior knowledge through retrieval.

2. We fully excavate the internal knowledge of LLM and use it as the initial value for training specific MNER tasks to quickly converge and obtain a new model suitable for new tasks. Our approach fully explores the knowledge of images, texts and LLM.

II. Related work

This section introduces the work related to MNER. The task of MNER can be categorized into two types: sequence label-based MNER and generation-based MNER.

Sequence label-based MNER

Many studies approach the task of MNER as a sequence labeling task, requiring the use of CRF layer for decoding. In this process, researchers must carefully consider how to effectively utilize features from the image.

ITA [12] overcomes the challenge of aligning multiple modalities by refraining from directly using raw visual characteristics and instead adopting Optical Character Recognition (OCR) and image captioning methodologies to convey the content of images. HVPNeT [13] conducts hierarchical processing at the visual level to enhance model robustness against image noise. MoRe [14] enhances its model by retrieving relevant vision and text data, utilizing MoE module to fuse image and text encoding, which is then input into Transformer and CRF layer for prediction in a sequence labeling task.

With recent advancements in LLMs, many studies incorporate LLM to further improve performance in MNER tasks. However, LLM is often used as an auxiliary tool while continuing with sequence labeling tasks. PGIM [15] obtains the caption of the image, inputs it into LLM to obtain auxiliary text, and inputs it together with the original text into Transformer and CRF layer for predictions.

Generation-based MNER

Some studies have explored the use of generative methods for NER, particularly in uni-modal named entity recognition and information extraction. Many examples exist of using LLMs to participate in generative work.

Ma et al. focus on the few-shot ability of fine-tuned SLMs and propose a method to complement SLM with LLM through appropriate incentive strategies, addressing Hard Samples that SLM cannot handle [16]. Wang et al. employ a method of altering task logic to generate special tags next to the entity to be extracted after detecting the entity, transforming the problem of text sequence labeling into a problem of plain text generation that can be easily adapted by LLM [17]. Ashok and Lipton concentrate on prompt learning for LLM. Through training with few-shot and entity definitions, LLM can generate a list of potential entities when facing text, striving for compatibility and interpretability with the entity definitions to improve NER task accuracy [18].

Similar results based on generation are still relatively rare in the field of multi-modal tasks. Some studies utilize the ICL ability of LLM to retrieve relevant image-text pairs as demonstrated in ICL and then treat MNER as a generative task [19]. This approach is relatively primitive in processing image and text features, making it more dependent on the strong performance of subsequent generative models.

III. Model Architecture

In this section, we commence by defining the task, subsequently providing a comprehensive description of the proposed model.

Task Formulation. Given a text $\boldsymbol{T}$ and its associated image $\boldsymbol{I}$ pair $\boldsymbol{(I,T)}$ as input, where $\boldsymbol{T}=\{t1, ..., tn\}$ and n denotes the length of the text, the goal of MNER is to extract a set of entities from $\boldsymbol{T}$ and classify each extracted entity into one of the predefined types with the assistance of image information, e.g. Person (PER), Location (LOC), Organization (ORG), and Miscellaneous (OTHER). Unlike most previous work, we treat this task as a generative task here. Specifically, taking $\boldsymbol{Y}$ as the final generated result of the model, there can be various styles of $\boldsymbol{Y}$, and we will extract the recognized entity content from $\boldsymbol{Y}$.

Fig. 1 illustrates the overall architecture of GMNER-LF, which consists of four major components: (1) Retrieval augment module, which retrieves relevant information to provide prior knowledge for the model. (2) Uni-modal feature extraction module, which encode individual modalities separately. (3) Multi-modal fusion module, which effectively filters and fuses semantic information from various modalities. (4) Large language model generation module, which apply the large language model module to generate prediction results.

A. Retrieval augment module

We can use external relevant knowledge to assist in recognizing entities. In MNER task, it is necessary to retrieve relevant image-text pairs as data augmentation. The retrieval task of image-text pairs can generally be achieved through text-text retrieval, image-image retrieval, or image-text retrieval.

Relatively speaking, combining image and text retrieval will be more accurate. Here, we use the similarity between images and text for retrieval. Specifically, for the retrieval base $\boldsymbol{U}$ and the target image-text pair sample $Target(Tar_i,\ Tar_t)$, we use the Vision encoder E_i and the Text Encoder E_t to encode the image and text respectively. For the sample $u(u_i,u_t)$ in the retrieval base, we calculate the similarity to the target sample image and the similarity to the target sample text, respectively. And the total similarity between u and the *Target* sample is:

$$Sim = \lambda \cos(E_i(Tar_i), E_i(u_i)) + \cos(E_t(Tar_t), E_t(u_t)). \quad (1)$$

Where λ is a hyper-parameter that assigns different weights to different modalities. Calculate the similarity between the target sample and the samples in the retrieval base and select the image-text pairs with Top K similarity as the knowledge enhancement samples.

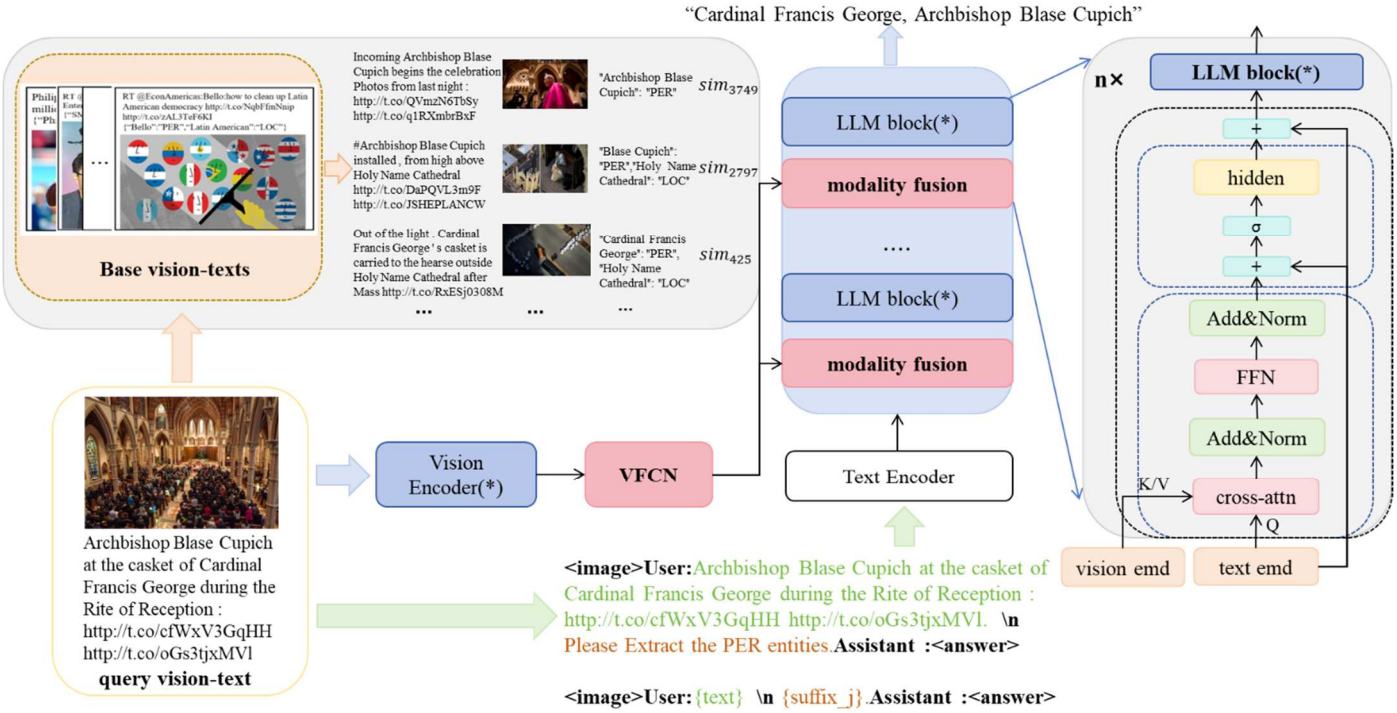


Fig. 1 The main architecture of GMNER-LF.

B. Uni-modal feature extraction module

We need to encode each modality, which involves encoding two categories: image and text.

Visual Encoder. We adopted a frozen CLIP vision transformer encoder that takes Image $I \in R^{C\times H\times W}$ as input, tokenises image into M patches, and compute dense visual representation, $Z = \Phi_I(I) \in R^{M\times d}$. Where d is the embedding dimension of the image mapping.

Text Encoder. The universal method is utilizing the Bert model for encoding and freezing the parameters of it. Here, we use the pre-trained transformer embedding layer in LLM for encoding to better adapt to the subsequent modules. What's more, we update the parameters of this module.

The text input prompt P_t is formed by combining image. Here, the text is converted into MRC format to make it easier for subsequent LLM to understand semantics and answer questions. We will analyze the impact of different prompt constructions on the results later. The prompt is denoted as:

$$P_t = <image>User:\{T\}.\{\text{suffix}_j\}Assistant:<answer>. \quad (2)$$

Where <image> and <answer> are special media tokens representing the image and answer introduction symbol in text format, respectively. In this task, $suffix_j$ represents the instruction for different entities, E is the number of entity types and $j \in \{1,...,E\}$. $\boldsymbol{T}$ is input text.

C. Multi-modal fusion module

Visual Feature Compression Net (VFCN). It models the relationship between visual features and compresses them into a bottleneck feature of a fixed length N, where typically N is significantly smaller than M. We can obtain learnable query token embedding $q \in R^{N\times d}$.

Multi-modal fusion module. This module is used to integrate visual and textual information. Visual information is processed through Visual Feature Compression Net to obtain fixed length query token embedding. Input text embedding and query token embedding into cross attention for information fusion. Due to some information in the image not being associated with the text, it is necessary to filter out this information in the fusion module, and a gating mechanism is added here. The formula for filter gate g is as follows:

$$g = \sigma(W_a A + W_b B), \quad (3)$$

$$K = g \otimes B, \quad (4)$$

$$H = A + K. \quad (5)$$

Where W_a and W_b are both weight matrices, $\boldsymbol{A}$ is the text embedding, $\boldsymbol{B}$ is the image-text cross embedding, σ is the activation function, $g \in R$ and it belongs to [0,1]. $\otimes$ represents multiplying point by point. We add $\boldsymbol{K}$ and $\boldsymbol{A}$ to obtain final image-text fusion representation $\boldsymbol{H}$.

D. Large language model generation module

In this module, we generate named entity recognition results by combining the LLM block and multi-modal fusion module mentioned earlier. Specifically, text embedding and image embedding are as input before entering the first layer. The text embedding is recognized as query, and the image embedding is recognized as key and value to input into the multi-modal fusion module of the first layer. Then, the output is input into the LLM block to obtain the latent vector, which is then input into the next multi-modal fusion module as a query input, the image embedding is still recognized as key and value, and so on. Finally, the output latent vector is input into the Linear layer to obtain the prediction. Choose the most likely generated sequence as the anticipated output:

$$\hat{y} = \max P_\theta(y \mid text, image). \quad (6)$$

Where θ represents the model parameters that need to be trained, y represents the label results, and $\hat{y}$ represents the generated results. Lastly, we decode the predicted output $\hat{y}$ into the specified form and perform post-processing on the results, thereby completing the MNER task.

E. Modal Training

As shown in Fig. 1, the parameters of the Vision Encoder model and LLM block model are frozen, while the parameters of Text Encoder, VFCN, and multi-modal fusion modules should be updated during training. The training objective aims to maximize the following conditional probability:

$$LM_loss = -E_{(I,T)} \log P(T_m \mid T_{<m}, I). \quad (7)$$

Where T_m and $T_{<m}$ denote the masked tokens and the tokens before them, respectively. And $\boldsymbol{I}$ denote the image in the sample. This loss leverages conditional causal masked language modeling to bolster the model's capacity for producing named entity recognition results that encompass various modalities.

IV. Experiments

A. Benchmark Datasets

We conducted experiments on Twitter-2017 [20] and Twitter-2015 [21], each sample consisting of image-text pairs. The specific statistics about two datasets are shown in Table I.

B. Evaluations

Precision, Recall, and F1 are used as the evaluation metrics in this work.

$$precision = num_{right} / num_{pred_entites} \quad (8)$$

$$recall = num_{right} / num_{gt_entites} \quad (9)$$

$$F1 = 2 \times recall \times precision / (recall + precision) \quad (10)$$

Where num_{right}, $num_{pred_entites}$, and $num_{gt_entites}$ represent the number of correctly predicted entities, the total number of predicted entities, and the total number of entities in the sample, respectively.

Table I: Statistics of Twitter-2015 and Twitter-2017.

Entity Label	Train	Dev	Test	Total
Twitter-2015				
Person	2217	552	1816	4585
Location	2091	522	1697	4410
Organization	928	247	839	2014
Miscellaneous	940	225	726	1891
Total Entity	6167	1546	5078	12791
Samples num	4000	1000	3257	8257
Twitter-2017				
Person	2943	626	621	4190
Location	731	173	178	1082
Organization	1674	375	395	2444
Miscellaneous	701	150	157	1008
Total Entity	6049	1324	1351	8724
Samples num	3373	723	723	4819

C. Baselines

We compare our approach against a generative method. It utilizes LLM and In Context Learning (ICL) to generate entity results. Specifically, it first retrieves relevant image-text pairs as demonstrations in ICL to guide entity recognition of the target sentence. For images, it uses a caption model to obtain image captions, combines them with the original text, and inputs them into LLM.

D. Experiment Details

Model Configuration. In the uni-modal feature extraction module, we use ViT-L/14 in CLIP as the Vision Encoder to encode the image. And we select the embedding layer in MPT-1B as the Text Encoder. The LLM block uses the decoder layer from MPT-1B. VFCN uses 6 layers of Transformer layers. The retrieval module uses the text encoder and image encoder of CLIP with ViT-B/32 for text and image embedding respectively and selects relevant candidates from Top 5 for subsequent retrieval augment.

Training Configuration. This work is based on the operating systems Ubuntu 18.04, Python 3.9.19, and PyTorch 2.0.1, and we conduct experiments on an NVIDIA A800 GPU. In addition, we use the AdamW optimizer to train the model in the experiment, and hyper-parameter learning rates selected within the range of {1e-3, 1e-4, 5e-5, 1e-5}. We use a learning

rate of 1e-4 and a batch size of 100 for fine-tuning the model. The model is trained for 5 epochs. And the similarity parameter λ of the retrieval module is set to 0.5. And We utilize PyTorch and Transformers for implementation.

Baseline Configuration. Referring to the original configuration, we still use its retrieval module and adopt the OFA model as the caption model. For LLMs, we choose Qwen-7B and Llama-7B for the experiments.

E. Results And Analysis

In Table II, we present the pivotal experimental outcomes and subsequently derive the ensuing conclusions.

Table II: Experiments Results on two datasets.

Methods	Single Type (F1)				Overall (%)		
	PER	LOC	ORG	OTH	P	R	F1
Twitter-2015							
ICL(Qwen)	0.49	0.28	0.24	0.05	60.57	22.09	32.37
ICL(Llama)	**0.70**	0.52	0.37	0.11	63.38	42.95	51.20
GMNER-LF	0.67	**0.72**	0.55	0.37	**72.47**	55.25	62.70
GMNER-LF(Aug)	0.66	**0.72**	**0.58**	**0.38**	71.99	**56.09**	**63.05**
Twitter-2017							
ICL(Qwen)	0.65	0.27	0.40	-	68.89	38.84	49.67
ICL(Llama)	**0.77**	0.33	0.49	-	71.53	50.64	59.30
GMNER-LF	0.61	**0.73**	0.72	-	75.67	59.80	66.80
GMNER-LF(Aug)	0.70	**0.73**	**0.74**	-	**80.09**	**65.20**	**71.88**

It is worth mentioning that due to image data issues in Twitter-2017, samples from the test dataset that could not be read were removed. Therefore, there is no OTHER type in the test data, so there is no F1 score for it.

From the results in Table I, it can be seen that compared to the baseline method, the F1 score of our method has been improved, which proves the effectiveness of this method. Unlike the baseline method, our method integrates the original image features into the model and trains them together with text features, rather than converting the image into text information, which enhances the image information that matches the text to a certain extent.

What's more, the baseline method has better recognition performance for PER Type, possibly due to the presence of more personal name information in internal knowledge of LLM, but it is poor to recognize other entities. Our method has comparable recognition performance for PER, LOC, and ORG. However, for OTHER, it has some noise, so it may be more difficult to recognize. Nevertheless, there has been a significant improvement compared to the baseline, demonstrating the robustness of our method.

In addition, we see that after adding augment part, almost all metrics have improved, indicating the effectiveness of the retrieval augment module in adding prior knowledge.

Different prompts and responses have a certain impact on the performance of the model. Therefore, we test two different prompt and response forms, one of which is dictionary format, where prompt is sentence and response is {entity1: type1, entity2: type2}, such as {'Jeff Lockie':'PER', 'Finals MVP': 'OTHER'}. Another format is MRC (machine reading comprehension), which is organized into a triplet of (sentence, query, answer). In our experiment, four types of queries were constructed for one sentence in the prompt, each of which recognizes entities of four types. The response is an answer, for example, if we want recognize 'PER', the response is 'Jeff Lockie'.

Table III: Compare with different prompts.

Prompt	Single Type (F1)				Overall (%)		
	PER	LOC	ORG	OTH	P	R	F1
Twitter-2015							
Dict	0.59	0.59	0.49	0.33	70.88	42.95	53.49
MRC	**0.67**	**0.72**	**0.55**	**0.37**	**72.47**	**55.25**	**62.70**
Twitter-2017							
Dict	0.51	0.55	0.56	-	65.68	45.28	53.60
MRC	**0.61**	**0.73**	**0.72**	-	**75.67**	**59.80**	**66.80**

We compared two types of prompts and response forms, and the experimental results are shown in Table III. From the data in the table, it is evident that the MRC form is more effective. Intuitively speaking, dictionary format requires extracting all entities from one sentence and outputting them in a fixed format, which is somewhat difficult for the model. On the other hand, MRC format only needs to output corresponding entities of one type, with shorter length and more flexible format, making it relatively less difficult.

V. CONCLUSIONS

In this paper, we propose a new paradigm for generative multi-modal named entity recognition based on LLM with information fusion (GMNER-LF). We add relevant image-text pairs to provide prior knowledge and construct prompts and responses in the form of MRC. At the network structure level, we designed a modal fusion of the gating mechanism and integrated it into LLM to fully tap its internal knowledge. This method has achieved great results. In addition, we found in the experiment that the recall of the large language model in the named entity recognition task is generally low. It is possible that the generation of the large language model may be somewhat conservative. Subsequent work can be carried out along the line of improving the recall.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (NSFC No.62076031).

REFERENCES

[1] Li, J., Sun, A., Han, J. and Li, C., 2020. A survey on deep learning for named entity recognition. IEEE transactions on knowledge and data engineering, 34(1), pp.50-70.

[2] Fei, H., Wu, S., Li, J., Li, B., Li, F., Qin, L., Zhang, M., Zhang, M. and Chua, T.S., 2022. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. Advances in Neural Information Processing Systems, 35, pp.15460-15475.

[3] Schäfer, Henning, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. "Cross-language transfer of high-quality annotations: Combining neural machine translation with cross-linguistic span alignment to apply NER to clinical texts in a low-resource language." In Proceedings of the 4th Clinical Natural Language Processing Workshop, pp. 53-62. 2022.

[4] Bień, Michał, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. "RecipeNLG: A cooking recipes dataset for semi-structured text generation." In Proceedings of the 13th International Conference on Natural Language Generation, pp. 22-28. 2020.

[5] Moon, Seungwhan, Leonardo Neves, and Vitor Carvalho. "multi-modal Named Entity Recognition for Short Social Media Posts." In Proceedings of NAACL-HLT, pp. 852-860. 2018

[6] Wu, Zhiwei, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. "multi-modal representation with embedded visual guiding objects for named entity recognition in social media posts." In Proceedings of the 28th ACM International Conference on Multimedia, pp. 1038-1046. 2020.

[7] Yu, Jianfei, Ziyan Li, Jieming Wang, and Rui Xia. "Grounded multi-modal named entity recognition on social media." In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9141-9154. 2023.

[8] Sui, Yuan, Fanyang Bu, Yingting Hu, Liang Zhang, and Wei Yan. "Trigger-gnn: a trigger-based graph neural network for nested named entity recognition." In 2022 International Joint Conference on Neural Networks (IJCNN), pp. 01-08. IEEE, 2022.

[9] Li, J., Li, H., Sun, D., Wang, J., Zhang, W., Wang, Z. and Pan, G., 2024. LLMs as Bridges: Reformulating Grounded multi-modal Named Entity Recognition. arXiv preprint arXiv:2402.09989.

[10] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Philip, S.Y., 2020. A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32(1), pp.4-24.

[11] Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R. and McHardy, R., 2023. Challenges and Applications of Large Language Models. arXiv e-prints, pp.arXiv-2307

[12] Wang, Xinyu, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. "ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition." In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3176-3189. 2022.

[13] Chen, Xiang, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. "Good Visual Guidance Make A Better Extractor: Hierarchical Visual Prefix for multi-modal Entity and Relation Extraction." In Findings of the Association for Computational Linguistics: NAACL 2022, pp. 1607-1618. 2022.

[14] Wang, Xinyu, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. "Named Entity and Relation Extraction with Multi-Modal Retrieval." In Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 5925-5936. 2022.

[15] Li, Jinyuan, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. "Prompting ChatGPT in MNER: Enhanced multi-modal Named Entity Recognition with Auxiliary Refined Knowledge." In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 2787-2802. 2023.

[16] Ma, Yubo, Yixin Cao, Yong Hong, and Aixin Sun. "Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples!." In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 10572-10601. 2023.

[17] Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J. and Wang, G., 2023. GPT-NER: Named Entity Recognition via Large Language Models. arXiv e-prints, pp.arXiv-2304.

[18] Ashok, D. and Lipton, Z.C., 2023. Promptner: Prompting for named entity recognition. arXiv preprint arXiv:2305.15444.

[19] Cai, Chenran, Qianlong Wang, Bin Liang, Bing Qin, Min Yang, Kam-Fai Wong, and Ruifeng Xu. "In-context learning for few-shot multi-modal named entity recognition." In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 2969-2979. 2023.

[20] Lu, Di, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. "Visual attention model for name tagging in multimodal social media." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1990-1999. 2018.

[21] Zhang, Qi, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. "Adaptive co-attention network for named entity recognition in tweets." In Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1. 2018.

Novel Estimators for the Number of Susceptible Individuals in SIR Models of Infectious Epidemics

Michaël Antonie van Wyk*, André Martin McDonald†, David M. Rubin* and Fangfang Zhang‡
* University of the Witwatersrand, Johannesburg, Braamfontein, South Africa
E-mail: {anton.vanwyk, david.rubin}@wits.ac.za Tel: +27-11-717-7267
† Council for Scientific and Industrial Research, Pretoria, South Africa
E-mail: amcdonald@csir.co.za Tel: +27-12-841-4349
‡ Qilu University of Technology, Jinan, China
E-mail: zhff4u@qlu.edu.cn Tel: +86-151-6916-3982

***Abstract*—Accurate estimation of model parameters early in infectious epidemics may improve planning and resource allocation in mitigating the adverse consequences on affected populations. By applying the Peano-Baker series formula and the Cauchy repeated integral formula, we present the development of three novel estimators which facilitate the estimation of the number of susceptible people as a function of time $S(t)$ for an SIR model of an infectious epidemic. Association of these three estimators by combining them produces new estimators. We present the case for the new estimator, $\hat{S}_{1,3}$, derived from the association of two of the original estimators, $\hat{S}_1$ and $\hat{S}_3$, to estimate $S(0)$. This produces an estimate based on the history of net infection rate, $I'(t)$, from time 0 to t. By assuming parameters from the literature for the spread of COVID-19 in Wuhan, we run numerical simulations starting with an infection rate $I'(t)$ and adding filtered Gaussian noise. Discretization produces inaccuracy in bias and variance, however $\hat{S}_1$ and $\hat{S}_3$ yield accurate figures for $S(t)$, despite noise contamination. $\hat{S}_{1,3}$ also yields accurate figures for $S(0)$, with improvements as more observations are accumulated. We plan to investigate other novel estimators and further study their performance with real-world data.**

I. Introduction

The ability to predict the likely course of an infectious epidemic is critical for planning and the allocation of resources to assist in mitigating the adverse effects on the population. This became abundantly evident during the COVID-19 pandemic, and is arguably the most important reason for the need to develop highly predictive models and techniques for the accurate, early estimation of model parameters.

Various deterministic compartmental models have been studied in mechanistic epidemiology. Arguably the simplest of all is the three-compartment Susceptible-Infected-Recovered (SIR) model, an early version of which was described in 1927 [1]. The equations for the SIR model are as follows:

$$S'(t) = -\frac{\beta}{N} S(t) I(t), \tag{1}$$

$$I'(t) = \frac{\beta}{N} S(t) I(t) - \gamma I(t), \tag{2}$$

$$R'(t) = \gamma I(t), \tag{3}$$

where the five parameters are infection rate (β), recovery rate (γ), total population size (N), initial number of susceptible individuals ($S_0 := S(0)$), and initial number of infected individuals ($I_0 := I(0)$).

Adaptations to the SIR model family requiring more than five parameters have been used in modeling epidemics. For example the use of a Susceptible-Exposed-Infected-Recovered (SEIR) model during the early spike in COVID-19 cases in Northern Italy in 2020, highlighted the importance of knowing the model parameters as accurately as possible [2]. Amiri et al. [3] performed parameter estimation on an extended SIR model (SQAIR) which included a quarantined (Q) compartment and a hidden asymptomatic (A) compartment.

While these elaborate SIR variants can provide more detailed modeling, they introduce additional system states that may not be directly observable. Thus, parameter values for these more complex models often need to be estimated through informed guesses, rather than direct measurement.

On one hand, these expanded model parameters allow for more nuanced representations of real-world dynamics. However, the reliance on inferred parameter values introduces greater uncertainty and potential for error compared to the parsimonious five-parameter basic SIR model. Even the simplest SIR model meets von Neumann's upper limit of five parameters and in order to preserve model simplicity, we take it as our point of departure.

In line with this argument, Shayak et al. [4] point out that a number of equally good fits may be achieved with a range of parameter values and argue that the standard deviation, mean time, and total number of cases represent the important characteristics in deciding goodness of fit. They go on to suggest that while greater compartmentalization beyond S, I and R may seem desirable, it potentially gives rise to overparameterization.

Numerous approaches have been implemented to estimate the simple SIR model parameters in the COVID-19 epidemic, for example [5]–[14]. Rica et al. [8] achieved an order-of-magnitude improvement in goodness of fit compared to a random search algorithm by applying a differential evolution optimisation algorithm to a single differential equation reformulation of the SIR model using COVID-19 data from Chile.

Liu at al. [13] developed a discrete-time Markov chain

approach to an SIR model to estimate the evolution of the COVID-19 pandemic in Wuhan. Despite the relative paucity of data, very early predictions of the course of the epidemic were feasible in terms of guiding public health decisions.

The difficulty in achieving adequate estimates for SIR parameters is not simply a function of data paucity. It is also determined by the type of data available. The I in SIR refers to infected, which is a binary category in that a person is either defined as infected or not.

It is not normally possible to know exactly when the infected state reverts to non-infected for any individual, as this would require a clear definition of what constitutes being "infected" which is often somewhat arbitrary. Also, even if this definition were less arbitrary, such as when it is based on the result of a polymerase chain reaction (PCR) test, it would require regular testing which is often not feasible.

For example, as a COVID-19 infection subsides, a time is reached when the PCR test becomes negative. However a positive test does not necessarily indicate active infection, and it is defined at a specific sensitivity of the cycle threshold of the PCR test [15], which itself may be somewhat arbitrary or be based on the notion of being infectious rather than infected. The additional issue of false positive and false negative tests adds another layer of complexity to the definition.

The incidence rate, namely the rate at which new infections occur per unit time, is a more attainable measure, however this does not provide the overall rate of change of the infected category, $I'(t)$, which, of course, depends on both the incidence rate as well as the recovery rate, which, as we have described, is fraught with ambiguity.

The need for parameter estimation as early as possible in an infectious epidemic is evident. In this paper, we describe the development of a novel family of estimators, and we show their utility in epidemiological parameter estimation in the context of an SIR model. We will consider the situation where the net number of infections per unit time is available. For several reasons, these numbers can never be known exactly and therefore we will view the reported net infection rate $\tilde{I}'(t)$ as a noise contaminated version of the true net number of infections, $I(t)$, expressed mathematically as follows,

$$\tilde{I}'(t) = I'(t) + n(t), \tag{4}$$

where $n(t)$ represent the unknown noise that accounts for the combined errors causing these inaccuracies.

II. Novel Estimators

In this section we present three of a novel family of estimators that we have derived which have application beyond just epidemiology, for example, in radar and sonar where nonlinear differential equations are encountered. However, in epidemiology, the SIR models provide an ideal vehicle for introducing and demonstrating these estimators. For this reason, we will be introducing these estimators using the simplest SIR model. Furthermore, these estimators can only be arrived at by considering continuous-time representations before introducing time-discretization. We begin by presenting the necessary mathematical devices.

A. Preliminaries

The two key analytical tools needed for the work presented here, are the Peano-Baker series and the Cauchy repeated integral formula. The Peano-Baker series provides a method for solving *linear, time-varying* (LTV) ordinary differential equations. For the simplest case, namely for first-order differential equations of the form,

$$x'(t) = a(t)\, x(t), \quad x(0) := x_0 \tag{5}$$

the Peano-Baker series takes the form,

$$x(t) = x_0\, \mathrm{e}^{\int_0^t a(\tau)\, \mathrm{d}\tau}. \tag{6}$$

The second of these analytical devices, is the Cauchy repeated integral formula,

$$\int_a^t \int_a^{\tau_1} \cdots \int_a^{\tau_{n-1}} f(\tau_n)\, \mathrm{d}\tau_n \mathrm{d}\tau_{n-1} \ldots \mathrm{d}\tau_1 = \frac{1}{(n-1)!} \int_a^t (t-\tau)^{n-1} f(\tau)\, \mathrm{d}\tau \tag{7}$$

Finally, since *digital signal processing* (DSP) algorithms are discrete by nature, we will resort to one of the simplest and most versatile methods for converting continuous-time expressions to equivalent discrete-time versions, namely the *forward difference approximation* to the derivative,

$$x'(t) \approx \frac{x(t+\Delta t) - x(t)}{\Delta t} \tag{8}$$

which yields the discrete-time expression

$$x[k] \approx x[k-1] + \Delta t\, x'[k-1], \tag{9}$$

where $x[k] := x(k\,\Delta t)$ and $x'[k] := x'(k\,\Delta t)$ with $k = 0, 1, \ldots$ the discrete-time equivalent of the continuous-time t. Below, we will use the standard notation and write $T := \Delta t$.

B. Estimator 1

As mentioned previously, we consider the situation where $\tilde{I}'(t)$, a noisy version of the net number of infections per unit time, is available and from which we can obtain the noisy version of its time-integral,

$$\tilde{I}(t) = \tilde{I}(0) + \int_0^t \tilde{I}'(\tau)\, \mathrm{d}\tau. \tag{10}$$

Considering the state equation (1) as a LTV with $I(t)$ contributing toward the coefficient of $S(t)$ there, the Peano-Baker series formula (5) followed by application of the Cauchy repeated integral formula (7), yields the expression,

$$\hat{S}_1(t) = S_0\, \mathrm{e}^{-\frac{\beta}{N}\left(I_0 t + \int_0^t (t-\tau)\tilde{I}'(\tau)\, \mathrm{d}\tau\right)} \tag{11}$$

Now we are in a position to convert this expression to discrete-time, using the forward difference method, to obtain

$$\hat{S}_1[m] = S_0\, \mathrm{e}^{-\frac{\beta T}{N}\left(I_0 m + T\sum_{n=0}^{m-1}(m-n)\tilde{I}'[n]\right)}, \tag{12}$$

which yields an estimator for $S_1[m]$ exclusively in terms of some parameters and of the past and present values of $\tilde{I}'$. In terms of its parameterization, (12) is linear in the primary parameter S_0 but nonlinear in the primary parameters β, N, I_0 and in the secondary parameter T that appears as a result of time-discretization. Observe that the estimator depends nonlinearly on the past and present values of $\tilde{I}'$.

Perhaps the most important observation is that there does not seem to exist any direct discrete-time equivalent method that would produce this estimator. Therefore, it was essential to work with the continuous-time version until a final expression for the estimator was obtained, before converting to discrete-time. Furthermore, even though we assumed uniform sampling here, the expression for the more general case of nonuniform sampling where $T_m := \Delta t_m$ follows easily. Another important aspect which we will discuss here, is that this estimator can be expressed in a computationally efficient recursive form where $\hat{S}_1[m]$ is expressed in terms of $\hat{S}_1[m-1]$.

C. Estimator 2

Once more, we assume we have only $\tilde{I}'(t)$, a noisy net number of infections per unit time and its time integral, and hence we can write (2) as

$$\tilde{I}'(t) \approx \frac{\beta}{N} S(t)\tilde{I}(t) - \gamma \tilde{I}(t), \tag{13}$$

from which $S(t)$ can be made the subject to obtain yet another estimator for $S(t)$,

$$S(t) \approx \frac{N}{\beta}\left(\frac{\tilde{I}'(t)}{\tilde{I}(t)} + \gamma\right) =: \hat{S}_2(t), \tag{14}$$

or, explicitly in terms of $\tilde{I}'$, namely

$$\hat{S}_2(t) = \frac{N}{\beta}\left(\frac{\tilde{I}'(t)}{I_0 + \int_0^t \tilde{I}'(\tau)\,\mathrm{d}\tau} + \gamma\right). \tag{15}$$

Using (9), we obtain the discrete-time estimator,

$$\hat{S}_2[m] = \frac{N}{\beta}\left(\frac{\tilde{I}'[m]}{I_0 + T\sum_{n=0}^{m-1} \tilde{I}'[n]} + \gamma\right). \tag{16}$$

This estimator depends nonlinearly on past and present values of $\tilde{I}'$ and, except for being independent of S_0, it depends nonlinearly on the same parameters as $\hat{S}_1[m]$ does, and also on γ.

D. Estimator 3

Under the same assumptions as for the previous two estimators, substituting (1) into (2) gives,

$$S'(t) \approx -\tilde{I}'(t) - \gamma \tilde{I}(t). \tag{17}$$

Integrating with respect to time, we get a third estimator for $S(t)$, namely.

$$\hat{S}_3(t) := S_0 + I_0 - \int_0^t \tilde{I}'(\tau)d\tau - \gamma\int_0^t (t-\tau)\tilde{I}'(\tau)\,\mathrm{d}\tau \tag{18}$$

from which we obtain the following discrete-time version of the estimator,

$$\hat{S}_3[m] = S_0 + I_0 - T\sum_{n=0}^{m-1} \tilde{I}'[n] - \gamma T^2 \sum_{n=0}^{m-1} (m-n)\tilde{I}'[n]. \tag{19}$$

As opposed to the previous two estimators, this estimator is linear in the present and past values of $\tilde{I}'$ and for fixed T, it depends linearly on S_0, I_0 and γ. As with estimator $\hat{S}_1$, this estimator can also be expressed recursively.

III. Estimators by Association

Since $S_0 = S(0)$, the initial number of susceptible individuals to an infectious disease in a given population, is extremely important for understanding the disease and for anticipating the impact of a new outbreak, we proceed to consider associations of the three estimators above. In this work, we studied all three possible associations of the above estimators but because of limited space, we only present the association of estimators 1 and 3. Technically the new estimator obtained through this association should be referred to as $\hat{S}_{1,3}(0|t)$ to emphasize that it produces an estimate based on the history of I' from time 0 up to and including t. However, to simplify notation we will simply write this as $\hat{S}_{1,3}(0)$

For the association of estimators $\hat{S}_1(t)$ and $\hat{S}_3(t)$, we assume that both $\hat{S}_1(t) = \hat{S}_3(t)$ and $\hat{S}_1(0) = \hat{S}_3(0)$ hold approximately, and define $\hat{S}_{1,3}(0) = (\hat{S}_1(0) + \hat{S}_3(0))/2$. The continuous-time estimator obtained from this is given by,

$$\hat{S}_{1,3}(0) = \frac{\tilde{I}(t) - \tilde{I}_0 + \gamma(\tilde{I}_0 t + \int_0^t (t-\tau)\tilde{I}'(\tau)\,\mathrm{d}\tau)}{1 - \exp\{-\frac{\beta}{N}(\tilde{I}_0 t + \int_0^t (t-\tau)\tilde{I}'(\tau)\,\mathrm{d}\tau)\}}. \tag{20}$$

Finally, using (9), the discrete-time version follows, namely

$$\hat{S}_{1,3}[0] = \frac{\tilde{I}[m] - \tilde{I}_0 + \gamma T(\tilde{I}_0 m + T\sum_{n=0}^{m-1}(m-n)\tilde{I}'[n])}{1 - \exp\{-\frac{\beta T}{N}(\tilde{I}_0 m + T\sum_{n=0}^{m-1}(m-n)\tilde{I}'[n])\}}, \tag{21}$$

where $\tilde{I}[m] = T\sum_{n=0}^{m-1} \tilde{I}'[n]$. This estimator is nonlinear in the history of I' as well as in its parameters, namely I_0, N, β and T.

We will now study these estimators numerically from an estimation theoretic perspective [16].

IV. Numerical Experiment

This section presents a numerical experiment conducted to characterize the novel estimators derived in section II. The experiment is described in section IV-A. The parameter values selected for the experiment are listed in section IV-B, and the results are presented in section IV-C.

A. Description of experiment

The numerical experiment uses the SIR model introduced in section I, with parameter values that are representative of a COVID-19 outbreak. In the experiment, the differential equations that define the SIR model are solved numerically for the susceptibility (S), infected (I) and recovered (R) counts as functions of time t (where $t \geq 0$ denotes the number of elapsed days since the start of the outbreak). These functions,

TABLE I
PARAMETER VALUES OF THE NUMERICAL EXPERIMENT.

Parameter	Symbol	Value
Infection rate	β	1/15 per day
Recovery rate	γ	4/10 per day
Total population size	N	7501 individuals
Initial number of susceptible individuals	S_0	7491 individuals
Initial number of infected individuals	I_0	10 individuals
Noise gain factor (low intensity)	κ	0.005
Noise gain factor (medium intensity)	κ	0.010
Noise gain factor (high intensity)	κ	0.015
Noise IIR filter pole	ρ	0.75
Sample period	T	1 day

referred to as the true counts, serve as the baseline (i.e., the ideal values) for comparison with the estimated values.

The experiment involves the computation of the true infected rate I' by substituting the true susceptibility count S and true infected count I into (2). An ensemble of B discrete-time noise-contaminated versions $\tilde{I}'_b[m] := I'(mT) + n_b[m]$, $b = 1, 2, \ldots, B$, of the true infected rate I' are computed. Each noise function n_b is a sample function of a zero-mean Gaussian random process filtered using an all-pole low-pass infinite impulse response (IIR) filter with a single real-valued pole $\rho > 0$. To simulate realistic noise-contaminated curves, where the intensity of the noise fluctuation is proportional to the count of infected individuals, the variance $\sigma^2[m]$ of the random variate $n_b[m]$ is selected to be proportional to $I(mT)$ according to the expression

$$\sigma^2[m] = \kappa I(mT), \tag{22}$$

where κ is called the *noise gain factor*. The power of the above noise process is proportional to the noise gain factor selected.

In the numerical experiment, the discrete-time estimators $\hat{S}_1$ and $\hat{S}_3$, and the discrete-time estimator $\hat{S}_{1,3}$ of S_0, are computed by substituting each noise-contaminated infected rate function $\tilde{I}'_b$ in the ensemble into (12), (19) and (21), respectively. The experiment characterizes each estimator according to its mean and variance, which are computed by averaging over the ensemble.

B. Experiment parameters

The parameters for the experiment are selected as outlined in Table I. The SIR model parameters selected were reported for the first COVID-19 outbreak in Wuhan, China [14]. In the experiment, all population counts are normalized such that a count of unity represents 10^3 individuals.

C. Results

Time plots of the susceptible (S), infected (I) and recovered (R) counts of the numerical example are plotted in Figure 1. The figure shows that count of infected individuals reaches its peak value on day 25, after which the infected count decays to zero. After 92 days, the number of infected individuals, as a fraction of the total population, falls below 1%.

The true infected rate I' and true infected count I of the numerical example are plotted as functions of time in Figure 2.

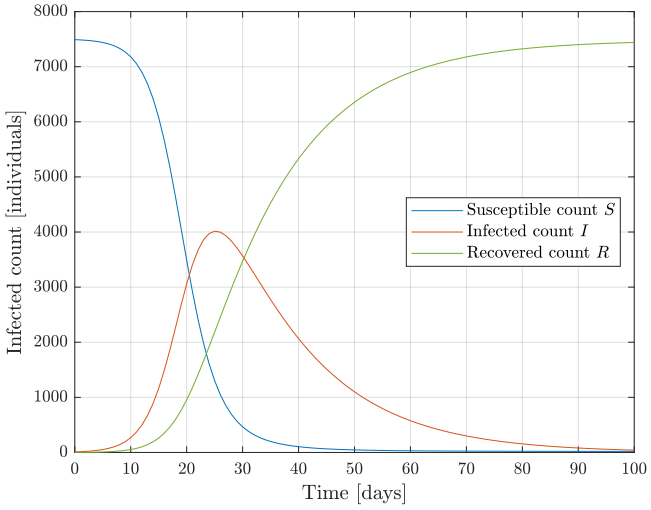


Fig. 1. Time plots of the susceptible (S), infected (I) and recovered (R) counts of the experiment before noise contamination.

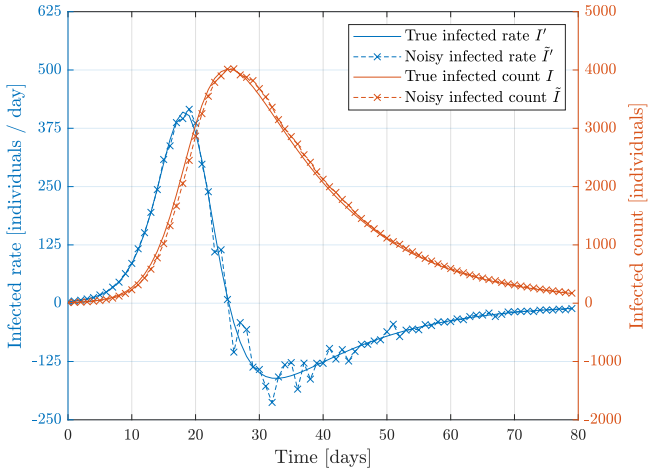


Fig. 2. Time plots of the true infected rate (I') and true infected (I) counts of the experiment, and noise-contaminated examples $\tilde{I}'$ and $\tilde{I}$.

The figure includes an example of a discrete-time noise-contaminated infected rate curve $\tilde{I}'$. This curve was generated using the process outlined in section IV-A, with a medium intensity noise gain factor $\kappa = 0.01$. The figure shows that the noise fluctuation is largest over the time period where the infection count reaches its peak value.

The noise-contaminated infected count curve $\tilde{I}$, obtained via discrete-time numerical integration of $\tilde{I}'$, is plotted in Figure 2. This curve exhibits a smaller noise fluctuation due to the low-pass filtering effect of integration. Furthermore, the noise-contaminated infected count curve exhibits a time delay relative to the ideal infected count curve. This time delay affects the accuracy of the estimators, as demonstrated below.

Figure 3 presents time plots of the bias and variance of the discrete-time estimator $\hat{S}_1$, as computed numerically for the low, medium and high intensity noise gain factors κ. The

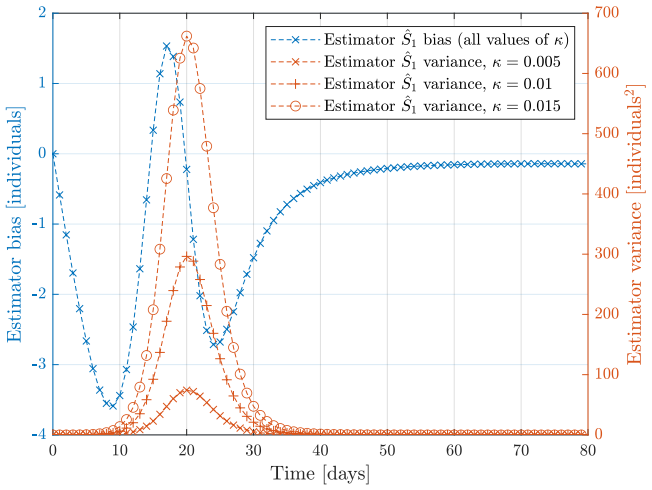


Fig. 3. Bias and variance of estimator $\hat{S}_1$, plotted as a function of time and for distinct noise gain factors κ.

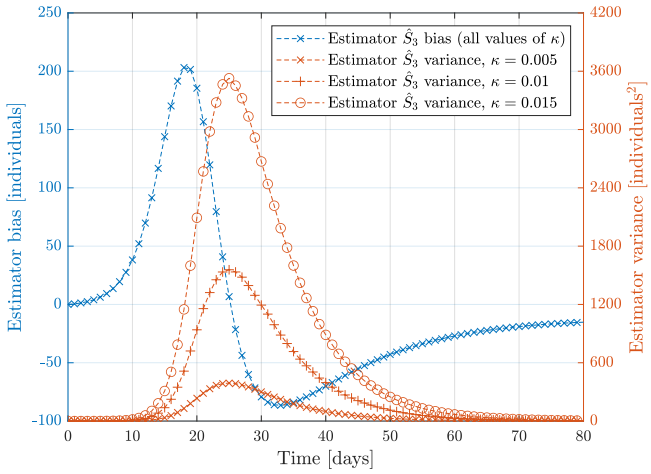


Fig. 4. Bias and variance of estimator $\hat{S}_3$, plotted as a function of time and for distinct noise gain factors κ.

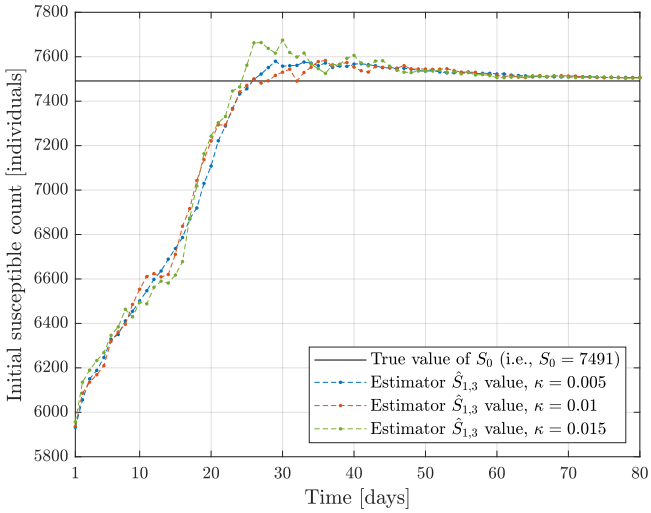


Fig. 5. Examples of time plots of the estimator $\hat{S}_{1,3} \approx S_0$, for distinct noise gain factors κ.

figure reveals that the bias of the estimator is small over the entire time period considered. Here, the peak bias value is equal to only 0.03% of the susceptible count (on day 17). This demonstrates the high accuracy of estimator $\hat{S}_1$.

The bias curve of estimator $\hat{S}_1$ in Figure 3 exhibits a large fluctuation over the time period that coincides with the peak of the infection rate curve $\tilde{I}'$ (days 10 to 30), but appears to converge slowly towards zero after day 30. It was found that this slow convergence is due to the time delay caused by discrete-time numerical integration of the noise-contaminated infected rate $\tilde{I}'$. Furthermore, the figure shows that the variance of estimator $\hat{S}_1$ reaches its peak on day 20. Larger estimator variance is observed as the noise gain factor κ is increased.

Figure 4 presents plots of the bias and variance of the discrete-time estimator $\hat{S}_3$, computed numerically for the low, medium and high intensity noise gain factors κ. A comparison of Figures 3 and 4 shows that estimator 3 exhibits a significantly larger bias and variance than estimator 1. However, the bias is not excessive; the peak bias value of estimator 3 is equal to 4.4% of the susceptible count (on day 18).

The bias curve of estimator $\hat{S}_3$ exhibits a large fluctuation over the same time period as estimator $\hat{S}_1$ exhibits maximum fluctuation. The figure shows that the bias of estimator $\hat{S}_3$ also exhibits slow convergence towards zero over time, due to the time delay caused by discrete-time numerical integration.

Figure 5 presents time plots of the discrete-time estimator $\hat{S}_{1,3}$ (here, the value of $\hat{S}_{1,3}$ on day m is an estimate of the initial susceptible count S_0 on day zero, as obtained by accumulating observations up to day m). Three curves, each corresponding to low, medium and high intensity noise gain factors, are plotted. The numerical results show that the estimator $\hat{S}_{1,3}$ converges to the true value of S_0 (plotted as the solid black line in the figure) over time, for each noise gain factor. Furthermore, an increase in the noise gain factor yields larger fluctuation in the estimator value over time.

Figure 6 presents time plots of the mean and variance of estimator $\hat{S}_{1,3}$, as computed numerically for the low, medium and high intensity noise gain factors κ. The figure shows that the bias of the estimator converges towards zero after day 33, and that a higher noise gain factor increases the variance of estimator $\hat{S}_{1,3}$. The variance of the estimator reaches its peak value on day 24. Thereafter, the estimator variance converges towards zero. The time convergence of the estimator bias and variance towards zero proves that the estimation accuracy improves as more observations are accumulated.

V. Conclusions

The numerical results for estimator bias and variance show that the novel estimators $\hat{S}_1$ and $\hat{S}_3$ yield accurate figures for the number of susceptible individuals over time, despite noise contamination of the observations. The novel estimator $\hat{S}_{1,3}$ also yields accurate figures for the initial number of susceptible individuals S_0. In particular, the results show that the accuracy

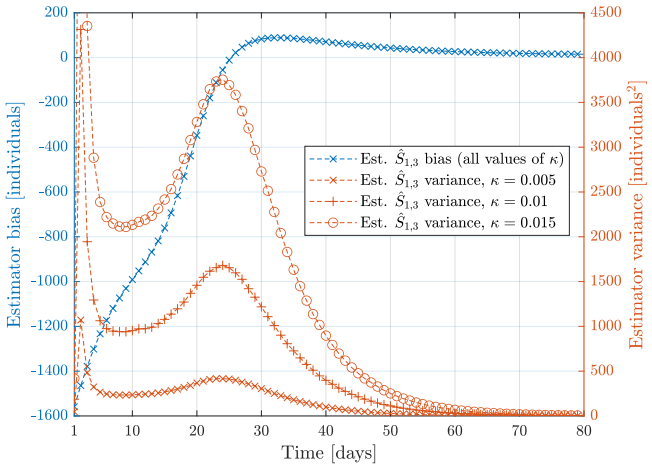


Fig. 6. Bias and variance of estimator $\hat{S}_{1,3} \approx S_0$, plotted as a function of time and for distinct noise gain factors κ.

of the latter estimator improves over time as more observations are accumulated.

While the time delay caused by discrete-time integration required to compute the estimates was identified as a source of estimator bias, further analysis showed that the impact of this error may be reduced by shortening the time between subsequent observations of the infected rate.

It is anticipated that the application of these and other estimators in the accurate determination of parameters at an early stage in infectious epidemics, will enhance the capacity to plan and allocate resources to mitigate the ravages on populations.

Planned future work on these estimators include detailed analysis and evaluation on real-world data.

Acknowledgment

We acknowledge the support of the Carl and Emily Fuchs Foundation.

References

[1] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, vol. 115, no. 772, pp. 700–721, 1927.

[2] J. M. Carcione, J. E. Santos, C. Bagaini, and J. Ba, "A simulation of a COVID-19 epidemic based on a deterministic SEIR model," *Frontiers in Public Health*, vol. 8, p. 230, 2020.

[3] A. H. Amiri Mehra, M. Shafieirad, Z. Abbasi, and I. Zamani, "Parameter estimation and prediction of COVID-19 epidemic turning point and ending time of a case study on SIR/SQAIR epidemic models," *Computational and Mathematical Methods in Medicine*, vol. 2020, no. 1, p. 1 465 923, 2020.

[4] B. Shayak, S. Jahedi, and J. A. Yorke, *Ambiguity in the use of sir models to fit epidemic incidence data*, 2024. arXiv: 2404.04181 [q-bio.PE]. [Online]. Available: https://arxiv.org/abs/2404.04181.

[5] M. Lounis and D. K. Bagal, "Estimation of SIR model's parameters of COVID-19 in Algeria," *Bulletin of the National Research Centre*, vol. 44, no. 1, p. 180, 2020.

[6] A. Alsayed, H. Sadir, R. Kamil, and H. Sari, "Prediction of epidemic peak and infected cases for COVID-19 disease in Malaysia, 2020," *International journal of environmental research and public health*, vol. 17, no. 11, p. 4076, 2020.

[7] D. Prodanov, "Analytical parameter estimation of the SIR epidemic model. applications to the COVID-19 pandemic," *Entropy*, vol. 23, no. 1, p. 59, 2020.

[8] S. Rica and G. A. Ruz, "Estimating SIR model parameters from data using differential evolution: An application with COVID-19 data," in *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, IEEE, 2020, pp. 1–6.

[9] N. A. Kudryashov, M. A. Chmykhov, and M. Vigdorowitsch, "Analytical features of the SIR model and their applications to COVID-19," *Applied Mathematical Modelling*, vol. 90, pp. 466–473, 2021, ISSN: 0307-904X. DOI: https://doi.org/10.1016/j.apm.2020.08.057. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0307904X20304984.

[10] C. S. Carlson, D. M. Rubin, V. Heikkilä, and M. Postema, "Extracting transmission and recovery parameters for an adaptive global system dynamics model of the COVID-19 pandemic," in *2021 IEEE AFRICON*, IEEE, 2021, pp. 455–458.

[11] D. Prodanov, "Analytical solutions and parameter estimation of the SIR epidemic model," *Mathematical Analysis of Infectious Diseases*, pp. 163–189, 2022.

[12] F. G. Schmitt, "An algorithm for the direct estimation of the parameters of the SIR epidemic model from the i(t) dynamics," *The European Physical Journal Plus*, vol. 137, no. 1, p. 57, 2022.

[13] T. Liu, J. Huang, Z. He, *et al.*, "A real-world data validation of the value of early-stage SIR modelling to public health," *Scientific Reports*, vol. 13, no. 1, p. 9164, 2023.

[14] R. Saxena, M. Jadeja, and V. Bhateja, "Propagation analysis of COVID-19: An SIR model-based investigation of the pandemic," *Arabian Journal for Science and Engineering*, vol. 48, no. 8, pp. 11 103–11 115, 2023.

[15] T. P. Velavan and C. G. Meyer, "COVID-19: A PCR-defined pandemic," *International Journal of Infectious Diseases*, vol. 103, pp. 278–279, 2021.

[16] H. L. van Trees, *Detection, estimation and modulation theory*, 2nd ed., New York, NY: John Wiley & Sons, 2001.

Peer Learning via Shared Speech Representation Prediction for Target Speech Separation

Xusheng Yang, Zifeng Zhao and Yuexian Zou*
School of Electronic and Computer Engineering, Peking University, Shenzhen, 518055, China
E-mail: yangxs@stu.pku.edu.cn, zifeng_z@foxmail.com, zouyx@pku.edu.cn

***Abstract*—Target speech separation approaches widely use the encoder-separation-decoder framework. Generally, the crucial idea of the framework is how to map the mixture of speech signals to a distinctive representation. However, existing methods fail to learn discriminative speech representations because they ignore the distinctive characteristics (tone and timbre) related to each target speaker, which are shared within the corresponding mixture samples. To acquire shared speaker characteristics and discriminative representations, we propose a novel "peer learning" method for this framework. Specifically, we construct a pair of mixture samples involving the same target speaker and design a two-branch speech representation prediction module. Then, the more discriminative representation can be determined via computing the signal-to-noise ratio and exchanged in the prediction module to improve separation performance. Experiments on the Libri2Mix dataset demonstrate the effectiveness of our method.**

I. Introduction

Target speech separation (TSS) is to separate the voice of a speaker of interest from an overlapped mixture of speech signals. It can be applied to hearing aids, mobile communication, and speech recognition systems [1]. Humans can naturally extract relevant information from the target speaker in noisy environments. Researchers have identified the importance of essential cues to a target speaker [2], [3], so the popular approaches usually transform the mixture into a separable space leveraging acoustic, visual and spatial information [4]–[7]. Recently, TSS methods of encoder-separation-decoder have gained more attention. Their separation performance heavily relies on mapping the mixture of speech signals to a distinctive representation, which is a still challenging research problem.

For a specific target speaker, the information in the different mixture signals (tone and timbre) helps learn the distinctive representations. By analyzing these characteristics, the model can learn to recognize the specific frequency patterns and harmonic structures that are unique to the target speaker. Recent works [8]–[10] on target speech separation retain the temporal architecture [11], [12] of the encoder-separation-decoder (speech extraction network) and incorporate a convolutional network (auxiliary network) for extracting the information of target speaker. The speech extraction network first encodes the mixture speech of signals by a 1-D convolutional layer, and subsequently separates and decodes them by several convolutional blocks and a decoder layer in representation space. Finally, it outputs a single signal corresponding to the target speech, effectively isolating a clean signal from the mixture.

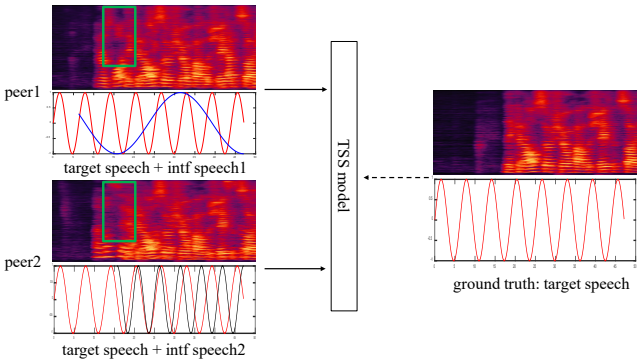


Fig. 1. The intuition of peer learning samples. The red sinusoidal signal represents the target speech, while the blue signal and black signal are interference speech at different frequencies. Peer1 represents a mixture of red sinusoidal signal and blue signal. Peer2 represents a mixture of red sinusoidal signal and black signal. The green-boxed regions indicate harmonic structures and frequency patterns where two samples can mutually enhance each other.

However, when speakers have similar voice characteristics, existing methods become challenging to identify a specific target speaker from a mixture of signals. These methods use each mixture sample of the target speaker in the training independently, so they cannot share enough useful characteristics on speech qualities (tone, timbre, and pitch) resulting in non-discriminative representations. Moreover, the conventional encoder-separation-decoder framework, while adept at distinguishing the target speech from a mixture of signals by estimating masks, is inherently constrained in capturing shared speech representations across diverse mixture samples. This limitation stems from the architecture's primary focus on differentiating the target signal rather than extracting commonalities that may exist within the mixture.

In this work, we introduce a novel learning paradigm termed "peer learning" to extract shared attributes and learn discriminative representations from diverse speech mixtures. Our approach is underpinned by the creation of a novel mixing dataset that pairs mixture samples centered around the target speaker. Our core concept is to explore the potential for two mixture samples to engage in a reciprocal learning process, akin to peers sharing insights, hence the term "peer learning" samples. As depicted in Figure 1, we present two mixture

*Corresponding author.

samples, peer1 and peer2, each a blend of the target speaker's speech with that of a different, non-target speaker or interfering voice (for instance, peer1 is a mix of Speaker 1 and Speaker 2, while peer2 combines Speaker 1 with Speaker 3). Notably, we propose that the harmonic structures and frequency patterns within the green-boxed regions of these mixture samples can be interchangeably leveraged to bolster the accuracy of target speech prediction. To optimize the utilization of "peer learning" samples, we have engineered a dual-branch architecture for speech representation prediction designed to impart distinctive characteristics. In this bifurcated approach, we derive predicted speech via both the representation prediction module and the decoder, subsequently quantifying the signal-to-noise ratio (SNR) based on the predicted speech and the actual target speech. The representation associated with a superior SNR is selected, as it is indicative of a higher-quality speech representation. This enhanced representation is then fed back into the other branch to refine the original, less discriminative representation. Through this mechanism, we facilitate the exchange of discriminative representations between branches, thereby enhancing the overall speech separation performance. Experimental results show that our peer learning framework outperforms the performance of baselines for the target speech separation on the Libri2Mix dataset [13].

II. Related Work

Target speech separation. TD-SpeakerBeam [8] acquires more discriminative speaker embedding vectors using an auxiliary convolutional network and a multi-task loss with speaker identification loss. It follows the Conv-TasNet to construct multiple convolutional filters to transform the temporal signals within a time slot into a learnable representation. SpEx+ [9] shares the same weight in the latent domain by twin speech encoder, importantly it encodes multi-scale information in a uniform latent feature space. Besides, the refining neural network [14] demonstrates that high-order embedding space can leverage the discriminative representation for speech separation. However, in a mixture speech, the same target speaker can be mixed with any speaker. These works do not consider shared knowledge such as different frequency patterns and harmonic structures by constructing richer training pairs.

The paradigms of learning. Self-supervised learning [15]–[20] constructs a pair of positive and negative samples by selecting one speaker's utterance as the anchor, and selecting speech from the same person but a different utterance as the interfering speech. To capture speech characteristics, they introduce a speaker consistency loss on the speech embeddings. The teacher-student network [21], [22] can unidirectionally transfer knowledge from the teacher model to the student model. Mutual learning [23], [24] approaches can use a group of students to learn and share knowledge simultaneously with each other during the training phase, but they require the selection of two models when the testing phase and the parameter of the threshold must be manually set. However, the mainstream frameworks focus on masking interference speech while ignoring learning the shared presentations to improve separation performance.

III. Methods

In this section, we first give a notation of target speech separation. We adopt the mainstream framework of encoder-separator-decoder as the backbone and introduce the proposed peer learning (PL) method to obtain shared attributes and learn discriminative representations. Fig.2 shows the overview of our PL method. Then we explain the speech representation prediction mechanism in detail.

A. Problem Formulation

Notations. Target speech separation is to isolate the speech of a target speaker from a mixture of multiple overlapping speakers and optionally an additional noise. First, X denotes the mixed speech signals, and S is reference clean sources. The adaptation utterance about the target speaker will be denoted r. We aim to separate the predicting source $\hat{S}$ for each speaker from X by leveraging the target speaker information r. To make it simple, a two-speaker setup is considered.

B. Peer Learning Method

Some advanced learning methods have provided performance improvements for target speech separation. However, these methods have certain limitations. These works ignore more discriminative representation can be beneficial to separation performance further. The works mentioned above do not leverage shared harmonic structures and frequency patterns in a pair of mixture samples of the same target speaker. Our intuition, as previously highlighted, posits that the introduction of a pair of mixture samples pertinent to the target speaker can unlock superior representations. This strategy is predicated on the notion that by harnessing the intrinsic relationships between samples, we can distill representations that are more conducive to effective speech separation.

Constructing "peer learning" samples. We construct pairs of training data that contain a mixture sample of the same target speaker and different interference speakers. Generally, there are n speakers. Suppose we have one speech segments of target speaker $s_i(t)$ for speaker i, two speech segments of interference speaker $s_j(t)$ and $s_k(t)$ for speaker j and k. Note that j and k may be the same speaker. We create two mixture signals $x_1(t)$ for speaker i and speaker j, $x_2(t)$ for speaker i and speaker k, which is defined in:

$$\begin{aligned} x_1(t) &= s_i(t) + s_j(t), (i \neq j, i, j \in n) \\ x_2(t) &= s_i(t) + s_k(t), (i \neq k, i, k \in n) \end{aligned} \tag{1}$$

Shared two-branch Framework. We adopt the general framework of neural target speech separation, mainly including a mixture encoder, a fusion layer, and a target extractor. Specifically, the peer mixture speech signals are then transformed into the embedding space. Then we use the fusion layer to introduce the information of the target speaker. Finally, we use the extractor module to estimate target speech. The entire pipeline of peer learning with shared speech representation

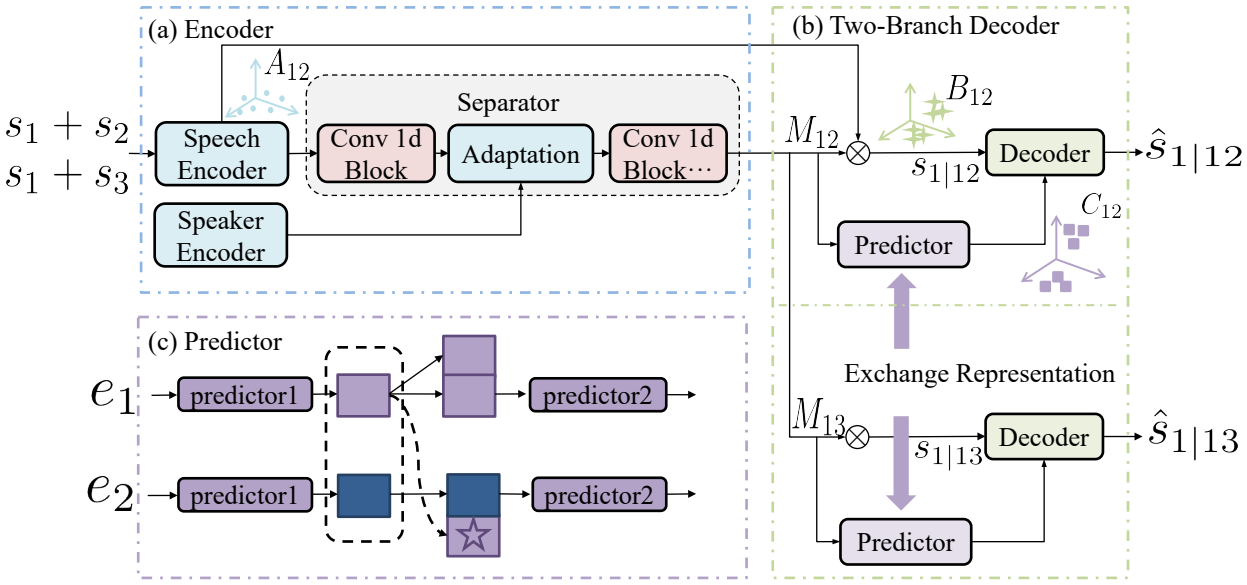


Fig. 2. The pipeline of peer learning via shared speech repressentation prediction. The purple bidirectional arrow illustrates the sharing discriminative representation in this space. (a) The encoder of our method with shared weights for "peer learning" samples (different mixture samples of the same target speaker). (b) The two-branch decoder for "peer learning" samples. (c) The detail of the predictor (the purple block).

prediction is as shown in Fig.2, which is designed based on the famous framework of TD-speakerBeam [8]. The proposed two-branch network employs an encoder with shared weights, indicating that identical network structures are deployed to handle different input datasets. This approach ensures that the same underlying model processes various inputs, promoting efficiency across the network's operations. The speech encoder denotes extracting the mixture representations by 1-D convolutional filters, speaker encoder uses the same 1-D convolutional encoder to accept the enrollment utterance of the target speaker. Then the separator transforms further the representation into the separable representation space incorporating the adaptation layer and the stacked 1-D convolutional blocks. They perform the following equations:

$$\begin{aligned} X_{emb} &= \text{Enc}(x; \theta^{enc}) \\ Spk_{emb} &= \text{SpkEnc}(r; \theta^{Spkenc}) \\ E_{emb} &= \text{Sep}(X_{emb}, Spk_{emb}; \theta^{Sep}) \end{aligned} \tag{2}$$

where Enc and SpkEnc respectively represent the convolutional network and each network with parameters θ^{Spkenc} and θ^{Spkenc}. For $X_{emb} \in R^{D \times L}$, D denotes dimension, L is the encoded temporal length. And $E_{emb} \in R^{D^s \times L}$ is the output of separator Sep in gray modules in Fig.2. Then, we define the mask branch of the representations in green space B, and the purple space C denotes the predictor branch of the representations. We discuss the details of the prediction branch in the following paragraphs.

Shared Speech Representation Prediction. Motivated by self-supervised learning methods BYOL [25], this work employs two different networks and adopts a stop-gradient strategy to introduce asymmetry for representation learning. Therefore, we also attempt to introduce asymmetry in the two-branch network to obtain discriminative representations. From this perspective, the reason why directly adding a speech representation prediction module is that the better representation can guide the worse one, which plays a role in improving the worse representation. We directly add a speech representation prediction module to guide the original representation with the better one. For a clearer presentation, suppose we have one speech segment of target speaker $s_1(t)$ for speaker 1, two speech segments of interference speaker $s_2(t)$ and $s_3(t)$ for speaker 2 and 3. We create two peer mixture signals: $x_1(t) = s_1(t) + s_2(t)$ and $x_2(t) = s_1(t) + s_3(t)$. We have obtained the E_{emb} from the shared two-branch network, at this point, so we denote them e_1 and e_2 for two peer learning mixtures. After the separator module, we generate a speech representation using a prediction network with the convolutional layer. Note that this representation is near to the ground truth of the target speech before the decoder, so the discriminative speech representation can be shared.

More discriminative representations can be understood as features that, after being mapped by the predictor, more closely align with the actual target speech. This enhanced alignment signifies a higher fidelity in the representation, enabling more accurate speech processing. Fig.2 (b) illustrates a predictor that introduces a novel mapping of the representation prior to the decoder. Since the representation at this stage is already quite close to the actual speech, it can be leveraged as new knowledge to share. In other words, a more discriminative speech representation is effectively evaluated through this comparison, highlighting the enhanced fidelity of the predictor's output, as shown in Fig.2 (c). The structure of the predictor is the linear neural network. The predicted output is then fed into

the decoder as input and compared with the true labels to calculate the SNR, which serves as a metric for assessing the quality of the speech representation.

We elaborate in detail on the speech representation prediction module in Fig.3. We first build a prediction branch to generate the representations $e_{11}, e_{22} = \mathrm{Pred}_1(\mathrm{e}_1), \mathrm{Pred}_1(\mathrm{e}_2)$. Further, we decide on the more discriminative representation by calculating SI-SNR with the target speech. Specifically, the representation through predictor 1 and input the decoder to acquire the predicted speech. We compare the SNR values of the two representations by predicted speech and the target speech. In the Fig.3, when SI-SNR1 is bigger than SI-SNR2, the better representation e_{11} with star marker is used to formulate for the other branch. Moreover, we concatenate the star marker representation with the original blue representation, after that, we can acquire the $e_1^{'}$, $e_2^{'}$ using a predictor 2.

$$e_1^{'}, e_2^{'} = \mathrm{Pred}_2(\mathrm{e}_{11}, \mathrm{e}_{11}), \mathrm{Pred}_2(\mathrm{e}_{22}, \mathrm{e}_{11}) \tag{3}$$

Then, we integrate them into the mask branch for recovering target speech combining with e_1 and e_2. In such a way, it can be viewed as a speech representation augmentation.

Training and Inference. During the training, we easily construct peer learning samples of the same target speaker. However, in the validation and inference, it is usually not feasible to create these samples. Therefore, during the phase of inference, we directly concatenate the mixture's representation itself. It means that we directly double the representation after predictor 1 to match the dimension.

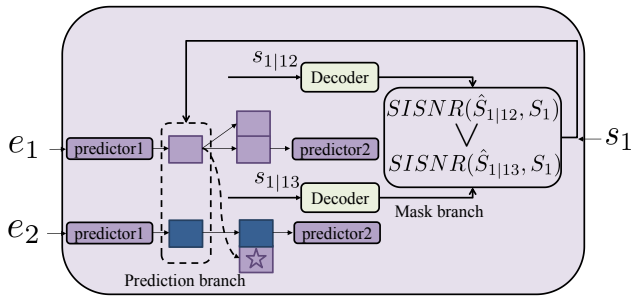


Fig. 3. The mechanism of the speech representations prediction by comparing the SNR.

IV. Experiments

A. Experimental Setup

Datasets. The Libri2Mix[13] dataset extends the utility of the LibriSpeech corpus into the domain of speech separation. Libri2Mix dataset [13] is introduced to address the generalization issue in WSJ0-2mix dataset[26]. The train-100 subset total has 13900 utterances with 58 hours of data from 251 speakers. The development and test subsets have 3,000 mixture utterances with 11 hours of data respectively, which is from 40 unseen speakers. The sampling rate of speech audio is $8\,\mathrm{kHz}$. We generate all mixtures using 'minimum' mode. We only construct peer samples in training data. During the mixing of data, the specific details have been thoroughly explained in the section III-A. The first speaker is chosen as the target speaker, the second speaker is regarded as the interference speaker. The reference speech of the target speaker must be a randomly chosen utterance distinct from the one in the mixture. The other peer sample has the same first speaker as mixtures and chooses a different utterance from the interference speaker. The development and test subsets remain consistent with the original dataset setting.

Baselines and Setup. We compare it with baseline approaches of target speech separation, including sDPCNN [27], TD-SpeakerBeam [8], SpEx+ [9]. Some of the results are obtained from the reported papers. Our all hyper-parameters selections follow the original TD-SpeakerBeam. The training process uses chunks with a 3.0 seconds duration. We adopt Adam optimizer with a learning rate of 1e-3. The batch size is 8. The epoch is set to 200. Early stopping is used in the training process when the validation loss doesn't decrease in 20 consecutive epochs. A gradient clip with a maximum L2-norm is used to avoid gradient exposure.

Evaluation Metrics. We train all time-domain models using negative SISNR loss. For evaluation metrics, we employ the scale-invariant SDR (SI-SDR) and PESQ metrics. SI-SDR assesses from a signal perspective, while PESQ evaluates them from perceptual quality. The higher evaluation value represents the better quality of speech separation.

B. Results Analysis

1) Performance Comparison: We explore the performance of different target speech separation methods on the Libri2Mix test dataset, the overall results are shown in Table.I. We denote the peer learning method as PL. In the last two rows of Table.I, our approach outperforms the sDPCNN method in the time-frequency domain, as well as the TD-SpeakerBeam method in the time domain. Our approach (PL) improves the performance to 1.0 dB and 0.5 dB SI-SDRi respectively compared with TD-SpeakerBeam and SpEx+. In the meantime, we obtain 0.13 and 0.32 PESQ improvement compared with TD-SpeakerBeam and SpEx+ respectively. Furthermore, it can be concluded that our approach can be integrated into existing time-domain frameworks easily. These findings suggest that PL can augment performance by learning more discriminative representation.

TABLE I
PERFORMANCE COMPARING ON THE LIBRI2MIX DATASETS.

Methods	SI-SDR	PESQ
Mixture	0.001	1.603
sDPCCN	11.65	2.74
TD-SpeakerBeam	12.86	2.75
SpEx+	13.41	2.94
TD-SpeakerBeam + PL	**13.94**	**3.07**

2) Speech Representation Prediction Analysis: We further investigate the usage strategy of the speech representation prediction. Specifically, there are three strategies. First, **cat itself**: both branches double their representation without knowledge

sharing. Secondly, **cat anyway**: both branches use a random representation from the other. Finally, **cat better**: we utilize a better representation by speech representation prediction module. We train the three strategies using SISNR as the same loss function. It can be observed that the performance progressively is boosted as shown in Table.II, indicating the effectiveness of our strategy via the speech representation prediction module.

TABLE II
RESULTS ON THE DIFFERENT CONFIGURATIONS.

Model	Operations	SI-SDR	PESQ
Mixture	-	0.001	1.603
TD-SpeakerBeam	-	12.86	2.75
TD-SpeakerBeam + PL	cat itself	12.84	2.74
TD-SpeakerBeam + PL	cat anyway	12.89	2.77
TD-SpeakerBeam + PL	**cat better**	**13.94**	**3.07**

3) Analysis of training process: To further validate how to share speech representation, we analyze the training process of the different representations. As shown in Fig.4, after adopting the asymmetric training method of peer learning, we have 4 outputs: the orange line represents the better peer, and the gray one represents the better mask. Besides, the yellow line denotes the worse peer while the blue one is a worse mask. Firstly, for the results of the mask branch, the better peer outperforms the worse peer by around 1.2 to 1.8 dB. For the map branch, we find that the gap between the representations of the better peer and the worse peer gradually decreases during the training process, from around 1.3 dB to 0.7 dB. The green line represents a remapped result obtained by using the better representation. We can see that its performance is not only better than the mapping performance of the worse peer but also surpasses that of the better peer itself. Upon closer examination, we observed that the representations of the two peers indeed have a synergistic effect.

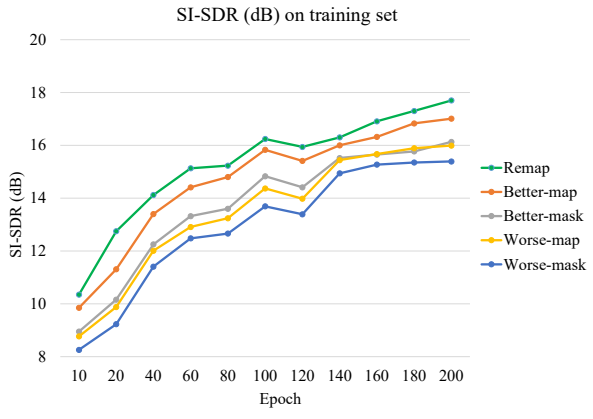


Fig. 4. SI-SDR value of through the five representations on training set in the training phase.

C. Complexity analysis

We calculate the increase in computational load brought by the additional predictive branch, as can be seen in the Table.III, there is only an increase of 0.7 M parameters, which is completely acceptable.

TABLE III
COMPLEXITY ANALYSIS ON THE DIFFERENT CONFIGURATIONS.

Model	Trainable Paras (M)	Total Paras (M)
TD-SpeakerBeam	7.1	18.3
TD-SpeakerBeam + PL	**7.8**	**19.0**

V. CONCLUSION

In this paper, we propose a "peer learning" method to acquire the shared characteristics and the discriminative representations through a two-branch speech representation prediction module. This scheme first constructs the two mixture samples related to the same target speaker for peer learning. A separable speech representation is generated by a predictor module, which is a discriminative representation for predicting the target speech waveform. Notably, the enhanced representation from one branch is leveraged to bolster the separation performance of its counterpart. We verify that the speech representations can be exchanged knowledge of each other in two branches. Extensive experiments have demonstrated the effectiveness of the proposed peer learning method. In the future, we plan to extend our methods to multi-channel and more speakers for target speech separation.

ACKNOWLEDGMENT

This paper was partially supported by NSFC (No:62176008).

REFERENCES

[1] R. Masumura, M. Ihori, A. Takashima, T. Tanaka, and T. Ashihara, "End-to-end automatic speech recognition with deep mutual learning," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2020, pp. 632–637.

[2] A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.

[3] S. Kindt, J. Thienpondt, and N. Madhu, "Exploiting speaker embeddings for improved microphone clustering and speech separation in ad-hoc microphone arrays," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[4] J. Chen, W. Rao, Z. Wang, *et al.*, "Mc-spex: Towards effective speaker extraction with multi-scale interfusion and conditional speaker modulation," *arXiv preprint arXiv:2306.16250*, 2023.

[5] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "Facefilter: Audio-visual speech separation using still images," *arXiv preprint arXiv:2005.07074*, 2020.

[6] R. Gu, S.-X. Zhang, L. Chen, *et al.*, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7319–7323.

[7] H. Taherian, A. Pandey, D. Wong, B. Xu, and D. Wang, "Multi-input multi-output complex spectral mapping for speaker separation," in *Proc. Interspeech*, 2023, pp. 1070–1074.

[8] M. Delcroix, T. Ochiai, K. Zmolikova, *et al.*, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 691–695.

[9] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," *arXiv preprint arXiv:2005.04686*, 2020.

[10] V. A. Kalkhorani, A. Kumar, K. Tan, B. Xu, and D. Wang, "Time-domain transformer-based audiovisual speaker separation," in *Proc. Interspeech*, 2023.

[11] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 46–50.

[13] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[14] Z. Yao, W. Pei, F. Chen, G. Lu, and D. Zhang, "Stepwise-refining speech separation network via fine-grained encoding in high-order latent domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 378–393, 2022.

[15] A. Mohamed, H.-y. Lee, L. Borgholt, *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[16] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning*, PMLR, 2020, pp. 9929–9939.

[17] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2021, pp. 15 490–15 500.

[18] J.-B. Grill, F. Strub, F. Altché, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[19] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[20] H. Nakamura, M. Okada, and T. Taniguchi, "Representation uncertainty in self-supervised learning as variational inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 484–16 493.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[22] R. Aihara, T. Hanazawa, Y. Okato, G. Wichern, and J. Le Roux, "Teacher-student deep clustering for low-delay single channel speech separation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 690–694.

[23] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4320–4328.

[24] H. M. Tan, D.-Q. Vu, C.-T. Lee, Y.-H. Li, and J.-C. Wang, "Selective mutual learning: An efficient approach for single channel speech separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 3678–3682.

[25] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*, PMLR, 2021, pp. 12 310–12 320.

[26] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 31–35.

[27] J. Han, Y. Long, L. Burget, and J. Černockỳ, "Dpccn: Densely-connected pyramid complex convolutional network for robust speech separation and extraction," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7292–7296.

Physics-Informed Neural Networks for Estimation of Scattered Sound Fields with Boundary Condition

Ryosuke Onizawa*, Gen Sato*, Izumi Tsunokuni* and Yusuke Ikeda*
* Tokyo Denki University, Tokyo
E-mail: {23fmi07, 24udc03, 21udc02}@ms.dendai.ac.jp, yusuke.ikeda@mail.dendai.ac.jp

***Abstract*—Sound field measurement helps us understand the sound fields that are difficult to understand only by hearing. Recently, methods for estimating sound fields using physics-informed neural networks (PINNs) based on the wave equation have attracted much attention, as they can reduce the measurement points. In this study, we propose a sound field estimation method for the scattered sound field with PINNs by introducing the boundary condition into the loss function as additional physical laws. From the simulation experiments, the proposed method can improve the estimation accuracies in the three-dimensional sound field around a scattering object with acoustic impedance.**

I. Introduction

In an ordinary room, sound propagates through reflections, diffraction, and scattering, depending on the materials and boundary shapes of the object such as walls and furniture. Analyzing such complex sound field is useful for various applications, including room acoustics design and acoustic rendering in virtual reality.

The acoustic scattering from walls and objects complicates the sound field and is especially challenging to model. In recent years, many modeling methods have been proposed to estimate the scattered sound fields. Based on the concept of compressive sensing [1], [2], sound fields can be modeled using only a small number of microphones. For example, using the superposition of point sources as a physical model of sound propagation [3], [4], the scattered sound field from a rigid sphere can be estimated at lower frequencies [5].

On the other hand, recently, estimation methods using deep learning have been proposed [6], [7]. In [8], the sound pressure level in a scattered sound field was estimated using a convolutional neural network (CNN) based on the geometry of scattering objects. However, the typical data-driven methods do not guarantee that the physical laws governing the outputs will hold, as the models are trained only on the dataset.

By contrast, physics-informed neural networks (PINNs) has been proposed in 2019 [9]. A notable feature of PINNs is that the output follows the laws of physics by introducing governing equations in the loss function. In acoustics, PINNs have been applied to estimate the sound field in various studies [10]–[13]. In these studies, the wave equation or Helmholtz equation is used as the governing equation for acoustics. Moreover, the scattered sound field around a rigid sphere was estimated using the wave equation and Neumann boundary condition using 32 microphones [14]. However, the case where the scattering object is not rigid, i.e., where the acoustic impedance of the object is finite, has not yet been investigated.

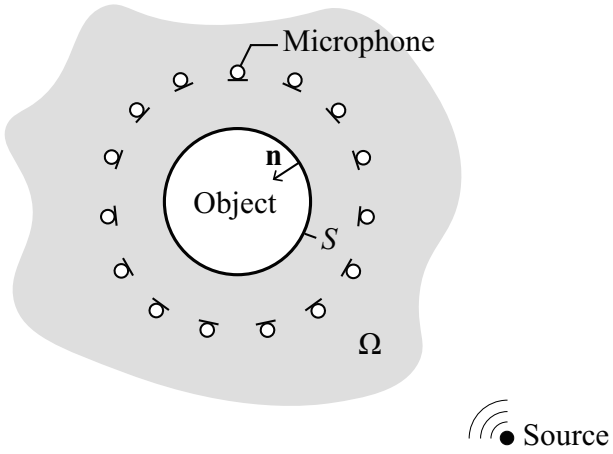


Fig. 1. Concept of estimation for scattered sound field Ω around a scattering object from an enclosing microphone array. The incident sound field is propagated from a sound source outside the microphone array. The boundary S of the scattering object has the acoustic impedance. $\mathbf{n}$ is normal vector on the boundary S.

In this study, we propose an estimation method of the scattered sound field around an object with acoustic impedance, based on the PINNs. The proposed method introduces the Helmholtz equation and the boundary condition into the loss function. Notably, we assume that the acoustic impedance of each boundary element varies smoothly with that of adjacent elements. To enforce this, we introduce a loss function that minimizes the differences in acoustic impedance between neighboring boundary elements. This allows us to model and estimate the scattered sound field around an object with arbitrary acoustic impedance. The key strength of the proposed method is its ability to estimate the sound field using a small number of microphones, even in cases where boundary conditions have sound absorption coefficients.

The rest of this paper is organized as follows. Section II describes the problem statement and proposed method. Section III-A describes the simulation conditions. Section III-B describes the results of three simulation experiments. Finally, we conclude this work in section IV.

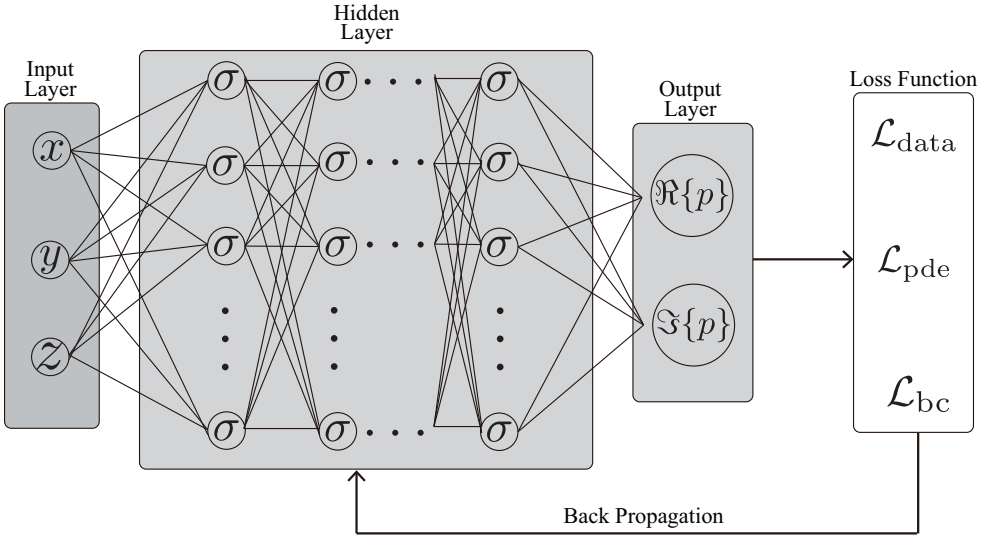


Fig. 2. Network architecture of proposed method. Input is three-dimensional coordinates, σ is the activation function, $\Re\{p\}$ and $\Im\{p\}$ are real and imaginary parts of complex sound pressure, respectively.

II. Proposed Method

A. Problem Statement

Fig. 1 shows the concept of the problem of scattered sound field estimation. This study estimates the three-dimensional scattered sound field Ω around a scattering object with spatially-smooth absorption coefficients. We assume that the sound sources are located outside the target region Ω and the microphone array. That is, since the sound field in Ω is homogeneous, the Helmholtz equation [15] holds as follows,

$$k^2 p(\mathbf{x}) + \nabla^2 p(\mathbf{x}) = 0 \quad (\mathbf{x} \in \Omega), \tag{1}$$

where k is wave number, ∇^2 is the Laplacian operator and $p(\mathbf{x})$ is the sound pressure $p(\in \mathbb{C})$ at the position $\mathbf{x}(\in \mathbb{R}^3)$. The N measurement points at position $\mathbf{x}_n (n = 1, 2, ..., N)$ are positioned around the scattering object in the region Ω.

An acoustic impedance at the boundary element of the scattering object is represented as follows.

$$Z(\mathbf{x}) = \frac{p(\mathbf{x})}{v(\mathbf{x})} \quad (\mathbf{x} \in S), \tag{2}$$

where v is normal particle velocity to the boundary S, $p(\mathbf{x})$ is sound pressure on the boundary. In this study, the acoustic impedance $Z(\mathbf{x})$ on S is assumed to be spatially smooth, for example, approximately uniform.

B. Proposed Method

Raissi *et al.* proposed PINNs, which introduce the governing equations into the loss function [9]. This study applies PINNs to an inverse problem, using the Helmholtz equation as the governing equation to ensure the output follows the physical laws of sound propagation. Additionally, the proposed method introduces boundary conditions into the loss function.

The proposed method employs the SIREN architecture [16] to perform higher-order derivatives of the loss function. Fig. 2 shows the network architecture of the proposed method. Using multilayer-perceptron (MLP), the function $\phi(\cdot)$ and the estimated sound field $\hat{p}(\mathbf{x})$ is represented as follows:

$$\hat{p}(\mathbf{x}) = \phi(\mathbf{w}, \mathbf{x}) = (f_L \circ f_{L-1}, ..., \circ f_2 \circ f_1), \tag{3}$$

where $\mathbf{x} = (x, y, z)$ is the coordinates as input and $\mathbf{w}$ is the weights. f_l is the function of the l-th layer of MLP. Then, the Helmholtz equation in (1) must be hold on the estimated sound pressures $p(\mathbf{x})$. In other words, the model should be trained so that the outputs satisfy the laws of acoustic physics.

As shown in (1), the Helmholtz equation has the second derivative of p. ReLU, one of the typical activation functions, is not appropriate for calculating the Helmholtz equation because it becomes zero after second-order differentiation. To obtain the derivatives by the automatic differentiation, it is required to choose the activation functions that are two-time differentiable. Thus, in this study, we use the sine function as an activation function. The MLP with sine activation function is known as the SIREN architecture [16]. Other examples include using hyperbolic tangent functions for activation functions [14].

By using the sine function, the l-th layer function f_l in (3) is as follows:

$$f_l(\mathbf{x}_l) = \sin\left(\omega(\mathbf{x}_l \mathbf{w}_l + \mathbf{b}_l)\right), \tag{4}$$

where $\mathbf{x}_l$ is the input of l-th layer, $\mathbf{w}_l$ is the weight parameter, $\mathbf{b}_l$ is bias of l-th layer and ω is hyperparameter.

To determine the weight parameter $\mathbf{w}$, the PINNs introduce a law of physics or a governing equation into the loss function. By also using the complex sound pressures measured at the microphone positions, the loss function $\mathcal{L}$ is defined by

$$\mathcal{L} = \mathcal{L}_{\mathrm{data}} + \lambda_1 \mathcal{L}_{\mathrm{pde}} + \lambda_2 \mathcal{L}_{\mathrm{bc}}, \tag{5}$$

$$\mathcal{L}_{\mathrm{data}} = \frac{1}{N} \sum_{n=1}^{N} \left| \hat{p}(\mathbf{x}_n) - p(\mathbf{x}_n) \right|^2 \ (\mathbf{x}_n \in \Omega), \tag{6}$$

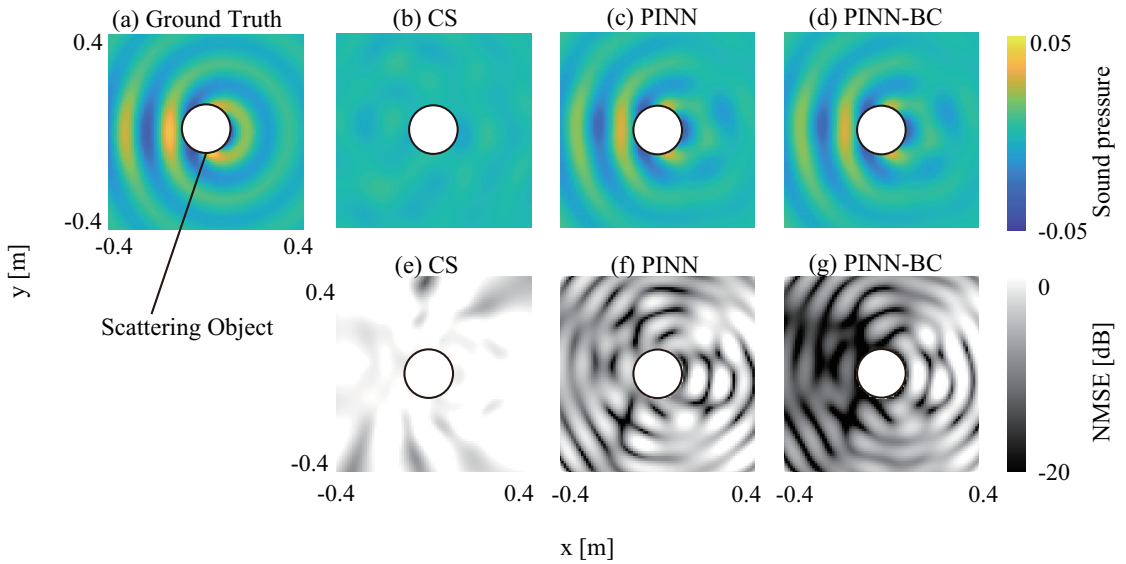


Fig. 3. Comparison of scattered sound fields and NMSE distributions ($z = 0$) at frequency is 2 kHz. (a) Ground Truth, (b) Estimated sound field by CS. (c) Estimated sound field by PINNs (without $\mathcal{L}_{\mathrm{bc}}$), (d) Estimated sound field by the proposed method, PINN-BC (with the loss of boundary condition $\mathcal{L}_{\mathrm{bc}}$), (e)-(g) are NMSE distributions by CS, PINN and PINN-BC, respectively. The number of microphone for the estimation was 32. The circle at the center indicates the scattering object.

TABLE I
SIMULATION CONDITION

Frequency [Hz]	2000
Absorption rate	0.5
Radius of scattering object [m]	0.1
Number of microphones	16×2, 25×2, 36×2
Radius of microphone array [m]	0.2, 0.25
Balance parameterλ_1, λ_2	10^{-4}, 1.0

$$\mathcal{L}_{\mathrm{pde}} = \frac{1}{M}\sum_{m=1}^{M}\left|k^2\hat{p}(\mathbf{x}'_m) + \nabla^2\hat{p}(\mathbf{x}'_m)\right|^2 \ (\mathbf{x}'_m \in \Omega), \quad (7)$$

$$\mathcal{L}_{\mathrm{bc}} = \sum_{e\in E}\sum_{e'\in E'}\left|Z(\mathbf{x}''_e) - Z(\mathbf{x}''_{e'})\right|^2 \ (\mathbf{x''}_e, \mathbf{x''}_{e'} \in S), \quad (8)$$

where $\hat{p}$ and p are estimated and measured sound pressures, respectively. λ_1, λ_2 are balance parameters. $\mathbf{x}_n$ is the microphone position, $\mathbf{x'}_m$ is three-dimensional coordinates selected randomly, $\mathbf{x''}_e$ is the point on the boundary of scattering object, $\mathbf{x''}_{e'}$ is a point neighbor to $\mathbf{x''}_e$.

$\mathcal{L}_{\mathrm{data}}$ is a loss function to calculate MSE between the estimated and measured sound pressures. $\mathcal{L}_{\mathrm{pde}}$ is a loss function to evaluate how well the estimated sound pressures satisfy the Helmholtz equation, as shown in (1). $\mathcal{L}_{\mathrm{bc}}$ is a loss function to evaluate the difference in acoustic impedance between the neighboring elements. Thus, by learning the weight parameter $\mathbf{w}$ so that $\mathcal{L}_{\mathrm{pde}}$ and $\mathcal{L}_{\mathrm{bc}}$ become small, the output of the MLP will more closely follow the Helmholtz equation and the boundary condition.

III. Simulation Experiment

A. Simulation Condition

We conducted the simulation experiment to evaluate the accuracy of estimating the scattered sound field around a sphere with a radius of 0.1 m in three dimensions. For the ground truth, the analytical solution [17] was calculated with spherical harmonics. The center of the scattering sphere object is the origin of the coordinate system. For $\mathcal{L}_{\mathrm{data}}$, the equally-spaced spherical microphone array had two layers with radii of 0.2 m and 0.25 m. The microphone array was positioned at the origin. The numbers of microphones were $M = 32, 50, 72$. The sound speed was 343 m/s, and the source position is placed at $(x, y, z) = (2, 0, 0)$ m. We employ the Adam optimizer with a learning rate of 1.0×10^{-5} for 15000 epochs. The number of randomly selected positions M for $\mathcal{L}_{\mathrm{pde}}$ is 10000 at each epoch. The number of boundary elements for $\mathcal{L}_{\mathrm{bc}}$ is 400 with equal spacing on the surface S. The weight parameters λ_1 and λ_2 are 1.0×10^{-4} and 1.0, respectively. The other details of simulation conditions are shown in Table I.

To evaluate the accuracy of the scattered sound field, we defined the normalized mean squared error (NMSE) as

$$\mathrm{NMSE} = 10\log_{10}\frac{\sum_{i=1}^{I}\left|\hat{p}(\mathbf{x}) - p(\mathbf{x})\right|^2}{\sum_{i=1}^{I}\left|p(\mathbf{x})\right|^2}. \quad (9)$$

where I is evaluation points.

B. Result

First, we compare the estimation accuracies of the proposed method (PINN–BC) and the two conventional methods: equivalent source method (CS) [4], [5] and PINNs with no boundary condition (PINN). In CS, the scattered sound field was modeled from the 32 microphones using the sparse equivalent source method. In simple PINN, the model was trained without using the loss of boundary condition $\mathcal{L}_{\mathrm{bc}}$ (8).

Fig. 3 compares scattered sound fields and NMSE distributions at 2 kHz by CS, simple PINN, and PINN–BC (proposed method) when 32 microphones were used. From Fig. 3 (a)–(d), simple PINN and PINN–BC were able to estimate the

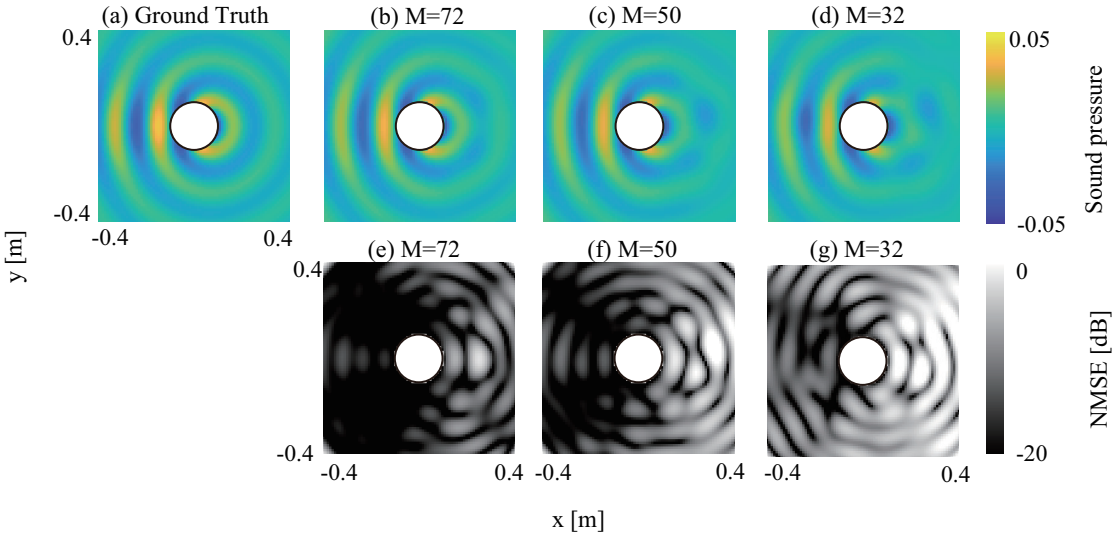


Fig. 4. Comparison of scattered sound fields and NMSE distributions ($z = 0$) by the number of microphones. (a) Ground Truth, (b) and (e) 72 microphones, (c) and (f) 50 microphones, (d) and (g) 32 microphones.

TABLE II
COMPARISON OF MEAN NMSES ON THE X–Y PLANE AND IN THREE-DIMENSIONAL EVALUATION POINTS

	CS	PINN	PINN–BC
Mean NMSE ($z = 0$)	0.2 dB	-5.1 dB	-8.3 dB
Mean NMSE (3D)	0.0 dB	-3.9 dB	-8.5 dB

scattered sound field significantly closer to the ground truth compared to CS. From Fig. 3 (e)–(g), PINN and PINN–BC significantly improved the estimation accuracies over the estimation region compared to CS. Furthermore, compared to simple PINN, PINN-BC has significantly improved the accuracy of estimating the scattered sound field, especially at locations opposite the sound source.

Table II shows the comparison of mean NMSEs on the x–y plane ($z = 0$) and in the three-dimensional evaluation points. For the $z = 0$ plane, the evaluation region was a square with 0.4 m on each side. For the 3D evaluation points, a spherical region between a radius of 0.1 m and 0.4 m was used as the evaluation region. On the $z = 0$ plane, the mean NMSEs of PINN and PINN–BC were improved by approximately 5.3 dB and 8.5 dB, respectively, compared to CS. Furthermore, in three-dimensional evaluation points, the estimation accuracy of PINN–BC was more stable than the other methods, with an estimation accuracy of -8.5 dB, approximately equivalent to the accuracy of the $z = 0$ plane. The proposed method improves the accuracy of 4.6 dB NMSE compared to PINN and 8.5 dB NMSE compared to CS. Therefore, the proposed method improved the estimation accuracies of scattered sound fields by introducing the boundary condition into the loss function.

Second, we compare the estimation accuracies by the number of microphones in the proposed method. Fig. 4 shows the estimated sound fields and NMSE distributions with 32, 50, and 72 microphones. As shown in Figs. 4(a)–(d), it was found that the scattered sound field can be estimated with high accuracy regardless of the number of microphones. From Figs. 4 (e)–(g), as the number of microphones decreased, the estimation accuracies were significantly degraded over the entire region. However, it is clear that even with 32 microphones, the high estimation accuracy was maintained on the opposite side of the sound source. Based on these results, the number of microphones should be increased according to the desired estimation accuracy.

IV. CONCLUSIONS

In this study, we proposed a deep-learning method for estimating the scattered sound field from a small number of microphones using the Helmholtz equation and the spatially smooth boundary condition. From the simulation experiments, the proposed method significantly improved the estimation accuracy by approximately 4.5 dB in NMSE when boundary conditions are introduced into the loss function as an additional physical model. From the other simulation experiments, it was found that the estimation accuracy significantly decreased with the number of microphones. However, even with a small number of microphones, it was found that it is possible to estimate a wide range of scattered sound fields.

In future work, we will study the estimation of scattered sound fields in the broad frequency band. Furthermore, we will estimate the scattered sound field around more complicated shapes of scattering objects. We also aim to evaluate the estimation accuracy in actual measurements and environments.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number 22K12099.

REFERENCES

[1] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[2] P. Gerstoft, F. C. Mecklenbräuker, W. Seong, and M. Bianco, "Introduction to compressive sensing in acoustics," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3731–3736, Jun. 2018.

[3] E. Fernandez-Grande, A. Xenaki, and P. Gerstoft, "A sparse equivalent source method for near-field acoustic holography," *The Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 532–542, Jan. 2017.

[4] I. Tsunokuni, K. Kurokawa, H. Matsuhashi, Y. Ikeda, and N. Osaka, "Spatial extrapolation of early room impulse responses in local area using sparse equivalent sources and image source method," *Applied Acoustics*, vol. 179, p. 108 027, 2021.

[5] R. Onizawa, I. Tsunokuni, and Y. Ikeda, "Visualization of scattered sound field by enclosing microphone array based on sparse equivalent source method," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, 2023, pp. 6771–6776.

[6] F. Lluís, P. Martínez-Nuevo, M. B. Møller, and S. E. Shepstone, "Sound field reconstruction in rooms: Inpainting meets super-resolution," *The Journal of the Acoustical Society of America*, vol. 148, no. 2, pp. 649–659, 2020.

[7] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft, "Generative models for sound field reconstruction," *The Journal of the Acoustical Society of America*, vol. 153, no. 2, pp. 1179–1190, Feb. 2023.

[8] Z. Fan, V.Vineet, H. Gamper, and N. Raghuvanshi, "Fast acoustic scattering using convolutional neural networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 171–175.

[9] M. Raissi, P. Perdikaris, and G. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[10] K. Shigemi, S. Koyama, T. Nakamura, and H. Saruwatari, "Physics-informed convolutional neural network with bicubic spline interpolation for sound field estimation," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[11] X. Karakonstantis and E. Fernandez-Grande, "Room impulse response reconstuction using physics-constrained neural networks," in *Proc. the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, 2023, pp. 3181–3188.

[12] M. Pezzoli, F. Antonacci, and A. Sarti, "Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses," in *Proc. the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, 2023, pp. 2177–2184.

[13] I. Tsunokuni, G. Sato, Y. Ikeda, and Y. Oikawa, "Spatial extrapolation of early room impulse responses with noise-robust physics-informed neural network," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. advpub, 2024EAL2015, 2024.

[14] X. Chen, F. Ma, A. Bastine, P. Samarasinghe, and H. Sun, "Sound field estimation around a rigid sphere with physics-informed neural network," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2023, pp. 1984–1989.

[15] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. 1999.

[16] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proc. NeurIPS*, 2020.

[17] N. A. Gumerov and R. Duraiswami, *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Elsevier Science, 2005.

WildPose: HRNet-based Lightweight and Efficient Wildlife Pose Estimation

Sibusiso Reuben Bakana[1], Yongfei Zhang[1,2*], Senior Member, IEEE and Bhekisipho Twala[3]

[1] School of Computer Science and Engineering, Beihang University, Beijing 100191, China

[2] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

[3] Faculty of Information and Communication Technology, Tshwane University of Technology, Tshwane, South Africa

*Corresponding Author: Yongfei Zhang (E-mail: yfzhang@buaa.edu.cn Tel: +86-1082314108)

***Abstract*— Lightweight and efficient wildlife monitoring algorithms, including pose estimation, that can operate on edge devices with limited computational resources are urgently needed for wildlife studies and protections. To reduce computational costs while maintaining accuracy in animal pose estimation, this paper introduces WildPose, an HRNet-w32 based model, designed, which integrates Efficient Channel Attention (ECA) to enhance important feature representations without complex operations, Non-Local Blocks (NLB) to capture long-range contextual information and handle occlusions, and Depthwise Convolutions (DWConv) to reduce computational complexity and parameters. Additionally, the Online Hard Keypoint Mining with Mean Squared Error (OHKMSE) loss function is employed to improve accuracy for occluded keypoints detection by focusing on harder keypoints. When evaluated on the largest wildlife dataset of Animal Kingdom, WildPose demonstrated a nice trade-off between accuracy and efficiency, 75% reduction in parameters and 65% reduction in GFLOPs, while achieving an increased Percentage of Correct Keypoints (PCK) for occluded keypoints such as the hip, ankle, and tail. Therefore, WildPose can serve as a practical solution for real-time wildlife monitoring and artificial intelligence-based ecological studies, especially in resource-constrained environments.**

Keywords: Wildlife Pose, Pose Estimation, Efficient, Lightweight

I. INTRODUCTION

Ecological environments are increasingly damaged due to human activities, leading to the extinction of animals in aquatic and savanna settings. Understanding animals through images and videos aids in evaluation behavioral information crucial for monitoring and conservation [1]. Continuous monitoring of an animal's state through pose estimation (PE) is critical and reflects dynamic behavior. For instance, detecting if an animal is limping helps understand its health. PE captures an animal's pose, making it essential for behavior analysis and highlighting the need for new technologies to recognize animal behavior quickly and intelligently.

Due to large datasets and consistent subjects, PE studies has mainly focused on humans, which can be roughly clarified in two categories, i.e., regression-based methods and heatmap-based models.

Regression-based methods: directly predict the coordinates of keypoints in an image using neural networks, offering a simpler and faster approach. For instance, DeepPose [2] performs regression learning on keypoints using a cascaded architecture. Combining Appearance and Regression (CAR) [3] employs a multi-stage approach for refining keypoint predictions through regression model that incorporate appearance features. DensePose [4] maps each pixel to a point on the body surface for detailed geometry understanding. While effective, these regression-based models often struggle with accuracy in complex poses, compared to heatmap-based methods.

Heatmap-based methods: Building on these regression-based methods, heatmap-based models have become popular due to their improved accuracy in locating keypoints. In these models, the input is passed through a network that involves high-to-low-resolution subnetworks connected in series to elevate resolution. For example, Hourglass network [5] recovers high resolution through a process of low-to-high-resolution transitions, while SimpleBaseline [6] implements a few transposed convolutions to generate high resolution. Additionally, an advanced architecture known as High-Resolution Net (HRNet) [7] maintains high resolution throughout the process.

HRNet has set a new standard for PE performance and has been enhanced for accuracy. For example, Li et al. [8] proposed the Dynamic Information Transfer (DIT) to improve accuracy. Bao et al. [9] proposed ECA-HRNet for high-speed sports motion analysis. Li et al. [10] introduced HRNeXt for occlusion in crowded scenes, demonstrating computational efficiency compared to baseline HRNet.

Although great progress has been made in human PE, they are not applicable for wildlife PE. Animals vary significantly in shape, size, and texture, making PE more challenging [11]. This inconsistency has led to less attention to computer vision-based ecological studies, particularly animal PE. Even as interest in animal PE grows, the focus has been on improving accuracy rather than developing lightweight and efficient models for resource-constrained (RC) environments [12]. The Animal Kingdom dataset [13] facilitates these studies by providing a diverse collection of animals across six major groups with 33K frames for PE tasks. Developing animal PE models by leveraging advanced human PE models often overlooks the intensive computational requirements.

PE faces challenges, particular with occluded keypoints like ankles and hips in wildlife settings. These keypoints are crucial for understanding movement but are often not visible due to large areas coverage limitations. This poses challenges in RC settings where edge devices must be computationally efficient (CE), require minimal power, and have less storage and memory. In such locations, technology must respond quickly for efficient monitoring, as wildlife is vulnerable to poaching and extinction. Therefore, lightweight, and efficient wildlife PE algorithms that can operate on edge devices with limited computational resources are urgently needed. Human PE, a great progress, but not applicable for wildlife PE as it faces significant challenges in balancing accuracy, computational efficiency, and speed, which are crucial for practical deployment in wildlife monitoring. Several models have been

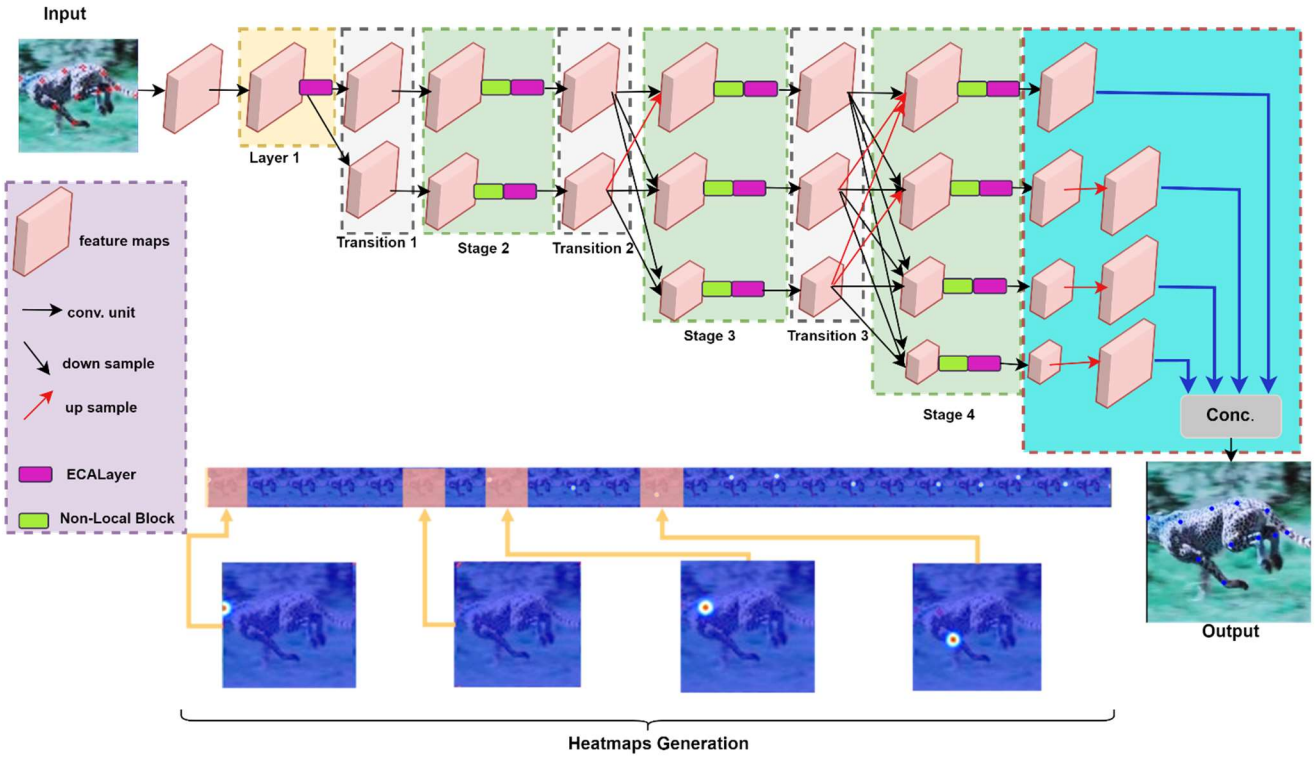


Fig. 1 WildPose – lightweight and efficient wildlife pose estimation model

proposed with varying degrees of success in addressing these issues. HRNet-w32 [7] while highly accurate, it suffers from high computational costs and significant parameter size, making it impractical for deployment on edge devices.

ViPNAS-Resnet50 [14] and HRFormer-w32 [15] models offer reduced parameters and GFLOPs but at the cost of decreased accuracy and significantly lower FPS, hindering their applicability in real-time scenarios. HRNeXt [10] and Lightweight-Cross Fussion [16], although these models improve accuracy and reduce computational demands, they face issues with speed and efficiency in processing high-resolution inputs. Lightweight SHN [17] and DSC-HRNet-w32-v3 [18] both achieves better accuracy and efficiency but still struggle to provide the necessary balance for deployment in resource-constrained environments. He et al. [19] addressed bird diversity challenge by proposing VHR-BirdPose, combining Vision Transformer (ViT) and HRNet. Wang et al. [20] used HRNet with Swin Transformer (HRST) for pig joint detection. Dai et al. [21] proposed a 3D constraint-based HRNet for wild animal settings. Liao et al. [22] introduced THANet, a cross-domain vision transformer for animal PE. Zhang et al. [23] proposed a cross-modal approach using contrastive learning. Chen et al. [24] combined object detection and PE for autopilot systems. However, existing animal PE models mainly focuses on improving the accuracy.

Recognizing above challenges, we proposed WildPose, an HRNet-based model designed to address the balance between accuracy, computational cost, and speed. Our model integrates several advanced components to overcome these issues. ECA [25] enhances feature representation by adaptively recalibrating channel-wise feature, focusing on crucial parts of the image without adding significant computational cost. NLB [26] captures long-range dependencies within feature maps, essential for accurately predicting dispersed and occluded keypoints. DWConv [27] replaces standard convolutions in the Basicblock from stage 2 to stage 4, significantly reducing computational cost and the number of parameters. OHKMSE improves the accuracy of harder-to-predict keypoints by dynamically focusing on keypoints with higher loss values.

Our evaluations demonstrate that WildPose achieves a significant reduction in GFLOPs and parameters, along with improved accuracy for occluded keypoints and a higher FPS, setting a new standard for efficient animal PE models.

II. Proposed WildPose Model

A. Model Architecture

The systematical architecture of the proposed WildPose model is depicted in Fig. 1, in which Efficient Channel Attention (ECA) is incorporated to enhance important feature representations without complex operations (Section II.B), Depthwise Convolution (DWConv) to reduce computational complexity and parameters (Section II.C), and Non-Local Blocks (NLB) to capture long-range contextual information and handle occlusions (Section II.D), the modules configurations illustrated by Fig. 2. Additionally, the Online Hard Keypoint Mining with Mean Squared Error (OHKMSE) loss function (Section II.E) is employed to improve accuracy for occluded keypoints detection by focusing on harder keypoints.

B. Efficient Channel Attention

To maximize the effectiveness of ECA, it was integrated after key convolutional layers in all stages of WildPose, refining feature maps to emphasize important channels and ensuring relevant features for keypoint allocation with minimal computational cost. Adjusting and rounding to odd numbers maintains convolutional symmetry, preserving feature integrity, and allowing symmetric padding, which maintains the spatial alignment of feature maps. The global average pooling

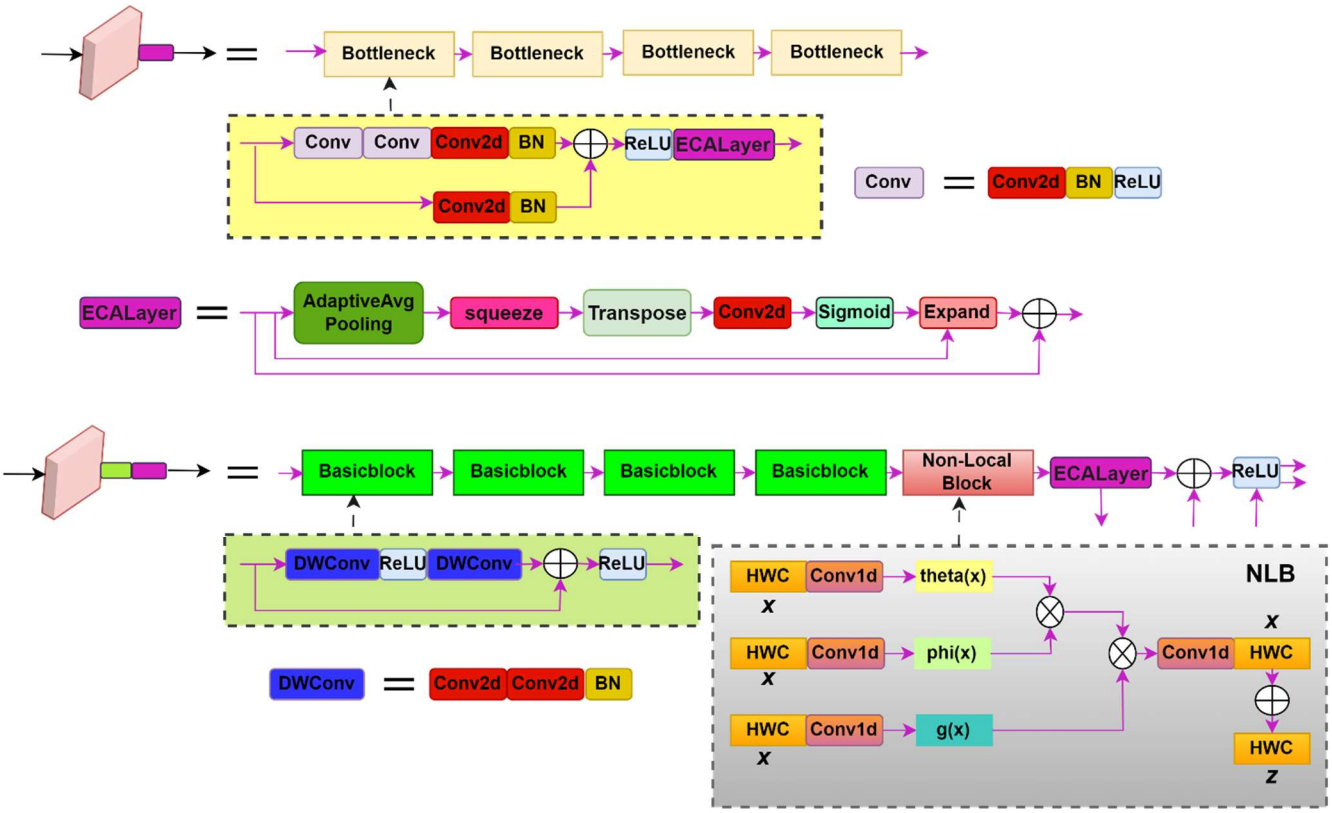


Fig. 2 Modules configurations for WildPose

operation reduces each channel of the input feature map X (with dimensions C, H, W) to a single value, resulting in a vector Z of size (C,1), as illustrated by Equation. (Eq.) (1)

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{c,i,j} \quad (1)$$

$$Y = \text{Conv1D}(Z, k) \quad (2)$$

The pooled vector is processed using 1D convolution with dynamically calculated k as in Eq. (2), as Y is a result of applying 1D convolution to Z and produce a vector of the same size (C, 1). Where z_c is the global averaged value for c^{th} of the input feature map X. $X_{c,i,j}$ being the value of the input feature map at channel, height index i, and width index j. The

Output Eq. (2) is passed via sigmoid activation σ to produce the final channel attention weights as in Eq. (3). S is the vector of (C, 1) and each element S_c is attention weight for corresponding channel.

$$S = \sigma(Y) \quad (3)$$

$$\tilde{X}_{c,i,j} = X_{c,i,j}.S_c \quad (4)$$

The original X is recalibrated using the channel attention weights S to give an output of the ECA as in Eq. (4). This enhances the model's ability to learn meaningful channel-wise feature at low computation.

C. *Depthwise Separable Convolution (DWConv)*

CNNs enhance PE but are often computationally inefficient due to their size. WildPose addresses this by using DWConv instead of standard convolution. Standard convolution processes all input channels together, while DWConv uses depthwise convolutions to operate on each channel individually and pointwise convolutions with 1x1 filters to merge these outputs. This approach preserves channel depth, efficiently integrates channel information, and modifies the channel count.

D. *Non-Local Block*

HRNet Basicblock modules are used in stage 2 to 4 to ensure that the network is maintaining strong performance, allowing the model to be both powerful and deployable, but not in edge devices as it uses standard convolution, which are computationally demanding. In the Basicblock we employed DWConv to make the model deployable to RC settings. We also incorporated NLB strategically to capture long-range dependencies within the feature maps. While this NLB comes with computational cost, it was insignificant because of the DWConv that accounted for it. NLB receives the input feature map X during channel reduction with dimensions (B, C, H, W), where B is batch size, C is number of channels, and H and W are height and width respectively. The input is processed to reduce the C to ($C/2$). This reduction is achieved by using three 1x1 convolutional layers namely $'g'$, $'theta' = \theta$, $and\ 'phi' = \phi$, multiplying Conv2d with the reduced C.

The output of these convolutional layers is reshaped and permuted to prepare for matrix multiplications. Therefore, an affinity matrix f is computed by matrix multiplication of θ_x and ϕ_x. This matrix is the representation of the pairwise relationship between all positions. It then normalised this affinity matrix, using a softmax function, then the information from all positions is then aggregated through an attention map operation. The results then get reshaped back to the original dimensions and processed by a final convolutional layer called $'w'$ indicated by Eq. (5) and (6) respectively, which is added to the input X to facilitate identity mapping through a residual connection as represented by Eq. (7).

$$y = y.view(B, inter_{channels}, H, W) \quad (5)$$

$$W_y = W(y) \quad (6)$$

$$z = \mathrm{W_y} + X \quad (7)$$

The NLB in WildPose enhanced model's ability to capture long-range dependencies, an important aspect to understand relations between distant keypoints in PE tasks.

E. Loss function

WildPose employs OHKMSE loss function, derived from the Mean Squared Error (MSE) loss used in the baseline model. OHKMSE is an improvement over standard MSE loss, it focuses on hard-to-predict keypoints and achieves that by selecting the hardest keypoints in each batch based on their loss values and computing the average loss. As Animal Kingdom has 23 key points, we selected top 12 hardest key points for WildPose. In animal PE, hardest keypoints are often those that frequently move or are occluded, such as ankles and hips.

The OHKMSE selects the top-k hardest keypoints based on their loss values as in Eq. (8), where *B* is the batch size, and *k* is the number of hardest keypoints based on their loss values. The final loss is computed by averaging the OHKMSE loss across *B* as in Eq. (9).

$$\mathcal{L}_{OHKM} = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{K}\sum_{k\in \text{Top-K}}\mathcal{L}_{MSE}, k\right) \quad (8)$$

$$\mathcal{L}_{final} = \frac{1}{B}\sum_{b=1}^{B}\mathcal{L}_{OHKM}, b \quad (9)$$

III. Experimental Results

This section provides the experimental results and performance comparisons of the proposed WildPose model against comparable schemes, with experiment settings in Section III.A, performance comparisons in III.B, visualization of keypoints predictions III.C and finally ablation studies in III.D.

A. Experimental Settings

- *Hardware settings*

We conducted our experiments on Windows 11 Pro (64-bit) system with an Intel Core i7-8565U CPU @ 1.80GHz x 16GB memory, remotely connected to an Ubuntu 18.04.6 LTS system with an NVIDIA GeForce GTX 1080 Ti GPU, using CUDA 10.2 CUDA10.2 for GPU parallelization. The computer vision library used was PyTorch.

- *Dataset*

The performance of WildPose was evaluated on a diverse Animal Kingdom dataset, which provides PE annotations to enhance understanding of natural animal behaviors. This dataset includes 33 000 manually annotated frames, covering 850 species across six major animal classes, recorded at different times of the day in various environments with diverse backgrounds, viewpoints, illumination, and weather conditions. For our evaluation, we utilized 413 species. The dataset was split into training and testing sets in 9:1 ratio, with 23 keypoints annotated for each animal class. Fig. 3 illustrates sample keypoints distributions and annotations.

- *Training Configuration*

The model was trained for 200 epochs using a batch size of 16 per GPU. The Adam optimizer was employed with an initial learning rate of 0.001, which was reduced by a factor of 0.1 at epochs 170 and 200. To mitigate overfitting, a weight decay of 0.01 was applied. Data augmentation techniques, including random rotation, scaling, and horizontal flipping, were used to enhance generalization. The model architecture was fine-tuned with pre-trained weights from pose_hrnet_w32 model,

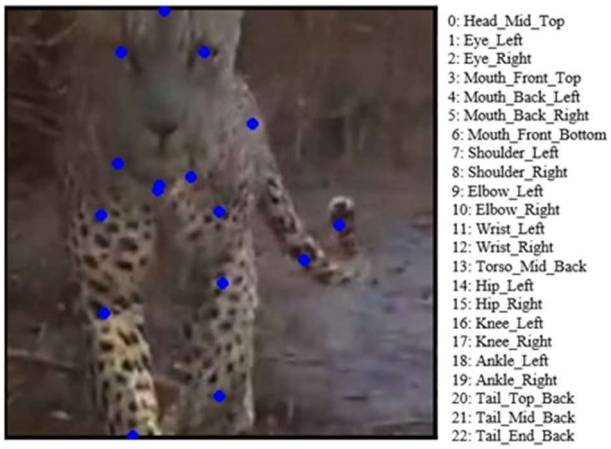


Fig. 3 Sample of keypoints annotation

originally trained on the MPII dataset. Keypoints were represented using heatmaps of 64 x 64 resolution, with a Gaussian sigma of 2.

- *Evaluation Metrics*

WildPose is evaluated with metrics that are essential, such as percentage of correct keypoints (PCK) as in Eq. (10). The mean PCK (mPCK) is computed as the weighted average of PCK score for different keypoints, considering the visibility of each keypoint as in Eq. (11). Where $\hat{y}_k$ being the predicted location of keypoint *k*, y_k being the ground truth location of keypoint *k*, *d* is the distance threshold, α is the distance threshold, and 1 is the indicator function that returns 1 if the condition is true or 0 otherwise. The PCK_k is the PCK score for keypoint *k*, and w_k is the weight for keypoint *k*, calculated based on its visibility.

$$\text{PCK} = \frac{1}{N_{keypoints}}\sum_{k=1}^{N_{keypoints}} 1\left(\frac{\|\hat{y}_k - y_k\|}{d} < \alpha\right) \quad (10)$$

$$\text{mPCK} = \sum_{k=1}^{N_{keypoints}} \text{PCK}_k \text{ x } w_k \quad (11)$$

The time taken for WildPose to estimate the test dataset is evaluated to gauge its real-time suitability. The inference time is then measured by recording the time to process each batch of the input during validation, with Equation (12), where *N* is the number of batches, and $time_t$ is the time taken for the ith batch.

$$\text{Inference Time} = \frac{1}{N}\sum_{i=1}^{N} time_t \quad (12)$$

To understand computational cost, as measured in terms of floating-point operation (FLOPs), is demonstrated by Eq. (13). Where $FLOPS_{sl}$ is the number of floating-point operations in layer *l*, and L is the total number of layers.

$$\text{FLOPs} = \sum_{l=1}^{L} FLOPS_{sl} \quad (13)$$

The model's complexity is measured through the number of parameters. The number of parameters is calculated as in Eq. (14), where $Parameters_{sl}$ is the number of parameters in layer *l*, and L is the total number of layers.

$$\text{Parameters} = \sum_{l=1}^{L} Parameters_{sl} \quad (14)$$

Furthermore, the balance must be measured also through models frames per second (FPS) that the model is capable to gauge its real-time suitability. Equation (15) reveals how WildPose processes images.

$$\text{FPS} = \frac{\text{Batch Size}}{\text{Time per Batch}} \quad (15)$$

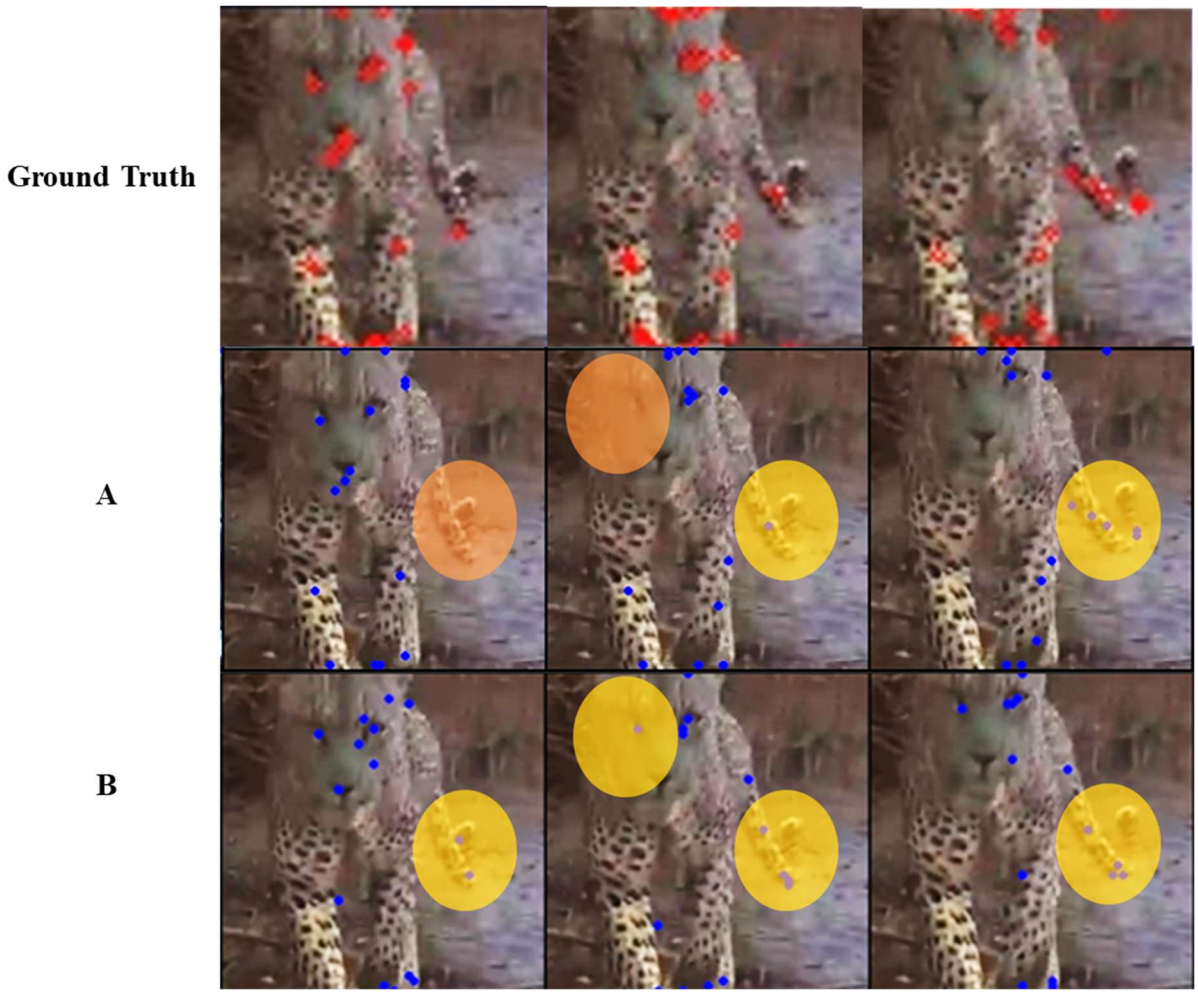


Fig. 4 Visual analysis of keypoint predictions for HRNet (A) and WildPose (B) for a walking leopard

B. Performance Comparison

Table 1 shows the SOTA comparison results. WildPose represents a 75% decrease in parameters compared to the baseline model, with a slight decrease of 5.7% and 11.3% compared to [14] and [15], respectively. WildPose has slightly more parameters than [17] and [18] by 5.7% and 12.2%. In terms of GFLOPs, WildPose shows a 65% reduction from HRNet-w32 and 20% fewer GFLOPs compared to [15], but 40% more than [14] and 92% less than [18]. WildPose achieves an mPCK of 53.281, which is 4.6% decrease compared to the baseline, 2.8% lower than [14], and 0.96% lower than [15], while 28% lower than [17]. WildPose significantly outperforms all models in FPS, achieving 581.818 FPS, a 113% increase over HRNet-w32's 272.804 FPS and vastly superior to [14], [15], and [18]. These improvements underscore WildPose's potential for efficient and accurate wildlife PE in surveillance videos and edge devices, contributing significantly to poaching mitigation efforts.

Table 2 compares the accuracy of keypoint detection using the PCK@0.5 metric between the baseline and our proposed model. WildPose achieves better accuracy results for hard-to-

TABLE. 1 COMPARISON OF WILDPOSE WITH SOTA LIGHTWEIGHT PE MODELS

Model	Input	Parameters	GFLOPs	mPCK	FPS
HRNet-w32	256x256	28.536 M	9.494	55.856	272.804
ViPNAS-Resnet50 [14]	256x256	7.29 M	**1.99**	54.8	22.90
HRFormer-w32 [15]	256x256	7.75 M	4.17	53.8	6.04
HRNeXt [10]	256x192	9.5 M	2.5	63.1	-
Lightweight-Cross Fusion [16]	368x368	-	18.02	44.4	-
Lightweight SHN [17]	256x256	6.48 M	8.23	**74.63**	-
DSC-HRNet-w32-v3 [18]	256x256	**6.03 M**	41.3	68.9	68
Ours	256x256	6.871 M	3.321	53.281	**581.818**

TABLE. 2 COMPARISON RESULTS OF PCK@0.5 ACCURACY

Keypoints	Head	Shoulder	Elbow	Wrist	Hip	knee	Ankle	Mouth	Tail	mPCK	FPS
HRNet	**66.790**	**51.155**	**47.859**	**52.387**	46.308	**42.908**	44.518	**73.651**	42.028	55.856	272.804
Ours	63.462	50.498	41.915	43.197	**46.852**	38.297	**46.650**	68.540	**44.179**	53.281	**581.818**

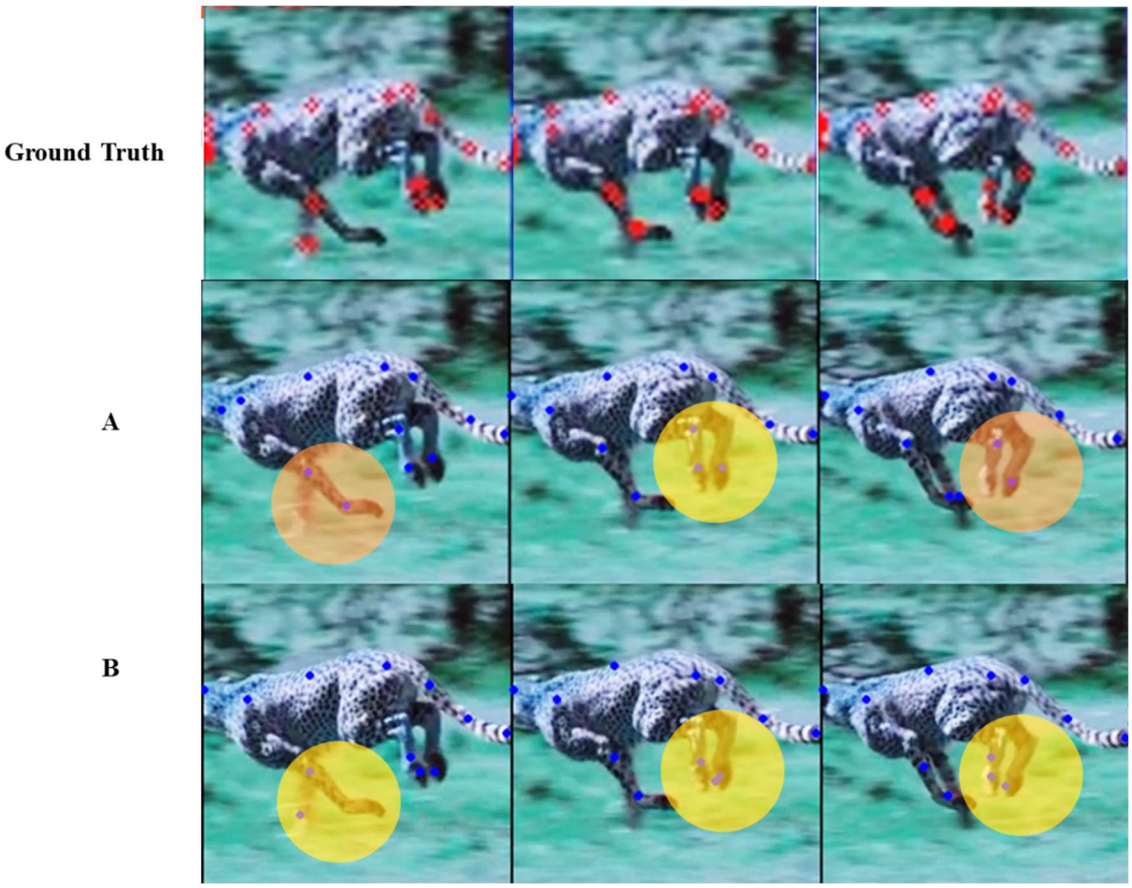


Fig. 5 Visual analysis of keypoint predictions for HRNet (A) and WildPose (B) for a running cheetah

predict keypoints such as the ankle, tail, and hip, which are critical for analyzing animal behavior and movement. The slight increase in accuracy for these keypoints (ankles by 4.79%, tail by 5.12%, and hip by 1.18%) indicates that WildPose can more reliably track these vital points, offering deeper insights into animal motions and interactions. While maintaining competitive mPCK for other keypoints, the advancement is noteworthy given that the Animal Kingdom dataset consist of frames extracted from video sequences. The usage of OHKMSE in our design significantly enhances the model's capability to predict these challenging keypoints accurately, even in motion. While HRNet-w32 achieves lower and more stable training loss values, indicating a robust and consistent training process, WildPose, despite having a higher training loss, excels in FPS and accuracy for hard-to-predict keypoints. This trade-off highlights WildPose's design optimization for real-time applications and specific keypoint detection, making it suitable choice for wildlife PE where efficiency and accuracy of critical keypoints are prioritized.

C. Visualization of Keypoints Predictions

Fig. 4 and 5 illustrate the keypoint predictions for a walking leopard and a running cheetah, respectively. Both HRNet-w32 and WildPose accurately predict keypoints across various animal motions. However, WildPose demonstrates improved accuracy and stability, particularly for challenging keypoints such as the ankles, hips, and tail. This enhanced performance is essential for detailed behavior and movement analysis in wildlife studies. The yellow circles in the figures indicate accurately predicted keypoints, while the orange circles represent missed predictions. The robustness and precision of WildPose make it essentially suitable for high-fidelity motion tracking applications, such as wildlife monitoring and real-time behavior analysis.

D. Ablation studies

Table 3 presents the results of our ablation studies, which systematically evaluate the impact of different architectural modifications and loss functions on model performance. The configurations are as follows:

- **Model A**: Baseline HRNet using the Mean Squared Error (MSE) loss function.
- **Model B**: HRNet with OHKMSE loss function (Top12).
- **Model C**: HRNet integrated with NLB, using the MSE loss function.
- **Model D**: HRNet with both NLB and ECA, using the OHKMSE loss function (Top12).
- **Model E**: Incorporates NLB, ECA, and DWConv for both Bottleneck and Basicblock modules, using the MSE loss function.
- **Model F**: Similar to Model E, but uses the OHKMSE (Top12) loss function.
- **Model G**: NLB, ECA, and DWConv applied at the Basicblock level with the OHKMSE (Top8) loss function.
- **Model H**: Combines NLB, ECA, and DWConv at the Basicblock level, with the OHKMSE (Top12) loss function.

The results show the following trends:

TABLE. 3 ABLATION STUDIES

Model	Loss Function	mPCK	FPS	GFLOPs	Parameters (M)
A	MSE	55.856	272.804	9.494	28.536
B	OHKMSE - Top12	55.872	283.186	9.494	28.536
C	MSE	56.584	270.27	9.697	29.241
D	OHKMSE-Top12	**58.107**	264.026	9.697	29.241
E	MSE	50.394	341.88	**2.768**	**6.727**
F	OHKMSE-Top12	43.450	303.03	**2.768**	**6.727**
G	OHKMSE-Top8	51.617	361.582	3.321	6.872
H	OHKMSE-Top12	53.281	**581.818**	3.321	6.872

- ✓ **Effect of OHKMSE Loss**: Incorporating the OHKMSE loss function (Models B, D, F, G, and H) consistently improves the mPCK compared to using MSE (Models A, C, E). This improvement is more pronounced when combined with NLB and ECA, as seen in Model D, which shows a significant increase in mPCK compared to Model C.
- ✓ **Impact of NLB and ECA**: The addition of NLB (Model C) and the combination of NLB and ECA (Model D) increase mPCK, indicating that these components help the model better capture long-range dependencies and channel-wise importance.
- ✓ **Effect of DWConv**: Models incorporating DWConv (Models E-H) show a substantial reduction in GFLOPs and parameters, demonstrating increased computational efficiency. However, there is a trade-off in mPCK, with some loss in accuracy, especially noticeable in Model F. This trade-off is mitigated in Model H, where a more balanced performance is achieved by applying DWConv at Basicblock level, yielding competitive mPCK and the highest FPS.
- ✓ **OHKMSE Top12 vs Top8**: Comparing Models G and H shows that the Top12 strategy yields better accuracy (mPCK), while the Top8 configuration provides higher efficiency (FPS). This highlights the trade-off between accuracy and speed, allowing for flexibility in choosing the model configuration based on application requirements.

Model H emerges as the most balanced configuration, offering competitive mPCK, reduced GFLOPs and parameters, and exceptionally high FPS, making it suitable for real-time wildlife PE tasks.

IV. CONCLUSION

We introduced WildPose, a lightweight and efficient PE model tailored for wildlife monitoring in resource-constrained environments. By integrating ECA, NLB, and DWConv, WildPose significantly reduces computational costs and parameters while maintaining competitive accuracy. The use of the OHKMSE loss function further enhances the model's ability to handle difficult-to-predict keypoints, such as ankles, hips, and tails.

Evaluations on the Animal Kingdom dataset show that WildPose achieves a 65% reduction in GFLOPs and a 75% reduction in parameters, while doubling the FPS compared to the baseline HRNet model. These improvements make WildPose suitable for real-time wildlife behavior analysis, particularly is RC settings and on edge devices with limited power consumption.

Despite these advancements, potential limitations exist. The model's generalization capabilities across different environmental conditions, backgrounds, or animal species not represented in the training data could impact performance. Future work will explore domain adaptation techniques and more diverse datasets to mitigate these challenges.

In terms of real-time applicability, WildPose demonstrates a significant increase in FPS, achieving a double to the baseline speed. This positions WildPose favorably against existing real-time systems like ViPNAS-Resnet50 [14] and HRFormer-w32 [15], which often struggle to balance accuracy and efficiency in RC environments. The model's lightweight nature makes it ideal for deployment on devices such as a CPU with a Quad-core Cortex-A72 (ARM v8) 64-bit SoC at 1.5GHz, 4GB or 8GB LPDDR4-3200 SDRAM, and a high-quality camera module powered by 5V/3A DC via USB-C.

WildPose sets a new standard for efficient animal PE, significantly contributing to ecological conservation efforts and enabling practical applications in real-time wildlife monitoring.

ACKNOWLEDGEMENT

This work was partially supported by the National Natural Science Foundation of China (No. 62072022), the Fundamental Research Funds for the Central Universities and the Council for Scientific and Industrial Research (CSIR) of South Africa.

REFERENCES

[1] S. Ison, C. Cvitanovic, G. Pecl, A. J. Hobday, and I. van Putten, "The role of visual framing in marine conservation communication," *Ocean Coast Manag*, vol. 248, p. 106938, Feb. 2024, doi: 10.1016/J.OCECOAMAN.2023.106938.

[2] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660. doi: 10.1109/CVPR.2014.214.

[3] X. Chen and A. Yuille, "Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations," *Adv Neural Inf Process Syst*, vol. 27, Jul. 2014, [Online]. Available: http://arxiv.org/abs/1407.3399

[4] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense Human Pose Estimation In The Wild," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.00434

[5] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in *ECCV*, Mar. 2016, pp. 483–499. [Online]. Available: http://arxiv.org/abs/1603.06937

[6] B. Xiao, H. Wu, and Y. Wei, "Simple Baselines for Human Pose Estimation and Tracking," in *European Conference on Computer Vision (ECCV)*, Apr. 2018. [Online]. Available: http://arxiv.org/abs/1804.06208

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in *Computer Vision and Pattern Recognition (CVPR)*, Feb. 2019. [Online]. Available: http://arxiv.org/abs/1902.09212

[8] Y. Li, Q. Shi, J. Song, and F. Yang, "Human Pose Estimation via Dynamic Information Transfer," *Electronics (Basel)*, vol. 12, no. 3, Feb. 2023, doi: 10.3390/electronics12030695.

[9] W. Bao, T. Niu, N. Wang, and X. Yang, "Pose estimation and motion analysis of ski jumpers based on ECA-HRNet," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-32893-x.

[10]Q. Li, Z. Zhang, F. Zhang, and F. Xiao, "HRNeXt: High-Resolution Context Network for Crowd Pose Estimation," *IEEE*

Trans Multimedia, vol. 25, pp. 1521–1528, 2023, doi: 10.1109/TMM.2023.3248144.

[11] L. Jiang, C. Lee, D. Teotia, and S. Ostadabbas, "Animal pose estimation: A closer look at the state-of-the-art, existing gaps and opportunities," *Computer Vision and Image Understanding*, vol. 222, 2022, doi: 10.1016/j.cviu.2022.103483.

[12] S. Liu, Q. Fan, S. Li, and C. Zhao, "MAPoseNet: Animal pose estimation network via multi-scale convolutional attention," *J Vis Commun Image Represent*, vol. 97, 2023, doi: 10.1016/j.jvcir.2023.103989.

[13] X. L. Ng, K. E. Ong, Q. Zheng, Y. Ni, S. Y. Yeo, and J. Liu, "Animal Kingdom: A Large and Diverse Dataset for Animal Behavior Understanding," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19001–19012. doi: 10.1109/CVPR52688.2022.01844.

[14] L. Xu *et al.*, "Vipnas: Efficient video pose estimation via neural architecture search," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16067–16076. doi: 10.1109/CVPR46437.2021.01581.

[15] Y. Yuan *et al.*, "HRFormer: High-Resolution Transformer for Dense Prediction," in *Advances in Neural Information Processing Systems*, 2021, pp. 7281–7293.

[16] X. Zhu, X. Zeng, and W. Ma, "Lightweight Cross-Fusion Network on Human Pose Estimation for Edge Device," *IEEE Access*, vol. 11, pp. 134899–134907, 2023, doi: 10.1109/ACCESS.2021.3065574.

[17] W. Zhang, Y. Xu, R. Bai, and L. Li, "Animal Pose Estimation Algorithm Based on the Lightweight Stacked Hourglass Network," *IEEE Access*, vol. 11, pp. 5314–5327, 2023, doi: 10.1109/ACCESS.2022.3231750.

[18] Z. Zhao, A. Song, S. Zheng, Q. Xiong, and J. Guo, "DSC-HRNet: a lightweight teaching pose estimation model with depthwise separable convolution and deep high-resolution representation learning in computer-aided education," *International Journal of Information Technology (Singapore)*, vol. 15, no. 5, pp. 2373–2385, Jun. 2023, doi: 10.1007/s41870-023-01297-5.

[19] R. He, X. Wang, H. Chen, and C. Liu, "VHR-BirdPose: Vision Transformer-Based HRNet for Bird Pose Estimation with Attention Mechanism," *Electronics (Basel)*, vol. 12, no. 17, Sep. 2023, doi: 10.3390/electronics12173643.

[20] X. Wang, W. Wang, J. Lu, and H. Wang, "HRST: An Improved HRNet for Detecting Joint Points of Pigs," *Sensors*, vol. 22, no. 19, Oct. 2022, doi: 10.3390/s22197215.

[21] X. Dai, S. Li, Q. Zhao, and H. Yang, "Animal Pose Estimation Based on 3D Priors," *Applied Sciences (Switzerland)*, vol. 13, no. 3, 2023, doi: 10.3390/app13031466.

[22] J. Liao, J. Xu, Y. Shen, and S. Lin, "THANet: Transferring Human Pose Estimation to Animal Pose Estimation," *Electronics (Switzerland)*, vol. 12, no. 20, 2023, doi: 10.3390/electronics12204210.

[23] X. Zhang, W. Wang, Z. Chen, Y. Xu, J. Zhang, and D. Tao, "CLAMP: Prompt-based Contrastive Learning for Connecting Language and Animal Pose," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogntion*, Jun. 2023, pp. 23272–23281. doi: 10.1109/CVPR52729.2023.02229.

[24] K. Chen, Y. Yu, and H. Zhou, "Animal Movement Prediction based on the Corroboration of Object Detection and Pose Estimation," 2023.

[25] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11531–11539. doi: 10.1109/CVPR42600.2020.01155.

[26] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local Neural Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[27] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Oct. 2017, pp. 1251–1258. [Online]. Available: http://arxiv.org/abs/1610.02357

Teager Energy Cepstral Coefficients for Audio Deepfake Detection

Ritik Mahyavanshi*, C.V. Mahesh Reddy†, Arth J. Shah*, and Hemant A. Patil*
* Dhirubhai Ambani Institute of Information and Communication Technology, Gujarat, India
* E-mail : {202001030, 202101154, hemant_patil}@daiict.ac.in
† Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana.
† E-mail : cvnreddym9@gmail.com

***Abstract*—Audio deepfakes have emerged as a significant concern, as these convincingly mimicked synthetic audio and can be exploited to manipulate targeted groups by altering speech information. No prior research has explored the use of Teager Energy Operator (TEO)-based features for ADD task. This paper discusses a novel technique for distinguishing deepfake audio from real ones using Teager Energy Cepstral Coefficients (TECC) features, which captures the energy fluctuations of audio signals, making it a promising approach for detecting real *vs.* deepfake audio. For classification, the ResNet-50 model is employed in combination with the FoR (Fake or Real) dataset. Experimental results demonstrate the efficacy of this method, achieving an Equal Error Rate (EER) of approximately 7.53 % and an accuracy of 92.55 % on static TECC feature sets. The findings indicate that TECC features outperform traditional techniques, such as Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC). This superior performance highlights the significant potential of TECC features for further research in deepfake detection.**

***Index Terms*- Audio Deepfake Detection, Teager Energy Cepstral Coefficients, FoR Dataset.**

I. Introduction

Deepfake is a type of deep learning technique to create or manipulate video and audio content, typically to depict someone saying or doing something that never actually happened. One real-life example of deepfake technology impacting individuals involves a case, where in 2019, criminals used Artificial Intelligence (AI)-generated audio to impersonate a CEO's voice and issue fraudulent instructions to transfer funds. It highlighted the importance of better security measures and awareness to prevent the misuse of deepfake technology in sensitive situations. The rise of deepfake technology has extended beyond visual media, posing significant challenges to the authenticity of audio recordings. Audio deepfakes, generated through sophisticated deep learning models, can convincingly mimic human speech, leading to potential misuse in various domains [1]. This study focuses on an innovative approach for Audio Deepfake Detection (ADD). It leverages Teager Energy Cepstral Coefficients (TECC) derived from Teager Energy Operator (TEO) in combination with a ResNet-50 classifier. The approach is evaluated using the standard Fake or Real (FoR) dataset [2].

Researchers previously used various techniques for ADD task. Many studies have applied different models to identify fake audio clips. In [3], they suggested using the Wav2vec2 model to train a siamese neural network for safeguarding anti-spoofing models from attacks.

The focus in ADD has been on identifying suitable audio features, broadly divided into are used and learning-based features. Features, such as Mel Frequency Cepstral Coefficients (MFCC) [4], [5] and Linear Frequency Cepstral Coefficients (LFCC) [6], [7] closely mimic human auditory traits and concentrate on low frequency information crucial for speech detection tasks. The emergence of learning-based audio features in recent years has garnered significant attention for ADD research. Utilizing whisper audio features [8] has shown promise in detecting synthetic speech, surpassing handcrafted features with the vast amount of supporting data available for the whisper system. Self-supervised learning-based audio features have also displayed benefits in ADD tasks by leveraging pre-trained data from diverse sources to help differentiate between genuine and fake speech effectively even in complex situations and thus, performing well on varied datasets.

In our work, we have proposed to use TECC features as in previously used methodology, such as MFCC and LFCC features, which captures various details in audio signals through different filter window lengths across frequency bands but as it is primarily used for music analysis and in comparison; these features which have good match with human auditory characteristics are more suitable for ADD systems have a only drawback of more sensitivity towards noise. In our approach of using TECC features having more resistant to noise as TECC, which is derived by TEO naturally having noise suppression capability making TECC a potential feature set for ADD systems. Previously TECC have also been used for vatious tasks such as dysarthic classification [9], and classification of normal *vs.* whispered speech [10]. In our approach, we utilize TECC features, which are inherently resistant to noise due to the mathematical structure of TEO from which they are derived. This noise resistance makes TECC a promising area for research in ADD systems. We use the extracted TECC features as inputs for the ResNet-50 classifier [11], which is adapted for audio classification tasks. ResNet-50's deep architecture, renowned for its success in image classification, is modified to process TECC features, enabling it to discern between genuine *vs.* fake audio recordings. The FoR dataset, comprising a diverse collection of real and deepfake audio samples, is used to train and validate the proposed detection

system. Experimental results demonstrate that the combination of TECC features with the ResNet-50 classifier achieves high accuracy in identifying audio deepfakes, surpassing traditional ADD methods.

The the rest of the paper is organised as follows: In Section II gives explanation about the TEO and TECC. Section III gives description of experimental setup used in this study, and Section IV presents experimental results of our research followed by summary and conclusions in the last Section V.

II. Proposed Methodology

TEO is employed to capture non-linear aspects of energy variations in audio signals, providing a means to extract proposed TECC features. These featuers are effective in identifying fluctuations in the running estimate of energy patterns of speech, which are indicative of nonlinearities in production of natural speech.

A. TEO

TEO is a nonlinear operator utilized in signal processing to estimate the instantaneous energy of a signal. Its effectiveness lies in its ability to detect changes in both amplitude and frequency, making it particularly useful for the non-linearity, which define the naturality of a speech production mechanism.

Mathematically, TEO for a discrete-time signal $x[n]$ is defined as [12]:

$$\psi(x[n]) = x^2[n] - x[n-1] \cdot x[n+1]. \quad (1)$$

Here, $\psi(x[n])$ represents the transformed signal via TEO containing the instantaneous energy information, $x(n)$ being the discrete-time signal, and n represents the sample index.

The TEO can estimate the amplitude envelope of amplitude modulated (AM) signals and the instantaneous frequency of frequency modulated (FM) signals. It can also be used for signals with changing amplitude and frequency (AM-FM signals). The TEO profile of a audio signal is shown in Fig. 1.

B. TECC Feature Extraction

The functional block diagram of TECC extraction is show in the Fig. 2. Here, the input speech signal is passed through a pre-emphasis highpass filter having a Z domain system function, $H(z) = 1 - \alpha z^{-1}$; with a typical value of α (i.e. α is zero in H(z)) is set to 0.97. It is used in speech processing to enhance the quality and intelligibility of the speech signal. It works by amplifying the high frequency components of the speech signal, which typically have lower energy compared to the low frequency components. This emphasis on high frequencies helps to improve the signal-to-noise ratio (SNR), making the speech signal clearer, especially in noisy environments.

Now, the signal is passed through the Gabor filterbank [13], which consists of two fuctions: a Gaussian function and a sinusoidal waveform. The Gaussian function, $\exp(-b^2t^2)$, forms a bell-shaped envelope around the input signal, creating a window for the sinusoidal waveform, $\cos(\omega_c t)$, to be modulated. This windowing effect allows a specific range of

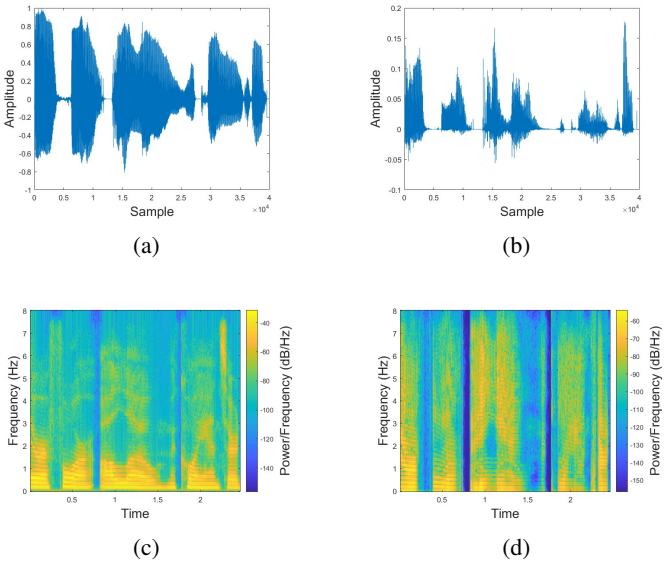


Fig. 1: (a) Time-domain waveform, (b) corresponding TEO profile of the input signal shown in Fig 1 (a), (c) spectrogram of the signal shown in Fig. 1 (a), and (d) TEO spectrogram of the Fig. 1 (b).

frequencies to pass through, effectively subband filtering the signal.

The factor b in the Gaussian function determines the spread of the bell-shaped envelope, impacting how much of the input signal is encompassed by the window. A larger b results in a wider spread, capturing more of the input signal, while a smaller b results in a narrower spread, capturing less of the input signal. The term $\exp(-b^2t^2)$ represents the Gaussian function, while $\cos(\omega_c t)$ represents the sinusoidal waveform. The factor b in the Gaussian function affects the bell-shaped envelope's spread over the input signal. The impulse response of the Gabor filter $h(t)$ is given by:

$$h(t) = \exp(-b^2t^2)\cos(\omega_c t).$$

In the context of deepfake audio detection, the Gabor filterbank can be used to extract specific features from the audio signal. By adjusting the parameters of the Gaussian function and the sinusoidal waveform, it is possible to isolate and analyze different frequency components of the signal, aiding in the identification of synthetic or manipulated audio.

Now, TEO is applied to the formed signal in order get the TEO profile of the input signal. Here, all the frequency fluctuations and the instantaneous energy of the input signal can be observed. TEO profile of both real and fake audio is shown in the Fig. 1.

The framing and averaging step is crucial for breaking down the continuous speech signal into smaller parts called speech framer. This process involves splitting the audio into short, overlapping frames. Each frame is looked at separately under the assumption of short-term stationarity (and assumption of linear time-invariant (LTI) system), making it easier to

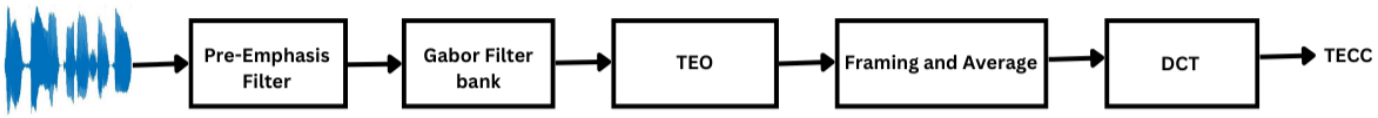


Fig. 2: Fuctional Block Diagram of TECC Feature Extraction After. [12].

spot subtle inconsistencies or anomalies typical of deepfake audio and efficienty exploit linear signal prediction algorithms. Averaging within these frames helps in standardizing the signal, reducing noise, and emphasizing important features that may show tampering. By studying these frames, we can catch abnormal variations and artifacts introduced by deepfake techniques that often struggle to maintain natural coherence over brief time periods.

The Discrete Cosine Transform (DCT) phase is crucial for enhancing feature extraction in ADD. The DCT is applied to the log filter-bank subband energies, converting them into the cepstral-domain. This transformation captures essential frequency-domain features while reducing redundancy. The resulting TECC are then utilized to identify abnormalities and artifacts unique to deepfake audio, as these coefficients are highly sensitive to subtle energy variations introduced by synthetic speech methods. The application of DCT ensures that the extracted features are decorrelated, enhancing their effectiveness in distinguishing between genuine and fake audio samples. In particular, DCT is applied for feature vector decorrelation, dimensionality reduction of underlying TECC features vector, and energy compaction. Ultimately, the TECC feature encapsulates the TEO profile of the input signal, providing a robust representation for detecting audio deepfakes.

In real audio, glottal closures [14] cause distinct and sharp spikes in the TEO profile, reflecting the natural rhythm of the speaker's vocal fold vibrations. The energy released during each closure significantly contributes to the clarity and naturalness of the speech. Conversely, fake audio often simulates these spikes with less pronounced or irregular energy-levels during glottal closures. Synthetic methods struggle to replicate the natural dynamics of glottal closures, resulting in less distinct energy spikes or piles in the TEO profile. Within the regions between closures, real audio exhibits smoother and more gradual energy fluctuations, reflecting the controlled damping characteristics of human speech. This stability in energy patterns over time underscores the consistent nature of genuine vocal fold vibrations and glottal airflow. In contrast, fake audio shows more erratic energy fluctuations within these regions, with rapid rises and falls that *disrupt* natural continuity. This inconsistency in energy transitions highlights the artificial nature of the generated speech and its inability to sustain energy-levels as effectively as real speech.

The differences in energy fluctuations and persistence between real *vs.* fake audio provide crucial indicators for detecting audio deepfakes. Analyzing TEO profiles allows for the identification of genuine speech characteristics, such as stable and persistent energy patterns in real audio, contrasting sharply with the erratic and rapidly decaying energy in fake audio. These distinctions are instrumental in developing effective deepfake detection algorithms, ensuring the integrity and authenticity of audio content.

III. Experimental setup

A. Dataset Used

Several datasets featuring synthetic speech have been previously published, yet there are compelling reasons for introducing the dataset outlined in this study. Existing datasets predominantly lack utterances generated from the latest deep-learning-based speech synthesis algorithms. Furthermore, their volume of utterances is often insufficient for training complex neural network models effectively. Moreover, most focus on spoofed utterance detection for speaker verification systems. The FoR dataset [2], tailored for ADD, is a meticulously curated collection of audio recordings aimed at advancing research in identifying synthetic or manipulated speech. It encompasses a diverse array of audio samples, including authentic human speech and various types of deepfake audio produced through techniques, such as Text-to-Speech (TTS) and voice conversion (VC).

TABLE I: Distribution of FoR-Norm Corpus. After [2].

	Fake	**Real**
Training	26,924	26,938
Testing	2,370	2,264
Validation	5,398	5,399

FoR Dataset , which contains more than 87,000 synthetic utterances as well as more than 111,000 real utterances. Such a dataset is fundamental for research in synthetic speech detection, since it contains enough data to train the most complex deep learning algorithms, which makes it more suited for Deep Neural Network (DNN), such as ResNet-50 classifier.

B. Pattern Classifier Used

We performed experiments using ResNet-50 as a two-class classifier, where the two classes are real and fake speech signals [11]. ResNet-50, a variant of the Residual Networks (ResNet) architecture, excels in various image recognition tasks due to its depth and sophisticated design. The architecture consists of four main blocks, each progressively deeper and more complex. The first block has three convolutional layers that capture low-level features, such as *edges* and *textures*. Each convolutional layer is followed by batch normalization

and a ReLU activation function to stabilize the learning process and introduce non-linearity.

The second block contains four convolutional layers, building on the features detected in the first block and identifying more complex patterns. The third block, with six convolutional layers, captures detailed and abstract features essential for distinguishing subtle differences in the data. The fourth and final block has three convolutional layers, consolidating the features learned in the previous blocks.

After these four blocks, the feature maps undergo global average pooling, reducing their spatial dimensions by computing the average of each feature map and summarizing the spatial information. This reduction helps prevent overfitting and reduces computational complexity. The architecture then includes a fully-connected layer that integrates the pooled features into a single vector, which is passed through a softmax activation function to output a probability distribution over the different classes, providing the final classification results.

ResNet-50's design, with its depth and residual connections that facilitate gradient flow, makes it highly effective for deep feature extraction and robust learning. This architecture has been successfully applied in numerous applications, including deepfake detection, where it benefits from its ability to learn and distinguish intricate patterns in visual data.

IV. Experimental Results

We focused on the TECC features and explored a thorough dimensional analysis comparison with existing approaches to evaluate its efficacy and efficiency. The dimensional analysis revealed that TECC offers superior performance in key metrics, such as accuracy, precision, and computational complexity. Additionally, a detailed latency period analysis was performed to assess the real-time applicability of TECC.

A. Effect of Dimension of TECC Feature Vector

To optimize audio classification models, we have thoroughly investigated both static and dynamic conditions for the TECC features. This exploration is crucial for refining model performance, as selecting the ideal conditions for feature extraction can significantly impact accuracy. Our analysis revealed that among various configurations, the static TECC features with dimension 30 provides the most effective results, achieving a remarkable accuracy rate of 92.55% and EER of 7.53%. Given its superior performance, we have considered this static configuration as the optimal condition for further classification purposes. The dynamic features show an accuracy rate of 91.45% with 28 dimension. While notable, this does not surpass the highest accuracy achieved under static conditions. Thus, only static conditions are preferred for classification.

B. Comparison with Cepstral Features

Analysis of different feature sets under these optimal conditions reveals that TECC leads with an accuracy of 92.55%, an F1-Score of 92.52%, and an EER of 7.53%. In comparison, MFCC achieves 66.61% accuracy, 63.05% F1-Score, and 32.63% EER, while LFCC records 49.87% accuracy, 35.33% F1, and 49.01% EER. This demonstrates that TECC outperforms both MFCC and LFCC, as shown in Table II.

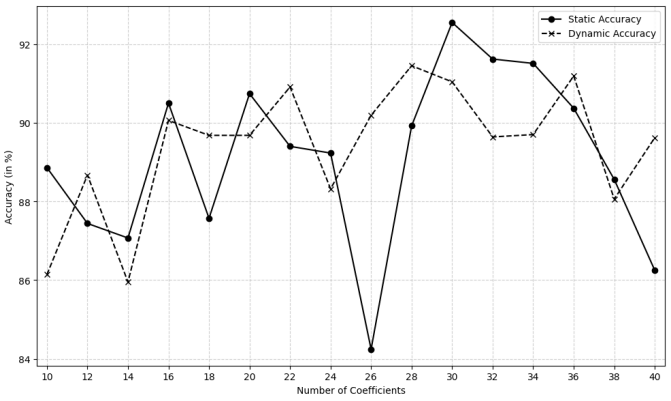


Fig. 3: Effect of Dimension of TECC Feature Vector.

TABLE II: Performance Comparison with Existing Features

Feature Set	Accuracy	F1-Score	EER
LFCC	49.87%	35.33%	49.01%
MFCC	66.61%	63.05%	32.63%
TECC	**92.55%**	**92.52%**	**7.53%**

C. Analysis of Latency Period

To analyze the proposed TECC feature vector and optimize speech frame length, we need to reduce storage requirements. Latency analysis helps to determine the minimum data needed to effectively train the model, thereby optimizing storage requirements. We explored different latency conditions for various features, including TECC, MFCC, and LFCC. The graph shown in Fig. 4 labels the accuracy performance of TECC, LFCC, and MFCC w.r.t. varying speech frame lengths.

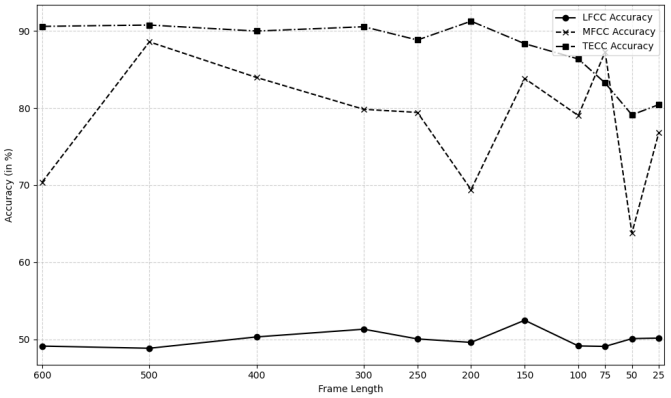


Fig. 4: Analysis of Latency Period for LFCC, MFCC, and TECC.

TECC maintains an impressive accuracy range between 79.11% and 91.26%. This high level of performance is especially notable at a frame length of 200, where TECC achieves

TABLE III: Comparison With Related Works in the Literature

Source	Feature Set	Classifier	Accuracy (in %)
[15]	MFCC-20	XGB	59
		KNN	62
		RF	62
		SVM	67
[2]	Timbre Model Analysis (Brightness, Hardness, Depth, Roughness)	NB	67.27
		DT	70.26
		RF	71.47
		SVM	73.46
[2]	STET, Mel-Spectrogram, MFCC, and CQT	VGG19	89.79
Proposed	**TECC**	**ResNet-50**	**92.55**

its highest accuracy of 91.26%. In contrast, MFCC, although capable of high accuracy, demonstrates unreliable performance with substantial drops, and LFCC consistently underperforms. The results demonstrated that TECC significantly reduces the latency period compared to existing methods, ensuring faster and more reliable outcomes. These findings highlight TECC's capability to improve performance and reduce delays, establishing it as a reliable solution for real-world applications. Therefore, TECC's ability to deliver robust and dependable accuracy across different frame lengths makes it the superior feature extraction technique for ensuring consistent and reliable results.

D. Comparison with Existing Studies

In this study, we assessed the performance of various classifiers and feature sets for a classification task. The results, as summarized in Table III, reveal a notable range in effectiveness across different combinations. Among the feature sets evaluated, the TECC feature set used with the ResNet-50 classifier achieved the highest accuracy of 92.55 %. This superior performance underscores the efficacy of the TECC feature set in conjunction with the ResNet-50 classifier compared to other combinations. The findings highlight the significant impact that the choice of feature set and classifier can have on classification accuracy, with the proposed TECC feature set demonstrating the best overall performance in our experiment.

V. Summary and Conclusions

The study underscores the superior performance of combining TECC features with a ResNet-50 classifier for detecting audio deepfakes. TECC's ability to capture nonlinear characteristics enhances the model's accuracy in differentiating between real *vs.* fake audio. Evaluation on the standard FoR dataset shows that TECC consistently achieves the lowest EER and highest accuracy compared to existing MFCC and LFCC features. Specifically, TECC's lower EER values indicate its superior performance in minimizing both false acceptance and false rejection rates, and its higher accuracy underscores its effectiveness in correct classification. This suggests that TECC, when used with a ResNet-50 classifier, not only reduces classification errors more effectively than MFCC and LFCC but also offers robust performance in real-world audio deepfake detection. Future research could build on this success by integrating additional features and classifiers to further enhance detection capabilities and address more advanced deepfake techniques.

Appendix: noise superrasion of TEO

The TEO has demonstrated its noise suppression capability in previous analyses, particularly for speech recognition tasks, including handling various noises and person recognition in noisy settings. Here, we examine TEO's effectiveness in suppressing additive noise. Let $s[n]$ represent the clean speech signal and $\hat{s}[n] = s[n] + w[n]$ denote the noisy speech signal, where $w[n]$ is a zero-mean additive noise component. We present the TEO profiles for both $s[n]$ and $w[n]$:

$$\begin{aligned} \Psi\{s[n]\} &= s^2[n] - s[n-1]s[n+1], \\ \Psi\{w[n]\} &= w^2[n] - w[n-1]w[n+1]. \end{aligned} \tag{2}$$

The TEO profile for the noisy speech signal $\hat{s}[n]$ is calculated as:

$$\begin{aligned} \Psi\{\hat{s}[n]\} &= \hat{s}^2[n] - \hat{s}[n-1]\hat{s}[n+1] \\ &= (s[n] + w[n])^2, \\ &\quad - (s[n-1] + w[n-1])(s[n+1] + w[n+1]) \\ &= s^2[n] + 2s[n]w[n] + w^2[n], \\ &\quad - s[n-1]s[n+1] - s[n-1]w[n+1], \\ &\quad - w[n-1]s[n+1] - w[n-1]w[n+1]. \end{aligned} \tag{3}$$

Moreover from [12], we know that:

$$E[\Psi\{\hat{s}[n]\}] \approx E[\Psi\{s[n]\}]. \tag{4}$$

The Eq. (4) indicates that when TEO is applied to a noisy signal with zero-mean additive noise, it can effectively suppress the noise. Thus, TEO exhibits a notable noise suppression capability.

Acknowledgements

This study has been funded by Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, under the sponsorship of project 'BHASHINI', (Grant ID: 11(1)2022-HCC.(TDIL)). Authors thank authorities of of DA-IICT Gandhinagar for their kind support and cooperation to carry out this research work.

REFERENCES

[1] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: Challenges and future directions," *Algorithms*, vol. 15, no. 5, 2022, ISSN: 1999-4893. [Online]. Available: https://www.mdpi.com/1999-4893/15/5/155.

[2] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, IEEE, 2019, pp. 1–10.

[3] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection," in *Proc. INTERSPEECH*, 2023, Dublin, Ireland, pp. 2808–2812.

[4] A. Hamza, A. R. R. Javed, F. Iqbal, *et al.*, "Deepfake audio detection via MFCC features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022, 9996362.

[5] A. Qais, A. Rastogi, A. Saxena, A. Rana, and D. Sinha, "Deepfake audio detection with neural networks using audio features," in *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*, 2022, Hyderabad, India, pp. 1–6.

[6] P. Kawa, M. Plata, and P. Syga, "Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection," *arXiv preprint arXiv:2206.13979*, 2022, {Last Accessed Date : $22^{nd} July, 2024$}.

[7] R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou, and X. Li, "Audio deepfake detection system with neural stitching for add 2022," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 9226–9230.

[8] Y. Yang, H. Qin, H. Zhou, *et al.*, "A robust audio deepfake detection system via multi-view feature," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, Seoul, Korea, pp. 13 131–13 135.

[9] A. Kachhi, A. Therattil, A. T. Patil, H. B. Sailor, and H. A. Patil, "Teager energy cepstral coefficients for classification of dysarthric speech severity-level," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, Chiang Mai, Thailand, pp. 1462–1468.

[10] K. Khoria, M. R. Kamble, and H. A. Patil, "Teager energy cepstral coefficients for classification of normal vs. whisper speech," in 28th *European Signal Processing Conference (EUSIPCO)*, 2021, Dublin, Ireland, pp. 1–5.

[11] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio deepfake approaches," *IEEE Access*, vol. 11, pp. 132 652–132 682, 2023.

[12] M. R. Kamble and H. A. Patil, "Analysis of Reverberation via Teager Energy Features for Replay Spoof Speech Detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2019, Brighton, UK, pp. 2607–2611.

[13] R. Hammouche, A. Attia, S. Akhrouf, and Z. Akhtar, "Gabor filter bank with deep autoencoder based face recognition system," *Expert Systems with Applications, pp. 11*, vol. 197, 2022.

[14] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2011.

[15] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian Journal for Science and Engineering, vol. 47, pp. 3447–3458*, 2021.

Parameterizing Hierarchical Particle Filters with Concept Drift for Time-varying Parameter Estimation

Joshua Murphy*, Conor Rosato†, Andrew Millard* and Simon Maskell*
* Dept. of Electrical Engineering and Electronics, University of Liverpool, United Kingdom
† Dept. of Pharmacology and Therapeutics, University of Liverpool, United Kingdom
E-mail: {joshua.murphy, cmrosa, andrew.millard, smaskell}@liverpool.ac.uk

***Abstract*—Models used in real-world applications can experience concept drift, where the deployment environment differs from the training environment due to previously unseen input data distributions or changes in the data generation process. Sequential Monte Carlo squared (SMC^2) is a Bayesian algorithm which infers the distribution of both the states and parameters of non-linear, non-Gaussian state space models. A Sequential Monte Carlo (SMC) sampler operates on the static parameter space and samples the parameters with which a number of particle filters (PFs), tackle the dynamic state inference problem. PFs often do not consider the influence of concept drift on the parameters of interest. Models that consider drift will do so in a purely online context and can suffer from slow convergence. By formulating PFs as hierarchical models, with explicit dynamics to describe parameter evolution, the use of batch parameter estimation techniques like SMC^2 is enabled. We demonstrate our approach by applying five concept drift models to an epidemiological disease model with time-varying transmission rates.**

I. Introduction

Bayesian inference is well-suited for quantifying uncertainty because it outputs distributions rather than single-point estimates. This approach allows for the integration of prior knowledge, the connection of observations to quantities through likelihood distributions, and the generation of posterior distributions. It can be utilized to infer unknown quantities in both static and dynamic contexts [1].

Particle filters (PFs) [2] are a widely used Bayesian method for tracking the dynamic evolution of state spaces in non-linear and non-Gaussian state space models (SSMs). They have achieved widespread usage in a variety of fields [3]–[5]. Typically, in PFs, the parameters governing the evolution of the dynamics are fixed. However, in some instances, PFs can be expanded to infer these parameters. One approach is to augment the state space with the parameters and consider their co-evolution [6], [7].

Bayesian parameter estimation using PFs can occur online [8]–[11] and offline. One such online recursive method is the nested particle filter (NPF) which uses one PF to estimate the parameters and an inner particle filter to track the states [12], [13]. Offline methods include particle marginal methods, such as particle Markov Chain Monte Carlo (p-MCMC) [14] and Sequential Monte Carlo squared (SMC^2) [15], which are batch methods that require access to the full data-set at each evaluation. Similarly to the NPF, an inner particle filter tracks the states but an outer static Monte Carlo method deals with parameter estimation. SMC^2 is similar to the NPF in that they both contain two Sequential Monte Carlo methods but SMC^2 possesses favourable convergence properties while the NPF has favourable time complexity [12].

Concept drift refers to changes in the statistical environment of the problem at hand, a term commonly used in classification contexts [16]–[18]. In the context of PFs, concept drift has been modelled in online streaming contexts [19]–[21] and rarely associated with batch inference [22]. A pertinent example of a changing statistical environment is the modelling of communicable diseases. It has been noted in [23] that the use of concept drift in compartmental disease modelling is under-utilized. Two related examples can be found in [22], [24] where the abrupt concept drift (see Section III) is used to define the sudden change in the transmissibility of Ebola and HIV. Particle-MCMC was employed to infer the parameters of the disease model as well as the hyper-parameters governing the sudden change in transmissibility. As interventions from policymakers are introduced, the disease dynamics changes over time. It is therefore common practice to define β as time-varying [25]–[28].

The contribution of this paper is to explicitly define an epidemiological model with frequently encountered types of concept drift and incorporate them as part of a hierarchical particle filter to perform batch Bayesian inference with SMC^2. The paper continues by describing the SMC^2 algorithm in Section II and defining the dynamic models used to represent concept drift in Section III. The compartmental epidemiological model is defined in in Section IV and numerical results are provided in Section VI. Conclusions and future work are given in Section VII.

II. SMC^2

SSMs have been used to model unobserved latent states of dynamic systems in a wide range of research fields [29]. An SSM consists of a state equation,

$$\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1} \sim p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}), \tag{1}$$

JM and AM were funded by a Research Studentship jointly funded by the EPSRC Centre for Doctoral Training in Distributed Algorithms EP/S023445/1; CR was funded by the Wellcome CAMO-Net UK grant: 226691/Z/22/Z and SM was funded by EPSRC through the Big Hypotheses under Grant EP/R018537/1. The authors would like to thank Molly Waring and Yiyi Whitchelo for their invaluable contribution.

which is parameterized here by $\boldsymbol{\theta}$ and models the Markovian movement of the dynamic system from the previous state, $\boldsymbol{x}_{t-1}$, to the current state, $\boldsymbol{x}_t$. SSMs include an observation equation

$$\boldsymbol{y}_t \mid \boldsymbol{x}_t \sim p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta}), \tag{2}$$

which describes how the observation, $\boldsymbol{y}_t$, is linked to state $\boldsymbol{x}_t$.

Analysing a set of observations $\boldsymbol{y}_{1:T} \in \mathbb{R}^{D_y}$, SMC^2 uses an inner PF to estimate the unobserved states of a system, $\boldsymbol{x}_{1:T} \in \mathbb{R}^{D_x}$, (Section II-A) and an outer SMC sampler to estimate the parameters, $\boldsymbol{\theta} \in \mathbb{R}^D$ (Section II-B).

A. SMC Sampler

At each iteration k, SMC^2 targets

$$\pi(\boldsymbol{\theta}) \propto p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{3}$$

via sequential importance sampling and resampling steps. An unbiased estimate of the marginal likelihood $p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta})$ is provided by the PF. In this section, the posterior distribution over the $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{1:T})$ is defined as $\pi(\boldsymbol{\theta})$.

The joint distribution of all states until $k = K$ is defined as

$$\pi(\boldsymbol{\theta}_{1:K}) = \pi(\boldsymbol{\theta}_K) \prod_{k=2}^{K} L(\boldsymbol{\theta}_{k-1}|\boldsymbol{\theta}_k), \tag{4}$$

where $L(\boldsymbol{\theta}_{k-1}|\boldsymbol{\theta}_k)$ is the L-kernel, which is a user-defined probability distribution. The choice of this distribution can affect the efficiency of the sampler [30].

At $k = 1$, N samples $\forall i = 1, \ldots, N$ are drawn from a prior distribution $q_1(\cdot)$ as follows:

$$\boldsymbol{\theta}_1^i \sim q_1(\cdot), \quad \forall i, \tag{5}$$

and weighted according to

$$\boldsymbol{w}_1^i = \frac{\pi(\boldsymbol{\theta}_1^i)}{q_1(\boldsymbol{\theta}_1^i)}, \quad \forall i. \tag{6}$$

At $k > 1$, subsequent samples are proposed based on samples from the previous iteration via a proposal distribution, $q(\boldsymbol{\theta}_k^i|\boldsymbol{\theta}_{k-1}^i)$, as follows:

$$\boldsymbol{\theta}_k^i \sim q(\cdot|\boldsymbol{\theta}_{k-1}^i). \tag{7}$$

The proposal is commonly chosen to be Gaussian with a mean of $\boldsymbol{\theta}_{k-1}^i$ and a covariance of $\boldsymbol{\Sigma} \in \mathbb{R}^{DxD}$, such that

$$q(\boldsymbol{\theta}_k^i|\boldsymbol{\theta}_{k-1}^i) = \mathcal{N}(\boldsymbol{\theta}_k^i; \boldsymbol{\theta}_{k-1}^i, \Sigma), \quad \forall i. \tag{8}$$

These samples are weighted according to

$$\boldsymbol{w}_k^i = \boldsymbol{w}_{k-1}^i \frac{\pi(\boldsymbol{\theta}_k^i)}{\pi(\boldsymbol{\theta}_{k-1}^i)} \frac{L(\boldsymbol{\theta}_{k-1}^i|\boldsymbol{\theta}_k^i)}{q(\boldsymbol{\theta}_k^i|\boldsymbol{\theta}_{k-1}^i)}, \quad \forall i. \tag{9}$$

A common, suboptimal approach to selecting the L-kernel in (9) is to choose the same distribution as the forwards proposal

$$L(\boldsymbol{\theta}_{k-1}^i|\boldsymbol{\theta}_k^i) = q(\boldsymbol{\theta}_{k-1}^i|\boldsymbol{\theta}_k^i), \quad \forall i, \tag{10}$$

so that, when $q(\cdot|\cdot)$ is symmetric, the weight update in (9) is

$$\boldsymbol{w}_k^i = \boldsymbol{w}_{k-1}^i \frac{\pi(\boldsymbol{\theta}_k^i)}{\pi(\boldsymbol{\theta}_{k-1}^i)}, \quad \forall i. \tag{11}$$

Estimates of the expectations of functions, such as moments, on the distribution are realised by

$$\tilde{\mathbf{f}}_k = \sum_{i=1}^{N} \tilde{\boldsymbol{w}}_k^i \boldsymbol{\theta}_{1:k}^i, \tag{12}$$

where the normalised weights $\tilde{\boldsymbol{w}}_k^i$ are calculated by

$$\tilde{\boldsymbol{w}}_k^i = \frac{\boldsymbol{w}_k^i}{\sum_{i=1}^{N} \boldsymbol{w}_k^i}, \quad \forall i. \tag{13}$$

SMC^2 computes the Effective Sample Size (ESS) as a measure of the efficacy of the sampler

$$N^{\text{eff}} = \frac{1}{\sum_{i=1}^{N} \left(\tilde{\boldsymbol{w}}_k^i\right)^2}. \tag{14}$$

As iterations continue, one weight tends to unity while all other to zero. This is known as particle degeneracy and can be mitigated by resampling. Resampling is undertaken if N^{eff}. In this paper, systematic resampling scheme outlined in [31] is employed. Samples are assigned an unnormalized weight of $\frac{1}{N}$ after resampling.

B. Particle Filter

The PF uses a set of N_x particles to recursively represent any nonlinear, non-Gaussian SSM as $N_x \to \infty$. At every time step t, particles are drawn from a proposal distribution, $q\left(\boldsymbol{x}_{1:t}|\boldsymbol{y}_{1:t}, \boldsymbol{\theta}\right)$ and weighted according to

$$\boldsymbol{w}_t^j = \boldsymbol{w}_{t-1}^j \frac{p\left(\boldsymbol{y}_t|\boldsymbol{x}_t^j, \boldsymbol{\theta}\right) p\left(\boldsymbol{x}_t^j|\boldsymbol{x}_{t-1}^j, \boldsymbol{\theta}\right)}{q\left(\boldsymbol{x}_t^j|\boldsymbol{x}_{t-1}^j, \boldsymbol{y}_t\right)}, \tag{15}$$

In this example, we set the transmission model to equal the proposal which simplifies the weight update in (15) to

$$\mathbf{w}_t^j = \mathbf{w}_{t-1}^j p\left(\boldsymbol{y}_t|\mathbf{x}_t^j, \boldsymbol{\theta}\right). \tag{16}$$

State estimates are made by a weighted sum of functions on the particles

$$\int p\left(\boldsymbol{x}_{1:t}|\boldsymbol{y}_{1:t}, \boldsymbol{\theta}\right) f\left(\boldsymbol{x}_{1:t}\right) d\boldsymbol{x}_{1:t} \approx \sum_{j=1}^{N_x} \tilde{\boldsymbol{w}}_t^j f\left(\boldsymbol{x}_{1:t}^j\right), \tag{17}$$

where the normalized weights are calculated as in (13) (with $\tilde{\boldsymbol{w}}_t^j$ in place of $\tilde{\boldsymbol{w}}_t^i$).

An unbiased estimate of the marginal likelihood can be obtained from an average of the unnormalized weights

$$p\left(\boldsymbol{y}_{1:t}|\boldsymbol{\theta}\right) = \int p\left(\boldsymbol{y}_{1:t}, \boldsymbol{x}_{1:t}|\boldsymbol{\theta}\right) d\boldsymbol{x}_{1:t} \approx \frac{1}{N_x} \sum_{j=1}^{N_x} \boldsymbol{w}_t^j. \tag{18}$$

As with the SMC sampler, the ESS in the PF can be calculated using (14) and resampling employed if N_x^{eff} is less than $N_x/2$. In this paper, multinomial resampling is chosen within the PF. To keep the total unnormalized weight constant (such that the approximation (18) is the same immediately before and after resampling), each newly-resampled sample is assigned an unnormalized weight

$$\frac{1}{N_x} \sum_{j=1}^{N_x} \boldsymbol{w}_t^j, \tag{19}$$

such that the normalized weights after resampling are $1/N_x$.

III. Concept Drift

Concept drift is a broad term that, in a general machine learning context, describes changes over time in the relationship between input data and the target variable. Survey papers have sought to characterize the various types of drift and the timescales on which they occur [16], [17].

A change in the parameters governing the dynamic system in (1) (i.e. $\boldsymbol{\theta}_{t-1} \neq \boldsymbol{\theta}_t$) can be thought of as an example of real concept drift. We define $\boldsymbol{\theta} = \{\boldsymbol{\phi}_t, \boldsymbol{\psi}\}$, with $\boldsymbol{\phi}_t$ denoting time-varying parameters and $\boldsymbol{\psi}$ non-time-varying parameters. Equation (1) can be extended by (20) to include effective dynamics on the parameter space

$$\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1} \sim p(\boldsymbol{\phi}_t \mid \boldsymbol{\phi}_{t-1}, \boldsymbol{\tau}). \tag{20}$$

We parameterise an SSM as hierarchical, with equations (1), (2) and (20), and perform inference on the now static set of parameters $\boldsymbol{\theta} = \{\boldsymbol{\tau}, \boldsymbol{\psi}\}$.

Reference [8] proposes the concept drift models defined in (22)-(25) as ways of evolving a parameter to generate synthetic data. These models are analogous to the aforementioned types of drift timescales and are presented below:

1) Drift-less:
$$\boldsymbol{\phi}_t = l, \tag{21}$$
2) Incremental:
$$\boldsymbol{\phi}_t = l + u\boldsymbol{\tau}t, \tag{22}$$
3) Recurring:
$$\boldsymbol{\phi}_t = \frac{u-l}{2}\mathrm{s}(\frac{2\pi t}{\boldsymbol{\tau}}) + \frac{u+l}{2}, \tag{23}$$
4) Outlier:
$$\boldsymbol{\phi}_t = \begin{cases} l, & t = \boldsymbol{\tau} \\ u, & else \end{cases}, \tag{24}$$
5) Abrupt:
$$\boldsymbol{\phi}_t = \begin{cases} l, & t < \boldsymbol{\tau} \\ u, & else \end{cases}, \tag{25}$$
6) Gradual:
$$\boldsymbol{\phi}_t = \frac{u-l}{2}\mathrm{s}(\frac{2\pi t}{\boldsymbol{\tau}_0 + \boldsymbol{\tau}_1 t}) + \frac{u+l}{2}, \tag{26}$$

where $s(t)$ is the square wave function, u and l are the upper and lower values the parameter may take, respectively and $\boldsymbol{\tau}$ is the drift parameter of interest. A graphical representation of how these concept drift models change a parameter can be seen in Figure 1. The red lines in Figure 1 show the different timescales on which concept drift occurs for $\boldsymbol{\beta}_t$ in the SEIS disease model outlined in Section IV. Figure 1(a) shows a non-time varying parameter (i.e. $\boldsymbol{\beta}_t$ is fixed for the duration of the simulation), Figure 1(b) depicts a sudden shift from one regime to another at a specific time point, while Figure 1(c) illustrates a gradual drift from one regime to another and back, with the parameter spending increasingly longer periods in the upper regime over time. In Figure 1(d), a slow linear variation in the parameter is displayed. Figure 1(e) depicts a sudden jump to another regime followed by an immediate return, while Figure 1(f) illustrates equal time spent in both regimes.

IV. SEIS model

The stochastic Susceptible, Exposed, Infected, Susceptible (SEIS) disease model utilized in this paper is a variant of the SIR model [32]. It simulates the movement of N_{pop} individuals between the Susceptible, Exposed and Infected latent state compartments which take values $\mathbf{S}_t, \mathbf{E}_t, \mathbf{I}_t \in [0..N_{pop}]$

The total count of individuals moving compartments is governed by the parameters $\boldsymbol{\theta} = \{\boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\sigma}\}$. We note that $\boldsymbol{\tau}$ is a static parameter which governs the rate of change of $\boldsymbol{\beta}_t$ i.e. $\phi_t = \beta_t$. The corresponding binomial distributions are

$$n(\mathbf{S} \to \mathbf{E}) \sim \text{Binomial}(\mathbf{S}_{t-1}, 1 - \exp\left(-\frac{\boldsymbol{\beta}_{t-1}\mathbf{I}_{t-1}\mathbf{S}_{t-1}}{N_{pop}}\right)), \tag{27}$$

$$n(\mathbf{E} \to \mathbf{I}) \sim \text{Binomial}(\mathbf{E}_{t-1}, 1 - \exp(-\boldsymbol{\gamma})), \tag{28}$$

$$n(\mathbf{I} \to \mathbf{S}) \sim \text{Binomial}(\mathbf{I}_{t-1}, 1 - \exp(\boldsymbol{\sigma})). \tag{29}$$

The complete discrete, stochastic SEIS model is

$$\mathbf{S}_t = \mathbf{S}_{t-1} - n(\mathbf{S} \to \mathbf{E}) + n(\mathbf{I} \to \mathbf{S}), \tag{30}$$

$$\mathbf{E}_t = \mathbf{E}_{t-1} + n(\mathbf{S} \to \mathbf{E}) - n(\mathbf{E} \to \mathbf{I}), \tag{31}$$

$$\mathbf{I}_t = \mathbf{I}_{t-1} + n(\mathbf{E} \to \mathbf{I}) - n(\mathbf{I} \to \mathbf{S}). \tag{32}$$

The likelihood is

$$\boldsymbol{y}_t | \boldsymbol{x}_t \sim \text{Poisson}\left(\boldsymbol{y}_t; \boldsymbol{I}_t\right). \tag{33}$$

V. Computational Setup

The experiments were conducted using the distributed memory SMC^2 framework outlined in [33], [34]. The code is provided here[1]. The analysis was performed on a distributed memory cluster equipped with two Xeon Gold 6138 CPUs, providing 384GB of memory and 40 cores.

The experimental configuration comprises $T = 150$ observations, $N_x = 4096$ particles in the PF, $N = 256$ samples in the SMC sampler and $K = 50$ iterations. The step-size for the RW proposal was 0.1. The non-time varying parameters are $\boldsymbol{\gamma} = 0.1$ and $\boldsymbol{\sigma} = 0.1$ with priors $q_1(\boldsymbol{\theta}_1^i) = \Gamma(1, 0.1)$. The drift parameters have bounds of $l = 0.1$ and $u = 0.8$ and priors of $q_1(\boldsymbol{\theta}_1^i) = \Gamma(1, 0.1)$ and $q_1(\boldsymbol{\theta}_1^i) = \Gamma(8, 0.1)$. The drift parameters have true values of $\boldsymbol{\tau} = 0.3, [\frac{1}{3}, 0.3], \frac{u-l}{T}, 0.3, \frac{1}{3}$ for abrupt, gradual, incremental, outlier and recurring scenarios, respectively. The prior distributions are $q_1(\boldsymbol{\theta}_1^i) = \Gamma(3, 0.1)$ for all scenarios except incremental drift which is given the prior $q_1(\boldsymbol{\theta}_1^i) = \Gamma(0.1, 0.1)$. The initial distribution of $N_{pop} = 1000$ individuals in each compartment at t_0 was drawn from

$$\text{Multinomial}(\boldsymbol{x}_0; N_{pop}, \boldsymbol{p}), \tag{34}$$

where $\boldsymbol{p} = [0.9, 0.05, 0.05]$.

VI. Results

Over 10 Monte carlo runs, all the parameters when including drift ($\boldsymbol{\theta} = \{\boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\sigma}, u, l\}$) and when drift-less ($\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\sigma}, l\}$) were estimated with SMC^2 for each concept drift scenario. Using the parameter estimates, the PF predicts the states of

[1] https://github.com/j-j-murphy/Explicit-Concept-Drift

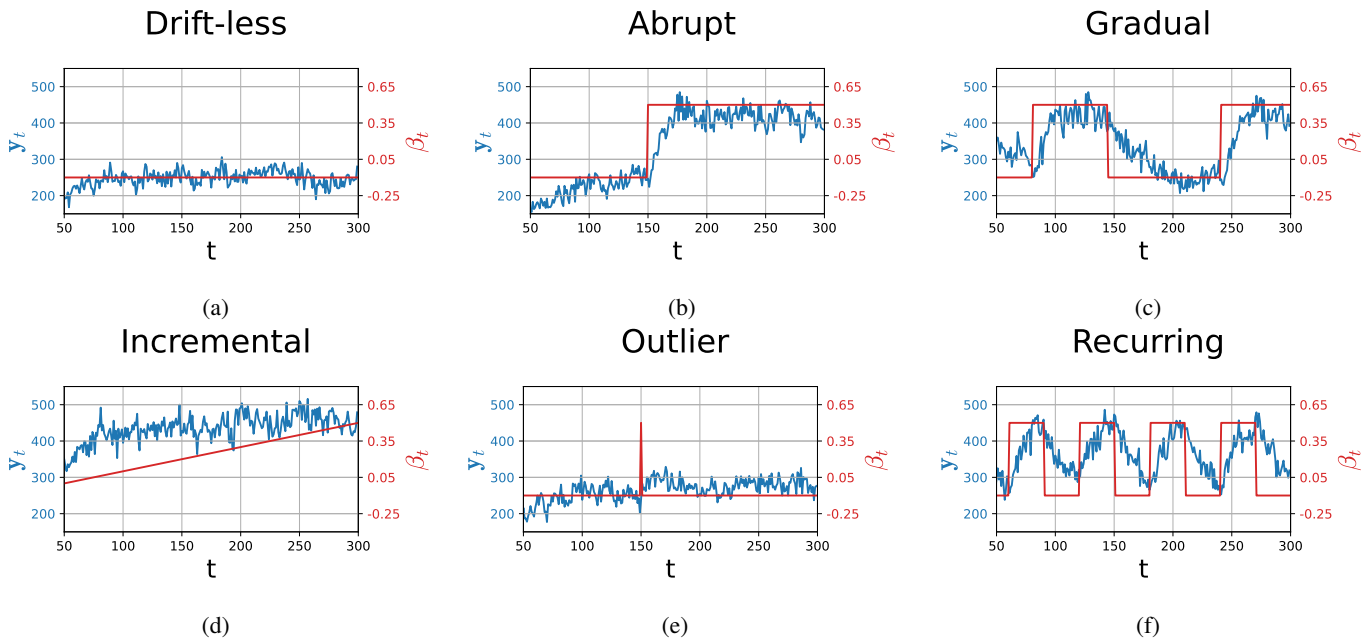


Fig. 1: Infection observations (blue) and corresponding β_t (red) generated for a trajectory of length $T = 300$ with $\gamma = 0.1$, $\sigma = 0.1$, $l = 0.2$ and $u = 0.8$. (a) In the drift-less scenario there is no τ (b) in abrupt $\tau = 0.5$, (c) in gradual $\tau = [0.2, 0.25]$, (d) in incremental $\tau = 0.005$, (e) in outlier $\tau = 0.5$ and (f) in the recurring drift scenario $\tau = 0.2$.

TABLE I: Mean Square Error (MSE) of different drift scenarios.

Scenario	Including Drift	Drift-less
Abrupt	1.655×10^3	5.502×10^4
Gradual	7.051×10^3	7.754×10^4
Incremental	5.791×10^4	3.426×10^4
Outlier	1.150×10^3	1.263×10^3
Recurring	1.922×10^4	6.745×10^4

the disease model, which can be seen in Figure 2. It is evident that when including drift, the PF can more accurately estimate the true states as outlined in Table I when comparing the Mean Square Error (MSE). However, this is not the case with incremental drift which actually performs worse than the drift-less PF. We hypothesis that the is due to the estimate of the susceptible compartment changing too quickly. This is likely due to the random walk proposal which is suitable for the other parameters but not the far finer-scaled τ parameter, giving an estimate an order of magnitude lower of 5.155×10^{-2} against a true value of 4.667×10^{-3}.

VII. Conclusion

In this paper, we explicitly define a hierarchical PF with time-varying parameters governed by concept drift models. We apply these methods to a compartmental disease model, demonstrating that the PF can track unobserved states and estimate parameters using SMC^2.

The scenarios in this paper are simplified, by assuming that the drift type is known apriori and considering only one type of drift at a time. Natural extensions include a model selection method like the likelihood ratio outlined in [35], modeling multiple types of drift simultaneously and exploring drift beyond the binary high-low framework. The SEIS example model is discrete so prevents the the use of improved, gradient-based proposals in the PF and SMC sampler as in [36], [37]. One method to overcome this would be to use HINTS [38]. Further verification of the approach could be found in applications to real world scenarios.

References

[1] A. Doucet, N. De Freitas, and N. Gordon, "An introduction to sequential monte carlo methods," *Sequential Monte Carlo methods in practice*, pp. 3–14, 2001.

[2] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 174–188, 2002.

[3] D. M. Sheinson, J. Niemi, and W. Meiring, "Comparison of the performance of particle filter algorithms applied to tracking of a disease epidemic," *Mathematical biosciences*, vol. 255, pp. 21–32, 2014.

[4] Q.-b. Zhang, P. Wang, and Z.-h. Chen, "An improved particle filter for mobile robot localization based on particle swarm optimization," *Expert Systems with Applications*, vol. 135, pp. 181–193, 2019.

[5] P. M. Djuric, M. Vemula, and M. F. Bugallo, "Target tracking by particle filtering in binary sensor networks," *IEEE Transactions on signal processing*, vol. 56, no. 6, pp. 2229–2238, 2008.

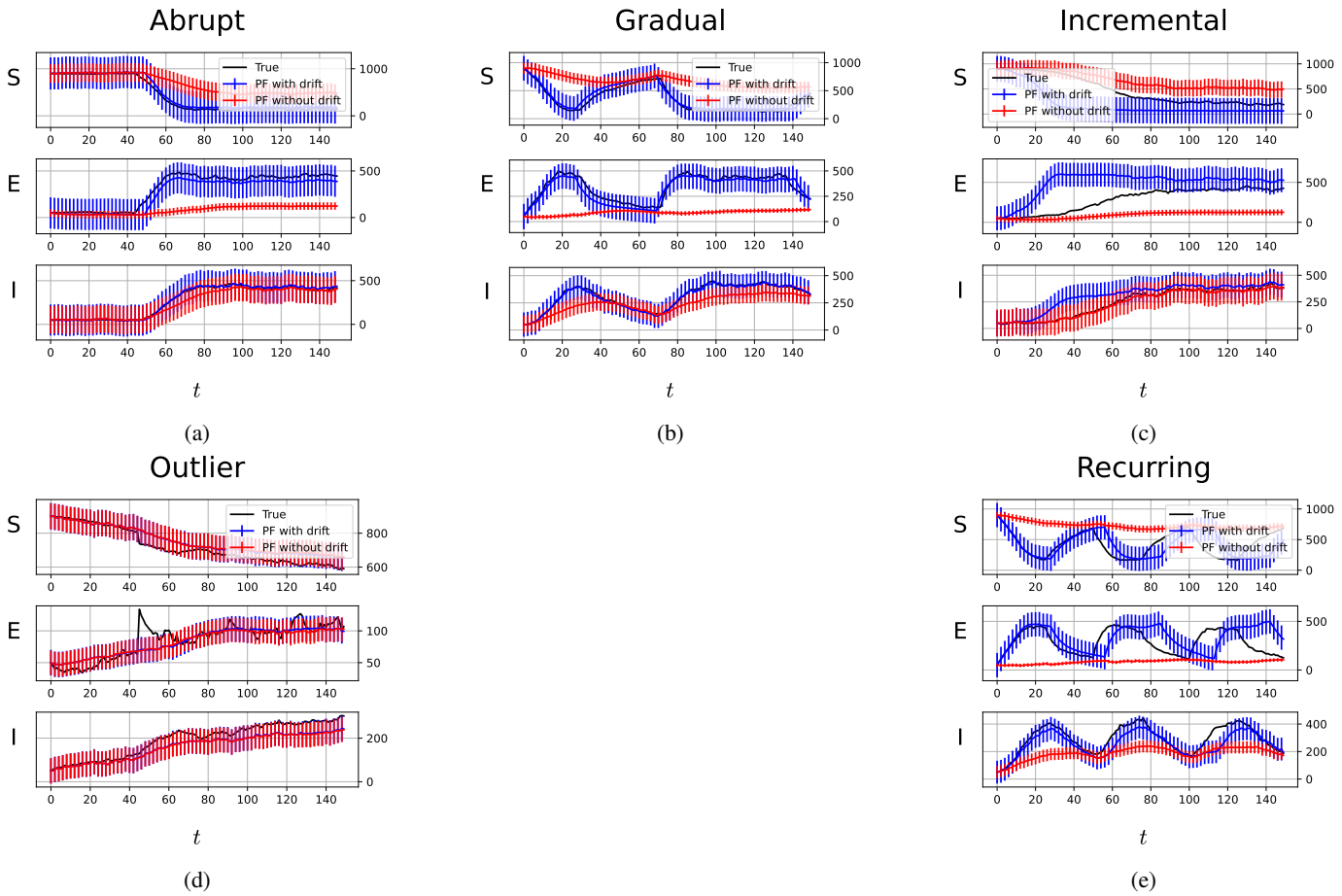


Fig. 2: State estimates of PFs over 5 Monte Carlo runs in (a) abrupt, (b) gradual, (c) incremental, (d) outlier and (e) recurring drift scenarios. The mean estimate is the blue line and the error bars give the standard deviation.

[6] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin, "On Particle Methods for Parameter Estimation in State-Space Models," *Statistical Science*, vol. 30, no. 3, pp. 328–351, 2015. DOI: 10.1214/14-STS511. [Online]. Available: https://doi.org/10.1214/14-STS511.

[7] G. Kitagawa, "A self-organizing state-space model," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1203–1215, 1998, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/2669862 (visited on 05/24/2024).

[8] A. Arnold, "When artificial parameter evolution gets real: Particle filtering for time-varying parameter estimation in deterministic dynamical systems," *Inverse Problems*, vol. 39, no. 1, p. 014 002, 2022.

[9] C. Nemeth, P. Fearnhead, and L. Mihaylova, "Sequential monte carlo methods for state and parameter estimation in abruptly changing environments," *IEEE Transactions on Signal Processing*, vol. 62, no. 5, pp. 1245–1255, Mar. 2014, ISSN: 1941-0476. DOI: 10.1109/tsp.2013.2296278. [Online]. Available: http://dx.doi.org/10.1109/TSP.2013.2296278.

[10] M. Satoh, P. J. Van Leeuwen, and S. Nakano, "Online state and time-varying parameter estimation using the implicit equal-weights particle filter," *Quarterly Journal of the Royal Meteorological Society*, 2024.

[11] T. Matsumoto and K. Yosui, "Adaptation and change detection with a sequential monte carlo scheme," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 3, pp. 592–606, 2007.

[12] D. Crisan and J. Miguez, *Nested particle filters for online parameter estimation in discrete-time state-space markov models*, 2017. arXiv: 1308.1883 [stat.CO].

[13] S. Pérez-Vieites, I. P. Mariño, and J. Míguez, "Probabilistic scheme for joint parameter estimation and state prediction in complex dynamical systems," *Phys. Rev. E*, vol. 98, p. 063 305, 6 Dec. 2018. DOI: 10.1103/PhysRevE.98.063305. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.98.063305.

[14] C. Andrieu, A. Doucet, and R. Holenstein, "Particle markov chain monte carlo methods," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 72, no. 3, pp. 269–342, 2010.

[15] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos, "Smc2: An efficient algorithm for sequential analysis of state space models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 75, no. 3, pp. 397–426, 2013.

[16] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, pp. 1–37, 2014.

[17] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.

[18] S. Liu, S. Xue, J. Wu, *et al.*, "Online active learning for drifting data streams," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 186–200, 2021.

[19] S. Mohamad, M. S. Mouchaweh, and A. Bouchachia, "Active learning for data streams under concept drift and concept evolution," in *STREAMEVOLV at ECML-PKDD*, 2016. [Online]. Available: https : / / api . semanticscholar.org/CorpusID:32173629.

[20] L. Chen, Y. Zhou, and J. Yang, "Object tracking within the framework of concept drift," in *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part III 11*, Springer, 2013, pp. 152–162.

[21] R. Fok, A. An, and X. Wang, "Mining evolving data streams with particle filters," *Computational Intelligence*, vol. 33, no. 2, pp. 147–180, 2017.

[22] D. J. Morris, "Inference on historical ebola outbreaks using hierarchical models: A particle filtering approach," Ph.D. dissertation, 2021.

[23] T. Lazebnik, "Computational applications of extended sir models: A review focused on airborne pandemics," *Ecological Modelling*, vol. 483, p. 110 422, 2023.

[24] A. Gill, J. Koskela, X. Didelot, and R. G. Everitt, "Bayesian inference of reproduction number from epidemiological and genetic data using particle mcmc," *arXiv preprint arXiv:2311.09838*, 2023.

[25] G. Storvik, A. Diz-Lois Palomares, S. Engebretsen, *et al.*, "A sequential monte carlo approach to estimate a time-varying reproduction number in infectious disease models: The covid-19 case," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 186, no. 4, pp. 616–632, 2023.

[26] X. Liu and P. Stechlinski, "Infectious disease models with time-varying parameters and general nonlinear incidence rate," *Applied Mathematical Modelling*, vol. 36, no. 5, pp. 1974–1994, 2012.

[27] P. Birrell, J. Blake, E. Van Leeuwen, N. Gent, and D. De Angelis, "Real-time nowcasting and forecasting of covid-19 dynamics in england: The first wave," *Philosophical Transactions of the Royal Society B*, vol. 376, no. 1829, p. 20 200 279, 2021.

[28] J. Dureau, K. Kalogeropoulos, and M. Baguelin, "Capturing the time-varying drivers of an epidemic using stochastic dynamical systems," *Biostatistics*, vol. 14, no. 3, pp. 541–555, 2013.

[29] A. Doucet, A. M. Johansen, *et al.*, "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of nonlinear filtering*, vol. 12, no. 656-704, p. 3, 2009.

[30] P. L. Green, L. Devlin, R. E. Moore, R. J. Jackson, J. Li, and S. Maskell, "Increasing the efficiency of sequential monte carlo samplers through the use of approximately optimal l-kernels," *Mechanical Systems and Signal Processing*, vol. 162, p. 108 028, 2022.

[31] A. Varsi, S. Maskell, and P. G. Spirakis, "An o (log2 n) fully-balanced resampling algorithm for particle filters on distributed memory architectures," *Algorithms*, vol. 14, no. 12, p. 342, 2021.

[32] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, vol. 115, no. 772, pp. 700–721, 1927.

[33] C. Rosato, A. Varsi, J. Murphy, and S. Maskell, "An o(log2 n) smc2 algorithm on distributed memory with an approx. optimal l-kernel," in *2023 IEEE Symposium Sensor Data Fusion and International Conference on Multisensor Fusion and Integration (SDF-MFI)*, 2023, pp. 1–8. DOI: 10.1109/SDF-MFI59545.2023.10361452.

[34] C. Rosato, J. Murphy, A. Varsi, P. Horridge, and S. Maskell, "Enhanced smc2: Leveraging gradient information from differentiable particle filters within langevin proposals," *arXiv preprint arXiv:2407.17296*, 2024.

[35] N. J. Gordon, S. Maskell, and T. Kirubarajan, "Efficient particle filters for joint tracking and classification," in *Signal and Data Processing of Small Targets 2002*, O. E. Drummond, Ed., International Society for Optics and Photonics, vol. 4728, SPIE, 2002, pp. 439–449. DOI: 10.1117/12.478524. [Online]. Available: https://doi.org/10.1117/12.478524.

[36] C. Rosato, J. Harris, J. Panovska-Griffiths, and S. Maskell, "Inference of stochastic disease transmission models using particle-mcmc and a gradient based proposal," in *2022 25th International Conference on Information Fusion (FUSION)*, IEEE, 2022, pp. 1–8.

[37] A. Varsi, L. Devlin, P. Horridge, and S. Maskell, "A general-purpose fixed-lag no-u-turn sampler for nonlinear non-gaussian state space models," *IEEE Transactions on Aerospace and Electronic Systems*, 2024.

[38] M. Strens, "Efficient hierarchical mcmc for policy search," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 97.

Developing a Multilingual Spontaneous L2 Speech Corpus for Automated Proficiency Assessment

Seunghee Han*, Sunhee Kim† and Minhwa Chung‡
* Seoul National University, Seoul
E-mail: seunghee.han@snu.ac.kr Tel/Fax: +82-10-7517-0101
† Seoul National University, Seoul
E-mail: sunhkim@snu.ac.kr Tel/Fax: +82-10-9402-0342
‡ Seoul National University, Seoul
E-mail: mchung@snu.ac.kr Tel/Fax: +82-10-3726-8496

***Abstract*—Currently, most accessible multilingual L2 spontaneous speech corpora primarily include L2 English from speakers of various L1 backgrounds, with few and often small-scale corpora available for non-English L2s. Annotated assessment data from expert raters is especially scarce. This study addresses this gap by constructing a large-scale dataset of spontaneous L2 speech in seven languages (Chinese, Japanese, English, French, German, Spanish, and Russian) from Korean L1 speakers accompanied by detailed assessments using a carefully designed rubric. The dataset includes extensive metadata analysis and validation processes to ensure the reliability of subjective assessments. To our knowledge, this is the first large-scale, comprehensive corpus of its kind, featuring diverse L2 spontaneous speech from Korean speakers with expert-annotated assessments.**

I. Introduction

Research on the automatic assessment of spontaneous speech has traditionally concentrated on language disorders, disfluency markers, and child language acquisition [1]. In the context of computer-assisted foreign language education, advancements in speech recognition technology have led to significant interest in pronunciation assessment. Meanwhile, natural language processing (NLP) developments have emphasized automated writing assessment. However, a predominant focus on English, driven by the availability of English language datasets, has created a major limitation. This focus presents challenges for expanding research to low-resource languages, where relevant corpora are often scarce and limited in size.

Recently, academia and industry have shown an increased interest in the automatic assessment of spontaneous speech for foreign language education [2], [3], [4], [5], [6], [7]. These assessments require not only advanced speech recognition technologies but also an interdisciplinary approach that integrates NLP, dialogue processing, language education, and linguistics. However, a significant challenge remains the lack of comprehensive training datasets. In response, there has been increasing attention on leveraging large language models (LLMs) and prompt engineering to generate synthetic data, which helps alleviate the shortage of real-world training data. The construction of large-scale, expert-annotated datasets is both time-consuming and expensive, leading to a rising interest in unsupervised learning methods. Nonetheless, ensuring the reliability and validity of automated assessment results is critical, particularly in foreign language education, where measuring instructional effectiveness is crucial.

This work was supported by the National Information Society Agency (NIA) of Korea. The corpus is publicly available on AI Hub.
- English: https://aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=71463
- Chinese and Japanese: https://aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=71464
- German, Spanish, French, Russian: https://aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=71466)

II. Related Work

A. Challenges in Automated Speech Assessment

Open-ended questions in speech and writing assessments typically do not have predefined correct answers, relying instead on rubrics for assessment. This complexity makes it challenging to recruit skilled raters for fair and accurate assessments, particularly in large-scale assessments, due to the significant time and labor involved. Recent NLP research has explored leveraging LLMs for zero-shot and few-shot learning through prompt engineering, even without annotated training data. However, LLMs often lack fine-tuning for specific tasks like speech assessment and may not fully capture the linguistic nuances and evaluation criteria relevant to L2 spontaneous speech. Consequently, responses generated by LLMs can be overly generalized, potentially missing atypical linguistic errors or patterns, challenging the consistency and fairness of such assessments.

Research highlights that LLM-generated responses, such as those using the Chain of Thought (CoT) framework, do not inherently adhere to learned scoring criteria. Instead, these responses are influenced by prompt engineering, raising concerns about consistency and transparency. Studies indicate that CoT can outperform Auto CoT generated by GPT-4 in assessing speaking proficiency. However, CoT may reflect researchers' subjective biases, while Auto CoT might incorporate biases from its training data.

TABLE I: Spontaneous Speech Corpora

Title	L1	L2	Duration	# of Speakers
J-AESOP corpus (Asian English Speech cOrpus Project) [8]	6 different languages including Japanese	English	approximately 1 hour per session/speaker.	154 (only Japanese)
LANGSNAP corpus [9]	11 different languages including English	Spanish, French	-	56
LINDSEI (The Louvain International Database of Spoken English Interlanguage) [10]	6 different languages including Bulgarian	English	130 hours	554
The Trinity Lancaster Corpus [11]	36 different languages including Spanish	English	4.2 M words	2,053
CBFC (The Corpus Bilangue Français Coréen) [12]	French, Korean	Korean, French	A mean value of the speech time - L2 Korean: 73.3 minutes - L2 French: 66 minutes	22
FLACGS (The French Learners Audio Corpus of German Speech) [13]	German, French	German	7 hours	40
ICNALE (International Corpus Network of Asian Learners of English) [14]	11 different languages including English	English	2.1 M words	1,525

Multiple trained human raters are thus suggested to mitigate these biases, offering more reliable and nuanced assessments. This human feedback is crucial for creating high-quality labeled datasets, which are essential for deep learning and machine learning applications and for fine-tuning LLMs. The growing necessity for such datasets underscores the importance of integrating expert human evaluation in developing automated assessment technologies.

B. Corpora for L2 Spontaneous Speech Assessment

Various corpora exist for L2 spontaneous speech assessment, such as ICNALE, Trinity Lancaster Corpus, LINDSEI, and the AESOP corpus, primarily focusing on L2 English from diverse L1 backgrounds. (See Table I.) However, corpora for non-English L2s like Korean, French, and German are limited in scope, with small-scale datasets ranging from 2 to 11.6 hours and involving only a few speakers (20 for Korean, 10 for French, and 40 for German). These datasets often contain responses to a small set of questions focused on a limited range of topics, highlighting the lack of large-scale corpora for non-English L2 spontaneous or semi-spontaneous speech that reflects the diversity of language use in real-world contexts. We thus took into account the representativeness of thematic domains. This study's development of a corpus for Korean L1 speakers across seven L2s (including English, French, and German) represents a pioneering effort in that regard.

III. Corpus Methodology

A. Task Design

The corpus aims to comprehensively evaluate L2 speaking abilities through diverse topics and five types of speaking tasks. (See Table III.)

1) Difficulty Categorization: Using the CEFR standards, vocabulary was categorized into three levels: High (B2 - C2), Medium (A2 - B1), and Low (Pre-A1 - A1), to create a diverse learning experience that mirrors real-life language usage. The distribution of questions by difficulty level was designed to align with the actual proficiency levels of L2 learners in Korea, consisting of approximately 20% advanced, 30% intermediate, and 50% beginner-level questions.

TABLE II: Questionnaire

Language	# of Unique Questions	# of Speakers per Question	Speech Duration per Question (second)
English	1,250	59.52	30.25
Japanese	850	52.12	25.15
Chinese	900	40.94	29.61
German	650	31.46	28.47
Spanish	600	32.00	29.92
French	550	30.36	33.17
Russian	650	26.92	33.02

2) Balanced Distribution: Participants encountered a balanced mix of difficulty levels because each questionnaire included 50 questions in total, with ten questions for each of the five task types, and the difficulty levels were evenly distributed across all questionnaires. Each question was designed to ensure a minimum minute of recorded speech, thus recording at least 25 minutes per participant.

3) Expert-Driven Script Development: Language experts with experience creating questions for certified language tests in Korea (such as TOEIC, OPIC, and FLEX) developed similar speaking tasks for this project. The experts designed the questions by considering the difficulty of vocabulary, syntactic structures, and the diversity of thematic domains. Table II details the number of questionnaires developed for each language, the number of unique questions, and the average number of speakers and speech duration per question.

4) Pilot Testing and Feedback Integration: Before finalizing questionnaires for the speech task, language education experts reviewed a prototype containing 50 questions per language. Trial recordings provided additional insights, leading to the finalization of questionnaires that met academic and pedagogical standards. (See Figure 1.)

B. Topic Selection

Existing speaking tasks have typically focused on specific everyday topics. In this study, the vocabulary and syntactic

TABLE III: Examples for Each Types of Speaking Tasks

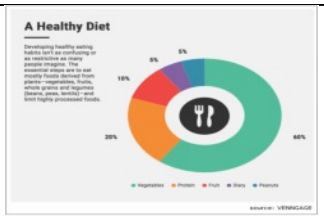
Task Type	Example	
	Direction	Question
Answering Questions	Reply to the following situation.	If neither the mobile phone nor the Internet had been invented, what would be different about our daily lives?
Sentence Creation	Translate each response into English.	A: Shouldn't you be packing for our trip? B: 이미 다 준비했어. 이제 출발할까? (Speak in English.)
Picture Description	Describe the picture on your screen in as much detail as you can.	
Chart Explanation	You will see a chart showing healthy eating habits. Describe and compare what you see in the chart.	
Repetition	Listen and Repeat.	If the following symptoms are detected in a high voltage battery, discharge it as soon as possible by submerging it in salt water.

TABLE IV: Scroing Rubric Development

Major Categories	Subcategories	Grade
Language use	Grammatical accuracy	1-9
	Lexical diversity	
Topic development	Topic relevance	
	Coherence	
Delivery	Pronunciation accuracy	
	Fluency	

structures are assumed to vary by thematic domain. Therefore, we combined the CEFR framework with the thematic matrix from the Korean National Curriculum to categorize topics into personal, public, occupational, and educational domains. In each domain, the questions were divided into the following subcategories: Locations, Institutions, Persons, Objects, Events, Operations, and Texts. We then provided diverse situational keywords to design tasks that encourage speech in contexts similar to real-life situations.

C. Scoring Rubric Development

The evaluation schema integrates standards from TOEFL and CEFR with adaptations for the Korean educational context. Major assessment criteria include language use, topic development, and delivery, with subcategories such as grammatical accuracy, lexical diversity, topic relevance, coherence, pronunciation accuracy, and fluency. (See Table IV.)

IV. Data Collection

A. Participant Recruitment

Recruiting thousands of learners for seven different foreign languages across Korea was challenging. We mainly targeted students specializing in these languages at foreign language universities. For English, 50% of the participants were middle and high school students, reflecting the standard educational curriculum. Our sample comprised both students with and without official language proficiency scores. Those with scores were divided into advanced, intermediate, and beginner levels according to the CEFR. Participants without scores were classified as Beginner for first and second-year university students, Intermediate for third and fourth-year students, and Advanced for those with over six years of continuous study or experience living abroad, aligned with the foreign language learning objectives set for each academic year in higher education. Gender imbalance among foreign language learners is a well-documented global phenomenon, and Korea was no exception. To minimize gender bias in model training, we aimed to achieve an overlap rate of over 50%. The formula is applied as follows:

$$Overlap \quad rate = \frac{MIN(goal \quad value, \quad actual \quad value)}{MAX(goal \quad value, \quad actual \quad value)}$$

B. Recording Sessions

We optimized recording protocols for Chrome on personal computers, instructing participants to use quiet environments and earphones to minimize background noise. To prevent truncation at the start of responses, participants were asked to wait 0.5 to 1.0 seconds after starting the recording before speaking, ensuring full response capture.

C. Data Overview

Our corpus includes about 1,800 hours of L2 speech data, with 300 hours each for English, Chinese, and Japanese and 150 hours each for French, German, Spanish, and Russian. An

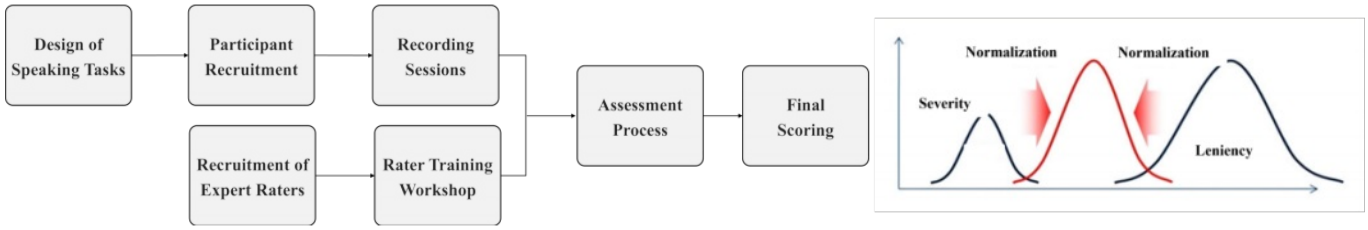


Fig. 1: Pilot Testing and Feedback Integration

TABLE V: Assessment Pannel

Language	# of Assessors	# of Groups of Three	Questions/Group	Inter-rater Agreement
English	43	29	2,564	0.7315
Japanese	14	12	3,690	0.5724
Chinese	25	17	2,166	0.5855
German	18	26	790	0.5304
Spanish	16	10	1,918	0.5972
French	12	9	1,860	0.6575
Russian	18	11	1,589	0.5511

additional 300 hours of English speech from middle and high school students were included, totaling 600 hours for English. All data were curated with consistent standards to maintain comparability. (See Table VI.)

V. Data Processing and Annotation

A. Data Preprocessing

The audio captured on the authoring platform was processed to remove background noise. A 0.5-second silence was then added at the beginning and end of the speech segments before the labeling process began.

B. Annotation Process

Collected responses were paired with corresponding questions and evaluated by expert raters using a detailed rubric. Each panel of three raters assessed the responses based on lexical diversity, grammatical accuracy, topic relevance, coherence, communicative effectiveness, and delivery. The raters assigned scores from 1 to 5 for each criterion, which were then averaged to ensure objectivity. These total scores were subsequently converted to a 1-9 scale. Multiple pilot tests were conducted before the formal evaluation to ensure further consistency in subjective evaluations aligning the raters' judgments and reducing discrepancies.

C. Quality Assurance

Traditionally, FACETS analysis has played a key role in foreign language pedagogy research, helping to identify assessors' biases, such as severity or leniency, and the factors affecting their evaluations. This method typically involves a small group of experts simultaneously scoring a limited sample set. However, our corpus required extensive qualitative assessments, making simultaneous evaluation impractical. We utilized Krippendorff's alpha, a statistical measure of coder agreement, to ensure reliability. Our study, involving 114 groups of three assessors, each evaluating an average of 2,082 questions, yielded a Krippendorff's alpha of approximately 0.629 across the entire dataset. While this indicates moderate agreement—falling below the typical threshold of 0.667 for strong agreement—it remains significant in the context of large-scale subjective evaluations, where perfect consensus is challenging to attain. This is particularly true given the number of assessor groups, the volume of utterances each group assessed, and the thematic diversity of the items evaluated. In such scenarios, a score of this magnitude suggests that the evaluation process captured a reasonable level of consensus despite the inherent challenges of diverse and subjective data. (See Table V.)

VI. Conclusions

This study introduces a large-scale multilingual L2 speech corpus for assessing Korean L1 learners' speaking proficiency in seven languages. The dataset, featuring expert-annotated assessments, addresses the scarcity of corpora for non-English L2s, making it a crucial resource for advancing both academic research and industry applications.

Potential applications include enhancing speech recognition technologies, intelligent language tutoring systems, and conversational AI, particularly for low-resource languages. The corpus enables more accurate automated evaluations of L2 proficiency, supporting innovations in language education and AI-driven interactions.

Future research will focus on comparative analysis across languages and within proficiency levels, examining how different linguistic features and fluency markers affect assessments. These insights will refine models to ensure accuracy and fairness in automated evaluations across diverse learner profiles.

In conclusion, this corpus is a critical tool for developing scalable solutions in speech recognition and automated language assessment. Its public availability is expected to promote further research, drive innovation in foreign language education technologies, and enhance automated assessments for multilingual populations.

Ethical Considerations

A specialized survey company handled participant recruitment to ensure professionalism and ethical standards. Participants were recruited only after providing informed consent through the "Copyright Usage Agreement for Data" and the

TABLE VI: Demographics and Speech Data Overview Across Languages

			English	Japanese	Chineses	German	Spanish	French	Russian
Demographics	# of Speaker		1,484	885	737	408	382	334	350
	Age	10s	57.21%	1.58%	8.01%	1.23%	1.05%	0.30%	0.29%
		20s	30.32%	71.75%	71.64%	85.29%	87.96%	88.62%	84.29%
		30s	10.11%	19.77%	14.52%	11.52%	8.64%	9.88%	14.00%
		+40s	2.36%	6.89%	5.83%	1.96%	2.36%	1.20%	1.43%
	Gender	M	35.85%	20.23%	33.65%	20.34%	27.23%	19.76%	46.00%
		F	64.15%	79.77%	66.35%	79.66%	72.77%	80.24%	54.00%
		Overlap Rate	74.82%	51.57%	71.33%	51.72%	61.58%	50.92%	92.30%
	Proficiency	High	13.81%	6.55%	8.82%	8.33%	9.95%	16.17%	14.57%
		Middle	35.92%	35.48%	24.02%	24.75%	27.75%	39.52%	28.29%
		Low	50.27%	57.67%	67.16%	66.91%	62.30%	44.31%	57.14%
Speech Characteristics	Duration (h)		600	300	300	150	150	150	150
	# of Samples		74,355	44,280	36,829	20,549	19,179	16,736	17,479
	# of Sentences		230,079	145,350	124,753	71,682	60,833	65,177	59,751
	# of Tokens/Characters		3,945,835	4,351,262	2,886,144	882,993	941,304	933,227	711,293
	Avg. Tokens/Characters		17.15	29.94	23.13	12.32	15.47	14.32	11.90
	Duration/Speaker (h)		0.42	0.35	0.41	0.4	0.42	0.46	0.46
Recording Standards	Format (WAV), Bit rate (256kb/s), Channels (1 channel), Sampling rate (16kHz), Bit depth (16 bits)								

"Consent Form for Personal Information Collection, Use, and Third-Party Provision." Formal consent from legal guardians was secured in cooperation with schools for minors involved in English speech recordings. Participants received compensation for adhering to ethical standards for research involving human subjects.

REFERENCES

[1] K. Horii, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, "Language modeling for spontaneous speech recognition based on disfluency labeling and generation of disfluent text," *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1851–1856, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265335598.

[2] B. Jadhav, A. Sawant, A. Shah, P. Vemula, A. Waikar, and S. Yadav, "A comprehensive study and implementation of the mock interview simulator with ai and pose-based interaction," *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, pp. 01–05, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269953413.

[3] R. Southwell, W. Ward, V. A. Trinh, *et al.*, "Automatic speech recognition tuned for child speech in the classroom," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12 291–12 295, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268525709.

[4] S. Bannò and M. Matassoni, "Proficiency assessment of l2 spoken english using wav2vec 2.0," *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1088–1095, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253098949.

[5] R. Al-Ghezi, Y. Getman, E. Voskoboinik, M. Singh, and M. Kurimo, "Automatic rating of spontaneous speech for low-resource languages," *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 339–345, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256353629.

[6] S. Park, J. Chun, M. H. Kim, H. Lee, S. H. Lee, and S. Kim, "Designing a korean french-learners' speech corpus (kflsc) for spoken language assessment," *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 1–6, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:255188356.

[7] P. Bamdev, M. S. Grover, Y. K. Singla, P. Vafaee, M. Hama, and R. R. Shah, "Automated speech scoring system under the lens: Evaluating and interpreting the linguistic cues for language proficiency," *arXiv preprint arXiv:2111.15156*, 2021.

[8] K. Yazawa, T. Konishi, R. Pérez-Ramón, and M. Kondo, "More foreign-accented but more comprehensible: Attrition and amelioration of l1 speech in proficient l2 learners," *F1000Research*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271652548.

[9] A. Huensch, "Effects of speaking task and proficiency on the midclause pausing characteristics of l1 and l2 speech from the same speakers," *Studies in Second Language Acquisition*, vol. 45, pp. 1031–1055, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260045257.

[10] I. Yakut and F. Yuvayapan, "Forms and functions of self-repetitions in spoken discourse: A corpus linguistics analysis of l1 and l2 english," *Topics in Linguistics*, vol. 23, pp. 83–96, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249960548.

[11] D. Gablasova, V. Brezina, and T. McEnery, "The trinity lancaster corpus: Development, description and application," *International Journal of Learner Corpus Research*, vol. 5, no. 2, pp. 126–158, 2019.

[12] H. Yoo and I. Kim, "Cbfc: A parallel l2 speech corpus for korean and french learners," in *International Conference on Language Resources and Evaluation*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:21702338.

[13] J. Wottawa and M. Adda-Decker, "French learners audio corpus of german speech (flacgs)," in *International Conference on Language Resources and Evaluation*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:26867813.

[14] T.-H. Lo, F.-A. Chao, T.-I. Wu, Y.-T. Sung, and B. Chen, "An effective automated speaking assessment approach to mitigating data scarcity and imbalanced distribution," *ArXiv*, vol. abs/2404.07575, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269043380.

Data-Driven Physics-Informed Neural Networks for Sound Field Estimation in Rooms of Arbitrary Size

Gen Sato* and Yusuke Ikeda*
* Tokyo Denki University, Japan
E-mail: 24udc03@ms.dendai.ac.jp, yusuke.ikeda@mail.dendai.ac.jp

***Abstract*—In recent years, many methods for simulating sound fields using deep learning have been proposed. In particular, neural networks using physical knowledge, known as Physics-Informed Neural Networks (PINNs), have been proposed to solve various acoustic problems with higher accuracy. Many of the proposed PINN-based methods have achieved the modeling of a single sound field. In this study, we propose data-driven PINN for estimating simulated sound fields in rooms of arbitrary size. By using the coordinates and room size information as the input of PINNs, the proposed method demonstrated a significant improvement in estimation accuracy by approximately 5.4 dB compared to the simple data-driven model.**

I. Introduction

Sound field simulation has applications in various fields, including spatial acoustics, architectural acoustics [1], virtual reality, and sound rendering in mixed reality [2].

Sound field simulation methods can be divided into two main categories: simulations based on geometric acoustics [3] and simulations based on the wave equation [4]–[6]. Ray tracing is a well-established simulation method based on geometric acoustics [7], where sound propagation is modeled as rays, enabling high-speed calculations of sound fields. In contrast, numerical simulations that approximate the wave equation provide highly accurate simulations of sound fields, such as physical wave phenomena such as diffraction and scattering. Many methods have been proposed, including finite-difference time-domain method [4], finite-element method (FEM) [5], and boundary-element method (BEM) [6]. Although these methods achieve high accuracy, they are known to require significant computation time and computational resources.

Consequently, many studies have been conducted to improve processing speed and reduce the computational resources required. For example, Yoshida *et al.* reduced the computational time by developing a method to minimize the dispersion errors in the time-domain FEM [8]. Wilkes *et al.* proposed a method for solving broadband sound fields in parallel using BEM with fast multipole solutions [9].

In recent years, data-driven deep learning methods for sound field simulation have been proposed. These methods learn the properties of partial differential equations (PDEs) from a large number of sound fields computed using conventional numerical simulations based on the wave equation. A trained model that approximates the PDE enables fast simulation of the sound field. Fan *et al.* proposed a method for estimating the sound pressures of a scattered sound field from the shape of the scattering object using convolutional neural networks, which are commonly used in tasks such as semantic segmentation. This method was demonstrated to be 100 times faster than conventional numerical simulations [10]. Borrel-Jensen *et al.* have proposed a room sound-field simulation using a deep operator network [11]. In this method, the user can freely select the location of the sound source and the listener. In addition, it can estimate sound pressure within 100 ms.

Furthermore, a deep learning method called a physics-informed neural networks (PINNs), which learns based on the governing physical equation, has been proposed [12]–[14]. In particular, many PINNs studies have been conducted in acoustics using the wave equation [15]–[19]. Karakonstantis *et al.* applied PINNs to sound field reconstruction and achieved more accurate sound field interpolation compared to compressed sensing [15]. Additionally, Chen *et al.* employed the PINNs approach to estimate the acoustic scattering from measured signals using a limited number of microphones [18]. Previous studies have shown that the PINN approach is effective in improving sound field estimation accuracy. However, the model must be modified when the arrangement or conditions vary, such as when the room shape changes.

In this study, we propose a sound field simulation using data-driven deep learning with PINNs. The input of PINN is a single coordinate and a vector indicating the room shape. The output is the sound pressure at the input coordinate for the given room shape. In particular, this study focuses on estimating the scattered sound field in rooms of arbitrary size. For the training dataset, sound fields of various room sizes are used. In the simulation experiments, we evaluate the estimation accuracy of the proposed method by comparing it to a model trained using only data.

The remainder of this paper is organized as follows. Section II introduces the problem statement and the proposed method. Section III presents the simulation experiments for evaluation in a two-dimensional sound field. Finally, we conclude the paper in Section IV.

II. Method

A. Problem Statement

We consider the two-dimensional sound fields with a Dirac impulse emitted from a sound source at position $\mathbf{x}_{\mathrm{src}} = (x_{\mathrm{src}}, y_{\mathrm{src}})$. Our study focuses on estimating complex sound

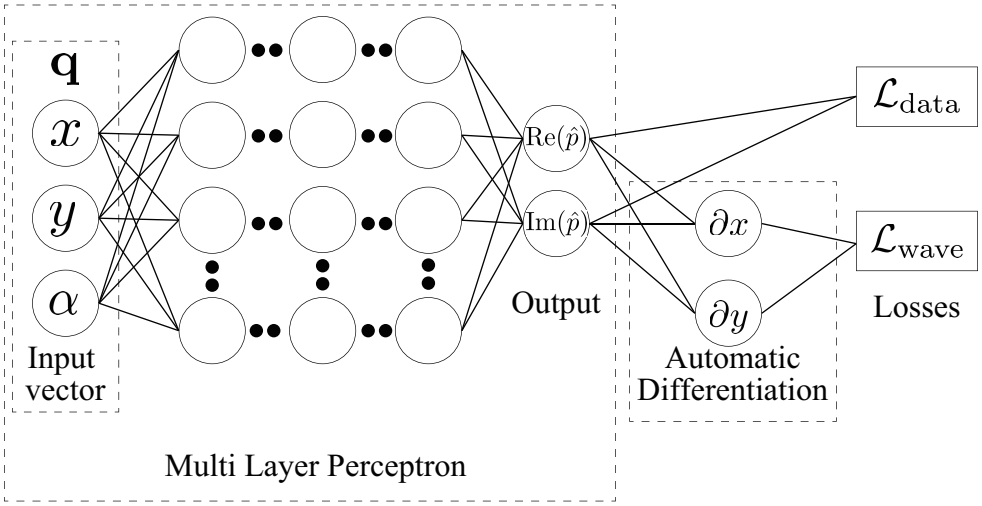


Fig. 1. Workflow of the proposed method. The real and imaginary parts of the complex sound pressure at the position (x, y) are estimated from the input vector $\mathbf{q}$ (x, y: two-dimensional plane coordinates, α: information on the room shape). $\mathcal{L}_{\mathrm{data}}$ calculates the mean squared error using the estimated complex sound pressure and training data. $\mathcal{L}_{\mathrm{wave}}$ calculates the Helmholtz equation using the Laplacian calculated by automatic differentiation and the estimated complex sound pressure. By using $\mathcal{L}_{\mathrm{data}}$ and $\mathcal{L}_{\mathrm{wave}}$, the backpropagation is conducted.

pressure $p_{\mathrm{all}}(\mathbf{x})$ at any position $\mathbf{x}$ in the sound field Ω of arbitrary room shape.

The sound pressure $p_{\mathrm{all}}(\mathbf{x})|\mathbf{x} \in \Omega$ at any position $\mathbf{x}$ in the room sound field Ω represents

$$p_{\mathrm{all}}(\mathbf{x}) = p_{\mathrm{dir}}(\mathbf{x}) + p_{\mathrm{ref}}(\mathbf{x}), \tag{1}$$

where, p_{dir} represents the complex sound pressure of the direct sound and p_{ref} represents the complex sound pressure of the reflected sound. That is, p_{all} can be represented as the sum of p_{dir} and p_{ref}.

Furthermore, since the reflected sound field p_{ref} is homogeneous, the following Helmholtz equation holds.

$$(\nabla^2 + k^2)p_{\mathrm{ref}}(\mathbf{x}) = 0, \tag{2}$$

where, $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the Laplacian operator, k is the wave number.

Here, when the sound source is assumed to be a line source in the two-dimensional sound field, the direct sound p_{dir} can be analytically represented by the Green function.In contrast, the reflected sound field p_{ref} is more difficult to obtain than p_{dir} because the sound pressure depends on the boundary conditions, i.e., the size and shape of the room and the sound absorption coefficients. Hence, in this study, we only estimate the reflected sound fields. To simplify the notation, p_{ref} will be denoted as p in the following sections.

B. Workflow and Input Vector

In this study, based on PINNs, we proposed an estimation method for the complex sound pressure $p(\mathbf{x})$ of the reflected sound field at coordinate $\mathbf{x}$ in an arbitrary-sized room. Fig. 1 shows the workflow in the proposed method.

In the proposed method, a multilayer perceptron (MLP) estimates the real $\mathrm{Re}(\hat{p}(\mathbf{q}))$ and imaginary $\mathrm{Im}(\hat{p}(\mathbf{q}))$ parts of the complex sound pressure $\hat{p}(\mathbf{q})$ corresponding to the input vector $\mathbf{q}$. PINN employs a PDE for the loss function to learn with the physical characteristics of the output. In this study, since the sound pressure is estimated in a frequency-domain sound field, the Helmholtz equation in (2) is used as the PDE. Therefore, MLP is learned using $\mathcal{L}_{\mathrm{data}}$, which evaluates the error of the estimated sound pressure, and $\mathcal{L}_{\mathrm{wave}}$, which evaluates the Helmholtz equation using the estimated sound pressure. This workflow is repeated to create a PINN model that estimates the complex sound pressure at an arbitrary point for an arbitrary-sized room.

PINNs are generally well known for estimating physical quantities such as sound pressure corresponding to input coordinates. Since the model learns the mapping between the coordinates and the data, a different model is required for other conditions, such as different room shapes. In this study, we aim to construct a single model that can handle different conditions by introducing a condition vector as the additional input. To simplify the problem, the experiment was conducted with the two-dimensional sound fields in a square room. Accordingly, the side length is employed as the value of α for the condition vector. This allows PINN to estimate the sound field for different room sizes. Furthermore, if the condition vector contains more information, it can be expected to handle a wider variety of room shapes.

C. PINN Architecture

In this study, the MLP for estimating the complex sound pressure $\hat{p}(\mathbf{q})$ at any points in any reflected sound fields can be represented as follows using the function $f(\cdot)$.

$$\hat{p}(\mathbf{q}) = f(\mathbf{w}, \mathbf{q}) = (\phi_N \circ \phi_{N-1} \circ \cdots \phi_n \circ \cdots \phi_1)(\mathbf{q}), \tag{3}$$

where, $\mathbf{q}$ represents the input vector and $\mathbf{w}$ represents the learnable weights. ϕ is a function in the N-th layer. In other words, MLP is represented by a composition function of ϕ.

The function ϕ in the n-th layer is

$$\phi_n(\mathbf{q}_n) = \sin(\omega_0(\mathbf{q}_n\mathbf{w}_n + \mathbf{b}_n)), \tag{4}$$

where, the input vector in the n-th layer is represented by $\mathbf{q}_n$, while the learnable weights and biases in this layer are represented by $\mathbf{w}_n$ and $\mathbf{b}_n$, respectively. ω_0 is a hyperparameter when a sine function is used as the activation function [20]. Since this network can calculate the second-order derivative through automatic differentiation, the Helmholtz equation is employed for the PINN loss function. Compared to other activation functions, such as hyperbolic tangent and rectified linear units, the sine function is known to be more accurate in high-order partial derivatives. Thus, using a sine function for the activation function, PINN improves the estimation accuracy [15]–[17].

The loss function $\mathcal{L}$ for each α is given by

$$\mathcal{L} = \mathcal{L}_{\mathrm{data}} + \lambda\mathcal{L}_{\mathrm{wave}}, \tag{5}$$

$$\mathcal{L}_{\mathrm{data}} = \frac{1}{M_{\mathrm{data}}} \sum_{m\in\mathcal{M}_{\mathrm{data}}} |\hat{p}_m - p_m|^2, \tag{6}$$

$$\mathcal{L}_{\mathrm{data}} = \frac{1}{M_{\mathrm{data}}} \sum_{m\in\mathcal{M}_{\mathrm{data}}} |\mathrm{Re}(\hat{p}_m) - \mathrm{Re}(p_m)|^2 + |\mathrm{Im}(\hat{p}_m) - \mathrm{Im}(p_m)|^2, \tag{7}$$

$$\mathcal{L}_{\mathrm{wave}} = \frac{1}{M_{\mathrm{data}}} \sum_{m\in\mathcal{M}_{\mathrm{data}}} |(\nabla^2 + k^2)\hat{p}_m| + \frac{1}{M_{\mathrm{wave}}} \sum_{m\in\mathcal{M}_{\mathrm{wave}}} |(\nabla^2 + k^2)\hat{p}_m|, \tag{8}$$

where λ is the balance parameter. $\mathcal{M}_{\mathrm{data}}$ is a set of sound pressure indices of training data for each α, $\mathcal{M}_{\mathrm{wave}}$ is a set of sound pressure indices that only calculates $\mathcal{L}_{\mathrm{wave}}$ for each α. p_m and $\hat{p}_m$ denote m-th ground truth and estimated complex sound pressure, respectively. M_{data} and M_{wave} are the number of elements in each set. $\mathcal{L}_{\mathrm{data}}$ calculates the mean squared error using training data and estimated data of sound pressure ($\hat{p}_m \in \mathcal{M}_{\mathrm{data}}$). $\mathcal{L}_{\mathrm{wave}}$ is a loss function for learning so that the sound pressure satisfies the laws of physics, based on the homogeneous Helmholtz equation in (2).

III. Experiments

In this experiment, we compare the estimation accuracies with the proposed method and simple MLP only using the loss function in (7).

A. Experiment Condition

The pair data of input vector $\mathbf{q}$ and complex sounds pressure $p(\mathbf{q})$ were used as the training dataset. For the training data, the reflected sound fields at 500 Hz were calculated by the BEM [21]. Table I shows the simulation conditions of the BEM. The spacing between boundary elements was 0.1 m (approximately $1/6$ of the wavelength) or less. The experiment was conducted with the sound fields in square rooms. The length of one side of the square room was 2–4.5 m. Note that the range utilized for the purposes of learning and estimation was limited to a 2 m square region centered on the origin, which is part of the sound field of the square room. The information α is the side length of the room. A total of 251 sound fields for training and evaluation were simulated by modifying α in increments of 0.01 m. Three different datasets were constructed, with α increments d being set to 0.03 m, 0.04 m, and 0.05 m, respectively. Experiments were then conducted to compare these data sets, which contained 84, 63, and 51 data numbers, respectively. The test dataset for evaluation is other than the data used for training.

TABLE I
Condition of BEM

Sound speed [m/s]	340
Frequency [Hz]	500
Sound source coordinates	$(0, 0)$
Sound absorption coefficient	0.5

For datasets, the input coordinates were chosen randomly, and the number $M_{\mathrm{data}}(\alpha)$ was $(\alpha/0.2)^2$ such that the density was 1/0.04 m^2 per the room side length α. In other words, each sound field was discretized by about $1/3$ of a wavelength. To compare methods of determining input coordinates, the other test dataset was created with a grid of 0.1 m intervals for the input coordinates. The number of coordinates M_{wave} was the same as M_{data}. The coordinates for calculating the loss $\mathcal{L}_{\mathrm{wave}}$ were randomly determined for each iteration. The input coordinates, complex sound pressure, and α were normalized within the range of -1 to 1.

The network was a five-layer MLP comprising 256 neurons in each layer. The optimization method employed Adam, with a learning rate of 1.0×10^{-5}. The hyperparameter ω_0 was set to 30, and the balance parameter λ was set to 1.0×10^{-3}. The batch processing was conducted for each room size α. The loss for validation was calculated for every 10 epochs on the test dataset with random coordinates. The model with the smallest loss value within 200 epochs was used for evaluation.

The accuracy of the estimated sound field was evaluated using the following signal-to-noise ratio (SNR).

$$\mathrm{SNR} = 10\log_{10}\frac{|p(\mathbf{x})|^2}{|p(\mathbf{x}) - \hat{p}(\mathbf{x})|^2} \tag{9}$$

where, $p(\mathbf{x})$ is ground truth at position $\mathbf{x}$. $\hat{p}(\mathbf{x})$ is the estimated sound pressure.

B. Results

Models were created using three training datasets with different increments d of α. Table II and Table III show the mean SNR for the test dataset.

From Table. II, the proposed method improved the estimation accuracy with all three models compared to MLP. When d = 0.04 and 0.05 m, the proposed method improved the estimation accuracy by 5.4 dB. When d = 0.03 m, the difference in estimation accuracy between MLP and PINN was

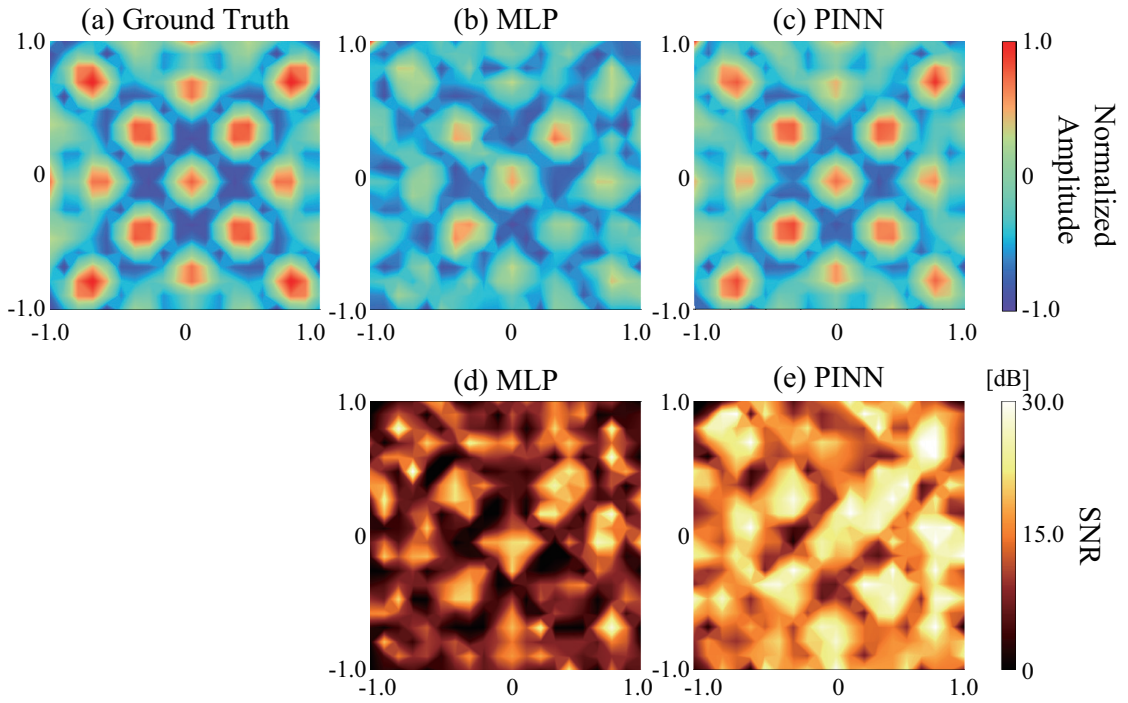


Fig. 2. Comparison of estimated sound fields at $\alpha = 2.350$ m with a model trained on a dataset with spacing $d = 0.04$ m. The estimation region is $[-1, 1]$ m. (a) Ground truth calculated by BEM, (b)–(c) and Estimated sound fields by MLP and PINN (proposed method), (d) and (e) SNR distributions by MLP and PINN.

TABLE II
MEAN SNR OF THE SOUND FIELDS ESTIMATED BY THE MODEL FOR A RANDOM COORDINATE DATASET

Increment d	0.03 m	0.04 m	0.05 m
MLP	9.6 dB	7.4 dB	5.8 dB
PINN(Proposal)	13.5 dB	12.8 dB	11.2 dB

TABLE III
MEAN SNR OF THE SOUND FIELDS ESTIMATED BY THE MODEL FOR A GRID COORDINATE DATASET

Increment d	0.03 m	0.04 m	0.05 m
MLP	9.0 dB	7.1 dB	5.4 dB
PINN(Proposal)	12.9 dB	12.2 dB	10.6 dB

3.9 dB smaller compared to d = 0.04 m and 0.05 m. Thus, the difference in estimation accuracy between MLP and PINN tended to become smaller as d became smaller, i.e., as the number of training data increased.

From Table. III, the estimation accuracy of the sound field in grid coordinates using the proposed method was slightly lower by 0.6 dB than the estimation accuracy in random coordinates in Table. II. However, the proposed method improved the accuracy by 3.9 dB, 5.1 dB, and 5.2 dB at d = 0.03, 0.04, and 0.05, respectively.

Fig. 2 shows the estimated sound field at $\alpha = 2.350$ m utilizing a model trained by the dataset with $d = 0.04$ m. From Figs. 2(a)(b), the estimated sound field by MLP, data-driven models, was significantly degraded over the estimation region. From Figs. 2(a)(c), the sound field estimated by PINN closely approximates the ground truth, compared to MLP. From Figs. 2(d)(e), the proposed method significantly improved the SNR distribution compared to MLP. In this condition, the mean SNRs of MLP and PINN were 7.4 dB and 14.7 dB, respectively. Thus, the PINN improved mean SNR by 7.3 dB than MLP.

IV. CONCLUSIONS

In this study, we proposed an estimation method for the sound fields in rooms of arbitrary size based on PINNs and data-driven MLPs. The experiments demonstrated that the proposed method can estimate the sound field in a room of a size not included in the training with higher accuracy than a model trained on data alone. We also found that PINNs improves estimation accuracy compared to MLPs when using smaller amounts of training data.

However, the proposed method was evaluated under fixed experimental conditions, such as the location of the sound source, the estimated region and sound absorption coefficients. To facilitate practical application, these conditions are required to be relaxed. In such cases, more sound field data would be required. Therefore, the future research will focus on developing methods to reduce the amount of data required by improving the model.

ACKNOWLEDGMENT

This work was partially supported by Tokyo Denki University Special Academy Research Fund 224S40.

References

[1] T. Yoshida, T. Okuzono, and K. Sakagami, "A parallel dissipation-free and dispersion-optimized explicit time-domain fem for large-scale room acoustics simulation," *Buildings*, vol. 12, no. 2, 2022.

[2] N. Raghuvanshi and J. Snyder, "Parametric directional coding for precomputed sound propagation," *ACM Trans. Graph.*, vol. 37, no. 4, Jul. 2018.

[3] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, Aug. 2015.

[4] T. Tsuchiya, "Recent techniques on sound field simulation," *Japanese Journal of Applied Physics*, vol. 61, no. SG, SG0801, 2022.

[5] L. L. Thompson, "A review of finite-element methods for time-harmonic acoustics," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1315–1330, Mar. 2006.

[6] S. Kirkup, "The boundary element method in acoustics: A survey," *Applied Sciences*, vol. 9, no. 8, 2019.

[7] A. Krokstad, S. Strom, and S. Sørsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, vol. 8, no. 1, pp. 118–125, 1968.

[8] T. Yoshida, T. Okuzono, and K. Sakagami, "Dissipation-free and dispersion-optimized explicit time-domain finite element method for room acoustic modeling," *Acoustical Science and Technology*, vol. 42, no. 5, pp. 270–281, 2021.

[9] D. R. Wilkes, A. J. Duncan, and S. Marburg, "A parallel and broadband helmholtz fmbem model for large-scale target strength modeling," *Journal of Theoretical and Computational Acoustics*, vol. 28, no. 03, p. 2 050 001, 2020.

[10] Z. Fan, V. Vineet, H. Gamper, and N. Raghuvanshi, "Fast acoustic scattering using convolutional neural networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 171–175.

[11] N. Borrel-Jensen, S. Goswami, A. P. Engsig-Karup, G. E. Karniadakis, and C.-H. Jeong, "Sound propagation in realistic interactive 3d scenes with parameterized sources using deep neural operators," *Proceedings of the National Academy of Sciences*, vol. 121, no. 2, e2312159120, 2024.

[12] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[13] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations," *arXiv preprint arXiv:1711.10561*, 2017.

[14] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics informed deep learning (part ii): Data-driven discovery of nonlinear partial differential equations," *arXiv preprint arXiv:1711.10566*, 2017.

[15] X. Karakonstantis, D. Caviedes-Nozal, A. Richard, and E. Fernandez-Grande, "Room impulse response reconstruction with physics-informed deep learning," *The Journal of the Acoustical Society of America*, vol. 155, no. 2, pp. 1048–1059, Feb. 2024.

[16] M. Pezzoli, M. Cobos, F. Antonacci, and A. Sarti, "Sparsity-based sound field separation in the spherical harmonics domain," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1051–1055.

[17] I. Tsunokuni, G. Sato, Y. Ikeda, and Y. Oikawa, "Spatial extrapolation of early room impulse responses with noise-robust physics-informed neural network," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. advpub, 2024EAL2015, 2024.

[18] X. Chen, F. Ma, A. Bastine, P. Samarasinghe, and H. Sun, "Sound field estimation around a rigid sphere with physics-informed neural network," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 1984–1989.

[19] K. Shigemi, S. Koyama, T. Nakamura, and H. Saruwatari, "Physics-informed convolutional neural network with bicubic spline interpolation for sound field estimation," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[20] V. Sitzmann, J. N. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proc. NeurIPS*, 2020.

[21] "OpenAcoustics,"
http : / / www . openacoustics . org / sample - page/. Last accessed 2024-8-5.

Towards a B-format Ambisonic Room Impulse Response Generator Using Conditional Generative Adversarial Network

Hualin Ren[1], Christian Ritz[1, *], Jiahong Zhao[2], Xiguang Zheng[1], Daeyoung Jang[3]
[1]School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, Australia
E-mail: hualin@uow.edu.au, E-mail: critz@uow.edu.au, E-mail: xiguang@uow.edu.au
[2]Institute of Sound and Vibration Research (ISVR), University of Southampton, Hampshire, UK
E-mail: Jiahong.Zhao@soton.ac.uk
[3]Media Coding Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea
E-mail: dyjang@etri.re.kr

***Abstract*— This paper proposes a B-format Ambisonic room impulse response (RIR) generator based on a conditional generative adversarial network (CGAN). The B-format RIR is a first-order Ambisonics (FOA) representation of the RIR used for creating spatial audio for virtual reality (VR) applications. The neural network produces FOA RIRs for specific VR rooms based on given receiver and source positions, and room dimension. The CGAN is trained using real-world FOA RIR recordings, with generative adversarial loss and other loss functions. These loss functions include the frequency-domain multi-resolution short-time Fourier transform (MRSTFT) loss function and various acoustic parameters, such as reverberation time (RT), early decay time (EDT), and direct-to-reverberant ratio (DRR). Adaptive weightings are applied to balance the importance of different loss functions. Objective evaluation metrics include mean squared error, MRSTFT, RT, EDT, DRR, DROQM for listening quality and direction of arrival accuracy. The results show that utilizing acoustic parameters as loss functions effectively controls these metrics, resulting in high-quality FOA RIRs.**

I. INTRODUCTION

In the application of virtual reality (VR), spatial audio is critical to ensure an immersive listening experience. Simulating spatial audio within rooms requires convolution of the sound source with the room impulse response (RIR) between the receiver and source positions. A first-order Ambisonic (FOA) RIR is a B-format array recording. The FOA RIR is beneficial for rendering spatial audio in VR environments corresponding to a listener's changing position and look direction relative to the source, known as six degrees of freedom spatial audio.

For streaming VR applications, it can be impractical to record and transmit the RIR for every possible listening position in a virtual room. Alternatively, RIRs for the chosen source and receiver positions can be synthesized based on acoustic signal processing models that utilize knowledge of the room geometry and wall reflection and absorption coefficients. The well-known image method [1] works well for shoebox room models while more complex methods are required for rooms with irregular geometries and containing reflective surfaces other than walls (e.g. furniture). However, these simulators can be too computationally complex to quickly update the RIR to match a listener's moving position in a virtual room. Alternatively, interpolation approaches applied to the FOA RIR can be used to produce an interpolated FOA RIR at a new position [2], [3], [4], [5] using the sound source direction information derived from the FOA RIR channel relationships. While promising, these approaches still require regular transmission of two or more FOA RIRs to facilitate the interpolation of new FOA RIRs matching the listener's position. In addition, accurately simulating the specular and diffuse reflections for a given room is difficult for conventional FOA RIR synthesis.

Recently, neural network approaches have been proposed for generating RIRs corresponding to arbitrary receiver and source positions in a room. Generative adversarial networks (GAN) [6] are widely applied in image generators, composed of two models, generator and discriminator, trained in a back-and-forth manner to challenge each other. WaveGAN is the first attempt to generate raw-waveform audio based on GAN [7]. As an extension, IR-GAN is proposed as a WaveGAN based RIR generator [8]. However, [8] does not control the acoustic characteristics of RIR (e.g. reverberation time 60 (RT60)) during the training session. Conditional generative adversarial net (CGAN) is introduced as an extension of GAN, to ensure that generated data (e.g. images) is constrained according to match certain desirable criteria [9]. For example, Image2Reverb synthesizes RIRs using a CGAN using images of acoustic environments as input [10]; FAST-RIR is a fast diffuse RIR generator with the inputs including receiver positions, source positions, room dimensions, and room acoustics parameters [11]; and MESH2IR generates RIRs using

a three-dimensional (3D) scene mesh of the room, based on the FAST-RIR architecture [12]. These methods focus on improving the augmentation of far-field automatic speech recognition. Considering RIR interpolation on a spatial grid application, SoundNeRirF adopts multi-layer perceptrons (MLPs) to predict a forward RIR for the target position based on the 3D position of reference and target receivers [13]. Existing approaches aim to generate a mono RIR, while the proposed approach fills the gap to generate FOA RIRs.

Main Contributions: (1) The CGAN-based neural FOA RIR generator is introduced to generate plausible FOA RIRs at the chosen receiver position, source position and room dimension in the given acoustic environment using various combinations of loss functions. Without transmitting the known FOA RIRs, this approach improves the transmission efficiency by simply inputting the receiver position and source position to generate a new FOA RIR. (2) Frequency-domain multi-resolution short-time Fourier transform (MRSTFT) and multiple key acoustic parameters including RT30, early decay time (EDT), and direct-to-reverberant ratio (DRR) are investigated to find the best performance of losses combination and acoustic parameter losses. (3) Adaptive weightings are implemented to balance the main loss and auxiliary losses. (4) The model is trained using recordings of FOA RIRs from a real-world room, while existing literatures in this field mostly work on synthetic mono RIR datasets only or their synthetic mixture with real-world recorded datasets [8], [11], [12], [13]. Evaluations compare the difference between the recorded FOA RIRs and the generated FOA RIRs at the same pair of receiver positions, source positions and room dimensions using measures of mean squared error (MSE), MRSTFT, RT30, EDT and DRR, along with DROQM (for listening quality (LQ)) [14] and direction of arrival (DOA) accuracy of direct sound (DS) [15].

II. Proposed System

The GAN has been shown to generate realistic audio waveforms [7]. Here, a CGAN is used as the basis for the proposed generator to learn the specular and diffuse reflections. The proposed network takes the receiver position, source position and room dimension, represented in 3D Cartesian coordinates (x_r, y_r, z_r),(x_s, y_s, z_s) and (l_{rm}, w_{rm}, h_{rm}), as the vector embedding $\varepsilon_p = (x_r, y_r, z_r, x_s, y_s, z_s, l_{rm}, w_{rm}, h_{rm})$. Each vector embedding is obtained directly from the real-world recorded dataset. The output is one-dimensional (1D) raw-waveform FOA RIR with four channels in Ambisonic channel number (ACN) order *W*, *Y*, *Z* and *X*. The sampling frequency of the generated FOA RIR is 16 kHz, with the length being 8192. The architecture of the proposed model is modified from FAST-RIR [11] that is based on the Stage-I structure of StackGAN [16]. Input of the proposed network is the vector embedding ε_p. The target real-world recorded FOA RIR is the 1D signal with four channels, similar to image inputs with RGB channels. The size of an input signal is set to (32, 4, 8192), corresponding to batch size, channel and height, so that the Dataloader is capable of inputting and outputting 1D FOA RIR (4, 8192), 4 arrays with each length being 8192. Another upsample layer of generator is also added, compared to the FAST-RIR architecture, to adapt the size of the input signal. To generate FOA RIRs in any specific position, the proposed network controls loss functions including the MRSTFT, RT, DRR and EDT.

A. Generator Modified CGAN Loss

Conditioned on ε_p, the generator G is trained through modified CGAN loss to minimise the overall error between the generated FOA RIR and real-world recorded FOA RIR. The D represents the discriminator, while the generated FOA RIR is $\tilde{\mathbf{p}}$ and the real-world recorded FOA RIR is $\mathbf{p}$.

$$\mathcal{L}_{ADV}(G) = \mathbb{E}[(D(\tilde{\mathbf{p}}) - 1)^2] \tag{1}$$

B. MRSTFT Loss

The multi-resolution short-time Fourier transform (MRSTFT) is widely utilized in audio processing [17], [18], [19] due to its effectiveness in capturing the time-frequency distribution of realistic waveforms [18]. This approach helps avoid the generator's overfitting to a single short-time Fourier transform (STFT) representation, which can otherwise degrade performance in the waveform domain [17]. In this study, MRSTFT is computed using four different STFT resolutions: 1024, 2048, 512, and 128. The hop size is set to half the length of the STFT to enable overlapping segments during the transformation process. A Hann window is applied in STFT, where $|\mathfrak{F}\{\cdot\}|$ is the magnitude of a STFT operator that converts time-domain sound pressure vectors $\mathbf{p}(t)$ and $\tilde{\mathbf{p}}(t)$, into spectrograms of dimension. The MRSTFT loss is calculated as the sum of two terms: spectral convergence $\mathcal{L}_{\mathrm{SC}}$ and spectral log-magnitude $\mathcal{L}_{\mathrm{SM}}$ across the various STFT resolutions. The channel is represented by c, from *W* to *Z*.

$$\mathcal{L}_{SC}(\mathbf{p}, \tilde{\mathbf{p}}) = \frac{\| \, |\mathfrak{F}\{\mathbf{p}\}| - |\mathfrak{F}\{\tilde{\mathbf{p}}\}| \, \|_F}{\| \, |\mathfrak{F}\{\mathbf{p}\}| \, \|_F} \tag{2}$$

$$\mathcal{L}_{SM}(\mathbf{p}, \tilde{\mathbf{p}}) = \frac{1}{N} \| \, log|\mathfrak{F}\{\mathbf{p}\}| - log|\mathfrak{F}\{\tilde{\mathbf{p}}\}| \, \|_F \tag{3}$$

$$\mathcal{L}_{MRSTFT}(\mathbf{p}, \tilde{\mathbf{p}}) = \sum_{c=1}^{4} \mathcal{L}_{SC}(\mathbf{p}, \tilde{\mathbf{p}}) + \mathcal{L}_{SM}(\mathbf{p}, \tilde{\mathbf{p}}) \tag{4}$$

C. RT Loss

The RT loss is based on the standard ISO 3382-1:2009. The absolute loss is computed on each channel, and the total loss is added together. RT30 is calculated as the time it takes for the sound level to decay by 30 decibels (dB) after the sound source has stopped emitting.

$$\mathcal{L}_{RT30}(\mathbf{p}, \tilde{\mathbf{p}}) = \sum_{c=1}^{4} |\mathbf{p}_{RT30} - \tilde{\mathbf{p}}_{RT30}| \tag{5}$$

D. DRR Loss

The DRR measures the energy ratio of sound arriving directly from a source to that arriving after one or more reflections from surfaces [20]. To compute the DRR, the direct impulse is first identified, and then the DS component is calculated by measuring the sound energy within 5 milliseconds (ms) around the direct impulse [21].

$$\mathcal{L}_{DRR}(\mathbf{p}, \tilde{\mathbf{p}}) = \sum_{c=1}^{4} |\mathbf{p}_{DRR} - \tilde{\mathbf{p}}_{DRR}| \quad (6)$$

E. EDT Loss

EDT is determined by fitting a straight line to the initial 10 dB of sound decay, indicating changes in perceived reverberance [22]. To predict reverberance ratings accurately, the average EDT is calculated across frequency bands from 125 Hz to 2 kHz [23], in accordance with ISO 3382-1:2009 standard.

$$\mathcal{L}_{EDT}(\mathbf{p}, \tilde{\mathbf{p}}) = \sum_{c=1}^{4} |\mathbf{p}_{EDT} - \tilde{\mathbf{p}}_{EDT}| \quad (7)$$

F. Total Loss Function

The total loss function for the FOA RIR generation integrates several distinct loss components to achieve the most accurate results. The weights assigned to the CGAN loss, MRSTFT, RT30, DRR, EDT losses are controlled by parameters λ_{ADV}, λ_{MRSTFT}, λ_{RT}, λ_{DRR} and λ_{EDT}, respectively. A weight of 0 excludes a loss function from the total loss, while non-zero weights determine the influence of each loss function.

To enhance the accuracy of RIR generation, additional auxiliary losses, such as acoustic parameters, are incorporated alongside primary loss MRSTFT. This paper investigates the use of adaptive weightings, as opposed to fixed weightings commonly used in prior work [11], which are set as hyperparameters and remain unchanged throughout training. Fixed weightings can be inadequate for managing multiple loss functions, often requiring numerous training processes to find the optimal values. In the absence of adaptive adjustments, there can be imbalances where auxiliary losses either dominate or fail to contribute effectively. To address this, adaptive weightings are employed based on the MetaBalance method [24]. This approach adjusts the gradient magnitudes of auxiliary losses (RT30, DRR and EDT losses) to better align with the target loss (MRSTFT loss), ensuring a balanced optimization process by maintaining comparable magnitudes for all losses involved.

$$\mathcal{L}_{\mathrm{G}} = \lambda_{ADV}\mathcal{L}_{ADV} + \lambda_{MRSTFT}\mathcal{L}_{MRSTFT} + \lambda_{RT}\mathcal{L}_{RT} + \lambda_{DRR}\mathcal{L}_{DRR} + \lambda_{EDT}\mathcal{L}_{EDT} \quad (8)$$

$$\mathcal{L}_{D} = \mathbb{E}\left[(D(\mathbf{p}) - 1)^2 + \left(D(\tilde{\mathbf{p}})\right)^2\right] \quad (9)$$

III. Experiments

The proposed model is trained via a GeForce RTX 3080Ti GPU over 600 epochs on the dataset demonstrated below. The optimizer is RMSprop, and the learning rate is 2×10^{-4}.

A. Datasets

The realistic FOA RIR dataset is recorded by using B-format microphones in a room of the Institute of Electronic Music and Acoustics in Graz, Austria [25]. The FOA RIR is carried out in a 10.5 m × 12 m × 5 m studio in ACN order before being normalized. The number of FOA RIRs is 720, while the whole dataset is split into training and testing datasets, with their percentages being 90% and 10%, respectively. Similar to [11], the real-world recorded FOA RIRs are resampled to 16 kHz, and 9600 samples is the length, i.e. 0.6 s. In the experiments, the recorded FOA RIR length is truncated to 8192 samples, corresponding to 0.512 s. The FOA RIRs are normalized by the maximum absolute value among the four channels to maintain the relationships within the channels.

B. Implementations Evaluated

Eight experiments are conducted to assess the performance of the proposed network with various loss combinations. Each experiment builds upon the base CGAN loss described in equation (1). The additional combinations tested include: CGAN loss combined with MRSTFT loss in (4); CGAN loss combined with MRSTFT and RT30 losses in (5), DRR loss in (6), EDT loss in (7); and CGAN loss combined with MRSTFT loss and various combinations of RT30, DRR and EDT losses in (8).

C. Objective Evaluation

The objective metrics are used to evaluate performance by several key measures: MSE over different lengths, MRSTFT, RT30, DRR, EDT, DOA, and DROQM. These metrics assess the impact of various loss functions on the generated FOA RIRs. Errors for MRSTFT, RT30, EDT, and DRR are calculated consistently with their corresponding loss functions, ensuring uniformity in evaluation. The MSE metric is computed over various lengths, including the entire signal, 5 ms and 50 ms for DS length. DROQM predicts the impact on LQ (0 to 1, where higher is better, with over 0.4 acceptable) of the DS part of FOA RIRs [14]. The errors presented in Table 1 are the average absolute errors between the real-world recorded FOA RIRs and the generated FOA RIRs averaged across all four channels for each pair of receiver and source positions in the testing dataset.

The DOA information is extracted from FOA RIR. The four channels *W*, *X*, *Y* and *Z* are derived from a B-format microphone positioned at a specific azimuth angle and elevation angle. The omnidirectional channel *W* represents the impulse response of the audio, while the three orthogonal directional components *X*, *Y*, and *Z* channels capture directional cues. After converting the signal to the time frequency domain using the STFT, the azimuthal DOA is estimated for each time-frequency, where n is the frame number and k is the frequency index, as [15],

$$p_w(n,k) = \frac{W(n,k) \times conj(W(n,k))}{\sum_{n,k}^{N,K} W(n,k) \times conj\big(W(n,k)\big)} \quad (10)$$

Table 1. The evaluation of generated RIR by different loss functions in CUBE dataset.

Loss function implementation	MSE↓ (DS, 5 ms, /)	MSE↓ (DS, 50 ms, /)	MRSTFT↓ (dB)	RT↓ (s)	EDT↓ (s)	DRR↓ (dB)	DOA↓ Azimuth (DS, radian)	DOA↓ Elevation (DS, radian)	DROQM↑ (DS, LQ, /)
MRSTFT	0.04	**0.01**	3.14	0.20	0.19	2.64	0.82	0.42	0.44
MRSTFT, RT30	0.03	0.02	3.44	0.16	0.12	2.75	0.80	0.48	0.45
MRSTFT, EDT	**0.02**	**0.01**	3.30	0.21	0.14	2.28	**0.69**	**0.36**	0.43
MRSTFT, DRR	0.04	**0.01**	3.32	0.23	0.19	2.20	0.84	0.45	0.42
MRSTFT, RT30, EDT	0.04	0.02	3.67	0.24	0.25	2.29	0.81	0.53	**0.46**
MRSTFT, RT30, DRR	0.03	**0.01**	**3.10**	0.16	0.16	2.61	0.78	0.47	**0.46**
MRSTFT, EDT, DRR	0.04	**0.01**	3.21	0.18	0.11	2.43	0.89	0.41	0.41
MRSTFT, RT30, EDT, DRR	0.05	0.02	3.42	**0.15**	**0.10**	**1.96**	0.78	0.43	0.41

* The bold and underlined text indicates the lowest error in Table 1. The "/" means no unit.

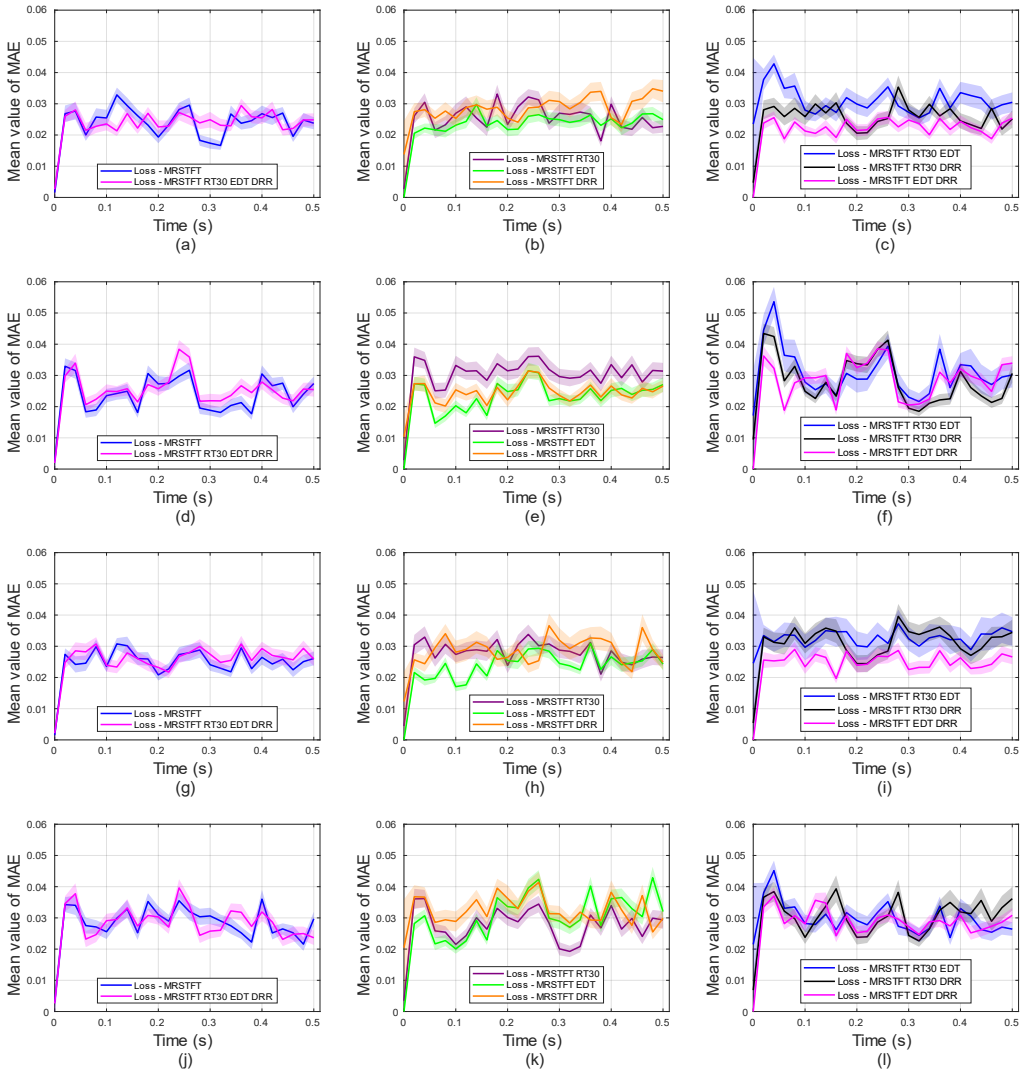


Fig. 1. Mean value of MAE with 95% confidence interval (shaded area) for the model using MRSTFT as the main loss function, combined with various acoustic parameters as auxiliary loss functions across time. (a) - (c) for *W* channel, (d) - (f) for *Y* channel, (g) - (i) for *Z* channel, (j) - (l) for *X* channel.

$$I_{C=X,Y,Z}(n,k) = Re(conj(W(n,k)) \times C(n,k)) \quad (11)$$

$$I_{azi}(n,k) = \frac{I_{C=Y}(n,k)}{I_{C=X}(n,k)} \quad (12)$$

$$I_{ele}(n,k) = \frac{-I_{C=Z}(n,k)}{\sqrt{\left(I_{C=X}(n,k)\right)^2 + \left(I_{C=Y}(n,k)\right)^2}} \quad (13)$$

$$DOA_{azi} = \sum_{n,k}^{N,K} p_w(n,k) \times Re\left(\tan^{-1}\left(I_{azi}(n,k)\right)\right) \quad (14)$$

$$DOA_{ele} = \sum_{n,k}^{N,K} p_w(n,k) \times Re\left(\tan^{-1}\left(I_{ele}(n,k)\right)\right) \quad (15)$$

where, $p_w(n,k)$ is the power weightings, while $I_{azi}(n,k)$ and $I_{ele}(n,k)$ are the azimuth ad elevation power intensity of each time-frequency instant, respectively. The $W(n,k)$, $C(n,k)$, where $C = X,Y,Z$ are the time-frequency instant in the four channels. The $conj$ represents conjunct, while Re demonstrates real part. The final DOA estimation is based on a weighted average of the values in Eq. (10), where the weights are based on the magnitude of the corresponding time-frequency component in the *W* channel (this reduces the impact on the DOA estimation of low magnitude components not corresponding to DS [15]). The segmental length of DS in time domain is 5 ms referenced by [21].

IV. Results

When comparing the MSE for different lengths of DS segments in Table 1, it is clear that the model incorporating MRSTFT and EDT losses achieves the lowest MSE for both 5 ms and 50 ms segment lengths. This model also records the lowest DOA azimuth and elevation errors, as DOA and MSE metrics evaluate the same DS section (5 ms). Furthermore, the model with MRSTFT, RT30, EDT, and DRR losses achieves the minimum errors in RT, EDT, and DRR measurements. The objective results in Table 1 also reveal that the model with MRSTFT and RT30 losses achieves the second-lowest RT30 errors, while the model combining MRSTFT and DRR losses achieves the second-lowest DRR error. Although the model with MRSTFT and EDT does not reach the lowest EDT error, the model combined MRSTFT, RT, EDT, and DRR losses does. Additionally, the model with MRSTFT, RT30, and DRR losses achieves the lowest MRSTFT error. The DROQM scores for all models exceed 0.4, indicating that the quality is acceptable.

Fig. 1 illustrates the mean value of MAE with 95% confidence intervals for the model using MRSTFT as main loss function, with various combinations of acoustic parameters as auxiliary loss functions over time for each channel. Comparing Fig. 1(a), 1(d), 1(g), and 1(j), the model combining MRSTFT, RT30, EDT, and DRR losses (represented by the magenta line) performs relatively stable to the model using only the MRSTFT loss (blue line). In Fig. 1(b), 1(e), 1(h), and 1(k), the model that combines MRSTFT and EDT losses (green line) consistently performs well compared to models that combine MRSTFT with only one other acoustic parameter. This model achieves the lowest mean value of MSE from 0 to 0.05 seconds, covering the DS and some early reflections, outperforming all other models in this time frame. This corresponds to the lowest MSE across different segment lengths and DOA errors. Additionally, when considering the MRSTFT loss combined with two acoustic parameter losses, the model with MRSTFT, EDT, and DRR losses (magenta line) shows greater stability with fewer errors throughout the entire signal duration, as shown in Fig. 1(c), 1(f), 1(i), and 1(l). This stability is more observable compared to the combination of MRSTFT, RT30, and EDT losses (blue line), as well as the combination of MRSTFT, RT30, and DRR losses (black line).

V. Discussion

Different acoustic parameters contribute to improvements in various aspects of the performance of models. Generally, loss combinations that include the acoustic parameter EDT consistently result in lower errors compared to those incorporating RT30 and DRR. When specific evaluations aim to achieve the lowest error values, selecting the corresponding acoustic parameter can reach the goal. A significant challenge in training a network with FOA RIR data is managing multiple channels to maintain the directional relationships among B-format channels. The experiments demonstrate that focusing on generation accuracy for each channel enhances the overall accuracy of generated FOA RIRs.

VI. Conclusions and Future Work

In this work, a novel CGAN-based network is proposed to generate FOA RIRs for specified receiver and source positions, as well as room dimension. The network employs a generator loss that includes adaptive weightings of various combinations of acoustic parameter loss functions, in addition to the CGAN loss. Objective results show that this network facilitates the generation of FOA RIRs with controlled MRSTFT and acoustic parameters through adaptive weightings. Specific acoustic parameters are utilized to manage the corresponding characteristics of the generated FOA RIRs. The models incorporating MRSTFT and EDT, as well as those combining MRSTFT, RT30, EDT, and DRR, outperform other loss combinations. The EDT acoustic parameter demonstrates better stability and fewer errors compared to models using alternative acoustic parameter losses. Future work will investigate alternative loss functions to further improve FOA RIR generation and explore extending this approach to produce higher-order Ambisonics using neural networks.

VII. Acknowledgment

This research was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded

by the Korean government, grant number 23ZH1200 (The research of the basic media contents technologies).

REFERENCES

[1] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979, doi: 10.1121/1.382599.

[2] J. Zhao, X. Zheng, C. Ritz, and D. Jang, "Interpolating the Directional Room Impulse Response for Dynamic Spatial Audio Reproduction," *Applied Sciences*, vol. 12, no. 4, Art. no. 4, Jan. 2022, doi: 10.3390/app12042061.

[3] A. Southern, J. Wells, and D. Murphy, "Rendering walk-through auralisations using wave-based acoustical models," in *2009 17th European Signal Processing Conference*, IEEE, 2009, pp. 715–719.

[4] J. G. Tylka and E. Y. Choueiri, "Soundfield Navigation using an Array of Higher-Order Ambisonics Microphones," *Los Angeles*, p. 10, 2016.

[5] T. McKenzie, N. Meyer-Kahlen, R. Daugintis, L. McCormack, S. Schlecht, and V. Pulkki, "Perceptually informed interpolation and rendering of spatial room impulse responses for room transitions," presented at the 24th International Congress on Acoustics (ICA), Gyeongju, Korea, Oct. 2022.

[6] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, in NIPS'14. Cambridge, MA, USA: MIT Press, Dec. 2014, pp. 2672–2680.

[7] C. Donahue, J. McAuley, and M. Puckette, "Adversarial Audio Synthesis," presented at the International Conference on Learning Representations, Feb. 2018.

[8] A. Ratnarajah, Z. Tang, and D. Manocha, "IR-GAN: Room Impulse Response Generator for Far-Field Speech Recognition," in *Interspeech 2021*, ISCA, Aug. 2021, pp. 286–290. doi: 10.21437/Interspeech.2021-230.

[9] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv.org. Available: https://arxiv.org/abs/1411.1784v1

[10] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 286–295. doi: 10.1109/ICCV48922.2021.00035.

[11] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "Fast-Rir: Fast Neural Diffuse Room Impulse Response Generator," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 571–575. doi: 10.1109/ICASSP43922.2022.9747846.

[12] A. Ratnarajah, Z. Tang, R. Aralikatti, and D. Manocha, "MESH2IR: Neural Acoustic Impulse Response Generator for Complex 3D Scenes," in *Proceedings of the 30th ACM International Conference on Multimedia*, Lisboa Portugal: ACM, Oct. 2022, pp. 924–933. doi: 10.1145/3503161.3548253.

[13] Y. He, J.-X. Zhong, Z. Dai, N. Trigoni, and A. Markham, "SoundNeRirF: Receiver-to-Receiver Sound Neural Room Impulse Response Field," Sep. 2022.

[14] H. Ren, C. Ritz, J. Zhao, and D. Jang, "Towards an Objective Quality Metric for Interpolated Directional Room Impulse Responses," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 8205–8209. doi: 10.1109/ICASSP48485.2024.10446507.

[15] X. Zheng, C. Ritz, and J. Xi, "Encoding and communicating navigable speech soundfields," *Multimed Tools Appl*, vol. 75, no. 9, pp. 5183–5204, May 2016, doi: 10.1007/s11042-015-2989-3.

[16] H. Zhang *et al.*, "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5908–5916. doi: 10.1109/ICCV.2017.629.

[17] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6199–6203. doi: 10.1109/ICASSP40776.2020.9053795.

[18] S. Joshi *et al.*, "Defense against Adversarial Attacks on Hybrid Speech Recognition System using Adversarial Fine-tuning with Denoiser," Sep. 2022, pp. 5035–5039. doi: 10.21437/Interspeech.2022-10977.

[19] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft, "Generative models for sound field reconstruction," *The Journal of the Acoustical Society of America*, vol. 153, no. 2, pp. 1179–1190, Feb. 2023, doi: 10.1121/10.0016896.

[20] Patrick A. Naylor and Nikolay D. Gaubitch, *Speech Dereverberation*. Springer London, 2010.

[21] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2110–2117, Oct. 2002, doi: 10.1121/1.1506692.

[22] J. S. Bradley, "Review of objective room acoustics measures and future needs," *Applied Acoustics*, vol. 72, no. 10, pp. 713–720, Oct. 2011, doi: 10.1016/j.apacoust.2011.04.004.

[23] M. Barron, "Subjective Study of British Symphony Concert Halls," *Acta Acustica united with Acustica*, vol. 66, no. 1, pp. 1–14, Jun. 1988.

[24] Y. He, X. Feng, C. Cheng, G. Ji, Y. Guo, and J. Caverlee, "MetaBalance: Improving Multi-Task Recommendations via Adapting Gradient Magnitudes of Auxiliary Tasks," in *Proceedings of the ACM Web Conference 2022*, in WWW '22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 2205–2215. doi: 10.1145/3485447.3512093.

[25] K. Müller, "CUBE B-format RIR dataset (Soundfield ST450 MKII)". Available: https://phaidra.kug.ac.at/o:104435

Improved Cassava Plant Disease Classification with Leaf Detection

Ming Xuan Chai*†, Yao Deng Fam*†, Quinito Norman Octaviano*, Chih-Yang Pee*‡,
Lai-Kuan Wong*, Mas Ira Syafila Mohd Hilmi Tan*, John See§
* Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Malaysia
§ School of Mathematical and Computer Science, Heriot-Watt University (Malaysia), 62200 Putrajaya, Malaysia
† Equal contribution (Emails: mingxuan.chai.cmx@gmail.com and yaodeng07@gmail.com)
‡ Corresponding author (Email: cypee@mmu.edu.my)

***Abstract*—Cassava (Manihot esculenta Crantz), a crucial food source for millions, has experienced significant crop yield losses due to various plant diseases. To address this issue, extensive research has been conducted on the Cassava 2020 dataset released by the Makerere AI Lab. Images in this dataset are captured in complex real-world environments, adding difficulty to the classification process. While several deep learning models have shown promising results, we believe that reducing interference from overlapping leaves and complex backgrounds could further enhance classification performance. To accomplish this, up to five leaves of interest have been identified and annotated in each image. These annotated leaves are then used to train a detection model: Faster R-CNN with ResNet-101, designed to automatically detect leaves of interest in the images. The identified leaves are separated from the background through a masking process, producing masked images (Mi), where i indicates the maximum number of leaves in the masked images, with i ranging from 1 to 5. The masked images M1 to M5 are then trained on various CNN models; i.e. EfficientNetB1, DenseNet121, ResNet50, and Xception, for Cassava disease classification. The results show that classifiers trained with M3 improve the accuracy by 2.13% to 3.06% compared to models trained with original images. The main contribution of this work is a novel technique capable of identifying leaves of interest from complex backgrounds to produce masked images, which improve the accuracy of various CNN-based classification models, and a novel leaf annotation where up to 5 leaves of interest have been annotated for each image in Cassava 2020 dataset. Annotated images can serve as a valuable resource for advanced cassava disease analysis, detection, and classification.**

I. Introduction

Plant diseases pose a serious threat to agricultural productivity, accounting for 20% to 40% of crop losses globally [1]. Therefore, improving plant yields through early detection of plant diseases using advanced artificial intelligence (AI) is crucial to minimize both crop and economic losses [2]. Early plant disease study focus mainly on relatively clean images where traditional machine learning (ML) like K-Means Clustering, Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), k-nearest neighbour(KNN), etc. are effective [3]–[6]. Although these ML algorithms offer promising solutions, effectively detecting diseases in real-world images remain tricky mainly because of mixed healthy and diseased leaves, invisible lesions in the early stages of infection, and complex backgrounds that introduce noise [7].

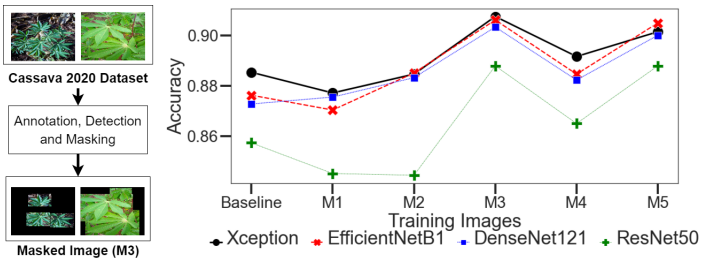


Fig. 1. Accuracy of Cassava disease classification using Xception, EfficientNetB1, DenseNet121, and ResNet50 classifiers trained on original images (Baseline) and masked images (M1-M5). Classifiers trained on M3, M4, and M5 show significant improvements over baseline, while M1 and M2 generally reduce performance, except for DenseNet121. Further details are discussed in Subsection III-B

In order to support advanced machine learning applications in agriculture, particularly for improving the diagnosis and management of cassava diseases, the Artificial Intelligence Lab of Makerere University published the Cassava 2019 dataset in Kaggle [8]. This dataset consists of 22,026 images, where only 5,657 images labeled with diseases. Surya et al. in [9] demonstrated that a pre-trained MobileNetV2 model could fairly classify the diseases but the performance was affected by imbalanced data. Ravi et al. [10] emphasized the disease features on leaves with an attention-based model using EfficientNet and classified the features by implementing meta-classifier (random forest + SVM for prediction and logistic regression for classification). Sambasivam and Geoffrey [11] built a low-cost method to detect cassava infections. The quality of the images was improved by Contrast Limited Adaptive Histogram Equalization (CLAHE), and the images were augmented by applying a combination of SMOTE, focal loss, and class-weight due to the imbalance of the data. In addition, Lilhore et al. [12] enhanced the CNN model by utilizing depth-wise separable convolution, de-correlation stretching with gamma correction, and batch normalization to improve feature count, image color segregation, and overall computational efficiency.

In 2020, the Makerere AI Lab released the Cassava 2020 dataset [13], which serves as an update to the Cassava 2019 dataset. The number of images labeled with diseases in the dataset has grown from 5,657 to 21,397. Utilizing this dataset,

Maryum et al. [14] studied the technique of utilizing U-Net, a deep learning model for leaf segmentation, with EfficientNetB4 (32 layers + 2 fine-tuned layers) for classification with complex backgrounds. The Color Index Vegetation method was applied to extract the vegetation parts from the leaves, providing ground truth for corresponding images. At the same time, Zhang et al. [15] explore a self-supervised learning approach, that allows for classification and training of a reliable model. Their methods utilized the SimCLR model based on ResNet architecture, which maximizes the similarity between the same images with the ResNet18 model to minimize recognition and classification loss during model training. However, self-supervised learning has a critical limitation in training ratios, achieving optimal performance (90%) only with a specific dataset ratio. Concurrently, Zhuang [16] addressed this challenge by utilizing ViT with data augmentation to handle the imbalanced dataset issue and loss function to detect cassava disease on leaves. Singh and the team [17] proposed a similar method using transfer learning techniques with DenseNet169 and the Adam optimizer, resulting in a less time-consuming implementation, though the results are less satisfactory compared to earlier approaches. Soon after, Thai et al. [18] enhanced the base vision transformer with a pruning algorithm to optimize self-attention heads in ViT layers, resulting in reduced model architecture implementation while attaining better results(96.82%) compared to earlier vision transformer approaches.

Inspired by Lilhore et al. [12] and Maryum et al. [14] on how these models reduce background noise in cassava disease classification, we propose a novel framework where an object detection algorithm is used to identify leaves of interest in images so that they can be isolated from images to form masked images. The masked images are then used for training of CNN based cassava disease classifiers like EfficientNetB1, DenseNet121, ResNet50, and Xception. Figure 1 illustrates the forming of masked images from object detection algorithm and performance of CNN based classifiers mentioned above, trained using original images as baseline and masked images (M1-M5). The graph shows compared to baseline, significant improvement in classification accuracy can be achieved when masked images M3 to M5 are used to train the classification models. The main contribution of this article are as follows:

- A novel technique that can accurately identifying leaves of interest from images with complex backgrounds to generate masked images. Classifiers trained using masked images M3 achieve significant improvement in accuracy by 2.13% to 3.06% compared to models trained on original images.
- A new leaf annotation has been created, featuring up to five leaves of interest per image from the Cassava 2020 dataset. This annotation can serve as a valuable resource for advanced disease leaf analysis, detection, and classification.

The remainder of this paper is organized as follows: Section II details the proposed leaf disease classification model, which includes leaf annotations, object detection, masked image generation, and the development of disease classifiers. This is followed by a discussion of experimental results and an ablation study. Finally, the paper concludes in Section IV.

II. Our Approach

In this subsection, the methodology of proposed method is presented. As shown in Figure 2, the process begins with data preparation, where leaf annotation is performed on each image in Cassava 2020 dataset to identify its leaves of interest. The dataset is stratified and split into training and testing sets with a ratio of 80:20. This is following by model building process where leaf detection model is built using the training dataset images together with their leaf annotations. The leaves detection model is then used to identify leaves of interest in each image to produce masked images which are then used to build disease classification system. Finally, the testing images are deployed on the developed leaf detection and disease classification system to evaluate its performance.

A. Cassava 2020 Dataset and leaf Annotation

Cassava 2020 Dataset [13] published by the Makerere Artificial Intelligence Lab in Kaggle, comprises 21,397 labeled images, each with dimensions of 800×600 pixels, a resolution of 96 dpi, and 24-bit RGB colors, stored in JPEG format. The included images consist of healthy (H) cassava plants and four types of diseases: Cassava Bacterial Blight (CBB), Cassava Brown Streak Disease (CBSD), Cassava Green Mottle (CGM), and Cassava Mosaic Disease (CMD). The distribution of the dataset is as shown in TABLE I.

Sample images for each class are shown in Figure 3. As the images are captured in a real-life environment, they are complicated and difficult to classify. Hence, in order to enhance the classification performance, up to five leaves of interest are labelled in each image. The leaf annotation, as shown in Figure 4, are performed by using LabelImg [19] following these guidelines:

1) The first annotated leaf in each image denotes the class to which the image is assigned.
2) Each image must have at least one annotated leaf and no more than five annotated leaves.
3) The order of labeling each leaf is based on its visual prominence.
4) A visually prominent healthy leaf in any disease image should be labeled as healthy (H). Likewise, a visually prominent unhealthy leaf in a healthy image should be labeled as unhealthy (U).

TABLE I
Training and Testing of Cassava 2020 Dataset

Category	Class	Number of Images		
		Training	Testing	Total
0	CBB	870	217	1,087
1	CBSD	1,751	438	2,189
2	CGM	1,909	477	2,386
3	CMD	10,526	2,632	13,158
4	H	2,062	515	2,577
	Total	17,118	4,279	21,397

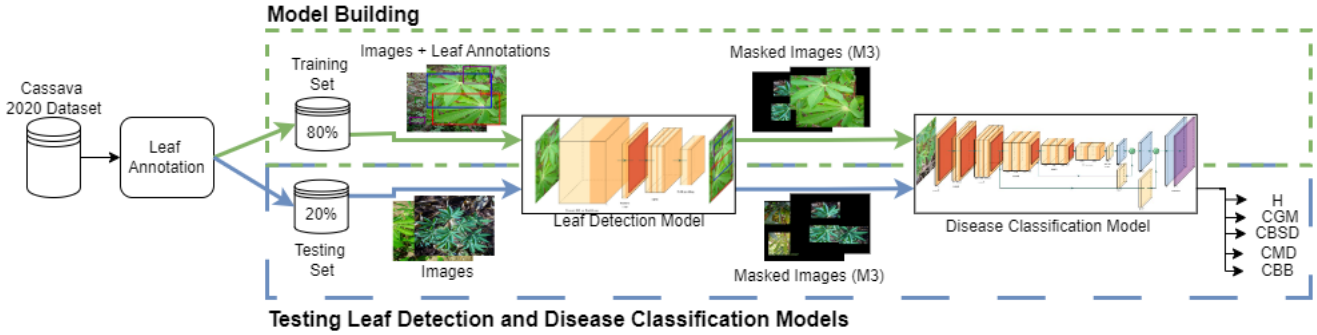


Fig. 2. Framework of Cassava Plant Diseases Classification



Fig. 3. Samples of Cassava Leaves by Classes

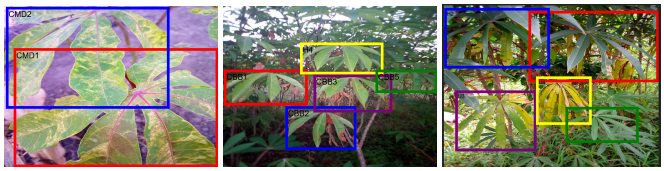


Fig. 4. Sample of Leaf Annotations Where Leaves Sequence are Visualized by Colors of Bounding Boxes: Red, Blue, Purple, Yellow and Green.

An annotated leaf in an image is enclosed by a bounding box and tagged with a label consists of a class: "CBB", "CBSD", "CGM", "CMD", "H", or "U", followed by a number from 1 to 5, indicating the sequence of the leaves. The additional class "U" is used in leaf annotation because prominent unhealthy leaves may appear in otherwise healthy images, as shown in Figure 4(c). Since unhealthy leaves can often appear on healthy plants, they should not be excluded from the image. Most images in the Cassava 2020 dataset focus on a specific leaf (Figure 4(a)) or a small group of leaves (Figures 4(b) and 4(c)). Thus, labeling the five most prominent leaves is typically sufficient to capture the leaves of interest in most Cassava 2020 images.

B. Cassava Leaf Detection and Masked Images

Continued from leaf annotation, three different object detection models; Faster R-CNN with ResNet-101 [20], EfficientDet [21] and Single-Shot Detector (SSD) MobileNetV2 [22]; were evaluated for the leaf detection task to automatically identify leaves of interests from images. Faster R-CNN ResNet101 is known for its object detection prowess. Pre-trained on COCO dataset, the model can generalize to detect various unseen objects in images. SSD EfficientDet and SSD MobileNetV2 offer a compelling balance of detection speed and accuracy, making them the preferred choice. The leaf detection models

TABLE II
MEAN AVERAGE PRECISION OF DETECTION MODELS

	mAP@0.5				
Detection Model	M1	M2	M3	M4	M5
SSD MobileNetV2	0.4817	0.4889	0.5134	0.5288	0.5349
SSD EfficientDet D1	0.5776	0.5982	0.6108	0.6091	0.6005
Faster R-CNN ResNet101	**0.6051**	**0.6136**	**0.6194**	**0.6163**	**0.6050**

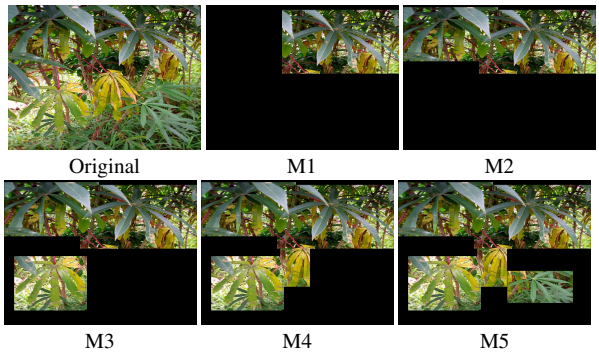


Fig. 5. A Sample of Original and Its Masked Images M1 to M5

were trained on images annotated with different numbers of leaves, ranging from one to five per image, for detecting the same number of leaves. Mean average precision calculated from intersection over union with threshold 0.5 (mAP@0.5) is used to evaluate the performance of leaf detection models. The results are shown in TABLE II where Mi, $i = 1, 2 \cdots, 5$ of columns 2 to 6 denote the scores of mAP@0.5 when up to i leaves of interest are detected by the models in each image.

As shown in the last row of TABLE II, Faster R-CNN ResNet101 outperforms all other models in mAP@0.5 scores in all one to five leaves of interest detection (M1 - M5). In addition, Faster R-CNN ResNet101 score highest in identifying three leaves of interest (M3). Consequently, the leaves detected by Faster R-CNN with ResNet101 are used to create masked images Mi, which are henceforth referred to as masked images containing up to i leaves per image. A sample of original and masked images is shown in Figure 5.

C. Cassava Plant Disease Classification

To evaluate the effectiveness of using masked images generated from leaf detection algorithm, the following CNN models are selected: DenseNet121 [23], EfficientNetB1 [24], ResNet50 [25], and Xception [26]. In this experiment, the CNN models

mentioned above are trained using original Casava training dataset, which serves as the baseline for evaluating the effectiveness of our proposed approach. All models are fine-tuned with pre-trained ImageNet weights [27]. The pre-trained models are connected to a 2D global average pooling layer, followed by the 2 fully connected (FC) layers of 256 and 128 nodes. A dropout rate of 0.4 is applied after each FC layer to allow the models to learn more robust features. A batch normalization layer is inserted after the first dropout layer to speed up the convergence of the model. The final layer of the network utilized Softmax activation function to perform classification of the 5 diseased classes.

III. Empirical Study

A. Experiment Settings

In this experiment, the Cassava 2020 dataset was stratified and split randomly into training and testing sets using 80:20 ratio, with a fixed seed for reproducibility. Input training images were resized to 300×300 pixels. The batch size and epoch were set to 32 and 100, respectively. The initial learning rate was set to 0.001, where it reduced by a factor of 0.1 when there is no improvements in loss for 5 epochs. Early stopping was also utilized to prevent the occurrence of overfitting. The models were trained on a workstation equipped with Intel Core i7-12700K processor running at 3.6 GHz, 64 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU.

B. Classification Performance of the Proposed Model

Our results demonstrate that isolating the leaves from complex backgrounds is effective for improving performance of leaf disease classification. Figure 1 illustrates the performance of Xception, EfficientNetB1, DenseNet121, and ResNet50 classifiers trained on original images (as baseline) and masked images (M1-M5). It can be observed from the results that regardless of the type of CNN models, there is a consistent performance improvement over the respective baseline for models trained with at least 3 leaves of interest. Notably, classifiers trained on M3 images give the best performance. As depicted in row 2 to 5 of TABLE III, the classification models trained on M3 images show significant accuracy improvements of 2.13% to 3.06% over their corresponding baseline models reported in row 6 to 9 of TABLE III respectively. A key highlight is that, across the four CNN models trained on M3, Xception achieved the highest score with 90.64%. Returning to Figure 1, performance of models trained with M5 images is close to M3, with only a slight overall drop in performance. Unexpectedly, there is a consistent performance drop for models trained with M4 images. An investigation into the annotated dataset reveals that some of the smaller leaves, coincidentally those annotated as the fourth leaf, may not accurately reflect the disease class that the image represents. As an example, the fourth labelled leaf (green) shown in Figure 4(b) does not resemble the neigbouring leaves in the same image and it is very likely that the leaf does not belong to the labeled class. This misrepresentation, which is natural in the real-world scenario where not all leaves of a diseased plant show the disease symptoms, is unavoidable and resulted in noise that reduces the models performance. On the other hand, except for DenseNet121, the performance of models trained with M1 and M2 images shows a dip in performance compared to the baseline models. This could be due to the smaller number of leaves of interest in M1 and M2 images, leading to insufficient diversity and information to support the models in learning to discriminate the diseased leaves.

TABLE III
Classification Performance Comparison

Paper	Methods	Train-Test	Accuracy (%)
Proposed Model	**Xception (M3)**	80 - 20	**90.64**
	EfficientNetB1 (M3)	80 - 20	90.61
	DenseNet121 (M3)	80 - 20	90.33
	ResNet50 (M3)	80 - 20	88.79
	Xception (Baseline)	80 - 20	88.51
	EfficientNetB1 (Baseline)	80 - 20	87.60
	DenseNet121 (Baseline)	80 - 20	87.28
	ResNet50 (Baseline)	80 - 20	85.73
Zhang et al. [15]	SimCLR + ResNet18	70 - 30	90.00
Maryum et al. [14]	UNet+ EfficientNetB4	85 - 15	89.90
Singh et al. [17]	DenseNet169	80 - 20	87.86
Zhong et al. [28]	Transformer-Embedded ResNet	80 - 20	**91.12**
Zhuang, L. [16]	Vision Transformer	80 - 20	90.02

Although the baseline models were fine-tuned with the same weights, slight differences in results were observed, especially with ResNet50, which had the lowest accuracy, trailing Xception by 2.78%. This may be due to weaker feature extraction of ResNet50, missing finer details in cassava images with complex backgrounds. In contrast, Xception's architecture, particularly its depthwise separable convolutions, allows it to capture subtle features more effectively. While EfficientNetB1 is highly efficient due to its compound scaling, it may sacrifice sensitivity to intricate details, giving Xception an edge in this task. DenseNet121, though achieving 87.28%, suffers from feature redundancy due to its densely connected architecture, which potentially hinders its ability to focus on new features.

Experiments are compared with existing approaches as detailed in rows 10 to 15 of Table III. Rows 10 to 12 present various CNN approaches trained with the same image size of 224×224 but with different train-test splits. Zhang et al. [15] proposed an ensemble model of SimCLR and ResNet18 with a 70-30 train-test split, achieving an accuracy of 90.00%. Maryum et al. [14] employed U-Net to segment the leaves before classifying the disease with EfficientNetB4, training their model with an 85-15 train-test split dataset, resulting in an accuracy of 89.90%. Furthermore, Singh et al. [17] utilized DenseNet169 to classify cassava leaf diseases, training their model with an 80-20 train-test split dataset, achieving an accuracy of 87.86%. By comparison, the proposed Xception model trained on M3 images demonstrated better performance. Additionally, Table III's rows 13 and 14 present Vision Transformer models trained on higher input image resolution and stratified 5-fold cross-validation, with an 80-20 train-validation split. Zhong et al. [28] achieved 91.12% accuracy using the Transformer-Embedded ResNet with input images sized at 512×512. Additionally, Zhuang [16] trained a ViT model on a dataset with an input image size of 384×384 incorporating image augmentation techniques, achieving 90.02%.

While the proposed Xception method demonstrates comparable performance, its accuracy lags slightly behind Transformer-Embedded ResNet approaches, potentially due to the difference in image resolution.

Overall, fair comparison with state-of-the-art methods is hindered by variations in experimental settings, including image resolution, train-test splits, and evaluation methodology. More crucially, comparison with the baseline demonstrates a considerable improvement in performance, demonstrating the efficacy of our proposed strategy of employing the leaf masks to direct the model's attention.

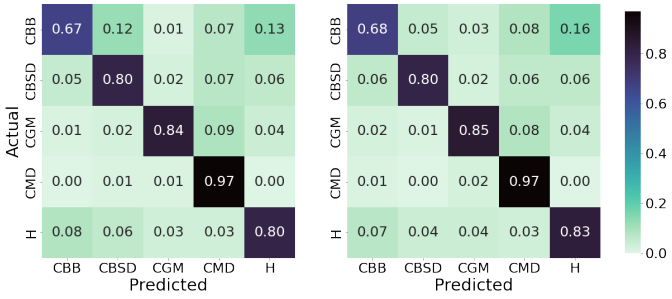
C. Xception and EfficientNetB1 Trained with M3 Images

In this subsection, we analyze the performance of the two best performing models trained on M3 images; Xception and EfficientNetB1, which achieve classification accuracy of 90.64% and 90.61% respectively. Table IV compares the precision, recall, and F1-Score for Xception and EfficientNetB1 models on the five disease classes. It can be observed that the performance of both models are very close, with Exception slightly outperformed EfficientNetB1 on the overall weighted (micro) precision, recall and F1-score. Xception achieves better precision for CBB and H classes, with significant improvements seen in the CBB class, where Xception achieved a precision of 66.06%, which is 3.98% better than EfficientNetB1. On the other hand, EfficientNetB1 yields better precision for CBSD, CGM and CMD. In terms of recall, Xception obtained better scores for CGM and CMD, while EfficientNetB1 scores higher for CBB, CBSD and H. It is interesting that EfficientNetB1 achieves better precision and recall for three of the five classes although it obtained lower scores for the overall metrics. Considering the total number of parameters for EfficientNetB1 (7.1M) is only one-thrid of Xception (22.9M), EfficientNetB1 model can be a better option for resource-constrained deployment.

TABLE IV
XCEPTION AND EFFICIENTNETB1 ON M3 IMAGES

	Xception			EfficientNetB1		
Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score
CBB	0.6606	0.6697	0.6651	0.6208	0.6835	0.6507
CBSD	0.8682	0.7968	0.8310	0.8794	0.7991	0.8373
CGM	0.8244	0.8658	0.8446	0.8269	0.8512	0.8388
CMD	0.9625	0.9761	0.9761	0.9644	0.9677	0.9660
H	0.8422	0.7965	0.8187	0.8356	0.8275	0.8315
Weighted	**0.9076**	**0.9082**	**0.9075**	0.9073	0.9061	0.9064

Figure 6 depicts the confusion matrix for Xception and EfficientNetB1. Results for Exception in Figure 6(a) indicate that the CMD class achieved the highest accuracy at 97.42%, while the CBB class had the lowest accuracy at 67.10%. This discrepancy may be attributed to the imbalanced dataset, with the CMD class being over-represented and the CBB class being underrepresented. Specifically, the CBB class comprises only 1,087 images—approximately 5% of the total dataset—whereas the CMD class includes 13,158 images, representing 60% of the dataset. Notably, the CBB class was misclassified as the Healthy class 13.10% of the time, a relatively high rate compared to other false predictions. This may be due to similarities in patterns between the CBB and Healthy classes, as depicted in Figure 4(b) and Figure 7 that may have contributed to this misclassification. Similarly, the performance of the EfficientNetB1 model, shown in Figure 6(b), reveals a similar pattern, with imbalance data size having an impact on accuracy. However, the misclassification rate for EfficientNetB1 in the CBB and Healthy classes is lower compratively, suggesting that EfficientNetB1 is less sensitive to dataset size variations when compared to Xception model.



(a) Xception (b) EfficientNetB1

Fig. 6. Confusion Matrices for Xception and EfficientNetB1 on M3



(a) H (b) CBB

Fig. 7. Samples of Healthy and CBB Images

D. Xception Classification Performance on Masked Images

Table V compares the results of the Xception model trained on the original images (baseline) and the five set of masked images (M1-M5), categorized by cassava disease classes. The baseline model yielded an overall accuracy score of 88.52%. In contrast, the model trained on M3 images demonstrated a 2.13% improvement, achieving 90.64% accuracy; the model trained on masked images M4, achieved 89.16%, reflecting a 0.65% improvement; and the model trained on masked images M5, improved by 1.61%, reaching 90.12% accuracy. However, models trained with masked images M1 and M2 resulted in slight regressions. This regression could be attributed to the reduced leaf representations, resulting in less diverse and robust leaf features. Notably, the results of all approaches are significantly biased toward CMD, the largest class. Interestingly, isolating leaves from complex backgrounds aids better prediction in minority classes such as CBB and the Healthy class. It can be observed that all models trained with masked images (M1-M5) outperformed the baseline for these two classes. Specifically, CBB achieved 67.10% accuracy on M3 images and the Healthy class achieved 81.01% accuracy with masked images M5. Overall, the proposed method helped to improve the prediction in minority classes within the range of 1.67% to 9.1%. This results reinforce that models trained with at least three leaves of interest (M3-M5) successfully capture sufficient diversity and distinct characteristics of diseased leaves, leading to improved disease classification over the baseline model.

TABLE V
CLASSIFICATION PERFORMANCE OF THE XCEPTION MODEL

	Accuracy(%)					
Class	Baseline	M1	M2	M3	M4	M5
CBB	57.34	63.30	60.55	**67.10**	62.39	61.47
CBSD	80.82	80.14	79.00	80.39	**83.11**	78.09
CGM	78.62	71.07	71.91	**84.41**	80.29	83.86
CMD	**97.45**	96.28	97.07	97.42	96.81	**97.45**
H	72.90	76.12	79.61	80.43	74.76	**81.01**
Overall	88.52	87.71	88.45	**90.64**	89.16	90.12

IV. CONCLUSION

This paper proposes a novel approach to enhance cassava disease classification by implementing the Faster RCNN ResNet101 leaf detection algorithm to identify and extract leaves of interest from images with complex backgrounds. Masked images form by the extracted leaves are subsequently utilized for the training and evaluation of CNN-based classification models, including Xception, EfficientNetB1, DenseNet121, and ResNet50. To train the Faster RCNN ResNet101 detection algorithm, leaf annotations were meticulously conducted on each image from the Cassava 2020 dataset, with up to five of the most prominent leaves being labeled. The experimental results indicate that masked images containing a maximum of three leaves (M3) yield the most significant improvement in disease classification performance, with an increase of approximately 2.13% to 3.06% in accuracy score. Thus, the proposed method effectively extracts local information while mitigating the impact of background complexity, enabling the model to learn more robust class-discriminative features. The leaf annotations can be a valuable resource for advanced cassava disease analysis, detection, and classification.

The proposed approach however faces challenges due to class imbalance in the Cassava 2020 dataset, where larger classes achieve better predictions and smaller classes yield poorer results. Hence, in future work, data augmentation can be applied to the Cassava 2020 images to mitigate the issue of data imbalance, which currently favors majority classes.

ACKNOWLEDGMENT

This work is supported by the Ministry of Higher Education, Malaysia through the Fundamental Research Grant Scheme (FRGS) FRGS/1/2021/ICT06/MMU/03/2.

REFERENCES

[1] D. Tilman, C. Balzer, J. Hill, and B. L. Befort, "Global food demand and the sustainable intensification of agriculture," Proceedings of the national academy of sciences, vol. 108, no. 50, pp. 20260–20264, 2011.

[2] H. Mishra and D. Mishra, "Artificial intelligence and machine learning in agriculture: Transforming farming systems," Research Trends in Agriculture Science, vol. 1, pp. 1–16, 2023.

[3] F. Faithpraise, P. Birch, R. Young, J. Obu, B. Faithpraise, and C. Chatwin, "Automatic plant pest detection and recognition using k-means clustering algorithm and correspondence filters," 2013.

[4] M. Ebrahimi, M. H. Khoshtaghaza, S. Minaei, and B. Jamshidi, "Vision-based pest detection based on svm classification method," Computers and Electronics in Agriculture, vol. 137, pp. 52–58, 2017.

[5] A. Ojha and V. Kumar, "Image classification of ornamental plants leaf using machine learning algorithms," in 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 834–840, 2022.

[6] S. M. Javidan, A. Banakar, K. A. Vakilian, and Y. Ampatzidis, "Diagnosis of grape leaf diseases using automatic k-means clustering and machine learning," Smart Agricultural Technology, vol. 3, p. 100081, 2023.

[7] L. Li, S. Zhang, and B. Wang, "Plant disease detection and classification by deep learning—a review," IEEE Access, vol. 9, pp. 56683–56698, 2021.

[8] E. Mwebaze, T. Gebru, A. Frome, S. Nsumba, and J. Tusubira, "icassava 2019 fine-grained visual categorization challenge," arXiv preprint arXiv:1908.02900, 2019.

[9] R. Surya and E. Gautama, "Cassava leaf disease detection using convolutional neural networks," in 2020 6th international conference on science in information technology (ICSITech), pp. 97–102, IEEE, 2020.

[10] V. Ravi, V. Acharya, and T. Pham, "Attention deep learning-based large-scale learning classifier for cassava leaf disease classification," Expert Systems, vol. 39, 11 2021.

[11] G. Sambasivam and G. D. Opiyo, "A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks," Egyptian Informatics Journal, vol. 22, no. 1, pp. 27–34, 2021.

[12] U. K. Lilhore, A. L. Imoize, C.-C. Lee, S. Simaiya, S. K. Pani, N. Goyal, A. Kumar, and C.-T. Li, "Enhanced convolutional neural network model for cassava leaf disease identification and classification," Mathematics, vol. 10, no. 4, p. 580, 2022.

[13] E. Mwebaze, J. Mostipak, Joyce, J. Elliott, and S. Dane, "Cassava leaf disease classification," 2020.

[14] A. Maryum, M. U. Akram, and A. A. Salam, "Cassava leaf disease classification using deep neural networks," in 2021 IEEE 18th international conference on smart communities: improving quality of life using ICT, IoT and AI (HONET), pp. 32–37, IEEE, 2021.

[15] H. Zhang, Y. Xu, and J. Sun, "Detection of cassava leaf diseases using self-supervised learning," in 2021 2nd International Conference on Computer Science and Management Technology (ICCSMT), pp. 120–123, IEEE, 2021.

[16] L. Zhuang, "Deep-learning-based diagnosis of cassava leaf diseases using vision transformer," in Proceedings of the 2021 4th Artificial Intelligence and Cloud Computing Conference, pp. 74–79, 2021.

[17] R. Singh, A. Sharma, N. Sharma, and R. Gupta, "Automatic detection of cassava leaf disease using transfer learning model," in 2022 6th International Conference on Electronics, Communication and Aerospace Technology, pp. 1135–1142, IEEE, 2022.

[18] H.-T. Thai, K.-H. Le, and N. L.-T. Nguyen, "Formerleaf: An efficient vision transformer for cassava leaf disease detection," Computers and Electronics in Agriculture, vol. 204, p. 107518, 2023.

[19] Tzutalin, "Labelimg." Free Software: MIT License, 2015.

[20] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "Revisiting rcnn: On awakening the classification power of faster rcnn," in Proceedings of the European conference on computer vision (ECCV), pp. 453–468, 2018.

[21] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10781–10790, 2020.

[22] Y.-C. Chiu, C.-Y. Tsai, M.-D. Ruan, G.-Y. Shen, and T.-T. Lee, "Mobilenet-ssdv2: An improved object detection model for embedded systems," in 2020 International conference on system science and engineering (ICSSE), pp. 1–5, IEEE, 2020.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708, 2017.

[24] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International conference on machine learning, pp. 6105–6114, PMLR, 2019.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, 2009.

[28] Y. Zhong, B. Huang, and C. Tang, "Classification of cassava leaf disease based on a non-balanced dataset using transformer-embedded resnet," Agriculture, vol. 12, p. 1360, 09 2022.

Audio Similarity Detection

Siddharth Harsh Y. Malhotra* and Sapan H. Mankad*
* Nirma University, Ahmedabad, India
23mce021@nirmauni.ac.in
* Nirma University, Ahmedabad, India
sapanmankad@nirmauni.ac.in

***Abstract*—Audio similarity detection plays a crucial role in various applications, such as music recommendations, speaker recognition, copyright infringement detection, and content-based retrieval. Various categories of similarity types such as Audio signal similarity, audio content similarity, audio perceptual similarity, speaker voice similarity, and audio caption similarity are covered by the researchers to identify similarity in audio signals. We have performed various experiments to investigate the appropriate audio embeddings for this task and to identify which feature and distance metrics are more appropriate for identifying speaker and content similarity. Various short-term spectral features have been extracted from the audio signal for feature extraction tasks which are then used for calculating a similarity matrix using Euclidean similarity and Cosine similarity. The performance of the fused features is better than the standalone features. Also, with the use of pre-trained features which are extracted from YAMNet, the model's performance has been improved.**

***Index Terms*—Audio similarity, Speaker-speaker similarity, Euclidean distance, Cosine similarity**

I. Introduction

Audio similarity detection is the process of identifying and quantifying the similarities between two or more audio signals. It can be applied in a variety of fields, including voice casting, speaker recognition, audio categorization, audio information retrieval, and audio captioning. Various characteristics of the audio signals, such as the signal waveform, the content or meaning, the perceptual quality, the speaker identity, or the caption text, can be used to determine audio similarity. Audio similarity can be measured and assessed using a variety of techniques and metrics, depending on the purpose and the data at hand.

Finding reliable and significant features that capture relevant aspects of similarity in audio signals is one of the challenges in audio similarity detection.

Algorithms that utilise deep learning have been proposed to automatically learn audio properties from data in order to get over such limitations. Multiple layers of nonlinear transformations make up deep learning models, which are able to develop intricate and sophisticated representations of the audio signals. These representations can serve as compact, fixed-length vectors called embeddings, which condense the properties of the audio signals. Simple metrics of similarity or distance, like contrastive loss, cosine similarity, or Euclidean distance, can be used to compare embeddings. Without using intermediary embeddings, deep learning models can also be trained end-to-end to directly optimise the similarity metric.

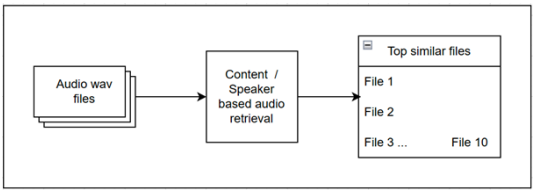


Fig. 1. Audio Similarity Model

Since the similarity of audio can be based on a lot of features such as speakers' voice, similar content, similar audio signals, perceptual similarity, similar melody, rhythm, beats in case of music, a lot of features extraction methods and features have been proposed by the researchers. The main objective of this study is to find which features are more suited to find which type of similarity in audio signals. In this study, we have experimented to find which features and distance metrics for calculating similarity are suited to find audio speaker and audio content similarity.

II. Related Works

Purwins et al. [9] provide an overview of the various deep learning methods for audio signal processing, considering the current explosion in deep learning research and development. The domains of speech, music, and environmental sound processing are compared side by side to identify similarities and distinctions, as well as to highlight common approaches, issues, important sources, and opportunities. A review is conducted on the popular feature representations used including log-mel spectra and raw waveform, as well as other deep learning methods, such as convolutional neural networks (CNNs), long short-term memory (LSTM) architecture variants, and more audio-specific neural network models. Along with that, well-known deep learning application domains using audio similarity are discussed including audio recognition (autonomous speech recognition, music information retrieval, ambient sound detection, localization, and tracking) and synthesis and transformation (source separation, audio improvement, generative analysis). The review discussed various methods for appropriate feature representation including mel frequency cepstral coefficients (MFCC), log-mel spectrum and spectograms along with their advantages. Various models such as CNNs, Recurrent Neural Networks (RNNs), Sequence-to-Sequence Models and Generative Adversarial Networks

TABLE I
LITERATURE SUMMARY TABLE

Author	Year	Title	Key methodology used
Y. Peng et al. [1]	2006	Audio similarity measure by graph modeling and matching	graph modeling and matching segment-based similarity
J. H. Jensen et al. [2]	2007	A framework for analysis of music similarity measures	Mel frequency cepstral coefficient Multiple signal classification
K. Seyerlehner et al. [3]	2008	Frame level audio similarity-a codebook approach	multi-level vector quantization approach
L. Lu and A. Hanjalic [4]	2008	Unsupervised anchor space generation for similarity measurement of general audio	Spectral clustering anchor space
X. Yu et al. [5]	2010	Audio similarity measure based on Renyi's quadratic entropy	Renyi's Quadratic Entropy Mel Frequency Cepstral Coefficients
Q. Wu et al. [6]	2012	Perceptual similarity between audio clips and feature selection for its measurement	acoustic features perceptual similarity
M. K. Singh et al. [7]	2019	Speaker's Voice Characteristics and Similarity Measurement using Euclidean Distances	Mel-frequency cepstral coefficient Euclidean distance
A. Gress et al. [8]	2019	Similarity Metric Based on Siamese Neural Networks for Voice Casting	Siamese Neural Networks
H. Purwins et al. [9]	2019	Deep Learning for Audio Signal Processing	Deep learning Computational modeling
P. Manocha et al. [10]	2021	CDPAM: Contrastive Learning for Perceptual Audio Similarity	Contrastive Deep Perceptual Audio Metric
P. Avgoustinakis et al. [11]	2021	Audio-based Near-Duplicate Video Retrieval with Audio Similarity Learning	Convolutional neural networks spectrogram
Z. Li et al. [12]	2021	Audio similarity detection algorithm based on Siamese LSTM network	Siamese LSTM network
W. Wang et al. [13]	2022	Similarity Measurement of Segment-Level Speaker Embeddings in Speaker Diarization	neural-network-based similarity segmental pooling strategy
S. Bhosale et al. [14]	2023	A Novel Metric For Evaluating Audio Caption Similarity	Text-to-Audio Grounding s2v embeddings

(GANs) are discussed along with appropriate loss functions to be used with them and phase modeling for the audio signal similarity detection.

Peng et al. in [1] suggest a novel method for ranking audio clips similarity using graph modelling and matching. A new approach, Segment-based similarity is suggested as a substitute for frame-based or salient-based criteria in assessing the acoustical similarity of audio clips. This approach uses segment-based representation, then it uses a similarity measure and ranking system based on four different types of similarity characteristics. Segments retain and display the temporal order and change relationship between audio components in addition to capturing the changing properties of the audio clip in segment-based representation. There are four types of similarity factors shared which are acoustical, granularity, temporal order and interference which are measured by optimal matching and dynamic programming progressively and jointly. A bipartite graph computed by Optimal matching method has been extended and used for similarity of audio with dynamic programming. After calculation of interference factor, Audio clip similarity is measured jointly by the degree of acoustical and granularity similarity, the temporal order of matching, and interference factor.

P. Manocha et al. [10] introduce the Contrastive Deep Perceptual Audio Metric (CDPAM), a revolutionary deep learning-based approach to voice processing techniques. This method improves generalisation to a wider range of audio disturbances by gathering human opinions on triplet comparisons. It has been demonstrated that CDPAM correlates well with human reactions in nine different datasets. The CDPAM method uses limited human-labeled data to supplement it with multi-dimensional, self-supervised learning. The training architecture shared in the paper is as follows. First an audio encoder is trained using contrastive learning, then trained the loss-net on JND data. Finally, the loss-net is fine tuned on the newly collected dataset of triplet comparisons.

K. Seyerlehner et al. in [3] present a novel multi-level vector quantization (ML-VQ) approach for modelling audio content based on local spectral properties in this research. By implying a normed vector space, the suggested histogram intersection distance makes it possible to use more potent search techniques and to enable content-based music retrieval for even bigger music archives. This paper also indicates that the generated acoustic similarity space does not suffer from the hub problem that affects Gaussian Mixture Models of songs.

P. Avgoustinakis et al. in [11] propose a unique method for audio-based near-duplicate video retrieval called Audio Similarity Learning (AuSiL). By extracting representative audio-based video descriptors by transfer learning from a CNN trained on a large dataset of audio events, the method identifies temporal patterns of audio similarity between video pairings. This CNN network extracts the temporal structures

in its content is fed the similarity matrix that is created by comparing these descriptors pairwise. After generating triplets, the network is trained by optimising the triplet loss function. Its model is as follows. Each video's spectrogram is fed into the feature extraction procedure, which pulls out the feature vectors for every audio clip. Next, the dot product between the feature vectors of the two videos is used to create a similarity matrix. To capture the temporal patterns of the segment-level within-video similarities, the resulting matrix is sent to Audio Similarity CNN. Chamfer Similarity is used to aggregate the final similarity score.

X. Yu et al. [5] present an innovative method for measuring audio similarity based on Renyi's quadratic entropy. Since, noise interference occurs during audio processing, standard distance metrics are not reliable for determining audio similarity. To represent each audio in their method, MFCCs are extracted. Subsequently, the probability density function (pdf) of MFCCs is computed using the entropy of audio samples, which can be approximated using the Parzen window. The main process in this paper's application of Renyi's Quadratic Entropy include feature extraction, similarity measure and smoothening parameter. Their results show that the proposed algorithm exceeds the Euclidean distance in accuracy.

Q. Wu et al. in [6] explore the retrieval of perceptually similar audio, focusing on finding sounds according to human perceptions which investigates the retrieval of perceptually comparable audio. Comparatively speaking, this method is more "human-centered" than earlier audio retrievals that sought to identify similar sounds. The authors measure perceptual similarity by utilising a wide range of auditory features. Their results show that Line Spectral Pairs (LSP), MFCC, and Perceptual Linear Prediction (PLP) are the three most effective acoustic features. Moreover, the optimal combination of features can improve the accuracy of similarity classification compared with the best performance of a single acoustic feature.

M. K. Singh et al. in [7] used the MFCC and the Euclidean distance to measure the similarity between speakers in their work. This approach is used in forensic statistics for voice similarity measurement. Euclidean distance is used to quantify the speaker voice similarity between two distinct sets of speakers: same speakers and dissimilar speakers. The voice of the speaker has a significant impact on the speech feature coefficient. When calculating speaker similarity, Euclidean distance is derived from the MFCC mean. The similarity index between same speakers and dissimilar speakers is found using the Euclidean distance difference. Their results show that the Euclidean distance is minimum for the same speakers who speak the same text or message and maximum when different speakers speak the same text or message.

A. Gresse et al. [8] discussed a novel approach to voice casting, a process in dubbing where the original voice in a source language is replaced by a new one in a target language. The authors have proposed a Siamese neural networks-based approach to model a voice similarity metric that captures all voice characteristics, including the observers' receptive interests. The paper hypothesizes that pairwise relationships between different voices can be used to learn a similarity metric for professional acted voices. The experimental results show that Siamese networks can model this abstract notion of similarity and provide better generalization results on unseen voices compared to classic architectures. The learnt metric can discriminate target and non-target pairs on new speakers/characters, highlighting the abstract information sought.

J. H. Jensen et al. in [4] covered a technique for utilising an anchor space to calculate similarity. Spectral clustering is used in this method to classify audio recordings into semantic groups based on shared low-level properties. Using the data corpus, the spectral clustering method first creates an affine matrix that compares the similarity of each pair of data points, in this example, one-second audio segments. After extracting the eigenvectors, the original data is mapped into a low-dimensional space that is simple to cluster using a Singular Value Decomposition (SVD). The authors suggest a simplification technique in which the mean vector of each audio document's feature vectors is used to represent it in order to lessen this. After that, the collection of mean vectors calculated for the complete data set is subjected to spectral clustering, and the resulting clusters are chosen to serve as anchors. The study also covers an estimating method based on the eigen-gap that spectral clustering suggests for figuring out how many clusters need to build.

J. H. Jensen [2] presented a MIDI-based test framework for analysing music similarity measures in their study. The framework is used to investigate how sensitive a music similarity measure is to transpositions and how dependent it is on a song's instrumentation as opposed to its melody. Three software programmes are analysed for their music similarity measures: Marsyas, MA toolbox, and Intelligent Sound Processing toolbox. Although the investigated timbral similarity measures are sensitive to transpositions and variations in sound fonts, they demonstrate good performance in instrument recognition. Sometimes the same melody played on different instruments is not recognised by the beat/rhythm/melody similarity metrics. According to the study, timbral similarity metrics did not translate well to other sound typefaces. In cases where certain smooth spectral adjustments, like changing the bass and treble, do not significantly impact the impression of timbre, it is hypothesised that timbral similarity measurements that additionally rely on the temporal envelope will better reflect human sound perception. The study also points out that melody recognition still needs work. Only a genre categorization experiment could not have produced the results that were obtained. The influence of melody and instrumentation have been successfully separated through the use of MIDI files, and a signal change called transposition has been included. The study makes the case that measuring how similarity metrics are affected by pace, instrument combinations, bandwidth, and audio compression. Various feature sets are tested for that in this paper which included Timbre, MFCC, Beat, Pitch, Sone,

Spectrum histogram, Periodicity histogram, fluctuation pattern and multivariate autoregressive model.

Z. Li and P. Song [12] proposed an algorithm for detecting audio similarity based on a Siamese LSTM network. The feature matching model and the selection of audio signal features are the main technologies. LSTM is used in the Siamese network's fundamental network segment as part of the approach. In order to compute audio similarity, the method starts with the extraction of filter banks characteristics from two audio signals. In order to compare the similarity of two samples, the Siamese LSTM network model outputs their high-dimensional space representation. The network is made up of two neural networks that share weights and have the same structure. The FBank feature of each audio segment is used as input, processed by LSTM, and then coupled to a multi-layer complete connection, replacing the base network of the Siamese neural network. In audio similarity calculations based on deep learning, the paper proves the usefulness of the FBank feature and claims that it outperforms the MFCC feature. The Siamese LSTM network and the FBank feature work together to precisely determine how similar two audio portions are.

S. Bhosale [14] presented in this study a new measure in order to assess Automatic Audio Captioning (AAC), a task that requires summarising an audio sample in natural language text. Acoustic semantics is needed for AAC in order to map natural language text to analogous sounds, in contrast to other natural language activities that use lexical semantic metrics for evaluation. The authors include these acoustic semantics into a unique metric that is based on Text-to-Audio Grounding (TAG). They point out that although tasks involving natural language, such as language translation and picture captioning, are anticipated to yield outputs that are semantically comparable, AAC activities do not follow this pattern. As an example, the terms "clock" and "car-turn indicator," which are semantically distant in the Word Embedding (WE) space, may sound identical and appear interchangeably in an audio caption. The research suggests a novel embedding space based on TAG that can map words that produce comparable sounds next to one another in order to fix this. The creation of an acoustic embedding (s2v) that produces similar embeddings for text corresponding to acoustically similar noises is the primary contribution. This metric is composed of two modules: one for Phrase Extraction (PE), which extracts phrases from text; and another for TAG, which creates s2v embeddings for each phrase. The trials show that the suggested evaluation metric outperforms other metrics that are currently in use in the research on AAC for natural language text and picture captioning.

W. Wang et al. [13] presented a Neural-network-based approach to measuring similarity between any two speaker embeddings that takes into account both past and future contexts. For similarity measurement, a segmental pooling technique is suggested, which gives the embeddings more context and hence better performance. To further improve performance, the speaker embedding network's parameters are unfrozen and trained in conjunction with the similarity measurement goal. This joint training approach provides a generic framework for segment-level target-speaker voice activity detection (TS-VAD) and similarity measuring. The authors extended the system to TS-VAD and investigate the impact of various pooling sizes. The findings demonstrate that the MISS in the Diarization Error Rate (DER) is considerably decreased by the TS-VAD approach.

III. Proposed Approach

In this work, we aim to investigate impact of different audio embeddings in evaluating the speaker-speaker similarity and content similarity. To evaluate our proposed approach, we use a subset of ASVspoof 2017 v 2.0 dataset which includes only genuine speech data.

In Fig. 2, a general architecture of the proposed system with feature extraction module and similarity calculation module is shown. First of all various short term spectral features are extracted from audio wav files. On these extracted features, similarity is calculated with distance metrics- Euclidean distance and Cosine similarity. The top similar files are returned as output to the user. The performance of these features is noted down.

In the next approach, the YAMNet model has been used to extract deep features from audio wav files. On these extracted features, similarity is calculated with distance metrics- Euclidean distance and Cosine similarity. The top similar files are returned as output to the user. The performance of these deep features is compared with that of the previous approach. If both the models performance is not satisfactory, then both of the features can be combined and then used to calculate the similarity of the audio wav files.

1) Extract features from each wav file and represent it in form of 13- dimensional feature vector. We have used various short-term spectral feature representations including MFCC. Total 3565 audio files are part of the study.
2) Compute Euclidean distance between each feature vector. This gives a 3565x3565 distance matrix. Further looking into and analysing this matrix shows us some insights.
    a) We found top 10 closest audio pairs for each feature.

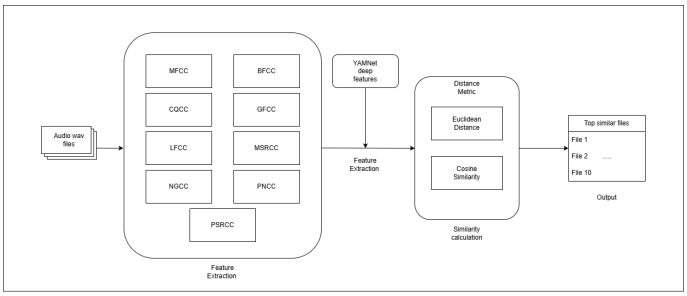


Fig. 2. Architecture of audio similarity detection system

b) We also found which two speakers are the most similar as per all the features.
c) Observing them showed that most of the features work well on speaker similarity rather than content similarity.

3) We repeat the same experiment with Cosine similarity based approach.

Following is the algorithm for generating feature matrix from the audio files.

Algorithm 1 Feature Extraction

```
Set file path
Require features not to be null
procedure FeatureExtraction
    for each metric do
        Take one feature
        Generate feature matrix
        Find top most similar files based on feature matrix
    end for
end procedure
```

IV. Experimental Results and Discussion

A. Dataset

To test the effectiveness of our proposed approach, we used ASVSpoof 2017 version 2.0 dataset [15]. This dataset comprises bonafide and spoofed audio samples from ten differnt speakers and phrased through repeated trials under various recording, and playback conditions.

There are 10 common phrases which are spoken by 42 individual speakers. Since, we have details about content spoken and speaker in the dataset, we can use that to identify which features are able to identify similar content or similar speaker voice.

We focus on examining speaker-speaker similarity and speech-content similarity through these experiments.

B. Features

We use short-term spectral features to represent the input. We use them as standalone first, and then we see the impact of fused features for the underlying task. In addition, pre-trained features are extracted from wav files with the help of the deep YAMNet model [16]. It is a pre-trained deep learning learning model which is trained on AudioSet corpus and can predict around 521 audio events. This model is available on TensorFlow Hub and can give three outputs which are class scores, embeddings and Log Mel Spectrograms.

C. Experimental Setup

To compare the performance of approaches firstly, Short term spectral features are extracted from 3565 audio wav files. These are used to find similar audio files based on content and speaker.

Then, YAMNet model is used to extract deep features from the same audio wav files. These also are then used to find similar audio files based on content and speaker. Finally, the short term spectral features and YAMNet features are then combined and used to find the similarity of the audio files using Euclidean distance and cosine similarity.

D. Evaluation Metrics and Results

To compare the performance of each approach in finding audio similarity, Top 1 and Top 5 metric has been used. To calculate similarity of audio files, for each approach they are sorted according to cosine similarity in descending order and similarly to calculate similarity using Euclidean distance, they are sorted in ascending order.

From this, the top 10 similar files are selected for each file and sorted according to similarity from highest to lowest ensuring that the most similar files appear first in list. This list is then used to compute Top 1 and Top 5 metric where top 1 metric shows how often the model's prediction is correct, while top 5 metric shows how often the file with similar speaker is among the top 5 predictions.

In Table II, the top 1 and top 5 metric for short term spectral features, YAMNet features and Merged features have been shown using both Euclidean distance and cosine similarity as the evaluation metric.

TABLE II
Comparison of Top 1 and Top 5 Metric using each approach for speaker similarity

Features	Euclidean (Top 1)	Euclidean (Top 5)	Cosine (Top 1)	Cosine (Top 5)
BFCC	0.62	0.83	0.60	0.81
CQCC	0.28	0.55	0.20	0.46
GFCC	0.74	0.90	0.73	0.88
LFCC	0.64	0.84	0.59	0.79
MFCC	0.67	0.86	0.63	0.83
MSRCC	0.61	0.82	0.55	0.78
NGCC	0.67	0.86	0.62	0.83
PNCC	0.70	0.88	0.69	0.87
PSRCC	0.40	0.62	0.37	0.60
YAMNet	0.51	0.79	0.58	0.84
Spectral Merged	0.83	0.94	0.89	0.97

E. Discussion

After getting top similar files for each file with previously mentioned distance metrics, the following observations were noted:

- When using Euclidean distance as distance metric, with all of the feature extraction techniques it was able to identify same speaker but not content/common phrases spoken.
- When using cosine similarity as distance metric, it was able to identify similar content in some of the wav files with features extraction techniques- MFCC, CQCC, MSRCC and NGCC.
- With CQCC feature extraction technique, it was able to identify same content in top similar files.

- For both distance metrics and all feature extraction techniques, it was able to find the same speakers in top most similar files.
- With the use of YAMNet deep features, the model is able to find more audio wav files with the similar speakers.
- The combined approach using both short term spectral features and YAMNet features showed significant improvement in finding audio wav files with similar speakers which can be seen with Top 1 and Top 5 metric results in Table II.

V. Conclusion and Future Work

The field of audio similarity detection is broad and diverse, and a lot of cutting-edge techniques are being created. These techniques, which have many uses, including as speaker identification and music retrieval, are always developing in tandem with advances in computational techniques and technology.

In this work, a study was done specifically to measure how effective these features are in identifying similar speaker voice and content similarity. It was concluded that with the distance metrics of Euclidean distance and Cosine similarity, these features identified similar speaker voice, but could not identify content similarity significantly. It was also concluded that deep learning models are more effective in capturing similarity as when also including features extracted from YAMNet to current model, performance improved in comparison to using short term spectral features.

In future work, this audio similarity detection can be extended to work on a broader range of audio types beyond speaker identification such as music signals, environmental sounds and other types of audio. More approaches for distance calculation will be examined to capture diverse characteristics specific to different types of audio could further lead to detect similarity in a wide audio variety, meaning more precise and better audio retrieval across multiple use case scenarios.

References

[1] Y. Peng, C.-W. Ngo, C. Fang, X. Chen, and J. Xiao, "Audio similarity measure by graph modeling and matching," in *Proceedings of the 14th ACM International Conference on Multimedia*, ser. MM '06, Santa Barbara, CA, USA: Association for Computing Machinery, 2006, pp. 603–606, ISBN: 1595934472.

[2] J. H. Jensen, M. G. Christensen, and S. H. Jensen, "A framework for analysis of music similarity measures," in *2007 15th European Signal Processing Conference*, IEEE, 2007, pp. 926–930.

[3] K. Seyerlehner, G. Widmer, and P. Knees, "Frame level audio similarity-a codebook approach," in *Proc. of the 11th Int. Conf. on Digital Audio Effects (DAFx-08)*, 2008, p. 31.

[4] L. Lu and A. Hanjalic, "Unsupervised anchor space generation for similarity measurement of general audio," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 53–56.

[5] X. Yu, X. Pan, W. Yang, W. Wan, and J. Zhang, "Audio similarity measure based on renyi's quadratic entropy," in *2010 International Conference on Audio, Language and Image Processing*, 2010, pp. 722–726.

[6] Q. Wu, X. Zhang, P. Lv, and J. Wu, "Perceptual similarity between audio clips and feature selection for its measurement," in *2012 8th International Symposium on Chinese Spoken Language Processing*, 2012, pp. 387–391.

[7] M. K. Singh, N. Singh, and A. K. Singh, "Speaker's voice characteristics and similarity measurement using euclidean distances," in *2019 International Conference on Signal Processing and Communication (ICSC)*, 2019, pp. 317–322.

[8] A. Gresse, M. Quillot, R. Dufour, V. Labatut, and J.-F. Bonastre, "Similarity metric based on siamese neural networks for voice casting," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6585–6589.

[9] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[10] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "Cdpam: Contrastive learning for perceptual audio similarity," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 196–200.

[11] P. Avgoustinakis, G. Kordopatis-Zilos, S. Papadopoulos, A. L. Symeonidis, and I. Kompatsiaris, "Audio-based near-duplicate video retrieval with audio similarity learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 5828–5835.

[12] Z. Li and P. Song, "Audio similarity detection algorithm based on siamese lstm network," in *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, 2021, pp. 182–186.

[13] W. Wang, Q. Lin, D. Cai, and M. Li, "Similarity measurement of segment-level speaker embeddings in speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2645–2658, 2022.

[14] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "A novel metric for evaluating audio caption similarity," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[15] T. Kinnunen, M. Sahidullah, E. Héctor Delgado, *et al.*, "The 2nd automatic speaker verification spoofing and countermeasures challenge (asvspoof 2017) database, version 2," 2018.

[16] C. Malmberg, *Real-time audio classification onan edge device: Using yamnet and tensorflow lite*, 2021.

Development and Evaluation of a Semi-autonomous Parallel Attentive Listening System

Divesh Lala, Koji Inoue, Haruki Kawai, Zi Haur Pang, Mikey Elmers and Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
E-mail: [lala][inoue][pang][elmers][kawahara]@sap.ist.i.kyoto-u.ac.jp

***Abstract*—Open domain spoken dialogue systems are still not at the level of human performance and understanding. In this work we propose that a remote human operator who can take over the conversation from an autonomous system would improve the quality of an attentive listening conversation. Furthermore this operator could also manage multiple users simultaneously. We describe and implement this as a semi-autonomous parallel system. Features of this system are detection of disengagement to let the operator know when to intervene, summarization of conversations using ChatGPT to allow the operator to manage multiple users, and conversion of the operator's voice to the agent's to make intervention less abrupt. We conduct an experiment to compare this system to a fully autonomous system and find that it improves performance for enjoyment and empathy.**

I. Introduction

Through the use of large language models (LLMs) and improved automatic speech recognition (ASR) technology, spoken dialogue chatting systems have advanced to become more human-like. However, there is still a clear gap between human-human and human-agent conversation.

Most commercial spoken dialogue systems, such as smart speakers, are primarily task-based. However, chatting systems are yet to reach this level of wide usage. This is arguably because such open-domain systems are more complex and require servicing a variety of individual users. LLM dialogue systems such as ChatGPT are now becoming popular, although using these in natural spoken conversation still requires real-time conversation functions such as human-like turn-taking and backchannelling.

LLM-based approaches are also not immune to making inappropriate responses (hallucinations) that arise as a result of either inaccurate ASR results or the model itself. Many dialogue systems do not understand when they make errors and so continue without acknowledgement. On the other hand, a human can recognize and rectify the situation through an apology or social behavior to express this to the user.

Recently there has been much work investigating how LLMs can show this type of empathy in a dialogue system, with some degree of success [1]–[3]. However LLMs and artificial systems can only show this in a synthetic way without being able to feel emotions. Furthermore, empathetic expression is still difficult to replicate through text-to-speech technologies.

To bridge this gap we propose to combine the conversation skills of humans with an autonomous system to make a semi-autonomous system. In this system a remote human operator can smoothly "take over" the conversation from an agent and provide human-level intelligence and empathy at certain times. This approach has been implemented in previous work for service agents [4], [5], but is redundant for open-domain conversation with only one user.

We therefore consider a schema where multiple users interact with separate spoken chat systems simultaneously. In this case, the semi-autonomous function is used by the operator to freely switch between each human-agent conversation and intervene where necessary. We define this setup with independent users as a parallel system. Fig. 1 shows a general outline of the semi-autonomous parallel approach.

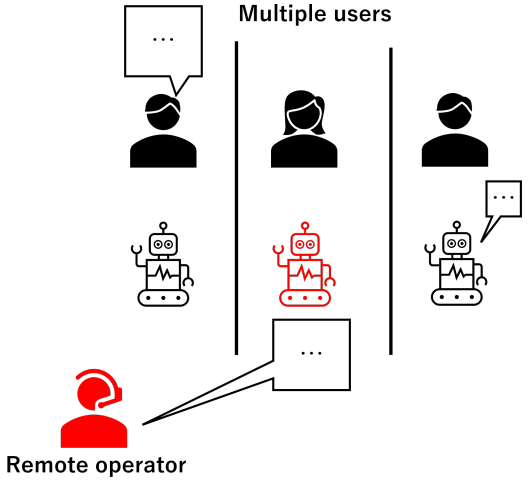


Fig. 1. Overview of the semi-autonomous parallel approach. A remote operator monitors several independent users and can intervene in a conversation, taking over from the autonomous agent.

This system needs to address several challenges such as deciding when the operator should intervene in the conversation and how to allow the operator to immediately understand the conversation in real-time. In this work we focus on attentive listening, an open-domain dialogue system. The semi-autonomous system can be implemented in areas where online agents are used for the benefit of multiple end-users, such as guidance, interviews and teaching.

We expect that a semi-autonomous system can outperform an autonomous system in attentive listening by exploiting a human's conversation skill. Our goal in this work is to implement the system and compare it against a fully autonomous dialogue system. The work in this study is conducted in the

Japanese language.

II. Related work

Fully tele-operated robots have been implemented in various service and therapeutic tasks, with either the operator communicating directly as a human being [6], [7] or attempting to imitate a robot [8], [9]. In these cases one operator is responsible for a single agent and not a parallel implementation.

The development of semi-autonomous systems, where a human can intervene on behalf of an agent, has been achieved with customer service-related tasks [4], [5] and chatting [10], although these are for text-based chatting and not spoken dialogue. A recent work implemented a semi-autonomous system in spoken dialogue but only in a dyadic context [11].

The parallel architecture has also been implemented in a semi-autonomous system with multiple mobile robots [12], although the operator did not directly speak to users, instead they communicated through a limited fixed set of responses. Additionally, [13] previously implemented a semi-autonomous parallel system for a question-answering task. However chatting has much different requirements in terms of handover. To our knowledge this work presents the first semi-autonomous parallel system for an open-domain spoken dialogue task.

III. Autonomous attentive listening system

The dialogue system used in this experiment is an attentive listening system where the agent acts as a listener towards the user's talk, such as a recent enjoyable experience. The objective is for the user to speak to the system and for the conversation to slowly encompass a range of topics. However, if the user becomes disengaged due to inappropriate or monotonous responses from the system they will find it difficult to speak for a lengthy period of time. Furthermore, since the dialogue system does not take initiative from the user it must produce meaningful questions and empathetic responses which stimulate more conversation.

This system uses a state-of-the-art transformer-based Japanese ASR model [14]. Turn-taking and backchannelling are handled by our previously implemented models [15]. Backchannels in response to an emotional state (e.g. "ah!", "oh!") are also generated as the system detects user sentiment.

The system has four types of responses - sentiments, elaborating questions, repeated responses and generic responses, summarized in Fig. 2.

Sentiments are a reaction containing some type of emotion (e.g. "Wow, that's great" (*ii desu ne*) or "That's a shame" (*zannen desu ne*)) and are prioritized over the other types of response. We use a simple lexical model to detect a sentiment.

A. Focus word detection

Our approach requires detecting the focus of the user's talk which we implemented in previous work [15]. For example, if the user says "Today I went to a restaurant and ate pasta", the focus word *pasta* would be detected.

The focus word is also used for repeated responses, which are simply the focus word with *desu ka* added to it. In

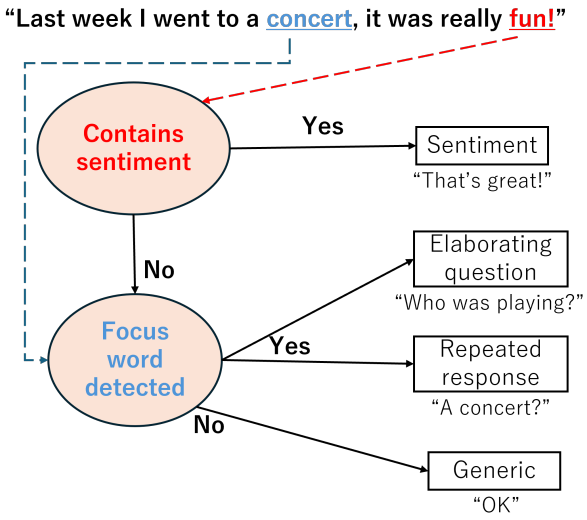


Fig. 2. Overview of the response types used in the attentive listening system.

Japanese, these types of responses act like affirmations or backchannels rather than direct questions. To ensure a variety of responses, the system will maintain a balance between elaborating questions and repeated responses.

Generic responses are used when no focus word can be detected and are simple statements such as "I see" (*sou desune*) or OK (*hai*). Our aim is to keep these to a minimum.

B. Response generation with T5 model

If a focus word is detected we generate an elaborating question using a specialized deep learning model distilled from ChatGPT. We first extracted data from RealPersonaChat [16], a corpus of 14,000 dialogues. The data was statements which preceded a question. These were then used with the following prompt to ChatGPT (translated from Japanese):

"Please respond to the statements with elaborating questions to elicit more conversation. The purpose of the elaborating questions is to get important information about the 5W1Hs: "why","what","where","who","when" and "how".

This prompt produces responses which are more ideal for attentive listening than the longer utterances used by naive ChatGPT. An example of the different responses is:

User input: I ate lunch outside but I had a bad feeling so I came back.

ChatGPT (naive) That is unfortunate. Was something not to your liking? Was there something wrong with the location or the food?

ChatGPT (prompt) Why did you have a bad feeling?

This resulted in training data consisting of around 16,000 statement and response pairs used to fine-tune a T5 model [17]. The model is smaller than a typical LLM (around 220 million parameters), but produces attentive listening-type dialogue. The resulting model is lightweight, produces responses suitable for attentive listening and can be used with a CPU.

The system also uses elaborating questions in situations when the conversation seems to be stalling. When there is a silence from the user of greater than 5 seconds, an elaborating question is uttered using a history of the dialogue. Elaborating questions are also generated when three or more generic responses are generated consecutively by the system in order to preserve response variety.

C. Agent interface

The agent we use in this work is Gene, based on the MMDAgent platform [18], shown in Fig. 3. Gene can use a variety of text-to-speech (TTS) systems. In this case we use a female voice. Gene has lip syncing capabilities and conveys emotional states through facial expressions. When sentiment responses or emotional backchannels are uttered by Gene, they are combined with the appropriate expression.



Fig. 3. Gene, the agent used in this work, with a sad expression.

D. Limitations of autonomous attentive listening

Although this system is robust, there are limitations which could be improved by human intervention. The elaborating questions are generally appropriate but quite limited. If a focus word cannot be found the system reverts to generic responses which do not stimulate more talk and it may become difficult for the user to speak fluently. We also showed that users still found autonomous attentive listening to be lacking in empathy and understanding compared to a tele-operated system [15].

These issues can be alleviated with human intervention. An operator can quickly come up with more probing questions to promote talk, quickly recognize dialogue breakdowns, and assist the user in a similar manner. This motivates our semi-autonomous approach.

IV. Semi-autonomous system

We now describe the main aspects of the semi-autonomous system which are disengagement detection (to determine when the operator should intervene) and dialogue monitoring (so the operator quickly understands the state of the dialogue).

A. Disengagement detection

The first technical challenge is detecting when the system needs human intervention because of dialogue breakdown or disinterest. Although this has been widely studied [19], it is more difficult in attentive listening since the user may try to keep talking even in an uninteresting conversation. In this work we estimate if the user may require intervention by continuously analyzing aspects of the attentive listening dialogue itself.

To approach this problem, we use the set of heuristics described in [11]. These are based on the assumptions that a "good" attentive listening dialogue involves the user talking continuously.

- User is silent for more than five seconds
- User has had two short turns (less than 20 characters) in a row
- System has not uttered anything except generic responses for three turns in a row
- System has not uttered a sentiment response in the last five turns

The first two heuristics estimate the user state in the dialogue. The final two heuristics relate to the performance of the system. The operator is provided information when one or more of the conditions are met, in the form of recommendations to intervene. The operator is not required to act upon these recommendations but they are useful for understanding what type of dialogue intervention may be needed to help maintain the conversation.

B. Dialogue monitoring

The operator must be able to monitor multiple conversations without too much cognitive load. If the operator has to read ASR results it will take too long to "catch up" to the conversation. For a parallel system this is even more critical as the operator has to deal with multiple users without delay.

To approach this problem we use dialogue summarization through ChatGPT as proposed in previous work [10]. The performance of ChatGPT in summarization tasks is comparable to humans [20] and output can be verified in real-time by the operator. We use few-shot prompting for ChatGPT using the following instruction (translated from Japanese):

"Below is part of a conversation between A and B. Summarize A's entire talk, their emotional state and provide simple short questions to ask A. Ignore any unnatural Japanese in the dialogue. Also please ignore B's dialogue. The output should be less than 100 Japanese characters. There are some examples below."

The prompt includes several examples and the most recent dialogue of the conversation from ASR results. The user and agent correspond to speaker A and B respectively. The output format from ChatGPT is in three parts - a short textual summary of the conversation, an estimation of the emotional state of the user, and questions that could be asked to the user. The

intention is that the operator uses these to formulate more in-depth questions with appropriate backchannels and sentiment reactions. Summarization is executed when the operator clicks on a user's video panel.

C. Seamless switching

Another major challenge is to match the voice of the operator and the agent's TTS to prevent an abrupt change when control is switched between them. We use voice conversion [21] to change the operator's voice to one which matches the agent's TTS in real-time. This allows the operator to speak as the agent while preserving the nuances of human speech.

V. Operator interface

Fig. 4 shows the operator interface. The operator can view ASR results and system dialogue in real-time. They select the conversation they would like to listen to by clicking on the video panel. The agent is not displayed. To interact with the user, the operator clicks a microphone button corresponding to the conversation in which they wish to intervene. This button is disabled when the agent is speaking to prevent interference between the operator and agent's voice. During an intervention the operator may also control particular facial expressions of the agent by clicking the corresponding emoji button.

We evaluated the interface in an initial feasibility study with two operators over 10 sessions. We used a 5-point Likert scale with six items and found that all items averaged at least 3. The usability of the interface (4.4) and the ability to talk spontaneously to the users (4.3) were rated the highest.

VI. Experiment methodology

We hypothesize that a semi-autonomous system with one user is more effective than one with multiple users and that summarization improves the system. We conducted experiments under four different conditions:

AUTO The autonomous attentive listening system.
DYADIC A semi-autonomous system where the operator only handled one user with no summarization.
SEMI-BASE A semi-autonomous parallel system where the operator handled three users with no summarization.
SEMI-FULL A semi-autonomous parallel system where the operator handled three users using summarization.

The **DYADIC** condition was implemented in [11] with the same attentive listening system and experimental procedure.

We first conducted an experiment with the **SEMI-BASE** system. There were six sessions with three simultaneous participants. We then conducted a within-subjects experiment with 11 participants who interacted with both the **SEMI-FULL** and **AUTO** systems. For the **SEMI-FULL** system we conducted four sessions. One session had two participants so we used a "dummy" participant, to ensure that the operator would still manage three users. The order of conditions was randomized.

Participants were given an explanation of attentive listening and time to consider what they would talk about. The motivation topic was something interesting that they had experienced recently. We informed participants that their conversations would be listened to by a remote operator who could intervene on behalf of the agent, although they would not know when it would happen. We used two different operators for these conditions. Each session was around eight minutes in duration. Experiments were conducted with both participants and operators in sound-proof cubicles in the same room and the operator cubicle was separated away from the participants so they could not see each other.

Operators in the semi-autonomous conditions were instructed to intervene in conversations by considering information provided by the interface, although they would make the final intervention decision themselves. They were instructed to only use short questions and statements suitable for attentive listening. Note there is no difference between the **SEMI-BASE** and **SEMI-FULL** systems from the user perspective, only the information the operator receives.

VII. Results

After each interaction participants answered a 7-point Likert survey consisting of 19 items based on those found in previous work [15]. We grouped these items to construct measures and used them as the basis of the statistical analysis.

Likert scale results are shown in Fig. 5, taking the average item for each measure. We could not outperform the **DYADIC** condition, which had the highest ratings across all measures, even without summarization. However **SEMI-FULL** is close to the level of the **DYADIC** system, showing that summarization is beneficial when handling multiple users. **SEMI-BASE**, which did not provide any summarization, performs worse than **SEMI-FULL** on all measures.

The **SEMI-FULL** system also outperformed **AUTO**, particularly for empathy and enjoyment. Paired t-test results for these measures showed p-values of 0.07 and 0.08 respectively. These are also the only two measures where the medians of **SEMI-BASE** are higher than that of **AUTO**.

On the other hand, the **SEMI-BASE** system was worse than the **AUTO** system in terms of naturalness and timing. From this result we propose that if the operator cannot provide a quality intervention then the users perceive it as upsetting the flow of the conversation. On the other hand the **SEMI-BASE** system was better in terms of enjoyment and empathy but there is a large amount of variation between users. The average number of interventions per user was around 3, but the number of interventions did not correlate with any measures.

VIII. Discussion

Results of our study indicated that the semi-autonomous approach improved enjoyment and empathy but naturalness and timing were perceived more negatively. One reason for this could be that the items which make up naturalness and timing are arguably more to do with conversational processes as opposed to how users feel about the agent.

The effect of human intervention changes with both the number of simultaneous users and the existence of summarization. Although intervention can be done successfully in a

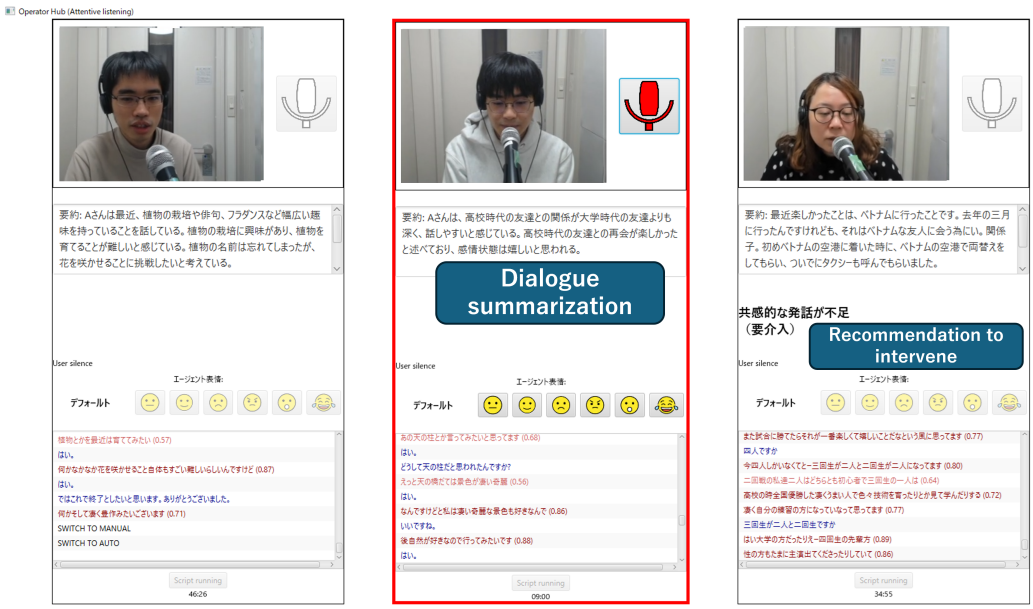


Fig. 4. User interface of the remote operator. The currently selected user is in the middle panel and the operator can speak to them directly. Intervention recommendations and dialogue summarization are labeled.

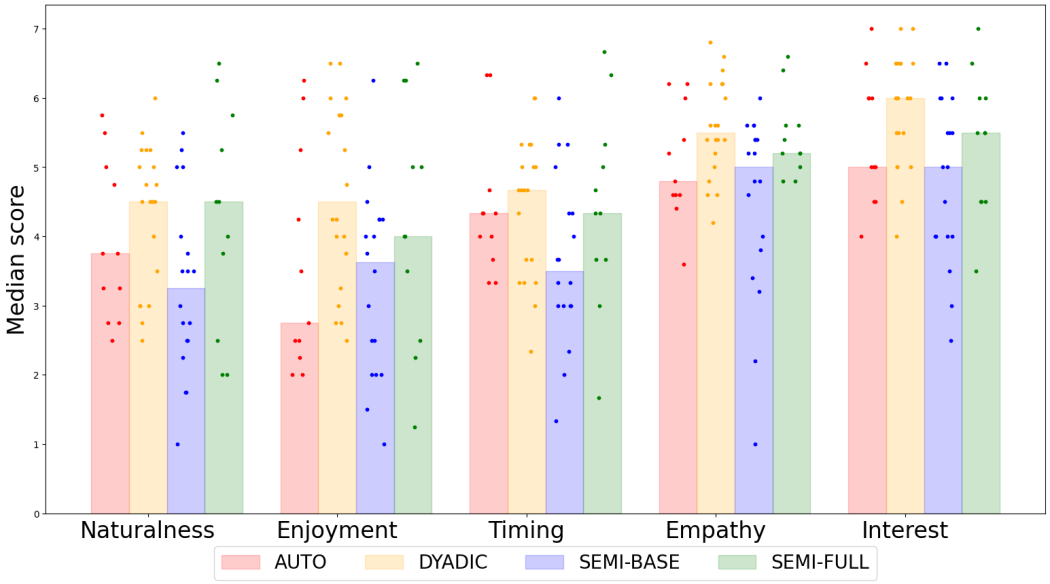


Fig. 5. Results of experiment. Bars indicate median values.

dyadic interaction, for multiple users summarization is necessary to quickly understand the state of the dialogue. Without summarization the *quality* of the intervention is degraded because they cannot successfully stimulate the conversation.

Other strategies which can assist with quality interventions should be considered. Our eventual aim is to at least match the performance of the **DYADIC** system, which we used as a comparative baseline. This would show that one operator can service the needs of multiple people simultaneously at the same level as a single user. Limitations of this work were the relatively low number of participants in the study and the quality of the voice conversion.

IX. Conclusion

In this work we introduced a semi-autonomous parallel system for an attentive listening spoken dialogue system. A remote operator intervenes and takes control of a human-agent conversation to help maintain user interest. The operator can manage several of these dialogues simultaneously through our system, by using disengagement detection and dialogue summarization. Our user study showed that it could improve user perception of enjoyment and empathy and that dialogue summarization improved the performance of the system.

ACKNOWLEDGMENT

This work was supported by JST Moonshot R&D Goal 1 Avatar Symbiotic Society Project (JPMJPS2011).

REFERENCES

[1] J. W. Ayers, A. Poliak, M. Dredze, *et al.*, "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum," *JAMA internal medicine*, vol. 183, no. 6, pp. 589–596, 2023.

[2] Z. Elyoseph, D. Hadar-Shoval, K. Asraf, and M. Lvovsky, "ChatGPT outperforms humans in emotional awareness evaluations," *Frontiers in Psychology*, vol. 14, 2023.

[3] Y. Fu, K. Inoue, D. Lala, K. Yamamoto, C. Chu, and T. Kawahara, "Dual variational generative model and auxiliary retrieval for empathetic response generation by conversational robot," *Advanced Robotics*, vol. 37, no. 21, pp. 1406–1418, 2023.

[4] M. Poser, T. Hackbarth, and E. A. Bittner, "Don't throw it over the fence! Toward effective handover from conversational agents to service employees," in *International Conference on Human-Computer Interaction*, Springer, 2022, pp. 531–545.

[5] M. Poser, S. Singh, and E. Bittner, "Hybrid service recovery: Design for seamless inquiry handovers between conversational agents and human service agents," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021, p. 1181.

[6] K. Kuwamura, S. Nishio, and S. Sato, "Can we talk through a robot as if face-to-face? Long-term fieldwork using teleoperated robot for seniors with alzheimer's disease," *Frontiers in psychology*, vol. 7, p. 1066, 2016.

[7] L. Chen, H Sumioka, L Ke, M Shiomi, and L Chen, "Effects of teleoperated humanoid robot application in older adults with neurocognitive disorders in Taiwan: A report of three cases," *Aging Medicine and Healthcare*, vol. 11, pp. 67–71, 2 2020.

[8] J. Baba, S. Sichao, J. Nakanishi, *et al.*, "Teleoperated robot acting autonomous for better customer satisfaction," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 2020, pp. 1–8.

[9] S. Song, J. Baba, J. Nakanishi, Y. Yoshikawa, and H. Ishiguro, "Teleoperated robot sells toothbrush in a shopping mall: A field study," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 2021, pp. 1–6.

[10] S. Yamashita and R. Higashinaka, "Data collection for empirically determining the necessary information for smooth handover in dialogue," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, 2022, pp. 4060–4068.

[11] H. Kawai, D. Lala, K. Inoue, K. Ochi, and T. Kawahara, "Evaluation of a semi-autonomous attentive listening system with takeover prompting," *arXiv preprint arXiv:2402.14863*, 2024.

[12] D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "Teleoperation of multiple social robots," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 3, pp. 530–544, 2012.

[13] Y. Muraki, H. Kawai, K. Yamamoto, K. Inoue, D. Lala, and T. Kawahara, *Semi-autonomous guide agents with simultaneous handling of multiple users*, 2023.

[14] M. Mimura, K. Inoue, T. Kawahara, T. Nakamura, and H. Saruwatari, "Construction of a corpus of spoken Japanese under real-world conditions and its use in speech recognition benchmarks," Tech. Rep. 12, 2023, (in Japanese).

[15] K. Inoue, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara, "An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, O. Pietquin, S. Muresan, V. Chen, *et al.*, Eds., Association for Computational Linguistics, 2020, pp. 118–127.

[16] S. Yamashita, K. Inoue, A. Guo, S. Mochizuki, T. Kawahara, and R. Higashinaka, "RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities," in *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, C.-R. Huang, Y. Harada, J.-B. Kim, *et al.*, Eds., Dec. 2023, pp. 852–861.

[17] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[18] A. Lee, *MMDAgent-EX*, version 1.0.0, 2023. [Online]. Available: https://github.com/mmdagent-ex/MMDAgent-EX.

[19] R. Higashinaka, L. F. D'Haro, B. Abu Shawar, *et al.*, "Overview of the dialogue breakdown detection challenge 4," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, Springer, 2021, pp. 403–417.

[20] J. Wang, Y. Liang, F. Meng, *et al.*, "Is ChatGPT a good NLG evaluator? A preliminary study," *arXiv preprint arXiv:2303.04048*, 2023.

[21] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 5530–5540.

Prediction of Negative User Reactions Towards System Responses During Attentive Listening

Divesh Lala, Koji Inoue and Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
E-mail: [lala][inoue][kawahara]@sap.ist.i.kyoto-u.ac.jp

***Abstract*—Detecting user dissatisfaction with system responses is an important task for situated spoken dialogue systems. In this work we train models to predict negative user reactions which display this dissatisfaction. The dialogue scenario used in this work is attentive listening and we use two distinct corpora for analysis. We take a multimodal approach and use audio, visual and linguistic features in our model. Results show that audio features are the most influential and multimodality improves model performance. The importance of visual and linguistic modalities differ according to the type of corpus. We identify the key features of each corpus and find that a model trained on both corpora with seven key features reaches a precision of 0.562, outperforming the baseline for this task.**

I. Introduction

Spoken dialogue systems have advanced in recent years with more accurate automatic speech recognition (ASR) and large language models (LLMs). However, in a situated environment there is still a need to gauge the reaction of the user towards system utterances, particularly negative reactions. We define negative reactions as user behaviors that indicate some kind of dissatisfaction with the response.

While user reactions to text-based chatbot systems can only be inferred based on text input, situated spoken dialogue systems provide information on the user state through modalities such as the tone of voice and facial expressions. By analyzing this multimodal data, we can potentially create a model which can detect negative reactions in real-time.

During robot conversation, system or dialogue model errors can produce negative reactions, so evaluation models could be used to detect these. However there are other reasons why users may be dissatisfied with the dialogue system, such as if the response is uninteresting as in Fig. 1. In this case there is nothing wrong with the dialogue itself, but the user still appears dissatisfied. Our target phenomena is related to the state of the user, not the performance of the system.

Another contribution of this work is to understand the relative importance of the individual features in our multimodal model. In this work we train models on two separate corpora with different user demographics and analyze feature importance for each of them. We then identify common important features for use in a more generalizable model. We require this type of analysis to understand what makes a reaction "negative" and to act as a basis for discussion and future research directions. The dialogue system and experiments described in this work were carried out in Japanese.

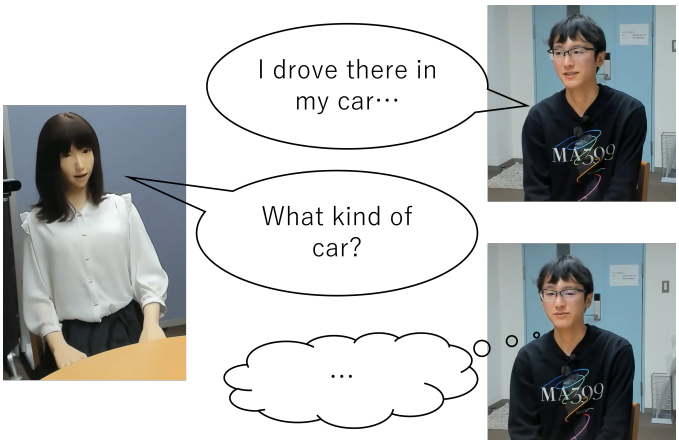


Fig. 1. Target of phenomena of this work. We study the reaction of the user after a system response and predict if they are dissatisfied even if the response is appropriate.

II. Related work

Detection of robot errors through multimodal social signals has been a focus of much previous work [1]–[4] with several of these being applied to collaborative tasks. Other research attempts to classify confusion during interactions [5]–[7] although some are unimodal or are not designed to be used in a real-time system. For spoken interaction, dialogue breakdowns are also a common research topic [8]–[10]. These works often attempt to use linguistic analysis to detect breakdowns, although other modalities have been tested [11], [12]. Other studies aim to detect disengagement [13]–[15] as a means to understand when the user has lost interest.

We use these previous works as grounding for this research, where user reactions are modeled in free-talk conversation. Our work is differentiated by the target phenomena and the scenario. Firstly, we do not aim to detect system errors in the conversation, only the reactions of users. There may be reasons why the user may be dissatisfied with the response for reasons unrelated to the system, that could be described as social errors [16]. Several works also elicit reactions by deliberately inserting system errors into task or conversation scenarios [5], [17], [18]. In our work the reactions in the corpus arise naturally during conversation and are not a result of deliberate manipulation.

Secondly, the nature of free talk is that interaction between the user and system is through dialogue alone, although in our

particular case the user does the majority of talking. Unlike other works [2], [17] there is no set task for participants in the conversation. The success of the conversation is based primarily on the user's perception and their behaviors are varied and in some cases the signals are subtle, arguably making this task more difficult than task-based scenarios.

III. Dialogue System

In this work we use attentive listening as the scenario for the spoken dialogue system. The goal is for the system to act as an empathetic listening partner by listening to the talk of the user and providing meaningful responses. In this work both casual and more serious talks are analyzed as two distinct corpora.

We use our previous attentive listening dialogue system, details of which can be found in previous work [19], [20]. In the attentive listening scenario, the user does the majority of talk while the system uses several types of responses. The general approach is to extract a focus word which is used to generate these responses. For example, if the user is talking about pasta the dialogue system may produce an elaborating question such as "What type of pasta?".

This type of dialogue can produce negative reactions if the dialogue system makes a mistake. There are also situations where the dialogue system produces a coherent response, but it is not engaging enough for the user. The reaction of the user can indicate whether the response is actually satisfactory.

The nature of attentive listening makes detecting reactions different than for task or collaboration-based agent systems. Users often continue with their talk even if the system makes a mistake, but signal their dissatisfaction through audio and facial cues. These dissatisfaction signals, although subtle, may be obvious for humans to understand but pose a challenge for autonomous models.

IV. Data Collection and Analysis

We collected sessions of users engaging with the attentive listening dialogue system. We compare two distinct corpora with differing subjects and topics - elderly and COVID. The first corpus is with elderly subjects and the topic of conversation was casual, such as a recent trip or an interesting episode in their life. The second corpus is comprised of university students, with the topic of conversation being their experiences during the COVID pandemic.

The interface is the android ERICA, a humanoid robot with a custom-built text-to-speech system. Users were instructed that they would be talking with ERICA and were given time to prepare their talk. Each session lasted for approximately 5-10 minutes and there were 56 sessions in total (36 from the elderly corpus and 20 from the COVID corpus).

Two experts viewed video of the corpora and independently annotated each of the adjacency pairs surrounding non-generic system responses as shown in Fig. 2. They then discussed their annotations together to generate the final labels.

We first annotated the appropriateness of the system response in the first adjacency pair to confirm that user reactions are not just correlated with the appropriateness of the actual

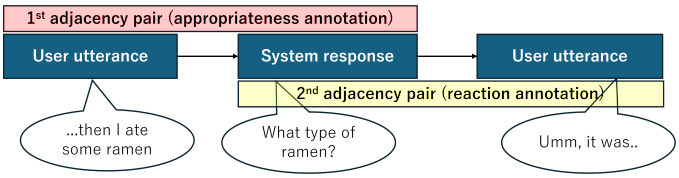


Fig. 2. Overview of annotation targets. Annotators used the 1st adjacency pair to assess appropriateness and the 2nd to evaluate user reaction.

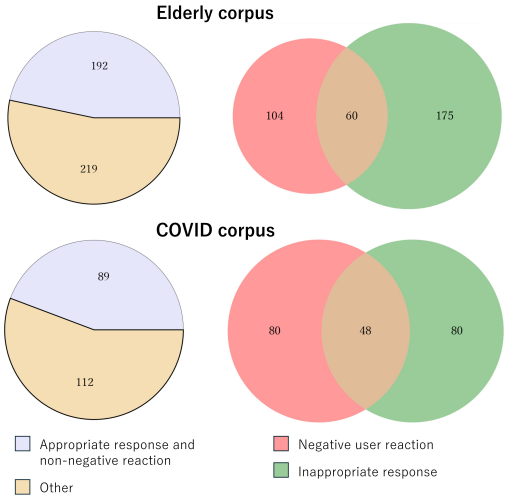


Fig. 3. Annotation results of the two corpora. The left side shows the distribution of system utterances which are both appropriate and followed by a user reaction which is not negative. The right side is the subset of "Other", showing utterances which have either a negative reaction, are inappropriate, or both.

system utterance. For attentive listening an appropriate utterance is a question or assessment which can stimulate more talk from the user. If an elaborating or repeated question was used, it was only deemed appropriate if the focus word was uttered by the user and that there was no other focus word which would have been more appropriate.

Annotation of user reaction is toward the second adjacency pair, without considering the content of the response. Annotators were asked to observe the user's reaction and speech to determine if they were dissatisfied with the preceding response.

The elderly and COVID corpora consists of 411 and 201 non-generic responses respectively. Fig. 3 summarizes the distributions of reactions and appropriateness for non-generic responses. Our analysis shows that there are still a significant number of utterances which are appropriate but are followed by a negative reaction by subjects. We justify our focus on negative reactions as not completely caused by dialogue mistakes and therefore should be studied as a separate phenomena.

V. Data Modalities

In this section we describe the modalities which will be investigated in this work as predictors of negative reactions.

A. Audio features

We extract prosodic pitch and power information using an online pitch tracker [21]. Audio is divided into system and user inter-pausal units (IPUs) with a 200ms segmentation for silence. All IPUs are transcribed and turn-taking was annotated to accurately handle backchannel utterances.

B. Facial action units and head pose

We also used the well-known facial action units (AUs) [22] which have also been used in previous research for error detection [2]. These are a taxonomy of specific movements or features in the face and are extracted and classified for each frame of the video using the OpenFace toolkit [23]. There are two types of output - the presence of an action unit (AUP) and the intensity of an action unit (AUI). AUPs are a binary value while AUIs are scored on a 5-point scale, higher values having more intensity. We also used OpenFace to extract head position and rotation information for each frame.

C. Linguistic features

We also used the capabilities of using automatic evaluation as a tool to predict reactions and appropriateness. For this task we used ChatGPT [24], asking the model to evaluate the system response on a scale of 1 to 10, using the following prompt with the previous user turn and system response:

"A and B are having a short dialogue in Japanese. Rate the appropriateness of B's response to A's talk on a scale of 1 to 10, with 1 being completely inappropriate and 10 being completely appropriate."

We use this numerical rating directly as a feature in the model to capture negative reactions which are a result of inappropriate or incoherent responses.

VI. Models

In this section we describe the construction of several models used in this work and their features sets.

A. Windows of analysis

First we decide the target time window for analysis for audio and visual features, shown in Fig. 4. We denote the beginning and end of the system response as s_0 and s_n and the next user IPU (not the entire turn) as u_0 and u_n. Note that in the case of overlapping turns it is possible that $s_n \geq u_0$.

Audio features are extracted from u_0 to u_n, as they are calculated only for the user's IPU. For the AU model the target window is s_0 to $s_n + 3$ to capture the visual features straight after the system begins its response.

B. Feature selection and extraction

For audio features we used turn switch time and the duration of the next user's utterance to capture hesitations and fillers which may indicate confusion. For prosodic features we extracted the difference in mean and median power of the next user IPU compared to the corresponding values over the entire session. The aim was to capture self-talk, which was found in negative reactions in previous work [5]. The percentage of the IPU that was unvoiced (no pitch detected) was also included to find disfluencies in user speech indicating confusion.

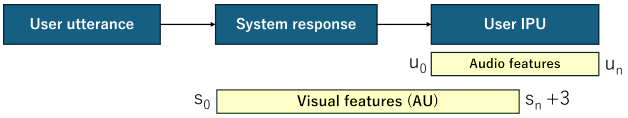


Fig. 4. Target time windows for extracting features.

TABLE I
Feature set used for models.

Feature type	Description and no. of features
Audio (IPU)	turn switch time, IPU length (2)
Audio (prosodic)	mean + median power diff., unvoiced % (3)
Visual (AUs)	AUP percentages, AUI averages (35)
Visual (pose)	Total head position and rotation change (2)
Linguistic	ChatGPT rating (1)

For action unit features, we used the the percentage of presence for each AUP and the average value of the AUIs over the target window. We also calculate the total movement of the head over the window (position and rotation) similar to previous work [5]. Table I provides an overview of the features used. In total there are 43 features.

We also attempted to use streamed audio as an input to fine-tuned transformer models, however these were unsuccessful. Therefore, in this work we focus on the relatively simpler statistical classifiers.

VII. Results

We implemented logistic regression, random forests and XGBoost classifier models using the Python toolkit scikit-learn [25]. Each model was trained using 10-fold cross-validation and 100 trials on the dataset were conducted to account for differences in the folds. We tested different subsets of features: audio, visual and linguistic. Hyper-parameter selection through grid search was used for the random forest and XGBoost models. Default values for random forest were satisfactory but for XGBoost we twe set the hyper-parameters of `learning_rate`= 0.3, `max_depth`= 4, `min_child_weight`= 7 and `colsample_bytree`= 0.5. In this work we focus on improving precision to increase our confidence in a negative reaction prediction.

Models were trained on each corpus. We compare them to a baseline which randomly predicts the user reaction based on the distribution of the corpus. Results for the elderly corpus are shown in Table II. Logistic regression with all features was the best performing in terms of recall and F-score. However random forest and XGBoost models are better for precision. Multimodality improves the audio-only models.

Results for the COVID corpus are displayed in Table III. For this corpus using the ChatGPT evaluation improves the audio-only model, but adding visual features does not.

To understand the contribution of the features, we analyzed the models using Shapley Additive Explanations (SHAPs)[26]

TABLE II
PERFORMANCE FOR MODELS TRAINED ON ELDERLY CORPUS.

Model	Precision	Recall	F-score
Baseline	0.253	0.253	0.253
Logistic Regression			
Audio	0.446	0.558	0.496
Visual	0.340	0.567	0.424
Audio + Visual	0.42	0.606	0.496
Audio + Linguistic	0.433	0.558	0.487
All	0.426	**0.606**	**0.500**
Random Forest			
Audio	0.595	0.423	0.494
Visual	0.563	0.173	0.265
Audio + Visual	0.630	0.327	0.430
Audio + Linguistic	0.568	0.404	0.472
All	**0.633**	0.298	0.405
XGBoost			
Audio	0.547	0.394	0.458
Visual	0.459	0.163	0.241
Audio + Visual	0.569	0.394	0.466
Audio + Linguistic	0.512	0.423	0.463
All	0.577	0.394	0.469

TABLE III
PERFORMANCE FOR MODELS TRAINED ON COVID CORPUS.

Model	Precision	Recall	F-score
Baseline	0.398	0.398	0.398
Logistic Regression			
Audio	0.554	0.575	0.564
Visual	0.361	0.438	0.395
Audio + Visual	0.479	0.562	0.517
Audio + Linguistic	0.540	**0.588**	0.563
All	0.495	0.575	0.532
Random Forest			
Audio	0.533	0.500	0.516
Visual	0.500	0.250	0.333
Audio + Visual	0.510	0.312	0.388
Audio + Linguistic	0.594	0.512	0.550
All	0.480	0.300	0.369
XGBoost			
Audio	0.590	0.488	0.534
Visual	0.490	0.312	0.381
Audio + Visual	0.523	0.425	0.469
Audio + Linguistic	**0.611**	0.550	**0.579**
All	0.531	0.425	0.472

on the XGBoost model which included all the features. We calculated the average SHAP values for every sample over 100 trials. We generated a plot which shows the SHAP value of every sample in a corpus and orders it by the contribution towards the model. The SHAP summary for both corpora are shown in Fig. 5, with only the top 15 features displayed.

For both corpora the feature which contributes the most is turn switching time, with a higher time more likely to be a negative reaction. However many other features are important in one corpus but not the other. For example, the average intensity of AU20 (lip stretcher) is a strong contributor to the elderly model, but not for the COVID model. Only in the COVID corpus is the GPT rating score a contributing feature.

For a combined model we decided to only use the features that were present in both corpora's top 15 SHAP features. This resulted in 7 features as shown in Table IV.

There are two audio features (switch time and utterance duration) and five visual features. The AUs themselves are

TABLE IV
FEATURES IN THE TOP 15 SHAP VALUES FOR BOTH CORPORA.

Feature	Rank (elderly)	Rank (covid)
Turn switch time	1	1
Utterance duration	3	7
AU4 (brow lowerer) average	6	4
AU9 (nose wrinkler) average	5	11
AU5 (upper lid raiser) average	12	8
AU6 (cheek raiser) average	15	12
AU17 (chin raiser) percentage	13	15

TABLE V
PERFORMANCE FOR XGBOOST MODELS TRAINED ON BOTH CORPORA.

Model	Precision	Recall	F-score
Baseline	0.301	0.301	0.301
Audio	0.510	0.402	0.450
Visual	0.547	0.223	0.317
Audio + Visual	**0.575**	**0.418**	**0.484**
Audio + Linguistic	0.511	0.391	0.443
All	0.545	0.397	0.459
Minimal	0.562	**0.418**	0.480

in all different parts of the face: eyebrow, eyelid, cheek, nose and chin. AU intensities over the target window appear to be more influential than just the presence of AUs.

Longer turn switching times are correlated with negative reactions, which is intuitive as users hesitate. However, for utterance duration the opposite pattern is true. This can be explained by the fact that many users acknowledged the system after it responded, with the first IPU being "*hai*". For negative reactions, it was more likely that this acknowledgment was not given as users hesitated before continuing their talk.

For the visual features, negative reactions were associated with a lower intensity of brow lowering (AU4), upper lid raising (AU5), and nose wrinkling (AU9). However, for the intensity of cheek raising (AU6) and the amount of chin raising (AU17) the patterns were different between corpora.

We trained XGBoost models on both corpora and included our minimal model which was trained using only the seven features shown in Table IV. Results are shown in Table V.

For the combined corpus, the model using all audio and visual features is the best performing. However the minimal model also shows similar performance with a fewer number of features. Fig. 6 also shows the generated SHAP summary plot for the minimal model. The order of feature importance of the previous SHAP analyses is maintained, with audio features being the most influential.

VIII. DISCUSSION

Our results showed that a multimodal approach is needed for negative reaction prediction and that the two corpora required different types of models for improved performance over an audio-only model. For elderly people visual features had a large effect, while for the university students ChatGPT evaluations were more useful for predicting negative reactions. This discrepancy could be a result of either the demographics of the subjects or the nature of the talk.

Our SHAP analysis identified a subset of features which could best contribute to the model. According to our minimal

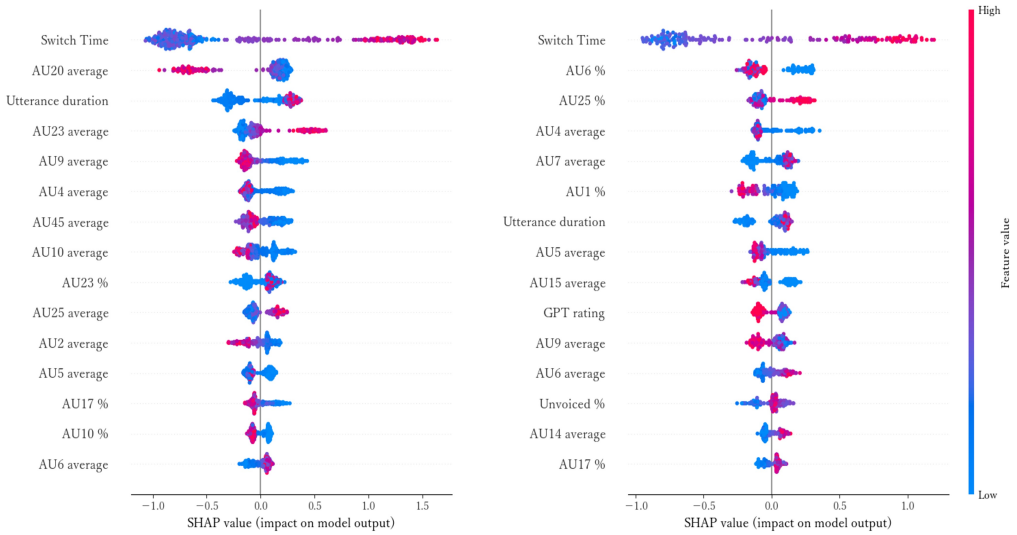


Fig. 5. SHAP value summary plots for elderly (left) and COVID (right) corpora. The top 15 features are listed in order of importance. Each point represents one sample with color signifying the relative value of the feature. Points towards the right side are more likely to be classified as negative reactions.

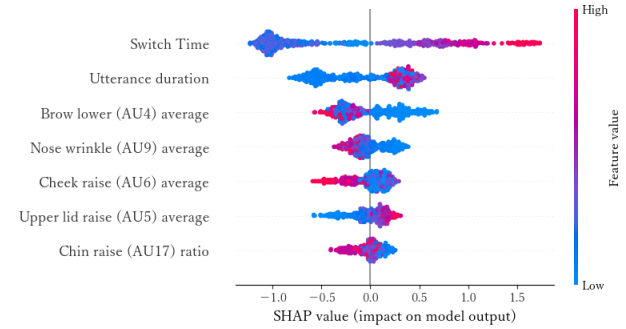


Fig. 6. SHAP summary for minimal model.

model, we can describe an exemplar of a negative reaction as when the user hesitates before responding to the system and continues with their talk without acknowledgment. They tend not to move the middle of their face, except to raise their eyelids. These features should not be seen as highly reliable, but it does allow us to provide explainability to our model and for future work this type of analysis is critical.

Our models were trained only on attentive listening data so we cannot prove that they would generalize to other open-domain dialogue scenarios. Attentive listening is quite different from mixed-initiative free talk where the user can ask questions to the system. It is possible that question-answering errors will be reacted to more strongly than those in our corpora.

There are several limitations in this work. Due to privacy concerns, the annotations of user reactions were conducted by only two experts. An approach with crowd-sourcing could produce different labels, although they may be more inconsistent. We also found that using powerful transformer models with audio streams was not successful even with fine-tuning, perhaps due to the comparatively low number of samples.

IX. Conclusion

We used multimodal features to detect negative reactions in two attentive listening corpora. We extracted audio, visual and linguistic features and trained models on each corpora, finding that facial action unit features were important in the corpus of elderly subjects, while ChatGPT evaluation was more influential for university students. We then used SHAP values to identify key features and trained a minimal combined model which had comparable performance to a larger model.

Acknowledgment

This work was supported by JST Moonshot R&D Goal 1 Avatar Symbiotic Society Project (JPMJPS2011).

References

[1] J. Ravishankar, M. Doering, and T. Kanda, "Zero-shot learning to enable error awareness in data-driven HRI," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 592–601.

[2] M. Stiber, R. H. Taylor, and C.-M. Huang, "On using social signals to enable flexible error-aware HRI," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 222–230.

[3] D. Kontogiorgos, M. Tran, J. Gustafson, and M. Soleymani, "A systematic cross-corpus analysis of human reactions to robot conversational failures," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 112–120.

[4] M. Giuliani, N. Mirnig, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi, "Systematic analysis of video data from different human–robot interaction studies: A categorization of social signals during error situations," *Frontiers in Psychology*, vol. 6, p. 931, 2015.

[5] M. Saeki, K. Miyagi, S. Fujie, *et al.*, "Confusion detection for adaptive conversational strategies of an oral proficiency assessment interview agent," English, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022-September, pp. 3988–3992, 2022.

[6] N. Li, J. D. Kelleher, and R. Ross, "Detecting interlocutor confusion in situated human-avatar dialogue: A pilot study," in *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Potsdam, Germany: SEMDIAL, Sep. 2021.

[7] R. Cumbal, J. Lopes, and O. Engwall, "Detection of listener uncertainty in robot-led second language conversation practice," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, ser. ICMI '20, Association for Computing Machinery, 2020, 625–629.

[8] Y. Tsuta, N. Yoshinaga, S. Sato, and M. Toyoda, "Rethinking response evaluation from interlocutor's eye for open-domain dialogue systems," *arXiv preprint arXiv:2401.02256*, 2024.

[9] A. Yamaguchi, K. Iwasa, and K. Fujita, "Dialogue act-based breakdown detection in negotiation dialogues," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 745–757.

[10] R. Higashinaka, L. F. D'Haro, B. Abu Shawar, *et al.*, "Overview of the dialogue breakdown detection challenge 4," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, Springer, 2021, pp. 403–417.

[11] K. Tsubokura, Y. Iribe, and N. Kitaoka, "Dialog breakdown detection using multimodal features for non-task-oriented dialog systems," in *2022 IEEE 11th Global Conference on Consumer Electronics (GCCE)*, IEEE, 2022, pp. 352–356.

[12] T. Mori, K. Jokinen, and Y. Den, "On the use of gestures in dialogue breakdown detection," in *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, IEEE, 2021, pp. 70–75.

[13] A. Ben-Youssef, C. Clavel, and S. Essid, "Early detection of user engagement breakdown in spontaneous human-humanoid interaction," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 776–787, 2021.

[14] N. Rollet and C. Clavel, "'Talk to you later': Doing social robotics with conversation analysis. towards the development of an automatic system for the prediction of disengagement," *Interaction Studies*, vol. 21, no. 2, pp. 268–292, 2020.

[15] A. Ben-Youssef, G. Varni, S. Essid, and C. Clavel, "On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks," *International Journal of Social Robotics*, vol. 11, no. 5, pp. 815–828, 2019.

[16] L. Tian and S. Oviatt, "A taxonomy of social errors in human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 2, pp. 1–32, 2021.

[17] D. Kontogiorgos, A. Pereira, B. Sahindal, S. van Waveren, and J. Gustafson, "Behavioural responses to robot conversational failures," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 53–62.

[18] D. E. Cahya, R. Ramakrishnan, and M. Giuliani, "Static and temporal differences in social signals between error-free and erroneous situations in human-robot collaboration," in *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*, Springer, 2019, pp. 189–199.

[19] D. Lala, P. Milhorat, K. Inoue, M. Ishida, K. Takanashi, and T. Kawahara, "Attentive listening system with backchanneling, response generation and flexible turn-taking," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, K. Jokinen, M. Stede, D. DeVault, and A. Louis, Eds., Association for Computational Linguistics, Aug. 2017, pp. 127–136.

[20] K. Inoue, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara, "An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, O. Pietquin, S. Muresan, V. Chen, *et al.*, Eds., Association for Computational Linguistics, 2020, pp. 118–127.

[21] C. T. Ishi, H. Ishiguro, and N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality," *Speech Communication*, vol. 50, no. 6, pp. 531–543, 2008.

[22] P. Ekman and W. Friesen, *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.

[23] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.

[24] OpenAI, *ChatGPT (GPT4, March 2024 version)*, 2024.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., Curran Associates, Inc., 2017, pp. 4765–4774.

Data Selection using Spoken Language Identification for Low-Resource and Zero-Resource Speech Recognition

Jianan Chen*, Chenhui Chu*, Sheng Li† and Tatsuya Kawahara*
* Graduate School of Informatics, Kyoto University, Kyoto, Japan
E-mail: {jichen, kawahara}@sap.ist.i.kyoto-u.ac.jp, chu@nlp.ist.i.kyoto-u.ac.jp
† National Institute of Information and Communications Technology (NICT), Kyoto, Japan
E-mail: sheng.li@nict.go.jp

***Abstract*—Large-scale pre-trained models have become common for Automatic Speech Recognition (ASR) tasks. They utilize large-scale, multilingual datasets to learn acoustic features and then are finetuned on downstream ASR tasks. However, their performance degrades when applied to low-resource and zero-resource languages lacking data. This paper introduces a new data selection method with Spoken Language Identification (SLI) models to bring non-target language speech data into training. With the help of phonetic labels as a universal intermediate representation to link low-resource languages to those with rich resources, we enhance the ASR system's performance on low-resource and even zero-resource languages. We conducted ASR experiments on Marathi, Assamese, and Panjabi with augmented non-target Hindi data in the CommonVoice corpora. The experimental results show that the proposed method can train ASR systems with little target language resources.**

I. Introduction

Multilingual pre-trained models have become common when dealing with multilingual Automatic Speech Recognition (ASR) tasks, because of their ability to transcribe speech into many languages with only one model. These models are usually pre-trained on large-scale datasets containing different languages. The pre-trained models can leverage shared acoustic features and improve downstream task performance [1]–[3]. However, these models still require a lot of annotated data for finetuning to perform well. It is still challenging to transfer them to low-resource or even zero-resource languages not included in the datasets used for pre-training [4].

Compared to the low-resource language, it is easier to get data from languages with rich resources. Several studies have investigated phonetic pre-training on various tasks, revealing that introducing non-target language data can improve performance across languages [5]–[8]. However, despite the success of introducing other languages to improve low-resource language performance, not all the data from other languages are useful for improving performance [9]. With low-quality data, the performance might even degrade. As a result, the challenge remains of how to select the optimal non-target language and data to improve the target ASR systems.

Several studies have investigated the similarities between languages [10]–[12]. However, their studies did not include acoustic features, which are essential for training ASR systems. Even if one language is similar to another, the detailed pronunciation and articulation can vary significantly. Several studies have investigated the semantic similarities between text sentences and sentences [13]–[15]. None of the studies focused on how similar one speech audio can be to another language.

Motivated by the fact that SLI is a task that classifies single speech audio into its languages, we propose to use SLI models to measure how similar one speech audio is to another language, including acoustic features. This will help us select the optimal non-target language data to assist the training of ASR systems for low-resource, even zero-resource languages. By identifying the language of each speech audio clip, even if different from the target languages, we can ensure the selected data reflects the acoustic features of the target languages to some extent.

Soky et al. (2023) [16] utilized SLI to enhance the performance of low-resource languages. That work involves using target language data from different domains and non-target language data from the same domains to supplement the original target language data, demonstrating that introducing non-target language data can improve the ASR performance of the target language. However, that work used all non-target language data, regardless of the quality of the data. Meanwhile, Ma et al. (2023) [17], and Wang et al. (2023) [18] used SLI to improve code-switch ASR. They used SLI as a router to select the neural network segments. They have shown that SLI can be used to code-switch ASR. However, they did not test the performance of SLI on multilingual ASR.

Samuel et al. (2016) [19] used SLI to cluster similar languages into groups from multilingual datasets. However, they also used all language data for training. Chuangsuwanich et al. (2016) [20] used SLI to select frames of non-target language audio clips that are similar to non-target languages. However, their work trained frame-level speech features for DNNs instead of end-to-end ASR. Since that work focused on the frame-level features, they might also lose some structured features in languages. Zhou et al. (2022) [21] used an SLI model to calculate the weights of the non-target samples to the final loss in the training process, making the model

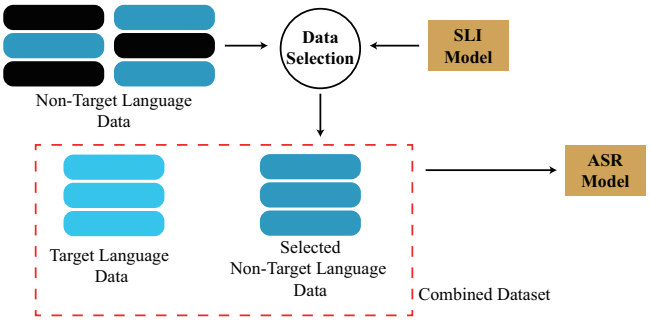


Fig. 1. Illustration of the pipeline to train ASR in this work. We used a data selection module to select non-target language data that sounds like the target language. Added to the original target language data, we obtained a combined dataset to train the ASR system for the target language.

more biased toward learning the target low-resource language. However, their method increased the complexity of training loss. Zhang et al. (2023) [22] also used the SLI model to filter similar non-target languages in the pre-training of multilingual ASR models. However, they did not use the selected data for the finetuning process.

To the best of our knowledge, this work is the first attempt to utilize SLI models to select non-target language data for augmentation in end-to-end ASR finetuning.

The experimental results show that non-target speech data of other languages can benefit the ASR training for low-resource languages. Even though much of the data is filtered out by the proposed SLI selection method, the performance can be better than the baselines using all non-target language data. In addition, it is shown that even without any target language data, we can build ASR systems for zero-resource languages with the help of high-quality non-target language data.

II. Data Selection with Spoken Language Identification

Spoken Language Identification (SLI) is a task that involves calculating the probabilities of the language to which a speech audio belongs. It can be utilized to measure how similar a speech audio clip is to another language. The probabilities given by SLI models can serve as objective metrics to measure how similar a speech audio clip can sound in one specific language. After selecting audio clips similar to the target language, we can use them to supplement the dataset for a low-resource target language.

We propose to use an SLI model to select non-target language data from different languages for the ASR training of the low-resource languages. As shown in Fig. 1, part of the non-target language data are selected with the SLI model, and in addition to the target language data, we obtain an augmented dataset for the following ASR training. We denote this method as **SLI Selection**. The process of selecting data from the original datasets is as follows:

1) **Train SLI model:** Train an SLI model with a multilingual dataset, including the target and the non-target languages.
2) **Calculate the Probabilities of Each Language:** Use the trained SLI model to predict the language of speech audio clips in the multilingual datasets. Here, $\mathbf{x}_i$ is the i-th speech clip in the speech dataset $\mathcal{X}$ = [$\mathbf{x}_1$, $\mathbf{x}_2$ ... $\mathbf{x}_N$], N is the size of the dataset; the probability vector of each language $\mathbf{p}_i^{(langs)}$ is calculated on each $\mathbf{x}_i$.

$$\mathbf{p}_i^{(langs)} = \mathrm{SLI}(\mathbf{x}_i) \tag{1}$$

3) **Select If Target Language in Top-K:** Sort the language candidates according to the probabilities. If the target language is in the top-k candidates, the SLI model considers this clip to sound similar to the target language. Thus, we select this clip for the new dataset.
We use the top-k selection because SLI models have high accuracies, causing the probability of the true language to be outstandingly higher than the rest of the probabilities and thus the probabilities of the other languages are not stable and reliable.

III. Experimental Setups

A. Dataset

Four Indian Languages of Hindi, Marathi, Assamese, and Panjabi from the CommonVoice corpora [23] are utilized for training and evaluations. CommonVoice is a large open-source voice dataset built by volunteers worldwide. The dataset is designed especially for the ASR task. It has many releases, and we use the 16.1 release for this work. The detailed information of the data used in this work is shown in Table I.

We fine-tune the XLSR-53 [24] on the downstream ASR task to transcribe Marathi, Assamese, and Panjabi. Hindi, one of India's most spoken languages and geologically near the regions where target languages are spoken, is used only in the training process as an extra language. We then implement the algorithms presented in Section II, selecting audio clips that are similar to the target languages from the Hindi dataset. The models trained with the proposed method are referred to as *top-k*, where k is the rank parameter used for data selection.

In the baseline setups, we build the ASR systems with 1) all the Hindi, 2) all the target language data, or 3) a combination of Hindi and the target language data. The baseline is referred to as *all*. In addition, to eliminate the effect of the quantity of data, we randomly select part of the data from the original Hindi dataset to build an additional baseline. This baseline is referred to as *random*.

We also conduct both few-shot and zero-shot experiments. For the few-shot experiments, we first combine the target language (Marathi Assamese or Panjabi) and the non-target language (Hindi) to obtain a combined training dataset. For the zero-shot experiments, we use the non-target language data and do not use the target language data in the training process.

B. Training SLI and data selection

We followed the ECAPA-TDNN architecture [25] provided by SpeechBrain [26] to build the SLI model in Fig. 2. The model mainly consists of four blocks. The TDNN (Time Delay

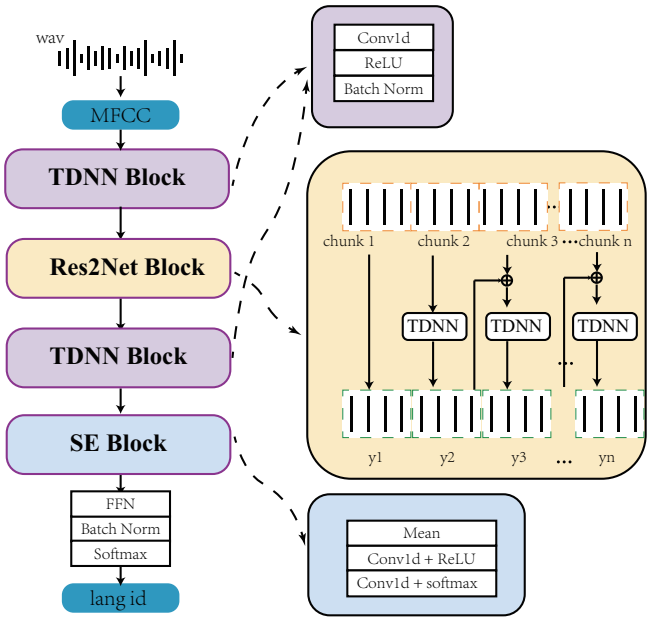


Fig. 2. Illustration of the SLI model architecture.

TABLE I
SPECIFICATIONS OF THE TRAINING DATASETS IN COMMON VOICE USED IN THE EXPERIMENTS.

Language	Audio clips	Duration (hours)
Hindi (hi)	4675	6.7
Marathi (mr)	2215	3.9
Assamese (as)	658	1.2
Panjabi (pa)	733	1.2

Neural Network) blocks are stacked with Conv1d, ReLU, and Batch Normalization layers. The Res2Net block groups the hidden representations into fixed-size chunks and has a recurrent architecture to capture the long-term temporal dependencies across chunks. The SE (Squeeze-and-Excitation) block first squeezes the representations and then expands them. Finally, there is a feedforward network and a softmax layer as the output layer that gives the predicted language id.

The model was trained on VoxLingua107 [27], which is a dataset designed for SLI tasks. The sampling rate of the single-channel speech audio clips is 16kHz. The input features are the 80-dimensional MFCCs calculated with a 25ms window and 10ms frameshift. The kernel sizes of Conv1d layers in TDNN blocks were set to 1, the chunk size of the Res2Net block was set to 8, and the kernel sizes in SE blocks were set to 8. We used the linear learning rate scheduler, and the maximum learning rate was set to 1e-3. We trained on one A6000 GPU until it converged.

C. Training ASR

We utilize the Wav2Vec 2.0 XLS-R 53 as the ASR encoder. The encoder's inputs are 16kHz single-channel speech audio clips, and its outputs are feature embeddings extracted by the encoder. Then, we added a linear projection layer after the encoder and fine-tuned the model with data with a CTC loss. We use the Adam optimizer with a 5e-5 final learning rate and the linear scheduler, setting the batch size to 16 and accumulating the gradient for two steps. The learning rate is set to 5e-5, and we utilize the linear learning rate scheduler. We train the models on the data for at most 200 epochs on 2 48GB A6000 GPUs.

IV. RESULTS AND ANALYSIS

This section presents the results of the experiments. There are mainly three parts of the experiments:

1) **SLI Selection on Hindi:** We first conduct the top-k SLI selections on the Hindi datasets and filter the audio clips similar to the three target languages.
2) **Few-Shot Experiments:** We conduct few-shot end-to-end ASR experiments in this subsection. Both the target language data (Marathi, Assamese, or Panjabi) and the extra non-target language data (Hindi) are utilized as training data.
3) **Zero-Shot Experiments:** We conduct zero-shot end-to-end ASR experiments in this subsection. Only extra non-target language data (Hindi) is utilized as training data.

A. SLI Selection on Hindi

This subsection presents the results of applying SLI selection on the Hindi train dataset in the CommonVoice Dataset. Following is the analysis of the results:

1) Number of audio clips after Selection: We applied the SLI selection on the dataset with k from 1 to 10 to select Hindi subsets that are similar to Marathi, Assamese, and Panjabi pronunciations, respectively. As shown in Table I, the original Hindi train dataset consists of about 4675 audio clips, and the total duration of the dataset is about 6.7 hours.

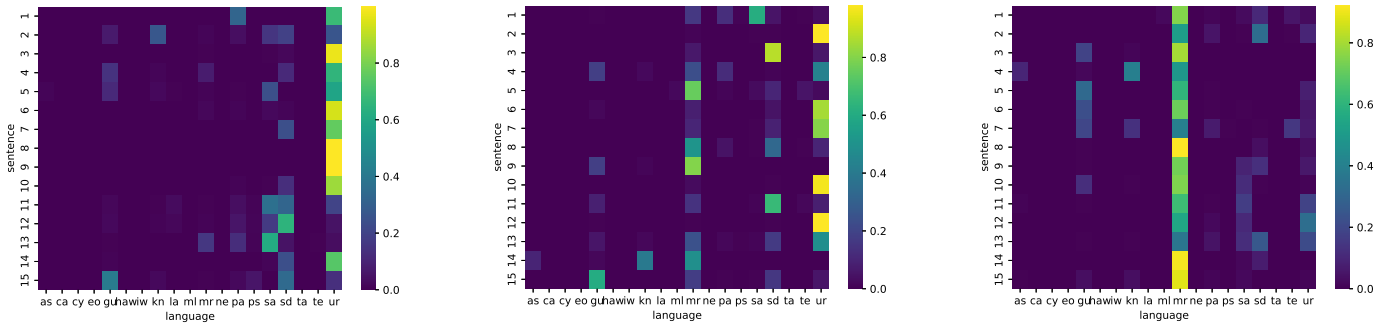
The number of audio clips after selection is shown in Table II. It can be seen that when k is set at 1, the selected non-target Hindi dataset contains only a small number of audio clips, which is neglectable compared to the original target language dataset size. As a result, we choose to conduct Marathi and Panjabi experiments with k=2 and 5 and Assamese experiments with k=7 and 10 in the following few-shot and zero-shot experiments to make a comparable amount of training data.

2) SLI Selection Filters More Target-like Audio Clips: Fig. 3 presents the probabilities of 19 top languages after data selection. We applied selection to the Hindi dataset and got a new dataset for training Marathi ASR. We selected 15 audio clips and used the SLI model again to predict the possible language IDs of the clips. We omitted the true language ID hi (Hindi) in the probability heatmap because Hindi's probabilities are outstandingly higher than those of the other languages.

Fig. 2a presents that the probabilities with random selection in ur (Urdu, an Indic language sounds similar to Hindi, but in a different writing system) are higher than the other languages. This indicates that the SLI model considers the Hindi speech clip to be very similar to Urdu, instead of the

TABLE II
NUMBER OF SELECTED AUDIO CLIPS AFTER APPLYING SLI SELECTION ON THE HINDI TRAIN DATASET IN COMMONVOICE. (BOLD FONTS ARE USED FOR ASR)

Top-k	k=1	2	3	4	5	6	7	8	9	10	all
Marathi (mr)	36	**497**	1154	1931	**2729**	3360	3743	4020	4198	4303	4675
Assamese (as)	3	21	67	123	215	319	**480**	692	950	**1224**	4675
Panjabi (pa)	83	**807**	2251	3168	**3670**	4000	4225	4366	4448	4493	4675



(a) random (b) top-3 (c) top-2

Fig. 3. Probabilities of the Hindi data with different data selection methods on 19 languages (as, ca, cy, eo, gu, haw, hi, iw, kn, la, ml, mr, ne, pa, ps, sa, sd, ta, te, ur). (a) is the probabilities for random selection; (b) is for SLI selection with k=3; (c) is with k=2.

target language, Marathi. Meanwhile, Fig. 2b and Fig. 2c show the probabilities of SLI selection with k equals 3 and 2, respectively. It is observed that with smaller parameter k=2, the proposed method can select more target-like data, showing higher probabilities in mr than ur.

B. Few-Shot Experiments

The few-shot experiments were conducted on three target languages: Marathi, Assamese, and Panjabi. The character error rate is calculated to measure the performance of phonetic label ASR systems. Table III presents the experimental results and following is the analysis:

1) Ratio of Target and Non-Target Language: One of our findings is that the ratio of the target and the non-target language data is essential in few-shot experiments. By comparing experiment no. 1 (only target language data) and experiment no. 2 (both target and non-target language data are used), we can see that introducing all target language data decreased the performance of the ASR system in the target language. However, by comparing experiment no. 2 (all non-target), experiment no. 4 to 6 (part of non-target), it can be observed that using fewer extra data might not only perform better than using only target language data but also perform better than using all the non-target language data. In addition, no. 6 (top-2) obtained the best result.

Thus, the ratio of the target and non-target language data is crucial because the non-target language data should not be too many or too few compared to the target language data amount. The same phenomenon can be seen in all three target languages. However, the best ratio of these data remains a problem that needs to be investigated.

2) Quality of Non-Target Language Data: Another finding is that the quality of the extra non-target language data is important. In the experiments, we used two methods to select subsets of the non-target language data for training. One is the random selection, and the other is the SLI selection. Despite the same amount of data for training, the experimental results differ depending on how we selected the data.

In Marathi, the result of experiment no. 6 (top-2) is about 14% better than no. 4 (randomly selected); In Assamese, no. 11 (top-10) is about 1% better than experiment no. 9 (randomly selected); In Panjabi, no. 18 (top-2) is 5% better than no. 16 (randomly selected). It can be concluded that even with the same amount of final training data, the data selected with the proposed method is better than random selection in quality.

3) Effect of the Proposed Method: Finally, from Table III, the proposed method achieved the best CER of 11.61 for Marathi by a 2% improvement; the best CER of 19.21 for Assamese by a 3% improvement; and the best CER of 27.14 of Panjabi by a 16% improvement compared to the baselines using only target language data (no. 1, 7, and 13, respectively). In addition, the proposed method outperformed all the corresponding baselines with random selection.

These experimental results showed that the proposed data selection with SLI models could select better non-target data for training the ASR systems, even though the amount of the selected data is much smaller (less than 1 hour).

One possible reason is that introducing low-quality non-target data might harm the training of ASR systems for target languages. In random selection, not only high-quality data but also low-quality non-target data are selected. The distribution of the data selected randomly should be similar to the original

TABLE III
RESULTS (CER%) OF THE ASR MODELS TRAINED WITH BOTH TARGET (TGT) AND NON-TARGET (EXTRA) SPEECH DATA (LOW-RESOURCE) AND PHONETIC LABEL SCRIPTS. THE BEST RESULTS ARE SHOWN IN BOLD.

Target	No.	Method	Duration (Hours)			CER%
			tgt	extra	total	
Marathi (mr)	1	all	3.9	-	3.9	11.81
	2	all	3.9	6.7	10.6	12.18
	3	random	3.9	3.9	7.8	13.80
	4	random	3.9	0.7	4.6	11.65
	5	top-5	3.9	3.9	7.8	11.90
	6	top-2	3.9	0.7	4.6	**11.61**
Assamese (as)	7	all	1.2	-	1.2	19.82
	8	all	1.2	6.7	7.9	22.00
	9	random	1.2	1.7	2.9	19.42
	10	random	1.2	0.7	1.9	19.80
	11	top-10	1.2	1.7	2.9	**19.21**
	12	top-7	1.2	0.7	1.9	19.71
Panjabi (pa)	13	all	1.2	-	1.2	32.48
	14	all	1.2	6.7	7.9	66.17
	15	random	1.2	5.3	6.5	28.86
	16	random	1.2	1.1	2.3	28.63
	17	top-5	1.2	5.3	6.5	27.28
	18	top-2	1.2	1.1	2.3	**27.14**

non-target dataset. As mentioned before, introducing non-target data directly might degrade the performance. Meanwhile, as shown in Fig. 3, the data selection method can select data more similar to the target language, helping the ASR models to learn more similar acoustic features.

4) Limitation of SLI Slection: Although SLI selection with k=2 leads to the best results in Marathi and Panjabi, a larger k was used in Assamese. In Assamese, the best result comes at k=10. The amount of the similar Hindi data is smaller than the other 2 languages, because the SLI model takes Hindi and Assamese not as similar as the other two languages. The best amount and the method to find it remain to be solved for future studies.

C. Zero-Shot Experiments

The zero-shot experiments were conducted on the three target languages. Hindi was used in the training process in the zero-shot experiments. The target languages are only used for the evaluations after the ASR systems are trained with the non-target language data. Table IV presents the experimental results. The table columns are similar to Table III, except that Duration is the non-target Hindi data hours. Following is the analysis for the results:

1) SLI Selection Shows Effectiveness in Zero-Shot: From Table IV, it can be seen that the proposed method has effect on the zero-shot experiments. Although the amount of data is much less than those in few-shot experiments, the systems can gain improvements on the baselines. In Assamese, no. 9 (top-10) obtained around 60% improvements from no. 7 (randomly selected). In Panjabi, no. 15 (top-2) obtained around 56% improvements from no. 13 (randomly selected).

Comparing Table III and Table IV, in Assamese and Panjabi, the proposed method achieved comparable zero-shot results to the few-shot results. This indicates that the proposed method can be utilized to construct ASR systems on zero-resource languages.

TABLE IV
RESULTS (CER%) OF THE ASR MODELS TRAINED WITH ONLY NON-TARGET SPEECH DATA (ZERO-SHOT) AND PHONETIC LABEL SCRIPTS. THE BEST RESULTS ARE SHOWN IN BOLD.

Target	No.	Method	Duration (Hours)	CER%
Marathi (mr)	1	all	6.7	34.03
	2	random	3.9	**33.15**
	3	random	0.7	98.61
	4	top-5	3.9	35.04
	5	top-2	0.7	33.63
Assamese (as)	6	all	6.7	73.01
	7	random	1.7	51.75
	8	random	0.7	66.72
	9	top-10	1.7	**20.74**
	10	top-7	0.7	22.85
Panjabi (pa)	11	all	6.7	88.60
	12	random	5.3	79.63
	13	random	1.1	75.90
	14	top-5	5.3	36.45
	15	top-2	1.1	**33.13**

2) Limitation of SLI Selection in Zero-Shot: Unfortunately, the proposed SLI selection method did not work consistently in the zero-shot experiments. In the zero-shot Marathi ASR experiments, the best result comes at no. 2 (randomly selected) and data selection does not show any effect. Moreover, there is a large gap between few-shot and zero-shot in Marathi. This suggests that the SLI selection is not stable. More stable data selection methods should be studied for future works.

V. CONCLUSIONS

This work has proposed a new data selection method with SLI models to select the suitable non-target language data to train ASR systems for target languages to augment low-resource and zero-resource languages. We trained the ASR systems in Marathi, Assamese, and Panjabi with the augmentation of the non-target language, Hindi. We concluded that 1) introducing non-target language data can improve the ASR results for low-resource languages; 2) SLI selection is more effective than the random selection baselines, especially in the zero-shot experiments.

The proposed SLI selection approach can help enhance the usage of non-target languages in training ASR systems. This work paves the way for future studies, particularly in utilizing SLI to select non-target language data for auxiliary training on low-resource languages. Future works can dive into finding the optimal amount non-target language data that can help improve the data selection process for better ASR performance. Moreover, future works on more stable data selection methods should be conducted.

ACKNOWLEDGMENT

This work was partly supported by JST BOOST, Grant Number JPMJBS2407.

References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] H. et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[3] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7967–7971.

[4] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv:2012.12121*, 2020.

[5] C. S. Anoop and A. G. Ramakrishnan, "Exploring a unified asr for multiple south indian languages leveraging multilingual acoustic and language models," in *Proc. IEEE SLT Workshop*, 2023, pp. 830–837. DOI: 10.1109/SLT54892.2023.10022380.

[6] Y. Qian, K. Yu, and J. Liu, "Combination of data borrowing strategies for low-resource lvcsr," in *2013 IEEE workshop on automatic speech recognition and understanding*, IEEE, 2013, pp. 404–409.

[7] Y. Qian and Z. Zhou, "Optimizing data usage for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 394–403, 2022.

[8] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," *arXiv:2109.11680*, 2021.

[9] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Language id-based training of multilingual stacked bottleneck features," in *Proc. Interspeech*, Citeseer, 2014, pp. 1–5.

[10] V. Beaufils and J. Tomin, "Stochastic approach to worldwide language classification: The signals and the noise towards long-range exploration," 2020.

[11] K. Batsuren, G. Bella, and F. Giunchiglia, "A large and evolving cognate database," *Language Resources and Evaluation*, pp. 1–25, 2022.

[12] J. Eronen, M. Ptaszynski, and F. Masui, "Zero-shot cross-lingual transfer language selection using linguistic similarity," *Information Processing Management*, vol. 60, no. 3, p. 103 250, 2023, ISSN: 0306-4573.

[13] L. Wang, Y. Ling, Z. Yuan, *et al.*, "Gensim: Generating robotic simulation tasks via large language models," in *Arxiv*, 2023.

[14] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Association for Computational Linguistics, Apr. 2017, pp. 427–431.

[15] C. Orašan, "Aggressive language identification using word embeddings and sentiment features," in *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, 2018, pp. 113–119.

[16] K. Soky, S. Li, C. Chu, and T. Kawahara, "Domain and language adaptation using heterogeneous datasets for wav2vec2. 0-based speech recognition of low-resource language," in *ICASSP 2023*, IEEE, 2023, pp. 1–5.

[17] G. Ma, W. Wang, Y. Li, Y. Yang, B. Du, and H. Fu, "Lae-st-moe: Boosted language-aware encoder using speech translation auxiliary task for e2e code-switching asr," in *ASRU 2023*, IEEE, 2023, pp. 1–8.

[18] W. Wang, G. Ma, Y. Li, and B. Du, "Language-routing mixture of experts for multilingual and code-switching speech recognition," *arXiv:2307.05956*, 2023.

[19] S. Thomas, K. Audhkhasi, J. Cui, B. Kingsbury, and B. Ramabhadran, "Multilingual data selection for low resource speech recognition.," in *Interspeech*, 2016, pp. 3853–3857.

[20] E. Chuangsuwanich, Y. Zhang, and J. Glass, "Multilingual data selection for training stacked bottleneck features," in *ICASSP 2016*, IEEE, 2016, pp. 5410–5414.

[21] Z. Zhou, W. Wang, W. Zhang, and Y. Qian, "Exploring effective data utilization for low-resource speech recognition," in *ICASSP 2022*, 2022, pp. 8192–8196. DOI: 10.1109/ICASSP43922.2022.9747543.

[22] Z. Zhang, W. Wang, and Y. Qian, "Fast and efficient multilingual self-supervised pre-training for low-resource speech recognition," *Proceedings of Interspeech. Dublin, Ireland*, 2023.

[23] R. Ardila, M. Branson, K. Davis, *et al.*, "Common voice: A massively-multilingual speech corpus," English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, *et al.*, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222, ISBN: 979-10-95546-34-4.

[24] A. Babu, C. Wang, A. Tjandra, *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv:2111.09296*, 2021.

[25] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *ArXiv*, vol. abs/2005.07143, 2020.

[26] M. Ravanelli, T. Parcollet, and P. Plantinga, *SpeechBrain: A general-purpose speech toolkit*, arXiv:2106.04624, 2021. arXiv: 2106.04624.

[27] J. Valk and T. Alumäe, "VoxLingua107: A dataset for spoken language recognition," in *Proc. IEEE SLT Workshop*, 2021.

What to Refer and How? - Exploring Handling of Auxiliary Information in Target Speaker Extraction

Tomohiro Hayashi*, Riku Ogino*, Kohei Saijo* and Tetsuji Ogawa*
* Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan
E-mail: thayashi@pcl.cs.waseda.ac.jp

***Abstract*—This study explored the appropriate auxiliary information for target speaker extraction (TSE). In TSE, the conventional method involves using pre-recorded speech as an auxiliary signal. However, in multi-channel environments where spatial information is available, alternative approaches that utilize a rough pre-extraction of the target speech as an auxiliary signal are being explored. Pre-recorded speech is typically clean but contains different content, whereas pre-extracted speech has content aligned with the target but includes various types of distortion depending on the extraction method used. These two approaches differ in the characteristics of the auxiliary signals and the conditioning methods applied. Previous research has not thoroughly examined the relationship between the nature of auxiliary signals and TSE model performance. This study aims to explore the effective use of auxiliary signals in TSE models and to provide a comprehensive analysis of their impact. Through TSE experiments conducted in a multi-channel environment, we investigated the effects of auxiliary signals from three key perspectives: the trade-off between content consistency and distortion, the level of referencing, and the type of distortion.**

I. Introduction

Speech separation has long served as a crucial front-end component for speech applications like automatic speech recognition (ASR) [1], [2]. Recently, efforts have focused on isolating the target source from mixtures containing various sources by incorporating auxiliary information about the target source into separation networks. This approach, known as target speaker extraction (TSE) [3], has seen significant advancements due to the adoption of neural networks. Alongside improvements in network architecture, considerable research has also been dedicated to enhancing the auxiliary information used to identify the target speaker [4], [5].

In TSE, an enrollment utterance, which is a pre-recorded sample of the target speaker's voice, is typically used as auxiliary information. The speaker embedding derived from this enrollment utterance via a speaker encoder is employed to condition the extraction network. Since the enrollment utterance and the target speech in the mixture differ, the speaker embedding is often transformed into a global feature, such as by applying mean pooling across time frames.

On the other hand, in some scenarios, it is possible to estimate the target speech without enrollment utterances, for instance, by leveraging the spatial position of the target speaker. For example, if multi-channel microphones are available and the direction of the target speaker is known, blind source separation (BSS) [6] can be used to roughly extract the target speech. In these situations, the pre-extracted signal can be employed as auxiliary information in place of enrollment utterances, which may be more advantageous for target speech extraction (TSE) because both the speech content and the recording environment are identical. However, this approach introduces a trade-off between content consistency and processing distortion, as the pre-extracted speech is likely to include various distortions. The nature and extent of these distortions, such as nonlinear artifacts and residual noise, depend on the specific signal processing methods used [7], [8].

In such scenarios, where the pre-extracted speech and the target speech in the mixture share the same content, it is possible to condition the model on a frame-by-frame basis, rather than relying solely on global features as in traditional TSE systems. In speech separation, various methods like Sequential Neural Beamforming [9] and Beam-guided TasNet [10] have been developed to use pre-separated speech to guide an additional separation network, thereby enhancing performance. This approach has also been extended to TSE, known as Beamformer-guided TSE [11], where the TSE network utilizes the pre-extracted beamforming output as frame-by-frame auxiliary information. However, as previously mentioned, the output from the pre-extraction stage often includes distortions, and it remains uncertain whether frame-by-frame conditioning is superior to systems employing global speaker features. Although Beamformer-guided TSE has demonstrated that frame-by-frame conditioning can surpass systems using global speaker features, the influence of distortions in pre-extracted speech on TSE has not been thoroughly investigated.

This raises several critical questions regarding the optimal auxiliary signals for TSE, particularly concerning the trade-offs between distortions and content alignment with the target source, the nature of distortions in the auxiliary signals, and the level at which these signals should be referenced:

- **Question 1**: When referencing signals at the statistical level, which is more effective: undistorted but unaligned enrollment utterances, or distorted but content-aligned pre-extracted signals?
- **Question 2**: When referencing distorted but content-aligned pre-extracted signals, which is more advantageous: statistical-level conditioning or frame-level conditioning?
- **Question 3**: When referencing distorted but content-aligned pre-extracted signals, which type of distortion is preferable: signals containing artifacts or those with

residual noise?

In this paper, we address these questions by exploring various aspects of providing auxiliary information to TSE models. Through experiments conducted in a multi-microphone environment, we aim to elucidate how different types of auxiliary signals and conditioning methods affect TSE performance. The findings from this study are expected to contribute valuable insights for the effective utilization of auxiliary information in TSE.

The remainder of the paper is organized as follows. Section II discusses the key technologies utilized in this study. Section III presents the experiments on target speaker extraction and summarizes insights related to the use of auxiliary signals. Finally, Section IV concludes with a summary of the findings.

II. Key Technologies

In this study, we address the three questions posed in the introduction by conducting experiments that compare five TSE models. These models utilize three types of auxiliary signals and two conditioning methods. The auxiliary signals include pre-recorded speech, which is commonly used in conventional TSE, as well as pre-extracted speech obtained using Independent Vector Analysis (IVA) [12] and Ideal Binary Masking (IBM) [13] in a multi-channel environment. The conditioning methods involve referencing the auxiliary signals either at the statistical level or at the frame level. This section provides an overview of the key technologies employed in our experiments.

A. Ideal Binary Masking (IBM)

IBM is a widely used technique for speech separation, where an ideal binary mask is generated to isolate target signals from interference. In this approach, each time-frequency bin is classified as binary: bins where the target signal dominates the interference are retained, while others are masked. Specifically, the spectral amplitude of the target signal $\mathbf{T}(t, f)$ is compared to that of the interference signal $\mathbf{I}(t, f)$, generating a mask that assigns a value of one to bins dominated by the target signal and zero otherwise. Applying this mask to the spectrogram of the mixture signal enhances the target signal while suppressing interference. The time-domain target signal is then reconstructed using the inverse Short-Time Fourier Transform (ISTFT) [14].

In this study, we estimated IBM using a Deep Neural Network (DNN) composed of convolutional layers and Long Short-Term Memory (LSTM) [15] layers. Multichannel mixture signals and corresponding clean target speech were used as ground truth. The DNN inputs the spectrogram of the mixture signal and outputs a binary mask that indicates whether each time-frequency bin corresponds to the target speech. During training, network parameters were optimized to minimize the error between the estimated and ideal masks using the binary cross-entropy loss function. Due to the non-linear nature of IBM processing, the output audio from DNN estimation contains artifacts.

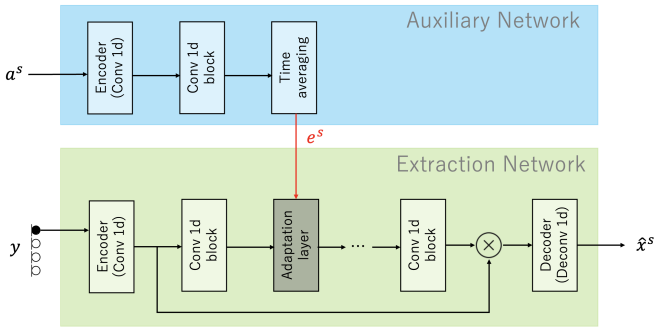


Fig. 1. Model structure for statistics-level referencing of auxiliary information. a^s can be either clean enrollment utterance or pre-extracted signal of speaker s.

B. Independent Vector Analysis (IVA)

IVA is a sophisticated Blind Source Separation (BSS) technique that leverages the statistical independence of multi-dimensional signals. Building upon Independent Component Analysis (ICA) [16], IVA excels in separating multiple sources captured by multiple microphones.

The core principle of IVA is based on a model where the observed signal vector $\mathbf{y}(t)$ is a linear mixture of unknown, independent source vectors $\mathbf{x}(t)$, indexed by time t. This relationship is mathematically represented as:

$$\mathbf{y}(t) = \mathbf{A}\mathbf{x}(t), \tag{1}$$

where $\mathbf{A}$ is the mixing matrix, with each column representing the propagation of signals from different sources to the microphones. IVA estimates this mixing matrix $\mathbf{A}$ from the observed signals, enabling the recovery of the original source signals $\mathbf{x}(t)$.

In IVA, the number of separated output signals typically matches the number of input channels. In this experiment, we assumed that the direction of arrival (DOA) of the target speaker was known in advance and used this information to perform the separation. From the separated signals, we identified the one corresponding to the target speaker by calculating the steering vector based on the speaker's DOA. We then assessed the similarity by computing the average absolute value of the inner product between the steering vectors of the separated signals and that of the target speaker. The signal with the highest similarity was selected as the target speech. However, due to the separation process, the pre-extracted speech obtained using IVA still contains residual noise.

C. Statistics-Level Referencing of Target Speaker Information

Although any existing TSE architectures could be employed for this investigation, we selected the time-domain implementation of SpeakerBeam (TD-SpeakerBeam) [17]–[21] due to its foundational role in many TSE models. Figure 1 provides an overview of TD-SpeakerBeam. This architecture consists of two networks: an auxiliary network [22] and an extraction network. The auxiliary network is designed to extract the

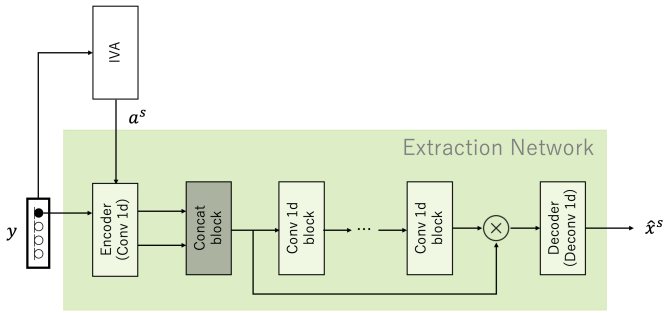


Fig. 2. Model structure of frame-level referencing of auxiliary information.

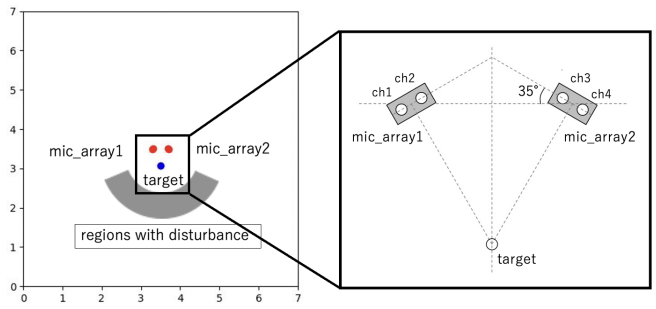


Fig. 3. Simulational acoustic field used in experiments.

global speaker embedding of the target speaker from either the enrollment utterance or the pre-extracted signal a^s. Meanwhile, the extraction network processes the mixture y and uses the speaker embedding to condition its internal features for extracting the target speech.

The auxiliary network includes an encoder layer and a single convolutional block to process the time-domain auxiliary signal. The output from the convolutional block is averaged over time frames to produce the global speaker embedding e^s.

The extraction network is configured similarly to Conv-TasNet [23]. The time-domain mixture y is input into the encoder layer, passes through several convolutional blocks, and generates a mask in the latent space. An adaptation layer, placed between the first and second convolutional blocks, incorporates the speaker embedding vector e^s from the auxiliary network to guide the mask estimation through element-wise multiplication. This approach ensures that only the target speech is extracted by tailoring the mask estimation based on the speaker embedding. The final mask, applied to the mixed signal and processed through the decoder layer, yields the time-domain extracted signal $\hat{x}^s$. Since this model uses the global speaker embedding averaged over time frames as statistical values, it exemplifies "referencing auxiliary information at the statistical level."

D. Frame-Level Referencing of Target Speaker Information

We investigate a framework in which the extraction network refers to the auxiliary information of the target speaker on a frame-by-frame basis, similar to the Beamformer-guided TSE approach. In scenarios where it is feasible to isolate and extract the target source from multi-channel observation signals, the extraction network can use the pre-extracted target source signal as auxiliary information. In this study, we employ IVA and IBM for pre-extraction to obtain this auxiliary information.

Figure 2 illustrates the structure of this model. Initially, signals recorded by multi-channel microphones are separated into source signals using IVA or IBM, and the extracted signal corresponding to the target source is referenced by the extraction network. The core extraction network retains the architecture of the SpeakerBeam. In the frame-level referencing approach, both the mixture and auxiliary signals are input into the encoder, and the outputs are concatenated along the feature dimension. Despite containing distortions, the auxiliary signals are aligned with the target source in the mixture. This allows the model to utilize the pre-extracted signal on a frame-by-frame basis, potentially leading to improved performance compared to statistical-level referencing due to the richer information available. However, when the pre-extracted signal contains significant distortion, global speaker embedding might offer greater stability. The trade-off between stability against distortions and content alignment remains an area for further investigation.

III. Target Speaker Extraction Experiment

To address the key questions concerning the trade-off between content consistency and processing distortion (**Q1**), the referencing level of auxiliary signals (**Q2**), and the type of distortion in auxiliary signals (**Q3**), we conducted target speaker extraction (TSE) experiments in a multi-channel input setting.

To achieve this, we compared the following models:

1) A model that refers to undistorted pre-recorded speech at the statistical level,
2) A model that refers to speech containing artifacts pre-extracted by IBM at the statistical level,
3) A model that refers to speech containing residual noise pre-extracted by IVA at the statistical level,
4) A model that refers to speech containing artifacts pre-extracted by IBM at the frame level, and
5) A model that refers to speech containing residual noise pre-extracted by IVA at the frame level.

Comparing Models 1, 2, and 3 addresses Question 1. To investigate Question 2, we compare Models 2 and 4, as well as Models 3 and 5. Finally, to explore Question 3, we compare Models 2 and 3, as well as Models 4 and 5.

A. Experimental Setups

Figure 3 illustrates the experimental simulation environment. The room dimensions are $7 \times 7 \times 3$ meters. Two microphone arrays, each with two channels, are arranged as shown in the figure. The microphone arrays are spaced 40 cm apart, with a 3 cm distance between microphones within each array. The

TABLE I
RESULTS OF TARGET SPEAKER EXTRACTION EXPERIMENTS IN TWO-SPEAKER MIXED ENVIRONMENT.

System	Target Speaker Extraction Models		Evaluation Results	
	Auxiliary Signal	Referencing Level	SISDR [dB]	STOI [%]
Observation	-	-	-0.5	70.3
IBM	-	-	7.3	86.5
IVA	-	-	7.8	88.4
TSE-1	Pre-recorded clean signal	Statistics	12.1	90.6
TSE-2	Pre-extracted signal by IBM	Statistics	12.6	91.3
TSE-3	Pre-extracted signal by IVA	Statistics	12.4	91.1
TSE-4	Pre-extracted signal by IBM	Frame-by-frame	15.2	94.5
TSE-5	Pre-extracted signal by IVA	Frame-by-frame	18.0	96.3

target source is placed in front of the microphone arrays, while the interfering sources are located one to two meters behind. This setup simulates the recording of mixed speech from two speakers [24]. The position of the target source remains fixed, whereas the interfering sources are positioned at angles ranging from 15° to 165° in 10° increments, allowing for interference from various directions. The number of interfering sources remains constant for each direction.

Both the target and interfering speeches are selected from the LibriSpeech corpus [25], using speaker combinations consistent with those in the Libri2Mix corpus [26]. The sampling frequency is set to 8 kHz, and the reverberation time is randomly chosen between 0.2 and 0.5 seconds. The dataset comprises 27000 samples for training, 6000 samples for validation, and 6000 samples for evaluation.

The model is trained for 200 epochs with a batch size of five, using the negative SI-SDR [27] as the loss function to measure the difference between the ground truth and the extracted speech. The Adam optimizer is employed with a learning rate of 0.001, and no weight decay is applied. To prevent convergence to local optima, the learning rate is halved if there is no improvement in validation performance for 20 epochs. If performance does not improve after 120 epochs, training is halted early.

B. Experimental Results

Table I presents the results from the target speaker extraction experiments. Based on these results, insights are organized to address the following three key questions: *i)* the trade-off between content consistency and distortion, *ii)* the approach to referencing auxiliary signals, and *iii)* the impact of different types of distortion in auxiliary signals.

1) Q1: Trade-off between Content Consistency and Distortion: We compared the use of undistorted pre-recorded speech (TSE-1), which differs in content from the target source, with pre-extracted speech containing distortions but aligned with the speech content (TSE-2 and TSE-3), referencing them at the statistical level. The results indicate that, in the trade-off between content consistency and distortion, the latter has a more favorable impact on model performance. This is likely because, when conditioning at the statistical level, the time-averaging process in the auxiliary network compresses the signal, thereby mitigating the adverse effects of distortion in the auxiliary signal.

2) Q2: Referecing Level of Auxiliary Singals: The results from using TSE-2 through TSE-5 indicate that models conditioned at the frame level outperform those conditioned at the statistical level when referencing pre-extracted speech, whether by IVA or IBM. This suggests that when the auxiliary signal is aligned with the target speech, learning local features that capture the temporal variations within the utterance is more effective than aggregating information to learn global features.

3) Q3: Type of Distortion in Auxiliary Singals: In a multi-channel environment, using IBM for source separation introduces specific artifacts due to nonlinear processing, whereas using IVA results in residual noise. The artifacts typically manifest as artificial, harsh sounds caused by missing spectral components of the target signal, while residual noise comprises leftover interference and environmental noise from the recording setup.

When comparing TSE-2 with TSE-3 and TSE-4 with TSE-5, it can be observed that at the statistical level, residual noise has a more detrimental impact on performance than artifacts, whereas at the frame level, artifacts are more problematic than residual noise. This suggests that for statistical-level conditioning, which captures global features of the auxiliary signal, it is crucial to ensure that no extraneous elements are present, leaving only the target speech. In contrast, for frame-level conditioning, which captures local features, it is more important that, even if extraneous elements are present, the target speech remains intact and clear.

IV. CONCLUSION

This study examined the effects of various types of auxiliary signals and conditioning methods on the performance of TSE models in a multi-channel environment. Our experiments revealed that pre-extracted signals with distortions, when aligned with the target speech, enhance TSE performance more effectively than undistorted but unaligned pre-recorded utterances. Frame-level conditioning proved particularly advantageous, as it captures local features and utilizes richer information compared to statistical-level conditioning. Furthermore, the type of distortion in the auxiliary signal was found to influence model performance: residual noise has a more adverse impact than artifacts in statistical-level conditioning, while the reverse holds

true for frame-level conditioning. These findings underscore the importance of maintaining the purity of auxiliary signals for global feature extraction and preserving the integrity of target speech information for local feature extraction, offering valuable insights for the design and optimization of TSE systems.

REFERENCES

[1] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 92–101, 2015.

[2] N. Kanda, C. Boeddeker, J. Heitkaemper, *et al.*, "Guided source separation meets a strong asr backend: Hitachi/paderborn university joint investigation for dinner party asr," in *CHiME-5 Workshop*, 2019.

[3] T. Nakatani, T. Yoshioka, M. Delcroix, and K. Kinoshita, "Neural target speech extraction: An overview," *arXiv preprint arXiv:2301.13341*, 2020.

[4] A. Ephrat, I. Mosseri, S. Lang, *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 11 792–11 802.

[5] J. Heitkaemper, T. Feher, M. Freitag, and R. Haeb-Umbach, "A study on online source extraction in the presence of changing speaker positions," in *International Conference on Statistical Language and Speech Processing 2019, Ljubljana, Slovenia*, 2019.

[6] P. Comon and C. Jutten, "Blind source separation: The state of the art," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 80–88, 2010.

[7] P. Bofill, J. M. Amigó, and E. Parra, "Residual crosstalk in blind source separation: Analysis and solutions," *Neurocomputing*, vol. 72, no. 1-3, pp. 1–13, 2008.

[8] K. Nishikawa and T. Yamasaki, "Nonlinear distortions in blind source separation algorithms: An experimental study," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 244–254, 2013.

[9] Z.-Q. Wang, H. Erdogan, S. Wisdom, *et al.*, "Sequential multi-frame neural beamforming for speech separation and enhancement," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021.

[10] H. Chen, Y. Yi, D. Feng, and P. Zhang, "Beam-guided tasnet: An iterative speech separation framework with multi-channel output," *arXiv preprint arXiv:2102.02998*, 2021.

[11] M. Elminshawi, S. R. Chetupalli, and E. A. P. Habets, "Beamformer-guided target speaker extraction," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023.

[12] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. ACSSC*, 2006, pp. 1393–1396.

[13] D. Wang, "Ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, pp. 181–197, 2005.

[14] J. B. Allen and L. R. Rabiner, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in *Neural computation*, MIT Press, vol. 9, 1997, pp. 1735–1780.

[16] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[17] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. ICASSP*, 2018, pp. 5554–5558.

[18] K. Žmolíková, M. Delcroix, K. Kinoshita, *et al.*, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[19] M. Delcroix, S. Watanabe, T. Ochiai, *et al.*, "End-to-end speakerbeam for single channel target speech recognition.," in *Proc. Interspeech*, 2019, pp. 451–455.

[20] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for speakerbeam target speaker extraction," in *Proc. ICASSP*, 2019, pp. 6965–6969.

[21] M. Delcroix, T. Ochiai, K. Zmolikova, *et al.*, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *Proc. ICASSP*, 2020, pp. 691–695.

[22] K. Vesely, S. Watanabe, K. Žmolíková, M. Karafiat, and L. Burget, "Sequence summarizing neural network for speaker adaptation," in *In Proc. ICASSP*, 2016.

[23] Y. L. *et al.*, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[24] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[26] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[27] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" In *Proc. ICASSP*, 2019, pp. 626–630.

Adaptive Spatial Re-sampling Method for Video Coding for Machines

Eun-Bin An*, Ayoung Kim*, Soon-heung Jung, Hyon-Gon Choo and Kwang-deok Seo*
* Yonsei University, Wonju, Korea
E-mail: {eunbin.an, aykim90, kdseo}@yonsei.ac.kr Tel: +82-33-760-2788
Electronics and Telecommunications Research Institute, Daejeon, Korea
E-mail: {zeroone, hyongonchoo}@etri.re.kr Tel: +82-42-860-6891

***Abstract*— As the performance of machine vision continues to improve, it is being used in various industrial fields to analyze and generate massive amounts of video data. Although the demand for and consumption of video data by machines has increased significantly, video coding for machines needs to be improved. Spatial re-sampling plays a critical role in video coding for machines because it reduces the volume of the video data to be processed while maintaining the shape of the data's features that are important for the machine to reference when processing the video. An effective method of determining the intensity of spatial re-sampling as an efficient coding tool for machines is still in the early stages. Here, we propose a method of determining an optimal scale factor for spatial re-sampling by collecting and analyzing information on the number of objects and the ratio of the area occupied by the object within a picture.**

I. INTRODUCTION

As deep-learning machine vision is now used widely in various industrial and research fields, the amount of video data being produced and distributed, not only by humans but also by machines, has been increasing rapidly. To manage these ever- growing volumes of data, many video codecs have been introduced, such as advanced video coding (AVC), high-efficiency video coding (HEVC), and versatile video coding (VVC) [1–2]. Research has also been progressing on perceptually optimized video coding. However, codecs based on human visual or sensory systems do not guarantee efficient compression of video data when used for deep-learning machine vision. There is therefore a need to study video coding that targets video data consumed by machines rather than humans [3-4]. An understanding of how machines interpret video data is needed to develop unprecedented coding technology for machines rather than for humans. However, a variety of machine vision systems have been designed, and how video data are under- stood differs among these system. It is essential to maintain as much as possible the contours and other common features that machines collect to interpret video while increasing compression efficiency. Spatial re-sampling can easily reduce the volume of video data while significantly maintaining the position or arrangement of objects and boundaries, which are common features in video data. In this paper, we apply a new spatial re-sampling method to each picture in the input video to maximize compression efficiency while maximizing retention of the common features that the machines need to interpret the video data. In the proposed method, the ratio of the area occupied by a detected object in a picture is calculated, and the optimal scale factor for spatial re-sampling for each picture is determined using an object occupancy distribution (OOD), which is generated based on the size and number of detected objects. Figure 1 is a flowchart for determining the optimal scale factor for spatial re-sampling based on the proposed OOD.

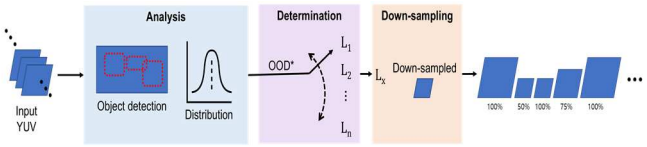


Fig. 1 Architecture of the proposed method for determining the optimal scale factor for spatial re-sampling based on OOD.

II. OOD FOR SPATIAL RE-SAMPLING

We used the test video sequences in the SFU-HW dataset, and explored the appropriate scale factors for spatial re-sampling for each video sequence through video data analysis and OOD generation.

A. Analysis of the impact of spatial re-sampling on machine vision Performance

Object detection was performed on the spatially up-sampled sequences at the same resolution as the original, and the performance results of the object detection accuracy (mAP) are organized by class and plotted in Figure 2. In Figure 2, several important phenomena are apparent. First, the scale factor for the spatial re-sampling increases as the object detection accuracy decreases because the boundaries of the image become increasingly blurred as the intensity of the spatial re-sampling strengthens.

Second, the machine accuracy drops significantly in a specific degree of re-sampling intensity depending on the class of the sequences, which are classified according to the resolution size.

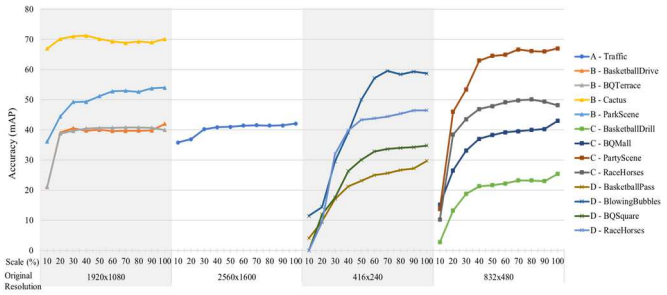


Fig. 2 Variation in object detection accuracy depending on scale factors for spatial re-sampling.

B. *OOD generation to determine an optimal re-sampling threshold*

In the analysis of Section A, the stronger the spatial re-sampling, the weaker the machine vision performance, and the object detection accuracy varies depending on the video resolution and the size and number of detected objects. The OOD is generated based on the detected object information including the size of detected objects within a single frame. The bounding box size of each detected object is divided by the original resolution size to obtain the object occupancy ratio (OOR). A histogram is then created depending on the OOR distribution. In simple terms, the OOR is the ratio of the detected object's size to the original resolution size, and the OOD is a histogram representing the distribution of the OOR. Figure 3 shows the experiment results related to OOD generation obtained by performing spatial down- sampling and object detection at increments of 10% from 10% to 100% of the original resolution for the 84th frame of the BasketballDrive sequence. Figure 3(a) depicts the results of object detection with bounding boxes for the spatial down-sampled pictures at various scales. Figure 3(b) depicts the generated OOD based on the detection results for each spatial down-sampled picture. It includes a histogram of object counts corresponding to various OOR ranges. Figure 3(c) is a graph expressing the kernel density estimation of Figure 3(b). Through Figure 3(c), the optimal re-sampling threshold can be explored visually. Figure 3(b) makes it clear that the distribution behaviors for the scale factors ranging from 50% to 100% are similar to each other. However, the distribution behaviors for scale factors below 50% are substantially different from those for the scale factors beyond 50%. We can therefore conclude that the optimal re-sampling threshold is 50%. Up to this threshold, we can keep the original OOD characteristics of the original picture with 100% resolution.

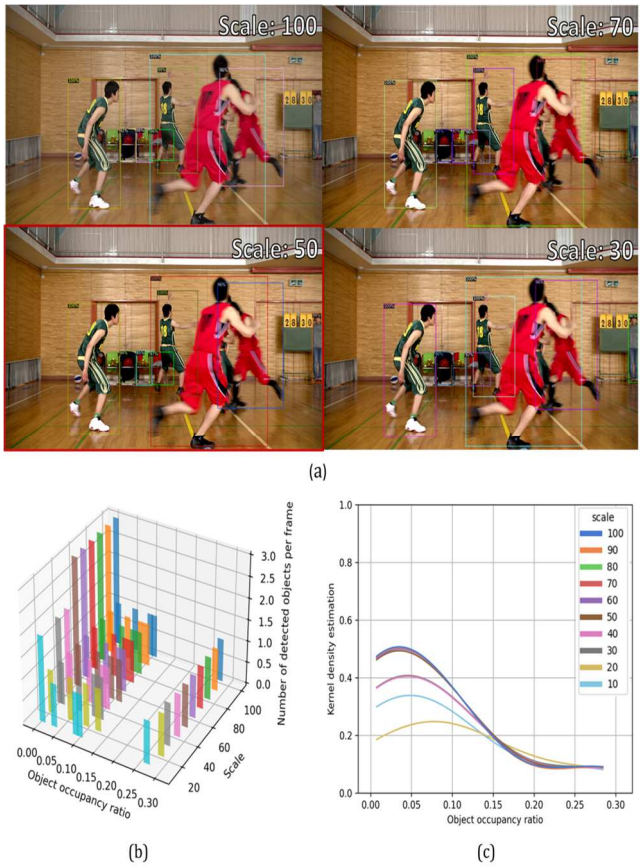


Fig. 3 Experiment results related to OOD generation for the 84th frame of the Bas- ketballDrive sequence: (a) Object detection results with bounding boxes for spatially down-sampled pictures with scale factors of 100%, 70%, 50%, and 30%, respectively, (b) Generated OODs and (c) Kernel density of the OOD.

III. DETERMINATION OF THE OPTIMAL SCALE FACTOR

Figure 4 is a flowchart of the process for determining the optimal scale factor for spatial re-sampling for each picture. The proposed process consists of four steps. First, the object detection network model generates object information, such as the bounding box size of detected object and the OOR for the original and spatial down- sampled pictures. Then, the OODs shown in Figure 4(b) are generated for all the candidate scale factors by using the object information. The next step is to analyze the correlation among the generated OODs to explore the similarity of the spatial down- sampled pictures to the original picture in terms of object detection performance. For analysis of similarity, we compared the correlation between the original picture with 100% resolution and the down-sampled picture with a scale factor x, where x is an element of the scale factor list SL which is {10%, 20%, 30%, . . . , 90%}. In this paper, the correlation was obtained using the OOD, which is the histogram representing the distribution of OOR.

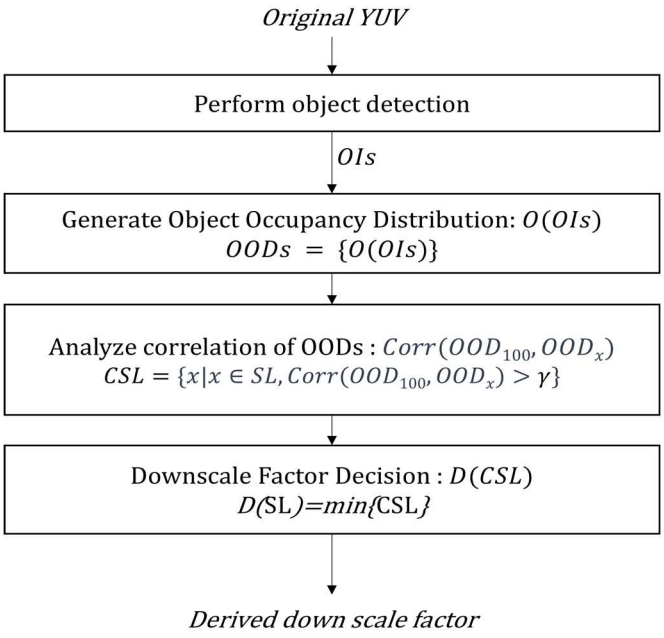


Fig. 4 Flowchart for determining the optimal scale factor for spatial re-sampling.

If the correlation exceeds the threshold value of $\gamma = 0.9$, the down-sampled picture is sufficiently similar to the original picture in terms of object detection. By gathering all the scale factors satisfying the threshold condition, we constitute the Correlated Scale- factor List (CSL). The CSL is a list of scale factors expected to have no significant difference in machine vision performance compared with the original picture. Among the CSL elements, the minimum scale factor is selected to be the optimal scale factor for re-sampling. If no element exists in the CSL, no definite optimal scale factor can be selected. In this case, the object detection performance even for the original picture with 100% resolution was found to be low. We applied the strongest spatial re-sampling ratio to achieve the utmost compression efficiency at the cost of slight machine performance loss.

Figure 5 is a conceptual diagram of an exemplary process for determining the opti- mal scale factor for spatial re-sampling among the candidate scale factors of 80%, 60%, 40%, and 20%. Because the OOD of the original picture and the OODs of the down-scaled pictures with 80% and 60% scale factors are estimated to be similar to each other, the scale factors of 80% and 60% are considered to be the candidates for an optimal scale factor for spatial re-sampling. Given the achievable compression efficiency, a scale factor of 60% corresponding to the lowest scale factor of the candidates was the optimal scale factor.

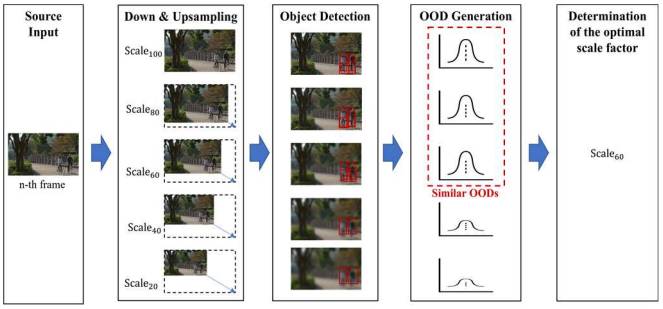


Fig. 5 Conceptual diagram of the proposed method for determining the optimal scale factor for frame-level spatial re-sampling.

IV. Experimental Results

This section describes the experimental verification of how much video data can be reduced and how much machine performance can be maintained by applying the proposed method. The experiment began by applying a re-sampling operation with a scale factor of {10%, 20%, 30%, . . . , 90%} to the original picture, after which object detection was performed using Faster-RCNN X101-FPN model. The proposed method determined the optimal scale factor for each picture. The machine accuracy error, ε, and the video data reduction ratio, C, were used to evaluate the performance of the proposed method. The machine accuracy error ε means the performance decrease in the object detection that occurs when the proposed spatial re-sampling with an optimal scale factor is applied compared with the object detection accuracy achieved for the original picture with 100% resolution (Equation 1). In this equation, $Accuracy_o$ represents the object detection accuracy for the original picture and $Accuracy_p$ is the object detection accuracy obtained for the down-sampled picture by the proposed method. The machine accuracy error ratio E_r is defined as the ratio between ε and $Accuracy_o$ as shown in Equation 2. C and C_r represent the data reduction ratio and rate, respectively, where C is the ratio of the original video data rate to the down-sampled video data rate, and C_r is the corresponding reduction rate expressed as a percentage. These are defined in Equations 3 and 4. From Equation 1 to 4 , the subscript o and p indicate the original picture and the down-sampled picture obtained by the proposed method, respectively

$$\varepsilon = \left|Accuracy_o - Accuracy_p\right| \geq 0 \quad (1)$$

$$E_r = \varepsilon / Accuracy_o \quad (2)$$

$$C = datarate_o / datarate_p \geq 1 \quad (3)$$

$$C_r = \left(1 - \frac{datarate_p}{datarate_o}\right) \times 100 \quad (4)$$

The data reduction–to–accuracy error ratio (*DRAER*) in Equation 5 is used to evaluate the achieved reduction in the video data volume against the obtained E_r.

$$DRAER = 10 * log_{10}(C/E_r) \quad (dB) \quad (5)$$

where the E_r excludes the case of value 0 indicating no change in the resolution of the video data after applying the proposed method, that cannot happen in practice.

Table I is the number of video frames classified to each scale factor for spatial down- sampling. The behavior of the classification depends highly on the characteristics of the video frames in terms of the size and the number of objects detected.

Table I. Number of video frames classified to each scale factor for spatial down- sampling.

Sequence		Number of video frames per each scale factor									
Class	name	100	90	80	70	60	50	40	30	20	10
A	Traffic	0	0	0	0	0	1	1	0	3	28
B	ParkScene	1	4	2	4	2	4	5	4	3	4
B	Cactus	13	11	1	5	8	0	0	4	10	45
B	BasketballDrive	8	13	3	5	11	9	11	22	15	0
B	BQTerrace	0	0	0	0	0	0	0	0	27	102
C	RaceHorsesC	62	12	11	6	2	2	1	1	0	0
C	BQMall	62	23	14	12	8	5	3	2	0	0
C	PartyScene	27	19	14	7	4	7	8	8	3	0
C	BasketballDrill	20	12	7	5	4	5	3	5	4	0
D	RaceHorsesD	40	19	12	8	4	4	5	4	1	0
D	BQSquare	21	31	9	11	5	11	39	2	0	0
D	BlowingBubbles	28	24	21	10	8	5	0	0	1	0
D	BasketballPass	19	10	6	7	6	7	6	4	0	0

Table II describes the experimental results obtained using the proposed method for the SFU-HW dataset. It compares the mAP for the original video with 100% resolution and the mAP for the down-sampled video by the proposed method, and shows the value of E_r, as well as *C*. In most cases, the E_r related to the ability to keep the object detection accuracy is less than 10%, which is a satisfactory performance. Only the Traffic and BQTerrace sequences have E_r values exceeding 10% (16.97% and 37.44%, respectively). Except for a few sequences, the E_r is stays within 10%, indicating that the machine vision performance can be stable for most test sequences. As for the exceptional cases in which the Traffic and BQTerrace sequences were used for the test, most of the video frames are classified into the 10% scale factor.

Table II. Experimental results of the proposed method for the SFU-HW dataset.

Sequence Name	Orig. (mAP)	Prop. (mAP)	E_r	*C*	C_r	*DRAER* (*dB*)
Traffic	42.13	34.98	16.97%	7.67	86.96%	16.55
ParkScene	54.07	49.14	9.11%	2.01	50.29%	13.44
Cactus	70.22	68.98	1.77%	2.44	59.07%	21.40
Basketball Drive	42.04	40.95	2.58%	1.91	47.71%	18.70
BQTerrace	40.04	25.05	37.44%	8.27	87.90%	13.44
RaceHorsesC	25.43	24.41	4.00%	1.09	8.60%	14.37
BQMall	43.02	41.39	3.80%	1.16	13.66%	14.84
PartyScene	67.05	65.67	2.06%	1.34	25.33%	18.13
Basketball Drill	48.22	49.77	3.23%	1.35	25.73%	16.20
RaceHorsesD	29.79	28.50	4.32%	1.21	17.30%	14.47
BQSquare	34.85	31.59	9.35%	1.46	31.42%	11.93
BlowingBubbles	58.73	59.72	1.69%	1.20	16.42%	18.50
Basketball Pass	46.51	44.93	3.40%	1.34	25.61%	15.97

V. CONCLUSIONS

In this paper, we devised an OOD to determine the optimal scale factor for spatial re-sampling to optimize video compression efficiency for machines. The OOD was created by applying the area information occupied by the detected object in the picture. By analyzing the similarity of OODs for various candidate scale factors between the down-sampled and the original pictures, we were able to obtain the optimal scale factor for the frame-level re-sampling.

VI. ACKNOWLEDGMENT

This work was supported by the Institute of Information and Communications Technology Planning & Evaluation (IITP) under grant number (2020-0-00011) and the Basic Science Research Program through the National Research Foundation of Korea under grant number (NRF-2021R1F1A1048404).

REFERENCES

[1] ISO/IEC: Information technology–coding of audio-visual objects–part 10: Advanced video coding. ISO/IEC 14496-10 (2022).

[2] ISO/IEC: Information technology–coded representation of immersive media–part 3: Versatile video coding. ISO/IEC 23090-3 (2022).

[3] H. Choi and I. Bajić, "Scalable image coding for humans and machines," IEEE Trans. Image Process. vol. 31, 2022, pp. 2739–2754.

[4] W. Yang, H. Huang, Y. Hu, L. Duan, and J. Liu, "Video coding for machines: Compact visual representation compression for intelligent collaborative analytics," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 46, July 2024, pp. 5174-5191.

A Multi-Perceptual Learning Network for Retina OCT Image Denoising and Classification

Zhe Xiao*, Zongqi He*, Zhuoning Xu*, Yunze Li*, Zelin Song*, Calvin Leighton†, Li Wang*, Shanru Liu*
and Shiun Yee Wong†, Wenfeng Huang†, Wenjing Jia†, and Kin-Man Lam*
* Hong Kong Polytechnic University, Hong Kong
† University of Technology Sydney, Australia

***Abstract*—Swept Source Optical Coherence Tomography (OCT), a non-invasive cross-sectional imaging technique, has been widely used in diagnosing and treating various vision-related diseases. However, OCT images often suffer from heavy noise issues, due to the limitations of imaging devices, making analysis and disease classification a great challenge. This paper proposes a Multi-Perceptual Learning Network (MPLN) for retina OCT image denoising and classification. We adopt a triplet cross-fusion GAN approach and use three unpaired OCT images to conduct perceptual learning. In addition, we integrate the Frequency Distribution Loss into GAN to preserve both the structural integrity and perceptual quality of the denoised OCT images, enabling better classification. The method can significantly reduce the noise of highly noisy images. Our proposed method is evaluated on the VIP Cup 2024 dataset in terms of the CNR, MSR, and TP scores. Our model achieves a CNR score of 6.351, and an MSR score of 11.573, which outperforms many existing methods on OCT images. In classification, our MPLN improves accuracy by more than one percent. These results demonstrate that our model can significantly enhance image quality and improve classification accuracy, highlighting its potential for clinical applications.**

I. Introduction

Optical coherence tomography (OCT), a non-invasive cross-sectional imaging technique, plays a crucial role in various medical applications, particularly retinal diagnostics [1]. However, two main issues hinder the development of OCT-based diagnosis. First, OCT images are inevitably corrupted by heavy speckle noise due to the low coherence interferometry imaging modality [2]. Second, OCT image analysis heavily relies on manual labor, which increases the burden on ophthalmologists and leads to potential bias due to subjective opinions. Therefore, an automatic and reliable OCT image analysis algorithm is essential for the efficient diagnosis of eye diseases, which will help alleviate the strain on medical resources.

Over the past decades, several approaches have been proposed to enhance the quality of OCT images. One approach to denoise OCT images involves capturing a sequence of repeated B-scans from the same position and then registering and averaging them to produce a less noisy image [3]. However, this method significantly increases the image acquisition time, making it impractical for clinical use. Another approach is model-based OCT denoising, where a single B-scan image is denoised using digital filters based on the statistical models of signal and noise [4], [5], or through deep/dictionary learning techniques [6], [7]. While these methods can reduce noise, they often result in over-smoothing or loss of fine details. Inspired by the remarkable success of deep learning in various vision tasks, deep learning-based methods for OCT denoising have shown superior performance in both speckle noise reduction and structure preservation. Researchers have improved denoising performance by optimizing deep network architectures [8], [9] and designing structure-sensitive loss functions [10], [11]. However, these methods require paired noisy-clean OCT images for training, which are difficult to obtain due to involuntary eye or body movements during scanning [12].

In addition to image reconstruction, OCT image analysis, especially classification, is also a highly important topic in clinical applications. This field has witnessed a growing number of diverse OCT classification methods over the past years [13]–[15]. Some of them are relatively straightforward, such as the utilization of Support Vector Machines (SVMs) and Random Forests [16], [17], while others are more effective and sophisticated, leveraging significant advancements in the computer vision community [18]. For example, as one of the most prominent deep learning techniques, Convolutional Neural Networks (CNNs) have been widely investigated and explored for OCT image classification. However, when CNNs are applied to noisy OCT images, their classification performance is notably degraded. Applying conventional image processing techniques, including denoising, to OCT images may also lead to the loss or alteration of critical details, ultimately affecting classification results. Additionally, it has been demonstrated that some OCT image classification algorithms can only perform well on specific datasets, and their performances heavily depend on the characteristics of the data. To address the abovementioned issues, we propose a Multi-Perceptual Learning Network (MPLN) for retina OCT image processing, which jointly tackles the challenges in denoising and classification of OCT images.

In the initial stage of our model, a Generative Adversarial Network (GAN), comprising a DnCNN [19] generator and a PatchGAN [20] discriminator, is employed. We incorporate the Frequency Distribution (FD) Loss [21] into the network. This FD Loss function introduces perceptual learning by converting image features into the frequency domain and computing the distribution distance. This process not only enhances the perceptual quality, but also safeguards the structural integrity of images. The generator loss function includes GAN loss, consistency loss, FD loss, PSNR loss, and SSIM loss. In ad-

dition, inspired by the Triplet Cross-Fusion Learning (TCFL) scheme [22], our model accepts a triplet combination of inputs, consisting of two noisy images and one clean image, which are unpaired. We enable perceptual learning between the input images to help the network extract the most relevant and significant features. This triplet input strategy significantly expands the training dataset by incorporating an arbitrary number of noisy and clean images. These OCT images do not need to be matched pairs and there is no restriction on the quantity. This broadens the scope and flexibility of our training mechanism.

In the second stage, for retina OCT image classification, we adopted and further improved the Lesion-Aware Convolutional Neural Network (LACNN) [13]. Inspired by the remarkable performance achieved by deep learning models in various vision tasks [23]–[27], we applied the Residual Network (ResNet) to the Lesion Detection Network (LDN) [13] to generate an attention map that highlights the lesion areas in each OCT image. This attention map is then used to compute a weighted feature map, which is subsequently combined with the original convolutional feature maps to obtain the lesion-aware feature map. This map serves as input to the subsequent convolutional layers, allowing the model to focus on clinically relevant lesion regions during classification.

In this study, we advance the LACNN architecture by replacing the original backbone of the LACNN with ResNet, which significantly enhances the model's ability to accurately classify OCT images. This improvement provides a valuable tool for the early detection and differentiation of ocular diseases in clinical practice. In addition, we demonstrate that the MPLN model enhances image perceptual quality. Our model facilitates multi-perceptual learning among images, providing effective denoising and improving the perceptual quality of images. FD loss serves as a crucial element for improving image quality, and this enhancement is validated through classification experiments. The results show that the model trained on denoised images by the MPLN model achieves higher accuracy, confirming the effectiveness of our approach.

In summary, our main contributions are as follows:

- We propose an MPLN framework that effectively addresses both OCT image denoising and classification tasks in a unified approach, leading to improved overall performance in image interpretation.
- We introduce the FD Loss into the GAN architecture, which helps preserve the structural integrity of OCT images during denoising. This facilitates multi-perceptual learning, enhancing both the quality of the denoised images and the classification accuracy.
- We incorporate LDN with ResNet to obtain the refined lesion-aware attention map, which is then combined with LACNN to classify the OCT images more accurately.

II. The Proposed MPLN Method

A. MPLN Denoising

1) Network Architecture: Our proposed MPLN method makes use of a triplet cross-fusion GAN [22] and employs three unpaired OCT images for perceptual learning. In this approach, a speckle noise-corrupted OCT image X can be expressed as the sum of its clean component C and its speckle noise Y, as follows:

$$X = C + Y. \tag{1}$$

Specifically, in our approach, the clean components and the noisy components for training consist of three unpaired images: two noisy OCT images X_1 and X_2, and one clean OCT image C_0. These three unpaired images undergo multi-perceptual learning to help the network extract and learn the speckle noise properties, thereby improving denoising accuracy even with a small dataset. We used a total of 5 clean images and 10 noisy images, all of which are unpaired. The total number of combinations of clean and noisy images is $\binom{5}{1} \times \binom{10}{2} = 225$, which significantly enlarges the training set.

In addition, our proposed approach employs a GAN-based denoising framework, which consists of three key components, as shown in Fig. 1: generating clean images from noisy images, synthesizing noisy images from clean images by adding estimated noises, and generating clean images from synthesized noisy images.

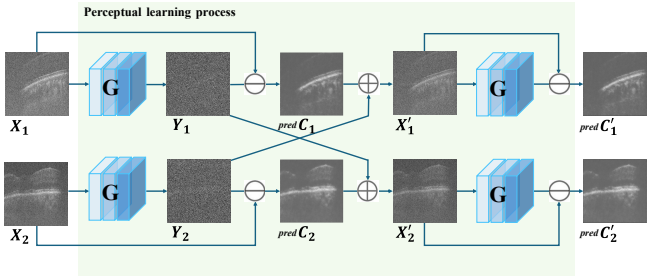
Fig. 1(a) shows the first component of the network, generation of predicted clean images from noisy images, where G represents the Generator. The generator first takes in the two noisy inputs X_1 and X_2. After passing through a perceptual learning process, the first component outputs predicted clean images predC_1' and predC_2'. In the perceptual learning process, the outputs of the generator Y_1 and Y_2 are the estimated noise component of the noisy inputs X_1 and X_2. Subtracting the estimated noise components from themrespectively produces the estimated clean images of the two noisy inputs predC_1 and predC_2. In the cross-fusion mechanism, the generated clean images predC_1 and predC_2 are then added with the estimated noise of the other noise components Y_2 and Y_1 respectively to produce two more synthesized noisy images X_1' and X_2', as the input for the following perceptual learning process.

Similar to the prediction from noisy images, we can add the estimated noise Y to the input clean image to synthesize a fake noisy picture X'. Then, a generator can produce their predicted clean images predC'. Fig. 1(b) illustrates the second component of the network, generation of predicted clean images from a clean image. The noises Y_1 and Y_2, predicted by the generator in the first component, are used in this component. The input is the clean image C_0, and the outputs are predC_3' and predC_4'.

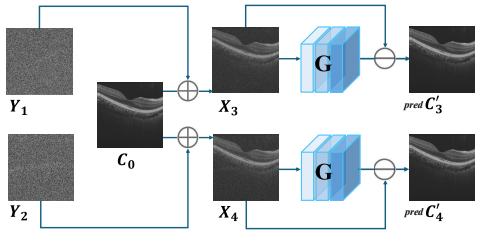
The Discriminator's role is to differentiate between generated images and clean images. As shown in Fig. 1(c), the predicted images from the first two components of the network, together with the clean image C_0, are inputted to the Discriminator D.

2) Loss Functions: The total loss of MPLN is defined as follows:

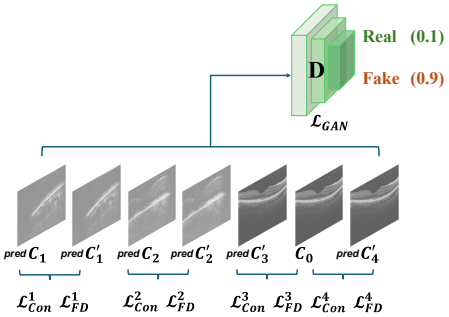
$$\mathcal{L}_{\text{Total}} = \alpha\mathcal{L}_{GAN} + \beta\mathcal{L}_{Con} + \gamma\mathcal{L}_{FD} + \delta\mathcal{L}_{PSNR} + \epsilon\mathcal{L}_{SSIM}, \tag{2}$$



(a) Generation of predicted clean images from noisy images.



(b) Generation of predicted clean images from a clean image.



(c) The Discriminator and the calculation of Consistency Losses.

Fig. 1. The MPLN triplet cross-fusion mechanism and GAN components.

where α, β, γ, δ, and ϵ are the weights for the respective loss functions. The GAN loss $\mathcal{L}_{GAN}$ attempts to minimize $\mathbb{E}\left[(D(c)-1)^2\right]$ for the generator and minimize $\mathbb{E}\left[(D(C_0)-1)^2+D(c)^2\right]$ for the discriminator, where c denotes an element within the set $\{\text{pred}C_1, \text{pred}C_1', \text{pred}C_2, \text{pred}C_2', \text{pred}C_3', \text{pred}C_4'\}$.

$\mathcal{L}_{Con}$ is the Consistency Loss; $\mathcal{L}_{FD}$ indicates the FD Loss [21] for computing the distribution distance in the frequency domain; $\mathcal{L}_{PSNR}$ and $\mathcal{L}_{SSIM}$ denote PSNR and SSIM losses, respectively.

The spatial Consistency Loss $\mathcal{L}_{Con}$ is designed to differentiate the real clean image C from the generated clean images $\text{pred}C_1$, $\text{pred}C_2$, $\text{pred}C_3'$ and $\text{pred}C_3'$, and is computed as follows:

$$\mathcal{L}_{Con} = \mathcal{L}^1_{Con} + \mathcal{L}^2_{Con} + \mu_{Con}(\mathcal{L}^3_{Con} + \mathcal{L}^4_{Con}). \quad (3)$$

where $\mathcal{L}^1_{Con}$, $\mathcal{L}^2_{Con}$, $\mathcal{L}^3_{Con}$ and $\mathcal{L}^4_{Con}$ are computed as follows:

$$\begin{aligned}
\mathcal{L}^1_{Con} &= \| \text{pred}C_1 - \text{pred}C_1' \|_1 \\
\mathcal{L}^2_{Con} &= \| \text{pred}C_2 - \text{pred}C_2' \|_1 \\
\mathcal{L}^3_{Con} &= \| \text{pred}C_3' - C_0 \|_1 \\
\mathcal{L}^4_{Con} &= \| \text{pred}C_4' - C_0 \|_1
\end{aligned} \quad (4)$$

where $\| \|_1$ denotes the $L1$ norm.

In addition to the spatial domain consistency, to achieve better perceptual quality, we adopt the FD Loss [21] to compute the distribution distances $\mathcal{L}_{FD}$ of the predicted clean images in the frequency domain, which can be calculated as:

$$\mathcal{L}_{FD} = \mathcal{L}^1_{FD} + \mathcal{L}^2_{FD} + \mu_{FD}(\mathcal{L}^3_{FD} + \mathcal{L}^4_{FD}), \quad (5)$$

where μ_{FD} is the loss weight, and $\mathcal{L}^1_{FD}$, $\mathcal{L}^2_{FD}$, $\mathcal{L}^3_{FD}$ and $\mathcal{L}^4_{FD}$ can be computed as follows:

$$\begin{aligned}
\mathcal{L}^1_{FD} &= \text{fdl}(\text{pred}C_1 - \text{pred}C_1'), \\
\mathcal{L}^2_{FD} &= \text{fdl}(\text{pred}C_2 - \text{pred}C_2'), \\
\mathcal{L}^3_{FD} &= \text{fdl}(\text{pred}C_3' - C_0), and \\
\mathcal{L}^4_{FD} &= \text{fdl}(\text{pred}C_4' - C_0),
\end{aligned} \quad (6)$$

where $fdl()$ represents the FD loss function, which uses a pre-trained feature extractor to map the predicted and target images to a feature space. Discrete Fourier Transform (DFT) is applied, followed by Sliced Wasserstein Distance to measure the distance among the target images in the frequency domain. This approach effectively mitigates the interference of spatial misalignment, thereby better capturing the perceptual properties of the images. The combinations of $\mathcal{L}^1_{FD}$, $\mathcal{L}^2_{FD}$, $\mathcal{L}^3_{FD}$, $\mathcal{L}^4_{FD}$, together with $\mathcal{L}^1_{Con}$, $\mathcal{L}^2_{Con}$, $\mathcal{L}^3_{Con}$ and $\mathcal{L}^4_{Con}$ facilitate the multi-perceptual learning of noises from different components.

B. Retina OCT Image Classification

To demonstrate the efficacy of the denoising model on the downstream OCT image classification task, we designed a classification model specially for denoised retina OCT images.

The core modules of our classification model, termed "LACNN-ResNet", are illustrated in Fig. 2. We adopted and enhanced the LACNN approach [13] by incorporating residual connections to learn more discriminative features for classification. Additionally, we embedded the Lesion Attention Network (LAN) [13] to strengthen the local lesion-related features while also considering the global structures of the OCT images.

For a lesion attention block, denote x as the input image tensor, $Con_{i,c}(x)$ as the output of a convolutional layer, and A_i as the attention map generated by LAN, where i represents the spatial position and c denotes the channel index. The weighted feature map, represented by $W_{i,c}$, is defined as follows:

$$W_{i,c}(x) = Con_{i,c}(x) \times A_i, \quad (7)$$

where $\times$ denotes the element-wise product of corresponding spatial positions from attention maps and feature maps.

In addition, the feature maps generated by convolution are superimposed with the weighted feature map of the corresponding spatial positions. This is based on the consideration that LAN can only mark lesion areas, while unmarked areas may also contain information useful for classification. Thus, the output of an LAN, denoted by $L_{i,c}(x)$, can be represented as an element-wise summation of the corresponding spatial positions from the attention maps and the weighted feature maps, as follows:

$$L_{i,c}(x) = W_{i,c}(x) + Con_{i,c}(x). \quad (8)$$

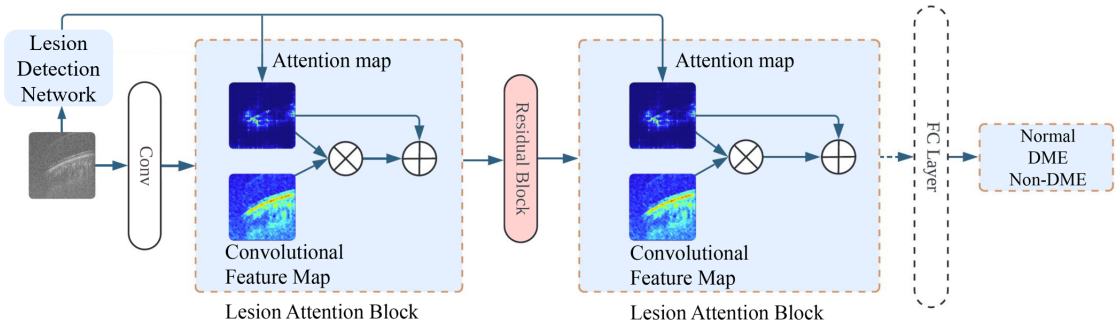


Fig. 2. The core modules of the proposed LACNN-ResNet classification model.

III. Experimental Setup and Results

A. MPLN Denoising

1) Experimental Setup: In the proposed MPLN method, we adopted DnCNN [19] as the generator and PatchGAN [20] as the discriminator. DnCNN is a classical denoising network and serves as a noise predictor in this GAN-based network.

We implemented MPLN with PyTorch and used an NVIDIA GTX 4090 GPU for training. The number of training epochs was set to 50. We opted for cosine annealing learning rate and set the cosine decay period to match the entire training cycle. An initial learning rate was set to 2e-5, gradually decaying to the final value of 1.2e-6. We empirically set α=1, β=3, γ=0.03, δ=0.5, and ϵ=1 as the weights for the generator loss function.

During testing, input images from the VIP Cup 2024 Test Set were normalized to the range $[1, 3]$. A dataloader was established to load a triplet set of input images. We evaluated the proposed MPLN method using the test set, which includes B-scans from 18 subjects. Each subject's dataset consists of 70 to 300 B-scans, which are noisy and have resolutions of either 300 × 150 or 300 × 200 pixels.

2) Evaluation Metrics: To assess the performance of denoising methods, we adopted the Contrast-to-Noise Ratio (CNR), Mean-to-Standard-deviation Ratio (MSR), and Texture Preservation (TP) as evaluation metrics.

CNR evaluates the contrast between the foreground and the background of generated images, which is calculated as follows:

$$S_{\text{CNR}} = 10 \log \left| \frac{\mu_f - \mu_b}{\sqrt{\sigma_f^2 + \sigma_b^2}} \right|, \tag{9}$$

where μ_f and σ_f denote the mean and the standard deviation of the foreground regions, respectively, and μ_b and σ_b denote the mean and the standard deviation of the background region, respectively. The Regions of Interest (ROIs) are chosen between different layers to show the change in contrast.

Similarly, **MSR** evaluates the concentration of pixel intensity values and is calculated as follows:

$$S_{\text{MSR}} = \frac{\mu_f}{\sigma_f}. \tag{10}$$

TP evaluates the conserving of texture between the noisy image and the denoised image and is calculated as follows:

$$S_{\text{TP}}^{(m)} = \frac{\sigma_{m(de)}^2}{\sigma_{m(noisy)}^2} \sqrt{\frac{\mu_{de}}{\mu_{noisy}}}, \tag{11}$$

where $\sigma_{m(noisy)}$ and $\sigma_{m(de)}$ denote the standard deviation of the m-th ROI in the noisy image and denoised image, respectively, and μ_{de} and μ_{noisy} denote their corresponding means. The ROIs encompass the intra-layer regions. We take the average of all the ROIs in the image as its TP score:

$$S_{\text{TP}} = \frac{1}{M} \sum_{m=1}^{M} S_{\text{TP}}^{(m)}. \tag{12}$$

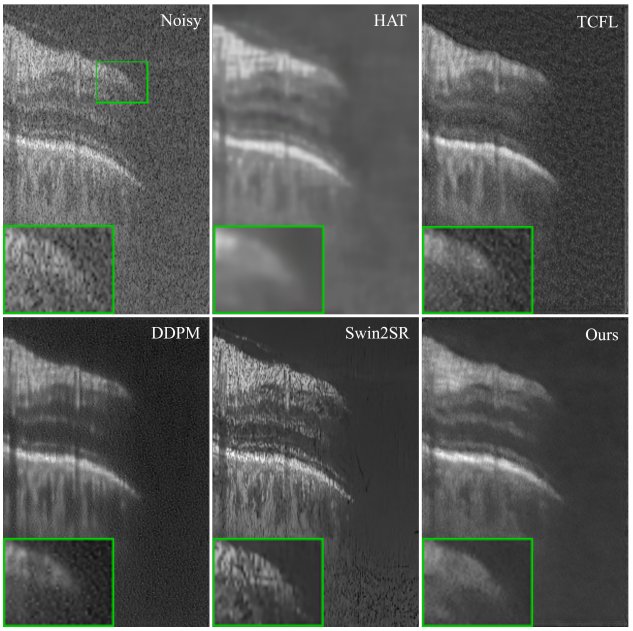


Fig. 3. Visual comparison of the denoising results of HAT [28], TCFL [22], DDPM [29], Swin2SR [30] and our MPLN.

3) Denoising Results and Comparison: We selected a representative retina OCT image from the VIP Cup 2024 Test Set and visually compared its denoising result with other existing approaches in Fig. 3.

Both DDPM and HAT have reduced the speckle noise to a certain extent. However, due to the heavy noise of the original noisy image, some noise still remains, as highlighted by the enlarged area in Fig. 3. While HAT produced smoother images compared to our MPLN method, they suffer from excessive

smoothness, which leads to the loss of critical structural details or blood flow structure, as indicated by the box in Fig. 3. The proposed MPLN effectively reduces the noise and improves perceptual quality of the image.

Table I shows the quantitative results comparing the denoising performance of our MPLN with other methods. As shown in this table, our MPLN obtained the best scores across CNR and MSR. Although TCFL, DDPM, and Swin2SR have marginally higher TP scores, they all obtain lower CNR and MSR scores. HAT yields a low TP score, indicating its over-smoothing problem. Our MPLN achieves balanced scores across the evaluation matrices, effectively reducing the noise.

TABLE I
QUANTITATIVE COMPARISON OF THE DENOISING PERFORMANCE OF OUR MPLN WITH SOTA APPROACHES HAT [28], TCFL [22], DDPM [29] AND SWIN2SR [30] ON THE VIP CUP 2024 TEST SET.

	Noisy	HAT	TCFL	DDPM	Swin2SR	**MPLN (ours)**
CNR↑	1.966	3.579	3.184	2.513	4.079	**6.351**
MSR↑	6.872	9.702	7.972	7.623	10.302	**11.573**
TP↑	1.000	0.342	**0.685**	0.670	0.578	0.554

B. MPLN-denoised Retina OCT Image Classification

1) Experimental Setup: The proposed MPLN classification was implemented using the same PyTorch framework and GPU as in the earlier stage. We evaluated the proposed LACNN-ResNet method utilizing the VIP Cup training dataset, which includes approximately 18,000 retina OCT B-scan images from 100 patients. We applied our denoising method to the original images and used them as training samples.

We used 90% of the VIP Cup 2024 training dataset for training and the remaining 10% for testing. We evaluated and analyzed the performance of our model under these two configurations. To further enhance the training process, we applied data augmentation to the dataset. Specifically, all input images were resized, horizontally flipped, and rotated within a range of -15 to +15 degrees, followed by normalization.

To further verify the improvement of the MPLN model over mainstream denoising models, we randomly selected a set of noisy images and used them as input for both the TCFL model and the MPLN model, generating two separate test sets. We then tested these sets using the LACNN-ResNet model, which had been trained on noisy images, and compared their accuracy.

2) Classification and Comparison: For each input image, the model outputs the predicted probability distributions across three classes: Normal, DME, and Non-DME. We determine the class label according to the highest proportion of predictions from the patient's OCT images.

Two models were trained for comparison. The first model on the original noisy VIP Cup 2024 training set, and the second model on the denoised version of the same dataset. After each epoch, testing was conducted on a test set approximately 1/9 the size of the training set (around 2,000 images). Results shows that the LACNN-ResNet model trained on the denoised dataset exhibited a 1.45% improvement in accuracy compared to the model trained on noisy data. In addition, the classification accuracy of MPLN denoised dataset is 5.82% higher than that of TCFL. These results demonstrate that the MPLN model effectively enhances classification performance for OCT images by reducing noise interference.

IV. CONCLUSION

In this paper, we have proposed a Multi-Perceptual Learning Network (MPLN) for denoising retina OCT images. To capture perceptually significant features for denoising, our MPLN model employs a cross-fusion GAN approach, utilizing three unpaired noisy and clean OCT images for multi-perceptual learning, while enforcing image consistency in both spatial and frequency domains by minimizing frequency distribution losses. Our MPLN has shown its ability to effectively reduce noise, enhance OCT image quality, and improve retina OCT image classification for eye disease diagnosis. Experimental results on the VIP Cup 2024 dataset have demonstrated that the proposed method outperforms state-of-the-art image denoising models both visually and quantitatively in this challenge.

ACKNOWLEDGMENT

We would like to extend our heartfelt thanks to Jun Xiao, Cong Zhang, and Cuixin Yang, whose expertise and willingness to share knowledge have greatly contributed to our understanding and application of the methodologies involved in this collaborative project. Additionally, we are immensely grateful to the organizers of the VIP Cup 2024 for providing a key dataset that enabled us to test our methods on real-world data. Finally, we acknowledge the support of our colleagues and peers who have provided constant encouragement and assistance throughout this journey.

REFERENCES

[1] W. Drexler and J. G. Fujimoto, "State-of-the-art retinal optical coherence tomography," *Progress in retinal and eye research*, vol. 27, no. 1, pp. 45–88, 2008.

[2] G. Gong, H. Zhang, and M. Yao, "Speckle noise reduction algorithm with total variation regularization in optical coherence tomography," *Optics express*, vol. 23, no. 19, pp. 24 699–24 712, 2015.

[3] A. W. Scott, S. Farsiu, L. B. Enyedi, D. K. Wallace, and C. A. Toth, "Imaging the infant retina with a hand-held spectral-domain optical coherence tomography device," *American journal of ophthalmology*, vol. 147, no. 2, pp. 364–373, 2009.

[4] M. H. Eybposh, Z. Turani, D. Mehregan, and M. Nasiriavanaki, "Cluster-based filtering framework for speckle reduction in OCT images," *Biomedical optics express*, vol. 9, no. 12, pp. 6359–6373, 2018.

[5] S. Aja-Fernández and C. Alberola-López, "On the estimation of the coefficient of variation for anisotropic diffusion speckle filtering," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2694–2701, 2006.

[6] L. Fang, S. Li, Q. Nie, J. A. Izatt, C. A. Toth, and S. Farsiu, "Sparsity based denoising of spectral domain optical coherence tomography images," *Biomedical optics express*, vol. 3, no. 5, pp. 927–942, 2012.

[7] L. Fang, S. Li, R. P. McNabb, *et al.*, "Fast acquisition and reconstruction of optical coherence tomography images via sparse representation," *IEEE transactions on medical imaging*, vol. 32, no. 11, pp. 2034–2049, 2013.

[8] M. Xu, C. Tang, F. Hao, and et al., "Texture preservation and speckle reduction in poor optical coherence tomography using the convolutional neural network," *Medical Image Analysis*, vol. 64, p. 101 727, 2020.

[9] Z. Shen, M. Xi, C. Tang, M. Xu, and Z. Lei, "Double-path parallel convolutional neural network for removing speckle noise in different types of OCT images," *Applied Optics*, vol. 60, no. 15, pp. 4345–4355, 2021.

[10] Y. Ma, X. Chen, W. Zhu, X. Cheng, D. Xiang, and F. Shi, "Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN," *Biomedical optics express*, vol. 9, no. 11, pp. 5129–5146, 2018.

[11] B. Qiu, Z. Huang, X. Liu, *et al.*, "Noise reduction in optical coherence tomography images using a deep neural network with perceptually-sensitive loss function," *Biomedical optics express*, vol. 11, no. 2, pp. 817–830, 2020.

[12] Y. Huang, W. Xia, Z. Lu, *et al.*, "Noise-powered disentangled representation for unsupervised speckle reduction of optical coherence tomography images," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2600–2614, 2020.

[13] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1959–1970, 2019.

[14] R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh, "Macular OCT classification using a multi-scale convolutional neural network ensemble," *IEEE Trans. Medical Imaging*, vol. 37, no. 4, pp. 1024–1034, 2017.

[15] X. He, Y. Deng, L. Fang, and Q. Peng, "Multi-modal retinal image classification with modality-specific attention network," *IEEE Trans. on Medical Imaging*, vol. 40, no. 6, pp. 1591–1602, 2021.

[16] B. Zagajewski, M. Kluczek, E. Raczko, A. Njegovec, A. Dabija, and M. Kycko, "Comparison of random forest, support vector machines, and neural networks for post-disaster forest species mapping of the krkonoše/karkonosze transboundary biosphere reserve," *Remote Sensing*, vol. 13, no. 13, 2021.

[17] A. Lang, A. Carass, E. Sotirchos, P. Calabresi, and J. L. Prince, "Segmentation of retinal OCT images using a random forest classifier," in *Medical Imaging 2013: Image Processing*, SPIE, vol. 8669, 2013, pp. 199–205.

[18] X. Huang, Z. Ai, H. Wang, *et al.*, "Gabnet: Global attention block for retinal OCT disease classification," *Frontiers in Neuroscience*, vol. 17, p. 1 143 422, 2023.

[19] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," en, *2017 ICCV*, Oct. 2017.

[21] Z. Ni, J. Wu, Z. Wang, W. Yang, H. Wang, and L. Ma, "Misalignment-robust frequency distribution loss for image transformation," *arXiv preprint arXiv:2402.18192*, 2024.

[22] M. Geng, X. Meng, L. Zhu, *et al.*, "Triplet cross-fusion learning for unpaired image denoising in optical coherence tomography," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3357–3372, 2022.

[23] J. Xiao, W. Jia, and K.-M. Lam, "Feature redundancy mining: Deep light-weight image super-resolution model," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 1620–1624.

[24] J. Xiao, Q. Ye, R. Zhao, K.-M. Lam, and K. Wan, "Self-feature learning: An efficient deep lightweight network for image super-resolution," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4408–4416.

[25] J. Xiao, X. Jiang, N. Zheng, *et al.*, "Online video super-resolution with convolutional kernel bypass grafts," *IEEE Transactions on Multimedia*, vol. 25, pp. 8972–8987, 2023.

[26] J. Xiao, Q. Ye, T. Liu, C. Zhang, and K.-M. Lam, "Deep progressive feature aggregation network for multi-frame high dynamic range imaging," *Neurocomputing*, vol. 594, p. 127 804, 2024.

[27] J. Xiao, Z. Lyu, C. Zhang, Y. Ju, C. Shui, and K.-M. Lam, "Towards progressive multi-frequency representation for image warping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2995–3004.

[28] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *2023 CVPR*, 2023, pp. 22 367–22 377.

[29] D. Hu, Y. K. Tao, and I. Oguz, "Unsupervised denoising of retinal OCT with diffusion probabilistic model," in *Medical Imaging 2022: Image Processing*, O. Colliot and I. Išgum, Eds., International Society for Optics and Photonics, vol. 12032, SPIE, p. 1 203 206.

[30] M. V. Conde, U.-J. Choi, M. Burchi, and R. Timofte, "Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration," in *ECCV*, Springer, 2022, pp. 669–687.

Source Attribution for Images Generated by Diffusion-Based Text-to-Image Models: Exploring the Forensics Approach

Xinqi Jiang[1], Lingyun Tian[1], Teng Sio Hong[2], Jinyu Tian[1]*
[1] Macau University of Science and Technology, Macau
E-mail: {3240003534, 2230015805}@student.must.edu.mo, jytian@must.edu.mo
[2] Faculty of Humanities and Social Sciences, Macao Polytechnic University, Macau
E-mail: p1308598@mpu.edu.mo

***Abstract*—As the image generation technology continues to advance, images generated by artificial intelligence have become ubiquitous on the internet, leading to a plethora of controversies regarding image copyright. Addressing this issue, tracing back to the source model of generated images has become a crucial approach to resolving it, aiding in establishing the legitimacy and ownership of images. Most existing approaches primarily focus on discerning image authenticity while overlooking the issue of source attribution for generated images. Few traditional image attribution methods are constrained by temporal limitations, only applicable to attributing past generative models such as GANs, while their effectiveness is limited for newer methods like diffusion generative models that have emerged in recent years. This paper proposes a novel method capable of attributing images generated by text-to-image diffusion models and maintaining effectiveness even for untrained models. We construct a semantic feature extraction network and train a feature-difference recognition model using a siamese network approach based on the semantic similarity of generated images. We collect and construct a dataset for training and testing, validating the outstanding performance of our method. Our method enhances compatibility with black-box models, ensuring effective source identification for images generated through text input. Through rigorous experimental validation, our method demonstrates significant progress, providing effective forensics approach for attributing generated images.**

I. Introduction

Over recent years, significant advancements have been observed in text-to-image generation models. Notably, models such as Stable Diffusion [1], DALL-E [2], Imagen [3], and others have demonstrated the capability to produce high-quality, lifelike images. A model in this domain receives a prompt, characterized as a textual segment, along with random noise as inputs. Subsequently, the model executes denoising of the image, guided by the provided prompt, ensuring the resultant image aligns with the given prompt.

With the proliferation of generated images flood the internet, there arises controversy concerning copyright issues and the difficulty in tracing their origins [4]. Some advertisers have used unauthorized models to generate commercial advertising images, and the authors of the models hope to receive payment from the advertisers. However, proving that these images were generated by their models is an extremely challenging task, as attributing the exact source of image generation can be restricted by technical limitations and data privacy [5]. The example highlights the complex challenges posed by copyright disputes and traceability issues in an era of AI-generated images. Consequently, research on attributing synthetic image sources is crucial for forensic investigations aimed at addressing infringement activities. Additionally, it plays a pivotal role in fostering the establishment of a more equitable and intellectual property-respecting society.

However, although the considerable work [6]–[10] devoted to identifying fake images, research on tracking remains limited [11]–[13], with the majority of existing studies focused on images generated by GAN [14]-based models. Most method for detecting images generated by GAN-based model exhibit limited efficacy when applied to diffusion-based generation models [15], rendering them inadequate for traceability.

In this paper, our main contributions are as follows:

- We propose a novel method for source attribution of synthetic images generated by diffusion-based text-to-image models, which can be naturally applied to real attribution scenarios.
- We propose a supervised contrastive loss that can effectively decouple the semantic features and style features of images.
- We constructed a strongly semantically related dataset, and our model performed well on the test dataset.

II. Related Work

A. Text-to-Image Generation

Text-to-image generation typically involves taking a text description (prompts) as input and producing an image that corresponds to the text. Early pioneering efforts in text-to-image generation [16] utilized Generative Adversarial Networks (GANs) [14]. These approaches combined a prompt embedding with a latent vector, aiming for the GANs to generate an image that illustrates the prompt. Such early work [17]–[19] has inspired further research in text-to-image generation using GANs. However, GAN-based models do not always deliver optimal performance in image generation [1]. More recently, significant advancements in text-to-image generation have been achieved with the introduction of diffusion models [1]–[3],

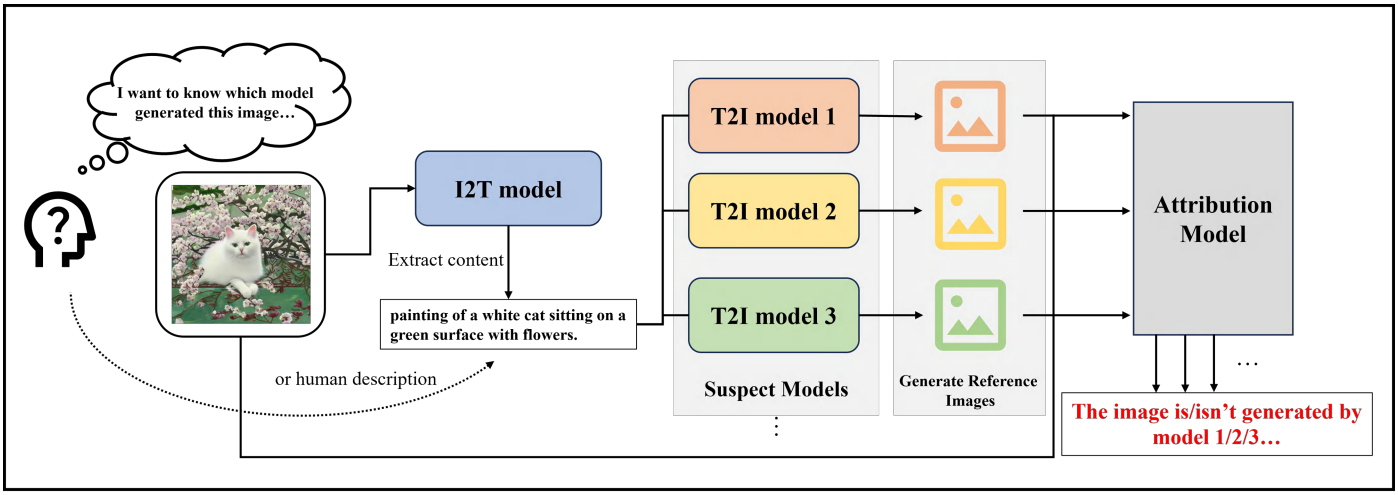


Fig. 1: Overview of Image Source Attribution Process: User first needs to describe the image to be attributed or automatically obtain descriptions using the I2T model. The obtained description is then input into the suspicion model to generate a reference image. These two images are then input together into the attribution model to output result.
I2T model: Image to Text model. **I2T model**: Text to Image model.

[20], [21]. These models typically start with random noise and prompts, then iteratively refine the noisy images into clear ones guided by the prompts. Modern diffusion model-based approaches, such as Stable Diffusion [1], DALL·E [2], Imagen [3], and GLIDE [21], have set new benchmarks in performance, surpassing earlier models. Consequently, our work focuses on these diffusion-based models.

B. Synthetic Image Attribution

Over the past year, research on the traceability of fake images has mainly focused on tracing images generated by GAN-based models. Representative works include: [11] propose a GAN fingerprinting technique based on representation mixing, which can matching images invariant to their semantic content. Their solution demonstrates robustness against benign transformations (e.g., changes in quality, resolution, shape) typically encountered during online image resharing. [12] present an iterative algorithm for discovering images generated from previously unseen GANs by exploiting the fact that all GANs leave distinct fingerprints on their generated images. [13] also uses the unique model fingerprint existing in the GAN, and will leave a stable fingerprint in the generated image, supporting the image attribution. Their experiments show that even minor differences in GAN training can result in different fingerprints, which enables fine-grained model authentication.

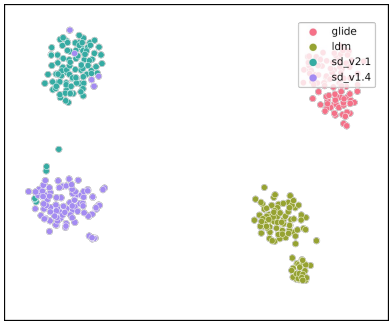
However, the results of these methods on the emerging T2I models are disappointing [15]. The literature concerning the tracing of Text-to-Image (T2I) generation models remains sparse, with only a handful of studies addressing this area. Notably, among the limited existing literature, the works of [15] and [22] stand out as significant contributions.

Sinitsa et al. [15] utilize the inductive bias of convolutional neural networks (CNNs) to develop a new detection method that requires a small amount of training samples and achieves accuracy that is on par or better than current state-of-theart methods.Based on the observation that images produced by each generative model exhibit distinct model fingerprints [13], [15] aims for the network to internalize this unique characteristic. Unlike analogous methodologies, the author utilizes a network to generate this Residual. Specifically, throughout model training, a similarity computation is conducted between the artifacts produced by the network and both synthetic and authentic images from the same source. The objective is to minimize the disparity between the artifacts and synthetic images while maximizing the gap between the artifacts and authentic images. Consequently, a novel loss function is introduced to facilitate this objective. During the inference phase, assessing the similarity solely between the acquired artifacts and the target image enables decision-making. Nonetheless, this approach encounters a notable challenge—its vulnerability to noise, as confirmed by the conducted experiments. Adding even a minimal amount of noise to the image is adequate to compromise the model's effectiveness.

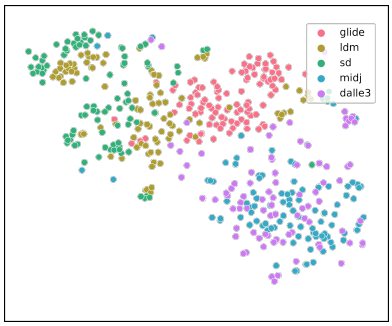
III. Method

A. Motivation

Our approach is based on the following hypothesis : A Text-to-Image (T2I) model is expected to exhibit a certain level of semantic and style similarity when generating images for the same prompt. Specifically, when employing the T2I model to generate images based on identical prompts, the resultant images demonstrate a notable correlation in content semantics and visual style, such as painting style, color brightness, etc. Furthermore, distinctions in content and visual semantics among various models arise, due to disparities in image databases and training methodologies employed during model



(a) Generated images from the same prompt but different models.



(b) Random prompts and different models.

Fig. 2: Utilize the image encoder of the CLIP model for extracting features from image data, then apply the t-SNE method to dimensionality reduction and visualization.

training. This unique and distinguishable similarity forms the basis for establishing a feature discrimination network, which aims to learn the semantic differences among images generated by different models under the same prompt.

To validate this hypothesis, we employed the Image Encoder of the CLIP [23] model for verification. Initially, we utilized a same prompt to generate 100 images for each of the four different T2I models (GLIDE [21], LDM [1], SDv1.4 [24], SDv2.1 [25]). Part of the sample image are shown in Fig.3. Subsequently, these images were input into the Image Encoder to obtain corresponding feature vectors. Finally, we employed the t-SNE [26] method for dimensionality reduction and visualization, as illustrated in Fig.2a. It is evident that, even with the use of non-finetuned pretrained weights, there is still a significant distinction among the models. Notably, the SDv1.4 and SDv2.1 models exhibit substantial separability, with SDv2.1 being fine-tuned on the basis of SDv1.4, demonstrating that their respective generated image features remain well-differentiated.

On the other hand, for each of the 5 different generative models (GLIDE, LDM, SDv2.1, Midjourney, DALLE·3), we randomly sampled 100 images from publicly generated datasets. Repeating the aforementioned procedure, the obtained results are depicted in Fig.2b. It is evident that their distributions are quite scattered and challenging to differentiate.

B. Process Design

Based on the above experimental observations, an apparent approach is to utilize identical prompts to generate multiple fake images, subsequently clustering them for tracing. However, this method is impractical in real-world scenarios due to the substantial time consumption involved in generating the images. The time cost required for image tracing through clustering is unacceptable. Therefore, we propose a feasible solution for real-world scenarios to tracing the origin of images generated by T2I models. When a user intends to trace the origin of a particular image, they can initially obtain the textual content of the image through self-description or an Image-to-Text (I2T) model (such as Blip [27], gpt2 [28] ...). Subsequently, this textual description is fed into the Text-to-Image(T2I) model under consideration for detection. The resulting newly generated image is then compared with the original image using a discriminant model, ultimately yielding a conclusion. If the image is generated by the T2I model under consideration, the newly generated image should exhibit strong semantic similarity to the original image. In such cases, the discrimination network would provide an affirmative decision. Conversely, if neither of them holds true, the discrimination network would render a negative decision for each pair. Thus, we convert the task of source attribution to a binary classification problem. The overall process of our approach is shown in Fig.1.

We believe that contrasting the input image with the newly generated image is more persuasive and practical than using a model to make direct judgments. After all, innocent until proven guilty.

C. Model Design

The above experiments have verified that the image encoder of CLIP possesses strong semantic feature extraction capabilities, and the provided features have separability at different generated model levels. Therefore, the detection model can be constructed by a feature extractor and a classification network. Meanwhile, considering our goal is to assess the similarity of two images in terms of content and visual semantics, we adopt the idea of Siamese networks [29]. The input images are separately processed by the encoder to obtain two features, and the distance between them is calculated to obtain the model loss. The specific model design is illustrated in Fig.4.

Here, we propose a Supervised Contrastive Loss [30], designed to decouple the semantic features and style features of images, defined as follows:

$$\mathcal{L}_{SupCon} = \frac{1}{|I|}\sum_{i\in I}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(d(\mathcal{F}_i,\mathcal{F}_p)/\tau)}{\sum_{a\in A(i)}\exp(d(\mathcal{F}_i,\mathcal{F}_a)/\tau)}, \tag{1}$$

where I denote the set of images generated using the same prompt, $P(i)$ the set of all images sharing the same source as i, $A(i)$ the set of all images from different model than i, $A, P \in I$. F_i denote the features of the image indexed by i, τ the scaling factor, and $d(,)$ represent the difference between two features.



Fig. 3: Image samples generated by different models from the same prompt: *'In a garden, a white cat rests beneath cherry blossoms, sunlight casting shadows amidst swirling petals'.*

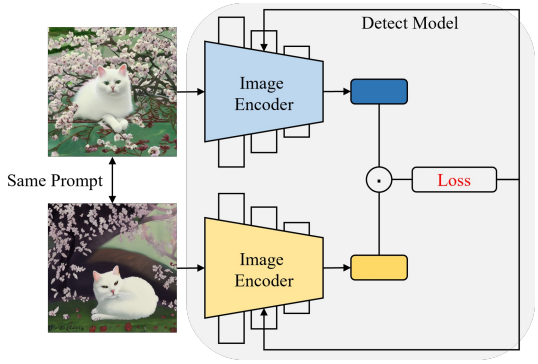


Fig. 4: Structure of the detect model.

IV. Experiment

A. Data Collection

We chose the Flickr [31] dataset as the real image training data. On the other hand, we utilize the generated images from LDM, SDv14 and SDv2.1 models for training. Specifically, we employ the LDM, SDv14 and SDv2.1 models to generate images corresponding to the text from the Flickr dataset. The final training dataset is shown in Table.I.

Here 6742×3 indicates that, for the same prompt, the generative model produced three synthetic images. The total number of images in our training data is 67,420, and testing data is 13,490.

B. Implementation Details

We decompose the attribution task into a binary classification problem, where given a pair of images as input, the model must determine whether the two images are generated by the same generative model. We use CLIP [23] and HR-Net [32] as backbones to extract image features. Each model was trained for 100 epochs using the Adam optimizer, with an initial learning rate set to 1e-4. The learning rate was scheduled to decay by a factor of 0.4 every 20 epochs to ensure stable convergence. The experiments were carried out on a single Nvidia A6000 GPU, leveraging the PyTorch framework for implementation.

TABLE I: Structure of dataset.

Type	Image Source	Image Size	Training Data	Testing Data
Real	Flickr	Not fixed	6742	1349
Fake	LDM	256 × 256	6742×3	1349×3
	SDv2.1	768 × 768	6742×3	1349×3
	SDv1.4	768 × 768	6742×3	1349×3
Total	-	-	67420	13490

C. Result

Table.II presents our experimental results. From the table, it is evident that CLIP and HR-Net without fine-tuning are ineffective; they tend to classify all image pairs as originating from the same source. In contrast, after fine-tuning CLIP and HR-Net using our method, the accuracy of provenance detection improves. This can be explained by the fact that images with the same prompt have strong semantic similarity, which CLIP and HR-Net can capture these similarities to make the same-source judgment. Our method further enables the model to focus more on the style features of the images, thus decoupling the original features. Notably, the provenance results with HR-Net as the backbone are significantly better than with CLIP. This is because the CLIP model aims to align image content with text, making it difficult to separate style and semantic features, whereas HR-Net focuses more on the image itself, thus better capturing the desired style features.

Fig.5 shows the probability distribution of predictions for the test data using HR-Net. It can be observed that, except for the image pairs consisting of real and sdv21 models, our method demonstrates high stability in other cases. For the real and sdv21 image pairs, we attribute this to the fact that the

TABLE II: Results of different backbones on the test sets.

Backbone	Model1	Model2	TP	FP	FN	TN	Recall	F1 Score	Accuracy	AUC
CLIP w/o fine-tune	LDM	SDv14	4842	7263	0	0	1	0.57	40.00%	0.82
	LDM	SDv21	4842	7263	0	0	1	0.57	40.00%	0.78
	SDv14	SDv21	4842	7263	0	0	1	0.57	40.00%	0.59
	Real	LDM	2421	2421	0	0	1	0.57	50.00%	0.89
	Real	SDv21	2421	2421	0	0	1	0.57	50.00%	0.87
	Real	SDv14	2421	2421	0	0	1	0.57	50.00%	0.83
CLIP	LDM	SDv14	3807	1414	1035	5849	0.78	0.75	72.92%	0.86
	LDM	SDv21	3833	1345	1009	5918	0.79	0.76	74.02%	0.86
	SDv14	SDv21	3832	5363	1010	1900	0.79	0.54	41.67%	0.54
	Real	LDM	1904	323	517	2098	0.79	0.82	85.50%	0.81
	Real	SDv21	1929	1467	492	954	0.80	0.66	56.80%	0.63
	Real	SDv14	1903	1324	518	1097	0.79	0.67	58.97%	0.66
HR-Net w/o fine-tune	LDM	SDv14	4842	7263	0	0	1	0.57	40.00%	0.79
	LDM	SDv21	4842	7263	0	0	1	0.57	40.00%	0.82
	SDv14	SDv21	4842	7263	0	0	1	0.57	40.00%	0.59
	Real	LDM	2421	2421	0	0	1	0.67	50.00%	0.88
	Real	SDv21	2421	2421	0	0	1	0.67	50.00%	0.88
	Real	SDv14	2421	2421	0	0	1	0.67	50.00%	0.78
HR-Net	LDM	SDv14	4576	37	266	7226	0.95	0.98	99.20%	0.99
	LDM	SDv21	4541	6	301	7257	0.94	0.97	99.87%	0.99
	SDv14	SDv21	4307	869	535	6394	0.89	0.86	83.21%	0.95
	Real	LDM	2405	15	16	2406	0.99	0.99	99.38%	0.99
	Real	SDv21	2136	557	285	1864	0.885	0.83	79.32%	0.90
	Real	SDv14	2171	244	250	2177	0.90	0.90	89.90%	0.96

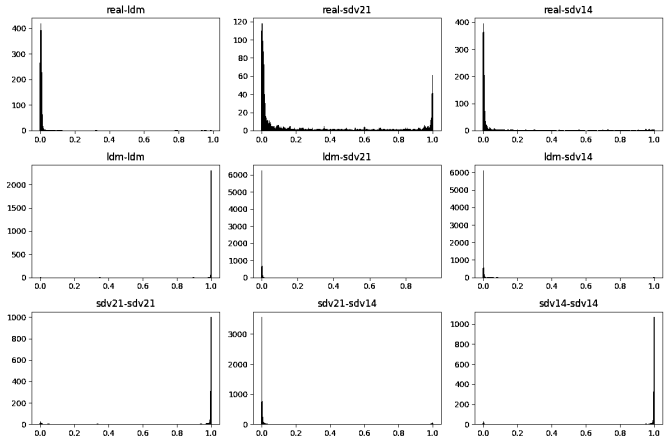


Fig. 5: Probability distribution of HR-Net backbone predictions on test data.

sdv21 model's training data is larger compared to the other generative models, resulting in generated images whose style more closely approximates real images, making them harder to distinguish from real images.

In Fig.6, we visualize the image features extracted by the two backbones using t-SNE. It is evident that HR-Net distinguishes the features of different models, with only few

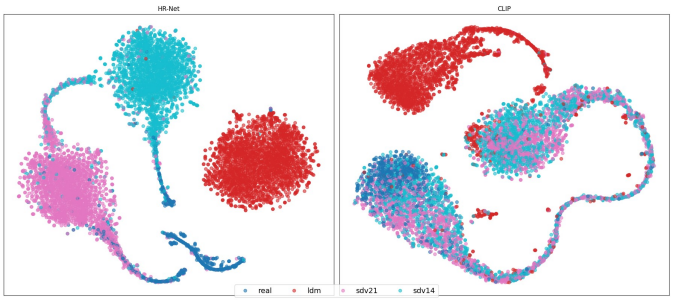


Fig. 6: T-SNE visualization of test image features extracted by CLIP and HR-Net.

features of real images overlapping. In contrast, CLIP performs rather poorly.

V. Conclusion

In this paper, We presents a novel approach for source attribution of generated images produced by diffusion-based text-to-image models. Compared to existing methods, our approach is more friendly towards black-box models. Users only need to input the same image descriptions into these models and then compare the generated images with the original ones. After experimental testing, our model exhibited strong performance in image source attribution task. This provides

a means of proof for copyright maintenance among image content creators.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62202009) and by the Macau Science and Technology Development Fund (Grant No. 0040/2023/ITP1, 0004/2023/RIB1).

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] J. Betker, G. Goh, L. Jing, *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, no. 3, p. 8, 2023.

[3] C. Saharia, W. Chan, S. Saxena, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[4] C. Campbell, K. Plangger, S. Sands, and J. Kietzmann, "Preparing for an era of deepfakes and ai-generated ads: A framework for understanding responses to manipulated advertising," *Journal of Advertising*, vol. 51, no. 1, pp. 22–38, 2022.

[5] S. Chesterman, "Good models borrow, great models steal: Intellectual property rights and generative ai," *Policy and Society*, puae006, 2024.

[6] H. Song, S. Huang, Y. Dong, and W.-W. Tu, "Robustness and generalizability of deepfake detection: A study with diffusion models," *arXiv preprint arXiv:2309.02218*, 2023.

[7] J. Lu, Y. Li, J. Zhou, B. Li, and S. Lyu, "Forensics forest: Multi-scale hierarchical cascade forest for detecting gan-generated faces," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2023, pp. 2309–2314.

[8] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8695–8704.

[9] J. Zhang, Y. Wang, H. R. Tohidypour, and P. Nasiopoulos, "Detecting stable diffusion generated images using frequency artifacts: A case study on disney-style art," in *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 1845–1849.

[10] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[11] T. Bui, N. Yu, and J. Collomosse, "Repmix: Representation mixing for robust attribution of synthesized images," in *European Conference on Computer Vision*, Springer, 2022, pp. 146–163.

[12] S. Girish, S. Suri, S. S. Rambhatla, and A. Shrivastava, "Towards discovery and attribution of open-world gan generated images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 094–14 103.

[13] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to gans: Learning and analyzing gan fingerprints," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7556–7566.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[15] S. Sinitsa and O. Fried, "Deep image fingerprint: Accurate and low budget synthetic image detector," *arXiv preprint arXiv:2303.10762*, 2023.

[16] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*, PMLR, 2016, pp. 1060–1069.

[17] N. Bodla, G. Hua, and R. Chellappa, "Semi-supervised fusedgan for conditional image generation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 669–683.

[18] Z. Wang, Z. Quan, Z.-J. Wang, X. Hu, and Y. Chen, "Text to image synthesis with bidirectional generative adversarial network," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2020, pp. 1–6.

[19] D. M. Souza, J. Wehrmann, and D. D. Ruiz, "Efficient neural architecture for text-to-image synthesis," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–8.

[20] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

[21] A. Nichol, P. Dhariwal, A. Ramesh, *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[22] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 3418–3432.

[23] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[24] *Stable diffusion v1-4 model.* https://huggingface.co/CompVis/stable-diffusion-v1-4.

[25] *Stable diffusion v2-1 model.* https://huggingface.co/stabilityai/stable-diffusion-2-1.

[26] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[27] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*, PMLR, 2022, pp. 12 888–12 900.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[29] G. Koch, R. Zemel, R. Salakhutdinov, *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, Lille, vol. 2, 2015.

[30] P. Khosla, P. Teterwak, C. Wang, *et al.*, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[31] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[32] J. Wang, K. Sun, T. Cheng, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

Automated Pseudo-Label Generation and Parallel Computing for Enhanced Few-Shot Medical Image Segmentation

Trong-Duc Nguyen* and Tien-Dung Do † and Thanh-Ha Do‡
* VNU University of Science
E-mail: nguyentrongduc_t66@hus.edu.vn
† Post and Telecommunication Institute of Technology (PTIT), Vietnam
E-mail: dungdt@ptit.edu.vn
‡ VNU University of Science
E-mail: dothanhha@hus.edu.vn

***Abstract*—Few-shot semantic segmentation is a technique with significant potential for medical image segmentation tasks. Most existing few-shot semantic segmentation methods require fully annotated labels for the training process. However, these methods may not be suitable for medical images, where data collection and labeling are challenging. To address this issue, this paper proposed an enhanced, few-shot semantic segmentation model with a new pre-processing step to generate pseudo-labels automatically. In this paper, parallel computing is also developed to accelerate image pre-processing. Experiments done on MRI image datasets present the effectiveness of the new approach since it outperforms conventional few-shot semantic segmentation methods.**

I. Introduction

Medical image segmentation classifies and locates areas in medical images, aiding in the diagnosis and treatment of diseases. Its applications are diverse, from detecting cancer and classifying organs over images to analyzing images of the brain and heart. Medical image segmentation presents a significant challenge for medical experts and image processing engineers because of the complexity, noise, and large size of medical images. Therefore, many types of research have been done to segment medical images. They are divided into two directions: traditional image segmentation methods and deep-learning-based segmentation methods.

Traditional image segmentation methods use images' characteristics, such as color and edge, to analyze the color values of pixels in the image over different color spaces, such as RGB, HUE, and HSV. Then, similar color values are grouped into regions with similar colors. Color-based image segmentation can be used for simple applications, but outliers, such as lighting, shadows, and transformations, can reduce accuracy.

Edge-based approaches detect edges and contours to segment the image into regions. These methods use edge detection algorithms such as Sobel, Canny, or Laplacian, followed by an algorithm such as RANSAC to detect points on the edge. After detecting points on the edge, clustering methods such as K-means and DBSCAN are applied to group the points on the edge into objects. In general, edge-based segmentation methods are often affected by noise and lighting variations, and segmenting objects of different shapes and sizes is challenging.

Threshold-based image segmentation methods set thresholds to divide pixels into regions of similar contrast. However, these methods also have limitations, such as setting thresholds, which can be challenging and require much experimentation to find the suitable threshold. Additionally, they could be more effective with low-contrast images or are affected by noise or uneven lighting.

Traditional image segmentation methods generally allow quick and easy implementation. Their fast processing speed is also advantageous in applications that require fast processing speed. However, they have some notable disadvantages, such as producing regions segmented incorrectly or confused with others in complex images or objects with complex shapes, being unstable to outliers such as noise or uneven lighting, and requiring adjusting many parameters to achieve the best segmentation results, which requires extensive experience and expertise.

Deep learning methods have recently developed efficient and accurate medical image segmentation solutions. Notable among these are the U-net [1], a convolutional neural network designed for biomedical image segmentation; the V-net, an extension of the U-net that incorporates volumetric data; and DeepLab, a state-of-the-art deep learning model for semantic image segmentation. In general, the performance of deep learning approaches is excellent. However, these networks all require labels for training. This is a big drawback because labeling medical images takes a lot of time and effort, and it requires well-qualified doctors to be able to label them correctly.

Recently, deep-learning-based approaches have combined the traditional model and deep learning architecture to form a self-learning method that takes advantage of the deep learning model and does not need to label the data (see Section II). Among approaches, the ALPNet represents a significant advancement in image segmentation by combining active learn-

ing with probabilistic deep learning models. It reduces the labeled data required for training models by iteratively selecting the most informative samples for labeling. By integrating with self-supervised learning (SSL) techniques, SSL-ALPNet proposed further efficiency and effectiveness than ordinary ALPNet since pre-training the network using self-supervised tasks on a large amount of unlabeled data can provide a strong initialization, reducing the number of learning cycles needed to achieve high performance.

Although SSL-ALPNet presents a promising approach, it still has room for improvement in performance, especially in computational complexity (iterative training, uncertainty estimation). Moreover, the risk of model overfitting to the noise in the pseudo-labels during the SSL phase can also be enhanced. Thus, this paper presents a new approach that improves the above limitations of SSL-ALPNet.

The paper's organization is as follows: Section II presents the state-of-the-art deep learning-based approach for medical segmentation. Section III presents the ALPNet and its enhanced version by investigating more on the clustering stage and time computing using parallel ideas. Section IV evaluates the proposed idea on the publication dataset, and the conclusion and future work are included in Section V.

II. Automated Pseudo-Label Generation for Medical Image Segmentation

There are medical image segmentation deep-learning-based methods for medical image segmentation, such as U-net, V-net, and DeepLab. While these networks require labels for training, which can be time-consuming for medical images, there is hope for the future. The self-learning method, pixel grouping, and deep-learning-based module are potential directions for improving the efficiency of image segmentation.

The authors in [2] presents an uncertainty-aware semi-supervised framework for 3D MR image segmentation with an uncertainty-aware scheme through the student and teacher models. The teacher model guide and target the student one to minimize segmentation loss and consistency loss. This uncertainty-aware scheme enables the student model to gradually learn from meaningful and reliable targets by exploiting the uncertainty information.

In [3], the authors proposed a multi-task deep network using a shareware semi-supervised segmentation strategy to leverage abundant unlabeled data and enforce a geometric shape constraint on the segmentation output. This network predicts semantic segmentation and signed distance maps of object surfaces. This paper also developed an adversarial loss to enhance the network's performance further. This loss function is designed to create a relationship between the predicted signed distance maps of labeled and unlabeled data. Doing so encourages the network to capture shape-aware features more effectively, thereby improving the quality of the segmentation output.

The authors in [4] present the MC-Net+ network, a potential approach in semi-supervised medical image segmentation. The MC-Net+ model introduces two designs. The first design features a shared encoder and multiple slightly different decoders. In contrast, the second design implements a mutual consistency constraint between one decoder's probability output and other decoders' soft pseudo labels. These two designs enable the MC-Net+ model to minimize the discrepancy of multiple outputs and generate consistent results in challenging regions, thereby promising to revolutionize model training.

In [5], Huimin Wu *et al.* propose a novel Compete-to-Win method (ComWin) to enhance the pseudo-label quality and reduce the critical limitation of the existing state-of-the-art semi-supervised approach cross with pseudo supervision. Their essential idea is to generate high-quality pseudo labels by comparing multiple confidence maps produced by different networks to select the most confident one. In addition, a boundary-aware enhancement module is integrated to refine pseudo labels further in near-boundary areas.

The SemiSL model [6] is a semi-supervised patch-based contrastive learning framework for medical image segmentation. In SemiSL, the pseudo-labels generated provide additional guidance, whereas discriminative class information learned in contrastive learning accurate multi-class segmentation. The novel loss in the SemiSL model also synergistically encourages inter-class separability and intra-class compactness among the learned representations.

Compared to the successful methods mentioned above, the model SSL-ALPNet [7] also significantly advanced medical image segmentation. In this paper, we research and then develop SSL-ALPNet [7] to improve performance, both in computational complexity and in reducing the over-fitting of the noise in the pseudo-labels during the SSL phase. On the other hand, our proposed method can be enhanced by incorporating unsupervised models in the SSL phase and adapted with other deep-learning models instead of ALPNet to boost semi-supervised medical image segmentation further.

III. Proposed Approach for Automated Pseudo-Label Generation

The proposed model is developed from the SSL-ALPNet model [7], which consists of two phases: offline pseudo-label generation and online training. The first phase directly affects the model's feature learning capability (see Figure 1). Therefore, we focus on this phase by proposing a novel pseudo-label generation algorithm instead of using Felsenszwalb's efficient graph-based [8] image segmentation algorithm as in the original SSL-ALPNet model. We also focus on optimizing prediction time to address more complex real-world challenges.

We selected the SLIC [9] algorithm to generate pseudo-segmentation labels to ensure segmentation accuracy and computational efficiency. Compared to the method used in [7], which employs a graph-based approach for segmentation, SLIC offers several advantages. Firstly, It is computationally more efficient due to its linear iterative clustering, with a computational complexity of $O(N)$, making it suitable for real-time applications. Secondly, SLIC produces more uniform

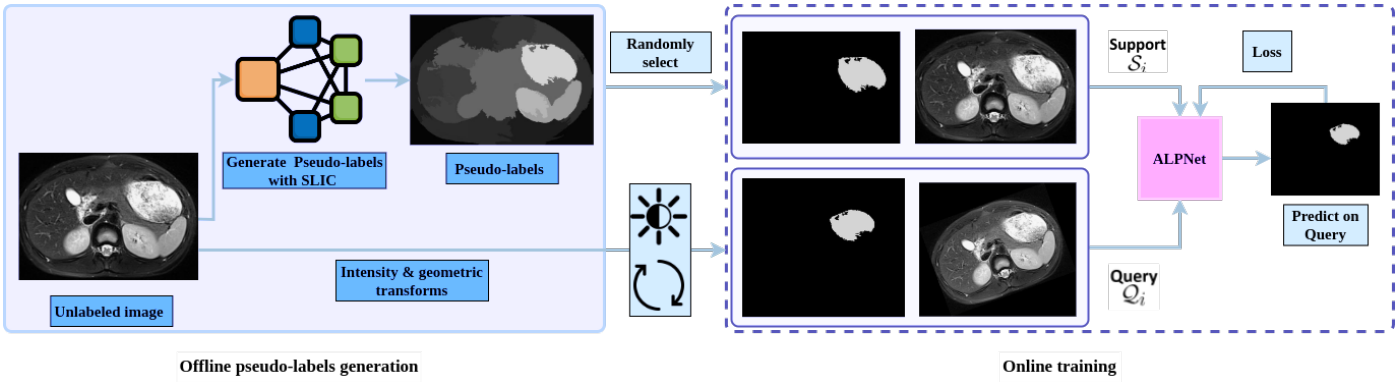


Fig. 1: Workflow of the superpixel-based self-supervised learning

superpixels that adhere closely to image boundaries, enhancing the quality of the segmentation. In contrast, the method used in [7], with a computational complexity of $O(N \log N)$, while effective in capturing irregular and varied segment shapes, can be more computationally intensive and may produce segments that do not align as precisely with image edges. These characteristics make SLIC a preferred choice for applications requiring both speed and accuracy in superpixel generation.

A. The SLIC Model

The Simple Linear Iterative Clustering (SLIC) algorithm is an efficient method for generating superpixels, which are compact and nearly uniform regions that adhere to image boundaries. Due to its simplicity and speed, It is particularly effective for image segmentation tasks, making it a popular choice in image processing applications.

SLIC operates by clustering pixels in the combined five-dimensional color and image plane space, each cluster representing a superpixel. The algorithm can be summarized in the following steps:

1) **Initialization:** Initialize cluster centers $C_k = [l_k, a_k, b_k, x_k, y_k]$ by sampling pixels at regular grid steps S. Perturb the centers in a 3×3 neighborhood to the lowest gradient position to avoid placing centers at edges.
2) **Assignment:** For each pixel p, assign it to the nearest cluster center whose search region overlaps the pixel. The distance measure D combines color similarity and spatial proximity:

$$D = \sqrt{\left(\frac{d_c}{m}\right)^2 + \left(\frac{d_s}{S}\right)^2} \tag{1}$$

where d_c is the color distance in CIELAB space, d_s is the spatial distance, and m is a compactness factor that balances the two distances.
3) **Update:** Update the cluster centers by computing the mean l, a, b, x, y values of all pixels assigned to each cluster.
4) **Iteration:** Repeat the assignment and update steps until convergence, typically defined by a minimal change in cluster centers.

B. Improving the Time Computing

Although SLIC has lower computational complexity than Felsenszwalb's graph-based image segmentation algorithm used in [7], we aimed to accelerate the pre-processing step further to reduce the time training model. By utilizing parallel computing on multiple CPU threads, we allow for concurrent image processing and significantly reduce the processing time by distributing the workload across several threads.

1) Parallel Computing Approach: Our approach to parallel computing is a testament to efficiency. By dividing the task into smaller sub-tasks and executing them simultaneously on multiple processors, we've significantly enhanced the efficiency of the pre-processing step. This allows us to process various images concurrently, assigning each to a different CPU thread for simultaneous processing.

- **Thread Allocation:** We allocate a specific number of CPU threads based on the available cores. Each thread processes a distinct image, ensuring balanced workload distribution and maximizing CPU utilization.
- **Synchronization:** While each thread processes its assigned image independently, synchronization mechanisms ensure that all threads start and complete their tasks coordinately. This avoids conflicts and ensures efficient resource management.
- **Load Balancing:** Dynamic load balancing techniques ensure that all threads complete their tasks approximately simultaneously. This avoids scenarios where some threads are idle while others are still processing.
- **Memory Management:** Efficient memory management is crucial in parallel processing to prevent bottlenecks. Shared resources are carefully managed to ensure that threads can access necessary data without significant contention.

2) Implementation Details: To implement parallel processing of multiple images, we used threading modules in Python, which provide high-level abstractions for parallel programming. The steps involved in our parallel implementation are as follows:

- **Image Allocation:** The set of images to be processed is divided among the available threads. Each thread is

assigned a subset of the images.

- **Parallel Processing:** Each thread independently processes its assigned images using the SLIC algorithm. This includes all steps of the SLIC process, from initialization to convergence.
- **Global Synchronization:** After processing all assigned images, threads synchronize to ensure that the entire batch of images has been processed before proceeding to the next batch.
- **Result Compilation:** The results from all threads are compiled into a unified output, maintaining the segmentation results for each image.

3) Performance Evaluation: We conducted extensive experiments to evaluate the performance of our parallel implementation. The results (see Section IV) show a significant reduction in processing time compared to the sequential version, especially when processing large batches of images. The algorithm's scalability was also tested, demonstrating that the parallel implementation scales well with the number of CPU cores. This parallelization approach accelerates the preprocessing step and enhances the overall efficiency of the image processing pipeline, making it highly suitable for real-time and high-resolution image processing tasks.

In conclusion, by utilizing parallel computing for the preprocessing step, we achieved significant improvements in processing time without altering the core SLIC [9] algorithm. This approach ensures that the algorithm can meet the demands of modern image processing applications, providing both speed and accuracy in superpixel generation.

IV. Experimental And Result

A. Dataset

We utilized the Abd-MRI dataset sourced from the ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge (Task 5) [10], comprising 20 3D T2-SPIR MRI scans. Each scan in the dataset was converted into 2D axial slices and resized to dimensions of 256 $\times$ 256 pixels. Standard pre-processing procedures were applied, including normalization and augmentation. Each 2D slice was replicated three times along the channel dimension to accommodate the network architecture.

B. Experimental Settings

The parameters are established in a similar experiment to [7]. We use the Dice score (0-100, with 0 indicating no overlap and 100 indicating perfect overlap), a standard medical image segmentation research metric, to assess the overlap between predicted and ground truth segmentation. For evaluating 2D segmentation on 3D volumetric images, we adhere to the protocol established by [11]. Within a 3D image, for each class $\hat{c}_j$, images between the top and bottom slices containing $\hat{c}_j$ are divided into C equally spaced chunks. The middle slice in each chunk from the support scan is the reference for segmenting all slices in the corresponding chunk in the query scan. In our experiments, C is set to 3. Notably, the support and query scans originate from different patients.

To evaluate generalization to unseen test classes, we ensure test classes are entirely unseen by removing any image containing a test class from the training dataset. We evaluate the model's ability to generalize to unseen semantic classes, testing its robustness in scenarios where new classes arise that were not present during training.

1) **Dataset Preparation:** Exclude any images containing the testing classes from the training dataset to ensure the model is trained without exposure to these specific classes.
2) **Training:** Train the model using only the remaining annotated data. The model learns to segment based on standard features and structures in the training data without explicit knowledge of the testing classes.
3) **Testing:** Evaluate the model on images containing the unseen testing classes. These classes are entirely novel to the model, having been excluded from the training phase. Measure how effectively the model can generalize its learned features and segmentation strategies to new, previously unseen classes.

The experiments in this section are conducted under a 1-way 1-shot setting to simulate the limited availability of labeled data in clinical practice.

The network is implemented in PyTorch based on the official PANet implementation [12]. To achieve high spatial resolutions in feature maps, $f_\theta(\cdot)$ is configured as an off-the-shelf fully convolutional ResNet101, pre-trained on a subset of MS-COCO to enhance segmentation performance [12], [13] (similar to the vanilla PANet in our experiments). It processes a $3 \times 256 \times 256$ image and generates a $256 \times 32 \times 32$ feature map. The local pooling window (L_H, L_W) for prototypes is set to 4×4 during training and 2×2 during inference on feature maps.

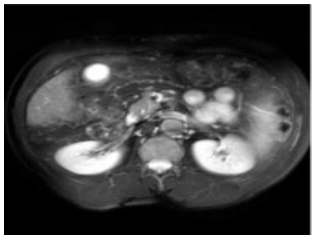
C. Comparative Experiments

Table I shows that the proposed SLIC [9] method is generally more stable and accurate than existing methods.

Tables I compare our method with vanilla PANet, one of the most advanced methods for natural images, as well as SE-Net [11], the latest approach for few-shot segmentation in medical images. Our proposed method, utilizing automatic label generation combined with ALPNet [7], achieves superior performance in these comparisons without manual annotation. Importantly, Table I highlights the strong generalization capability of our approach to unseen classes. Our proposed self-supervised learning method based on automatic label generation has successfully trained the network to learn diverse and generalizable image representations from unlabeled data.

The ALPNet model has demonstrated strong generalization capabilities for object features compared to existing methods. Figure 2 shows that the newly proposed pseudo-label generation method improves over the technique used in [11], [12]. SLIC generates segments closer to the ground truth labels compared to [8], [14].

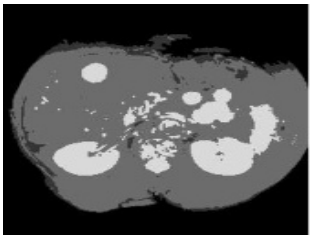
In addition, SLIC can speed up the pre-processing process compared to the method used in [7]; the results can be tracked



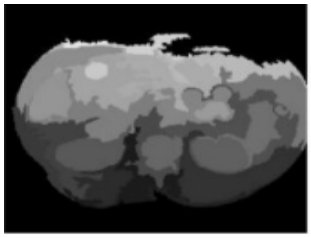
Original image



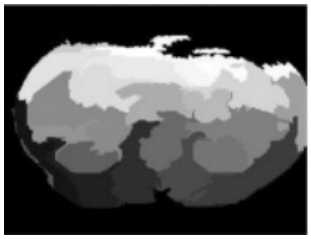
Truth ground



MeanShift



SLIC



Felzenszwalb

Fig. 2: Results of different pseudo-label generation methods

TABLE I: Experiment results (in Dice score) on abdominal images.

Method	Manual Annotation	Left kidney	Right kidney	Spleen	Liver	Mean
SE-Net [11]	✓	62.11	61.32	51.80	27.43	50.66
Vanilla PANet [12]	✓	53.45	38.64	50.90	42.26	46.33
MeanShift-ALPNet	×	50.03	46.25	56.44	42.78	48.88
SSL-ALPNet [7]	×	73.63	78.39	67.02	73.05	73.02
Proposed method	×	**82.53**	**82.04**	**80.61**	**80.35**	**81.38**

in Table II. The execution time is significantly reduced, mainly when parallel computing is applied.

TABLE II: Time computing in pre-processing process done on Intel Core i7-10750H CPU, 6 Core(s)

Method used in offline pseudo-label generation	Frames/s
Felzenszwalb [7]	13
SLIC [9]	28
SLIC-Parallel	91

To further elucidate the impact of the size and shape of pseudo-labels on the model, we experimented with another pseudo-label generation method, MeanShift [14]. MeanShift can generate pseudo-labels with more accurate contours for smaller objects than SLIC. However, it tends to produce more significant segments for larger organs than usual. As shown in Table I, the model's accuracy significantly decreased.

While manual annotations are precise, they are typically limited in quantity due to the labor-intensive nature of labeling medical images. On the other hand, SSL methods generate pseudo-labels from unlabeled data, allowing the model to learn from more diverse examples and generalize better to unseen cases. Additionally, pseudo-labeling may introduce beneficial regularization, forcing the model to label noise and variations across datasets robustly.

The results show that the size of the pseudo label significantly impacts the model's accuracy, not only if the labels are too coarse. Alternatively, pseudo labels that are too fine-grained might divert the granularity of clusters in the learned representation space from that of actual semantic classes.

V. Conclusion

The paper introduces a method for pseudo labels automatically combined with existing deep learning models, forming a self-supervised few-shot segmentation framework. The proposed method has outperformed the modern method; the state-of-the-art method requires no manual training. It also represents a significant improvement over existing supervised self-learning models. The ability to generalize to unseen classes outperforms its predecessor. Furthermore, we optimized the pre-processing process through parallel computing, enabling the model to solve more complex problems. Importantly, this optimization also allows the model to operate in real-time, demonstrating the practical applicability of our method and its potential for the future of supervised self-learning models.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

[2] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, *Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation*, 2019. arXiv: 1907.07034 [cs.CV].

[3] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3d semantic segmentation for medical images," in *Lecture Notes in Computer Science*. Springer International Publishing, 2020, pp. 552–561, ISBN: 9783030597108.

[4] Y. Wu, Z. Ge, D. Zhang, *et al.*, *Mutual consistency learning for semi-supervised medical image segmentation*, 2022. eprint: 2109.09960.

[5] H. Wu, X. Li, Y. Lin, and K.-T. Cheng, "Compete to win: Enhancing pseudo labels for barely-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 11, pp. 3244–3255, Nov. 2023, ISSN: 1558-254X.

[6] H. Basak and Z. Yin, "Pseudo-label guided contrastive learning for semi-supervised medical image segmentation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 786–19 797.

[7] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-supervision with superpixels: Training few-shot medical image segmentation without annotation," *arXiv preprint arXiv:2007.09886*, 2020.

[8] P. F. Felsenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," in *International Journal of Computer Vision*, Springer, vol. 59, 2004, pp. 167–181.

[9] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels," *Computer Vision–ECCV 2010*, pp. 849–862, 2010.

[10] A. Kavur, N. Gezer, M. Barı¸s, *et al.*, "Chaos challenge–combined (ct-mr) healthy abdominal organ segmentation," *arXiv preprint arXiv:2001.06535*, 2020.

[11] A. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "'squeeze excite' guided few-shot segmentation of volumetric images," *Medical Image Analysis*, vol. 59, p. 101 587, 2020.

[12] K. Wang, J. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," pp. 9197–9206, 2019.

[13] H. Shin, H. Roth, M. Gao, *et al.*, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[14] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, IEEE, 2002, pp. 603–619.

Enhanced Sparse Convolutional Detection Model for 3D Object Detection in Autonomous Vehicles Adapted to Traffic Conditions in Vietnam

Vu Hoang Dung*, Nguyen Trung Kien†, and Do Thanh Ha ‡
* Phenikka-X Company, Hanoi, Vietnam
E-mail: dungvh@phenikaa-x.com
† Hanoi University of Science and Technology
E-mail: kien.nt200303@sis.hust.edu.vn
† VNU University of Science
E-mail: dothanhha@hus.edu.vn

***Abstract*—The paper introduces a novel deep-learning model for enhancing 3D object detection in autonomous vehicles tailored to Vietnamese traffic conditions. The model improves efficiency by computing features only for relevant voxels, reducing computational costs and accelerating convergence through direct point cloud transformations. To address the limitations of the KITTI 3D dataset, a new proprietary dataset was created using Lidar16 and Lidar32 sensors, capturing specific Vietnamese traffic conditions. Experimental results on both the KITTI 3D and Phenikaa-X datasets show promising improvements in object detection performance, advancing techniques in 3D object detection and enhancing safety and reliability in diverse traffic environments.**

I. Introduction

In autonomous driving, 3D objects refer to entities detected and classified within three-dimensional space. These objects include vehicles, pedestrians, and static objects. Detecting and classifying these objects is crucial for the autonomous driving system to comprehend its surroundings and facilitate safe decision-making.

The computer vision approach for 3D object detection is mainly based on analyzing the image data resource or the characteristics of the image data. Various types of camera sensors usually capture image data. It comprises 2D images acquired from monocular cameras or 2D images with depth information obtained from stereo cameras. For image data in the form of 2D, the 3D object detection approaches analyze 2D images and then infer 3D information about objects [1][2]. Generally, the approach based solely on image data yields promising results but is constrained by the absence of depth information. Despite advancements in in-depth recovery from images, it remains an inherently unstable inverse problem. Variations in the 3D pose of the same object can result in distinct visual appearances on the image plane, complicating representation learning. Additionally, cameras are passive sensors, so images are susceptible to environmental factors such as varying lighting conditions (e.g., nighttime) or adverse weather like rain.

Point cloud data is gathered from 3D sensors like Lidar, capturing detailed 3D information about objects [3][4][5][6]. Point cloud data represents the surrounding 3D environment as a set of points, each with specific positional information in 3D space. This data aids in precisely determining the object's location in 3D space without additional transformations. However, detecting 3D objects from point clouds remains challenging due to several factors. Point clouds are often sparse, irregular, and unordered. Additionally, they may contain missing points or suffer from self-occlusion, further complicating accurate object detection data from images and point clouds are collected simultaneously from camera sensors and Lidar, providing comprehensive object characteristics (such as color) and precise object positions in 3D space based on point cloud information.

Processing combined data can help alleviate the limitations of each image data type [7]. However, establishing correspondence between the point cloud and the pixels presents a challenging problem that requires resolution. Thus, the prominent research direction receiving significant attention is enhancing point clouds-based approaches. One direction is representing the point cloud using voxels [8][9], which can improve resolution and accuracy but at the cost of increased computational complexity and memory usage. Another direction involves representing the point cloud as a point set, which can improve algorithm performance but add complexity to machine learning algorithms.

This paper researches and develops a 3D object detection method based on Lidar information. In more detail, we develop a sophisticated sparse convolutional detection framework to enhance the performance of 3D detection in autonomous. The proposed approach improves detection performance, reduces computational complexity and inference latency, and offers a more efficient and reliable solution for autonomous driving applications. In addition, in this paper, we propose to enhance the KITTI 3D published dataset by overcoming its restriction of a 180-degree viewing angle and labeling the front portion within the camera's field of view to a comprehensive 360◦ view of the surrounding objects.

The rest of this paper is organized as follows: Section II

provides a comprehensive review of related work and combined data from multiple sensors. Section III explores the Sparse Convolutional Detection (SECOND) model, incorporating enhancement to conventional convolutional network architectures. Section IV outlines the experimental setup and obtained results showcasing our method's performance improvements and advantages compared to baseline approaches. Finally, Section V concludes the paper with a summary of our contributions and critical outcomes.

II. Deep Learning Approaches in 3D Object Detection for Autonomous Driving

Over the past decade, we have witnessed significant advancements in 3D object recognition techniques, particularly those based on monochrome or dual-camera images [10], [11]. However, detecting 3D objects from monochrome images remains challenging, primarily due to the limited depth information available from a single image. Various methods, such as depth inference from images, leveraging geometric constraints, and shape assumption, have been proposed to tackle this issue, but they only offer partial solutions. One promising direction is a regression of 3D box parameters from images using a convolutional neural network (CNN). This approach, whether it focuses on predicting all 3D box parameters at once or initially predicts 2D boxes and then upgrades them to 3D boxes, holds the potential in 3D object detection.

3D object detection based on dual cameras is perspective direction since it uses two cameras positioned at a distance from each other to capture the same scene. While dual camera-based algorithms benefit from using geometric constraints to infer depth information, they also face limitations due to the need for precise calibration and synchronization.

Multi-view 3D detection is a direction that utilizes multiple cameras to capture the same scene from different angles. These images are then projected into a unified Bird's Eye View (BEV) space for processing by a BEV-based detector[12], [13]. The advantage of multi-view 3D detection is that it provides better detection than approaches relying on a single or dual camera, as multiple camera views offer more information about objects in the scene. However, multi-view 3D detection faces challenges in accurately transforming between camera views and BEV space without precise depth information, leading to misalignment between image pixels and their BEV positions. In addition, collecting and processing image data from multiple cameras can be costly and complex.

Unlike images, where pixels are evenly distributed on a plane, point clouds represent a sparse and irregular 3D data structure. Researchers have developed numerous 3D object detection models to explore pixel-level information based on point cloud data collected. The authors in Lidar [14],[15] proposed a method that uses a voxel grid to divide the point cloud into small cells, allowing for more efficient feature extraction.

Point-based 3D object detection relies on deep learning networks to detect objects within point clouds [16]–[18]. These methods, adapted from deep learning techniques used for images, involve two primary components: point cloud sampling and feature learning. Point cloud sampling, a critical step, involves selecting a subset of points from the original cloud using random, weighted, and structural sampling methods. This process proposes to reduce data size and enhance learning efficiency. Feature learning, the second component, entails extracting features from sampled points using point cloud operators. These features are then utilized for object classification within the point cloud.

To facilitate feature extraction, mesh-based 3D object detection involves converting point clouds into discrete grid representations, such as voxels, pillars, or top-view (BEV) maps [19]. Then, a deep learning model extracts features from these meshes. The methods in this direction involve two key components: grid-based representation and grid-based neural networks. Grid-based representation converts the point cloud into a discrete grid by dividing 3D space into small cells. Grid-based neural networks process these grid representations effectively in 3D space. Generally, mesh-based 3D object detection outperforms point-based approaches.

3D object detection using a combination of point and voxel data involves a hybrid architecture with two main frameworks: the one-stage and the two-stage. The one-stage framework directly processes the point cloud using points and voxels. Points are utilized for local feature extraction, while voxels contribute to global feature extraction. These extracted features are then employed for object detection within the point cloud. The two-stage framework operates in two distinct steps: the first one employs point- or grid-based methods for feature extraction, and the second one utilizes voxel-based methods specifically for object detection.

Another direction in 3D object detection involves integrating information from cameras and Lidar sensors, known as Early-Fusion-based 3D object detection [20]. Cameras can extract semantic features with color information, while Lidar excels in 3D positioning and structural information. The early fusion-based 3D object detection approach initiates with a 2D detection or segmentation network that extracts information from the image. This information is then transmitted to the Lidar point cloud. Depending on the approach to information fusion, early fusion-based 3D object detection methods are classified into region-level knowledge fusion and point-level knowledge fusion. Region-level knowledge fusion utilizes image-derived information to localize regions in the Lidar point cloud likely to contain objects. Point-level fusion methods enhance points in the Lidar point cloud using semantic features extracted from images. While Early-Fusion-based 3D object detection approaches show promise, they have the drawback of increased computational costs and inference time latency.

Late-Fusion-based 3D object detection approaches [16], [21], [22] leverage outputs from separate Lidar-based and camera-based object detectors. The Lidar detector provides 3D spatial information (position and size), while the camera detector contributes shape and color information. These methods often have good results by capitalizing on the strengths of both

approaches. However, the challenges are in fully integrating semantic information from both modalities, which can lead to inaccurate detection outcomes in specific scenarios.

III. Proposed Approaches for 3D Object Detection

This section explores the Sparse Convolutional Detection (SECOND) model, incorporating enhancement to conventional convolutional network architectures. One of the paper's contributions is adopting spatially sparse convolutional networks, which compute features only for voxels containing points, significantly reducing computational costs in the SECOND model. In addition, another contribution to adapting a dataset that can adjust to Vietnamese conditions traffic is presented in this section.

A. Enhanced Sparse Convolutional Detection Model

The architecture of the SECOND Network comprises the following components: Point Cloud, Voxel Feature Extractor, Sparse Convolution Layers, and Region Proposal Network (RPN). Point cloud grouping organizes point cloud data into coordinate tables, encoding voxel positions and actual point cloud data. The voxel feature extractor converts raw point clouds into voxel features, a crucial step in the process. Voxels, 3D cells that divide space into grids, are the backbone of the SECOND Network, which employs two Voxel Feature Extractor (VFE) layers to extract features from point clouds within voxels and encode them into voxel representations [14]. The sparse convolution layers, designed to operate on predominantly empty or sparse data spaces, such as those found in point cloud datasets, are the workhorses of the Network, extracting higher-level features from voxel representations, aiding in object identification within the point cloud. The region proposal network (RPN) generates bounding boxes around potential objects within the point cloud, a practical application of the Network's capabilities. These bounding boxes are subsequently utilized for object classification and localization, further demonstrating the real-world relevance of the SECOND Network.

In SECOND network architecture, computing sparse convolutions for feature extraction has significant challenges. In [23], the author uses regular sparse convolution and subarray sparse convolution (SSCN) for feature extraction. However, this approach has a limitation: it overlooks feature extraction from surrounding data and focuses only on activated inputs to produce activated outputs. Recognizing this, the paper [24] proposes a solution by modifying the model backbone. This modification involves integrating more targeted sparse convolution techniques, specifically focused sparse convolution (FSCN), to enhance the model's performance.

FSCN operates under the principle of fine-grained sparsity, where sparsity is distributed across the network. Fine-grained sparsity allows for more nuanced control over which weights contribute to the output, potentially leading to more efficient computations.

Conducting convolutions, mainly extracting voxel features, is the most time-intensive aspect of model execution. However, sparse convolutions significantly accelerate these calculations and processing, meeting practical demands. The paper investigates the potential of the SECOND model of replacing the backbone with FSCN to assess model accuracy and processing speed.

The total loss used during training enhanced SECOND model is a combination of the discussed loss functions:

$$L_{total} = \beta_1 L_{cls} + \beta_2 (L_{reg-\theta} + L_{reg-other}) + \beta_3 L_{dir} \quad (1)$$

in which L_{cls} is the classification loss, $L_{reg-other}$ is the regression loss for position and size, $L_{reg-\theta}$ is the angle loss, L_{dir} is the direction classification loss, β_1 = 1.0, β_2 = 2.0, β_3 = 0.2 are constant coefficients in this loss formula.

B. Constructing the dataset

The KITTI 3D dataset is a published dataset for computer vision in general and autonomous driving systems in particular. This dataset, derived from self-driving vehicles, includes images, point clouds, and associated information. However, when deploying the KITTI dataset in Vietnam presents several challenges. The first dataset reflects European (specifically German) road conditions characterized by well-demarcated roads conducive to object identification. The second one is the representation of vehicles. The KITTI dataset predominantly features cars, while Vietnamese roads are dominated by motorbikes and bicycles. This disparity can significantly impact the dataset's applicability in Vietnam. The third one is that the Lidar 64 sensor's point cloud resolution in the dataset yields high point density but requires a costly sensor. The last one is that the dataset labels only objects within the camera's visible range, resulting in partial vehicle annotations that enclose the front part of the vehicle with bounding boxes around the corresponding point clouds of identified objects.

Restricting the viewing angle is a notable limitation, as it only labels the front portion within the camera's field of view. This limitation impacts recognizing specific objects crucial for autonomous vehicles, necessitating a comprehensive 360° view of surrounding objects. To overcome this limitation, we proposed to enhance the KITTI 3D dataset's perspective as follows:

1. Utilize a deep learning model with pre-trained weights to re-identify all surrounding objects, extending beyond the camera's limited field of view. The focus is on achieving high accuracy rather than processing speed, aiming to reduce the need for post-calibration adjustments.
2. Employ the open-source SUSTECH tool [24], facilitating the visualization and information editing of 3D objects within a user interface. This tool aids in refining object-related data to enhance accuracy in the object recognition process.
3. Iteratively apply Steps 1 and 2 to expand the dataset's coverage and refine object labels, progressively enhancing the dataset's quality for subsequent training purposes.

To ensure a proposed model can work well in actual road conditions in Vietnam, we also train the model on a real

dataset with 3500 kilometers of data from various rod types in Vietnam. The data is collected using Lidar16 and Lidar32 sensors since these sensors are more cost-effective and suitable for Vietnam's conditions. In addition, this dataset is standardized with labels following a uniform structure compatible with various published data sources. Finally, the data collection is aligned with the characteristics of Vietnam's roads, focusing on critical vehicles such as cars, motorbikes, bicycles, and pedestrians. Figure I presents the data using Lidar 16 and Lidar 32 in the street in Vietnam.

IV. Experiment Results

Besides using the published dataset Kitti3D, we also use the dataset described in Table II.

In Table II, Kit3D-1 and Kit3D-2 having 500 and 7,481 samples Kitti3D dataset, respectively. These samples were collected using a Lidar Velodyne 64 sensor and labeled for objects within a 180-degree frontal view of the vehicle. In addition, Kit3D-2 has added 500 individually-developed samples generated as described in Section III.B. Kit3D-3 is a dataset enhanced from Kit3D-2 with more than 1000 individually-developed samples and manually labeled within a 360-degree frontal view of the vehicle. It contains 1000 customized manually labeled samples, also collected using a Lidar Velodyne 64 sensor.

The CUS-1 and CUS-2 datasets also take 2000 and 1000 enhanced samples from Kitti3D with labels that cover a 360-degree view of the vehicle. Notably, the CUS-1 dataset has 1000 individually developed samples using Lidar 16 sensors. In addition, the CUS-2 dataset has 500 individually developed samples using Lidar 16 sensors and 1000 individually developed samples using Lidar 64 sensors. All samples in CUS-1 and CUS-2 are customized to cover a 360-degree view of the vehicle.

The performance of the deep-learning model is evaluated using AP and $AP|RN$ (with N = 40) measures. They are widely adopted in object recognition studies over 3D datasets. These metrics comprehensively evaluate object recognition model performance, encompassing accuracy and completeness. AP is typically employed to assess overall model performance, while $AP|RN$ (see Equation 2) helps evaluate model performance under conditions where high precision is crucial.

$$AP|_{R_N} = \frac{1}{N} \sum_{r \in R} P_{interpolare}(r), \tag{2}$$

Given $R = [r_0, r_0 + \frac{r_1 - r_0}{N-1}, r_0 + \frac{2(r_1 - r_0)}{N-1}, ..., r_1]$, the precision corresponding to each recall level r is interpolated. This interpolation is achieved by identifying the maximum precision value for which the recall value is greater than or equal to r, as expressed by Equation 3.

$$P_{interpolare}(r) = \max_{\tilde{r}:\tilde{r} \geq r} P(\tilde{r}) \tag{3}$$

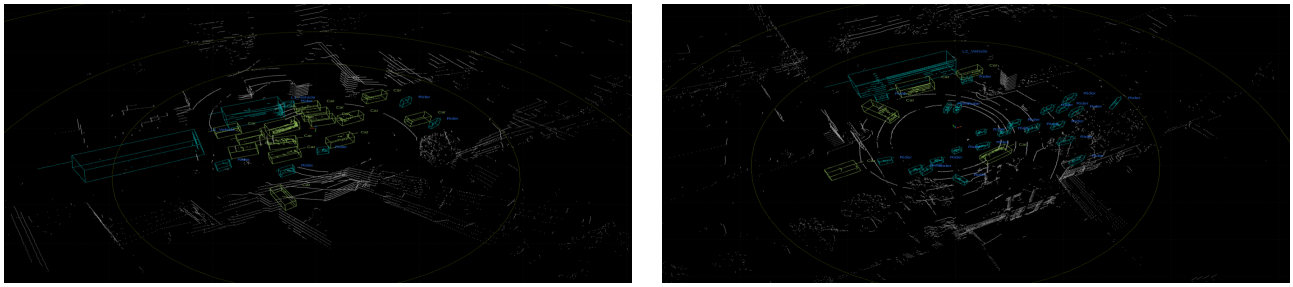
In the experiment, the rotated optimal threshold IoU_{3D} is 0.7 for cars and 0.5 for pedestrians and cyclists found after conducting many experiments. The IoU_{BEV} is computed by projecting the 3D bounding box onto the ground plane for BEV detection.

The first experiment evaluates the performance of using backbone FSCN instead of voxel backbone in 3D detection. The deep learning model with different backbones is trained on the KITTI dataset and then tested on Kit3D-1. Table III presents the obtained data and indicates that adopting the FSCN convolution backbone enhances model accuracy. However, utilizing the focal backbone prolongs model processing time since it takes 19.57s on average to detect objects, while the voxel backbone uses 17.49s.

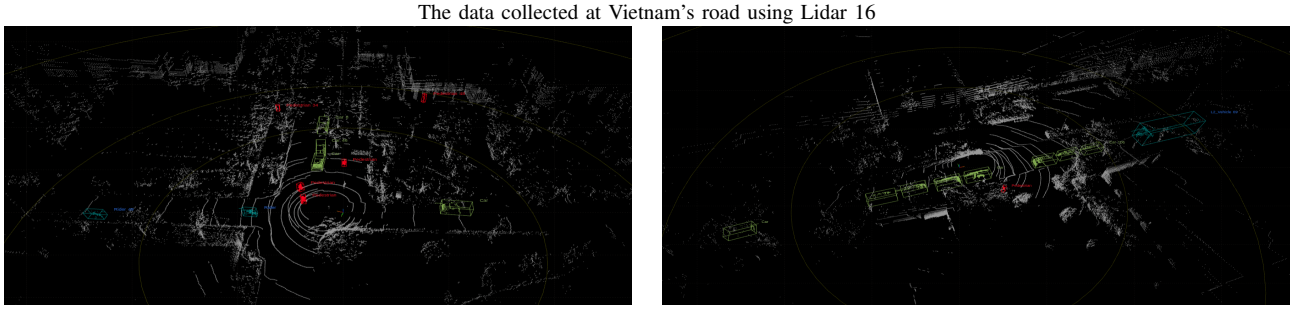
From the first empirical findings, we observe that modifying the model yields minimal changes. Therefore, we focus on leveraging diverse datasets in the following experiment to enhance the model's adaptability to Vietnamese traffic conditions. In more detail, we focus on data processing techniques as follows. First, training models on KIT3D-1, KIT3D-2, and KIT3D-3 datasets will be used, and then the performance on real datasets in Vietnamese traffic conditions will be verified. This real dataset belongs to the Phenikaa-X corporation. Due to confidentiality requirements, this dataset is not publicly accessible and is solely permitted to support the implementation of this paper. The dataset comprises samples collected using Lidar 16 and 32 sensors to capture 3000 kilometers of data from various road types in northern Vietnam, including highways, interprovincial roads, urban areas, and residential streets and 500 kilometers within Phenikaa University's campus and surrounding areas. This dataset is directly developing for Phenikaa's autonomous vehicles.

Table IV clearly illustrates the challenges we face. When the models are trained on the public dataset, they struggle to recognize 3D objects in Vietnamese traffic conditions, achieving only 0.94% accuracy with the Kit3D-1 dataset and 17.77 % with the Kit3D-2 training dataset and 18.43 % with the training dataset Kit3D-3. The low accuracy is a direct result of the limitations in the Kit3D-1 dataset, which only includes labels for the 180-degree frontal view of the vehicle. This results in the model's inability to identify surrounding objects (covering a 360-degree view of the vehicle). Moreover, as these datasets are public and not specifically tailored to Vietnamese environmental conditions, identifying accurate vehicle objects becomes challenging, thus leading to lower accuracy rates.

Table IV also indicates that the quality of data collection sensors significantly impacts dataset quality in terms of point cloud density and object distribution upon detection. It's crucial to emphasize that employing a combination of public and individually developed datasets is a step in the right direction to enhance model adaptability. However, these promising results are unsatisfactory due to these datasets' sparse point cloud density. Sparse point cloud density refers to a situation where the Lidar sensors have not captured enough data points in a given area, leading to gaps in the data. The Kit3D-3 dataset, despite being an extension of the public KITTI dataset that was individually developed for improved accuracy, still falls



The data collected at Vietnam's road using Lidar 16



The data collected at the campus of Phenikka University in Vietnam using Lidar 32

TABLE I
EXAMPLE COLLECTED DATA USING LIDAR 16 AND LIDAR 32, RESPECTIVELY IN VIETNAM

	KITTI dataset			**mixed dataset**	
	Kit3D-1	**Kit3D-2**	**Kit3D-3**	**CUS-1**	**CUS-2**
Lidar	64	64	64	64 + 16	64 + 32 + 16
Angle	180^0	180^0 + 360^0	360^0	360^0	360^0
Number of samples	500	7481 + 500	7481 + 1500	2000 + 1000	1000 + 1000 + 500

TABLE II
ALL DATASET IN THE PAPER

Measure	**Voxel**	**Voxel Focal**
AP	81.765 %	82.395 %
APR40	84.1 %	84.29 %

TABLE III
PERFORMANCE OF THE SECOND MODEL WITH DIFFERENT BACKBONE.

Measure	**KIT3D-1**	**KIT3D-2**	**KIT3D-3**	**CUS-1**	**CUS-2**
AP	0.9 %	9.09 %	9.09 %	28.64 %	40.09 %
APR40	0.94 %	17.11 %	18.43 %	48.72 %	47.77 %

TABLE IV
PERFORMANCE OF THE MODEL TRAINED ON DIFFERENT DATASETS AND VERIFYING ON PHENIKAA'S DATASET.

short of practical deployment requirements due to the sparsely distributed point cloud density resulting from using Lidar16 sensors for data collection. This underscores the pressing need to enhance data quality, even when collected under Vietnam's traffic conditions, and emphasizes the necessity for sensor improvement.

The model trained on the CUS-1 and CUS-2 datasets achieved accuracies of 48.72% and 47.77%, respectively. These training models combine public data with individually developed samples, addressing limitations present in both datasets, such as unsuitable environmental conditions for Vietnamese traffic and insufficient point cloud density for effective object recognition. However, it's crucial to stress the need for continuous improvement. Our work is not done yet. Training additional datasets under Vietnamese environmental conditions using advanced Lidar sensors is recommended to improve model accuracy further and ensure the safer deployment of autonomous vehicles in traffic scenarios. This emphasis on continuous improvement underscores the importance of our collective efforts in this field.

V. CONCLUSION

This paper presents a deep learning-based model designed to enhance 3D object detection capabilities in autonomous vehicles operating within the unique traffic conditions prevalent in Vietnam. By focusing on the computation of features for only the most relevant voxels, the model significantly reduces computational overhead and accelerates the convergence of the detection process through direct point cloud transformations. Addressing the shortcomings of the widely used KITTI 3D dataset, we evaluated our approach over an innovative proprietary dataset (Phenikaa-X datasets) utilizing Lidar16 and Lidar32 sensors to accurately capture the nuances of Vietnamese traffic scenarios. The experimental validation of

the model across both the KITTI 3D and Phenikaa-X datasets demonstrates substantial advancements in object detection performance. These findings underscore the potential of the proposed model to significantly improve safety and reliability in various traffic environments, particularly in regions characterized by complex driving conditions, such as Vietnam.

ACKNOWLEDGMENT

This research was funded by the research project QG.23.71 of Vietnam National University, Hanoi. The dataset used in this paper is supported by Phenikka-X Company in Hanoi.

REFERENCES

[1] S. Song and J. Xiao, *Deep sliding shapes for amodal 3d object detection in rgb-d images*, 2016. eprint: 1511.02300.

[2] X. Shen and I. Stamos, *Frustum voxnet for 3d object detection from rgb-d or depth images*, 2023.

[3] H. Ran, J. Liu, and C. Wang, "Surface representation for point clouds," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 920–18 930.

[4] J. Chen, B. Lei, Q. Song, H. Ying, D. Z. Chen, and J. Wu, "A hierarchical graph network for 3d object detection on point clouds," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 389–398.

[5] B. Cheng, L. Sheng, S. Shi, M. Yang, and D. Xu, "Back-tracing representative points for voting-based 3d object detection in point clouds," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8959–8968.

[6] P.-S. Wang, "Octformer: Octree-based transformers for 3d point clouds," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–11, Jul. 2023, ISSN: 1557-7368.

[7] Y. Zhang, J. Yu, X. Huang, W. Zhou, and J. Hou, *Pcr-cg: Point cloud registration via deep explicit color and geometry*, 2023.

[8] X. Du, M. H. A. J. au2, S. Karaman, and D. Rus, *A general pipeline for 3d detection of vehicles*, 2018. eprint: 1803.00387.

[9] V. A. Sindagi, Y. Zhou, and O. Tuzel, *Mvx-net: Multimodal voxelnet for 3d object detection*, 2019. eprint: 1904.01649.

[10] F. Ma and S. Karaman, *Sparse-to-dense: Depth prediction from sparse depth samples and a single image*, 2018. eprint: 1709.07492.

[11] X. Cheng, P. Wang, C. Guan, and R. Yang, *Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion*, 2019. eprint: 1911.05377.

[12] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6526–6534.

[13] Z. Wang, W. Zhan, and M. Tomizuka, "Fusing bird's eye view lidar point cloud and front view camera image for 3d object detection," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1–6.

[14] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, *Voxel r-cnn: Towards high performance voxel-based 3d object detection*, 2021. eprint: 2012.15712.

[15] Y. Zhang, W. Feng, Y. Quan, G. Ye, and G. Dauphin, "Dynamic spatial-spectral feature optimization-based point cloud classification," *Remote. Sens.*, vol. 16, p. 575, 2024.

[16] Z. Yin, H. Sun, N. Liu, H. Zhou, and J. Shen, *Fgfusion: Fine-grained lidar-camera fusion for 3d object detection*, 2023. eprint: 2309.11804.

[17] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85.

[18] L. Ma, Y. Li, J. Li, W. Tan, Y. Yu, and M. A. Chapman, "Multi-scale point-wise convolutional neural networks for 3d object segmentation from lidar point clouds in large-scale environments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 821–836, 2021.

[19] Y. Cui, R. Chen, W. Chu, *et al.*, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, Feb. 2022, ISSN: 1558-0016.

[20] X. Cheng, Y. Zhong, Y. Dai, P. Ji, and H. Li, "Noise-aware unsupervised deep lidar-stereo fusion," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6332–6341.

[21] T. He, P. Sun, Z. Leng, C. Liu, D. Anguelov, and M. Tan, *Lef: Late-to-early temporal fusion for lidar 3d object detection*, 2023. eprint: 2309.16870.

[22] Y. Li, A. W. Yu, T. Meng, *et al.*, *Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection*, 2022. eprint: 2203.08195.

[23] Y. Yan, Y. Mao, and B. Li, "SECOND: sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[24] E. Li, S. Wang, C. Li, D. Li, X. Wu, and Q. Hao, "Sustech points: A portable 3d point cloud interactive annotation platform system," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 1108–1115.

Enhancing Cell Segmentation using Deep Learning Models by Custom Processing Techniques

Van-De NGUYEN*, Minh-Huong Hoang DANG†, Quang-Huy NGUYEN†, Manh-Cuong DINH †, Thanh-Ha DO †
* Department of Pathological Anatomy, Central Military Hospital 108
E-mail: doctorde@gmail.com
† VNU University of Science
E-mail: danghoangminhhuong_t65, nguyenquanghuy1_t66, dinhmanhcuong_t66, dothanhha (@hus.edu.vn)

***Abstract*—This study presents a comprehensive workflow for cell segmentation using the Monuseg dataset. Our method integrates several advanced techniques to enhance the performance of segmentation models. We apply Principal Component Analysis (PCA) and Gaussian Mixture Model - Expectation Maximization (GMM-EM) clustering for data pre-processing. The training and evaluation significantly improved the mean IoU metric, demonstrating the enhanced performance of Language meets Vision Transformer (LViT) compared to other deep-learning networks. This constructed workflow shows substantial potential in improving the accuracy of cell segmentation tasks, which is crucial and beneficial for various biomedical applications.**

I. Introduction

Cytopathology, the study of cellular-level diseases, is a pivotal tool in cancer screening. Artificial intelligence can potentially revolutionize diagnostics, especially segmenting cell images from micrographs.

Accurate cell segmentation is critical, the cornerstone of medical image analysis. It enables the precise identification and classification of cellular structures within microscopic images. It is significant for various biomedical applications, such as computational pathology, disease diagnosis, prognosis, and computer-aided diagnosis.

For the classical approach, the Split and Merge Watershed (SM-Watershed) method [1] for cell segmentation in fluorescence microscopy images. Initially, the Marker-Controlled Watershed (MC-Watershed) algorithm provides preliminary segmentation. The Split phase separates clusters using cell characteristics like size and convexity, while the Merge phase refines these segments by eliminating over-segmentation. This method effectively balances segmentation accuracy without needing labeled data.

The segmentation method that combines the K-L transform and the OTSU method is not just a standard cell segmentation method. It is a proven and effective technique. The K-L transform is utilized to select the most informative channel from the image, followed by the OTSU method to determine an optimal automatic threshold value for segmentation. This approach has been applied to various types of cell images, demonstrating its practicality and effectiveness.

While effective in specific scenarios, classical approaches such as Watershed, OTSU, and K-L transform often need help with the complexity and variability of all cell segmentation tasks. These methods can result in over-segmentation, under-segmentation, or mis-identification of cells. Consequently, recent research has increasingly focused on deep learning techniques, which offer improved accuracy and robustness in handling diverse and complex cell images (see detail in Section II).

Deep Learning-based innovative algorithms generally archive state-of-the-art performance in medical imaging segmentation. Among the models, we are inspired by the LViT (Language meets Vision Transformer) model [2] since it tackles the challenge of limited labeled data in medical image segmentation by incorporating medical text annotations. This text information supplements the image data, allowing the model to learn even with limited labeled examples.

We improve the performance of the LViT model by ensuring high-quality data input. In more detail, we integrate Principal Component Analysis (PCA) for dimensionality reduction and Gaussian Mixture Models with Expectation-Maximization (GMM-EM) in the pre-processing pipeline to optimize the clustering process. This has resulted in improved preliminary segmentation maps, a promising outcome that paves the way for further advancements in cell segmentation.

The following parts are organized as follows: Section II shows some base deep learning models applied in cell segmentation. Section III focuses on our contribution to the combination of GMM-EM and LVit model. Moreover, Section IV indicates the experiment results, and Section V determines this method's conclusion and discussion.

II. Deep Learning Approaches for Cell Segmentation in Microscopic Images

This section reviews some deep-learning approaches that have had good results in the last few years, particularly on the MoNuSeg dataset, a widely used benchmark dataset in cell segmentation. Understanding the performance of these approaches on this dataset will help us decide the direction of continuing research in this field.

In 2020, Jha *et al.* [3] proposed a Double U-Net model by adding another U-Net at the bottom of the network to capture supplementary semantic information efficiently. Further, Atrous Spatial Pyramid Pooling (ASPP) was adapted to capture contextual data, and the post-processing techniques

significantly improved the result of automatic polyp detection. However, since this network uses two Unet models, the increase in the number of parameters is a limitation of this model.

Wang *et al.* [4] with a Bending Loss Regularized (BLR) model is successful in tackling the challenge of segmenting overlapped nuclei in histopathology images. This model applied high and low penalties to contour points with large and small curvature. In addition, the bending loss helps to avoid the generation of boundaries for two or more nuclei that are touching. The BLR model performs better than other models. However, there is still a problem with the segmentation of overlapping nuclei.

Hassan *et al.* [5] proposed a Pyramid Scene Parsing with SegNet (PSPSegNet) to identify and delineate the boundaries of nuclei. The experiment indicates that the PSPSegNet model is effective with F1-Score and AJI at 0.8815 and 0.7080, respectively. Concerning the object level, the PSPSegNet model relies on training data and, therefore, is unsuitable for exacting cell shapes.

The author Lagree *et al.* [6] proposed a gradient-boosting U-Net(GB U-Net) to segment breast tumor cell nuclei. This research shows that deep convolutional neural networks are suitable for training with transfer learning on a set of histopathological images independent of breast tissue to segment or not tumor nuclei of the breast.

In the same year, 2021, Li *et al.* [7] proposed the Bagging Ensemble Deep segmentation (BEDs) model, which aggregates self-ensemble learning and testing stage augmentation to improve the robustness of nucleus segmentation. However, this model needs to segment better when the images are complicated in shape and structure.

One year later, Qin *et al.* [8] proposed the REU-Net model to improve segmentation accuracy by focusing on region-specific features within images. It leverages enhanced feature extraction techniques to identify better and segment nuclei in medical images. However, while REU-Net boosts performance, it may introduce additional computational complexity due to its region-focused approach. Liang *et al.* [9] also use region information to build a model that integrates Guided Anchoring (GA) into the Region Proposal Network (RPN) and using a fusion box score (FBS) with soft non-maximum suppression (SoftNMS). This model improves accuracy over traditional CNN-based approaches and enhances cell-level analysis in digital tissue images.

In 2023, the large-scale model named Segment Anything Model (SAM) was introduced [10]. This model is trained on 11 million images with over 1 billion masks for general-purpose segmentation. Although SAM doesn't initially provide high-quality segmentation for medical images, its masks, features, and stability scores are valuable for improving medical image segmentation models. This model can be applied to augment inputs for models like U-Net, with experiments on three tasks that show its effectiveness.

In recent years, U-Net and its variants have been widely used in pathology image segmentation, leveraging skip connections to recover detailed information. However, the semantic gap between encoder and decoder can hinder performance. To address this, the FusionU-Net [11] incorporates a fusion module to reduce semantic gaps by exchanging information between skip connections. Our two-round fusion design considers local relevance and bi-directional information exchange across layers. In addition, another model, named BiU-Net [12], combines CNNs and transformers using a two-stage fusion strategy. The Single-Scale Fusion (SSF) stage integrates local and long-range features, while the Multi-Scale Fusion (MSF) stage eliminates the semantic gap between deep and shallow layers. Additionally, a Context-Aware Block (CAB) in the bottleneck enhances multi-scale features in the decoder, improving segmentation performance.

III. The Proposed Approach: Combination LViT Model with GMM-EM Methods

Compared to the successful methods mentioned in Section II, the LViT (Language meets Vision Transformer) model [2] overcomes these limitations by significantly enhancing context awareness and reducing dependency on extensive labeled data, making it more suitable for cell segmentation tasks. LViT effectively captures global relationships within images, leading to more accurate and efficient segmentation results than Unet++. Thus, we enhanced this model in this paper by applying the pre-processing stage to improve performance (see Figure 2).

Before describing our contribution in more detail, we revise the main idea of the LViT model, as illustrated (Figure 1): LViT uses text information to supplement the image data, allowing the model to learn even with limited labeled examples. Furthermore, LViT leverages semi-supervised learning, utilizing text data to generate high-quality pseudo-labels for unlabeled images. An Exponential Pseudo Label Iteration (EPI) mechanism assists the Pixel-Level Attention Module (PLAM) preserve local image features during this process. The Exponential Pseudo Label Iteration (EPI) mechanism is a critical component in LViT, designed to enhance the quality of pseudo-labels for unlabeled image iterative. EPI refines the pseudo-labels in each iteration by incorporating feedback from the model's predictions, which become increasingly accurate over time. This iterative process improves pseudo-labels' reliability, facilitating more effective semi-supervised learning. Complementing EPI, the Pixel-Level Attention Module (PLAM) ensures that fine-grained details are preserved during segmentation. PLAM assigns attention scores to each pixel, allowing the model to focus on critical regions within the image and maintain local feature integrity. This dual mechanism of EPI and PLAM enables LViT to achieve high precision in cell segmentation tasks, even with limited labeled data.

Compared to models like U-Net++ and BiO-LinkNet, which rely solely on convolutional operations, LViT's transformer architecture allows it to capture local and global information

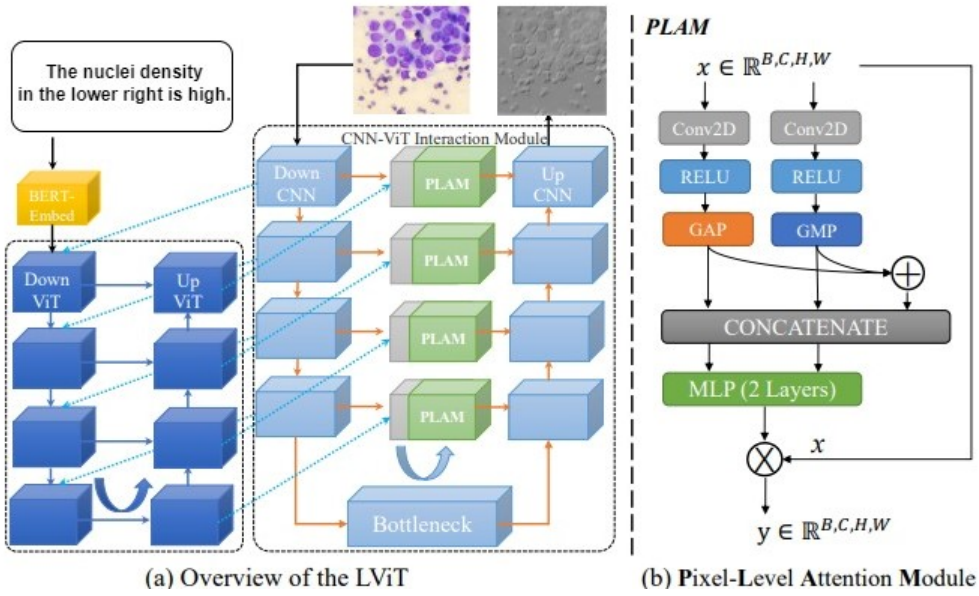


Fig. 1. Illustration of (a) the proposed LViT model and (b) the Pixel-Level Attention Module (PLAM). The proposed LViT model is a Double-U structure that combines a U-shape CNN branch with a U-shaped ViT branch [2].

more effectively. While MaxViT-UNet achieves high performance using a combination of convolutions and transformers, LViT's advantage lies in its ability to leverage image data and text annotations. This adaptability makes it more effective in cell segmentation tasks.

Our pre-processing stage prepares the data and ensures a robust and reliable process for LViT optimally. We use Principal Component Analysis (PCA) to reduce the data's dimensionality, followed by Gaussian Mixture Models (GMMs) with Expectation Maximization (EM) to group images into distinct clusters.

The Principle Component Analysis was first applied in our pre-processing pipeline to reduce the feature vectors' dimensions, including pixel intensity values and texture features extracted from the cell images. The PCA transformation, a crucial step in our pipeline, is defined in Equation 1.

$$\mathbf{z} = \mathbf{W}^T(\mathbf{x} - \boldsymbol{\mu}) \tag{1}$$

Where $\mathbf{x}$ is the original feature vector, $\boldsymbol{\mu}$ is the mean of the feature vectors, $\mathbf{W}$ is the matrix of eigenvectors of the covariance matrix, and $\mathbf{z}$ is the transformed feature vector in the principal component space.

Following dimensionality reduction, GMM-EM was applied for image clustering. GMM-EM was chosen due to its ability to effectively process diverse and complex cellular data, its flexibility in modeling complex distributions, and its capacity to provide soft probabilities for each cluster. The GMM model is defined as in Equation 2.

$$p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2}$$

Where $\mathbf{z}$ is the PCA-transformed feature vector, K is the number of clusters, π_k is the weight of the k-th cluster, and $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the probability density function of the Gaussian distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

An iterative experimental process meticulously determined the number of clusters K. This process was designed to optimize the separation between clusters and the compactness within each cluster, ensuring the highest quality results. Upon completion of the clustering process, each pixel in the image was labeled according to the cluster with the highest posterior probability, generating a preliminary segmentation map.

Our research has achieved several notable outcomes:

- Developed a pre-processing pipeline that effectively integrates PCA for dimensionality reduction with GMM-EM for clustering, providing high-quality input for the subsequent LViT model.
- Our research has led to a significant achievement-the combination of PCA and GMM-EM has optimized the clustering process. This has resulted in improved preliminary segmentation maps, a promising outcome that paves the way for further advancements in cell segmentation.

IV. Experiments Results

The experiment is evaluated using the highly regarded MICCAI MoNuSeg dataset. This dataset comprises 44 images, each sized at 1000×1000 pixels with $28,846$ labeled cell nuclei distributed across nine organs: breast, liver, kidney, prostate, bladder, colon, stomach, lungs, and brain. The dataset is organized into 24 images for training, 6 for validation, and 14 reserved for testing.

Our approach involved extensive experimentation with various sizes of image patches in images of the MoNuSeg2018

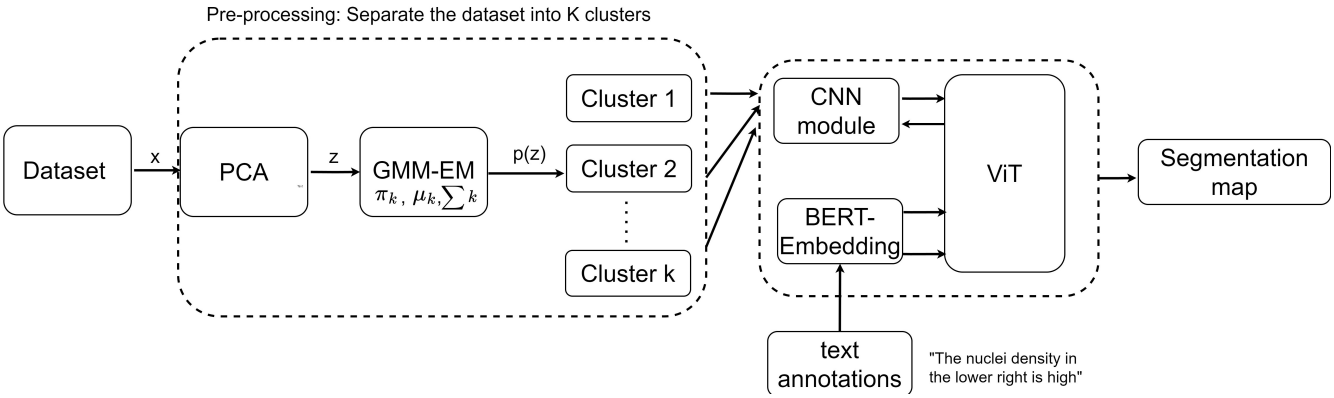


Fig. 2. Proposed Model for Cell Segmentation over Microscopic Images combining GMM-EM to LViT Model

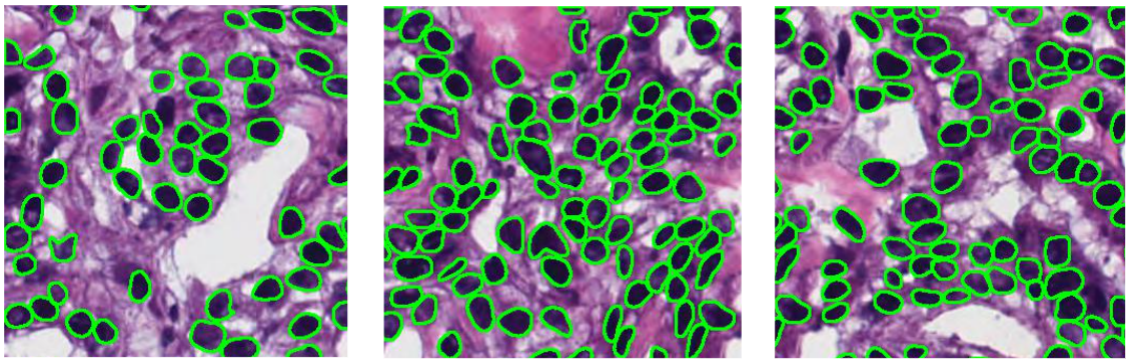


Fig. 3. Cell segmentation results on MoNuSeg2018 dataset [13] using the proposed model, achieving a Dice score of 0.80, IoU of 0.66, and a runtime of 17155.2 seconds on $K = 2$ clusters, trained on an NVIDIA A100 (Google Colab Pro).

Method	LVit [2]	Proposed model	MaxViT-UNet [14]	Dice Unet [15]	Unet++ [16]	BiO-LinkNet [17]	LinkNet [17]	R2U-Net [18]
Dice	0.78	**0.80**	0.83	0.76	0.77	0.77	0.77	0.80
IoU	0.65	**0.66**	0.72	0.62	0.63	0.62	0.63	0.68

TABLE I
SEMANTIC SEGMENTATION RESULTS ON THE MONUSEG2018 DATASET.

dataset. We were cropping images to 256×256 pixels produced trade-off results. We also addressed memory constraints by adopting overlapping patches with a 70-pixel overlap while maintaining the original organ distribution.

Given the relatively small dataset size of $1,100$ patches, this paper employs data augmentation techniques to enhance model training and accuracy, effectively expanding the dataset to $2,200$ images. These techniques, including random horizontal flipping, rotating, and adding a Gaussian filter with a random parameter, were developed to address the challenges of working with a small dataset, demonstrating our commitment to overcoming research obstacles.

The semantic segmentation model's accuracy is evaluated using the Intersection over Union (IoU) measure and the Dice coefficient. For the individual segmentation problem, the model's accuracy is assessed based on the Score value. The IoU, Dice, and Score measurements have values in the range [0, 1]; when the value is close to 0, the model's accuracy is low; the closer it is to 1, the higher its accuracy.

Using the LViT method, each image in the Monuseg2018 dataset is accompanied by a text passage providing a specific description and evaluation of that particular image. In more detail, each image describes the characteristics of the nuclei, such as evenly/sparse distribution and higher/lower density areas, etc. There may also be several images with the same text passage. We pre-processed the images using the GMM-EM (Gaussian Mixture Model with Expectation-Maximization) technique to effectively process the visual data before applying the LViT method. This process clusters only the image data, creating groups of visually similar images while maintaining the individual text descriptions for each image. We conducted experiments with different K values to determine the optimal number of clusters (K) for the image data. We adjusted based on experimental results and dataset

characteristics to optimize model performance. This approach allows us to leverage both the clustered visual information and the unique textual descriptions for each image in our dataset, enhancing the overall performance of the LViT method.

K	Dice Score	IoU Score	Execution Time (seconds)
1	0.7659	0.6307	13449.25
2	**0.8015**	**0.6566**	17155.2
3	0.7039	0.5522	23643.5

TABLE II
PERFORMANCE METRICS WITH DIFFERENT NUMBERS OF CLUSTERS ON THE TEST DATASET.

Following the data pre-processing with GMM-EM, we experimented with different numbers of clusters K, starting from 1 and incrementally increasing. After training the LViT model from scratch on each clustered dataset, we selected $K = 2$ as it yielded the highest accuracy compared to other configurations. Choosing $K = 2$ allows the model to capture enough variance to segment the cells accurately while maintaining robustness against noise or unnecessary complexity. This balance likely leads to better generalization, as evidenced by the superior performance metrics. Table II shows the performance metrics for each K. The model's performance was then evaluated on the test set using metrics such as IoU and the Dice coefficient. The results (see Table I) are a testament to our progress, with the LViT method combined with GMM-EM achieving an IoU of 0.66 and a Dice coefficient of 0.8, surpassing the baseline LViT model (IoU = 0.65, Dice = 0.78). Compared with standard LViT and U-Net++, our proposed method demonstrates that our approach is competitive and a significant step forward in the microscopy image processing field.

Table I illustrates our results compared to another method. The results of the proposed method show significant improvement in cell segmentation performance on the MoNuSeg2018 dataset compared to other methods. Our approach achieved the second highest Dice score of 0.80 and an IoU of 0.66, demonstrating better accuracy and reliability. In contrast, the standard LViT method scored a Dice of 0.78 and IoU of 0.65, indicating a noticeable enhancement with improvements. The proposed method outperforms both metrics compared to Dice Unet and Unet++, which scored Dice values of 0.76 and 0.77 and IoU values of 0.62 and 0.63, respectively.

Although MaxViT-UNet achieves a higher Dice score of 0.83, its superior performance can be attributed to its multi-axis attention mechanism, which better captures local and global contexts. This, together with combining convolutional layers and transformers, enhances its segmentation accuracy. However, its increased complexity and computational demands highlight the trade-offs, whereas the proposed method offers a strong balance between performance and efficiency.

Other methods, like BiO-LinkNet and LinkNet, with Dice scores of 0.77 and IoU scores of 0.62 and 0.63, respectively, also fall short of our results. The R2U-Net method also performs well, with a Dice score of 0.80 and an IoU of 0.68. However, our method remains robust and consistent across both evaluation metrics, reinforcing its effectiveness for accurate and efficient cell segmentation.

Figure 3 presents three cropped images from the original MoNuSeg2018 dataset. The figure indicates that using the GMM-EM and LViT effectively segments overlapping cells and distinguishes between cells and other components, such as blood, which are not the focus of this study.

V. CONCLUSIONS AND DISCUSSIONS

This paper presented an enhanced cell segmentation approach by integrating the LViT model with GMM-EM preprocessing. The use of clustering is crucial in improving the learning process, as it allows the model to capture more distinct and characteristic features from the data. By applying PCA for dimensionality reduction, followed by GMM-EM clustering, the data is effectively segmented into meaningful groups, making it easier for the model to focus on relevant patterns. This structured input significantly boosts the accuracy and reliability of the model. This combination achieved superior performance when evaluating the MICCAI MoNuSeg dataset. In the future, we plan to collect data from diverse sources and further optimize the model parameters to enhance the accuracy of our approach significantly.

ACKNOWLEDGMENT

This research was funded by the research project QG.23.71 of Vietnam National University, Hanoi.

REFERENCES

[1] M. Gamarra, E. Zurek, H. J. Escalante, L. Hurtado, and H. San-Juan-Vergara, "Split and merge watershed: A two-step method for cell segmentation in fluorescence microscopy images," *Biomedical signal processing and control*, vol. 53, p. 101 575, 2019.

[2] Z. Li, Y. Li, Q. Li, *et al.*, "Lvit: Language meets vision transformer in medical image segmentation," *IEEE transactions on medical imaging*, 2023.

[3] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, *Doubleu-net: A deep convolutional neural network for medical image segmentation*, 2020. arXiv: 2006.04868 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2006.04868.

[4] H. Wang, M. Xian, and A. Vakanski, *Bending loss regularized network for nuclei segmentation in histopathology images*, 2020. arXiv: 2002.01020 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2002.01020.

[5] L. Hassan, A. Saleh, M. Abdel-Nasser, O. A. Omer, and D. Puig, "Promising deep semantic nuclei segmentation models for multi-institutional histopathology images of different organs," 2021.

[6] A. Lagree, M. Mohebpour, N. Meti, *et al.*, "A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, p. 8025, 2021.

[7] X. Li, H. Yang, J. He, *et al.*, *Beds: Bagging ensemble deep segmentation for nucleus segmentation with testing stage stain augmentation*, 2021. arXiv: 2102 . 08990 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2102.08990.

[8] J. Qin, Y. He, Y. Zhou, J. Zhao, and B. Ding, "Reu-net: Region-enhanced nuclei segmentation network," *Computers in Biology and Medicine*, vol. 146, p. 105 546, 2022.

[9] H. Liang, Z. Cheng, H. Zhong, A. Qu, and L. Chen, "A region-based convolutional network for nuclei detection and segmentation in microscopy images," *Biomedical Signal Processing and Control*, vol. 71, p. 103 276, 2022.

[10] Y. Zhang, T. Zhou, S. Wang, P. Liang, Y. Zhang, and D. Z. Chen, "Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, pp. 129–139.

[11] Z. Li, H. Lyu, and J. Wang, "Fusionu-net: U-net with enhanced skip connection for pathology image segmentation," in *Asian Conference on Machine Learning*, PMLR, 2024, pp. 694–706.

[12] Z. Huang, Y. Zhao, Z. Yu, *et al.*, "Biu-net: A dual-branch structure based on two-stage fusion strategy for biomedical image segmentation," *Computer Methods and Programs in Biomedicine*, vol. 252, p. 108 235, 2024.

[13] A. Goodman, A. Carpenter, E. Park, *et al.*, *2018 data science bowl*, Kaggle Competition, 2018. [Online]. Available: https : / / kaggle . com / competitions / data - science-bowl-2018.

[14] A. R. Khan and A. Khan, "Maxvit-unet: Multi-axis attention for medical image segmentation," *arXiv preprint arXiv:2305.08396*, 2023.

[15] K. C. T. Nguyen, "Segment automatically cells over microscopic images based on deep learning techniques," *Undergraduate thesis in VNU University of Science*, 2021.

[16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, Springer, 2018, pp. 3–11.

[17] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE visual communications and image processing (VCIP)*, IEEE, 2017, pp. 1–4.

[18] M. Z. Alom, C. Yakopcic, T. M. Taha, and V. K. Asari, "Nuclei segmentation with recurrent residual convolutional neural networks based u-net (r2u-net)," in *NAECON 2018-IEEE National Aerospace and Electronics Conference*, IEEE, 2018, pp. 228–233.

Marker-Aware Ovarian Tumor Segmentation from Ultrasound Images

Hoang-Son Bui*, Sy-Hoang Tran*, Thuy-Binh Nguyen†, Thanh-Hai Tran*, Hai Vu*, Thi-Lan Le*
* School of Electrical and Electronics Engineering
Hanoi University of Science and Technology, Hanoi, Vietnam
† University of Transport and Communications
Corresponding author e-mail: lan.lethi1@hust.edu.vn

***Abstract*—Ovarian cancer remains one of the leading causes of cancer-related deaths among women, with early detection being pivotal for successful treatment. Accurate segmentation of ovarian tumor regions in ultrasound images is essential to assist clinicians in the effective diagnosis and treatment of ovarian cancer. However, many previous studies have struggled to achieve high segmentation accuracy due to the complex nature of ultrasound images. These images often contain inherent noise, and symbols marked by sonographers, which pose significant challenges for segmentation tasks. In this study, before feeding the images into the segmentation model, we employ an in-painting method to remove the symbols and markers from the ultrasound image. Then a segmentation model based on UNet3+ with ResNet50 as its encoder is introduced. We evaluated the performance of our model on the OTU2D dataset and demonstrated an improvement over the existing models with the Dice, IoU, Recall and Precision scores of 86.44%, 77.05%, 86.31%, and 89.18%, respectively. Moreover, when testing on a subset of clean images, the pre-processing technique based on in-painting helps to increase the Dice and Precision metrics from 90.63% and 86.23% to 96.64% and 96.08%. This study presents a novel approach that enhances image segmentation capabilities by introducing an in-painting pre-processing method and a UNet3+ model with ResNet50 as its encoder. Our method significantly improves segmentation accuracy and offers clinicians more reliable data for better-informed clinical decisions.**

***Index Terms*—Ovarian Tumor Segmentation; In-painting; Marker Removal; OTU2D Dataset; ResNet50-UNet3+;**

I. Introduction

Ovarian cancer progresses silently and is often diagnosed at late stages, making it a leading cause of cancer death in women. Previous studies show that if cancer is diagnosed early, the five-year survival rate exceeds 90% [1]. Early detection and accurate diagnosis of the tumors are crucial for timely treatment, which can significantly improve patient outcomes and maintain women's health and fertility. Identifying the tumor area can assist doctors in predicting and deciding appropriate treatment methods, thereby improving the overall quality of healthcare services.

Determining the tumor region of interest from ultrasound images is often formulated as a segmentation task. Segmentation tasks in medical imaging often utilize CNNs with encoder-decoder architectures to extract relevant features from images [2], [3]. Typically, the input of these models is an ovarian ultrasound image, and the output is a binary image in which tumor areas are separated from non-tumor areas. However, the complex nature of ultrasound images, with their inherent noise and the symbols marked by sonographers, poses significant challenges for ovarian tumor segmentation.

In our study, to overcome the challenges posed by noise and symbols in ultrasound images, we utilized the LaMa in-painting technique [4] as a pre-processing step. This approach effectively eliminates artifacts such as sonographer markers, measurement rulers, and patient details, enhancing the image quality and facilitating more accurate segmentation when applied to deep learning models. Then a segmentation model that couples the Resnet50 encoder with the architecture of UNet3+ is introduced. The results of our model on the OTU2D dataset [5] demonstrate an improvement over the existing models with the Dice, IoU, Recall, and Precision scores of 86.44%, 77.05%, 86.31%, and 89.18%. Moreover, when testing on a subset of 50 clean images, the pre-processing technique based on in-painting helps to increase the Recall and Precision metrics from 89.32% and 86.23% to 97.78% and 96.08%.

In summary, the contribution of the paper is two-fold. First, we propose to apply an in-painting technique in the pre-processing step to remove the unwanted markers and symbols in ultrasound images. Second, a ovarian tumor method that incorporates the ResNet50 encoder and the architecture of Unet3+ with full-scale skip connections is proposed.

The structure of this paper is as follows: Section II provides a brief review of existing in-painting techniques and ovarian tumor segmentation methods. Section III describes the main components of our proposed framework. The evaluation results are discussed in Section IV. Finally, Section V offers discussions and conclusions.

II. Related Work

Segmentation models have seen significant advancements recently, with a notable increase in research outcomes. Ronneberger et al. introduced U-Net [6], which is a widely adopted architecture for biomedical image segmentation. Its architecture includes a contracting path for capturing contextual information and an expanding path for precise localization. Skip connections between corresponding layers to preserve fine details. U-Net excels at handling small datasets and accurately segmenting medical images by capturing intricate details and spatial context simultaneously.

Building upon [6], U-Net++, an extension of UNet, was developed by Zhou et al. in [7]. UNet++ aims to improve segmentation accuracy. It integrates nested and densely connected pathways to capture hierarchical features across multiple scales. U-Net++ demonstrates enhanced performance by leveraging nested skip pathways to refine segmentation boundaries and details. Additionally, U-Net3+ [8], incorporates attention mechanisms and dense connectivity blocks to enhance feature reuse and integration across scales. This architecture achieves superior segmentation results by focusing on effective feature aggregation and refinement at different hierarchy levels.

To enhance segmentation capabilities in ultrasound images, in [9], a method named CR-Unet was proposed. CR-Unet integrates a spatial recurrent neural network (RNN) with a U-Net architecture, forming a composite network. CR-Unet addresses several typical challenges of ultrasound images, such as poor image quality, low contrast, and complex anatomical shapes.

Recent studies have tried to enhance UNet by [6] not only optimizing skip connections but also improving the encoder branches. For instance, the study in [2] replaced the encoder branch by MobileNetV2 to enhance feature extraction from ultrasound images. This approach led to a significant increase in the Dice coefficient to 79.00% on the OTU2D dataset [5]. It shows promising results in semantic segmentation tasks, demonstrating the effectiveness of integrating advanced encoder architectures within the UNet framework. Additionally, it highlights the potential of using more powerful architectures to further enhance model performance.

Several studies focus on addressing noise in ultrasound images, such as patient details and markers. In [1], Coburn et al. employed a CNN-CAE approach to diagnose ovarian tumors. They initially faced challenges with misaligned activation areas in their CNN model due to marks present in the images. By incorporating a denoising convolutional autoencoder (CAE) to remove these marks before training the CNN, they achieved improved classification results. This adjustment enabled more accurate alignment of activation areas with relevant regions in the images, thereby facilitating more reliable tumor diagnosis.

In [10], Chen et al. focused on enhancing classification and segmentation accuracy by addressing symbols and marks in medical images using image in-painting techniques. They introduced a framework called mask-guided generative adversarial network (MGGAN) designed to remove these symbols. MGGAN utilizes attention mechanisms to enhance the realism of lesion boundaries, thereby improving segmentation accuracy. Their approach significantly boosted segmentation performance, increasing accuracy from 71.51% to 76.06% for the Unet model and from 61.13% to 66.65% for the PSPnet model when applied to clean images. These studies underscore the importance of robust pre-processing techniques in medical image analysis, demonstrating how advancements in image cleaning and enhancement can lead to more accurate and reliable AI-based diagnostic tools.

III. Proposed framework for Ovarian Tumor Segmentation

A. Overall framework

The main target of this study is to propose an effective framework for the ovarian tumor segmentation task, illustrated in Fig. 1. This framework consists of two crucial components that are pre-processing and image segmentation steps. In fact, ultrasound images usually contain inherent noise and symbols marked by sonographers, leading to significant challenges for segmentation tasks. Consequently, the purpose of the preprocessing step is to eliminate noise and provide higher-quality images for the subsequent segmentation step. For the image segmentation task, we leverage the abilities of both the ResNet50 [11] and U-Net3+ [6] models by replacing the encoder of U-Net3+ with ResNet50 for the feature extraction step. The ResNet50 model enhances the segmentation capability of U-Net3+, significantly improving both Precision and Recall in the ovarian tumor segmentation problem. The details of these two crucial steps are provided in the following sections.

B. Symbol removing by in-painting technique

With the purpose of eliminating noise and providing higher images for the segmentation step, in this study, we propose to use the Large-Mask in-painting technique (LaMa) [4] in the pre-processing step. The LaMa technique is performed to remove and fill up large missing areas within a given image; from this, the input images become clearer and sharper. However, this process is time-consuming, particularly during the labeling phase, where annotating noise markers in ovarian ultrasound images. The Fast Fourier Convolution (FFC) block is a key component in this model, designed to split channels into local and global information branches in images. While real FFT (Fast Fourier Transform) is utilized in the global branch, enabling the model to take the entire image into account when generating convolutions, it is applied to real-valued signals, ensuring that the output remains real-valued. A local branch utilizes conventional convolutions to process information within specific regions of the image. By combining the outputs from their local and global branches, FFCs effectively leverage both local details and global context for accurate image in-painting. This approach provides an image-wide receptive field, enabling the generator to integrate global context from the initial layers, which is crucial for ensuring the filled areas remain consistent with the overall structure and appearance of the original image. The architecture of the LaMa model is described in Fig. 2.

C. Ovarian Tumor Segmentation

The main contribution of the proposed framework is demonstrated in the image segmentation component, where the ResNet50 [11] model is incorporated with U-Net3+ to enhance its segmentation capability. Fig.3 illustrates the modification in the architecture of U-Net3+, with ResNet50 serving as the encoder of the U-Net3+ model. By utilizing ResNet50

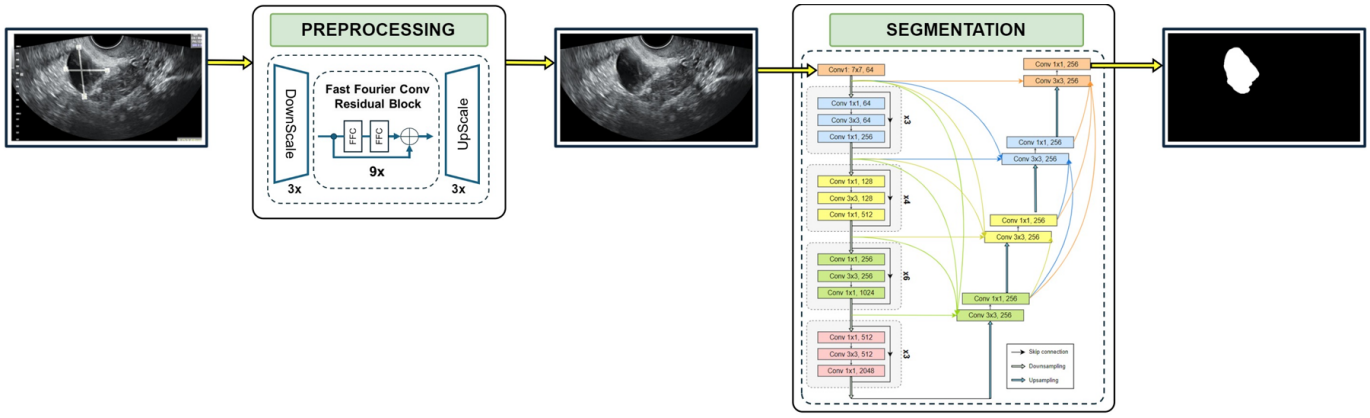


Fig. 1. The proposed framework for ovarian tumor segmentation.

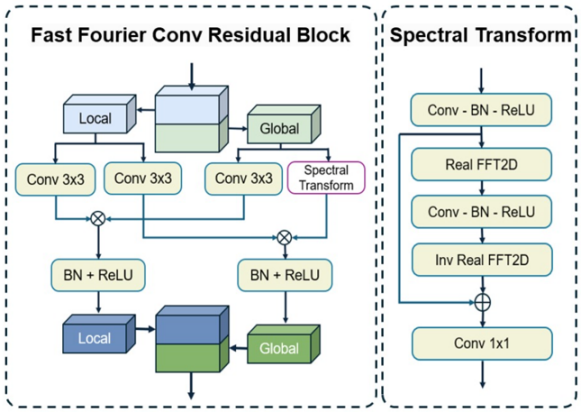


Fig. 2. Large Mask In-painting architecture.

as a pre-trained encoder, we leverage its ability to efficiently extract meaningful features from input images, significantly reducing both training time and computational demands compared to building a model from scratch. This approach is particularly advantageous because training from scratch can be resource-intensive and expensive. By leveraging the pre-trained ResNet50 encoder, we benefit from its proven performance and robustness, thereby improving the overall effectiveness and efficiency of the ovarian tumor segmentation task.

The decoder remains consistent with the original U-Net architecture. Integrating the U-Net architecture with ResNet50 has proven to be highly effective for segmentation tasks compared to the standard U-Net. The combination of U-Net and ResNet50 leverages the power of both models: the detailed spatial information captured by U-Net and the deep feature extraction capability of ResNet50. This synergy enhances the accuracy and efficiency of the segmentation process, making it particularly suitable for complex medical imaging tasks such as ovarian tumor detection and delineation. By maintaining the foundational elements of U-Net while incorporating the advanced features of ResNet50, this approach not only improves performance but also ensures that the segmentation process is more reliable and precise.

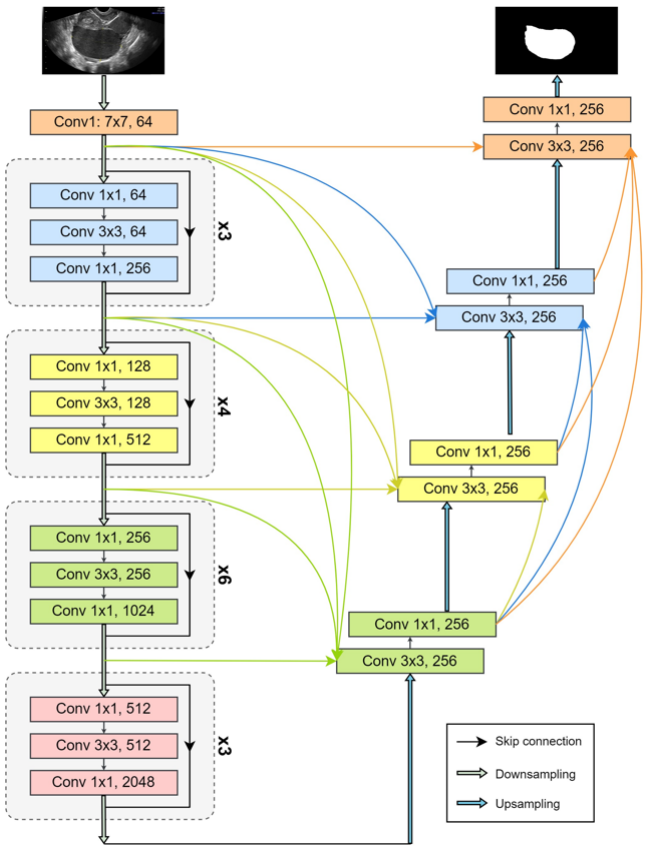


Fig. 3. The segmentation method incorporating ResNet50 and U-Net3+ architecture.

D. Loss function

For this study, in our experiment, we employ Binary Cross Entropy (BCE) [12], which is particularly suitable for binary segmentation tasks. The BCE is defined as follows:

$$\mathrm{BCE}(y, \hat{y}) = -\left(y \log(\hat{y}) + (1-y) \log(1-\hat{y})\right) \qquad (1)$$

where y represents the ground truth area of the ovary in ultrasound images, and $\hat{y}$ denotes the region of the ovary predicted by the model.

IV. Experiments and Results

A. Dataset

We utilized a benchmark dataset, OTU2D, which comprises 1469 ovarian ultrasound images collected from 294 patients at Beijing Shijitan Hospital and Capital Medical University, China [5]. However, in this study, we focus on binary segmentation to identify the presence or absence of tumors. To ensure a fair evaluation, we followed the split ratio protocol 8:1:1 images for training, testing, and validation, respectively. Figure 4 illustrates several sample images from the OTU2D dataset, showcasing the diverse types of tumors.

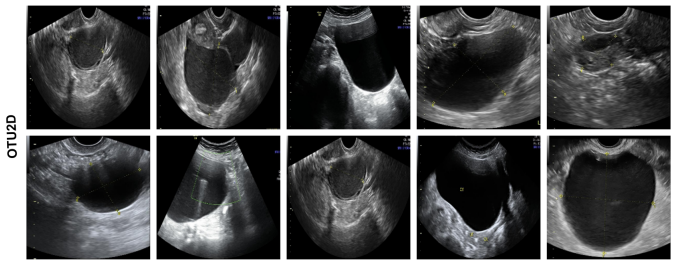


Fig. 4. Illustration of several ovarian tumor ultrasound images.

B. Experimental results

1) Qualitative evaluation of preprocessing step: Figure.5 illustrates some results of the preprocessing step. We can observe that markers, symbols, and other annotations have been removed. Although in some parts the contrast level is decreased, we can still see the detailed information of different parts of the ovarian area.

2) Ablation Study: To evaluate the role of each component in the proposed method, in our study, we have conducted three experiments. The first experiment aims to evaluate the performance of UNet3+, while the second experiment tried to assess the robustness of the encoder part. The final experiment is to compare the BCE loss function with other loss functions.

Table I shows the results of the first experiment. Among the three architectures that are UNet, UNet++, and UNet3+, UNet3+ demonstrates the best performance when using skip connections with Dice (84.6%), IoU (74.32%), and Recall (87.42%). Although UNet3+ outperforms UNet++ and UNet in terms of dice, IoU, and recall, UNet demonstrates higher precision (86.43%) in predicting the pixels of the regions belonging to ovarian tumors compared to UNet3+ (0.54%), with precision lower by 0.54% compared with UNet. The findings highlight a notable advantage when utilizing skip-connections to interconnect layers within UNet3+, leading to significantly superior performance compared to both UNet and UNet++ architectures.

In the second experiment, we compare the performance of three different encoders, which are VGG16, MobileNetV2, and ResNet50. Results are shown in Tab. II. ResNet50 outperforms VGG16 and MobileNetV2 in Dice (86.44%), IoU (77.05%), and Recall (89.18%), highlighting its effectiveness for high-quality segmentation. VGG16 has the highest Recall (88.11%) but is missing many important information. Fig. 6 shows results from different backbones, demonstrating that the proposed method closely matches the ground truth.

TABLE I
Comparison of Unet, Unet++ and Unet3+ backbones for ovarian tumor segmentation. Best results are in bold (%).

Model	Dice	IoU	Recall	Precision
UNet [6]	81.02	69.44	83.76	**86.43**
UNet++ [7]	73.68	59.34	79.07	82.55
UNet3+ [8]	**84.62**	**74.32**	**87.42**	85.89

TABLE II
Comparison of encoders when incorporating with UNet3+ architecture. Best results are in bold (%).

Model	Dice	IoU	Recall	Precision
MobileNetV2 + UNet3+	77.74	64.52	78.32	82.08
VGG16-Net + UNet3+	85.36	75.74	**88.11**	87.15
ResNet-50 + UNet3+	**86.44**	**77.05**	86.31	**89.18**

Loss function is also important in the segmentation model. Our third experiment aims at assessing the performance of loss functions. In the literature, there are several loss functions proposed for ovarian tumor segmentation. Table III highlights a comparison of the performance of various loss functions. The proposed method utilizing BCE loss stands out with Dice (86.44%) and IoU (77.05%), reflecting the effectiveness of the proposed model with BCE. In comparison, Dice loss performs well but is slightly lower than that achieved with BCE loss. IoU Loss shows the highest recall (89.29%) but a low precision (85.51%), indicating that it misses many important regions within the ovarian tumor region. Focal loss excels in precision (90.90%) but low recall (83.24%) and IoU (70.17%) show that it misses many other important regions that we need to segment. The Hybrid loss [2] and Joint loss [3] methods provide a balanced performance but do not surpass BCE loss in any metric. These results in Fig. 7 suggest that BCE loss is the most effective for our proposed method, achieving a good balance between precision and recall while maintaining high accuracy.

TABLE III
Comparison of loss functions. Best results are in bold (%).

Loss	Dice	IoU	Recall	Precision
BCE	**86.44**	**77.05**	86.31	89.18
Dice [13]	85.70	76.11	84.83	88.97
IoU [14]	86.14	76.61	**89.29**	85.51
Focal [15]	81.71	70.17	83.24	**90.90**
Hybrid [2]	85.67	75.89	83.97	89.89
Joint [3]	85.03	75.19	84.85	88.90

3) Comparison with the state-of-the-art methods: Table IV compares the segmentation results of the proposed method with the state-of-the-art methods on the OTU2D dataset. We can observe that the proposed method using BCE Loss achieves better performance than MU-Net [2].

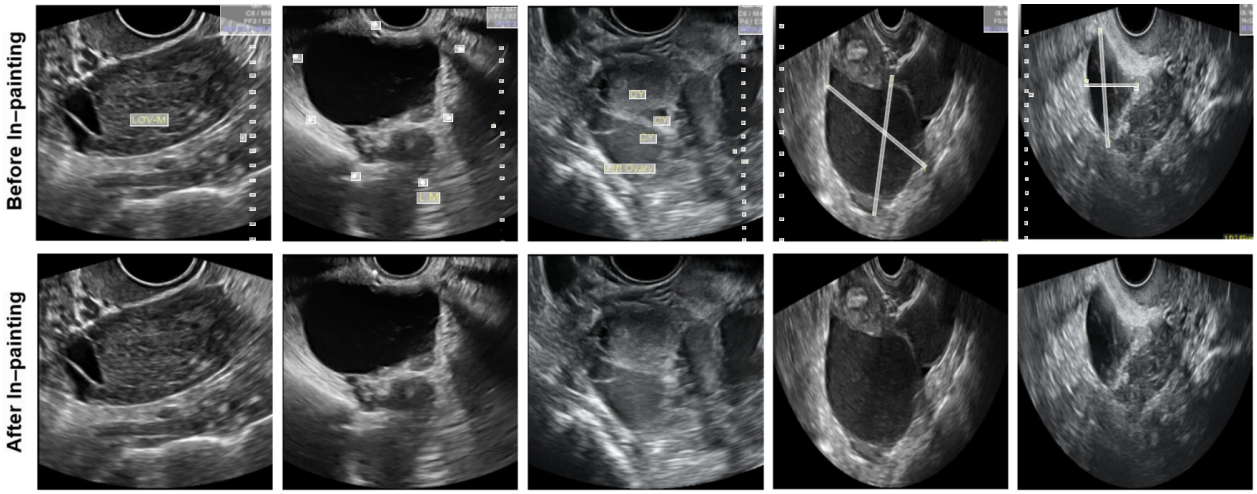


Fig. 5. Example of ultrasound images before an after preprocessing step in the OTU2D dataset. The first row shows the ultrasound images with markers highlighted by bounding boxes, while the second row shows the preprocessed images where markers and symbols are removed.

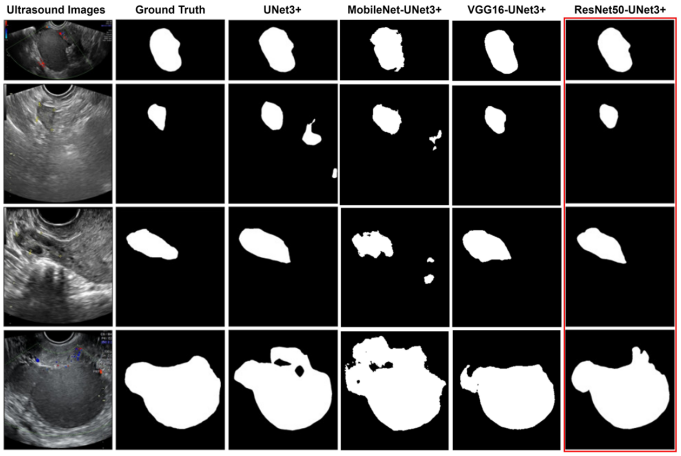


Fig. 6. Example of segmentation results from difference models. Results of the proposed result are highlighted in a red rectangle.

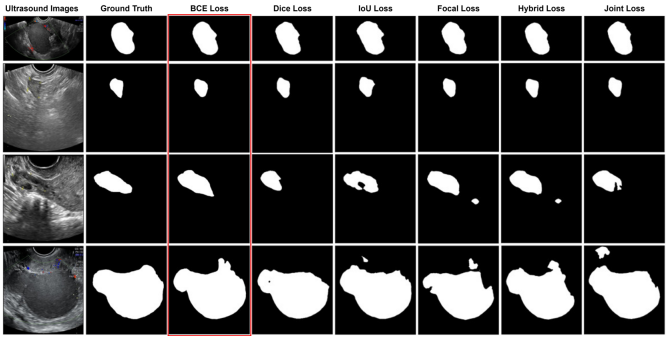


Fig. 7. Example of segmentation results from difference metrics loss. Results of the proposed method are highlighted in a red rectangle.

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON OTU2D DATASET (%)

Model	Loss	Dice	IoU	Recall	Precision
MU-Net [2]	BCE	56.70	39.80	77.20	82.70
MU-Net [2]	Hybrid	79.00	65.00	80.00	82.00
Proposed Method	BCE	**86.44**	**77.05**	**86.31**	**89.18**

To evaluate the effect of marker removal, 50 clean images from the OUT2D dataset were used. Two scenarios are defined using 1419 marker images for training. In the first scenario (without preprocessing), we employ the original images, while in the second scenario (with preprocessing), output images of LaMa models are utilized. Experimental results are reported in Tab. V. The results confirm the impact of symbol and marker removal in the preprocessing step, with the values of Dice, IoU, Recall, and Precision metrics being 96.6%, 93.57%, 97.78%, and 96.08%, respectively. Fig.8 illustrates some segmentation results.

TABLE V
COMPARISON RESULTS OF USING THE PROPOSED METHOD OF TESTING ON THE CLEAN IMAGES SUBSET IN THE OTU2D DATASET (%)

Scenario	Dice	IoU	Recall	Precision
Without preprocessing	90.63	83.02	89.32	86.23
With preprocessing	**96.64**	**93.57**	**97.78**	**96.08**

V. CONCLUSIONS AND FUTURE WORKS

In this study, we have proposed an effective framework for ovarian tumor segmentation by incorporating a pre-processing step and a segmentation network based on ResNet50 and U-Net3+. Several experiments were conducted on a benchmark dataset, demonstrating significant improvements, with the Dice Coefficient getting 86.64% on the OTU2D dataset and increasing from 90.63% to 96.64% when using the LaMa in-painting technique. Compared to previous studies, which often focus

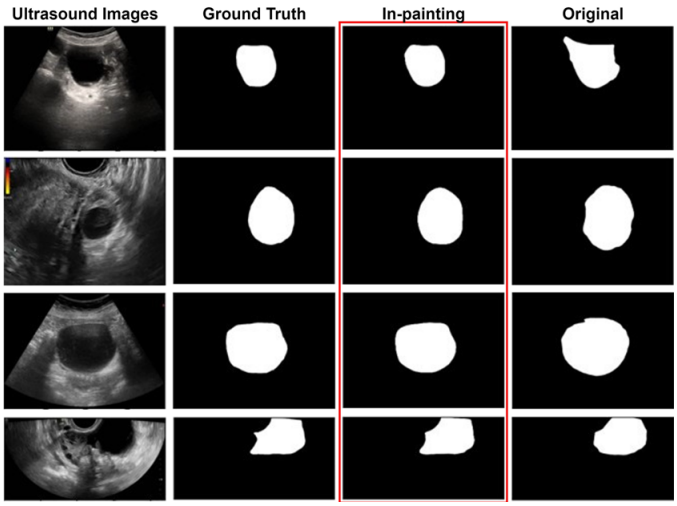


Fig. 8. Example of segmentation results with and without preprocessing step. Results with preprocessing step are highlighted in a red rectangle.

on enhancing segmentation performance without addressing the impact of annotations and noise, our work offers a more comprehensive approach by combining pre-processing with segmentation techniques. Our study contributes to the field by addressing this challenge, thus improving the performance of the segmentation tasks. However, despite the notable performance, our model still has limitations in the boundary. The future work will focus on refining these boundaries and leveraging the segmented regions to compute clinically relevant parameters, such as orthogonal diameters or boundary features.

ACKNOWLEDGEMENT

This research is funded by the Ministry of Science and Technology of VietNam (MOST) under grant number KC-4.0-45/19-25 **"Research and development of a computer-aided support system for ovarian cancer diagnosis using ultrasound images"**.

REFERENCES

[1] S. B. Coburn, F. Bray, M. E. Sherman, and B. Trabert, "International patterns and trends in ovarian cancer incidence, overall and by histologic subtype," *International Journal of Cancer*, vol. 140, no. 11, pp. 2451–2460, 2017. DOI: 10.1002/ijc.30676.

[2] T.-H. Nguyen, T.-L. Pham, Q.-V. Tran, *et al.*, "Systematic evaluation of loss functions for ovarian tumors segmentation from ultrasound images," in *2023 1st International Conference on Health Science and Technology (ICHST)*, 2023, pp. 1–6. DOI: 10.1109/ICHST59286.2023.10565336.

[3] H.-P. Luong, H.-S. Bui, N.-K. Nguyen, *et al.*, "Sovasegnet: Scale invariant ovarian tumors segmentation from ultrasound images," in *2024 IEEE International Conference on Image Processing (ICIP)*, 2024, pp. 2081–2087. DOI: 10.1109/ICIP51287.2024.10647995.

[4] R. Suvorov, E. Logacheva, A. Mashikhin, *et al.*, "Resolution-robust large mask inpainting with fourier convolutions," *arXiv preprint arXiv:2109.07161*, 2021.

[5] Q. Zhao, S. Lyu, W. Bai, *et al.*, "A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation," *arXiv preprint arXiv:2207.06799*, 2022.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention 2015*, Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[7] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," pp. 3–11, 2018.

[8] H. Huang, L. Lin, R. Tong, *et al.*, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1055–1059. DOI: 10.1109/ICASSP40776.2020.9053405.

[9] H. Li, J. Fang, S. Liu, *et al.*, "Cr-unet: A composite network for ovary and follicle segmentation in ultrasound images," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 4, pp. 974–983, 2020. DOI: 10.1109/JBHI.2019.2946092.

[10] L. Chen, C. Qiao, M. Wu, *et al.*, "Improving the segmentation accuracy of ovarian-tumor ultrasound images using image inpainting," *Bioengineering*, vol. 10, no. 2, p. 184, 2023, ISSN: 2306-5354. DOI: 10.3390/bioengineering10020184.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[12] S. Jadon, "A survey of loss functions for semantic segmentation," in *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2020, pp. 1–7. DOI: 10.1109/CIBCB48159.2020.9277638.

[13] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Cham: Springer International Publishing, 2017, pp. 240–248, ISBN: 978-3-319-67558-9.

[14] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proceedings of the 24th ACM international conference on Multimedia*, ACM, 2016, pp. 516–520.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017. DOI: 10.48550/arXiv.1708.02002. arXiv: 1708.02002 [cs.CV].

Enhancing Shear Wave Propagation Analysis in Tissue with Directional Filtering of Reflected Waves

Quang Hai Luong[1*], Duc-Nghia Tran[2], Nguyen Sy Hiep[3], Lam Sinh Cong[4], Duc-Tan Tran[5]
[1*]Le Quy Don Technical University, Hanoi, Vietnam
https://orcid.org/0009-0001-8860-4735
Email: luonghai@lqdtu.edu.vn
[2]Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi City, Vietnam
Email: nghiatd@ioit.ac.vn
[3]University of Information and Communication Technology, Thai Nguyen University, Thai Nguyen, Vietnam
Email: nshiep@ictu.edu.vn
[4]VNU University of Engineering and Technology, Hanoi City, Vietnam
Email: congls@vnu.edu.vn
[5]Faculty of Electrical and Electronic Engineering, Phenikaa University, Hanoi, Vietnam
https://orcid.org/0000-0002-7673-388X
Email: tan.tranduc@phenikaa-uni.edu.vn

***Abstract*— Shear wave elastography (SWE) is a rapidly developing imaging technique that enables disease diagnosis by estimating tissue elasticity and viscosity (collectively referred to as CSM). SWE relies on measuring the velocity of shear waves propagating through the tissue. This paper proposes improvements to a directional filter for removing the impact of reflected noise on shear wave propagation. Firstly, the measured velocity signal (a combination of incident and reflected waves) was transformed using a discrete Fourier transform and then multiplied by a normalized directional filter mask. In the next step, signal processing techniques were applied to improve the quality of filtering and eliminate undesired components. Subsequently, the processed signal was inverse discrete Fourier transformed. The obtained signal after this transformation was the filtered velocity of the shear wave. The results of the simulation demonstrated the effectiveness of the proposed solution.**

Keywords: shear wave; directional filter; reflected waves; elastography

I. INTRODUCTION

Shear wave elastography imaging (SWEI) has been widely used in medical image diagnosis, especially in the early detection of tumors [1]. This technique was also integrated in modern ultrasound devices for diagnosis liver diseases [2], [3]. Scientists are still actively studying SWEI for enhancing the quality of SWEI and expanding the range of application of SWEI (breast, thyroid, prostate tissues, ...).

Zheng et al., in 2007, proposed using a Kalman filter to estimate the phase difference. This allows the determination of shear wave speeds and dispersion curves with different frequencies [4]. However, their approach requires many complex calculations with many frequencies to reconstruct the shear module. This can potentially lead to artifacts. Additionally, the authors did not mention the reflected noise in their study.

In the study [5], Hengzhao et al. calculated the attenuation of shear wave to determine the viscosity and elasticity of tissues. They used a probe that created a non-focused ultrasound beam (4.1 MHz with 128 elements and 400s duration) to generate shear waves. Another probe which is placed on the opposite was used to measure signals. The values of c_s and α_s were estimated by 2D Fourier transformation (K-space). The authors used concurrently the peak extraction method and the full-width-at-half-maximum (FWHM) method. The important factors in evaluating tissue properties are the viscosity and elasticity values. However, the authors did not mention the types of noise generated in their study. The phantom sample which used in their study had homogeneous viscosity and elasticity.

Thomas deffieux et al. did research on the impacts of reflected noise in transient shear wave elastography [6]. Next, they proposed using a simple directional filter which is possible to mostly remove the reflected noise and reduce those artifacts in the shear modulus maps of a stiff inclusion. Simulation results proved the quality of their proposed solution.

In a recently study [7], the authors used two methods as the algebraic inversion of differential equation (AIDE) and the local frequency estimate (LFE) to calculate stiffness values in 2 case (use and without the directional filter). They used Magnetic resonance elastography (MRE) to observe the propagation of shear wave in tissues. Finally, they made a table for comparison of stiffness values acquired from all above cases.

In this paper, the authors focus on enhancing the directional filter for removal of the reflected noise which appears when shear wave propagates to the bounder of heterogeneous medium layers. In addition, the Finite Difference Time Domain (FDTD) technique was used to simulate the propagation of shear waves in tissues.

II. PROPOSED METHOD

In this section, authors present about the principle of shear wave generation and measurement, the mathematical model of the propagation of shear in tissues, simultaneously proposing the method for removal of the reflected noise.

A. Shear wave generation and measurement

Fig 1 illustrates the shear wave generation and measurement. A needle which is made from stainless-steel or hard plastic, with diameter of 1.5 mm and long of 12 cm is controlled by a mechanical actuator to vibrate along the z-axis at a frequency in the range of 100 Hz to 500 Hz. Consequently, the shear wave is generated and propagates in tissue in the perpendicular plane, while the particle velocity measurement of shear wave is performed by using a Doppler ultrasound system [8].

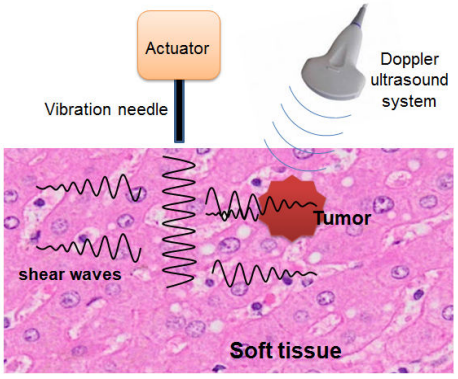


Fig 1. The incident shear wave propagates in tissue [9].

B. Modeling shear wave propagation in tissue

When shear wave propagates in tissue following the x-direction in Cartesian coordinates, its particle velocity vector and the stress tensor are related with each other. This relationship is represented by equations (1) and (2) [10].

$$\rho \partial_t v_z = \partial_x \sigma_{zx} \tag{1}$$

$$\partial_t \sigma_{zx} = (\mu + \eta \partial_t) \partial_x v_z \tag{2}$$

where η and μ are the viscosity and elasticity values of tissue, respectively; ρ is the density of tissue; ∂t stands for the partial derivative operator ∂/∂t of expressions on its right, while ∂x stands for the partial derivative operator ∂/∂x of expressions on its right.

The authors used the FDTD method [11] to expand equations (1) and (2) into equations (3) and (4), respectively.

$$v_z^{n+1}|_i = v_z^{n}|_i + \frac{\Delta_t}{\rho \Delta_x}\left(\sigma_{zx}^{n+\frac{1}{2}}|_{i+\frac{1}{2}} - \sigma_{zx}^{n+\frac{1}{2}}|_{i-\frac{1}{2}} \right) \tag{3}$$

$$\begin{aligned} \sigma_{zx}^{n+\frac{1}{2}}|_{i+\frac{1}{2}} &= \sigma_{zx}^{n-\frac{1}{2}}|_{i+\frac{1}{2}} + \frac{\mu \Delta_t}{\Delta_x}\left(v_z^{n+1}|_{i+1} - v_z^{n+1}|_i \right) \\ &+ \frac{\eta}{\Delta_x}\left(v_z^{n+1}|_{i+1} - v_z^{n+1}|_i \right) - \frac{\eta}{\Delta_x}\left(v_z^{n}|_{i+1} - v_z^{n}|_i \right) \end{aligned} \tag{4}$$

where Δx is distance between consecutive two spatial points (spatial step), Δt is the sampling cycle (time step), "i" and "n" are indexes for spatial and temporal, respectively.

The shear waves are also mechanical waves. When the shear waves strike a barrier or the bounder of heterogeneous medium layers, at least part of the energy of the waves will be reflected back into the medium from which they came. The reflected and incident waves are in opposite phase with each other. The propagation model of the reflected wave is similar to the incident waves.

C. Using the directional filter for removal of reflected noise

The measured signal is the combined waves of incident waves and reflected waves. Firstly, the measured signal is transformed Discrete Fourier (DFT). Then, the signal which is output of DFT is multiplied by a directional filter mask. After this process, Fourier image of incident waves are selectively retained, while others are removed. The result after

multiplication with mask is transformed Inverse Discrete Fourier Inverse (IDFT). The output signal of IDFT is the filtered signal. These procedures are our proposed solution for removal the reflected noise. The procedure for removal of reflected noise is demonstrated in Fig 2.

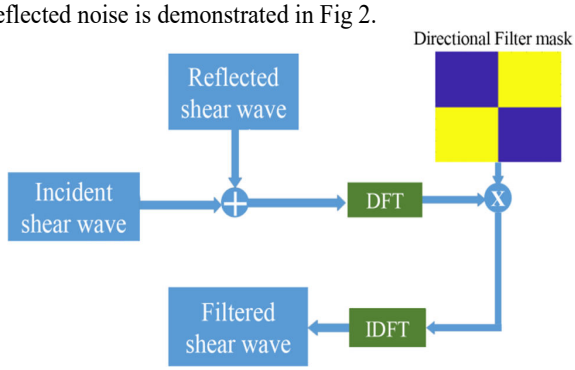


Fig 2. The procedure of filtering the reflection noise.

In fact, to perform DFT and IDFT, we use the functions *fftshift*(*fft2*(...)) and *ifft2*(*ifftshift*(...)) in the MATLAB environment. For two-dimensional signals, *fftshift* shifts the zero-frequency component to the center of both dimensions. This is particularly useful in image processing, where the frequency domain representation is often more interpretable when the low-frequency components are centered. When working with inverse FFT (IFFT), using *ifftshift* is often necessary to reverse the shift applied by *fftshift*, ensuring that the spatial or temporal domain signal is correctly reconstructed. This is an improvement over our previous works.

Fig. 3 shows the magnitude of a 2D Fourier transform applied to time-space signals. The presence of multiple peaks suggests that the original summed signal might contain both incident and reflected waves. The smaller peaks around the main peaks could be noise resulting from the interaction of the incident wave with the tumor. The reflected noise components, which are highlighted in red ovals, need to be removed (see Algo. 1).

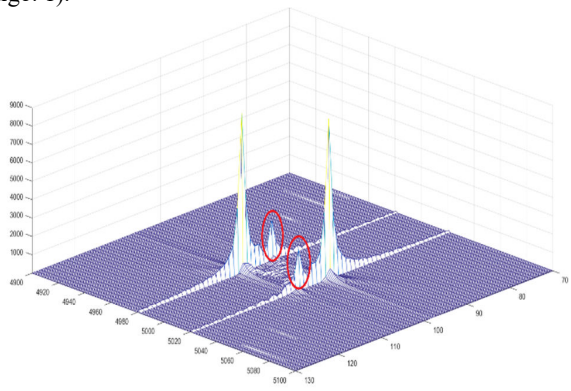


Fig 3. The 2D-Fourier applied to the time-space signal

Algorithm 1: Proposed Algorithm for CSM Estimation

Step 1: Determine the sampling frequency of the discrete signal. Let's denote the sampling frequency as *Fs* (in Hz).
Step 2: Collect the filtered discrete time-domain signal $v_z(x,y,n)$ whose interest frequency of *f*. Let's denote the signal length as *N* (in Hz), and *n* is the discrete time index.
Step 3: Determine the frequency resolution of the DFT bins. The frequency resolution, Δf, is given by $\Delta f = Fs / N$
Step 4: Determine the spatial resolution Δx, Δy
Step 5: Compute 2D-FFT of the time - space signal
Step 6: Design the directional filter mask to be the same size as the signal in step 5. The value of the mask component should be assigned to 0 if the location is within the red ovals. Otherwise, it should be assigned to 1.
Step 7: Multiple the the directional filter mask with the signal in step 5
Step 8: Compute 2D-IFFT of the frequency - *k*space signal
Step 9: Estimate the CSM as presented in [9]

III. RESULTS AND DISCUSSION

We set up a simulation scenario to evaluate the proposed method. The one-dimensional medium is 2 cm in size and digitalized by 200 pixels (respectively, the distance between two consecutive pixels is 0.1 mm). The mass density of medium (tissue) is 1000 kg/m³. The viscosity and elasticity values of medium are shown in Table 1. A needle which contacts with the tissue is controlled to vibrate with a vibration frequency of 200 Hz and amplitude of 2 mm. The shear wave is generated and propagates in tissue. The reflected noise appears at pixel 79. There the viscoelastic value of the tissue changes.

Table 1: The elasticity and viscosity values of tissue.

Pixels	Elasticity value (Pa)	Viscosity value (Pa.s)
1 to 78	1000	0.1
79 to 200	1200	0.2

Fig 4 illustrates the incident wave in space propagating in tissue, while Fig 5 shows the reflected wave in space. The both (incident and reflected wave) have an attenuated sinusoidal shape. They were simulated by using the FDTD technique.

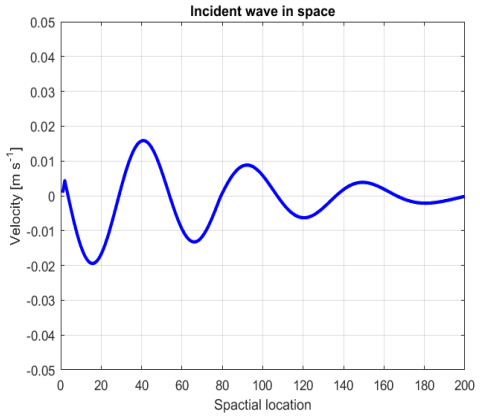


Fig 4. The incident shear wave propagates in tissue.

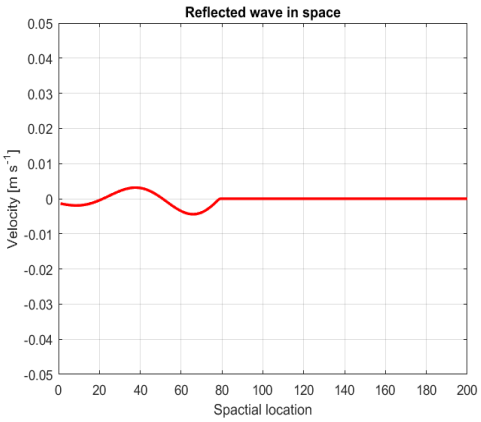


Fig 5. The reflected shear wave propagates in tissue (noise).

Shear wave is a matrix 2D (10.000, 200) with 200 points in space and 10.000 sample in time for each point. So Fourier transformation of shear wave is also matrix 2D (10.000, 200). Fig 6 is image of Fourier transformation of the incident shear wave. To see clearly central points, we only observe image in range of (4800:5200, 60:140).

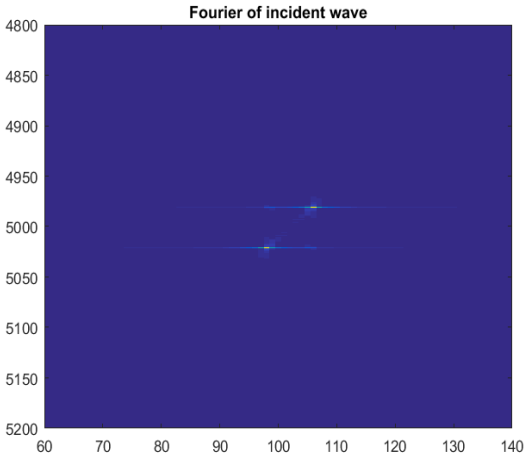


Fig 6. The image of Fourier transformation of incident wave.

Fig 7 illustrates image of Fourier transformation of the reflected shear wave. It can be seen that image of Fourier transformation of the reflected wave opposed to of the incident wave.

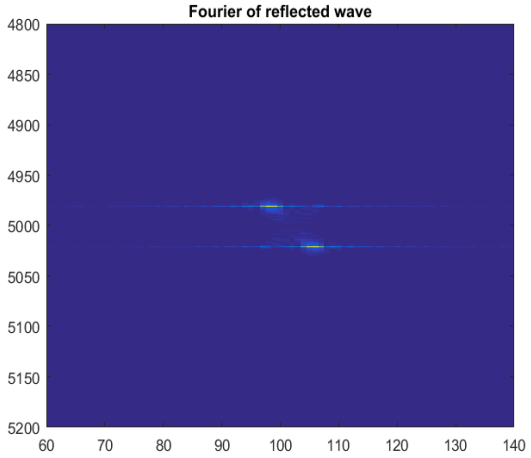


Fig 7. The image of Fourier transformation of reflected wave.

Fig 8 illustrates image of Fourier transformation of the combined wave. In fact, the combined wave is the measured signal which is combination of incident wave and reflected wave (noise).

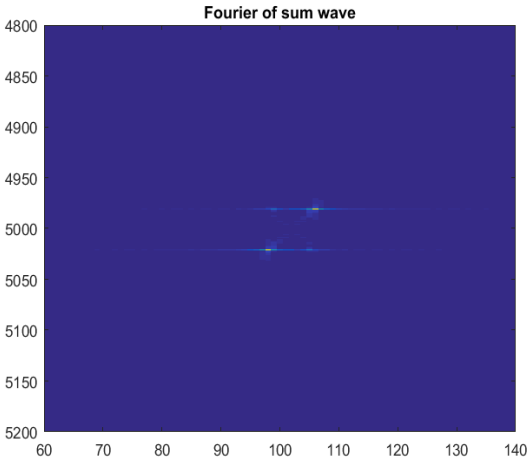


Fig 8. The image of Fourier transformation of sum wave.

After applying the directional filter, the image of Fourier transformation of the filtered signal is shown in Fig 9. It is almost similar to Fig 6. This means that the noise component has been removed from the measured signal.

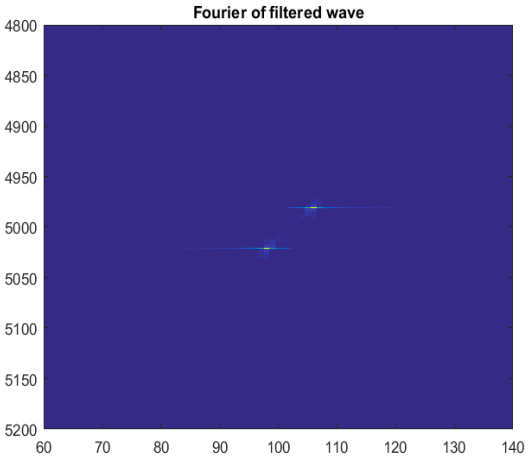


Fig 9. The image of Fourier transformation of filtered wave.

Fig 10 shows the particle velocity of shear wave in space. More detail, blue graph represents the ideal velocity, black graph represents the velocity with noise and red one is the filtered velocity. From point 1 to 78, the measured particle velocity of shear wave is affected by reflected noise, so its amplitude is reduced. However, after applying directional filter, the velocity of shear wave is restored and it is similar to ideal velocity.

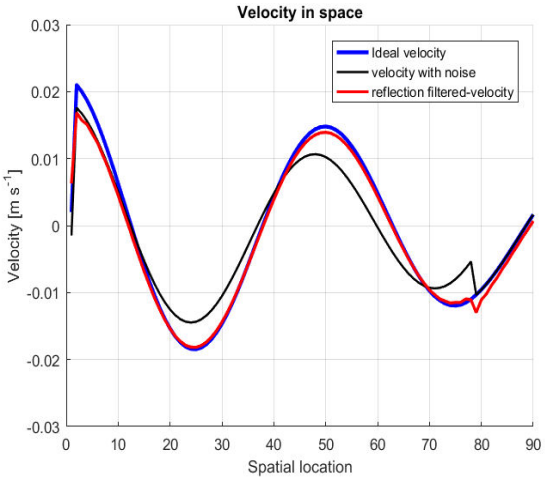


Fig 10. The particle velocity of shear wave in space.

Fig 11 shows the particle velocity of shear wave in time at location 50. This point is in range of [1, 78], so the particle velocity in time at here is affected by reflected noise. The particle velocity in time is a sinusoidal harmonic oscillation. The amplitude of measured particle velocity is reduced comparing to ideal particle velocity, while the filtered velocity is almost similar to the ideal velocity. This proves the effectiveness of the directional filter.

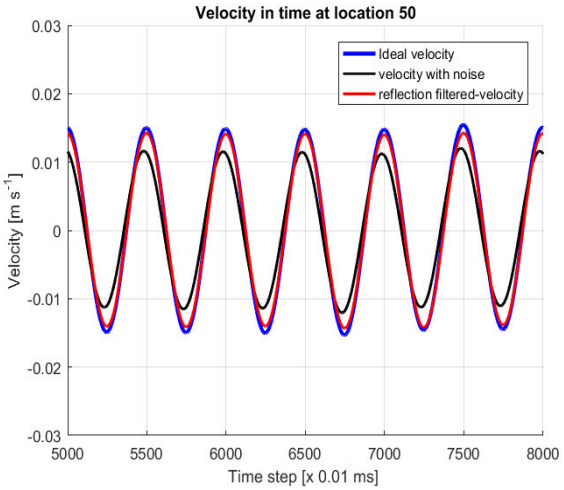


Fig 11. The particle velocity of shear wave in time.

To quantify the efficiency of the proposed method, the authors used the normalized error which be defined as equation (5). We perform to compare the normalized error between the filtered particle velocity and the ideal to the normalized error between the particle velocity with noise and the ideal, over the entire space and time domain.

$$\in_v = \frac{1}{X \times N} \sum_{x}^{X} \sum_{t}^{N} \frac{\left| v_{ideal(x,t)} - v_{(x,t)} \right|}{v_{ideal(x,t)}} \tag{5}$$

where X is the number point in space, N is number sample in time, v_{ideal} is the ideal particle velocity, v is the filered particle velocity/ the particle velocity with noise. The results corresponding to the changes of the amplitude of noise are presented in Table 2. This table presents normalized error (%) of $\in_v$ for without and with proposed directional filter at different Signal-to-Noise Ratios (SNR), highlighting the effectiveness of filtering in noise reduction. The SNR values are obtained by changing proactively the amplitude of the reflected noise. As the SNR decreases, the percentage error in the velocity without noise significantly increases, ranging from 1.926% at an SNR of 11.63 dB to 3.576% at an SNR of 3.67 dB. Conversely, the filtered velocity shows a much smaller increase in percentage error, from 1.193% to 1.264% over the same SNR range. This demonstrates that the filtering process substantially mitigates the impact of noise, maintaining lower and more stable error percentages compared to the unfiltered velocity. The data suggests a clear correlation between higher SNR values and reduced normalized errors in velocity

measurements, emphasizing the importance of effective noise filtering techniques in achieving accurate and reliable results.

Table 2: The normalized error (%) of $\in_v$

	SNR (dB)	Without directional filter	With proposed directional filter
1	11.63	1.926	1.193
2	10.22	2.110	1.197
3	7.89	2.498	1.211
4	5.61	3.007	1.232
5	3.67	3.576	1.264

IV. CONCLUSIONS

In this work, the authors wish to improve the directional filter to remove the reflected noise which appears when shear wave propagates to the bounder of heterogeneous medium layers. Therefore, we only focus on simulating the impact of the reflected noise on the propagation of shear waves in tissue. The FDTD technique is used to simulate the propagation of the incident and reflected waves. Simulation results are very good with the normalized error of about 1.2%.

REFERENCES

[1] J. Bercoff, A. Criton, C. Bacrie, J. Souquet, M. Tanter, J. Gennisson, T. Deffieux, M. Fink, V. Juhan, A. Colavolpe et al., "ShearWave Elastography A new real time imaging mode for assessing quantitatively soft tissue viscoelasticity," in *Ultrasonics Symposium*, 2008. IEEE, 2008, pp. 321–324.

[2] G. Ferraioli, P. Parekh, A. B. Levitov, and C. Filice, "Shear wave elastography for evaluation of liver fibrosis," *Journal of Ultrasound in Medicine*, vol. 33, no. 2, pp. 197–203, 2014.

[3] Wang, Jing-Houng. "Application of ultrasound liver elastography to the diagnosis and monitoring of liver disease." *Journal of Medical Ultrasound* 27.1 (2019): 1-2.

[4] Y. Zheng, S. Chen, W. Tan, R. Kinnick and J. F. Greenleaf, "Detection of tissue harmonic motion induced by ultrasonic radiation force using pulse-echo ultrasound and kalman filter," in *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 54, no. 2, pp. 290-300, February 2007, doi: 10.1109/TUFFC.2007.243, 2007.

[5] Zhao, Heng & Qiang, Bo & Amador Carrascal, Carolina & Song, Pengfei & Urban, Matthew & Kinnick, Randall & Greenleaf, James & Chen, Shigao. (2012). "Measure Elasticity and Viscosity Using the Out-of-plane Shear Wave". *IEEE International Ultrasonics Symposium*, IUS. 10.1109/ULTSYM.2012.0053.

[6] T. Deffieux, J. -l. Gennisson, J. Bercoff and M. Tanter, "On the effects of reflected waves in transient shear wave elastography," in *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control,* vol. 58, no. 10, pp. 2032-2035, doi: 10.1109/TUFFC.2011.2052, 2011.

[7] Maharjan, Surendra, et al. "Directional filter, local frequency estimate and algebraic inversion of differential equation of psoas major magnetic resonance elastography." *Open Journal of Medical Imaging* 10.1 (2020): 1-16.

[8] M. Orescanin and M. F. Insana, "Model-based complex shear modulus reconstruction: A Bayesian approach," in IEEE Ultrasonics Symposium (IUS). IEEE, 2010, pp. 61–64.

[9] Luong, Quang-Hai, et al. "Estimation of elasticity and viscosity in heterogeneous medium using FDTD method and AHI algorithm." *2016 International Conference on Advanced Technologies for Communications (ATC)*. IEEE, 2016.

[10] M. Orescanin, Y. Wang and M. F. Insana, "3-D FDTD simulation of shear waves for evaluation of complex modulus imaging," in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 58, no. 2, pp. 389-398, February 2011, doi: 10.1109/TUFFC.2011.1816, 2011.

[11] Schneider, J. B. "Introduction to the Finite-Difference Time-Domain Method: FDTD in 1D." *Understanding the Finite-Difference Time Domain Method, Seattle* (2016): 33-35.

Structural Analysis of Asian and African Rice Panicles via Transfer Learning

Cong Hieu Le[1], Lam Thai Nguyen[2], Trung Kien Pham[2], Le Khanh Nguyen[3], Tran Hiep Dinh[2,*], Stefan Jouannic[4], Helene Adam[4], Pierre Duhammel[5], Hoang Trong Minh[6], and Nguyen Linh Trung[1]
[*]*Corresponding author: Tran Hiep Dinh (tranhiep.dinh@vnu.edu.vn)*
[1]Advanced Institute of Engineering and Technology, VNU University of Engineering and Technology (VNU-UET), Vietnam
[2]Faculty of Engineering Mechanics and Automation, VNU-UET, Vietnam
[3]Faculty of Agriculture Technology, VNU-UET, Vietnam
[4]DIADE, University of Montpellier, IRD, CIRAD, 34394 Montpellier, France
[5]L2S, CNRS, CentraleSupelec, Paris-Saclay University, France
[6]Faculty of Telecommunications 1, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam

***Abstract*—Rice panicle architecture affects yield and grain quality in both Asian (*Oryza sativa*) and African rice (*Oryza glaberrima*). Manual low-throughput phenotyping methods are inaccurate and non-repeatable. This study introduces a novel image-based method for automated junction detection using transfer learning and advanced image processing, leveraging visual similarities between pavement cracks and rice panicles. Seven pre-trained pixel-level crack segmentation models are employed on the panicle images to extract their morphology. Skeletonization and some image processing techniques are then applied to detect axis junctions from the resulting binary images. Experiment results revealed U-Net's robust performance with nearly 90 % accuracy in segmentation tasks. Zhang-Suen thinning combined with Crossing Number reached about 80 % accuracy for main axis junctions detection. High-order junction detection was challenging, especially for the complex Asian panicles. This study shows a correlation between pavement cracks and rice panicles, demonstrating the potential for automated rice phenotyping. Future efforts will focus on enhancing detection accuracy and extending the approach to include grain detection.**

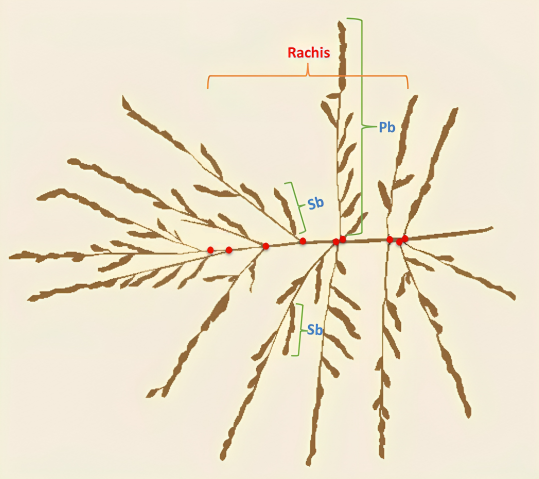


Fig. 1. Structure of a rice panicle.

I. Introduction

Rice inflorescense (panicle) architecture plays a pivotal role in rice breeding for yield and grains quality. In both Asian rice (*Oryza sativa*) and African rice (*Oryza glaberrima*), panicle structure, including size, shape, and branching patterns, directly affects grain number and uniformity, impacting overall yield [1]–[3]. Amid ongoing climate change, breeding for sustainable genetic improvement is crucial to maintain rice yield and quality [4]. Despite advances in sequencing and genotyping, progress is hindered by manual, low-throughput phenotyping, which is time-consuming, inaccurate, and non-repeatable for dry rice panicles [1]. Therefore, high-throughput phenotyping methods are needed to address these challenges.

Plant phenotyping involves measuring observable plant traits. For rice inflorescence, this includes panicle length, branch number and order, and grain count and size [1]. These insights help researchers explore rice diversity and optimize breeding and yield. Rice panicle branching is shaped by a complex interplay of environmental factors like climate, soil, and light, along with genes and hormonal regulation. The rice panicle structure includes the main axis (Rachis), primary branches (Pb), secondary branches (Sb), tertiary branches (Tb) and higher-order branches, with a central main axis bearing lateral primary branches that further support secondary and higher-order branches. Nodes, shown as red dots in Fig. 1, along with contact points of different order branches, are hereinafter referred to as junctions.

Detection of rice panicle junctions serves as an intermediary step in uncovering new traits and collecting phenotypic data to identify key genes linked to panicle diversity. The detected junctions facilitate the estimation of crucial traits like primary branch length, the length and count of higher-order branches, and their positions on predecessor branches. By employing automated techniques for junction detection, we can streamline the phenotyping process, addressing a need in rice research.

Interestingly, the process of detecting junctions in rice panicles shares notable similarities with crack detection in materials. Both tasks involve identifying thin, continuous branching patterns within complex images, and both can benefit from

advanced image analysis techniques. This parallel suggests a promising avenue for cross-domain methodologies, where successful crack segmentation models could be adapted to enhance the accuracy and efficiency of rice panicle structural analysis. Notable segmentation models that have been applied in crack detection include DCrack [5], FCN [6], SegNet [7], U^{2}CN [8], U-Net [9], ACS [10], RUC-Net [11], among others [12], [13]. Advancements in crack segmentation can improve the analysis of rice panicle structures, helping to overcome key challenges in phenotyping and genetic improvement efforts.

In this paper, we introduce a novel cross-domain method to automate the junction detection process in rice phenotyping. The proposed pipeline involves several key steps: segmentation via transfer learning, skeletonization, and junction detection. The contributions of this paper are summarized as follows.

1) Developing a novel method using transfer learning to fully automate the detection of rice panicle junctions, previously unattainable by existing platforms.
2) Conducting comprehensive experiments with various crack segmentation models and image processing algorithms on three datasets to test the feasibility of a general approach to both Asian and African rice panicles.
3) Addressing a crucial need in rice phenotyping, thus resolving a significant bottleneck in rice breeding and enabling precise high-throughput rice phenotyping.

The rest of this paper is organized as follows. Section II reviews the related high-throughput phenotyping tools. Section III outlines our pipeline for detecting rice panicle junctions. Section IV presents the experimental results. Section V discusses the limitations. Finally, Section VI concludes the paper.

II. Related Works

Notable high-throughput phenotyping tools, such as P-TRAP [1] and PANorama [2], have demonstrated satisfactory results in rice panicle structural analysis when compared to expert verification and traditional academic methods.

P-TRAP is an open-source tool for analyzing rice panicle images via a user-friendly Graphical User Interface (GUI). It offers three main functions: analyzing panicle structure, counting grains, and analyzing grain shape. P-TRAP outputs traits like axis lengths, diameters, grain numbers, areas, and perimeters. Despite its comprehensive functionalities, P-TRAP's manual steps - cropping images, setting a pixel-to-unit ratio, and defining generating points - are tedious and prone to errors. Additionally, these manual steps are needed for each image and require domain knowledge.

PANorama, another open-source phenotyping platform, excels in runtime and offers an additional functionality of detecting awns from branches, which can interfere with accurate phenotyping. PANorama achieves a more accurate skeleton compared to P-TRAP and can extract a similarly comprehensive set of structural features from given panicle images. However, user intervention is required for the specification of three junctions: the exsertion point, the initial main axis point, and the final main axis point. While these user inputs facilitate comprehensive processing, they fail to fully automate the process and remain prone to errors.

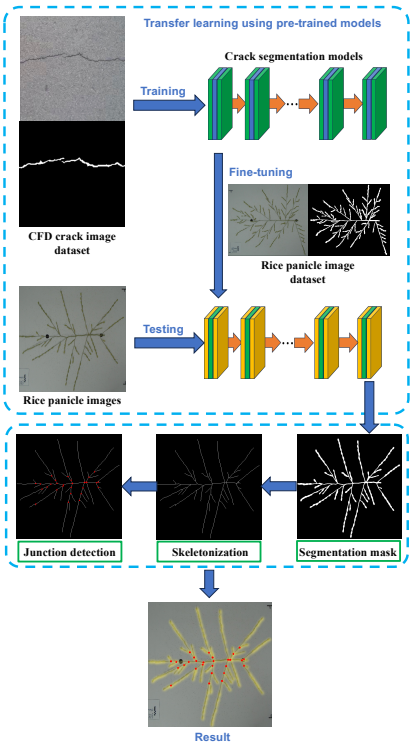


Fig. 2. The proposed pipeline.

Given the limitations of current tools, a fully automated process for rice panicle phenotyping is essential. Automation would eliminate tedious manual steps, reduce the potential for errors, and accelerate analysis, leading to more efficient and accurate phenotyping that enhances rice breeding and yield.

III. Methodology

A. Our proposed method

We present a novel method to automatically locate junctions in rice panicle images, using 7 state-of-the-art pre-trained segmentation models on an augmented version of the CrackForest Dataset (CFD) [14], combined with thinning and clustering algorithms. The proposed pipeline, as illustrated in Fig. 2, consists of the following steps:

- Step 1 (Segmentation): Rice panicles are segmented via transfer learning by employing a pre-trained model on the CFD dataset that has been fine-tuned with our dataset.
- Step 2 (Skeletonization): Rice panicles are transformed into unitary thickness skeletons, preserving shape and connectivity while eliminating most original white pixels.
- Step 3 (Junction detection): Clustering and image-processing methods are used to detect junctions, which are then compared to pre-labelled junctions for evaluation.

B. Segmentation

The segmentation models have been pre-trained on the CFD dataset, containing over 9000 road crack images of size

256×256, reflecting urban road surface condition [15]. A brief description of each model is given as follows.

- DCrack [5]: A trainable convolutional neural network for crack detection, using an encoder-decoder architecture with skip connections for improved information flow.
- FCN [6]: A concrete crack detection method employing VGG16 as its backbone for the encoder network.
- SegNet [7]: Optimized for pixel-wise semantic segmentation, SegNet utilizes an encoder-decoder architecture for effective road scene and indoor analysis.
- U-Net [9]: Initially designed for biomedical image segmentation, U-Net features a dual-path architecture for context capture and precise localization, making it versatile for medical and non-medical segmentation tasks.
- U^{2}CN [8]: A two-level nested U-structure for automatic pavement crack detection.
- ACS [10]: ACS incorporates an adaptive cost-sensitive loss function to address imbalanced data issues, resulting in improved performance in detecting road cracks.
- RUC-Net [11]: A semantic segmentation network that merges U-Net and ResNet architectures for pixel-level pavement crack image segmentation.

C. Skeletonization

Two skeletonization methods, including Zhang-Suen (ZS) [16] and a method proposed in [17], are used to transform the segmentation mask's foreground regions into a skeleton, maintaining the rice panicle's shape and connectivity while reducing the number of original white pixels. The ZS thinning method removes all contour points except those essential for creating an unitary thickness skeleton. This classical algorithm has proven effective, even in complex tasks like estimating the 3D skeleton, shape, and camera viewpoints of wild animals [18]. Widely adopted across various fields, this method promises to enhance the effectiveness of our proposed pipeline. The skeletonization method introduced in [17], presented by the author as the first three-dimensional algorithm compatible with gradient-based optimization, also ensures the preservation of an object's topology. This method, hereinafter referred to as GBO, involves two key steps for a binary image: (1) identification of simple points and (2) removal of simple points. By comparing these two methods, we can determine the most robust approach for thinning the rice panicle structure, ensuring reliable subsequent processing steps.

D. Junction detection

The assumption that the densest areas of a skeleton lie around intersections of different branches suggests a strategy leveraging clustering to detect these subsets. In this context, a sufficiently small subset can strongly indicate the position of the junction, typically located at the cluster center. For the task of junction detection, we employ three techniques:

- DBSCAN [19]: A foundational method renowned for detecting clusters of various shapes in noisy datasets.

TABLE I
PERFORMANCE (%) OF FINE-TUNED MODELS.

Model	DCrack	FCN	SegNet	U-Net	U^{2}CN	ACS	RUC-Net
AsP							
Pr	89.46	86.80	77.78	89.60	**91.61**	88.21	86.02
Rc	89.73	88.35	**94.95**	90.18	86.81	90.23	92.07
F_1	89.53	87.41	85.26	**89.85**	89.10	89.18	88.89
AfP							
Pr	89.14	86.09	77.18	89.26	**90.93**	87.95	86.00
Rc	89.24	88.49	**94.72**	89.78	86.72	89.89	91.40
F_1	89.08	87.04	84.84	**89.42**	88.69	88.83	88.52
AsAfP							
Pr	90.10	86.95	77.80	89.05	**91.72**	88.26	86.38
Rc	88.73	89.13	**95.41**	90.91	86.66	90.22	91.68
F_1	89.34	87.86	85.51	**89.91**	89.05	89.17	88.88

- FINCH [20]: Utilizes a single equation to generate hierarchical partitions based on an adjacency link matrix without the need for hyperparameters.
- Crossing Number (CN) [21]: Analyzes each pixel in a binary image and classifies it based on neighboring pixels.

E. Dataset and evaluation metrics

We test the effectiveness of our method on 200 images of rice panicles, consisting of 100 samples each of domesticated Asian and African rice. Each rice panicle is meticulously captured in an extended position, secured against a white background, and positioned to fit within the frame. Input images are in Red Green Blue format and vary in size.

We assess the accuracy of both segmentation and junction detection using the F_1 score, the harmonic mean of precision (Pr) and recall (Rc). For segmentation, we compare ground truth images with segmentation masks, while for junction detection, we assess pre-labeled junctions against our predictions. The F_1 score provides an overall assessment, with high precision indicating fewer incorrect detections and high recall reflecting effectiveness in capturing most target junctions.

IV. RESULTS

We evaluate our method on 200 rice panicle images from Asia and Africa, using the F_1 score for both segmentation and junction detection accuracy. We also perform 5-fold cross-validation for thorough segmentation evaluation and model selection, ensuring a reliable measure of each model's effectiveness and generalization to new data. The assessments are based on three experiments: (1) Asian rice panicles (AsP), (2) African rice panicles (AfP), and (3) mixed Asian and African panicles (AsAfP), accounting for structural differences, with Asian rice having more complex branching, allowing us to evaluate whether a general approach applies to both species.

A. Rice panicle segmentation performance

Seven pre-trained segmentation models on the CFD dataset are fine-tuned using 80 images of Asian rice panicles for the first test, 80 images of African rice panicles for the second, and 160 combined images for the third. After fine-tuning, they are tested with the remaining 20 images of Asian rice panicles, 20 of African rice panicles, and 40 combined images. The average performance of each model is recorded in Table I.

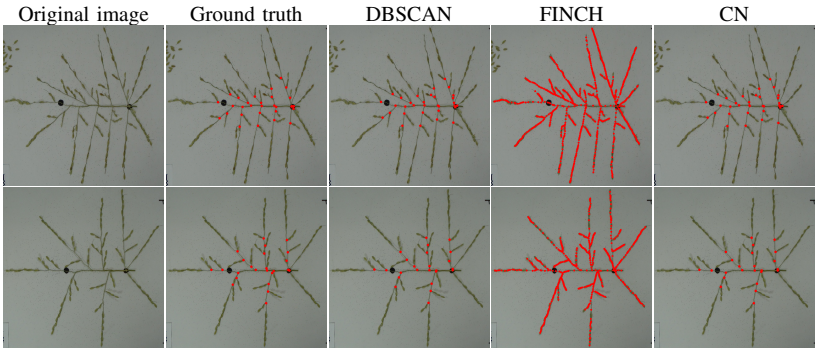


Fig. 3. Qualitative junction detection results.

The results across the three datasets show similar performances among the participating models. For Asian rice panicles, U-Net achieves the highest average F_1 score of 89.85 %, while SegNet records the lowest at 85.26 %. Precision and recall are relatively balanced for most models, with U^{2}CN and DCrack exhibiting slightly better precision. For African rice panicles, U-Net again achieves the highest F_1 score of 89.42 %, whereas SegNet scores 84.84 %. SegNet shows a high recall of 94.72 % but a lower precision at 77.18 %, indicating possible over-segmentation. In the combined dataset, U-Net demonstrates the best performance with the average accuracy of 89.91 %, while SegNet records at 85.51 %, despite having the highest recall of 95.41 %. Overall, U-Net, U^{2}CN and DCrack emerge as the most consistent performers. U-Net's highest F_1 score indicates robust performance and good generalizability, while DCrack and U^{2}CN excel in precision. Although SegNet has high recall, its lower precision suggests it may mis-identify too many background pixels. For balanced performance, U-Net is the best choice, but for scenarios prioritizing high precision, DCrack and U^{2}CN are reliable options.

B. Accuracy of junction detection in rice panicles

We conduct a comprehensive evaluation of junction detection accuracy using different combinations of skeletonization and clustering algorithms. The parameters choices for the methods that require them are listed below.

- GBO: $\beta = 0.33$, $\tau = 1.0$, simple_point_detection = Boolean, num_iter = 5.
- DBSCAN: eps = 1.5, minPts = 4.

Fig. 3 presents the qualitative junction detection results of each clustering algorithm applied to the ZS skeleton, where red dots represent the predicted junctions. DBSCAN and CN accurately identify junctions where branches intersect, while FINCH clusters data points effectively but overestimates the number of junctions, resulting in many false predictions. This discrepancy arises because FINCH's hierarchical agglomerative approach merges nearby points based on first neighbor relationships rather than precisely identifying junctions. The absence of strict parameterization, like DBSCAN's eps, leads FINCH to form chains of points that often include non-junctions. Thus, while FINCH excels in general clustering tasks, its traits make it less suitable for junction detection.

Along with evaluating the three specified datasets, we measure accuracy at three junction levels: (1) main axis junctions, (2) high-order junctions (i.e., secondary junctions and beyond), and (3) all junctions. This approach ensures a thorough assessment by capturing performance metrics across various junction types and algorithm combinations. The results, computed on top of the U-Net segmentation masks, exhibit similar patterns across datasets, with slight variations in the performance of the different combinations of algorithms.

For Asian rice panicles, as recorded in Table II, performance varies across junction levels. The GBO-CN combination achieves the highest F_1 score (82.56 %) and precision (81.73 %) for main axis junctions. ZS-CN also performs well, with an F_1 score of 82.02 % and a recall of 84.04 %. FINCH exhibits the highest recall at 96.99 % but its lower precision and F_1 score limit its effectiveness for junction detection. While DBSCAN shows intuitive results, its neighbor-based averaging may affect accuracy. For high-order junctions, CN with both ZS and GBO shows low precision due to the complexity of Asian rice overlapping branches. For all junctions, ZS-based combinations generally offer more balanced results, with ZS-CN achieving the highest F_1 score at 62.60 %. Overall, while detecting main axis junctions is feasible, accuracy drops notably for higher-order junctions, with CN combined with ZS skeletonization offering the most balanced performance.

For African rice panicles, as shown in Table III, performance varies across algorithm combinations but remains consistent across junction levels due to the simpler structure. The GBO-CN combination performs well for main axis junctions, with an F_1 score of 81.44 %, precision of 87.81 % and recall of 78 %. ZS-CN shows stronger results for high-order junctions, achieving the highest precision (69.97 %) and F_1 score (72.54 %). FINCH maintains the highest recall across all levels, similar to its performance with Asian rice. While DBSCAN is effective, it has lower accuracy than CN. For all junctions, ZS-CN leads with an F_1 score of 75.31 %, highlighting its effectivenes. The simpler architecture of African rice panicles results in less performance variability compared to Asian rice, highlighting ZS-CN's consistent superiority across all junction levels.

For the combined case of Asian and African rice panicles, as shown in Table IV, performance patterns differ from the individual cases. ZS-CN consistently outperforms others, achiev-

TABLE II
JUNCTION DETECTION PERFORMANCE (%) ON ASP.

Thinning	ZS			GBO		
Clustering	DBSCAN	FINCH	CN	DBSCAN	FINCH	CN
Main axis junctions						
Pr	77.63	27.64	81.66	78.41	34.41	**81.73**
Rc	72.33	90.23	84.04	77.45	**96.99**	85.04
F_1	73.94	41.71	82.02	77.16	49.96	**82.56**
High-order junctions						
Pr	48.19	7.46	**49.78**	44.97	7.46	45.55
Rc	60.36	**91.06**	61.64	61.97	86.40	62.51
F_1	53.18	13.48	**54.63**	51.64	13.64	52.23
All junctions						
Pr	55.67	9.46	**57.69**	52.61	10.64	53.75
Rc	66.04	**94.66**	69.13	67.97	93.42	70.21
F_1	60.09	16.91	**62.60**	59.01	18.78	60.59

TABLE III
JUNCTION DETECTION PERFORMANCE (%) ON AFP.

Thinning	ZS			GBO		
Clustering	DBSCAN	FINCH	CN	DBSCAN	FINCH	CN
Main axis junctions						
Pr	83.73	31.18	**88.00**	84.39	38.23	87.81
Rc	59.25	83.88	73.40	67.54	**96.20**	78.00
F_1	68.17	44.43	78.61	73.79	54.20	**81.44**
High-order junctions						
Pr	69.69	5.92	**69.97**	60.20	6.36	61.19
Rc	75.55	**96.11**	75.79	75.57	95.09	76.78
F_1	72.25	11.10	**72.54**	66.68	11.88	67.79
All junctions						
Pr	73.73	7.99	**75.33**	66.51	9.15	68.50
Rc	70.72	97.84	75.65	73.30	**98.30**	77.69
F_1	72.04	14.68	**75.31**	69.54	16.67	72.59

TABLE IV
JUNCTION DETECTION PERFORMANCE (%) ON ASAFP.

Thinning	ZS			GBO		
Clustering	DBSCAN	FINCH	CN	DBSCAN	FINCH	CN
Main axis junctions						
Pr	76.38	22.11	**79.95**	75.23	26.94	78.04
Rc	65.61	87.14	80.36	72.91	**94.66**	82.14
F_1	69.27	34.62	**78.85**	72.75	41.23	78.63
High-order junctions						
Pr	55.50	6.53	**56.51**	50.20	7.13	50.62
Rc	65.73	**97.65**	66.78	66.26	95.48	66.93
F_1	59.58	12.18	**60.61**	56.58	13.19	57.11
All junctions						
Pr	62.26	8.62	**64.43**	57.88	10.02	59.24
Rc	68.04	**99.23**	72.99	70.75	98.98	74.29
F_1	64.49	15.78	**67.91**	63.21	18.11	65.45

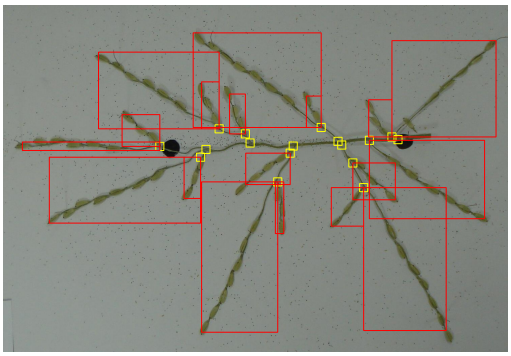


Fig. 4. Bounding boxes for rice panicle grains and junctions.

ing the highest F_1 scores for main axis (78.85 %), high-order (60.61 %), and all junctions (67.91 %). This success is due to its ability to handle the diverse structures in the mixed dataset. GBO-based combinations generally lag behind, with GBO-CN showing lower but stable performances. The combined dataset highlights the strengths of ZS-based methods, especially ZS-CN, while GBO-based methods struggle with consistency.

In conclusion, the ZS-CN combination delivered the best results across all junction levels and species, showing adaptability and accuracy. Its robust performance suggests it can serve as a general approach for various rice panicle architectures, though fine-tuning may be needed for unique species.

C. Comparison with P-TRAP

Obtained experimental results suggest an optimal pipeline for the structural analysis of rice panicles, using U-Net for segmentation, ZS for skeletonization, and CN for junction detection. This fully automated approach achieves high accuracy across Asian and African rice without human interaction. In contrast to P-TRAP, an effective tool developed by our co-authors, manual intervention is less required in this approach. P-TRAP employs ZS for skeletonization and converts the skeleton to a mathematical graph. Subsequently, it involves vertex classification, where users must identify generating vertices via P-TRAP's GUI, and the software classifies other vertices accordingly. Following vertex classification, the software proceeds with graph quantification, which involves quantifying rice panicle traits. Despite the manual initiation of vertices, potential post-manual correction may be required, which can introduce errors in the structure detection and classification process. Overall, these potential errors highlight the benefits of transitioning to our fully automated approach.

V. DISCUSSION

In this study, we compared the performance of crack segmentation models and combinations of thinning and clustering algorithms. While we achieved satisfactory results, some limitations remain. U-Net performed well, with nearly 90 % accuracy across all datasets, but occasionally misidentified critical junction pixels, leading to fragmented segmentation. This highlights the difficulty of analyzing complex structures like rice panicle junctions. Our approach consistently detected main axis junctions with around 80 % accuracy across datasets, but high-order junctions were harder to detect in Asian panicles due to their intricate branching. In contrast, the simpler African panicle architecture allowed more reliable detection. While our method shows promise for main axis junctions, further refinements are needed for high-order junction detection, especially in complex structures like Asian rice panicles.

Looking ahead, future developments in this field are promising. One avenue for improvement involves designing a more robust method with fewer steps. By generating bounding boxes around junctions and grains, the proposed rice panicle morphology analysis becomes a combined segmentation and detection problem, where advanced detection models are expected to further improve accuracy. Fig. 4 illustrates this vision, with red bounding boxes capturing grains and yellow bounding boxes capturing junctions. This approach enables

further detection and accurate counting of single spikelets, improving our understanding of rice panicle structure and supporting rice breeding and yield optimization efforts.

VI. Conclusion

This paper introduces a cross-domain approach for automated detection of junctions in Asian and African rice panicles, using transfer learning with 7 crack segmentation models, leveraging visual similarities between pavement cracks and rice panicles. U-Net performs best, achieving nearly 90 % accuracy across datasets. The Zhang-Suen thinning and Crossing Number (ZS-CN) combination reliably detects main axis junctions with about 80 % accuracy and offers insights into high-order junction detection. This study provides a high-throughput phenotyping tool for analyzing rice panicle structures, aiding researchers in improving crop morphology. Future efforts will aim to refine high-order junction detection, extend the method to include grain detection, and agricultural advancements.

Acknowledgement

This work is the output of the ASEAN IVO http://www.nict.go.jp/en/asean$_ivo/index.html$ project "Agricultural IoT based on Edge Computing" and financially supported by NICT http://www.nict.go.jp/en/index.html.

References

[1] F. Al-Tam, H. Adam, A. d. Anjos, *et al.*, "P-trap: A panicle trait phenotyping tool," *BMC Plant Biol.*, vol. 13, pp. 1–14, 2013.

[2] S. Crowell, A. X. Falcão, A. Shah, Z. Wilson, A. J. Greenberg, and S. R. McCouch, "High-resolution inflorescence phenotyping using a novel image-analysis pipeline, panorama," *Plant Physiol.*, vol. 165, no. 2, pp. 479–495, 2014.

[3] A. K. Parida, S. Sekhar, B. B. Panda, G. Sahu, and B. P. Shaw, "Effect of panicle morphology on grain filling and rice yield: Genetic control and molecular regulation," *Front. Genet.*, vol. 13, p. 876 198, 2022.

[4] W. Tanaka, T. Yamauchi, and K. Tsuda, "Genetic basis controlling rice plant architecture and its modification for breeding," *Breed. Sci.*, vol. 73, no. 1, pp. 3–45, 2023.

[5] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, "Deepcrack: Learning hierarchical convolutional features for crack detection," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1498–1512, 2018.

[6] C. V. Dung *et al.*, "Autonomous concrete crack detection using deep fully convolutional neural network," *Autom. Constr.*, vol. 99, pp. 52–58, 2019.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[8] P. Shi, F. Zhu, Y. Xin, and S. Shao, "U2cracknet: A deeper architecture with two-level nested u-structure for pavement crack detection," *Struct. Health Monit.*, vol. 22, no. 4, pp. 2910–2921, 2023.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Med. Image Comput. Computer-assisted Intervention–MICCAI 2015: 18th Int. Conf., Munich, Germany, October 5-9, 2015, Proc., part III 18*, Springer, 2015, pp. 234–241.

[10] K. Li, B. Wang, Y. Tian, and Z. Qi, "Fast and accurate road crack detection based on adaptive cost-sensitive loss function," *IEEE Trans. Cybern.*, vol. 53, no. 2, pp. 1051–1062, 2021.

[11] G. Yu, J. Dong, Y. Wang, and X. Zhou, "Ruc-net: A residual-unet-based convolutional neural network for pixel-level pavement crack segmentation," *Sensors*, vol. 23, no. 1, p. 53, 2022.

[12] T. H. Dinh, C. H. Le, and Q. Ha, "UAV imaging: Correlation between contrast and F1-score for vision-based crack detection," in *2024 IEEE/SICE Int. Symp. Syst. Integr. (SII)*, IEEE, 2024, pp. 657–662.

[13] J. Wang, Z. Zeng, P. K. Sharma, *et al.*, "Dual-path network combining CNN and transformer for pavement crack segmentation," *Autom. Constr.*, vol. 158, p. 105 217, 2024.

[14] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3434–3445, 2016.

[15] T. H. Dinh, V. T. T. Anh, T. Nguyen, *et al.*, "Toward vision-based concrete crack detection: Automatic simulation of real-world cracks," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023.

[16] T. Y. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," *Commun. ACM*, vol. 27, no. 3, pp. 236–239, 1984.

[17] M. J. Menten, J. C. Paetzold, V. A. Zimmer, *et al.*, "A skeletonization algorithm for gradient-based optimization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 21 394–21 403.

[18] C.-H. Yao, W.-C. Hung, Y. Li, M. Rubinstein, M.-H. Yang, and V. Jampani, "Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 4853–4862.

[19] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, 1996, pp. 226–231.

[20] S. Sarfraz, V. Sharma, and R. Stiefelhagen, "Efficient parameter-free clustering using first neighbor relations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8934–8943.

[21] R. Bansal, P. Sehgal, and P. Bedi, "Minutiae extraction from fingerprint images-a review," *Int. J. Comput. Sci. Issues (IJCSI)*, vol. 8, no. 5, p. 74, 2011.

New approach for Alzheimer's disease classification using topographic maps and deep learning model

Quoc Anh Le* and Hong Thinh Nguyen†
* AVITECH, VNU University of Engineering and Technology, Vietnam
E-mail: lqanh@vnu.edu.vn
† VNU University of Engineering and Technology, Vietnam
E-mail:hongthinh.nguyen@vnu.edu.vn

***Abstract*—With an aging global population, age-related illnesses like Alzheimer's disease (AD) are a growing research priority. Electroencephalogram (EEG) offers a low-cost and safe way to collect brain activity for AD detection. However, analyzing EEG is labor-intensive, and experts' visual inspection is subjective and prone to errors, especially with lengthy recordings containing signal artifacts. This study proposes a novel method for diagnosing AD, Frontotemporal Dementia (FTD), and healthy control (HC) subjects using brainwave patterns from EEG. Our approach involves analyzing the different frequencies within the EEG signal. We then create topographic maps, which are special images that capture the strength of these brainwaves across different brain areas. Finally, we apply a deep learning method to these images to distinguish between AD, FTD, and HC subjects. The results demonstrate an improvement in classification performance compared to traditional hand-crafted feature methods, achieving accuracy of 83.0%, 85.5%, and 87.3% for AD/FTD, AD/HC, and FTD/HC, respectively. Our approach also achieves results comparable to state-of-the-art methods on the same dataset, highlighting its potential as an effective tool for the diagnosis of AD and FTD.**

I. Introduction

With an aging global population, age-related illnesses like Alzheimer's disease (AD) are a growing research priority. While various clinical methods to diagnose Alzheimer's, such as positron emission tomography (PET) or magnetic resonance imaging (MRI), are expensive, EEG offers a low-cost and safe way to collect brain activity for Alzheimer's Disease detection [1]. However, analyzing EEG is labor-intensive, and experts' visual inspection is subjective and prone to errors, especially with lengthy recordings containing signal artifacts. Consequently, extracting features from EEG signals and applying supervised learning classifications for recognizing AD patients has been commonly used in the last decade [2]. These automated methods aim to support clinicians in the demanding task of early AD diagnosis.

Raw EEG signals are generally collected based on placing electrodes on the scalp, typically to a 10–20 standard. Prepossessing procedures then clean the raw signals from artifacts and apply band-pass filtering to remove noise. Features are extracted from the cleaned EEG to prepare the input for the machine learning model. This work involves analyzing the information within the signal, either in terms of time domain, frequency domain, or a combination of both. In the time domain, Tzimourta et al. successfully used mean, variance and inter-quartile range (IQR) features to classify AD and HC [3].

On the other hand, Fourier transform (FT) is a commonly used technique to extract frequency domain features for AD detection. The cleaned EEG signal shows different brain wave patterns and is classified according to their frequency. There are five main bands: delta (1-4 Hz, low frequency and high amplitude, often detected when people fall into deep sleep), theta (4-8 Hz, associated with drowsiness), alpha (8-13 Hz, usually appears in adults, when people are comfortable and close their eyes), beta (13-30 Hz, low amplitude but relatively high frequency, often appearing when people are concentrating or highly worried) and gamma (30-40 Hz, the highest frequency and lowest amplitude, a type of wave closely related to high-level human tasks such as cognitive functions). Based on FT analysis, several frequency-related features are applied for each frequency band, for example, spectral power, band power, entropy [4], [5].

Besides, some studies proposed using the Wavelet analysis method to exploit information in the frequency domain and showed much better results [4], [6]. Fiscon et al. [7] proposed a time-frequency analysis by applying both the Fourier and Wavelet to exploit feature extraction and to classify AD, Mild Cognitive Impairment (MCI), and healthy control (HC) samples. Recently, some projects have applied deep learning models on appropriate sets of EEG signal features to classify AD EEG patterns [8]–[10]. Fouladi et al. (2022) [11] proposed the use of two different deep learning (DL) architectures, including (CNN) and convolutional autoencoder (Conv-AE) neural networks, for classifying subjects into AD, MCI, and HC. Ieracitano et al. (2020) proposed a deep learning model that directly uses raw EEG signals as input instead of features extracted from EEG [9].

Although these methods yield promising results, they overlook the spatial information of the EEG signal from different brain locations. Typically, EEG features from all electrodes are extracted and concatenated into a single vector. This vector is then used for classification with supervised machine-learning algorithms or deep-learning models. Consequently, the spatial relationships and interactions between signals at various locations are not taken into account. In this paper, we propose a new method, exploiting spatial information about signal energy at different frequencies as well as the relationship between them.

II. Materials and Methods

A. Methods

1) Relative band power (RBP): In EEG analysis, RBP reveals energy distribution across different frequency bands within the signal, is calculated as the ratio between the power within a specific frequency band and the total power of the entire signal. This feature helps us understand the dominant frequencies and their significance in brain activity. To calculate RBP, we use the equation:

$$RBP_{band} = \frac{\sum_{f_{band}} P(f)}{\sum_{f=f_1}^{f_2} P(f)} \quad (1)$$

where $P(f)$ represents the Power Spectrum Density (PSD) of the EEG signal, f_1 and f_2 represent the minimum and maximum frequencies of the EEG signal, respectively, in the band. Research on general EEG analysis and Alzheimer's Disease (AD) specifically highlights RBP as a valuable feature [12]. The RBP features are often used in conjunction with machine learning algorithms to classify AD.

2) Topographic maps: Brain mapping is a technique that creates visual representations of brain activity. These are called "topographic maps". These maps reveal more than just the raw EEG data. They can show brainwaves' strength (voltage) or derived measures like energy, power, or peak latency [13]. Software libraries like EEGLab (MATLAB) and MNE (Python) can generate EEG topographic maps. These visualizations depict the distribution of brain activity across the scalp based on a standardized electrode placement system (10-20 system). The software interpolates between electrode measurements, considering the signal's magnitude, energy, power, or power distribution to create a continuous map. This map is then displayed as a heat map, where color intensity reflects the activity level in different brain regions.

3) Convolution neural networks: Deep learning or Convolution Neural Network (CNNs) is a powerful supervised machine learning for data processing; used to uncover hidden patterns within massive datasets. Its ability to extract meaningful insights from complex data makes it particularly effective for tackling image-related tasks or computer vision. This is due to its use of convolution layers, which are specifically designed to analyze and exploit the spatial relationships between pixels in images. In this article, the topographic maps for each type of brainwave are input into Convolution Neural Networks. CNNs are used to automatically extract relevant features from the topographic maps, capturing the essential characteristics of the brainwave patterns, especially the spatial information between brain waves at different locations in the brain.

4) Pipeline: The pipeline of the proposed method is illustrated in Figure 1. Initially, EEG signals are collected as time series data from scalp electrodes. These signals are segmented into time slices (epoch), and we calculate the relative band power (RBP) for each frequency band and each EEG channel. Since the RBP feature is related to the power density, we use RBP features to plot topographic maps. The topographic maps represent the power density for Alpha, Beta, Delta, Gamma, and Theta brainwaves at each location in the brain. Each map is then used as input for Convolutional Neural Networks (CNNs) to extract key features for classification. The features are then concatenated and undergo Global Average Pooling to reduce dimensionality. Finally, the aggregated features are fed into a neural network for classification.

In this study, we use three CNN backbones to evaluate the efficiency of feature extraction and provide optimal classification results: Vanilla, ResNet34, and EfficientNetV2.

- *Vanilla* is a simple yet effective CNN backbone for feature extraction from input images [14]. It comprises three convolutional layers, ReLU activation functions, max-pooling, and dropout layers. ReLU activations introduce non-linearity after each convolution, enhancing the model's ability to learn complex patterns. Max pooling layers reduce the feature map dimensions and increase translation invariance, while dropout layers with a probability of 0.5 improve generalization by randomly dropping some units to reduce overfitting.
- *ResNet34* is based on residual learning. This architecture addresses the problem of vanishing gradients in deep networks by introducing shortcut connections that skip one or more layers, maintaining gradient flow and enabling the training of much deeper networks [15]. ResNet34 comprises 34 layers, making it highly effective for extracting complex image features. Additionally, batch normalization and ReLU activations are employed after each convolutional layer to stabilize and accelerate training. ResNet34 has demonstrated excellent performance in various image recognition benchmarks.
- *EfficientNetV2* is part of the EfficientNet family, aiming to achieve higher accuracy with fewer parameters [16]. EfficientNet models are designed using a compound scaling method that uniformly scales the network's depth, width, and resolution. EfficientNetV2 offers a good trade-off between accuracy and efficiency, making it suitable for many applications, especially those requiring deployment on resource-constrained devices.

B. Data description

We use the OpenNeuro ds004504 dataset, recorded at the 2nd Department of Neurology of AHEPA General Hospital, Thessaloniki [12]. This EEG dataset is recorded from 88 subjects, including people with Alzheimer's disease (AD) (36 people), people with frontotemporal dementia (FTD) (23 people), as well as healthy control subjects (HC) (29 people). Each subject's cognitive and neuropsychological status was assessed using the International Mini-Mental State Examination (MMSE). The MMSE scores range from 0 to 30, with lower scores indicating more severe cognitive decline. Table I shows statistical information about participants.

The EEG recordings captured brain waves using electrodes placed at 19 specific locations on the scalp (including Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4,

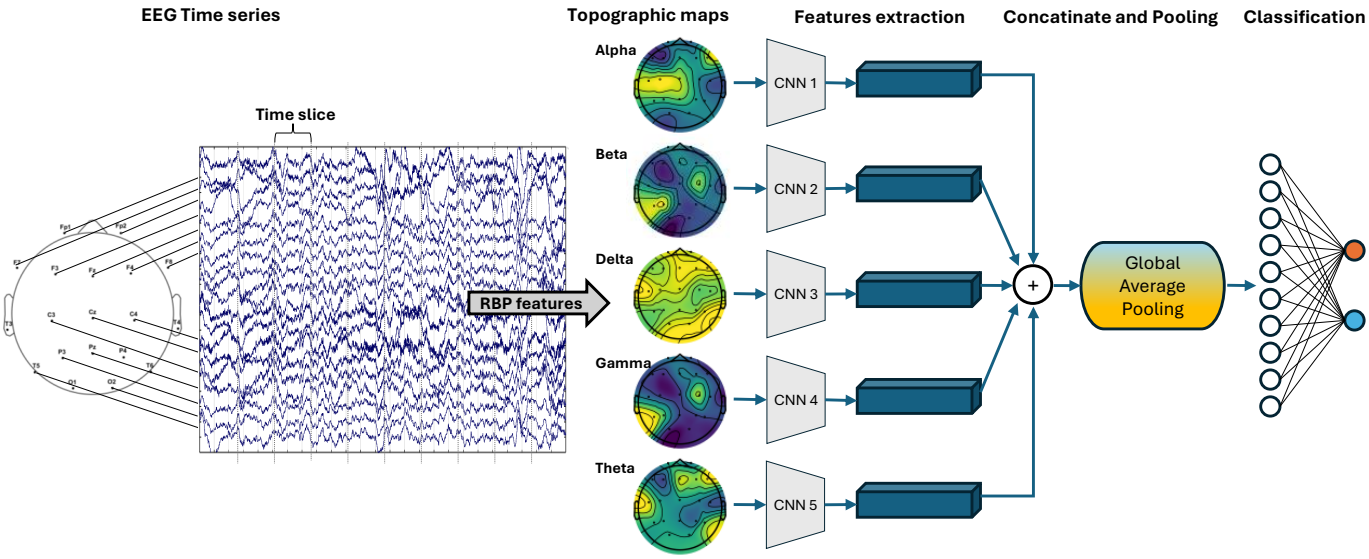


Fig. 1. The proposed pipeline for EEG signal analysis and classification. Five separate CNN models are used to extract features from topographic maps.

TABLE I
PARTICIPANT DESCRIPTION

	AD	FTD	HC
Number of subject	36	23	29
Ages	66.4(49-79)	63.6(44-78)	67.9(57-78)
Gender(M:F)	(12:24)	(14:9)	(18:11)
MMSE score	17.75	22.17	30

T6, O1, and O2). All recordings were conducted with closed eyes in a sleeping state. The recording montage was referential, using Cz for common mode rejection and then sampling at 500 Hz with a 10 uV/mm resolution.

III. EXPERIMENTS AND EVALUATION

A. Data preprocessing

1) Preprocessing: Before running feature extraction algorithms, the raw data underwent a multi-step process to prepare it for analysis. The pre-processing steps involved:

- Filtering: A Butterworth band-pass filter was applied to remove unwanted frequencies outside the range of 0.5-45 Hz. This frequency range focuses the analysis on the relevant brain activity.
- Re-referencing: The recordings were re-referenced to the A1-A2 electrodes, which helps to remove common background noise.
- Artifact removal: Artifact Subspace Reconstruction (ASR) was employed to identify and remove eliminate persistent or large-amplitude artifacts, which then removed bad data periods [12].
- Independent Component Analysis (ICA): The data was decomposed into 19 independent components using ICA. This method helps to separate brain signals from other sources of noise.
- Automatic artifact rejection: The researchers implemented automatic detection and removal of eye and jaw movement artifacts.

2) RBP feature extraction: We segment each channel's signal into fixed-length time slices called *epochs* to analyze the EEG data in detail. Each epoch captures a specific time window of the recording. Consecutive epochs overlap slightly, ensuring no data is lost during the analysis. This overlap allows for continuous monitoring of events and brain reactions within the recording.

The ideal epoch length depends on the analysis goals. For instance, some studies recommend a suggestion of 5 seconds [9], while in [12] the authors used 4 seconds long for each epoch. Considering these factors, we opted for a 4-second epoch length in this experiment. Since we use five frequency bands (delta, theta, alpha, beta, and gamma), so for each epoch of EEG, we get five values of RBP. A higher value of RBP indicates a greater concentration of energy. Combining this information with electrode locations allows us to create topographic maps – a heat map or energy distribution maps – for each frequency band.

Figure 2 showcases the results of using RBP-based topographic maps from subjects with Alzheimer's Disease (AD), Frontotemporal Dementia (FTD), and Healthy Controls (HC). The differences in energy distribution between these groups suggest the potential of RBP topographic maps in distinguishing neurological conditions.

B. Implementation detail

We used the MNE software (Python) to read EEG signals and create topographic maps. We extracted RBP features based on the power spectrum density calculated using the Welch method from the *scipy* [1] library.

[1]https://scipy.org/

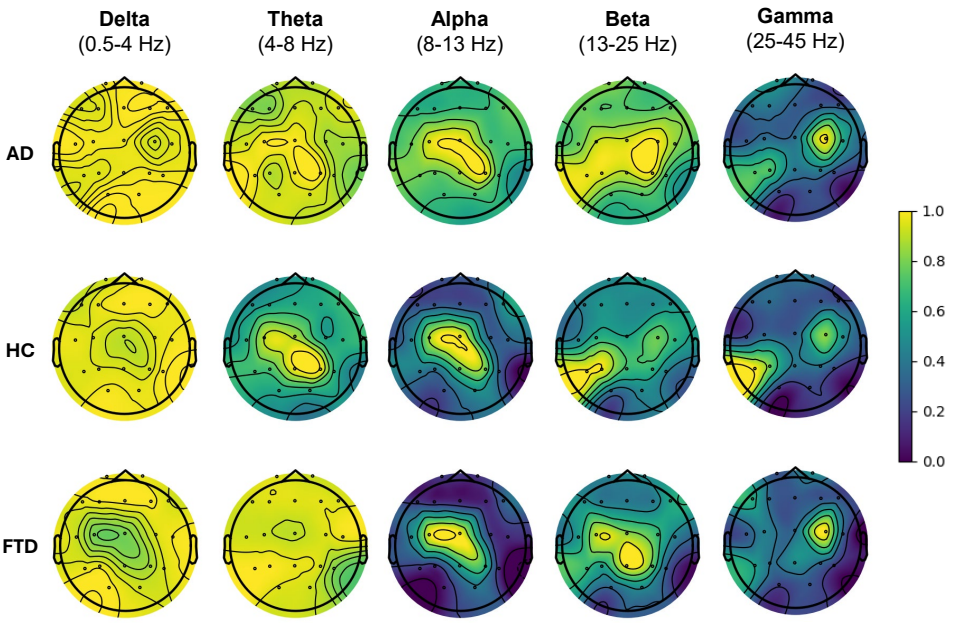


Fig. 2. RBP values were used to create topographic maps, which illustrate power distribution across the scalp in the AD, FTD, and HC groups.

CNN models are implemented using PyTorch 1.12. The ResNet34 and EfficientNetV2 backbones were constructed using the $timm$ [2] library. We trained the deep learning models using cross-entropy loss with the Adam optimizer, with a learning rate set of 0.001 and a batch size of 24. The CNN models were trained for 50 epochs using pre-trained weights from the ImageNet dataset.

We conducted experiments using Multi-layer Perceptron (MLP) and RandomForest (RF) classifiers with the RBP features to compare with deep learning methods. The RBP features were estimated for each epoch data, divided into 5 bands corresponding to 5 brain waves. Each channel had 5 RBP features, resulting in a 95-element vector for each epoch data used for classification. These classifiers were implemented using the $sklearn$ [3] library.

We employed 5-fold cross-validation for all experiments to ensure that the choice of a particular training and testing split did not bias our evaluation. The dataset is divided into five equal parts. Each part is used as a test set once, while the remaining four parts are used for training. This process is repeated five times, and the average performance across five iterations is reported.

C. Metrics

This paper uses evaluation criteria such as Precision, Recall, F1-score, Accuracy and AUC score. Precision measures the accuracy of positive predictions (true positives (TP) divided by all positive classifications)

$$Precision = \frac{TP}{TP + FP}. \tag{2}$$

Recall, also known as sensitivity, reflects the model's ability to identify all relevant instances (true positives (TP) divided by all positive cases).

$$Recall = \frac{TP}{TP + FN}. \tag{3}$$

F1-score combines these for a single performance metric. An ideal classifier achieves a perfect score of 1 for Precision and Recall.

$$F1\text{-}score = \frac{2 \times Recall \times Precision}{Recall + Precision}. \tag{4}$$

We also assess the performance of the proposed classification method using Accuracy and AUC score. Accuracy measures the ratio of correctly predicted instances to the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FT + FN}. \tag{5}$$

The AUC score represents the area under the ROC curve and evaluates the trade-off between the True Positive Rate and the False Positive Rate. AUC score compares two or more models as well as evaluates the same model's performance across different thresholds.

D. Evaluation results

Table II displays the classification performance metrics for various models using two types of inputs: RBP features (hand-crafted features with traditional machine learning) and Topographic maps (CNN-based). The Random Forest (RF) and MLP models, which use RBP features, exhibit comparable performance in hand-crafted methods. Both models show similar Precision and Recall rates, with RF generally performing slightly better. For CNN-based approaches, ResNet34 with topographic map inputs consistently achieves

[2]https://github.com/huggingface/pytorch-image-models
[3]https://scikit-learn.org/

TABLE II
COMPARISON OF CLASSIFICATION PERFORMANCE FOR CNN-BASED METHODS AND HAND-CRAFTED FEATURE-BASED METHODS

Input type	Model	Accuracy (%)	Percision (%)	Recall (%)	F1-score (%)	AUC score (%)
RBP features	**MLP**	79.5 / 82.1 / 82.1	80.2 / 83.3 / 73.8	90.1 / 84.1 / **87.1**	84.8 / 83.7 / 79.9	86.5 / 91.3 / 90.7
	RF	80.1 / 80.9 / 82.4	78.0 / 79.7 / 80.6	**95.7** / 87.3 / 74.8	86.0 / 83.4 / 77.6	87.0 / 88.3 / 91.5
Topographic maps	**Vanilla**	77.0 / 75.9 / 74.8	77.1 / 79.1 / 72.0	91.0 / 76.1 / 62.4	83.5 / 77.6 / 66.9	79.5 / 83.6 / 79.2
	ResNet34	**83.0** / **85.5** / **87.3**	**82.5** / **85.7** / **85.4**	93.1 / 88.2 / 83.1	**87.5** / **86.9** / **84.2**	**89.1** / **92.0** / **94.1**
	EfficientNetV2	78.2 / 81.0 / 80.8	78.3 / 77.9 / 77.6	91.0 / **91.3** / 74.3	84.2 / 84.0 / 75.9	83.7 / 89.9 / 88.6

The slash is used to separate classification results for three groups: AD/FTD, AD/HC, and FTD/HC.

TABLE III
COMPARE THE CLASSIFICATION ACCURACY OF THE PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON THE DS004504 DATASET

Study	Methodology	Accuracy (%)
Miltiadous et al. (2023a) [12]	RBP features, Random Forests, LOSO validation	- - / 77.0 / 72.0
Miltiadous et al. (2023b) [17]	RBP and SCC features, DICE-net, LOSO validation	- - / 83.3 / 75.0
Si et al. (2023) [18]	FMN features, Gaussian Naive Bayes, 5-fold cross-validation	81.1 / **86.2** / 84.7
This work	Topographic maps, ResNet34, 5-fold cross-validation	**83.0** / 85.5 / **87.3**

The slash is used to separate classification results for three groups: AD/FTD, AD/HC, and FTD/HC.

the highest performance across nearly all metrics and groups. The ResNet34 model improves Accuracy by over 3% for all three problems—AD/FTD (+3.5%), AD/HC (+3.4%), and FTD/HC (+4.9%)—compared to traditional machine learning methods. It also leads in Precision, F1-score, and AUC score across all groups. The only exception is Recall for the AD/FTD classification, where it performs slightly worse than the RF algorithm. Nevertheless, ResNet34 outperforms by 0.9% in AD/HC classification and by 8.3% in FTD/HC classification in terms of Recall. On the other hand, not all deep learning models perform equally well; the EfficientNetV2 model shows moderate performance, comparable to traditional machine learning methods, while the Vanilla model achieves the lowest performance among the models tested.

IV. DISCUSSION

In this paper, we proposed a method for classifying EEG signals using topographic maps. Table II indicate that the proposed CNN-based approach yields promising results. Specifically, ResNet34 with topographic maps input demonstrates the highest overall performance, achieving accuracy values of 83.0%, 85.5%, and 87.3% for AD/FTD, AD/HC, and FTD/HC, respectively. It also excels in precision, recall, F1 score, and AUC score across most groups, highlighting its robustness and reliability in classification tasks. Two models using RBP features show comparable performance. EfficientNetV2 results are similar to hand-crafted feature-based methods and show no significant improvement. The Vanilla model performs the worst, likely due to its shallow network (3 convolution layers), which limits its ability to extract complex features and results in lower classification performance. In contrast, ResNet34 performs relatively well, likely due to its residual learning mechanism that enables the model to be deeper and extract more complex features, thus improving classification effectiveness. EfficientNetV2, designed to reduce the number of parameters for hardware constraints, may trade off classification performance in favor of parameter efficiency. We also observe that AD and FTD are often misclassified as each other [18], explaining why the classification performance for these two diseases is generally lower than the AD/HC and FTD/HC classifications.

Several studies on the OpenNeuro ds004504 dataset have been published. Miltiadous et al. (2023a) evaluated classification performance using RBP features combined with Random Forests [12]. Leave-One-Subject-Out (LOSO) validation is performed and achieved accuracies of 77.0% and 72.0% for AD/HC and FTD/HC, respectively. Miltiadous et al. (2023b) further improved classification results by using more information for classification [17]. Besides RBP features, Spectral Coherence Connectivity (SCC) features were used to create two inputs for a CNN network called Dual-Input Convolutional Encoder Network (DICE-net). The results showed improvement using LOSO validation with accuracies of 83.3% and 75.0% for AD/HC and FTD/HC, respectively. However, LOSO is ideal for small datasets, but it is computationally expensive and can have high variance. Si et al. (2023) proposed a method using a frequency-based multilayer network (FMN) to extract features from EEG signals, which were fed into a Gaussian Naive Bayes classifier. Five-fold cross-validation was used to evaluate the method's performance. The results showed classification accuracies of 81.1%, 86.2%, and 84.7% for AD/FTD, AD/HC, and FTD/HC, respectively. According to Table III, the proposed method demonstrates comparable accuracy across all classification types compared to other state-of-the-art methods. The results show the effectiveness of using CNN-based topographic maps in enhancing classification performance for EEG signal analysis.

Our study has some limitations. First, our experiments were conducted on the OpenNeuro ds004504 dataset, which has a relatively limited number of subjects. We hypothesize that using a larger dataset for training would significantly improve the performance of the proposed method. Secondly, we utilized

available tools to generate topographic maps with predefined parameters, which may not be optimal for different disorders. Future studies should consider a more tailored approach to creating topographic maps for classification. Further research is needed to explore the impact of varying these parameters and develop more sophisticated techniques for generating topographic maps better suited to the specific characteristics of different neurological conditions.

V. Conclusion

In this study, we introduce a method for classifying EEG signals. Our approach utilizes feature extraction from topographic maps and harnesses the capabilities of CNN architectures. Unlike traditional methods that depend on hand-crafted features, our method uses RBP-topographic maps, which capture spatial information about EEG power density in various brain regions. The results demonstrate that using topographic maps markedly enhances classification performance compared to traditional hand-crafted feature extraction methods.

References

[1] J. Dauwels, F. Vialatte, and A. Cichocki, "Diagnosis of alzheimer's disease from eeg signals: Where are we standing?" *Current Alzheimer Research*, vol. 7, no. 6, pp. 487–505, 2010.

[2] C. Lehmann, T. Koenig, V. Jelic, *et al.*, "Application and comparison of classification algorithms for recognition of alzheimer's disease in electrical brain activity (eeg)," *Journal of neuroscience methods*, vol. 161, no. 2, pp. 342–350, 2007.

[3] K. D. Tzimourta, N. Giannakeas, A. T. Tzallas, *et al.*, "Eeg window length evaluation for the detection of alzheimer's disease over different brain regions," *Brain sciences*, vol. 9, no. 4, p. 81, 2019.

[4] P. Durongbhan, Y. Zhao, L. Chen, *et al.*, "A dementia classification framework using frequency and time-frequency features based on eeg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 5, pp. 826–835, 2019.

[5] K. AlSharabi, Y. B. Salamah, M. Aljalal, A. M. Abdurraqeeb, and F. A. Alturki, "Eeg-based clinical decision support system for alzheimer's disorders diagnosis using emd and deep learning techniques," *Frontiers in Human Neuroscience*, vol. 17, p. 1 190 203, 2023.

[6] N. Kulkarni and V. Bairagi, "Extracting salient features for eeg-based diagnosis of alzheimer's disease using support vector machine classifier," *IETE Journal of Research*, vol. 63, no. 1, pp. 11–22, 2017.

[7] G. Fiscon, E. Weitschek, A. Cialini, *et al.*, "Combining eeg signal processing with supervised methods for alzheimer's patients classification," *BMC medical informatics and decision making*, vol. 18, pp. 1–10, 2018.

[8] F. C. Morabito, M. Campolo, C. Ieracitano, *et al.*, "Deep convolutional neural networks for classification of mild cognitive impaired and alzheimer's disease patients from scalp eeg recordings," in *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*, IEEE, 2016, pp. 1–6.

[9] C. Ieracitano, N. Mammone, A. Hussain, and F. C. Morabito, "A convolutional neural network based self-learning approach for classifying neurodegenerative states from eeg signals in dementia," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–8.

[10] M. Amini, M. M. Pedram, A. Moradi, and M. Ouchani, "Diagnosis of alzheimer's disease by time-dependent power spectrum descriptors and convolutional neural network using eeg signal," *Computational and Mathematical Methods in Medicine*, vol. 2021, no. 1, p. 5 511 922, 2021.

[11] S. Fouladi, A. A. Safaei, N. Mammone, F. Ghaderi, and M. J. Ebadi, "Efficient deep neural networks for classification of alzheimer's disease and mild cognitive impairment from scalp eeg recordings," *Cognitive Computation*, vol. 14, no. 4, pp. 1247–1268, 2022.

[12] A. Miltiadous, K. D. Tzimourta, T. Afrantou, *et al.*, "A dataset of scalp eeg recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine eeg," *Data*, vol. 8, no. 6, p. 95, 2023.

[13] L. S. Hooi, H. Nisar, and Y. V. Voon, "Tracking of eeg activity using topographic maps," in *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, IEEE, 2015, pp. 287–291.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[17] A. Miltiadous, E. Gionanidis, K. D. Tzimourta, N. Giannakeas, and A. T. Tzallas, "Dice-net: A novel convolution-transformer architecture for alzheimer detection in eeg signals," *IEEE Access*, 2023.

[18] Y. Si, R. He, L. Jiang, *et al.*, "Differentiating between alzheimer's disease and frontotemporal dementia based on the resting-state multilayer eeg network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

M-IRRA: A Multilingual Model for Text-based Person Search

Hung-Phong Tran*, Thi-Hoai Phan*, Thuy-Binh Nguyen‡, Thi-Ngoc-Diep Do*
Hong-Quan Nguyen§, Thanh-Hai Tran*, Hien-Thanh Duong¶, Thi-Lan Le*
* SigM lab, School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam
† Faculty of Information Security, Academy of People Security, Hanoi, Vietnam
‡ Faculty of Electrical-Electronics Engineering, University of Transport and Communications, Hanoi, Vietnam
§ Faculty of Information Technology, Viet-Hung Industrial University, Hanoi, Vietnam
¶ Dept. Information Technology, Hanoi University of Mining and Geology, Vietnam
Corresponding author e-mail : lan.lethi1@hust.edu.vn

***Abstract*—Text-based person search (TBPS) is a crucial task in computer vision and natural language processing, with various applications such as security surveillance, social media analysis, and multimedia database management. Thanks to advancements in deep learning, TBPS has made significant progress in handling English queries. However, existing methods often struggle with multilingual queries due to the diverse linguistic structures, vocabularies, and cultural nuances. To overcome this challenge, in this paper, we introduce a multilingual model for text-based person search, named M-IRRA, built upon the IRRA (Cross-Modal Implicit Relation Reasoning and Aligning) model - proposed for text-based person search in English [1]. In the M-IRRA model, the sentence transformer Multilingual CLIP is utilized to extract the feature embedding from text. Then, a dense layer is added to match and align the features extracted from text and image. The proposed model can support queries in multiple languages, ensuring strong performance across various linguistic contexts. Extensive experiments are performed on 3000VnPersonSearch dataset for TBPS in Vietnamese and its derived version in English to show the effectiveness of the proposed framework on both monolingual and multilingual search scenarios. The results obtained for monolingual search in Vietnamese are 38.60%, 67.4% and 77.62% at rank 1, 5 and 10 respectively, showing a significant improvement over the existing models.**

I. Introduction

Text-based person search (TBPS) is a critical task at the intersection of computer vision and natural language processing, with wide-range applications in security, social media, and multimedia database management. It involves retrieving images based on a natural language description, where the input is a text query, and the output is a list of images of individuals that match the provided description. Due to the diversity in natural language, a major challenge in TBPS is learning to map visual and textual modalities into a shared latent space. As a result, most existing research has concentrated on achieving visual-textual alignment. With the substantial advancements in deep learning models, significant progress has been made in TBPS. However, most current TBPS systems are optimized primarily for a single language, usually English, creating a notable gap in their ability to accommodate diverse linguistic communities.

Recently, significant efforts have been made to handle TBPS in lower-resource languages, such as Vietnamese [2], [3]. To tackle Vietnamese TBPS, a Vietnamese dataset called 3000Vn-PersonSearch was built and introduced in [2]. Although the existing models in the literature perform well in English, they yield poor results with Vietnamese queries. Furthermore, in an increasingly interconnected world, the ability to build a TBPS system across language barriers is becoming crucial. To handle this issue, it is essential to develop TBPS systems with multilingual capabilities to effectively serve a broader and more diverse population.

Inspired by the work of [1] and the potential of sentence transformer multilingual CLIP [4] in cross-lingual transfer technique, we propose a multilingual text-based person search named M-IRRA. M-IRRA is built upon the IRRA (Cross-Modal Implicit Relation Reasoning and Aligning) model proposed for TBPS in English in which the original CLIP-based text encoder is replaced by a sentence transformer Multilingual CLIP, enabling efficient cross-lingual understanding and processing. By leveraging the strengths of both token-level and holistic image-text relationships, the proposed framework overcomes the drawbacks of existing TBPS methods. Moreover, when replacing the CLIP encoder by multilingual CLIP, the visual and textual descriptions have different dimensions. Therefore, a dense layer is adopted to align these descriptions. Extensive experiments have been conducted to evaluate the proposed M-IRRA model on two datasets including: 3000VnPersonSearch and a new version of 3000VnPerson-Search named 3000Vn-V2E in which Vietnamese description sentences are translated into English. The experimental results demonstrate that M-IRRA, an CLIP-based TBPS framework significantly outperforms the state-of-the-art methods while maintaining robust performance across languages.

The remainder of this paper is organized as follows: Section II provides a brief review of related works. Section III introduces the M-IRRA model structure. Section IV presents detailed experiments and our analysis on those experiments. Finally, Section V concludes the paper.

II. Related Work

Text-to-image Person Retrieval. The field of Text-Based Person Search (TBPS) has evolved significantly since Li et al. [5] introduced the CUHK-PEDES dataset. Early research, such as Wang et al. [6] and Chen et al. [7], primarily relied on CNN and RNN networks for feature extraction. As the field progressed, focus shifted from global to local feature analysis, aiming for more precise image-text alignment. However, challenges in effectively leveraging these local features persisted. More recently, transformer-based approaches [8] have demonstrated improved contextual understanding compared to traditional RNNs. Additionally, CLIP-based studies by Li et al. [9] and Gao et al. [8] have leveraged large-scale datasets to enhance search outcomes. Bai et al. [10] and Jing et al. [1] both aimed to learn multi-granularity alignment between text and image, with the former developing correlation-focused and context-sensitive training objectives, and the latter introducing implicit relation reasoning. Despite these advancements, most works have focused solely on English datasets, limiting model robustness across regions. M-IRRA addresses this gap by pioneering the development of multilingual and robust solutions for TBPS.

Multilingual Visual-Language Pretraining. While visual-language pretraining (VLP) models have made significant strides [11], multilingual capabilities remain less explored. Recent approaches include creating multilingual image-text pairs through translation and using them as input for Transformer-based encoders, as seen in M3P (Learning Universal Representations via Multitask Multilingual Multimodal Pre-training) [12] and UC2 (Universal Cross-lingual Cross-modal Vision-and-Language Pre-training) [13]. Another approach involves knowledge distillation, exemplified by mCLIP (Multilingual clip via cross-lingual transfer) [14] and certain Sentence Transformers [4] models which utilize CLIP's image encoder to train multilingual text encoders. MURAL (Mul- timodal, multitask representations across languages) [15] and SigLIP (Sigmoid loss for language image pre-training) [16] alternatively employ from-scratch training with augmented multilingual image-text pairs and parallel text corpora. These advancements in multilingual pretraining set the stage for models like M-IRRA to effectively handle diverse linguistic inputs in TBPS tasks.

III. Methodology

A. Overall proposed framework

In this study, we focus on building a multilingual TBPS system to provide a scalable deep-learning model across various linguistic contexts. The proposed model is considered as an improved version of IRRA model [1] by replacing CLIP-based text encoder by the sentence transformer Multilingual CLIP. Figure 1 illustrates the overall of M-IRRA model consisting of two main components: Dual-encoder for feature extraction task and Cross-Transformer for feature alignment. For the feature extraction task, while CLIP-based image encoder is used to pick up useful visual information on both patch- and image-level, sentence transformer multilingual CLIP is utilized in the textual branch. As seen in Fig. 1, sentence transformer multilingual CLIP operates in a 768-dimensional vector space for the token embedding, leading to a dimensional mismatch between visual and textual features for alignment. Consequently, we introduce a new dense layer to further compress the token embedding into the dimension of the visual descriptions for matching tasks. Likewise, the Implicit Feature Alignment via Cross-Transformer can effectively bridge the gap between the text and image modalities through the cross-attention mechanism. These improvements in the M-IRRA model help achieve a higher performance in multilingual TBPS. The details of the two components are presented below.

B. Dual-Encoder for Feature Extraction

CLIP Image Encoder. In the image branch, CLIP pre-trained Vision Transformer (ViT) model is exploited for image embedding. A given input image $I \in R^{H \times W \times C}$, I is divided into several non-overlapping patches. These patches are then linearly projected to generate a sequence of 1D tokens. Additionally, the positional encoding and extra $[CLS]$ token are injected to the image embedding to provide the order of image patches in a sequence. The final sequence is treated as the input of the CLIP-based transformer to extract both local and global image features. Note that, the $[CLS]$ token is linearly projected into the joint image-text embedding space, serving as the global image representation. This informative embedding aligns with the text embedding space, facilitating efficient cross-modal comparison in text-based image retrieval tasks.

Multilingual Sentence Transformer Text Encoder. For the textual branch, M-IRRA employs the sentence transformer multilingual CLIP to extract both token- and sentence-level features. This model is built up on Multilingual Knowledge Distillation [17], by utilizing parallel data across multiple languages. For this target, the original CLIP-ViT-B32 plays a role as a teacher model, guiding the training of a multilingual DistilBERT [18] model, serving as a student. Furthermore, to exploit fine-grain information for learning the similarity between visual and textual representation, in the original work [1], MLM (Masked Language Modeling) is used. For given an input text description, the text tokens are randomly masked out with a probability of 15% and replace them with the special token [MASK]. The main target of MLM is to predict masked textual tokens relied not only on the rest of unmasked textual tokens but also on the visual tokens. These masked sequences also are fed to the input of the text encoder to generate the local representation at token-level.

By using DistilBERT for token-level embeddings, each token is represented by a 768-dimensional feature vector. This results in a dimensional mismatch when handling cross-modality alignment between visual and textual descriptions. It is necessary to adopt a dense layer after the DistilBERT block to compress the token embedding to the same size as the patch-level image features. The architectural change in M-IRRA demonstrates a thoughtful approach to integrating advanced language models into multi-modal framework. By introducing

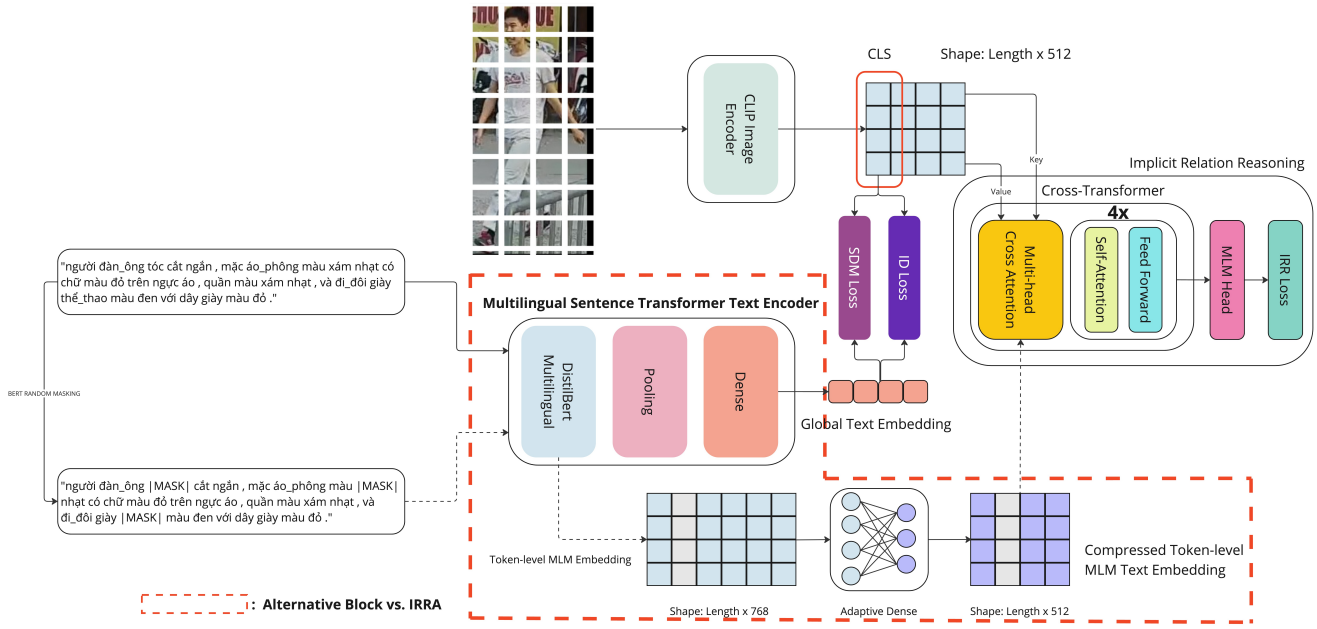


Figure 1. Overall of the M-IRRA framework for multilingual text-based person search.

this dense layer, M-IRRA gains the ability to work with different text encoders without major architectural changes, potentially allowing for easier adaptation to future advances in NLP models. This flexibility is particularly valuable in the rapidly evolving field of multi-modal learning, where new language models are frequently introduced.

C. Implicit Relation Reasoning

As aforementioned, the masked input sequences are fed forward to DistilBERT text encoder to generate token-level features. Furthermore, to implicitly align visual and textual information, a token-level feature is considered as an input Query $\mathcal{Q}$ of Cross Attention Block. Likewise, patch-level image features encoded by CLIP Image Encoder will also be used as Key $\mathcal{K}$ and Value $\mathcal{V}$. These will be utilized to feed into a Transformer with multi-head cross-attention (MHCA) in the next step. Noted that to perform visual-textual alignment, token-level features should have the same dimension with image representation. Consequently, a dense layer is adopted into framework to transform the original token-level features into 512-dimension vectors, the same size with patch-level representation. The output of MHCA block is presented in the following Equation as below [1]:

$$\mathrm{MHCA}(\mathcal{Q},\mathcal{K},\mathcal{V}) = \mathrm{softmax}\left(\frac{\mathcal{Q}\mathcal{K}^\top}{\sqrt{d}}\right)\mathcal{V} \tag{1}$$

with d being the embedding dimension of masked tokens.

D. Loss function

Similar to the IRRA model, for training M-IRRA, ID loss (Instance loss) and SDM (Similarity Distribution Matching) are used in the global branch, while IRR loss is exploited in the local branch. The overall optimization objective for the training phase is aggregated as in the following function:

$$\mathcal{L} = \mathcal{L}_{irr} + \mathcal{L}_{sdm} + \mathcal{L}_{id}. \tag{2}$$

The more detailed of these loss functions can be found in [1].

ID loss: This loss function (i.e. $\mathcal{L}_{id}$) is defined in [6]. It is known as the softmax loss that classifies an image or text into distinct groups based on their identities. ID loss is computed on both visual and textual branch.

SDM loss: This loss function denoted $\mathcal{L}_{sdm}$ is computed based on the probability of matching pairs between a mini-patch image and a token. These probability is relied on the softmax function of the similarity between local features of a mini-patch and a token, computed via the Cosine similarity function. The SDM loss is finally computed as:

$$\mathcal{L}_{sdm} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i} \tag{3}$$

where

$$\mathcal{L}_{i2t} = KL\left(\mathbf{p_i}\|\mathbf{q_i}\right) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p_{i,j}\log\left(\frac{p_{i,j}}{q_{i,j}+\epsilon}\right) \tag{4}$$

with

$$p_{i,j} = \frac{\exp\left(\mathrm{sim}\left(f_i^v, f_j^t\right)/\tau\right)}{\sum_{k=1}^{N}\exp\left(\mathrm{sim}\left(f_i^v, f_k^t\right)/\tau\right)} \tag{5}$$

where $p_{i,j}$ represents the matching probability between the i-th mini-patch visual feature f_i^v and the j-th token feature f_j^t, computed using softmax with temperature τ over N elements. The loss $\mathcal{L}_{i2t}$ is calculated using Kullback-Leibler divergence between $p_{i,j}$ and a reference distribution $q_{i,j}$, with ε added for numerical stability. The loss $\mathcal{L}_{t2i}$ is calculated in a similar way to $\mathcal{L}_{i2t}$.

IRR loss: This loss function, $\mathcal{L}_{\text{irr}}$, is used in III-C as the optimization objective by calculating the average cross-entropy loss by comparing the predicted probabilities p_j^i for classes to the true labels y_j^i, encouraging the model to assign higher probabilities to the correct classes.

$$\mathcal{L}_{\text{irr}} = -\frac{1}{|\mathcal{M}||\mathcal{V}|} \sum_{i \in \mathcal{M}} \sum_{j \in |\mathcal{V}|} y_j^i \log p_j^i \tag{6}$$

in which, $\mathcal{M}$ is the set of masked tokens, $|\mathcal{V}|$ is the size of the vocabulary and p_j^i is the probability of each masked tokens i^{th} belonging the class j^{th}.

IV. Experiments

A. Datasets and evaluation metrics

The dataset we use to evaluate the proposed M-IRRA model is 3000VnPersonsearch [2]. It comprises 6,302 images of 3,000 individuals (IDs). Each image has two descriptions, resulting in a total of 12,604 sentences. From the original 3000Vn-PersonSearch dataset, a derived dataset named 3000Vn-V2E is created by translating the Vietnamese description sentences into English using the VinAI [19] translation tool. The quality of the translation scripts has been evaluated in [3]. Two datasets have the same way to split training and testing sets. The training set has 2000 IDs (4302 images, 8604 descriptions), and the testing set has 1000 IDs (2000 images, 4000 descriptions).

The objective of TBPS is to produce a ranked list of person images from the database, ordered by their similarity to the given textual description, with the highest similarity first. To assess TBPS's performance, the correct matching at rank-k is commonly used as a metric. This metric represents the ratio of correct matches to the total number of description sentence queries. Typically, values are measured at rank 1, 5, 10 (R@K, K =1, 5, 10). Higher values indicate a more effective TBPS method.

B. Implementation details

M-IRRA consists of a pretrained image encoder - CLIP-ViT-B32 and a Sentence Transformers text encoder based on DistilBert. In our implementation, the MHCA Transformers has a depth of 4 layers, training on the 3000VnPersonsearch dataset with input images resized to 384x128 pixels. The Adam optimizer, configured with a learning rate of *1.0e-05* and weight decay of *4.0e-05*, is employed alongside a cosine annealing schedule and a 5-epoch warm up period. Training proceeds with a batch size of 24 for 120 epochs and is optimized by combining SDM, MLM, and ID losses equally. Key hyperparameters include a momentum of 0.9, β of 0.999, α of 0.9, γ of 0.1, and τ of 0.02. Augmentations such as random horizontally flipping, random crop with padding, and random erasing are applied, and a random sampler is used to select the samples. All experiments are conducted on a single RTX 3060 12GB and the hyperparameters as well as other setups are taken from the original IRRA [1]. Following this setup, Fig. 2 shows the training loss of M-IRRA model trained on three datasets over 120 epochs, indicating the patterns of rapid initial decrease followed by gradual convergence. The model trained on the combined dataset (3000Vn-PersonSearch + 3000Vn-V2E) consistently outperforms the models trained on individual dataset, achieving the lowest loss and fastest convergence.

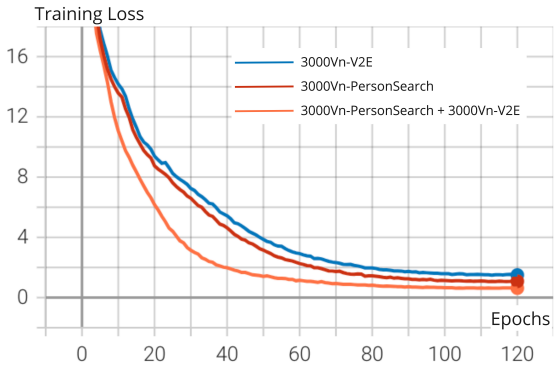


Figure 2. M-IRRA training loss on different dataset with respect to the number of training epochs.

C. Experimental Results

In this experiment, we investigate the effectiveness of the proposed framework in both monolingual and multilingual text-based person search.

1) Monolingual performance: We evaluate first the monolingual performance of the proposed M-IRRA. Results are reported in Table I. When trained and tested on the original Vietnamese dataset (3000VnPersonSearch), M-IRRA achieves a rank 1 accuracy of 38.65%. Similarly, when trained and tested on the English-translated dataset (3000Vn-V2E), it reaches a Rank-1 accuracy of 37.07%. This demonstrates that M-IRRA maintains consistent performance across both languages, with only a slight decrease (1.58 percentage points) when switching to English. The reason is that the maximum number of tokens in English descriptions (118) is significantly lower than in Vietnamese (187). This discrepancy may contribute to the performance differences between languages. The Vietnamese descriptions generally use a lot of complex adjectives to describe things so it could potentially provide more detailed information, which could explain the slightly better performance of M-IRRA on the Vietnamese dataset (38.65%) compared to the English dataset (37.07%).

2) Multilingual performance evaluation: We access the multilingual performance of M-IRRA. Results are reported in Table II. In this evaluation, M-IRRA is trained on both 3000VnPersonSearch and 3000Vn-V2E. Testing on the combined dataset of English and Vietnamese descriptions yields a Rank-1 accuracy of 37.32%. More impressively, when tested separately, it achieves 38.45% on 3000VnPersonSearch for Vietnamese descriptions only and 36.20% on 3000Vn-V2E for English descriptions only. This indicates that multilingual training enables M-IRRA to perform consistently well in both

Table I
PERSON SEARCH RESULTS OF OUR M-IRRA MODEL IN MONOLINGUAL SCENARIO. BEST VALUES ARE IN BOLD.

Train Set	Train Language	Test Set	Test Language	R@1
3000VnPersonSearch	Vietnamese	3000VnPersonSearch	Vietnamese	**38.65%**
3000Vn-V2E	English	3000Vn-V2E	English	37.07%

Table II
PERSON SEARCH RESULTS OF M-IRRA IN MULTILINGUAL SCENARIO. BEST VALUES ARE IN BOLD.

Train Set	Train Language	Test Set	Test Language	R@1
3000VnPersonSearch + 3000Vn-V2E	Vietnamese + English	3000VnPersonSearch + 3000Vn-V2E	Vietnamese + English	37.32%
3000VnPersonSearch + 3000Vn-V2E	Vietnamese + English	3000VnPersonSearch	Vietnamese	**38.45%**
3000VnPersonSearch + 3000Vn-V2E	Vietnamese + English	3000Vn-V2E	English	36.20%

languages without significant performance degradation. Some qualitative results are illustrated in Table III.

Table III
SEARCH RESULTS RETURNED BY THE PROPOSED MODEL M-IRRA FOR QUERY IN ENGLISH AND THAT IN VIETNAMESE.

Description	Top five matched images
VN: Nữ tóc dài màu đen để xoã, mặc áo phông cổ tròn cộc tay màu đỏ, đeo khẩu trang màu xanh da rời, mặc quần bò màu xanh sẫm, đi giầy màu trắng, trước ngực áo có hoạ tiết màu đen, tay phải co ngang ngực.	
EN: Women with long black hair wear a red short - sleeved round - neck T - shirt, a blue mask, dark blue jeans, white shoes. The front of the shirt has a black pattern. The right hand is drawn across the chest.	

3) Comparison with the state-of-the-art methods: In this section, we present comparison results with state-of-the-art methods for Vietnamese Text-based Person Search, in which all models are trained and tested on the same dataset 3000VnPersonSearch. All of the results have been shown as Table IV, the best results are in bold. Our proposed M-IRRA model significantly outperforms existing state-of-the-art methods across all evaluation metrics. At R@1, M-IRRA achieves an accuracy of 38.65%, which represents a substantial improvement over both GNA-RNN (26.85%) and ViTAA (27.08%). This corresponds to a relative improvement of 11.8% and 11.57% over GNA-RNN and ViTAA, respectively. The performance gains are even more pronounced at higher ranks. At R@5, M-IRRA reaches an accuracy of 67.4%, compared to 52.47% for GNA-RNN and 51.38% for ViTAA. This represents a relative improvement of 14.93 percentage point over GNA-RNN and 16.02 percentage point over ViTAA. Similarly, at R@10, M-IRRA achieves 77.62% accuracy, significantly surpassing both GNA-RNN (66.40%) and ViTAA (63.00%). These results correspond to relative improvements of 11.2 percentage point and 14.62 percentage point over GNA-RNN and ViTAA, respectively.

It's worth noting that M-IRRA consistently outperforms both baseline models across all ranks, demonstrating its superior ability to match textual descriptions with corresponding images in the Vietnamese language context. The substantial improvements, particularly at Rank-1, highlight the effectiveness of our multilingual approach in addressing the challenges of text-based person search in non-English languages. These results underscore the potential of M-IRRA to advance the state-of-the-art in multilingual text-based person search, offering a more robust and accurate solution for diverse linguistic scenarios.

Table IV
THE COMPARISON BETWEEN THE OBTAINED RESULTS OF M-IRRA MODEL AND THOSE OF THE SOTA METHODS ON 3000VNPERSONSEARCH.

Model	R@1	R@5	R@10
GNA-RNN[20]	26.85%	52.47%	66.40%
ViTAA[21]	27.08%	51.38%	63.00%
M-IRRA	**38.65%**	**67.40%**	**77.62%**

V. CONCLUSION AND FUTURE WORK

In this paper, we presented M-IRRA, a new method for multilingual text-based person search. M-IRRA is built upon IRRA framework by incorporating a Sentence Transformer Multilingual CLIP model. Our experiments on the 3000VnPersonSearch dataset and its English-translated version demonstrate M-IRRA's superior performance, achieving a rank-1 accuracy of 38.65% on the Vietnamese dataset and outperforming current state-of-the-art methods. Notably, M-IRRA maintains consistent accuracy across both Vietnamese and English queries, underscoring its robustness in multilingual contexts. This work marks the first attempt to address multilingual text-based person search, paving the way for more inclusive and global person search systems. Future research could focus on developing more sophisticated masking strategies to better capture features, optimizing description lengths to enhance cross-lingual performance, and exploring natively multilingual models for improved generalizability. Additionally, finding alternative methods to align embedding dimensions without information loss and expanding multilingual datasets for comprehensive evaluation are crucial steps forward. In conclusion, M-IRRA represents a significant advancement in multilingual text-based person search, providing a strong foundation for future research aimed at creating more effective and versatile systems for global use.

ACKNOWLEDGMENT

This research was funded by the Vietnam Ministry of Education and Training under grant number B2024-GHA-11.

REFERENCES

[1] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[2] T. T. T. Pham, H.-Q. Nguyen, H. Phan, *et al.*, "Towards a large-scale person search by vietnamese natural language: Dataset and methods," *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 27 569–27 600, 2022.

[3] T.-H. Phan, H.-P. Tran, H.-S. Bui, *et al.*, "Systematic evaluation of language characteristic and data enrichment in text-based person search," in *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2024.

[4] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2019.

[5] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[6] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 2, May 2020, ISSN: 1551-6857. DOI: 10.1145/3383184.

[7] Y. Chen, G. Zhang, Y. Lu, Z. Wang, and Y. Zheng, "Tipcb: A simple but effective part-based convolutional baseline for text-based person search," *Neurocomputing*, vol. 494, pp. 171–181, 2022, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2022.04.081.

[8] L. Gao, K. Niu, B. Jiao, P. Wang, and Y. Zhang, "Addressing information inequality for text-based person search via pedestrian-centric visual denoising and bias-aware alignments," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7884–7899, 2023.

[9] S. Li, M. Cao, and M. Zhang, "Learning semantic-aligned feature representation for text-based person search," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 2724–2728.

[10] Y. Bai, M. Cao, D. Gao, *et al.*, "Rasa: Relation and sensitivity aware representation learning for text-based person search," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, ser. IJCAI '23, Macao, P.R.China, 2023, ISBN: 978-1-956792-03-4. DOI: 10.24963/ijcai.2023/62. [Online]. Available: https://doi.org/10.24963/ijcai.2023/62.

[11] J. Cha, J. Mun, and B. Roh, "Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] M. Ni, H. Huang, L. Su, *et al.*, "M3p: Learning universal representations via multitask multilingual multimodal pre-training," *arXiv*, vol. abs/2006.02635, 2020.

[13] M. Zhou, L. Zhou, S. Wang, *et al.*, "Uc2: Universal cross-lingual cross-modal vision-and-language pre-training," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, Jun. 2021.

[14] G. Chen, L. Hou, Y. Chen, *et al.*, "mCLIP: Multilingual clip via cross-lingual transfer," in *Proceedings of ACL*, Toronto, Canada, 2023.

[15] A. Jain, M. Guo, K. Srinivasan, *et al.*, "MURAL: Multimodal, multitask representations across languages," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3449–3463. DOI: 10.18653/v1/2021.findings-emnlp.293.

[16] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, *Sigmoid loss for language image pre-training*, 2023. arXiv: 2303.15343 [cs.CV].

[17] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2020.

[18] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

[19] T. H. Nguyen, T.-D. H. Nguyen, D. Phung, *et al.*, "A Vietnamese-English Neural Machine Translation System," in *Proceedings of the 23rd Annual Conference of the International Speech Communication Association: Show and Tell (INTERSPEECH)*, 2022.

[20] T. T. T. Pham, D.-D. Nguyen, B. H. P. Ta, T.-B. Nguyen, T.-N.-D. Do, and T.-L. Le, "Person search by queried description in vietnamese natural language," in *Intelligent Information and Database Systems: 12th Asian Conference, ACIIDS 2020, Phuket, Thailand, March 23–26, 2020, Proceedings 12*, Springer, 2020, pp. 469–480.

[21] T. T. T. Pham, V.-T. Nguyen, H.-Q. Nguyen, *et al.*, "Person search by natural language description in vietnamese using pre-trained visual-textual attributes alignment model," in *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, 2021, pp. 1–6.

Data Augmentation and Assessment for Enhanced Ovarian Tumor Classification

Thi-Loan Pham*†, Gia-Minh Pham*, Tien-Dat Nguyen*, Van-Hung Le§, Thi-Lan Le*,
Duy-Hai Vu*, Hai Vu*, Chi-Mai Pham‡, Thanh-Hai Tran*
* School of Electrical and Electronic Engineering
Hanoi University of Science and Technology, Hanoi, Vietnam
† Hai Duong University, Vietnam
‡ National Hospital of Obstetrics and Gynecology, Hanoi, Vietnam
§ Tan Trao University, Vietnam
E-mail: hai.tranthithanh1@hust.edu.vn

***Abstract*—Deep learning based automatic classification of benign and malignant ovarian tumors from ovarian ultrasound images brings many significant benefits in women's health care. Nonetheless, due to privacy concerns associated with medical data, acquiring adequate data for training deep learning models is not straightforward, and sharing data for research purposes is not publicly feasible. Data augmentation helps address this issue by generating new diverse data from existing datasets. In this paper, we explore various data augmentation techniques, ranging from simple methods like geometric transforms and intensity adjustments to advanced techniques using generative AI models such as Deep Convolutional Generative Adversarial Networks (DCGAN). We then introduce a method to automatically assess the quality of generated images. The augmented data is then used to train four state-of-the-art image classification models: VGG16, ResNet34, DenseNet201 and Inception-v3. Experimental results achieved with the OTU2D dataset show an accuracy of 92.54%, increase of 1.92%. On the OvaTUS dataset, the recall reached 93.2%, increase of 5.29%. When combining the OTU2D and OvaTUS datasets, the accuracy was 91.69%, increase of 4.59%. Experimental results show that the generated images significantly improve the classification accuracy for all tested models on three datasets (OTU2D, OvaTUS, combined OTU2D with OvaTUS).**

I. Introduction

Ovarian tumors are abnormal growths that develop in the ovaries of women. These tumors can be either noncancerous (benign) or cancerous (malignant). Early detection of those tumors is crucial as it provides significant health, psychological, and economic benefits for patients. Ultrasound examination is now a primary method for assessing ovarian pathology [1]. However, the accuracy of this method greatly depends on the specialist's skills [2]. Consequently, missed or inaccurately detected tumors can result in incorrect treatment.

With the increasing advances in technology, artificial intelligence, and deep learning have emerged as cutting-edge and valuable approaches in modern medicine. In the field of ovarian tumor detection from ultrasound images, several studies have been proposed. Jose et al. [3] conducted a comparative study to evaluate the performance of four machine learning techniques (K-Nearest Neighbors, Linear Discriminant, SVM, and Extreme Learning Machine) to classify ovarian tumors using Fourier descriptors computed on the Region Of Interest. They utilized a real patient dataset of 348 IOTA tumor images. Their results indicated that classifiers based on Support Vector Machines and Ensemble Learning Machines are particularly robust for this type of biomedical classification. They achieved more than 85% of accuracy. With traditional machine learning, Lotfi et al. [4] introduced a novel prognostic method for predicting the time to tumor recurrence in ovarian cancer, utilizing genes as the independent variables. The authors performed automatic analysis and classification of ovarian tumors directly from B-mode ultrasound images appears limited.

Recently, significant advancements have been made in the diagnosis of ovarian tumors using Convolutional Neural Network (CNN) models, enhancing both diagnostic accuracy and speed. Shih-Tien et al. [5] conducted experiments with ten well-known CNNs using transfer learning. They found that ResNet-50 achieved the highest accuracy (90.51%), DenseNet-201 attained the highest sensitivity (89.77%), and MobileNetv2 exhibited the poorest performance. To attain such high performance, it is required to train CNN models with a large amount of images, but ultrasound images are often scarce due to privacy concerns. Therefore, data augmentation becomes crucial to enrich the dataset.

In this study, we employ various data augmentation techniques, including geometric transformations, intensity adjustments, and generative adversarial networks (GANs), specifically DCGAN [6]. While geometric transformations and intensity adjustments are commonly used and generally do not significantly alter image quality, GANs can produce unexpected results that contain unrealistic errors or unwanted details. To address this, we introduce a novel method for objectively evaluating the quality of GAN-generated samples. Finally, we use the augmented data to train existing state-of-the-art classification models: ResNet34 [7], VGG16 [8], DenseNet201 [9], and Inception-v3[10], to classify benign and malignant ovarian tumors in ultrasound images. Our results demonstrate that data augmentation improves classification accuracy, and the samples generated by DCGAN contribute more effectively than traditional augmentation techniques.

In summary, our contributions are threefold: First, we ex-

plore various image augmenting combination for ultrasound images. Second, we introduce a novel method to evaluate the quality of generated images. Lastly, we conduct a comparative study of classification networks, identifying the most effective augmentation techniques to train the CNN models on three datasets for classifying benign and malignant tumors.

II. Proposed Method

A. General framework

In this study, we propose a framework for data augmentation, data assessment, and image classification as shown in Fig. 1. It is composed of three main blocks: i) data augmentation block includes various modules (geometric transformation, intensity adjustment, and DCGAN); ii) data assessment using an objective metric of image quality; iii) a benign and malignant classification module. In the following, we will present in detail each component in our framework.

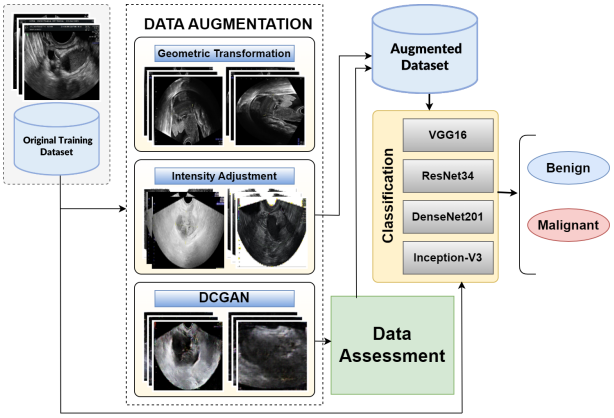


Fig. 1. The framework of ovarian tumor classification from ultrasound images with data augmentation.

B. Data augmentation techniques

1) Geometric transformation: Geometric transformations have been extensively applied to augment datasets effectively. For example, in [11], geometric augmentation demonstrated a noteworthy enhancement, improving CNN performance by up to 5%. In this work, we enhance our datasets by applying various transformations, including a 15° rotation, random zoom between 0.5 (reduces 1/2 size) to 2 (zooms 2 times), and vertical and horizontal flipping.

2) Intensity adjustment: Image augmentation through intensity transformation is an important technique in digital image processing that aims to vary image brightness or contrast. In this study, we use two methods: Contrast stretching (often called normalization) is a simple image enhancement technique that attempts to improve the contrast in an image by 'stretching' the range of intensity values it contains to span a desired range of values, e.g. the the full range of pixel values that the image type concerned allows, and intensity smoothing reduces noise and blurs the image by applying a 5×5 mean filter. Fig. 2 illustrates some examples.

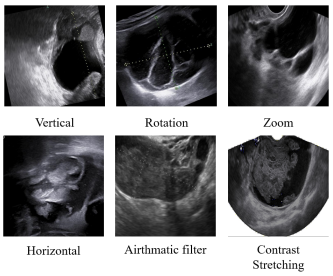


Fig. 2. Images generated by geometric transformation & intensity adjustment.

3) DCGAN model: GAN comprises two distinct networks: a generator network G and a discriminator network D. The generator G is designed to learn the data distribution by taking random noise z (a random number) and producing images from it. Conversely, the discriminator D's role is to assess whether an image is "real" or fabricated [12]. Compared to the original GAN, DCGAN model incorporated convolutional layers in the discriminator network and convolutional-transpose layers in the generator network [6]. As a result, it enhanced the performance compared to the original one. Generator and Discriminator in DCGAN are described as follows.

a) Generator: The generator starts with a 100×1 random noise vector input, reshaped by a dense layer to $4 \times 4 \times 1024$. Then, convolution-transpose layers upsample this to an image of $128 \times 128 \times 3$, using ReLU activation except for the output layer which uses Tanh activation. Batch normalization is applied to all layers except the output layer to stabilize the learning process.

b) Discriminator: The discriminator's goal is to classify images as real or fake. It takes $128 \times 128 \times 3$ images, both real from the dataset and generated by the generator, and processes them through a series of convolutions, ending with a sigmoid activation to determine authenticity. Each convolution block is followed by a LeakyReLU activation and batch normalization is applied to all layers except the input layer.

The model's loss function is a combined function that incorporates the objectives of both the Discriminator and the Generator as in the equation below:

$$\min_G \max_D V(D,G) =$$
$$E_{x \sim p_{\text{data (x)}}}[\log[D(x)]] + E_{Z \sim P_{(z)}}[\log[1 - D(G(z))]] \quad (1)$$

where x represents the sample image data, $P(z)$ represents the noise distribution, Z denotes the number of random vectors in $P(z)$, and E is the expectation of $P(z)$. $G(z)$ is the image generated by the Generator. $D(x)$ is the Discriminator's prediction value indicating whether the image x is real, and $D(G(z))$ is the prediction value indicating whether the image generated by the Generator is real. Through the training process, both Generator and Discriminator enhance their capabilities. The goal is for the Generator to create more realistic data, for the Discriminator to become proficient at classifying real data and data generated by the Generator, and ideally, for the Generator to produce data that the Discriminator can not distinguish from

real data, as illustrated in Fig 3 and Fig 4 illustrations of the results of ovarian ultrasound image generation using the DCGAN model.

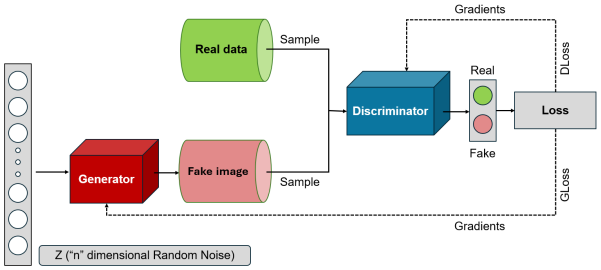


Fig. 3. The training process of the DCGAN model.

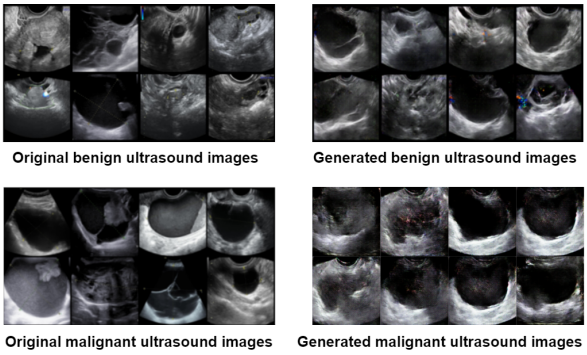


Fig. 4. Illustration of the results of ovarian ultrasound image generation using the DCGAN model.

C. Generative images assessment

To evaluate the quality of images generated by GANs, several techniques are commonly used, such as Inception Score (IS) [13] and Fréchet Inception Distance (FID) [14]. However, these metrics focus on the assessment of the generative models, which also cannot be used to assess the quality of each single generated image. Our objective, in contrast, is to rank the generated images in terms of quality score to take the best images for further processing. In [15], the authors proposed three algorithms for quantitatively evaluating the quality of each image. Among them, GMM-GIQA achieved promising results. GMM-GIQA estimated the real image distribution by a GMM with various numbers of Gaussian Components. The quality score is considered the probability of occurring a generated image in that distribution.

In this paper, we draw inspiration from the FID method to evaluate generated images. First, we input the original dataset into a ResNet50 model, retaining all layers except the final one. Each image passing through the network produces a feature vector of size 2048×1. Similarly, we pass the images generated by the generator network through the ResNet50 network to obtain corresponding feature vectors (size 2048×1) for each image (Fig. 5). The Euclidean distance between the mean vector of the multivariate Gaussian distributions of the real image set and the feature vector computed from the generated image serves as the criterion for evaluating the quality of the generated images. The greater the distance, the lower the quality of the generated image.

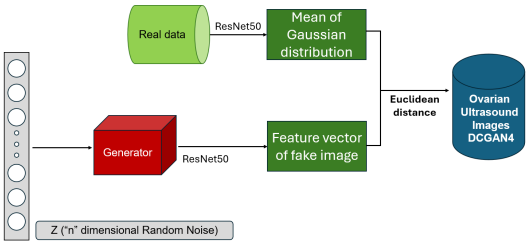


Fig. 5. Quality evaluation process of generated images by DCGAN.

Fig. 6 illustrates the best and worst images generated in the two classes ranked by the smallest and largest Euclidean distances. It is noted that the high-quality generated images seem to closely resemble original images, while the low-quality generated images exhibit unusual textures and details. We then sort the generated images based on their quality scores and retain only a predefined number of highest-scoring images. These selected images are combined with the original images to create augmented datasets, as presented in Section III.

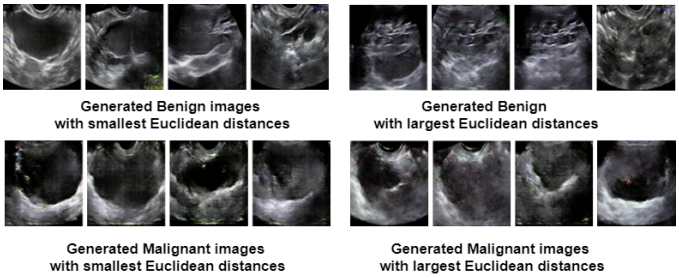


Fig. 6. Illustration of the results of the quality evaluation process of DCGAN generated images.

D. Classification models

The CNN model is widely used in ovarian tumor classification due to its effective image-processing capabilities. Among CNN models, ResNet34 stands out with its deep architecture that helps reduce the vanishing gradient problem. VGG16, with its simple design and densely packed convolutional layers, provides high accuracy in recognition tasks. DenseNet201 connects all layers, improving learning ability and resource efficiency. Finally, Inception-v3, with its technique of combining multiple convolutional sizes in a single layer, enhances accuracy and minimizes errors in diagnosis. So We use four CNN models, namely ResNet34, VGG16, DenseNet201, and Inception-v3, to conduct experiments on classification on both original and augmented datasets.

ResNet34: The ResNet-34 network is primarily built using residual building blocks, which serve as its core infrastructure. ResNet is a subset of the broader CNN family, characterized by a typical architecture that includes 34 layers, comprising

convolutional layers, one MaxPool layer, and one global average pooling layer. ResNet-34 has 34 weighted layers in its initial form. It also has 512 cores, 11x11 cores, and layers with kernel sizes of 9×9, 12×12, 18x18, and 9x9 [7].

VGG16: VGG16 has gained popularity among CNN architectures and is often adopted as a foundational framework for developing more complex CNN models through transfer learning. Its deep yet simple structure makes it an excellent choice for developing complex CNNs. The design, featuring a total of 16 layers and accepting 224x224 pixel input images, is predominantly composed of convolutional layers, complemented by fully connected layers [16].

Inception-V3: A sophisticated convolutional neural network with 48 layers comprises the Inception-V3 model. Its pretraining utilized an extensive ImageNet collection, which includes over one million distinct photographic images [10]. Inception-V3 introduces improvements to its forebear within the Inception-architecture to improve computing efficiency.

DenseNet201: The DenseNet201 architecture facilitates the creation of simple and straightforward models. Moreover, it allows for the reuse of features across layers, thereby enhancing parameter efficiency and enabling greater variation and improved performance in subsequent layers. The architecture employs a feed-forward approach, connecting each layer to every other layer [9].

III. Experiments

A. Datasets

1) OTU2D dataset: This dataset comprises 1,469 2D ultrasound images belonging to eight categories [17]. Initially, each image is annotated by one expert and subsequently verified by another to ensure high annotation quality. Among the eight categories, the first six (Chocolate cysts, Serous cystadenoma, Teratoma, Theca cell tumor, Simple cysts, and Normal ovary) represent benign ovarian tumors, while the last two (Mucinous cystadenoma and High-grade serous cystadenoma) indicate malignant ovarian tumors. In this paper, we classify the ultrasound images into two classes: malignant and benign, without specifying the individual categories of each.

2) OvaTUS dataset: The OvaTUS dataset was collected by our research team in collaboration with the National Hospital of Obstetrics and Gynecology (NHOG) in Hanoi, Vietnam. This dataset comprises 892 ultrasound images of benign ovarian tumors and 309 ultrasound images of malignant ovarian tumors. To ensure consistency in imaging and annotation, the ultrasounds were performed and labeled by the same sonographers who conducted the examinations.

3) OvaTUS+OTU2D dataset: From the OTU2D and OvaTUS datasets, we have created a third dataset, named OvaTUS+OTU2D. Our objective is to compile a larger and more diverse dataset with images taken from different settings. This dataset contains 2670 images (2204 benign, 466 malignant).

4) Augmented datasets: From three described datasets, we create new augmented datasets using data augmentation techniques presented in section II. It is important to note that two DCGANs were trained independently for two classes (malignant and benign).

TABLE I
THE NUMBER OF IMAGE SAMPLES IN DATASETS.

Datasets	Train	Test	Benign Tumors	Malignant Tumors
OTU2D (original set)	1000	469	1312	157
OTU2D-DCGAN1	2500	469	2312	657
OTU2D-DCGAN2	1500	469	1312	657
OTU2D-DCGAN3	4000	469	3312	1157
OTU2D-DCGAN4	2500	469	2312	657
OTU2D-GI	2500	469	2312	657
OvaTUS (original set)	804	397	892	309
OvaTUS-DCGAN1	2304	397	1892	809
OvaTUS-DCGAN2	1304	397	892	809
OvaTUS-DCGAN3	3804	397	2892	1309
OvaTUS-DCGAN4	2304	397	1892	809
OvaTUS-GI	2304	397	1892	809
OvaTUS+OTU2D (original set)	1804	866	2204	466
OvaTUS+OTU2D-DCGAN1	3304	866	3204	966
OvaTUS+OTU2D-DCGAN2	2304	866	2204	966
OvaTUS+OTU2D-DCGAN3	4804	866	4204	1466
OvaTUS+OTU2D-DCGAN4	3304	866	3204	966
OvaTUS+OTU2D-GI	3304	866	3204	966

1) -GI: We simply apply geometric transformation and intensity adjustment to generate an additional 1000 samples for the benign class and 500 samples for the malignant class in datasets.
2) -DCGAN1: We generate an additional 500 samples for the malignant class and 1,000 samples for the benign class using DCGAN. Our goal is to augment the overall dataset by increasing number of samples in both classes.
3) -DCGAN2: We only generate an additional 500 samples for the malignant class using DCGAN. We aim to specifically increase the number of samples in the minority class while maintaining the overall balance in the dataset.
4) -DCGAN3: We apply geometric transformation, intensity adjustment, and DCGAN to generate 2000 samples for the benign class and 1000 samples for the malignant class. So -DCGAN3 is a combination of GI and DCGAN1. We would like to evaluate the combined augmentation data impact on final classification performance.
5) -DCGAN4: We generate an additional 4,000 samples for each malignant and benign class using DCGAN. We then evaluate the quality of these images using our image assessment technique and retain only the 1,000 best benign images and 500 best malignant images.

Table. I summarize the original and augmented datasets with corresponding training and testing samples.

B. Evaluating of ovarian tumors classification

To evaluate the ovarian tumor classification, we employ the common metrics of image classification which are recall (R), precision (P), accuracy (Acc), and F_1 score.

1) Ovarian tumor classification results on OTU2D dataset: The classification results in Table II across various models indicate that DenseNet201 consistently outperforms the other

TABLE II
OVARIAN TUMOR CLASSIFICATION RESULTS ON OTU2D DATASET.

Dataset	Measure (%)	ResNet34	VGG16	Dense Net201	Inception _v3
OTU2D	*Acc*	90.62	90.83	93.18	92.11
	P	88.78	89.04	92.51	91.09
	R	90.62	90.83	93.18	92.11
	F_1	89.02	88.81	92.56	90.84
OTU2D-GI	*Acc*	92.32	91.26	92.10	91.04
	P	91.61	90.09	92.51	90.28
	R	92.32	91.26	92.18	91.04
	F_1	**91.83**	90.40	92.30	90.58
OTU2D -DCGAN1	*Acc*	**92.54**	91.26	93.39	91.47
	P	**91.68**	89.78	93.03	90.41
	R	**92.54**	91.26	93.39	91.47
	F_1	91.46	89.68	92.32	90.47
OTU2D -DCGAN2	*Acc*	92.11	**92.11**	**94.24**	92.32
	P	91.24	**92.13**	**93.82**	91.38
	R	92.11	**92.11**	**94.24**	92.32
	F_1	91.45	90.01	**93.76**	91.40
OTU2D -DCGAN3	*Acc*	91.47	91.68	92.54	**92.54**
	P	90.18	90.46	91.75	**91.68**
	R	81.47	91.68	92.54	**92.54**
	F_1	90.31	**90.49**	91.91	**91.46**
OTU2D -DCGAN4	*Acc*	91.47	91.47	92.11	91.47
	P	90.24	90.12	91.24	90.13
	R	91.47	91.47	92.11	91.47
	F_1	90.45	90.02	91.45	90.17

TABLE III
THE OVARIAN TUMOR CLASSIFICATION RESULTS OF THE OVATUS DATASET.

Dataset	Measure (%)	ResNet34	VGG16	Dense Net201	Inception _v3
OvaTUS	*Acc*	87.91	87.15	89.42	87.10
	P	87.64	86.89	89.22	87.69
	R	87.91	87.15	89.42	87.91
	F_1	87.48	86.97	89.13	87.75
OvaTUS -GI	*Acc*	88.66	**89.17**	89.45	89.92
	P	88.44	**88.96**	89.22	89.92
	R	88.66	**89.17**	89.45	89.92
	F_1	88.29	**88.85**	89.33	89.92
OvaTUS DCGAN1	*Acc*	90.43	86.65	88.67	88.92
	P	90.27	86.61	88.57	88.73
	R	90.43	86.65	88.66	88.92
	F_1	90.20	86.63	88.61	88.52
OvaTUS -DCGAN2	*Acc*	87.91	87.41	**90.68**	**91.69**
	P	87.93	87.17	**90.53**	**91.57**
	R	87.91	87.41	**90.68**	**91.69**
	F_1	87.21	87.24	**90.54**	**91.60**
OvaTUS -DCGAN3	*Acc*	88.16	88.45	89.43	87.40
	P	87.89	87.25	89.33	87.45
	R	88.16	88.45	89.43	87.41
	F_1	87.90	87.80	89.20	86.61
OvaTUS -DCGAN4	*Acc*	**93.20**	87.66	89.92	88.66
	P	**93.14**	87.38	89.74	88.42
	R	**93.20**	87.66	89.92	88.66
	F_1	**93.08**	87.43	89.69	88.33

architectures, achieving an accuracy of 94.24% and an F1 score of 93.76% for the OTU2D-DCGAN2 dataset, making it the most effective choice for this task. ResNet34 shows competitive performance with an accuracy of 92.54% and an F1 score of 91.46%, particularly when combined with DCGAN augmentations, but it does not achieve the same level of robustness as DenseNet201. The introduction of DCGAN-generated images, particularly in the datasets like OTU2D-DCGAN2, generally improves model performance, as seen in DenseNet201's results. Overall, the findings underscore importance of selecting appropriate models and augmentation techniques to optimize outcomes in medical image classification, with DenseNet201 demonstrating the highest effectiveness.

2) Ovarian tumor classification results of OvaTUS dataset: The ovarian tumor classification results in Tab. III show that the OvaTUS-DCGAN4 model achieves the highest performance with an accuracy (Acc) of 93.20%, precision (P) of 93.14%, recall (R) of 93.20%, and F1 score of 93.08% on OvaTUS-DCGAN4 dataset using ResNet34 model. Additionally, the DenseNet201 model on the OvaTUS-DCGAN2 dataset shows strong results, with a recall of 90.68% and an F1 score of 90.54%, outperforming other models on the same dataset. Models based on ResNet34 and VGG16 generally have lower performance, with accuracy ranging from 87.15% to 88.66%. Overall, the combination of DCGAN and deep learning models shows significant improvement in performance compared to the baseline models.

Confusion matrix of ovarian tumor classification results is shown in Fig. 7, it can be seen that the proposed method increases the ability to predict malignant ovarian cancer cases while maintaining the predictive capability for benign cases (R is increasing from 69.81% to 81.13% for malignant predictions). Enhancing prediction of malignant cases is a crucial criterion in healthcare.

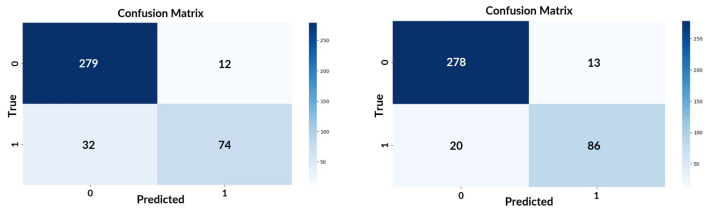


Fig. 7. The confusion matrix of ovarian tumor classification results on the OvaTUS (on the left) and OvaTUS-DCGAN4 datasets (on the right).

3) Ovarian tumor classification results on the OTU2D + OvaTUS dataset: The classification results on the OTU2D + OvaTUS dataset are shown in Tab. IV. The highest *Acc* 92.15% was achieved on the OTU2D+OvaTUS-DCGAN2 dataset using the DenseNet201 model. Fig. 8 presents the results on the metrics of the DenseNet201 model, which are the best. These results show that data augmentation has significantly improved performance compared to using the original dataset. The results for using the CNN ResNet model show an increase in accuracy from 89.95% with the original dataset to 91.22% with the -DCGAN4 dataset.

IV. CONCLUSIONS

In this paper, we explored various data augmentation techniques for ovarian ultrasound images, including geometric transformations, intensity adjustments, and deep convolutional generative adversarial networks (DCGANs). We introduced two evaluation methods for assessing the quality of generated

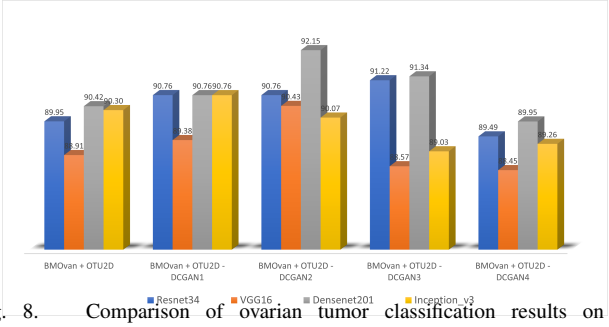


Fig. 8. Comparison of ovarian tumor classification results on the OTU2D+OvaTUS datasets using CNNs model.

TABLE IV
THE OVARIAN TUMOR CLASSIFICATION RESULTS OF THE OTU2D + OVATUS DATASET.

Dataset	Measure (%)	ResNet34	VGG16	Dense Net201	Inception _v3
OvaTUS +OTU2D	*Acc*	89.95	88.91	90.42	90.30
	P	89.46	88.34	89.96	89.84
	R	89.95	88.91	90.42	90.30
	F_1	89.13	88.50	90.04	89.94
OvaTUS +OTU2D -GI	*Acc*	89.84	88.57	90.48	90.42
	P	89.45	87.81	89.96	89.93
	R	89.84	88.57	90.48	90.42
	F_1	89.59	87.89	90.24	89.81
OvaTUS +OTU2D -DCGAN1	*Acc*	90.76	89.38	90.76	**90.76**
	P	90.34	**88.73**	90.38	**90.34**
	R	90.76	**89.38**	90.76	**90.76**
	F_1	90.42	**88.61**	90.48	**90.42**
OvaTUS +OTU2D -DCGAN2	*Acc*	90.76	**90.43**	**92.15**	**90.07**
	P	90.33	88.60	**91.91**	**89.55**
	R	90.76	89.26	**92.15**	**90.07**
	F_1	90.17	88.50	**91.98**	**89.35**
OvaTUS +OTU2D -DCGAN3	*Acc*	89.49	88.45	89.95	89.26
	P	88.86	87.63	89.44	88.71
	R	89.49	88.45	89.45	89.26
	F_1	88.83	87.58	89.53	88.20
OvaTUS +OTU2D -DCGAN4	*Acc*	**91.22**	88.57	91.34	89.03
	P	**90.83**	87.83	90.98	88.80
	R	**91.22**	88.57	91.34	89.03
	F_1	**90.78**	87.44	90.83	88.90

images: a subjective method involving doctor participation via a web-based interface, and an objective metric based on the distance of generated samples from the mean of the real dataset in feature space. Through this evaluation process, we identified the most effective augmented dataset and used it to train four state-of-the-art classification networks—ResNet, VGG, DenseNet, and Inception. Experiments conducted on three datasets demonstrated that our data augmentation approach improved performance compared to not using augmentation. We also identified optimal data augmentation settings for achieving the highest accuracy in classifying benign and malignant tumors. Future work will focus on exploring additional models for data augmentation and analyzing tumors at the regional and pixel levels.

ACKNOWLEDGMENT

This research is funded by Ministry of Science and Technology of VietNam (MOST) under grant number KC-4.0-45/19-25 **"Research and development of a computer-aided support system for ovarian cancer diagnosis using ultrasound images"**.

REFERENCES

[1] C. L. Ong, "The current status of three-dimensional ultrasonography in gynaecology," *Ultrasonography*, vol. 35, no. 1, p. 13, 2016.

[2] J. et al, *Gynecological Ultrasound Scanning: Tips and Tricks*. Cambridge University Press, 2020.

[3] M.-M. et al, "Evaluation of machine learning methods with fourier transform features for classifying ovarian tumors based on us," *PLoS One*, vol. 14, no. 7, 2019.

[4] L. et al, "Prediction of time to tumor recurrence in ovarian cancer: Comparison of three sparse regression methods," in *Bioinformatics Research and Applications: 13th International Symposium*.

[5] H. Shih-Tien and Y.-J. et al, "Automatic ovarian tumors recognition system based on ensemble cnn with ultrasound imaging," *BMC Medical Informatics and Decision Making (Web)*, vol. 22, no. 1, pp. 1–12, 2022.

[6] R. et al, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*,

[7] H. et al, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[8] L. et al, "Very deep convolutional neural network based image classification using small training sample size," in *2015 IAPR Asian conference on pattern recognition*.

[9] H. et al, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[10] S. et al, "Rethinking the inception architecture for computer vision," in *CV and pattern recognition, IEEE*.

[11] M. Z. Islam and al, "Static hand gesture recognition using convolutional neural network with data augmentation," in *2019 Joint 8th ICIEV and 2019 3rd icIVPR*.

[12] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[13] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[14] H. et al, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[15] S. Gu, J. Bao, D. Chen, and F. Wen, "Giqa: Generated image quality assessment," in *CV–ECCV 2020: 16th European Conference, Glasgow, UK, Proceedings, Part XI 16*, Springer, 2020, pp. 369–385.

[16] R. et al, "A deep learning-based framework for automatic brain tumors classification using transfer learning," *Circuits, Systems, and SP*,

[17] Q. Z and L. et al, "A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation," *arXiv preprint arXiv:2207.06799*,

A Solution For Anomaly Detection of Red Beans In A Product Processing Line

Duc Hai Nguyen*, Trong Hiep Do*, Hoang Linh Phuong Nguyen*, Quoc Khanh Nguyen*
Duc Tan Tran†, Tien-Son Bui‡, Van-Toi NGUYEN† §
* Faculty of Computer Science, PHENIKAA University, Hanoi 12116 Vietnam
† Faculty of Electrical and Electronic Engineering, PHENIKAA University, Hanoi 12116, Vietnam
‡ Hanoi University of Industry, Vietnam
§ Corresponding author. Email: toi.nguyenvan@phenikaa-uni.edu.vn

***Abstract*—Anomaly detection (AD) plays a crucial role in ensuring the quality and safety of processed products. With the rapid growth of machine learning, several AD frameworks and datasets have been introduced in the past decades. However, the ability of those models to adapt to real-world settings and overcome practical hindrances remains an issue. In this paper, we introduce a solution for abnormal red beans detection on a conveyor belt of the product processing line. We aim to provide deeper insights into the features and challenges of this task and introduce an end-to-end method which, to our best knowledge, is the first model to be applied to our regarded settings. The experimental results of our model show its promising potential for real-world industrial AD, both in terms of accuracy and latency.**

***Index Terms*—anomaly detection, processing line, benchmark, dataset**

I. Introduction

Anomaly detection (AD) in manufacturing processes plays an important role in identifying the defective and abnormal traits of products, therefore maintaining their quality. Product anomalies include a wide range of types: improperly assembled components, faulty parts or damage inflicted by external forces. Historically, AD has been carried out by human [1], which is inevitably susceptible to human weaknesses. However, the rapid development of machine learning (ML) and especially deep learning (DL) has paved the way for much more efficient AD approaches. DL-based methods for anomaly detection have garnered significant attention and have been widely studied, as the demand for automation in industrial applications becomes higher.

The primary challenge shifts to aligning the current AD frameworks with real-world challenges that a manufacturing system poses. For example, the ideal condition of one supervising camera capturing image of only one product at a time with desirable resolution cannot be met, as there can be multiple objects sliding on a conveyor belt.

In this paper, we investigate AD with real-world production line conditions, integrating environmental factors such as varying light condition, blurring object due to motion, diverse image resolutions and natural anomaly regions. These factors are expected when deploying AD models on a small to medium-scale agricultural production line. Additionally, we propose a joint object-anomaly detection framework as the solution for the aforementioned problems.

Our contributions are summarized as follows:

- We inspect a real-world anomaly detection situation in production lines, together with a specialized dataset.
- We introduce a DL-based joint object-anomaly detection pipeline for our situation and provide following experimental results.

II. Related works

A. Visual anomaly detection

Anomaly detection is one of the most prominent domains of computer vision and, to a greater extent, artificial intelligence. Anomaly detection enjoys a wide range of practical uses, ranging from intrusion inspection [2], anti-money laundering [3] to medical image analysis [4]. The diversity of anomaly detection applications has led to a remarkable number of different techniques being introduced throughout the last few decades. Anomaly detection algorithms can be divided into three categories [5]: supervised, semi-supervised and unsupervised anomaly detection.

As the major challenge of anomaly detection lies in the rarely observed, sometimes unseen and unknown, defects; unsupervised anomaly detection (UAD) has been top-priority. Primarily, UAD models belong to one of the following three types:

- Synthesizing-based methods put an emphasis on synthetically anomalous patterns imprinted on anomaly-free images. A simple convolutional neural network (CNN)-based structure is then used as a discriminator between abnormal and normal samples. Historically, synthetic anomalies has been created by cutting random patches of an image and pasting them elsewhere [6], [7] or mutating latent features with noises [8]–[10]. The main drawback of this genre is that real-world anomalies are much more diverse and unpredictable in terms of appearance and distribution than synthetic ones. Moreover, collecting a completely sufficient dictionary of anomalous features is practically infeasible.
- Embedding-based methods achieve state-of-the-art results lately. Visual representations are compressed into a latent

space using CNN backbones pre-trained with natural image datasets, most notably ImageNet [11]. Anomalous features lying in that latent space is easier to differentiate as they reside more distant from the embedding clusters. Previous studies incorporate normalizing flows [12]–[15] to train a bijection which maps pre-trained feature distribution to robust nominal data density used for identifying anomalies. k-Nearest Neighbor are also used [16]–[18] to search for out-of-distribution (OOD) data as anomalies. Additionally, knowledge distillation trains a student model to mimic a pretrained, frozen CNN used as teacher model. Anomaly detection lies in the difference between student and teacher output. Various teacher-model methods has been proposed [19]–[21] with competitive inference time.

B. Anomaly detection dataset and metrics

As anomaly detection research rapidly develops, the need for efficient datasets also grows fast. Synthetic dataset [22]–[24] have somewhat bolstered the diversity of AD data resource, though they cannot bridge the gap between industrial processes and experimental studies. Most notably, the advent of MVTEC data family, including MVTEC AD dataset [25], VisA [26] and MVTEC LOCO [9] has heralded the era of unsupervised anomaly detection [1], as they provide a formidable collection of anomalous features and encourage unsupervised methods. However, the construction of these datasets are mostly with favourable situations: individual objects captured in a static background, high resolution and detailed images. In reality, not many production lines satisfies these conditions, especially agricultural ones.

In terms of robustness and accuracy, the most prominent metrics used to measure performances of anomaly detection models are F1 score, AU-ROC (area under the Receiver Operating Characteristic curve) and AU-PR (area under the precision-recall curve) [1]. However, it is worth noting that these metrics do not consider the sparsity of abnormalities in relation to normal data, resulting in detrimentally high false positive rate (FRP) and loss of important information [1]. Latency (ms) is the predominant choice for computational efficiency evaluation, as competitive inference speed are required in industrial processes.

III. Methodology

As mentioned above, anomaly detection models tend to be designed based on the assumption of specific conditions. In the scope of our research, we would like to accurately simulate a real-world settings of anomaly detection, which shall require more than traditional anomaly detection frameworks to handle. Particularly, we divide our overall process into two phases: object phase and anomaly detection phase. Our proposed model locates and extracts region-level images of the red beans from the original conveyor belt image in the former and performs image-level anomaly detection on the extracted images in the latter phase (Fig 1).

A. Data collection

Commonly, capturing images of individual agricultural products for inspection in a conveyor system is not feasible. We consider a situation in which a number of red beans are passed through a conveyor belt.

TABLE I
Information about the data collection process

Parameter	Value
Belt Width	40 cm
Belt Color	Dark Green
Camera Resolution	1920×1080
Camera to Belt Distance	50 cm
Capture Speed	1 fps
Normal Set	974 frames
Abnormal Set	100 frames

Our system consists of a conveyor belt with a dark blue surface, 40 cm wide, and an Intel Realsense camera with a resolution of 1920×1080 mounted on a stand, positioned 50 cm above the conveyor surface and directed perpendicularly downward. The details are described as the table I. During operation, the conveyor moves steadily, with red beans scattered sparsely on the surface. The camera captures one frame per second. As a result, we obtained 1074 frames, with 974 frames containing images of normal red beans and 100 frames containing images of abnormal red beans (Figure 2).

B. Dataset description

The final outcome comprises of a training and a testing set. The training set includes 5275 images of normal red beans, while the testing set consists of both normal and abnormal images, each category comprising 712 images. Normal red beans have intact red skins and typically exhibit a small white spot at the tip, referred to as the 'bean eye'. Abnormalities include cracked, broken, sprouted, insect-infested, or moldy beans. Moldy beans range from small patches to fully covered beans, with two types of mold: white and green. Insect damage on red beans is relatively small, typically ranging from 1 to 2 holes, which can be confused with 'bean eyes'. In addition to natural abnormalities, other anomalies such as sprouting and breakage were manually introduced to diversify the dataset and simulate real-world agricultural scenarios. We provide an overview in figure 3

C. Object detection phase

For the object detection phase, we propose to use YOLO-V9 [27], a version of the YOLO [28] model family. YOLO (You Only Look Once) is an object detection model that operates by dividing the input image into a grid, predicting bounding boxes and object probabilities for each cell in the grid in a single processing pass. This increases the speed and efficiency of object detection, making it suitable for real-time applications. With real-time object detection, YOLO-V9 [27] excels with an innovative approach to overcome the challenge of information loss inherent in deep neural networks.

We annotated the coordinates of frames captured by the camera to train the YOLO-V9 model for detecting red beans.

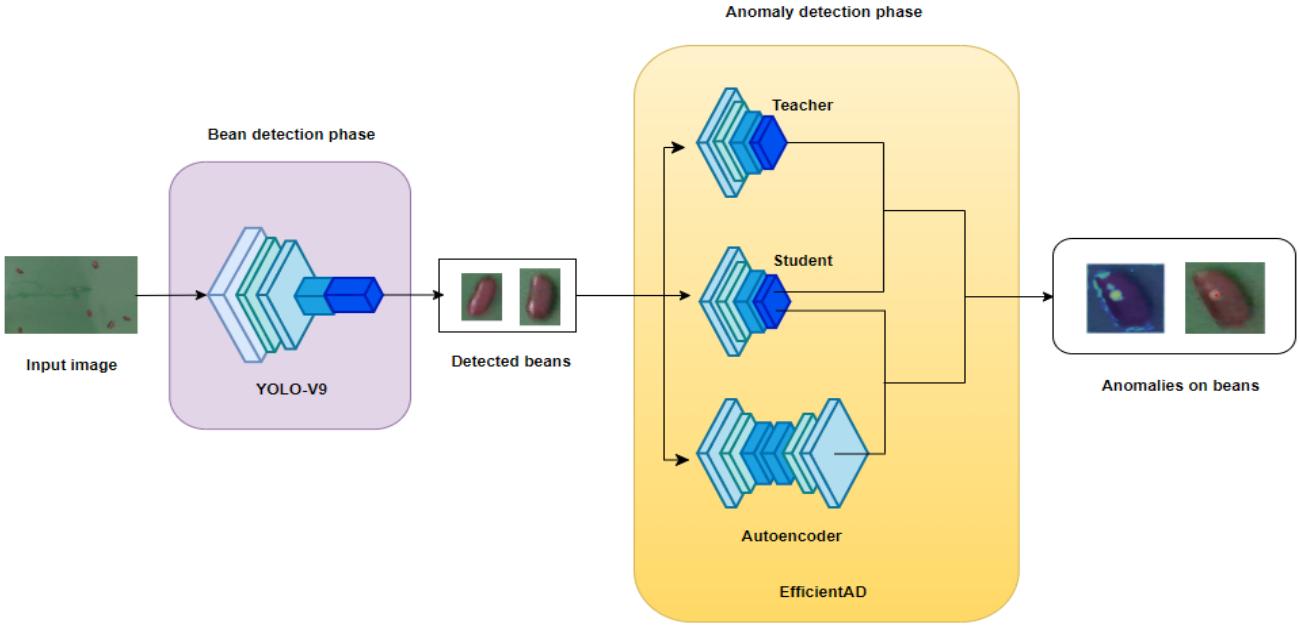


Fig. 1. The proposed process. First, we extract small frames of individual red beans from an input conveyor belt image with low resolution. Then, anomaly detection is performed given the extracted images as input. We also present the architecture of EfficientAD: student-teacher module combined with an autoencoder for detecting logical anomalies.

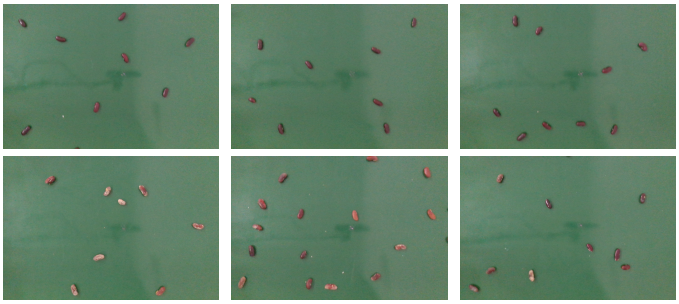


Fig. 2. Frames of images captured from the conveyor belt. The top row shows images of normal red beans, while the bottom row shows images of abnormal red beans.

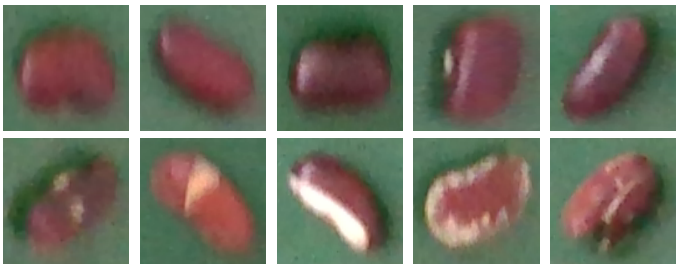


Fig. 3. Images of red beans in the dataset. Top row: normal red beans; bottom row from left to right: infested, cracked shell, broken, moldy, sprouting.

D. Anomaly detection phase

We propose to utilize EfficientAD [29] for our second phase. We shall provide the metric results of this model compared to other anomaly detection models in our experiments.

EfficientAD is one of the state-of-the-art anomaly detection models with a reconstruction-based approach. The model consists of a lightweight feature extractor specifically designed and a student-teacher structure (see Figure 1) to detect anomalous features. The feature extractor is a simple convolutional neural network (CNN) structure whose weights are distilled from a WideResNet-101 [30]. The student network is trained to predict the extracted features of anomaly-free training images, and EfficientAD relies on it failing to predict the abnormal features to detect anomalies at test time. The authors proposed a novel loss function based on Online Hard Example Mining [31] to hamper the student's ability to imitate the teacher feature extractor beyond normal images, significantly enhance the computational efficiency of the entire student-teacher framework while bettering the localization of anomalous features.

Another notable attribute of EfficientAD is its capability of handling logical anomalies, i.e incorrect combinations of normal local features that violate product standards. This is done by integrating an autoencoder that provides an intentionally faulty feature map. As the autoencoder cannot reconstruct the logical defects, its output differs from the student.

IV. Experiments

A. Experimental details

We provide the settings for code implementation, training and inference as in Table II.

We shall compare the experimental results of YOLO-V9 and EfficientAD with other state-of-the-art models in their respective domains, namely object detection and anomaly

TABLE II
ENVIRONMENTAL AND HARDWARE CHARACTERISTICS

Unit	Description
Processor	Intel(R) Xeon(R) W-2123 CPU @ 3.60GHz
RAM	Quadro RTX 8000 - 48GB
Operating System	Ubuntu 20.04.6 LTS
Packages	Numpy, Matplotlib, Pytorch, Anomalib

detection. The metrics chosen for the evaluation of YOLO-V9 are mAP50 and mAP50-95, while the overall framework shall be assessed using AUROC and F1 score.

In the first phase, we train 500 epochs for YOLO-V9 with batch-size of 16 and input image size of 640×640.

For anomaly detection, we use the same training schemes as the authors. All models share input image size of 256×256, Further details can be found in the respective papers. For EfficientAD [29], we set the learning rate to 10^{-4} with 10^{-6} weight decay in 70000 epochs. EfficientAD comes with 2 versions: EfficientAD-S and EfficientAD-M. Our choice is EfficientAD-S.

B. Results

The training results of different object detection models the first phase were evaluated and shown in Table III. Moreover, we evaluate and compare the anomaly detection models (see Table IV).

TABLE III
PERFORMANCES OF OBJECT DETECTION MODELS

Model	mAP50	mAP50-95
YOLO-V9	0.995	0.986
Watershed algorithm [32]	0.981	0.062
DETR [33]	0.987	0.812

TABLE IV
PERFORMANCES OF ANOMALY DETECTION MODELS WITHIN PIPELINES

Model	AUROC (%)	F1 score (%)
YOLO-V9 + DRAEM	64.19	53.85
YOLO-V9 + PatchCore	87.28	78.77
YOLO-V9 + EfficientAD-S	**94.37**	**83.39**
Watershed + DRAEM	63.27	51.04
Watershed + PatchCore	92.45	80.54
Watershed + EfficientAD-S	94.21	82.07
DETR + DRAEM	69.83	61.74
DETR + PatchCore	89.11	82.32
DETR + EfficientAD-S	93.95	81.26

As in Table III, YOLO-V9 performs exceptional in the object detection task. Table IV shows the superiority of EfficientAD when compared to the two other models, topping the board in both AUROC and F1 score. This is similar to results of EfficientAD on popular anomaly detection datasets such as MVTEC-AD and VisA [29].

V. CONCLUSIONS

In this paper, we introduce an anomaly detection setting which incorporates real-world conditions of an agricultural product processing line. We also propose a straightforward, easy-to-implement framework to tackle the challenges our setting poses and provide experimental results for different modifications, which shows significant room for further improvements.

Considering limitations, our anomaly detection pipeline is rather simple and can not effectively bridge the gap between object and anomaly detection. The baseline anomaly detection algorithms used are also not designed to handle blurry input data.

REFERENCES

[1] J. Liu, G. Xie, J. Wang, *et al.*, "Deep industrial image anomaly detection: A survey," *Machine Intelligence Research*, 2024.

[2] A. Lazarevic, L. Ertöz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," vol. 3, May 2003. DOI: 10.1137/1.9781611972733.3.

[3] M.-C. Lee, Y. Zhao, A. Wang, *et al.*, "Autoaudit: Mining accounting and time-evolving graphs," Dec. 2020, pp. 950–956. DOI: 10 . 1109 / BigData50022 . 2020 . 9378346.

[4] Y. Zhao, X. Hu, C. Cheng, *et al.*, "Suod: Accelerating large-scale unsupervised heterogeneous outlier detection," Jan. 2021.

[5] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, "Adbench: Anomaly detection benchmark," in *Neural Information Processing Systems (NeurIPS)*, 2022.

[6] V. Zavrtanik, M. Kristan, and D. Skocaj, "Draem - a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 8330–8339.

[7] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Apr. 2021.

[8] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "Simplenet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 402–20 411.

[9] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization," *International Journal of Computer Vision*, vol. 130, no. 4, pp. 947–969, 2022.

[10] V. Zavrtanik, M. Kristan, and D. Skoaj, "Dsr - a dual subspace re-projection network for surface anomaly detection," in *European Conference on Computer Vision*, 2022.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[12] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but differnet: Semi-supervised defect detection with normalizing flows," Aug. 2020.

[13] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, Lille, France: PMLR, Jul. 2015, pp. 1530–1538.

[14] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Fully convolutional cross-scale-flows for image-based defect detection," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1829–1838, Oct. 2021.

[15] Jiawei, Y. Zheng, X. Wang, *et al.*, "Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows," *Computer Vision and Pattern Recognition*, Nov. 2021.

[16] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, *Towards total recall in industrial anomaly detection*, 2021. arXiv: 2106.08265 [cs.CV].

[17] P. Napoletano, "Anomaly detection in nanofibrous materials by cnn-based self-similarity," 2018.

[18] S. Lee, S. Lee, and C. Byung, "Cfa: Coupled-hypersphere-based fwature adaptation for target-oriented anomaly localization," *IEEE Access*, 2020.

[19] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Asymmetric student-teacher networks for industrial anomaly detection," *Machine Learning (cs.LG); Artificial Intelligence (cs.AI); Computer Vision and Pattern Recognition (cs.CV)*, Oct. 2022. DOI: 10.48550/arXiv.2210.07829.

[20] T. Bao, J. Chen, W. Li, *et al.*, "Miad: A maintenance inspection dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 993–1002.

[21] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. Rabiee, "Multiresolution knowledge distillation for anomaly detection," *Computer Vision and Pattern Recognition*, Nov. 2020.

[22] L. Bonfiglioli, M. Toschi, D. Silvestri, N. Fioraio, and D. De Gregorio, "The eyecandies dataset for unsupervised multimodal anomaly detection and localization," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3586–3602.

[23] S. Tang, F. He, X. Huang, and J. Yang, "Online pcb defect detector on a new pcb defect dataset," *arXiv preprint arXiv:1902.06197*, 2019.

[24] M. Wieler and T. Hahn, "Weakly supervised learning for industrial optical inspection," in *DAGM symposium in*, vol. 6, 2007.

[25] P. Bergmann, X. Jin, D. Sattlegger, and C. Steger, "The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization," *arXiv preprint arXiv:2112.09045*, 2021.

[26] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *European Conference on Computer Vision*, Springer, 2022, pp. 392–408.

[27] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, *Yolov9: Learning what you want to learn using programmable gradient information*, 2024. arXiv: 2402.13616.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

[29] K. Batzner, L. Heckler, and R. König, "Efficientad: Accurate visual anomaly detection at millisecond-level latencies," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2024, pp. 128–138.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[31] A. Shrivastava, A. Guvta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on computer vision and pattern recognition (CVPR)*, 2016.

[32] J. Cousty, G. Bertrand, L. Najman, and M. Couprie, "Watershed cuts: Minimum spanning forests and the drop of water principle," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1362–1374, 2009.

[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoryuko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020.

An isolated Vietnamese Sign Language Recognition method using a fusion of Heatmap and Depth information based on Convolutional Neural Networks.

Xuan-Phuoc NGUYEN*, Thi-Huong NGUYEN†, Duc-Tan Tran‡, Tien-Son Bui §, Van-Toi NGUYEN‡,¶

* Faculty of Computer Science, PHENIKAA University, Yen Nghia, Ha Dong, Hanoi 12116, Vietnam
† Hanoi College of High Technology, Hanoi, Vietnam
‡ Faculty of Electrical and Electronic Engineering, PHENIKAA University, Yen Nghia, Ha Dong, Hanoi 12116, Vietnam.
§ Hanoi University of Industry, Vietnam.
¶ Corresponding author. Email: toi.nguyenvan@phenikaa-uni.edu.vn

***Abstract*—In recent years, interest in sign language recognition has continuously increased. However, recognition methods for exploiting the combination of RGB and depth data are limited, especially applied to Vietnamese sign language. This paper presents an isolated Vietnamese sign language recognition method using a novel streams-enhanced 3D ConvNet. The experimental results demonstrate the superiority of the proposed method over other methods using variations from RGB, depth, and RGB-D data. The speed and accuracy of our method are better than those of previous methods.**

***Index Terms*—Keywords: sign language recognition**

I. Introduction

As we know, deaf and hard-of-hearing people use sign language to communicate. They use their upper body to express speech, including hand gestures, facial expressions, and body language. The need for research into sign language recognition stems from its potential to empower the deaf and hard-of-hearing community, facilitating seamless communication with technological interfaces. The development of accurate and efficient algorithms in this field promises to bridge the communication gap, promote communication accessibility for the deaf, and promote inclusivity in technological development.

Sign language recognition presents challenges within computer vision. Existing studies have explored various modalities for feature representation, such as RGB frames [1], [2], [3], optical flows [4], audio waves [5], and human skeletons [6], [7]. Among these modalities, skeleton-based action recognition has received increasing attention in recent years due to its action-focusing nature and compactness.

In practice, human skeletons in a video are mainly represented as a sequence of joint coordinate lists, where pose estimators extract the coordinates. Since only the pose information is included, skeleton sequences capture only action information while being immune to contextual noises, such as background variation and lighting changes. Among all the methods for skeleton-based action recognition [8], [9], [10], graph convolutional networks (GCN) [7] have been one of the most popular approaches. However, these methods are limited in aspects such as: A small perturbation in the coordinates often leads to completely different predictions [11]; difficult to combine with varying modality datasets such as RGB, depth, RGB, optical flows, and skeletons.

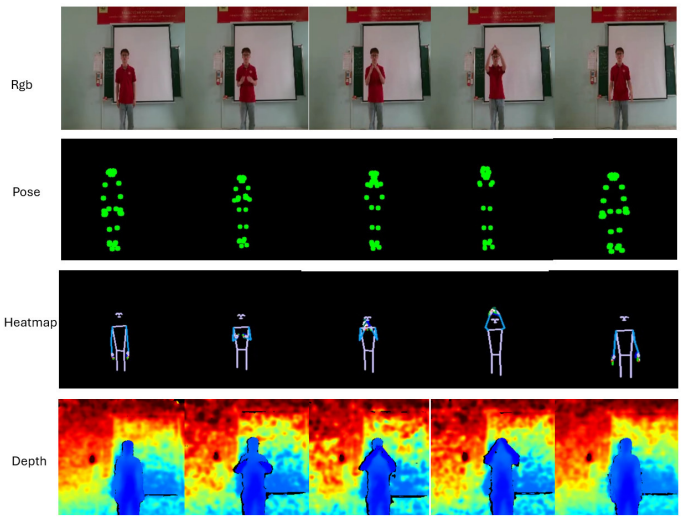


Fig. 1. Pose estimators produce 2D poses that are represented by stacks of heatmaps of skeleton joints.

In this paper, we propose a novel framework, Ad2C (Advantage 2 Convolution), which can solve the difficulties mentioned above. Specifically, Figure 1's present pose estimators produce 2D poses that Ad2C accepts as input. Rather than operating on coordinates on a human skeleton graph, 2D poses are represented by stacks of heatmaps of skeleton joints. A 3D heatmap volume will be created by stacking the heatmaps at various timesteps along the temporal dimension. To identify

actions, Ad2C based on I3D architecture and advandge it, uses two flow 3D-CNN for the 3D heatmap volume and depth data. Table 1 summarizes the key distinctions between Ad2C and the most popular skeleton-based recognition method.

TABLE 1
SUMMARIZES THE KEY DISTINCTIONS BETWEEN AD2C AND GCN-BASED TECHNIQUES.

	skeleton-based recognition	Ad2C
Input	Skeleton	Varied
Format	Coordinates	Heatmap Volumes Depth Volumes
Architecture	GCN	3D-CNN

The weaknesses of some of the methods can be solved with Ad2C. Initially, applying 3D heatmap volumes is kinder to the original posture estimate; our experience indicates that Ad2C performs brilliantly across input acquired by various methods. Furthermore, Ad2C benefits from the latest developments in 3D convolutional neural networks and is simpler to combine other modalities in multi-stream convolutional networks. This feature creates a lot of chances for improvement in terms of recognition performance through advantages architecture. We carry out extensive experiences over a number of our datasets, including RGB, Depth, RGB-D, and heatmap-depth, to confirm the efficacy and efficiency of Ad2C.

The main contributions of our study are folders.

- We propose a new model, called Ad2C, which combines the generated Heatmap-Depth data and gives good performance.
- We experimentally the performance of fusion methods in our dataset, including the evaluation of RGB, Depth, RGB-D, and Heatmap- Depth discrete data.

The remainder of this article is organized as follows. Part II discusses our related work. Section III presents details of our proposed method. Part IV describes the experiment result. Finally, there are conclusions and future development directions in Part V. Our code is available at https://github.com/phu2007vis/FeaturesFusionRGBDepth

II. RELATED WORK

I3D. [41] The video classification model, Two-Stream 3D ConvNets, is a published basic method. This method combines two I3D networks, one trained on RGB inputs and another on optical flow inputs. The two streams capture complementary information about the appearance and motion of the video frames. The I3D networks are based on inflating the filters and pooling kernels of the Inception v1 model into 3D, allowing them to leverage the ImageNet pre-trained weights and architectures. The two-stream I3D model achieves state-of-the-art performance on several action recognition benchmarks, such as UCF-101 and HMDB-51, after pre-training on the Kinetics dataset [1]. I3D was initially applied to Kinetics data about activities. However, I3D was later applied to other datasets on sign language recognition [38], [39], [40]

S3D. [42] S3D (Separable 3D Convolutional Neural Networks) is a variant of I3D that aims to improve computational efficiency while maintaining high performance in video understanding tasks. S3D is designed to reduce the computational cost of 3D convolutions by decomposing them into separate spatial and temporal convolutions. This reduces the number of parameters and increases efficiency without significantly sacrificing performance. By leveraging separable convolutions, S3D models can process videos faster and with less computational resources compared to traditional 3D ConvNets.

Applications in Sign Language Recognition: S3D's efficiency makes it suitable for applications where real-time processing is crucial. Its ability to handle spatio-temporal data effectively helps in capturing the dynamic nature of sign language gestures. S3D has been used in various research works involving large-scale sign language datasets, helping to achieve high accuracy in recognizing complex gestures and movements.

3D-CNN for RGB-based action recognition. In applied to learning spatial features to be included in videos, 3D-CNN is a natural expansion of 2D-CNN. Action recognition has continuously made use of it [12] [2]. Due to its numerous parameters, 3D-CNN requires an enormous number of videos to learn suitable representation. After I3D [1], 3D-CNN has emerged as the mainstream method for action recognition. Subsequently, the action recognition community presented various advanced 3D-CNN architectures [13], [14], [15], [16], which exceed I3D with regard to accuracy and efficiency.

GCN for skeleton-based action recognition. In skeleton-based action recognition, graph convolutional networks are frequently used [17], [18], [19], [20], [21], [7]. It represents sequences of the human skeleton as spatiotemporal graphs. A popular baseline for GCN-based techniques is ST-GCN [7], which combines interleaving temporal convolutions with spatial graph convolutions for spatiotemporal modeling. While self-attention mechanisms enhance the modeling capacity [24], [25], adjacency powering is used for multiscale modeling upon the baseline [23], [22]. Although GCN has achieved considerable success in skeleton-based action recognition, its scalability and reliability are restricted [11]. Furthermore, careful design may be required for GCN-based systems that fuse features from skeletons and other modalities [26].

CNN for skeleton-based action recognition. Convolutional neural networks are used in another line of research for skeleton-based action recognition. 2D-CNN-based methods start by modeling skeleton sequences using manually produced modifications to create a pseudo-image. Heatmaps are aggregated along the time dimension in one line of work [27], into a 2D input with color encodings, or learned modules [28], [29]. Despite meticulous design, information loss still happens during aggregation, resulting in subpar recognition performance. By using transformations, the coordinates in a skeleton sequence can be immediately converted to a pseudo image in other works [30], [31], [32], [33], [34]. This usually results in a 2D input with the shape K ×T, where K represents the number of joints and T is its temporal length.

These approaches are not as competitive as GCN on widely used benchmarks because such input cannot make use of

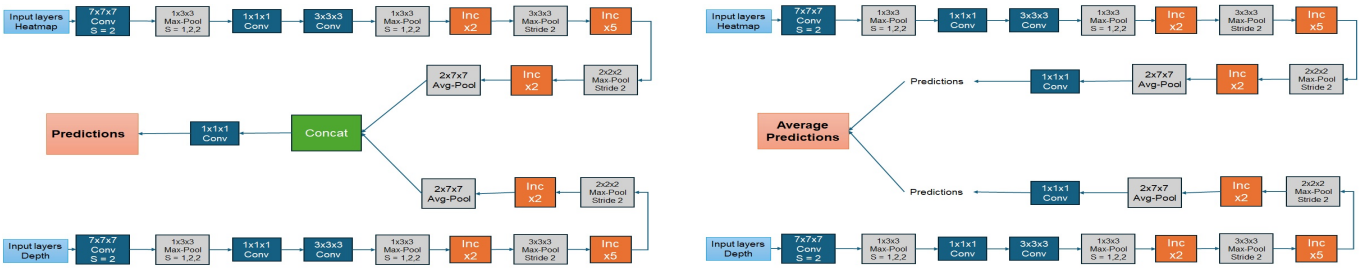


Fig. 2. Ad2C framework

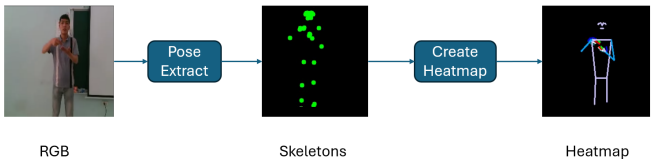


Fig. 3. The process of creating heat maps from RGB images.

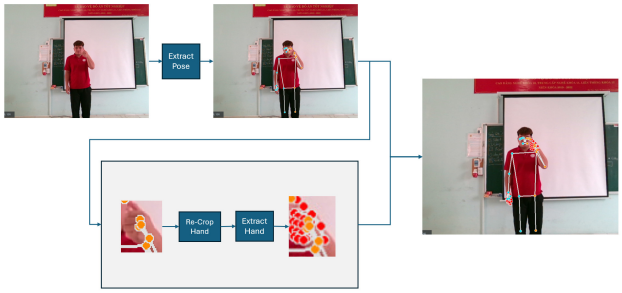


Fig. 4. Overlay joint or limb heatmaps along the time dimension and apply preprocessing that creates the heatmap volume.

the locality nature of convolution networks [30]. 3D-CNNs have only been used in a small number of earlier studies for skeleton-based action recognition. They either directly sum up the 3D skeletons into cuboids [37] or stack the pseudo images of distance matrices [36], [35] to create the 3D input. These methods likewise experience significant information loss and achieve performance that is much below the state-of-the-art.

III. PROPOSED METHOD

A. Ad2C architecture

We propose Ad2C structural designs, an overview of the Ad2C architecture is depicted in (Figure 2). Ad2C is the approach of a 3D-CNN for action recognition, which involves improving and combining two CNN networks based on the I3D structure. We further improved the model by splitting the Heat Map and Depth data into two separate training streams. Each stream is passed through Conv3D layers to extract features before combining. The combination of information can then be fed into another Conv3D layer for further processing to enhance the feature combination from the previous layer or no processing may be required. Additionally, we also train a 3D-CNN enhancement that applies a combination of Heatmap and Depth data without going through two streams. However, this test gives results that are not as feasible as the previous method. The variant method combines two CNN networks with Heatmap and Depth input layers to demonstrate the flexible interaction capabilities of Ad2C.

B. Training Objective

We start by looking at skeleton extraction, which is the basis of skeleton-based action recognition. We point out some

aspects that need to be considered such as choosing a skeleton extraction tool and promoting using 2D skeletons. Therefore, we introduce the Heatmap Volume (Figure 3), representing the 2D skeleton sequences used in the training model. In terms of storing estimated heatmaps, they are typically saved as a set of three coordinates (x;y;c), where c marks the maximum point of the heatmap and (x;y) is the coordinate corresponding to c. In our experiment, we found that the triple coordinates (x; y; c) save a large amount of storage space without sacrificing performance.

The training objective is a multi-task optimization problem. Let s_i be the maximum IoU score between the i-th default span and any ground truth span, and let $x_{ij}^k = \{1, 0\}$ be an indicator for matching the i-th default span to the j-th ground truth span of category $k \in [1, K]$. The total objective loss function consists of three weighted components: the localization loss (loc), the class confidence loss (conf), and the activity confidence loss (act):

$$\text{Loss} = L_{loc}(x, t, g) + \alpha L_{conf}(x, c) + \beta L_{act}(s, c) \quad (1)$$

where α and β are the weight terms used to balance each component of the loss function, the localization loss is a Smooth L1 loss between the predicted temporal offsets (t) and the ground truth span parameters (g). The class confidence loss (c) is a softmax loss over multiple class confidences. The activity confidence loss is a binary classification loss using sigmoid cross-entropy.

C. Dataset colection

Our dataset is captured by RGB-D cameras (Intel RealSense Depth Camera D456) with the following key specifications:

- Depth technology: Stereoscopic;
- RGB sensor technology: Global shutter;
- RGB and depth frame rate: 10-30fps;

- RGB and depth resolution: 1280 x 720;
- RGB sensor resolution: 1MP.

We collected a dataset from 110 volunteers who, although not native sign language users, learned the movements by following instructional videos provided by the Deaf-Mute Council. The dataset includes 120 isolated words, each performed by a minimum of 20 individuals. Detailed information can be found in Table 2. The volunteers' gestures were captured at an angle achieved by placing the camera facing the person at a distance of about 125cm to record the person's upper body gesture. Volunteers stood about 40cm away from the background to separate people from the background by depth easily. The recording occurred in normal lighting conditions (office), and the volunteers wore clothes freely.

TABLE 2
DESCRIPTION OF DATASET

Videos format	AVI
Resolution	640 x 480
Frames per second	30
Videos after preprocessing	18948
Signers	110
Words	120
Videos for each word	150-210
Average time for each video	03.21 second

D. Proposed setting

- *Data splitting and preprocessing*

Data splitting: We present an analysis of a dataset from our collection, reflecting real-world scenarios. This dataset includes 120 isolation signs, each performed by 20 volunteers, each volunteer performing 10 times per sign. For the Ad2C model, we split the dataset into three sets: training, validation, and testing, based on the 20 signers for a sign. This process ends with a total of 16x120 signers for training, 2x120 signers for validation, and 2x120 signers for testing; this corresponds to 80% for training, 10% for validation, and 10% for testing. The number of videos per set is as follows: the training set has 18948 videos, the validation set has 2369 videos, and the testing set has 2369 videos. The dataset is divided in this manner for both RGB and depth data.

Preprocessing:

Heatmap inputs: Use RGB data to convert into Heatmaps through skeleton extraction. In this way, we have eliminated environmental factors such as lighting, background, or blind spots in the image.

First, We resize the frame to 250 x 250 pixels and apply a center crop to 224 x 224 pixels. We resize the frame to 250 x 250 pixels and apply a center crop to 224 x 224 pixels. For data augmentation, we apply random rotations between -15 degrees and 15 degrees and adjust brightness by up to 10%. Additionally, the data undergoes horizontal and vertical translation. This augmentation is suitable because, in real-world scenarios, people may stand off center, and the camera setup may experience rotations within this range.

Then extracting the skeleton, we use the Holistic Pose method of media pipe. The two extraction stages are shown in Figure 4:

- Stage 1: preliminary extraction of the body and face skeleton.
- Stage 2: focus on each area of the hands.

Because the size of the body and face is large, we can detect it directly from large images. On the contrary, the fingers in the hand are much smaller but the information about the movements is very huge. Therefore, we will approximately cut the hand based on the previously detected wrist points and include them in the hand detection model for more accurate extraction. There will be 52 points in total

- 32 face and body points
- 20 finger points

Heatmap is created from line segments with skeleton points corresponding to accompanying colors. Heatmaps help the model learn better and more naturally with spatial information like RGB and the same skeleton- avoiding being influenced by the environment's same skeletons. Sign language recognition is greatly influenced by the fingers and wrist. Therefore, we use the colors of different wrist and finger segments, including 5 creating colors: green, white, dark blue, red, and light green. To better represent the data about the relative position of the hand and body parts, we used light purple for the body and blue for the arms. The head is quite far from the body, so we chose a color similar to the body color. The foot part carries no information, so it can be omitted.

Depth inputs: To enable the integration of Depth data with Heatmap data, resizing to 224 x 224 is necessary. Given that Depth data values range up to 1300 (equivalent to 130 cm), the process involves normalization and resizing. Initially, the Depth data is normalized to the range of 0-1. Subsequently, to fit within the 0-255 range typical for images, the values are scaled by multiplying by 255. Following this normalization step, standard resizing techniques are applied. Once the Depth data has been resized to the desired dimensions, normalization to the range of 0-1 is necessary. This method ensures that the depth values retain meaningful depth information throughout the process.

Time augmentation: In practical scenarios, sign language practitioners do not consistently execute gestures at standardized speeds. Thus, the input data necessitates speed variation achieved through random frame removal or repetition, while maintaining chronological order. This approach ensures the dataset's realism and enhances its robustness for model training. Furthermore, in practical scenarios, gestures are sometimes not fully executed to swiftly express language concepts. To equip the model with the ability to detect signs even in such instances, we initially segmented the video into seven parts, randomly selecting start indices from the first three segments and end indices from the last three. The rationale behind this data augmentation strategy stems from the observation that signers often abbreviate the beginning or end of gestures to increase communication speed while ensuring that the

intended meaning remains comprehensible to the interlocutor. Before temporal data augmentation, Heatmap and Depth data are combined to ensure that each frame corresponds to its respective pair.

IV. EXPERIMENTAL RESULTS

A. Experiment

As illustrated in the proposed settings, different colors are assigned to various wrist and finger segments for clarity. These include green, white, dark blue, red, and light blue. To accurately depict the relative positions of hands and body parts, light purple is used for the body and blue for the arms. Given that the head is positioned relatively far from the body, a color similar to the body color is chosen. The footer part does not convey any significant information and can be disregarded. The outcomes of these settings are shown in Figure 5.

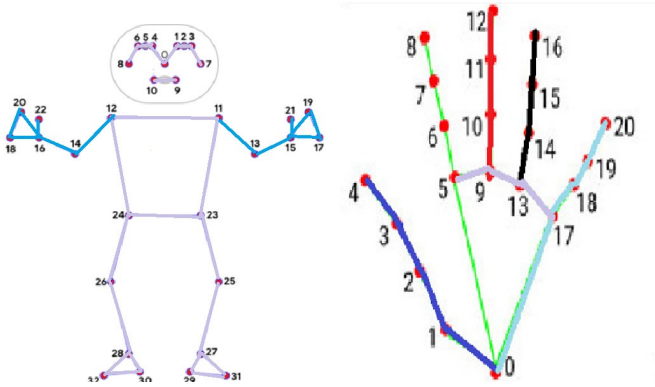


Fig. 5. Heatmap line segments with multiple colors

To achieve effective model training, we employed the RTX 3090 24GB hardware platform, utilizing a batch size of 80 and executing the process for 300 epochs. We also implemented Weight Decay with a rate of 0.00005 and Dropout with a value of 0.3 in the final Conv3D layer to mitigate overfitting. Additionally, each training sample comprised 72 frames. Due to the significantly larger parameter count, we initially disabled the Heatmap branch to train the Depth branch with a learning rate of 0.001 over 100 epochs. Subsequently, we enabled all layers of the model and continued training with a reduced learning rate below 0.0005.

When testing improvements in two streams of I3D for our Ad2C model, we tested the use of an Early Fusion model (which combines Heatmap and Depth data immediately after the input layers). With this early merging, training only takes place on one 3D-Conv stream. We used this model to pre-train the Kinetics dataset and fine-tune all its layers. For both Mid-Fusion (Ad2C architecture) and Late-Fusion (combines two streams at the end), we also used the Kinetics dataset to conduct training for both branches of 3D-Conv and fine-tune its inner layers. For these experiments, we proposed the improved model Ad2C as Mid-Fusion and started the experiments with our Heatmap-Depth dataset.

Finally, we start testing Ad2C and compare the training results of Ad2C on different input data types of our dataset. We tested Ad2C with Heatmap-Depth and RGB-D inputs. The results show that Heatmap-Depth input has better accuracy than RGB-D on the Ad2C model. To further demonstrate the superiority of Ad2C, we continue to compare the accuracy of the Ad2C model with the original I3D on the same RGB-D input data type. Because the original I3D only has 1 stream, to be able to use RGB-D, we must use the early fusion method (after the input layer). We also tested I3D with discrete data types including RGB and Heatmap. We presented and analyzed the test results in detail below.

B. Results

Table 2 presents the performance comparison of our dataset. From the table, it is clear that the model performance varies significantly depending on the input data type and the model type.

TABLE 3
MODEL PERFORMANCE

Model	Input	Accuracy	mAP	Validation Loss
I3D	RGB	59.35%	0.581	1.65
	Heatmap	53.33%	0.548	1.85
	RGB-D	64.21%	0.633	1.52
Ad2C	RGB-D	79.80%	0.804	0.77
	Heatmap- Depth	**80.15%**	**0.825**	**0.74**

When the I3D model is trained with discrete data (RGB or Heatmap), accuracy is higher with RGB data (59.53%) than with Heatmap (53.33%). This shows that using RGB can higher model performance in terms of accuracy. Additionally, mAP is also quite large in RGB (0.581) compared to Heatmap (0.548), indicating better performance when used only on RGB data. However, in these cases, Validation Loss is still large, with RGB and Heatmap data being 1.65 and 1.85 respectively.

Interestingly, when the I3D model is trained with RGB-D data, combining both RGB and Depth data, the accuracy is higher (64.21%). This shows that combining different types of data can improve model performance. mAP is approximately equivalent (0.633) to that in discrete RGB (0.581), suggesting a trade-off between precision and mAP. Validation loss is still large, although it is slightly reduced compared to RGB or Heatmap data but still at the threshold of 1.52

With the advanced model Ad2C, using a combination of data types is necessary. Therefore, Ad2C only uses input layers of RGB-D or variants such as Heatmap-Depth without using separate data. Comparing Ad2C and I3D on the same RGB-D input data set, it is easy to see that Ad2C's accuracy is much better (79.80%) than (64.21%). Ad2C continues to improve when using the Heatmap-Depth input layer, and this model's accuracy continues to increase from (79.80%) to (80.15%). More specifically, in the case of using the Ad2C model, the Validation Loss of the model has decreased significantly. When Ad2C trains on the RGB-D dataset, the Loss is at 0.77 and this number continues to decrease to 0.74 when using Heatmap-Depth as the input layers

Table 3 compares different data-trained Ad2C networks regarding the number of parameters, input frame sizes, and inference time on a machine with GPU RTX 3090 24GB. From the table, we can observe that the inference time on the Ad2C network increases significantly with 72 input frames sized 224 x 224.

TABLE 4
NUMBER OF PARAMETERS, INPUT SIZES OF THE MODELS, AND INFERENCE TIME.

Network	Input data	Number of parameters	GPU inference time
I3D	RGB	12.06M	12.36(ms)
I3D	RGB-D	24.77M	51.39(ms)
Ad2C	Heatmap-Depth	25.09M	51.11(ms)

The results presented in Table 3 show that input data type and model structure can significantly impact the performance of networks on the same datasets we collected. However, there seems to be a trade-off between accuracy and mAP on RGBD and Heatmap-Depth data that results in improved performance. Further research is needed to explore these relationships in more detail and optimize model performance under different conditions.

V. CONCLUSION AND FUTURE WORK

In our research, we improve current modern recognition methods to produce better performance and quality than previously published methods based on the heatmap-depth dataset. The recognition method is suitable for device conditions, does not consume resources, and runs well on GPU and in real-time. Our experimental implementation yields good, scalable results on larger datasets. The effectiveness of the dataset and training method has been confirmed by multiple experiments.

Our next research work in the future is to continue improving training methods to achieve higher performance. This requires meticulous methodological research to deliver our model suite of the highest quality.

ACKNOWLEDGMENT

We extend our gratitude to the volunteers for their contributions to data collection and to Phenikaa University's High-Performance Computing System (PHPC) for processing the data.

REFERENCES

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, pages 6299–6308, 2017. 1, 2, 4, 5, 9, 12, 13

[2] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In ICCV, 2015. 1, 2, 4, 10

[3] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In ECCV, 2016. 1, 3

[4] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. arXiv:1406.2199, 2014. 1

[5] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. arXiv:2001.08740, 2020. 1

[6] Philippe Weinzaepfel and Gregory Rogez. Mimetics: To- ´ wards understanding human actions out of context. IJCV, 129(5):1675–1690, 2021. 1

[7] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In AAAI, volume 32, 2018. 1, 2, 3, 5, 7, 14

[8] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In CVPR, pages 1110–1118, 2015. 1

[9] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In CVPR, pages 588–595, 2014. 1

[10] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In CVPR, pages 1290–1297. IEEE, 2012. 1

[11] Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In KDD, pages 1399–1407, 2019. 2

[12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. TPAMI, 35(1):221–231, 2012. 2

[13] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In CVPR, pages 203–213, 2020. 2, 4, 10

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In ICCV, pages 6202–6211, 2019. 2, 3, 4, 5, 10, 13

[15] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In ICCV, pages 5552–5561, 2019. 2

[16] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In CVPR, pages 6450– 6459, 2018. 2

[17] Jinmiao Cai, Nianjuan Jiang, Xiaoguang Han, Kui Jia, and Jiangbo Lu. Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition. In WACV, pages 2735–2744, 2021. 2

[18] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In ICCV, pages 13359–13368, 2021. 2

[19] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla. Quo vadis, skeleton action recognition? IJCV, 129(7):2097–2112, 2021. 2

[20] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In MM, pages 1625–1633, 2020. 2

[21] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. arXiv:2106.15125, 2021. 2

[22] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In CVPR, 2019. 2, 7

[23] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In CVPR, 2020. 2, 5, 7, 11, 13

[24] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. Spatio-temporal graph routing for skeleton-based action recognition. In AAAI, volume 33, pages 8561–8568, 2019. 2

[25] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Twostream adaptive graph convolutional networks for skeletonbased action recognition. In CVPR, pages 12026–12035, 2019. 2, 7, 8

[26] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In ECCV, pages 72–90. Springer, 2020. 2

[27] Vasileios Choutas, Philippe Weinzaepfel, Jer ´ ome Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In CVPR, pages 7024–7033, 2018. 2, 8, 12, 13

[28] Sadjad Asghari-Esfeden, Mario Sznaier, and Octavia Camps. Dynamic motion representation for human action recognition. In WACV, pages 557–566, 2020. 2

[29] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Poseaction 3d machine for video recognition. In CVPR, pages 7922–7931, 2019. 2, 8, 12, 13

[30] Carlos Caetano, Jessica Sena, Franc¸ois Bremond, Jefers- ´ son A Dos Santos, and William Robson Schwartz. Skelemotion: A new represen-

tation of skeleton joint sequences based on motion information for 3d action recognition. In AVSS, pages 1–8. IEEE, 2019. 2

[31] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In CVPR, pages 13289–13299, 2020. 2

[32] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In CVPR, 2017. 2

[33] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Cooccurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. arXiv:1804.06055, 2018. 2

[34] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In CVPR, pages 5137–5146, 2018. 2

[35] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017. 12

[36] Zeyi Lin, Wei Zhang, Xiaoming Deng, Cuixia Ma, and Hongan Wang. Image-based pose representation for action recognition and hand gesture recognition. In FG, pages 532– 539. IEEE, 2020. 2

[37] Hong Liu, Juanhui Tu, and Mengyuan Liu. Two-stream 3d convolutional neural network for skeleton-based action recognition. arXiv:1705.08106, 2017. 2

[38] Wong, Ryan, Necati Cihan Camgöz, and Richard Bowden. "Hierarchical i3d for sign spotting." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[39] Maruyama, M., et al. "Word-level sign language recognition with multi-stream neural networks focusing on local regions. arXiv 2021." arXiv preprint arXiv:2106.15989.

[40] Zhu, Yu, and Tiantian Yuan. "Continuous Sign Language Recognition Based on I3D-TCP Network Structure." 2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE). IEEE, 2023.

[41] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[42] Zhang, Da, et al. "S3d: single shot multi-span detector via fully 3d convolutional networks." arXiv preprint arXiv:1807.08069 (2018).

GILED: Lesion Detection of Gastrointestinal Tract from Endoscopic Images and Medical Notes

Vu-An Hoang*, Minh-Hanh Tran*, Viet Hang Dao†‡, Thanh-Hai Tran*
* School of Electrical and Electronics Engineering
Hanoi University of Science and Technology, Hanoi, Vietnam
† Hanoi Medical University Hospital, Hanoi, Vietnam
‡Institute of Gastroenterology and Hepatology, Hanoi, Vietnam
E-mail: hai.tranthithanh1@hust.edu.vn

***Abstract*—Existing methods for lesion detection in medical images typically use images as the primary training source for deep models. However, in addition to annotated images, the doctor provides his observations and conclusions about the lesion in a medical note, which remains underutilized. This paper presents a method for automatic detection of gastrointestinal tract lesions, named GILED, which combines endoscopic images and medical notes. Specifically, we leverage the medical notes accompanying each medical image of a patient to enhance the representation of the image encoder. Once pre-trained on an image-text dataset, this enhanced image encoder can be integrated into any model for detection and segmentation. In our work, we deploy the image encoder in the YOLOv8 model for the lesion detection task. To demonstrate the effectiveness of this image representation learning approach, we conduct experiments on a benchmark dataset Kvasir-SEG and two self-collected datasets. Thank to the integration of expert knowledge from medical notes, GILED achieves higher performance even with limited image-text data sources. On the Kvasir-SEG dataset, compared to the original Yolov8, GILED improved the mAP50 from 91.5% to 92.8%; on our datasets, the mAP50 increased from 83.6% to 85.1% and 53.0% to 54.0%. It shows the potential to improve lesion detection using additional medical notes.**

I. Introduction

Gastrointestinal tract diseases are prevalent worldwide. According to GLOBOCAN, in 2022 there were nearly 970,000 new cases of stomach cancer and more than 510,000 new cases of esophageal cancer globally [1]. In Asia, Vietnam ranks 6th in stomach cancer cases and 7th in esophageal cancer cases. A 2019 survey indicated that endoscopy centers in Vietnam had to handle over 300 cases daily. Due to hospital overload and the inexperience of doctors at remote hospitals, the rate of missed diagnoses is high, placing a significant burden on the healthcare sector. As a result, implementing automatic detection tools is essential to reduce overall treatment and healthcare costs.

A conventional approach for lesion detection and segmentation in endoscopic images is to use deep learning models, which usually require a large amount of labeled data for training [2] [3]. However, data annotation is time-consuming and costly, as it necessitates the expertise of specialist doctors. Self-supervised learning, which uses image pairs as input, is an alternative method [4]. While these methods have achieved impressive results, many do not leverage the rich sources of medical notes alongside the image data. With the emergence of large language models, combining images and text now offers improved efficiency in many chest X-ray image analysis tasks [5] [6].

In this paper, we propose a new method called GILED (**G**astro**I**ntestinal Tract **Le**sion **D**etection) for detecting lesions in the gastrointestinal tract by combining endoscopic images with medical notes provided by doctors. Specifically, we utilize the medical notes associated with each patient's medical image to enhance the representation of the image encoder. Once pre-trained on a sufficient image-text dataset, this enhanced encoder can be integrated into any model for detection and segmentation. In our work, we deploy the image encoder in the YOLOv8 model for lesion detection. This method draws inspiration from the MCGA framework proposed by Wang et al. [5] for dealing with text and X-ray images of chests. Our work differs in three key aspects. First, we employ a more efficient encoder in the YOLOv8 model to learn the image representation. Second, we adapt the framework for endoscopic images, which have different image and text characteristics. Finally, we investigate various strategies to train the model for the specific downstream task (lesion detection).

To demonstrate the effectiveness of this image representation learning approach, we train an encoder from image-text data, followed by transfer learning to YOLOv8 on the Kvasir-SEG dataset [7] and our two self-collected datasets. On the Kvasir-SEG dataset, the mAP50 increased from 91.5% to 92.8%; on the IGHEndoLesion-SEG-Cancer dataset, the result increased from 83.6% to 85.1%; and on the IGHEndoLesion-SEG, the result increased from 53.0% to 54.0%. The remainder of this paper is structured as follows: In Section II, we present related works. Section III describes our datasets and proposed frameworks. We analyze experimental results in Section IV. Section V concludes the paper and provides ideas for future work.

II. Related Work

A. Vision-and-Language Pre-training (VLP)

Recognizing the limitations of learning in a single modality and optimally leveraging the relationship between image and text data, VLP has garnered widespread attention and interest [8] [9] . Among these, the CLIP model [10] stands out as a breakthrough in integrating vision and language modalities.

The highlight of CLIP is its ability to link and understand both images and text by naturally approaching the problem of finding their similarity. Additionally, methods such as ALIGN [11], LiT [12], and FLIP [13] have demonstrated significant progress in improving joint representation learning for images and natural language, despite requiring substantial training data and computational resources [14], [15]. Therefore, the method of freezing the language model has been introduced to save costs and reduce computational load [16], [17], and has shown outstanding results in visual question answering tasks.

B. Medical Vision-and-Language Pre-Training (Med-VLP)

In the medical field, there is still a significant lack of research due to the complexity of medical text and the limited availability of large-scale paired medical image-text datasets. Previous models such as ConVIRT [18] used contrastive learning methods to improve representation and understanding of medical images, while GLoRIA [19] employed a global-local representation framework to efficiently identify medical images with minimal labeled data. More recently, MGCA [5] used fine-grained disease-level annotations for alignment, and MedKLIP [20] adopted a method of predicting masked values based on available information. Newer models based on pre-trained MGCA, such as M-FLAG [6], utilized language freezing and latent space optimization, and Med-UniC [21] minimized biases and improved model generalization on a multilingual platform.

III. Materials and Proposed Framework

A. Datasets

Regarding the analysis of endoscopic images, there is a lack of public datasets due to privacy concerns. Most researchers evaluate their work using datasets such as Kvasir [22] and ClinicDB [23]. However, these datasets do not include accompanying texts. Consequently, to the best of our knowledge, there is no existing work that integrates text analysis with endoscopic image analysis. In this work, we have built a multimodal dataset (text and image) called IGHEndoLesion-Image-Report to train the image encoder. Additionally, we train and test our lesion detection model on three other datasets, including the Kvasir-SEG dataset and two self-collected datasets (IGHEndoLesion-SEG and IGHEndoLesion-SEG-Cancer).

1) IGHEndoLesion-Image-Report dataset: This dataset includes 2,667 image-report pairs, with 1,383 related to esophageal cancer and 1,284 related to stomach cancer, where each patient may have more than one image-report pair. Each image is accompanied with a detailed report on the patient's condition, provided by the doctors who directly performed the test. These reports contain important information about the diagnosis and severity of the injury. Each report includes three main descriptive pieces of information: characteristics at the cancer site; detailed descriptions of lesions such as size, shape, and location; and annotations about the disease status. Specifically, the reports on stomach cancer have an average length of 232 characters per report and describe features such as gastric fluid, mucosa, curvature, antrum, and cardia. The reports on esophageal cancer have an average length of 218 characters per report and describe features such as veins, esophageal lumen, Z-line (the mucosal transition zone), and mucosa. Medical images and reports are provided by the Institute of Gastroenterology and Hepatology - Hanoi Medical University. It is important to note that, at the time of writing this paper, we have only been able to collect data on two severe diseases—esophageal cancer and stomach cancer—due to difficulties in accessing medical data in hospitals.

As the medical report is provided under PDF format, we can extract related disease text while removing all personal and unnecessary information. We employ a Vietnamese-English translator given by [24] to convert all medical texts into English.

2) IGHEndoLesion-SEG dataset: This dataset contains 4161 images of six lesion categories, including Esophagitis, HP-negative gastritis, HP-positive gastritis, Esophageal cancer, Stomach cancer, and Duodenal ulcer. The number of images in each category is quite imbalanced. It is provided by the Institute of Gastroenterology and Hepatology at Hanoi Medical University and labeled at the pixel level by doctors with over five years of experience. This dataset is utilized to validate the detector model for the lesion detection task.

3) IGHEndoLesion-SEG-Cancer dataset: This dataset is derived from the IGHEndoLesion-SEG dataset, exclusively featuring images related to Esophageal cancer and Stomach cancer. Since the IGHEndoLesion-Image-Report contains only samples of esophageal cancer and stomach cancer, the image encoder will be trained solely on the characteristics of these cancers. Our objective is to investigate how the performance of the encoder will be when it is used exclusively for cancer detection compared to detecting a variety of diseases as in the IGHEndoLesion-SEG dataset.

4) Kvasir-SEG: The final dataset we aim to experiment with is the Kvasir-SEG benchmark dataset, which includes 1,000 images containing polyps. Kvasir-SEG is widely used in the literature to evaluate polyp detection and segmentation models. Experimentation on this dataset will be conducted to test cross-domain learning. Table I summarizes the datasets used in this study. Medical records in the IGHEndoLesion-Image-Report dataset are extracted from Portable Document Format (PDF) and translated using VinAI_Translator [24]. IGHEndoLesion-SEG and Kvasir-SEG datasets with mask labels are converted to bounding box labels.

B. Overall framework

In this paper, we propose a framework that consists of three stages:

- Image representation training: this stage trains an image encoder from a text-image dataset. It aims to build an image representation that explores the textual information accompanying images.
- Detector training: this stage utilizes the pre-trained encoder from the pre-training stage, and continues to train the model using datasets of images for a specific task (i.e. lesion detection).

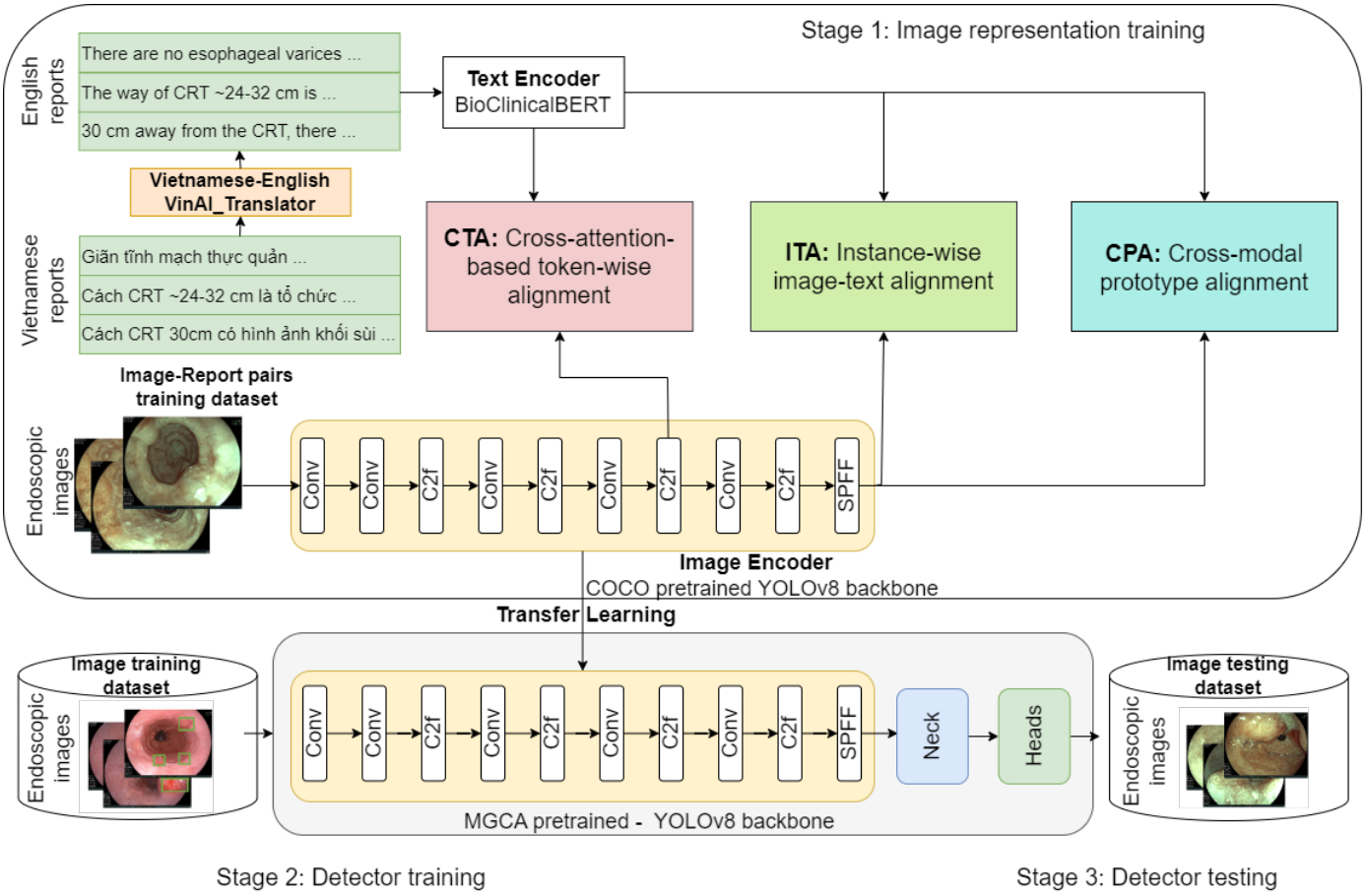


Fig. 1. Proposed framework for lesion detection of the gastrointestinal tract from endoscopic images and medical notes.

TABLE I
SUMMARY OF EXPERIMENTED DATASETS

Dataset	Lesion type	#Images	Split ratio	Data type
IGHEndo Lesion-Image-Report	Esophageal cancer	1383	8:2	Image (1280x995) Medical Report
	Stomach cancer	1284		
	Total	**2667**		
IGHEndo Lesion-SEG	Esophagitis	1421	8:2	Image (1280x995)
	Stomach cancer	426		
	Hp-positive gastritis	729		
	Duodenal ulcer	746		
	HP-negative gastritis	410		
	Esophageal cancer	429		
	Total	**4161**		
IGHEndo Lesion-SEG-Cancer	Stomach cancer	426	8:2	Image (1280x995)
	Esophageal cancer	429		
	Total	**855**		
Kvasir-SEG	Polyp	1000	8:2	Image (1280x995)

- Detector testing: this stage evaluates the model's performance on a testing dataset of images. We investigate various training strategies to reveal the best model for a given task.

The framework is illustrated in Fig.1. We describe each stage in detail in the following subsections.

1) Image representation training: At this stage, we train an image representation that combines both text and image inputs to integrate the semantics of both modalities. As highlighted by Wang et al. [5], medical images and reports exhibit a multi-granularity property, with semantic correspondences at three levels: pathological region-level, instance-level, and disease-level. Consequently, we aim to build the image representation using the MGCA framework [5] to fully exploit these semantic correspondences. It is important to note that the original MGCA was trained on medical images and radiology reports, which have different characteristics compared to endoscopic data.

Given a set of N samples of image-report pairs $\mathcal{D} = \{(x_{i,1}, x_{t,1}), (x_{i,2}, x_{t,2}), ..., (x_{i,1}, x_{t,N})\}$. We employ YOLOv8 backbone as image encoder f_i and BioClinicalBERT [25] as text encoder f_t. These encoders map $\mathcal{D}$ into a latent space $\mathcal{S} = \{(\mathbf{i}_1, \mathbf{t}_1), (\mathbf{i}_2, \mathbf{t}_2), ..., (\mathbf{i}_N, \mathbf{t}_N)\}$ where $\mathbf{i}_k = f_i(x_{i,k})$, $\mathbf{t}_k = f_t(x_{t,k})$, $k \in [1, N]$. The image encoder generates a global representation $\mathbf{i}_k$ and a sequence of visual tokens $\mathbf{R}_k = \{\mathbf{r}_k^1, \mathbf{r}_k^2, ..., \mathbf{r}_k^S\}$. Similarly, the text encoder generates a global representation $\mathbf{t}_k$ and a sequence of textual tokens $\mathbf{Z}_k = \{\mathbf{z}_k^1, \mathbf{z}_k^2, ..., \mathbf{z}_k^L\}$. The pre-training stage composes of three levels of granularity:

- At instance level: The ITA (instance-wise image text alignment) module is implemented to optimize the agreement between true image-report pairs $(\mathbf{i}_k, \mathbf{t}_k)$ versus random pairs. First both globale visual vector $\mathbf{i}_k$ and $\mathbf{t}_k$ are projected into a lower dimensional space using non-

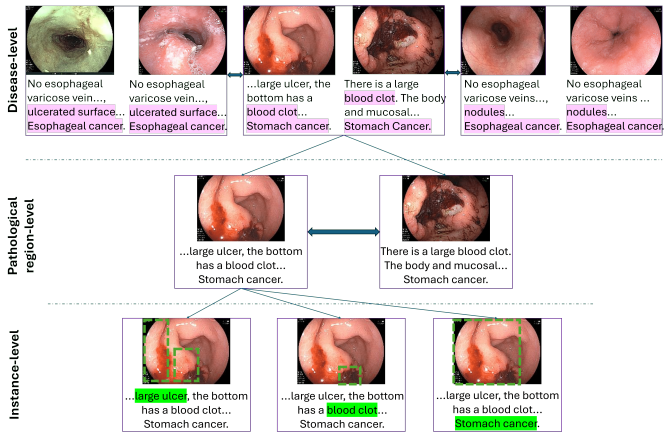


Fig. 2. Three levels of semantic correspondences between image and text. From top to bottom: disease level, pathological region level, and instance level.

linear mappings. Then the similiarity of two new features vectors $\tilde{\mathbf{i}}_k$ and $\tilde{\mathbf{t}}_k$ are computed to produce two contrastive image-to-text l_k^{i2t} and text-to-image l_k^{t2i} losses. The ITA module loss function is described as eq.(1):

$$\mathcal{L}_{ITA} = \frac{1}{2N}\sum_{k=1}^{N}(l_k^{i2t} + l_k^{t2i}) \tag{1}$$

- At pathological level: The CTA (Cross-attention based Token-wise Alignment) module is implemented to match and align the cross-modal local presentations between medical images and reports based on cross-attention mechanism. Specifically, the sequences of visual tokens $\mathbf{R}_k$ and textual tokens $\mathbf{Z}_k$ are projected in lower dimensional spaces. Then the soft matching between the generated visual and text tokens is computed with cross-attention mechanisms. The loss function in the CTA module is composed of Image-To-Text Alignment loss L_{LIA} and Local Text-to-image alignment loss L_{LTA} as in eq.(2):

$$\mathcal{L}_{CTA} = \frac{1}{2}(L_{LIA} + L_{LTA}) \tag{2}$$

- At disease level: The CPA (Cross-Modal Prototype Alignment) module is implemented to leverage the cross-modal correspondences between medical images and reports. From $\tilde{\mathbf{i}}_k$ and $\tilde{\mathbf{t}}_k$, a clustering technique is used to assign them into K clusters, each is represented by a corresponding code $\mathbf{q}_{t,i}$ and $\mathbf{q}_{i,t}$ respectively.

$$L_{CPA} = \frac{1}{2N}\sum_{k=1}^{N}(l(\tilde{\mathbf{i}}_k, \mathbf{q}_{t,i}) + l(\tilde{\mathbf{t}}_k, \mathbf{q}_{i,t})), \tag{3}$$

Total Loss:

$$\mathcal{L} = \lambda_1 \times \mathcal{L}_{ITA} + \lambda_2 \times \mathcal{L}_{CTA} + \lambda_3 \times \mathcal{L}_{CPA} \tag{4}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters. Detail of computation of these loss funtions is referered to the original paper [5].

We pre-train MGCA framework on IGHEndoLesion-Image-Report dataset for 300 epoches on RTX 4090 GPU with batch size of 8. The optimizer is AdamW with learning rate of $2e{-}5$ and weight decay of 0.05. We use a linear warmup with cosine annealing scheduler and initialize learning rate at $1e{-}8$ and warmup period of 20 epochs. The number of prototypes is $K = 50$ and $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 1$.

2) Detector training: After training the image representation with the MGCA framework, we obtained an image encoder that can deeply understand endoscopic images. This is achieved by incorporating expert knowledge through the text encoder, which serves as a guide and complement to the image encoder. The image encoder will be extracted and integrated into the YOLOv8 detection model as its backbone. Next, we train YOLOv8 as object detector and test with the labeled training and testing datasets.

Unlike the conventional method of training both text and image encoding blocks in the original MGCA framework, we propose freezing the text encoder to focus training on the image encoder. This approach enables the model to converge earlier when using the early stopping strategy, thereby reducing the number of parameters that need to be trained. By freezing the text encoder, the number of epochs required is reduced from 75 to 42, and the training time is decreased from 9 hours to 4 hours.

IV. Experiments

A. Evaluation metrics and experimental setups

We evaluate the performance YOLOv8 detector on three datasets: IGHEndoLesion-SEG, IGHEndoLesion-SEG-Cancer, and Kvasir-SEG using mAP50, mAP75, mAP50:95, and the area under curve of the precision-recall (AUC). We compare three strategies of training the detection model:

1) **YOLOv8**: We train YOLOv8 using the pre-trained model from ImageNet on the corresponding training sets. Our goal is to compare our proposed framework with the original YOLOv8.
2) **YOLOv8 + MGCA**: We train YOLOv8 using the backbone obtained from the first stage of the MGCA framework.
3) **GILED:** In this setup, we train the MGCA framework while freezing the text encoder, BioClinical-BERT, which has been trained on an extensive medical text corpus. By doing this, we use BioClinical-BERT as a text extractor without retraining it on a smaller dataset, as in the second training setup.

B. Experimental results

Table II presents the performance of the YOLOv8 model alongside our two proposed models, YOLOv8+MGCA and YOLOv8+MGCA with Frozen Text Encoder (called GILED model). On the IGHEndo Lesion-SEG-Cancer dataset, our first model, YOLOv8+MGCA, achieved a higher AUC (84.2%) compared to the original YOLOv8 (83.6%). The GILED further increased the AUC to 85.0%.

Similarly, on the larger IGHEndo Lesion-SEG dataset, which includes various lesion categories, the detection task is more challenging due to many less severe lesions being small and often confused with normal tissues. As a result, YOLOv8 achieved an AUC of only 53.0%. However, our proposed models, YOLOv8+MGCA and GILED, improved the AUC by 0.8% and 1.1%, respectively. On the Kvasir-SEG dataset, the improvement of our two proposed models is consistent, with increases of 1% and 1.2%. This indicates that the encoder learned from the MGCA framework effectively extracts semantic meaning from both images and text.

TABLE II
RESULTS OF DETECTION MODELS ON THREE DATASETS. VALUES IN BOLD AND UNDERSCORED INDICATE THE BEST AND SECOND-BEST PERFORMANCE, RESPECTIVELY

Dataset	Detection Model	mAP50	mAP75	mAP50:95	AUC
IGHEndo Lesion-SEG-Cancer	YOLOv8	0.836	**0.617**	0.555	0.836
	YOLOv8+MGCA	0.843	0.602	0.550	0.842
	GILED	**0.851**	0.615	**0.556**	**0.850**
IGHEndo Lesion-SEG	YOLOv8	0.530	**0.272**	0.292	0.530
	YOLOv8+MGCA	0.538	0.267	0.293	0.538
	GILED	**0.540**	0.267	**0.297**	**0.541**
Kvasir-SEG	YOLOv8	0.915	0.824	0.746	0.914
	YOLOv8+MGCA	0.917	0.838	0.741	0.915
	GILED	**0.928**	**0.845**	**0.759**	**0.926**

Figure 3 illustrates the detection results produced by the original YOLOv8 and our best proposed model, GILED, across three datasets. The first row of the figure represents the ground truth. The second and last rows show the detection results from the original YOLOv8 and GILED, respectively.

Notably, without MGCA, the original YOLOv8 model detects a larger bounding box for the IGHEndoLesion-SEG-Cancer dataset, whereas GILED detects a box that fits the ground truth perfectly. On the IGHEndoLesion-SEG dataset, where the lesion is less severe and confused with normal tissues, the original YOLOv8 model fails to detect the small lesions, while GILED successfully detects it, albeit with a smaller bounding box than the ground truth. For the Kvasir-SEG example, the original YOLOv8 model produces a small bounding box, considered a false positive, whereas GILED accurately detects the lesion.

Table III compares mAP50 results on the Kvasir-SEG dataset across different existing models: Faster R-CNN, RetinaNet, YOLOv3+spp, YOLOv4, and ColonSegNet [22]. It is noticed that among the reported models, RetinaNet obtained the highest mAP50 (91.0%) but can not surpass YOLOv8 (91.5%) while YOLOv8 (150 FPS on GTX 1080Ti) achieved FPS comparable to ColonSegNet (180 FPS on GTX 2080Ti). In addition, our proposed model GILED achieved the highest mAP50 among all compared models.

V. CONCLUSIONS

This paper presented an initial investigation into combining endoscopic images with their corresponding medical reports. To achieve this, we employed the MGCA framework to develop an image representation that leverages the multi-granularity relationship between text-image pairs. This image representation was then used as the backbone in a state-of-the-art detection model, YOLOv8. We compared various strategies for training YOLOv8 and found that the best performance was achieved when the image representation was trained while freezing the text encoder in the MGCA framework. Our proposed GILED model achieved the highest AUC scores of 84.2%, 54.1%, and 92.6% on the IGHEndoLesion-SEG-Cancer, IGHEndoLesion-SEG, and Kvasir-SEG datasets, respectively. Notably, the GILED model achieved the best mAP50 score (92.8%) among six state-of-the-art models on the Kvasir-SEG dataset. This demonstrates that exploring the semantic relationship between text and images is crucial and beneficial for improving detection performance. In the future, we plan to expand the multimodal dataset of text-image pairs and integrate multiple languages into the MGCA to ensure language independence.

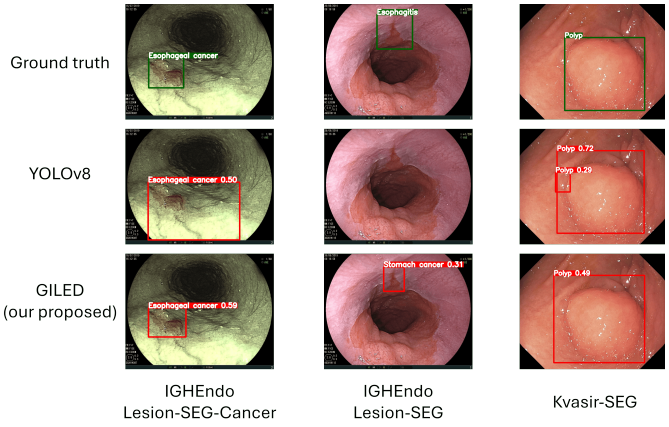


Fig. 3. Illustration of detection results obtained by three models on three dataset.

TABLE III
COMPARISION OF THE PERFORMANCE OF MODELS ON THE KVASIR-SEG DATASET

Detection Model	mAP50	FPS
Faster R-CNN	0.842	8.00
RetinaNet	0.910	16.20
YOLOv3+spp	0.853	45.01
YOLOv4	0.823	48.00
ColonSegNet	0.817	**180.00**
YOLOv8	0.915	**150.00**
GILED (our proposed)	**0.928**	**150.00**

ACKNOWLEDGEMENT

This research is funded by Vietnam Ministry of Science and Technology under grant number KC-4.0-17/19-25.

REFERENCES

[1] [Online]. Available: https://gco.iarc.fr/en.
[2] A. Nogueira-Rodríguez, M. Reboiro-Jato, D. Glez-Peña, and H. López-Fernández, "Performance of convolutional neural networks for polyp localization on public colonoscopy image datasets," *Diagnostics*, vol. 12, no. 4, p. 898, 2022.

[3] Y. Ma, X. Chen, K. Cheng, Y. Li, and B. Sun, "Ldpolypvideo benchmark: A large-scale colonoscopy video dataset of diverse polyps," in *MICCAI*, Springer, 2021, pp. 387–396.

[4] R. Hirsch, M. Caron, R. Cohen, *et al.*, "Self-supervised learning for endoscopic video analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, pp. 569–578.

[5] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, "Multi-granularity cross-modal alignment for generalized medical visual representation learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 536–33 549, 2022.

[6] C. Liu, S. Cheng, C. Chen, *et al.*, "M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, pp. 637–647.

[7] D. Jha, P. H. Smedsrud, M. A. Riegler, *et al.*, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, Springer, 2020, pp. 451–462.

[8] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CsUR)*, vol. 51, no. 6, pp. 1–36, 2019.

[9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[10] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[11] C. Jia, Y. Yang, Y. Xia, *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*, PMLR, 2021, pp. 4904–4916.

[12] X. Zhai, X. Wang, B. Mustafa, *et al.*, "Lit: Zero-shot transfer with locked-image text tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 123–18 133.

[13] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling language-image pre-training via masking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 390–23 400.

[14] C. Schuhmann, R. Vencu, R. Beaumont, *et al.*, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[15] C. Schuhmann, R. Beaumont, R. Vencu, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[16] A. Taleb, C. Lippert, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," in *International conference on information processing in medical imaging*, Springer, 2021, pp. 661–673.

[17] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Zero-shot video question answering via frozen bidirectional language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 124–141, 2022.

[18] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*, PMLR, 2022, pp. 2–25.

[19] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3942–3951.

[20] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 372–21 383.

[21] Z. Wan, C. Liu, M. Zhang, *et al.*, "Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[22] D. Jha, S. Ali, N. K. Tomar, *et al.*, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *Ieee Access*, vol. 9, pp. 40 496–40 510, 2021.

[23] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[24] T. H. Nguyen, T. D. H. Nguyen, D. Phung, *et al.*, "A vietnamese-english neural machine translation system," in *Annual Conference of the International Speech Communication Association (was Eurospeech) 2022*, International Speech Communication Association (ISCA), 2022, pp. 5543–5544.

[25] E. Alsentzer, J. R. Murphy, W. Boag, *et al.*, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

Does Brain Atlas Choice Matter? An Empirical Study in Alzheimer's Diagnosis Using FDG-PET Images

Pham Minh Tuan*†§, Mouloud Adel*‡, Nguyen Linh Trung§, and Eric Guedj*†
* Institut Fresnel, Aix-Marseille université, Marseille, France
E-mails: {minh-tuan.pham, mouloud.adel, eric.guedj}@univ-amu.fr
†CERIMED, Aix-Marseille université, Marseille, France
‡Galatasaray University, Istanbul, Turkey
§University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam
Correspondence: Nguyen Linh Trung (linhtrung@vnu.edu.vn)

***Abstract*—Brain mapping is essential for diagnosing neurological conditions and can greatly influence analytical outcomes. This study evaluates the performance of various brain atlases in classifying Alzheimer's disease (AD) and mild cognitive impairment (MCI) using fluorodeoxyglucose positron emission tomography (FDG-PET) images, focusing on the impact of atlas choice and granularity. Our analysis reveals substantial performance differences among atlases, with data-tailored atlases demonstrating notably higher effectiveness. These findings highlight the importance of choosing atlases that align with specific tasks and datasets, offering valuable insights for improving diagnostic accuracy.**

I. Introduction

Alzheimer's Disease (AD) is a widespread neurodegenerative condition predominantly affecting the elderly, with projections indicating an increase from 50 million cases in 2018 to 150 million by 2050 [1]. In efforts to distinguish AD from non-AD conditions, there is increasing interest in using machine learning techniques alongside fluorodeoxyglucose positron emission tomography (FDG-PET) for early diagnosis [2]–[5].

FDG-PET studies generally apply either voxel-based or region-based approaches. The voxel-based approach considers every brain voxel as a feature [6], [7], whereas the region-based method divides the brain into regions of interest (ROIs) and extracts features from these segments [4], [8]. The region-based approach is often favored for its ability to lower data dimensionality, reduce noise, and streamline analysis. This approach typically involves using predefined brain atlases such as the Anatomical Automatic Labeling (AAL) [9], the LONI Probabilistic Brain Atlas (LPBA40) [10], the Schaefer atlas [11], and the Harvard-Oxford atlas [12]. Given the variety of available brain atlases, selecting the appropriate one is crucial for analyzing and interpreting neuroimaging data, especially in studies of brain disorders. There is no one-size-fits-all atlas [13]; each atlas has unique advantages that make it more or less suitable depending on the type of neuroimaging data and the specific analytical objectives.

The question of which atlas to use is not new and has been a topic of considerable debate in the research community [14]–[16]. The selection of an atlas can significantly impact the outcomes of a study, influencing the final analysis and conclusions drawn from the data. Studies have shown that the suitability of an atlas can vary depending on the context of the research, indicating that an atlas that performs well in one scenario might not be as effective in another [15]. Therefore, the choice of atlas should be carefully considered within the specific context of each study.

In this study, we aim to explore the impact of atlas choice on the diagnosis of AD across different conditions, including Normal Control (NC), mild cognitive impairment (MCI), and AD stages using FDG-PET images. Our study focuses on two main questions: *Does the choice of atlas matter in the classification performance of AD diagnosis? And, is there a correlation between the scale of the atlas and the performance obtained?* By addressing these questions, we seek to provide insights into how different brain mapping strategies can affect the analysis of brain disorders and to identify the most effective atlases for specific neuroimaging applications. More specifically, our contribution lies in two key areas: 1) We compare the performance of Alzheimer's disease diagnosis under various conditions using different types of atlases. 2) We examine the correlation between diagnostic performance and atlas properties such as the number of regions, type of atlases (functional/structural atlas).

The remaining of this paper is organized as follows. Section II describes the proposed method and the used dataset from the ADNI. Section III presents experimental results and discussions, followed by conclusions and a plan for future work in Section IV.

II. Materials and Method

A. Dataset

Data used in the experiment is 18F-FDG PET images downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). After the downloading

TABLE I: Demographic and clinical information of subject

Characteristic	NC	sMCI	pMCI	AD
Number of subjects	242	360	209	237
Female/Male	122/120	153/207	87/122	97/140
Age (Mean ± std.)	73.66 ± 5.66	71.73 ± 7.66	73.89 ± 6.88	75.00 ± 7.91
MMSE (Mean ± std.)	29.03 ± 1.20	28.20 ± 1.59	27.13 ± 1.71	23.19 ± 2.12

process, images are then spatially normalized to Montreal Neurological Institute template with a voxel size of $2 \times 2 \times 2mm^3$ and dimensions of $91 \times 109 \times 91$. Next, the intensity normalization is then performed by dividing each voxel intensity by the average value of the cerebellum region. Finally, images are smoothed by a Gaussian kernel with a full width at half maximum of 8 mm. After preprocessing phase, we have a total 1048 FDG-PET images for later experiments, including 237 AD subjects, 360 stable MCI (sMCI) subjects, 209 progressive MCI (pMCI) subjects and 242 NC subjects. The demographic and clinical information of subjects is provided in Table I, in which MMSE stands for the Mini-Mental State Examination [17].

B. Methods

a) Atlas selection: There are many types of brain atlases, each with its own properties. These atlases may vary in resolution, construction criteria, and the methodologies used for their creation, which can include both manual and automated approaches. Such variations result in significant differences in the structure and interpretation of the atlases.

In this study, rather than conducting an exhaustive comparison, we focus on the most commonly used and standardized atlases. We include some common atlases that have been standardized with a voxel size of $2\times2\times2mm^3$ and dimensions of $91 \times 109 \times 91$, as detailed in previous research [14], [15]. The use of these standardized atlases simplifies our workflow and facilitates more straightforward comparisons among them. Since our study focuses on FDG-PET images, and most of the atlases used in [14] are based on MRI images, we have also included atlases specifically designed for FDG-PET images. One such atlas is C-Atlas, a PET-driven brain mapping tool specifically developed for the diagnosis of AD [18]. In total, we incorporate 28 atlases into our analysis. These atlases are categorized into several groups based on criteria such as modality (MRI-based or PET-based) and type (structural or functional), as shown in Table II.

b) Feature Extraction and Classifier Evaluation: For each atlas, the brain images are segmented into regions based on the atlas. Then, the mean voxel intensity within each region is used as a feature for analysis. This approach is well-established and commonly used in previous studies [4], [8], [19]. Given an atlas with n regions, feature extraction from each image results in a vector $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, where x_i represents the mean intensity value corresponding to region i in the image. These features can then be used to train a linear Support Vector Machine (SVM) classifier [20] for binary classification tasks, including NC-AD, sMCI-pMCI. Here, we use a linear SVM due to its simplicity and its prevalence in previous studies [2].

c) Experimental Design and Evaluation Metrics: The effectiveness of each atlas choice is evaluated based on the performance of the classifier. We design experiments using the ADNI dataset for AD classification tasks (NC-AD, sMCI-pMCI) and assess the performance metrics of the classifiers. The performance is then measured based on ACCuracy, SPEcificity, SENsitivity and Area Under Curve (AUC) metrics where:

$$\begin{aligned} ACC &= \frac{TP+TN}{TP+TN+FP+FN}, \\ SEN = \frac{TP}{TP+FN}, SPE &= \frac{TN}{TN+FP} \end{aligned} \tag{1}$$

AUC is the Area Under the Curve ROC (Receiver Operating Characteristics) determined by SEN and $1 - SPE$. The higher value of these metrics indicates the better performance. For classification tasks, we use 5-fold cross-validation. All experiments are repeated 40 times, and the average results are reported. The experiments are implemented using the scikit-learn library [21]. Default values set by the library are used for experimental parameters unless otherwise specified.

d) Comparing Atlas Choice: We evaluate the performance of each atlas based on ACC and AUC, which are commonly used metrics for comparing classifier performance. This evaluation is performed on two classification tasks: NC-AD and sMCI-pMCI. Additionally, we also analyze the correlation between performance metrics and atlas properties, such as the type of atlas and the number of regions.

III. Results and Discussion

A. Results

a) Performance Variation by Atlas Choice and Task: Table III and Fig. 1 illustrate the performance of each atlas on two classification tasks: NC-AD and sMCI-pMCI classification. In general, we observe notable differences in performance across atlases. By comparing the best and worst-performing atlases, we find that performance improvements can be up to approximately 22%, 18% for ACC and AUC for NC-AD task; and 15% and 31% for the sMCI-pMCI task, respectively. This highlights that the choice of atlas significantly impacts classifier performance. Additionally, we find that the performance of all atlases is generally higher for the NC-AD task compared to the sMCI-pMCI task, likely because distinguishing between sMCI and pMCI is more challenging due to the subtle differences between these groups.

b) Best and Worst Performances: In a more detailed comparison, for the NC-AD classification task, atlases from the C-Atlas group achieved the highest performance, while

TABLE II: List of included atlases and their properties

Atlas	Short name	#ROIs	Imaging Modality	Type
AAL	AAL	116	MRI	Structural Atlas
AICHA	AICHA	384	MRI	Functional Atlas
Brodmann	BrodM	41	MRI	Structural Atlas
CAPRSC	CAPRSC	333	MRI	Functional Atlas
DKT atlas	DKT	83	MRI	Structural Atlas
Hammersmith atlas	HamSmith	83	MRI	Structural Atlas
Harvard Oxford cortical atlas	HOCort	48	MRI	Structural Atlas
Harvard Oxford subcortical atlas	HOSub	21	MRI	Structural Atlas
Hemispheric atlas	Hemisphere	2	MRI	Structural Atlas
JHU	JHU	48	MRI	Structural Atlas
Juelich	Juelich	103	MRI	Structural Atlas
LONI Probabilistic Brain Atlas	LPBA40	56	MRI	Structural Atlas
Schaefer200	SC200	200	MRI	Functional Atlas
Schaefer300	SC300	300	MRI	Functional Atlas
Schaefer400	SC400	400	MRI	Functional Atlas
Schaefer1000	SC1000	1000	MRI	Functional Atlas
Talairach	Talairach	959	MRI	Structural Atlas
Tissue	Tissue	3	MRI	Structural Atlas
Yeo-17-liberal	Yeo17l	17	MRI	Functional Atlas
Yeo-17	Yeo17	17	MRI	Functional Atlas
Yeo-7-liberal	Yeo7l	7	MRI	Functional Atlas
Yeo-7	Yeo7	7	MRI	Functional Atlas
C-Atlas100	C100	100	PET	Functional Atlas
C-Atlas200	C200	200	PET	Functional Atlas
C-Atlas300	C300	300	PET	Functional Atlas
C-Atlas400	C400	400	PET	Functional Atlas
C-Atlas500	C500	500	PET	Functional Atlas
C-Atlas1000	C1000	1000	PET	Functional Atlas

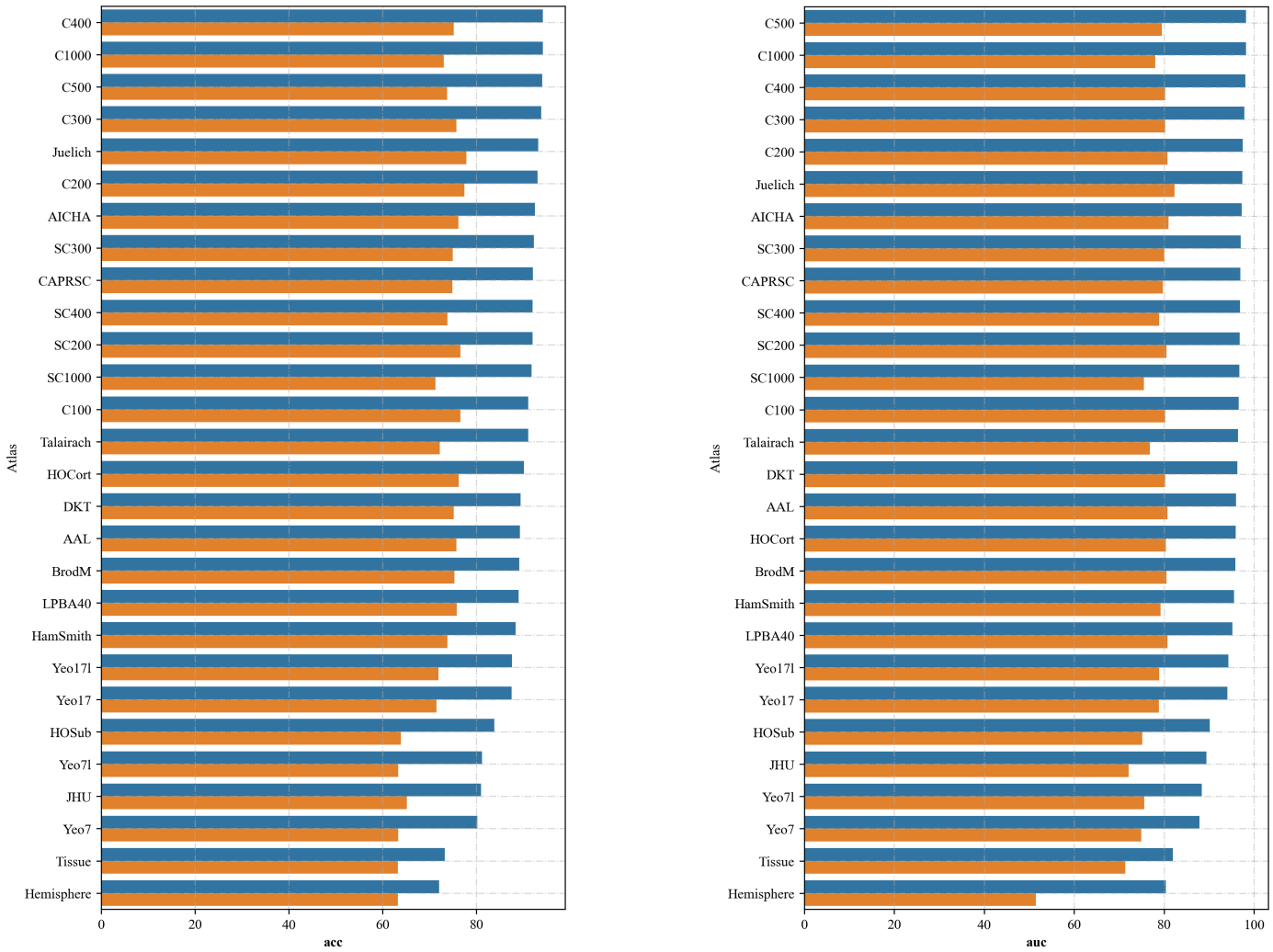
TABLE III: Compare Performance of various Brain Atlases on AD Classification Tasks

Atlas	NC vs. AD		sMCI vs. pMCI	
	ACC	AUC	ACC	AUC
AAL	89.258	96.009	75.751	80.780
AICHA	92.468	97.259	76.148	81.031
BrodM	89.154	95.869	75.298	80.585
CAPRSC	92.024	96.971	74.869	79.708
DKT	89.383	96.314	75.157	80.249
HOCort	90.126	95.923	76.226	80.432
HOSub	83.773	90.199	63.867	75.123
HamSmith	88.333	95.538	73.841	79.185
Hemisphere	*72.069*	*80.387*	*63.270*	*51.465*
JHU	80.996	89.401	65.177	72.157
Juelich	93.142	97.411	**77.887**	**82.336**
LPBA40	88.966	95.193	75.812	80.745
SC200	91.945	96.817	76.570	80.580
SC300	92.274	97.079	74.979	80.030
SC400	91.967	96.888	73.810	78.961
SC1000	91.712	96.773	71.224	75.488
Talairach	91.028	96.443	72.184	76.871
Tissue	73.243	81.935	63.270	71.377
Yeo17	87.490	94.111	71.446	78.840
Yeo17l	87.569	94.306	71.885	78.961
Yeo7	80.221	87.862	63.305	74.927
Yeo7l	81.176	88.421	63.314	75.595
C100	91.054	96.622	76.568	80.289
C200	93.000	97.505	77.456	80.746
C300	93.804	97.855	75.694	80.222
C400	**94.186**	98.083	75.132	80.265
C500	94.018	**98.257**	73.735	79.495
C1000	94.143	98.246	73.012	78.000

the Juelich atlas was the top performer for the sMCI-pMCI task. Moreover, as shown in Fig. 1, focusing on the top 10 atlases with the highest and lowest performance reveals that atlases such as C200, C300, Juelich, and SC300 performed better on both NC-AD and sMCI-pMCI tasks. Conversely, atlases like Hemisphere, Tissue, and Yeo7 performed poorly compared to others. Additionally, some atlases demonstrated strong performance in one task but weaker performance in another; for instance, LPBA40 performed better on the sMCI-pMCI task but showed lower performance on the NC-AD task compared to other atlases. Once again, these observations underscore the significant variability in performance based on the choice of atlas, highlighting the importance of selecting the most appropriate atlas for each specific classification task.

c) Relationship Between Atlas Properties and Performance: Table II and Fig. 2 show the properties of atlases and their correlation with performance, respectively. We observe that, in general, the performance of functional atlases (represented by orange points) is better than that of structural atlases (represented by blue points). This is reasonable, given that our study uses FDG-PET images, a functional imaging modality, making functional atlases more appropriate. Additionally, the C-Atlas demonstrates superior performance compared to other atlases, particularly on the NC-AD task, for both ACC and AUC metrics. This is likely because the C-Atlas is constructed from FDG-PET images and is specifically adapted for AD diagnosis, making it more suitable for FDG-PET data and AD classification. These observations emphasize the importance of selecting an atlas that aligns with the data properties and underscore the role of adapted atlases in improving classifier performance.

Fig. 3 illustrates the correlation between the number of regions in each atlas and their AUC performance on the NC-AD and sMCI-pMCI tasks. We observe a weak linear correlation between the total number of regions and performance, especially with the C-Atlas group for the NC-AD task. Generally, increasing the number of regions improves



(a) ACC

(b) AUC

Fig. 1: Comparison of performance across various brain atlases for Alzheimer's disease classification tasks: NC-AD (blue) and sMCI-pMCI (orange). Performance metrics include ACC (a) and AUC (b).

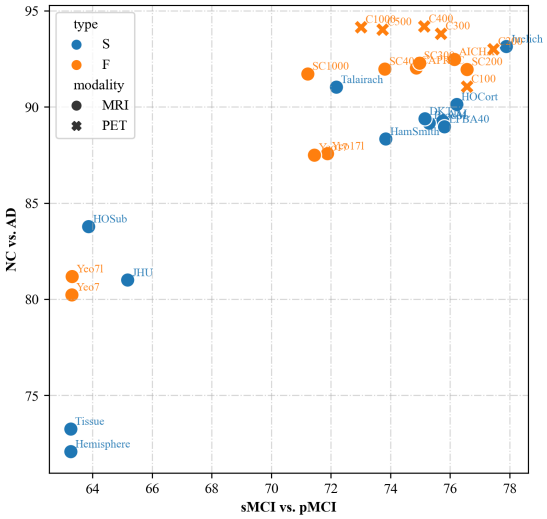
classifier performance, particularly within the range of 1 to 300 regions. However, beyond this range, adding more regions does not further enhance performance and may even reduce it. This trend is observed for both NC-AD and sMCI-pMCI tasks. These observations suggest that selecting an atlas with an optimal number of regions is crucial for better performance, as it ensures an effective representation of regional features.

Collectively, these observations highlight that the choice of atlas significantly impacts classifier performance. It is essential to select an atlas that aligns with the nature of the data to achieve better results. Additionally, these findings suggest the benefits of using customized atlases tailored to the data and specific tasks.

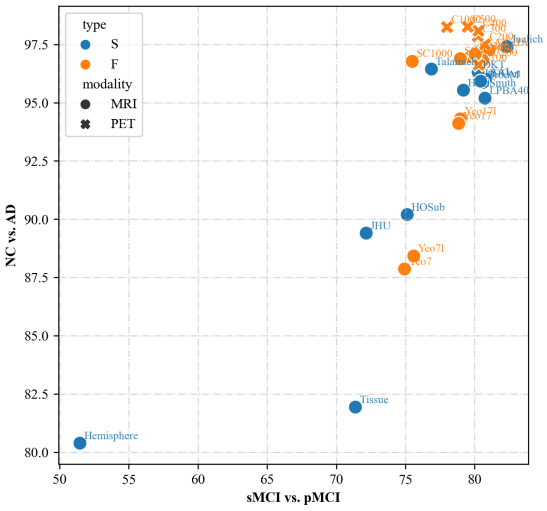
B. Discussion

a) Factors Influencing Atlas Performance: Our analysis shows that the choice of brain atlas significantly impacts classifier performance in AD classification tasks. Experimental results suggest that atlases aligning with the nature of the imaging data generally outperform those that do not; for instance, in our case, functional atlases are more suitable for FDG-PET images, a type of functional imaging. Additionally, the number of regions in an atlas affects performance, as an atlas with a suitable number of regions can ensure an effective representation of regional features, thereby enhancing classifier performance. These findings highlight the importance of selecting an atlas that aligns with the nature of the data and has an optimal number of regions to effectively represent relevant features.

b) Necessity of Customized Atlases: In our study, C-Atlas demonstrated better performance and effectiveness in AD classification tasks. This suggests the critical role of customized atlases tailored to specific imaging modalities and classification tasks. Customized atlases align better with the inherent properties of the data, thereby enhancing classification accuracy. Moreover, as described in [18], one of the advantages of customized atlases is their ability to be specified to various

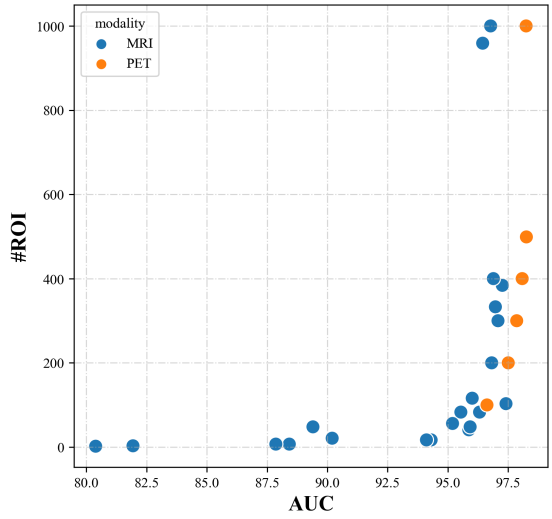


(a) ACC

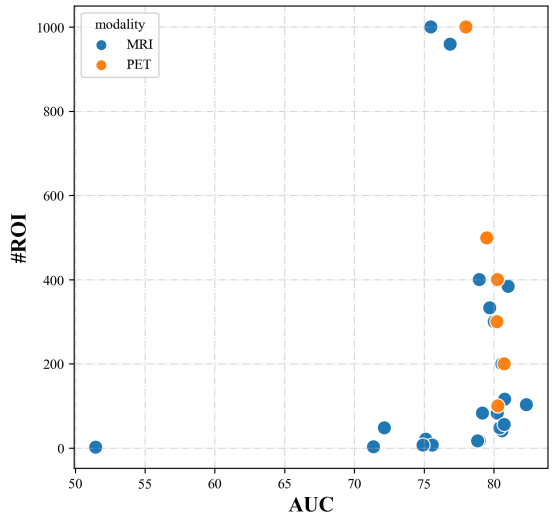


(b) AUC

Fig. 2: Relationship between atlas properties and AD classification performance. Metrics include ACC (a) and AUC (b). Structural atlases are 'S' and functional atlases are 'F'.



(a) NC vs. AD



(b) sMCI vs. pMCI

Fig. 3: Relationship between the number of atlas regions and AUC performance in AD classification tasks: NC-AD (a) and sMCI-pMCI (b).

scales, providing flexibility and facilitating future analyses, such as those involving brain network analysis.

c) Future Work: While this study provides valuable insights, it has several limitations. First, it uses a specific set of atlases and focuses exclusively on FDG-PET imaging and the SVM classifier. Future research should explore a broader range of atlases, including those for different imaging modalities and conditions, to better understand how atlas properties impact performance. Additionally, experiments with other classifiers, such as decision trees, logistic regression, and neural networks, as well as investigations into other tasks, such as graph analysis and dimensional reduction, are needed. Furthermore, studying the correlation between the structure of atlases (including the properties of each region) and performance could yield valuable insights. Lastly, further advancements in atlas customization and optimization, as well as the investigation of multi-modal atlases, could improve classification outcomes.

IV. Conclusions

In this study, we examined the performance of various brain atlases in AD classification tasks using FDG-PET images. Our experiments revealed significant performance differences across atlases, with the adapted atlas (C-Atlas) achieving notably higher results. These variations are primarily attributed to factors such as data types and the number of brain regions. Our findings emphasize the importance of selecting brain atlases carefully and highlight the crucial role of customizing atlases to align with specific tasks and datasets. Atlases tailored to the characteristics of the target dataset can significantly enhance classification outcomes. For future work, it will be valuable to explore a broader range of atlases and imaging modalities, test

additional classifiers, and examine the relationships between atlas structure and performance. Such efforts will help in drawing more comprehensive conclusions and generalizing results across different contexts.

ACKNOWLEDGMENT

The research of Nguyen Linh Trung and Pham Minh Tuan was funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2021.55.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

REFERENCES

[1] C. Patterson, "World Alzheimer Report 2018: The state of the art of dementia research," *New frontiers. London: Alzheimer's Disease International.*, 2018.

[2] A. Alberdi, A. Aztiria, and A. Basarab, "On the early diagnosis of Alzheimer's Disease from multimodal signals: A survey," *Artificial intelligence in medicine*, vol. 71, pp. 1–29, 2016.

[3] P. M. Tuan, T.-L. Phan, M. Adel, E. Guedj, and N. L. Trung, "AutoEncoder-based feature ranking for Alzheimer Disease classification using PET image," *Machine Learning with Applications*, vol. 6, p. 100 184, Dec. 2021.

[4] K. R. Gray, R. Wolz, R. A. Heckemann, *et al.*, "Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease," *NeuroImage*, vol. 60, no. 1, pp. 221–229, 2012.

[5] P. M. Tuan, N. L. Trung, M. Adel, and E. Guedj, "AutoEncoder-based Feature Ranking for Predicting Mild Cognitive Impairment Conversion using FDG-PET Images," in *2023 IEEE Statistical Signal Processing Workshop (SSP)*, Hanoi, Vietnam: IEEE, Jul. 2023, pp. 720–724.

[6] C. Cabral, P. M. Morgado, D. C. Costa, M. Silveira, A. s. D. N. Initiative, *et al.*, "Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages," *Computers in biology and medicine*, vol. 58, pp. 101–109, 2015.

[7] C. Hinrichs, V. Singh, L. Mukherjee, *et al.*, "Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset," *Neuroimage*, vol. 48, no. 1, pp. 138–149, 2009.

[8] K. Ota, N. Oishi, K. Ito, H. Fukuyama, S.-J. S. Group, A. D. N. Initiative, *et al.*, "Effects of imaging modalities, brain atlases and feature selection on prediction of Alzheimer's disease," *Journal of neuroscience methods*, vol. 256, pp. 168–183, 2015.

[9] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, *et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, Jan. 2002.

[10] D. W. Shattuck, M. Mirza, V. Adisetiyo, *et al.*, "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, Feb. 2008.

[11] A. Schaefer, R. Kong, E. M. Gordon, *et al.*, "Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI," *Cerebral Cortex*, vol. 28, no. 9, pp. 3095–3114, Sep. 2018.

[12] N. Makris, J. M. Goldstein, D. Kennedy, *et al.*, "Decreased volume of left and total anterior insular lobule in schizophrenia," *Schizophrenia Research*, vol. 83, no. 2-3, pp. 155–171, Apr. 2006.

[13] M. Salehi, A. S. Greene, A. Karbasi, X. Shen, D. Scheinost, and R. T. Constable, "There is no single functional atlas even for a single individual: Functional parcel definitions change with task," *NeuroImage*, vol. 208, p. 116 366, Mar. 2020.

[14] R. M. Lawrence, E. W. Bridgeford, P. E. Myers, *et al.*, "Standardizing human brain parcellations," *Scientific Data*, vol. 8, no. 1, p. 78, Mar. 2021.

[15] A. Y. Revell, A. B. Silva, T. C. Arnold, *et al.*, "A framework For brain atlases: Lessons from seizure dynamics," *NeuroImage*, vol. 254, p. 118 986, Jul. 2022.

[16] J. W. Bohland, H. Bokil, C. B. Allen, and P. P. Mitra, "The Brain Atlas Concordance Problem: Quantitative Comparison of Anatomical Parcellations," *PLoS ONE*, vol. 4, no. 9, O. Sporns, Ed., e7200, Sep. 2009.

[17] A. J. Mitchell, "A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment," *Journal of Psychiatric Research*, vol. 43, no. 4, pp. 411–431, Jan. 2009.

[18] P. M. Tuan, T.-L. Phan, M. Adel, S. Bourennane, N. L. Trung, and E. Guedj, "C-Atlas: A Brain Mapping based on FDG-PET Images for Alzheimer's Disease Diagnosis," in *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, Ho Chi Minh City, Vietnam: IEEE, Dec. 2022, pp. 150–155.

[19] Y. Asim, B. Raza, A. K. Malik, S. Rathore, L. Hussain, and M. A. Iftikhar, "A multi-modal, multi-atlas-based approach for Alzheimer detection via machine learning," *International Journal of Imaging Systems and Technology*, vol. 28, no. 2, pp. 113–123, 2018.

[20] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*, 2011.

Confidence-Aware Learning for Person Re-identification with Noisy Labels

Duhyun Kim[1,2] and Jae-Young Sim[1,†]
[1]Ulsan National Institute of Science and Technology, Republic of Korea
[2]S-1 CORPORATION, Republic of Korea
E-mail: {duhyunkim, jysim}@unist.ac.kr

***Abstract*—Accurately labeling a large amount of data for person re-identification is a significant challenge. In this paper, we introduce a technique to effectively perform person re-identification even in the dataset with noisy labels. Leveraging the widely observed phenomenon that data with wrong labels tends to have large loss values, we fit the Gaussian mixture model(GMM) to estimate confidence which is the probability of the sample being noise-labeled. We propose confidence-aware learning that appropriately reflects confidence to balance between mitigating the impact of samples with noisy labels and guiding anchors to the complete positive and negative samples. Additionally, we refine the GMM to enhance the accuracy of confidence for each data sample even in a lack of data situation. Experimental results demonstrate that our methods are effective techniques for handling noisy labels in person re-identification.**

I. Introduction

Person re-identification (Re-ID) aims to identify the same person across various viewpoints obtained from different cameras. This technology has progressed because it is crucial in surveillance for safety and lifelogging. Existing supervised Re-ID performed well using many person image datasets and high-quality identity labels during training. However, accurately labeling large amounts of frames is challenging due to the minor variations in inter-class such as appearance, and significant variations in intra-class such as viewpoints and resolution. In addition, establishing guidelines for perfect labeling and hiring personnel with expertise in labeling can be time-consuming and expensive. Therefore, even if there is noise in the labeled data, low-cost labeling methods like crowdsourcing or annotation tools can still be utilized. Unfortunately, when trained with noisy labels, various supervised methods suffer from performance degradation as the model learns to consider different individuals as the same person.

Various research efforts are emerging to address the impact of noisy labels, but applying these directly to Re-ID is hard because of the large number of ID classes and severe class imbalance. Fig. 1 shows the comparison of the amount of data within the classes. In the classification dataset [1][2], all classes have sufficient data and a low degree of class imbalance. In contrast, Re-ID datasets typically exhibit severe class imbalance, making distinguishing noise within ID classes difficult with tale class data. Moreover, the large number of ID classes in person re-identification datasets complicates the identification of a sample's correct ID.

Existing studies on learning with noisy labels [3][4][5] did not handle these issues. They commonly utilize predicted ID labels during training by adopting a soft label strategy. In other words, they believe that using predicted ID labels is more reliable than relying on noisy labels. However, Re-ID datasets have many ID classes, making it even more challenging to predict IDs during training accurately. Consequently, using predicted ID classes may introduce further inaccuracies that interfere with model training. Therefore, we propose a new learning strategy that does not rely on a soft label strategy.

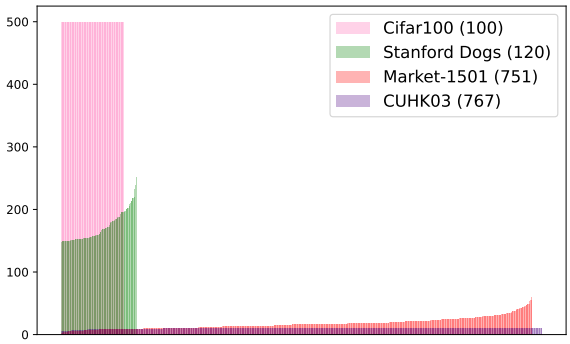


Fig. 1. Comparison of the amount of data within the class by four datasets. Cifar100 [1] for classification, Stanford Dogs [2] for fine-grained classification, Market-1501 [7] and CUHK03 [8] for re-identification. The number in the legend means the number of classes in the dataset.

Additionally, the data imbalance issue implies that the likelihood of noise is different according to ID class. Individuals from frequently appearing IDs in the dataset will have more variance and, consequently, a higher probability of having noise labels. This highlights the need to adequately consider whether a sample within a class is noisy or not. However, image recognition studies [5][3][6] assume that there is the same probability of the noise for each class, so they do not handle the issue of inter-class data imbalance. To tackle these challenges, we exploit the observation that samples with noisy labels have higher loss values than those with clean labels. We utilize a two-component Gaussian Mixture Model (GMM) for loss values to assess the label confidence of each instance, based on the probability that a sample's loss value belongs to a clean distribution component. In this phase, we devise GMM refinement to obtain more reliable confidence for each sample. This is crucial because GMM modeling tends to perform poorly when there is insufficient data within the ID class. After

that, we incorporate label confidence into model training by preventing the model from being trained on incorrectly labeled samples when access to the correct ID class is unavailable during training. The main contributions of this paper are as follows.

- In this paper, we devise the GMM refinement to get more reliable confidence for the label considering the inter-class data imbalance.
- We propose a new learning method to reflect the confidence of labels for each instance by reducing the impact of incorrect labels and helping the model optimize by guiding input to completely positive and negative samples.
- We demonstrate our proposed method effectively handles the issue caused by noisy labels on Market-1501 and CUHK03.

II. Related Works

A. Person Re-ID

Many supervised methods achieve the highest performance thanks to datasets with clean labels. TransReID [9] proposed a transformer-based strong baseline and a jigsaw patch module to handle occlusion and misalignment issues. CDNet [10] introduced combined blocks to extract features more suitable for retrieval by constructing two branches with different kernel sizes. However, these supervised methods rely on the correct labels with expansive labeling cost and show performance degradation under the noisy labels setting.

B. Image Classification with noisy labels

Co-teaching [11] trained the model with the small loss value first and prevented overfitting by exchanging results between two identical models. JoCoR [12] also adopted a two-identical model strategy and trained the model only when the predictions of the two models were the same. TCL [6] employed a GMM for features to obtain reliability scores and utilized contrastive learning, assuming that the two samples are always considered positive when creating mixup views. SNSCL [5] proposed a solution to address the issue of noisy labels in fine-grained classification, where the data for each class exist in a closer feature space. They estimated the reliability score of the data using a two-component GMM of loss values and leveraged a soft label strategy between the model's predictions and the ground truth. However, these methods also use a soft-label strategy without concerning the large number of ID classes. And they do not handle the issue of severe class imbalance because they assume that the number of noisy labels in all classes is similar.

C. Person Re-ID with noisy labels

There are existing Re-ID methods designed to address the challenges of noisy labels. DNet [13] assigned uncertainty into features to prevent the model from optimizing by noise samples. PurifyNet [4] trained model with soft label strategy using predicted ID classes and introduced a technique to assign lower weights to samples located away from the center of each class. CORE [3] introduced a co-refining strategy in which two identical structures collaboratively refine each other using soft label cross entropy by exchanging their predicted results during the training. TSNT [14] also adopted a co-refining strategy and introduced a more suitable triplet loss to handle noisy labels by selecting more reliable hard-positive and negative samples. [4][3][14] utilize a soft label strategy assuming that predicted ID class during the training is more reliable than noisy labels. However, it is hard to get the exact correct ID class during the training because there are a lot of ID classes in the re-identification. LRP [15] evaluates label confidence by determining how many close samples share the same ID class. LRNI-HSR [16] refines labels before training by relying on the judgments of a pretrained network. ICLR [17] refined label samples when they were far from the ID's center feature and adjusted the training contribution to give higher attention to samples estimated to be noisy in the early epochs.

III. Proposed Methods

We reflect the confidence of the ID label to minimize the model's performance degradation caused by incorrect labels. This process is composed of two steps and is repeated for every epoch. First, we estimate the sample's confidence, representing the probability of being the correct label, with GMM refinement to get a more reliable GMM even when there is a lack of data within the class. Second, the model should be trained with the Re-ID loss to suppress low-confidence samples and Instance loss to make low-confidence samples closer to the complete positive samples and far from negative samples. Fig. 2 illustrates these two steps, where the blue dashed line represents the confidence estimation step with frozen model weights, and the red line represents confidence-aware learning.

A. Confidence Estimation

We employ a two-component Gaussian Mixture Model(GMM) for each class to get the probability of a sample having the correct labels [5]. This is a common approach based on the observation that samples with noisy labels tend to have higher loss values than samples with clean labels. As Fig. 3, this knowledge can also adapted to the Re-ID task. We can regard the scale of loss as a clue to identifying samples with a noise label. After fitting the GMM, we estimate the posterior probability of belonging to the clean distribution of all the person instances in the training dataset, denoting confidence γ. The confidence is renewed before every training epoch.

However, because of the data imbalance in the Re-ID dataset, all the ID classes do not always have sufficient samples with noisy labels to get reliable GMM. For example, if one class is composed of only samples with correct labels, the second component, which should represent the distribution of noise loss values, encounters an issue where it is composed of clean samples rather than noisy samples. Furthermore, when the number of noise samples is extremely small, proper modeling may not occur because the noise samples do not necessarily have larger loss values than clean samples. To

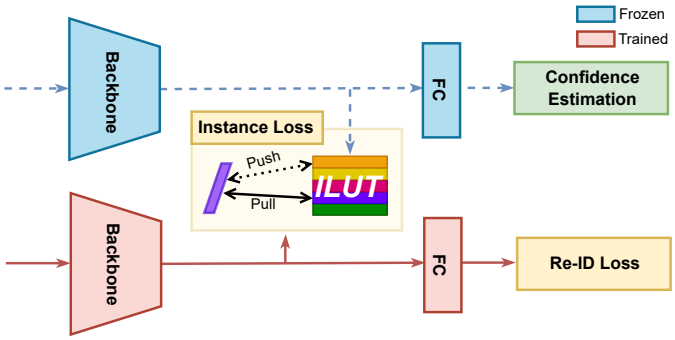


Fig. 2. Overview of our proposed methods.

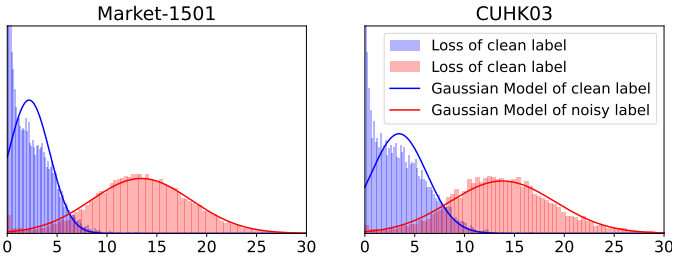


Fig. 3. Histogram of loss values calculated by correct label and incorrect label, and Gaussian distribution model with loss values in two datasets Market-1501 and CUHK03.

address this, we devised GMM refinement (GR), which is the trick to getting a more reliable GMM for our re-ID situation. When modeling the GMM for a specific ID class, we utilize the loss values of samples within the ID class and samples randomly selected from another class, representing the role of fake samples with noisy labels. As a result, the second component of the GMM can always represent the distribution of the noise loss value.

B. Confidence-Aware Learning

Cross-entropy loss and triplet loss are used to train the Re-ID model generally. However, they are ineffective in noisy label settings because of their strong dependence on the ID labels. To address this issue, we introduce a learning strategy for generating robust models from noisy labels using estimated confidence.

1) Re-ID Loss: If we train the model with the general cross-entropy loss in a noisy situation, samples with noisy labels are guided to the wrong ID class, harming the model training. To tackle this issue, we adaptively adjust training loss, ensuring that samples with low confidence exert a lesser impact on the model compared to samples with high confidence by

$$\mathcal{L}_{\mathrm{id},i} = -\gamma_i \log \frac{\exp(\boldsymbol{x}_{i,y_i})}{\sum_{j=1}^{K} \exp(\boldsymbol{x}_{i,j})}, \tag{1}$$

where γ_i is the confidence of i-th input, $\boldsymbol{x}_{i,j}$ is the prediction score for i-th input being in class j, y_i is the ground-truth label of i-th input, K is the number of ID class.

2) Instance Loss: Furthermore, we introduce a new loss function that aims to bring features closer to those in the Instance Look-Up Table (ILUT). The ILUT is maintained based on the same instance number, which refers to the unique number assigned to each frame. We especially store normalized features on ILUT before starting every epoch. Our concept is that a complete positive sample in the dataset is identical to itself. Therefore, the instance loss aims to bring the input features closer to the features obtained from the same instance after the previous epoch. We also regard the farthest features in ILUT from the input's features as complete negative samples. Therefore, we argue that it is a more effective approach that low-confidence samples learn with complete positive and complete negative samples rather than only suppress noisy data by our $\mathcal{L}_{\mathrm{id}}$. In this term, we add the weighting term $1-\gamma$, which means the probability of a sample with a noisy label. So we train a model by

$$\mathcal{L}_{\mathrm{in},i} = w_{\mathrm{in}}(1-\gamma_i)\log(1+\exp(d(\boldsymbol{f}_i, \boldsymbol{g}_i^p) - d(\boldsymbol{f}_i, \boldsymbol{g}_i^n))), \tag{2}$$

where $\boldsymbol{f}_i$ is feature vector of i-th input, $\boldsymbol{g}_i^p$ is i-th feature vector of ILUT, and $\boldsymbol{g}_i^n$ is the features in ILUT most far from the $\boldsymbol{f}_i$, w_{in} is the hyper parameter, $d(\cdot,\cdot)$ means euclidean distance. Using a complete positive and negative sample is considered the most reasonable choice in settings with noisy labels, where accessing the actual correct label is challenging.

IV. Experimental Results

A. Experimental Settings

1) Benchmark Datasets with Noisy labels: We conduct experiments on three benchmark datasets, Market-1501 [7], CUHK03 [8] in various noisy label settings. We generate noisy labels in the same way as previous studies [4][3]. For the random noise setting, we exploit noisy labels by replacing the IDs of some percentages of labeled samples with randomly selected other IDs using a fixed random seed. For the pattern noise setting, which is more similar to real-world conditions, we assign the secondary closest class as noisy labels.

2) Implementation Details: Our experiments are conducted on a single NVIDIA RTX-3090 GPU using PyTorch. We utilize a simple baseline, which is composed of ImageNet pre-trained ResNet50 [18] backbone with two fully connected layers and one batch normalization while training with general cross-entropy loss. The initial learning rate is set to $10-2$, adjusted to $10-3$ at epoch 20, and then decreased to $10-5$ beyond epoch 50. The batch size is set to 32. The weighting parameter w_{id} is set to $10^{-}3$. Data transformations include random crop to 256×128 and random horizontal flip with a probability of 0.5.

We also conduct experiments on the two supervised methods as a strong baseline, TransReID [9] and CDNet [10], by exchanging their all of cross-entropy losses and triplet losses with our $\mathcal{L}_{\mathrm{id}}$ and $\mathcal{L}_{\mathrm{in}}$ respectively. The initial learning rate is set to $10^{-}1$ and decays by multiplying $10^{-}2$ for every 20 epochs. Other settings are the same as their paper.

Method	Random Noise						Pattern Noise			
	10%		20%		30%		10%		20%	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
DNet [13]	82.3	61.5	77.0	53.4	77.1	44.3	52.4	27.0	49.3	24.4
PurifyNet [4]	84.2	64.3	83.1	63.1	81.4	60.2	81.8	63.2	77.8	56.2
CORE [3]	85.5	67.7	84.1	66.2	81.2	60.6	83.2	66.0	81.3	62.4
LRP [15]	88.9	71.4	88.8	70.5	86.3	67.0	88.5	70.7	86.7	67.6
LRNI-HSR [16]	86.9	68.8	84.5	65.6	81.9	62.2	85.3	67.7	84.1	65.6
ICLR [17]	88.0	70.8	87.2	70.3	86.4	69.7	86.8	68.0	83.6	64.9
TSNT [14]	-	-	90.3	76.0	-	-	-	-	90.9	76.4
Proposed	90.4	69.0	90.0	68.3	88.0	61.6	89.0	64.9	86.3	59.3

TABLE I
COMPARISON OF EXISTING METHODS AND OUR PROPOSED METHODS ON MARKET-1501 WITH RANDOM NOISE AND PATTERN NOISE.

Method	Market-1501		CUHK03	
	R-1	mAP	R-1	mAP
CDNet [10]	86.0	63.1	27.6	27.2
+Proposed	87.4	69.4	29.9	29.4
TransReID [9]	87.9	70.8	45.6	41.6
+Proposed	**91.8**	**79.2**	**50.3**	**47.4**

TABLE II
COMPARISION OF EXISTING METHODS AND OUR PROPOSED METHODS ON STRONG BASELINE [10][9] WITH 20% NOISE LEVEL.

Method	Random Noise						Pattern Noise			
	10%		20%		30%		10%		20%	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
DNet [13]	32.3	31.8	24.3	24.2	16.8	17.4	10.5	10.9	8.8	9.5
PurifyNet [4]	32.8	32.8	30.2	29.2	26.6	26.4	33.6	32.9	29.2	29.2
CORE [3]	40.4	39.6	34.4	35.0	34.6	34.7	38.6	37.2	34.4	33.2
TSNT [14]	-	-	49.4	48.3	-	-	-	-	50.2	49.0
Proposed	41.1	39.1	39.7	37.6	31.0	30.5	36.6	33.4	30.6	28.1

TABLE III
COMPARISON OF EXISTING METHODS AND OUR PROPOSED METHODS ON CUHK03 WITH RANDOM NOISE AND PATTERN NOISE.

Method			Market-1501		CUHK03	
$\mathcal{L}_{\text{id}}$	$\mathcal{L}_{\text{in}}$	GR	R-1	mAP	R-1	mAP
			86.1	56.7	23.4	23.3
✓			88.7	64.9	34.4	32.6
✓	✓		88.6	63.7	34.4	32.8
✓		✓	88.8	66.8	32.1	31.9
✓	✓	✓	**90.0**	**68.3**	**39.7**	**37.6**

TABLE IV
ABLATION STUDY OF OUR PROPOSED METHODS WITH 20% NOISE LEVEL.

3) Evaluation Metrics: To evaluate the model's Re-ID performance, we utilize the Rank-k(R-k) metric and Mean Average Precision(mAP). R-k indicates how many nearest k persons are in the true positive case. The mAP metric represents the mean of average precision, denoting the ratio of true positives among the model predictions.

B. Comparison with Robust Re-ID with Noisy Labels

TABLE I, TABLE III illustrates that our approach is competitive compared to other state-of-the-art methods on two datasets, Market-1501 and CUHK03. In the table, red indicates the highest performance, while blue indicates the second highest performance. Some mAP performances are lower than those of existing methods compared to the Rank-1 score relatively. Our $\mathcal{L}_{\text{in}}$ learns to make i-th feature behave similarly to the information stored in i-th feature in ILUT during training. As a result, if the query and gallery images have the same viewpoint or pose, our loss effectively improves the Rank-1 score. In terms of mAP, it may lead to incorrect matching results because the model can be fitted to specific viewpoints and poses. However, we argue that treating each sample as a complete positive sample is the best choice under the noisy label setting.

C. Flexibility of Usability

We also apply our proposed methods to state-of-the-art supervised Re-ID networks, such as those presented in [9] and [10]. Despite the strong performance of these networks, ID noise degrades their performance. However, our proposed method significantly improves performance, as shown in TABLE II. Some existing methods that utilize the co-teaching [3][14] approach have the drawback of doubling the parameters when applied to these supervised techniques. However, the proposed method has the advantage of not requiring any additional parameters. Additionally, using TransReID along with the proposed method outperforms existing methods in TABLE I and TABLE III. It is important to note that our method does not require additional parameters, making it easy to apply to any re-identification networks.

D. Ablation Study

1) Quantitative Results: Table IV is the ablation study of each component of our proposed methods. First of all, our $\mathcal{L}_{\text{id}}$ suppress low-confidence samples when model training, so it is effective under the data containing some incorrect labels. Furthermore, we conducted experiments to demonstrate the effectiveness of the $\mathcal{L}_{\text{in}}$, refer to the third low of the table. Our confidence-aware learning adjusts the two loss functions by balancing the loss scale adaptively utilizing the probability of the sample having the correct label. This allows the model to be optimized correctly even if the sample has low confidence. GMM refinement(GR) enhances confidence estimation in situations where few noise samples are included within an ID class. Therefore, it improves the performance of the loss that reflects confidence, denoted as $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{in}}$.

2) GMM Refinement: Our GMM refinement overcomes situations where the intra-class has significantly different proportions of correct labels and noisy labels. When there are no noisy samples within a class, a two-component GMM is created based on the loss values of clean samples. This implies a tendency that outlier samples within the intra-class will have lower confidence values, leading to GMM's second component representing clean sample loss values. In Fig. 4, the third sample among the women wearing white t-shirts has a dark background, making it seem different from other samples. Before applying the proposed technique, this clean sample has a low value of 0.02 despite having a correct label. However, after applying GMM refinement as shown on the right, it can obtain a high confidence value of 0.95, which can correctly

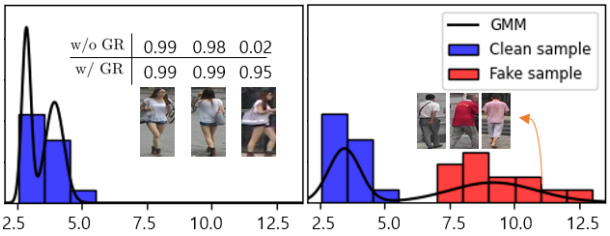


Fig. 4. Loss value distribution and visualization of correctly labeled samples with estimated confidence in the same ID class on Market-1501.

influence the training process through our confidence-aware learning.

V. Conclusion

In this paper, we propose a universal technique that can be adapted to any re-identification network without additional learning parameters. To alleviate the confusion caused by noisy labels, we reflect the sample's confidence to model training with our confidence-aware learning. We also devise the GMM refinement to get a more reliable model for loss values even if the intra-class has few samples for noise distribution. Experimental results show that our proposed method, when applied to a strong baseline, outperforms existing methods that learn with noisy labels.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00392536), and by the Institute of Information & communications Technology Planning & Evaluation(IITP) grants funded by the Korea government(MSIT), (RS-2020-II201336, Artificial Intelligence Graduate School Program(UNIST)) and (RS-2021-II212068, Artificial Intelligence Innovation Hub).

References

[1] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," *Toronto, ON, Canada*, 2009.

[2] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Worksh.*, 2011, pp. 806–813.

[3] M. Ye, H. Li, B. Du, J. Shen, L. Shao, and S. C. Hoi, "Collaborative refining for person re-identification with label noise," *IEEE Trans. Image Process.*, vol. 31, pp. 379–391, 2021.

[4] M. Ye and P. C. Yuen, "Purifynet: A robust person re-identification model with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2655–2666, 2020.

[5] Q. Wei, L. Feng, H. Sun, R. Wang, C. Guo, and Y. Yin, "Fine-grained classification with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 11 651–11 660.

[6] Z. Huang, J. Zhang, and H. Shan, "Twin contrastive learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 11 661–11 670.

[7] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[8] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 152–159.

[9] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 15 013–15 022.

[10] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 6729–6738.

[11] B. Han, Q. Yao, X. Yu, *et al.*, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8536–8546.

[12] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 13 726–13 735.

[13] T. Yu, D. Li, Y. Yang, T. M. Hospedales, and T. Xiang, "Robust person re-identification by modelling feature uncertainty," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 552–561.

[14] M. Liu, F. Wang, X. Wang, Y. Wang, and A. K. Roy-Chowdhury, "A two-stage noise-tolerant paradigm for label corrupted person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, pp. 4944–4956, 2024.

[15] Y. Chen, M. Liu, X. Wang, F. Wang, A.-A. Liu, and Y. Wang, "Refining noisy labels with label reliability perception for person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 9479–9490, 2023.

[16] X. Zhong, S. Su, W. Liu, X. Jia, W. Huang, and M. Wang, "Neighborhood information-based label refinement for person re-identification with label noise," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2023, pp. 1–5.

[17] X. Zhong, X. Han, X. Jia, *et al.*, "Iclr: Instance credibility-based label refinement for label noisy person re-identification," *Pattern Recog.*, vol. 148, p. 110 168, 2024.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

Test-Time Optimization for Post-Processing of Compressed Videos

Hongil Kim*, Changwoo Han*, Donghyun Kim†, Sung-Chang Lim† and Seung-Won Jung*
* Department of Electrical Engineering, Korea University, Seoul
E-mail: {yourhong1, hcwoo329, swjung83}@korea.ac.kr
† Hyper-Reality Metaverse Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon
E-mail: {kimddng, sclim}@etri.re.kr

***Abstract*— Lossy compression is widely used for video compression, but it often introduces compression artifacts that degrade the visual quality of compressed videos. Consequently, numerous deep learning-based methods have been developed to post-process compressed videos. However, previous post-processing models often encounter difficulties when there is a domain gap between the training and test datasets. Test-time optimization (TTO), a technique that finetunes the model during the test stage, has been considered an effective solution to address the domain gap problem. In this paper, we introduce a novel TTO method specialized for compression artifacts reduction. Specifically, we propose using image pairs available on the decoder-side, *i.e.*, the images before and after the adaptive loop filtering of the versatile video coding standard, as input and target of TTO such that the post-processing model can be adapted to the characteristics of test data. Experimental results on several baseline models and test datasets demonstrate the effectiveness of the proposed method in post-processing compressed videos.**

I. Introduction

With the increasing demand for multimedia content, video traffic occupies the majority of worldwide internet traffic. To address the exponential growth of video traffic, the Joint Video Experts Team (JVET) established the latest video coding standard called Versatile Video Coding (VVC) [1] as a successor to High Efficiency Video Coding (HEVC) [2]. These video coding standards aim to reconstruct high-quality video frames while requiring fewer bits for encoding. However, the block-based coding structure inevitably leads to compression artifacts, such as blocking and ringing artifacts, which significantly degrade image quality.

To alleviate compression artifacts, VVC adopted several filters, including deblocking filter (DBF) [3], sample adaptive offset (SAO) [4], and adaptive loop filter (ALF) [5]. DBF attenuates blocking artifacts across block boundaries primarily caused by block-wise transform and quantization. Furthermore, SAO and ALF perform adaptive adjustments to the decoded frame to reduce the difference between the decoded and uncompressed frames. Although these filters effectively reduce compression artifacts, there is still ample room for improving the quality of decoded frames. Inspired by the recent success of deep learning in low-level vision tasks [6], [7], [8], [9], [10], [11], [12], many researchers have explored convolutional neural network (CNN)-based compression artifacts reduction filters [13], [14], [15], [16], [17], [18]. CNN-based filters have demonstrated higher performance in reducing compression artifacts than handcrafted filters due to their strong ability to capture non-linear degradation in compressed frames.

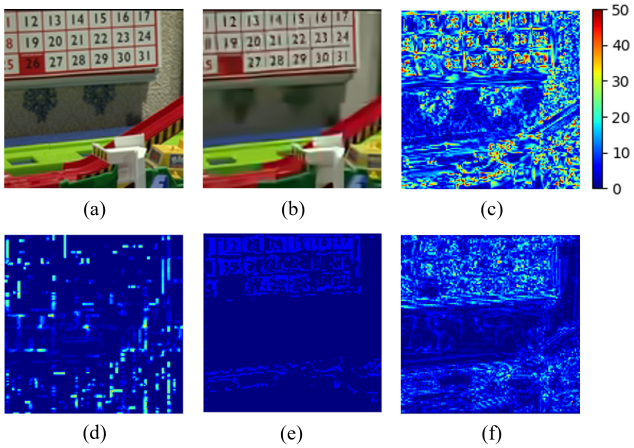


Fig. 1. Comparisons of the difference signals: (a) Original image, (b) decoded image (after ALF), (c) difference between the original and decoded images, (d) difference of before/after DBF, (e) difference of before/after SAO and (f) difference of before/after ALF. Absolute difference values are colored for visualization.

Nonetheless, CNN-based filters are inherently sensitive to the domain gap between training and test datasets. In particular, since compression artifacts depend on video coding parameters, such as the quantization parameter (QP), as well as video characteristics, it is very challenging to handle images with diverse distributions of compression artifacts using CNNs trained on specific training datasets. To this end, several methods have attempted to train individual CNNs for a set of QPs [19], [20], train a CNN using images compressed with different QPs [21], or even combine multiple CNNs [22]. However, such trained CNNs may still require more ability to handle images with different compression artifacts at the test stage.

Test-time optimization (TTO) has been considered one of the solutions for improving the performance of pre-trained CNNs for test data. Since ground-truth (GT) labels do not exist for test data, several novel ideas have been introduced in classification [23], human pose estimation [24], super-resolution [25], video frame interpolation [26], video segmentation [27], video object segmentation [28] tasks to finetune the CNN

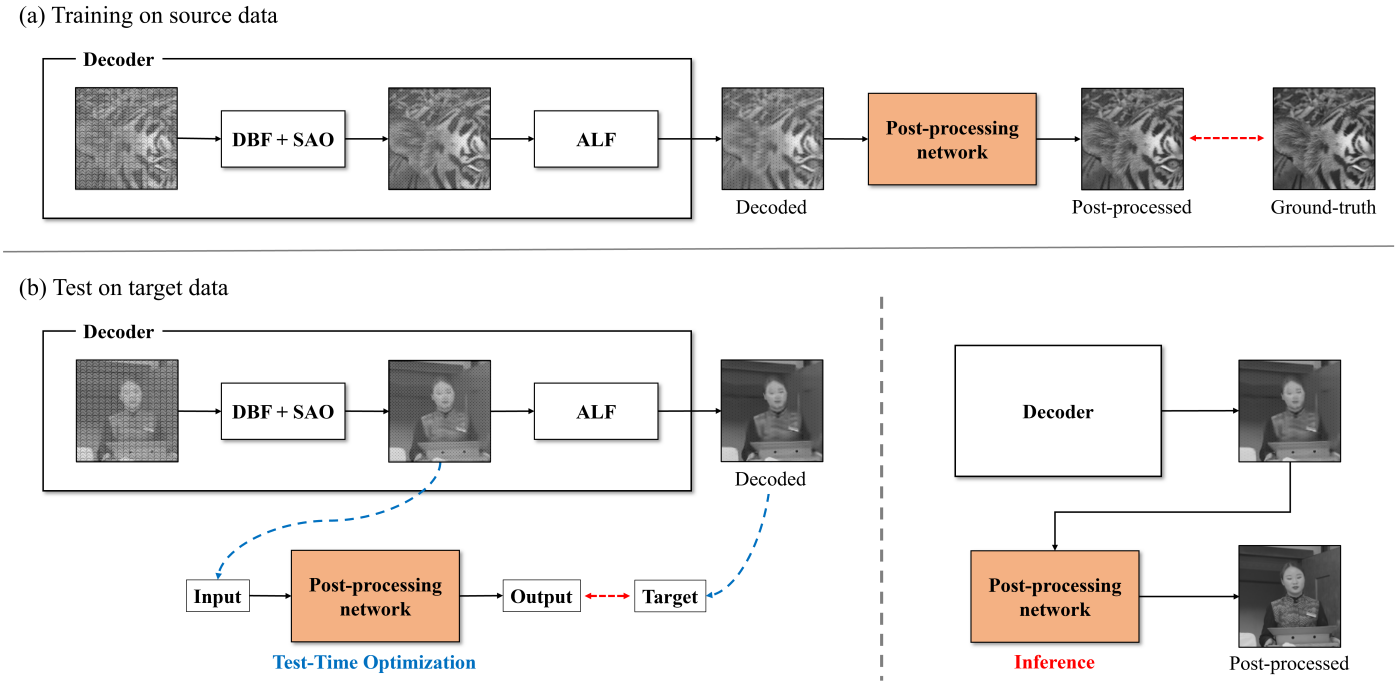


Fig. 2. The overall pipeline of the proposed method: (a) The post-processing network is trained on the source data using GT labels; (b) The post-processing network is first finetuned on the test data (left) and then applied to obtain artifacts-reduced video frames (right).

without GT labels. However, these methods [23], [24], [25], [27], [28] incorporate image generation or rely on self-training and pseudo-labeling, which demand significant computational capacity or suffer from the challenges posed by noisy pseudo-labels [29].

In this paper, we make the first attempt to apply TTO for compression artifacts reduction. In our case study of artifacts reduction in videos compressed by the VVC standard [1], we notice that the input and output of the ALF can effectively serve as the input and target for TTO of the CNN, respectively, to overcome the aforementioned challenges. Specifically, using a few input and output image pairs of the ALF, we update the parameters of the CNN for a few steps such that the CNN can better handle test data. From extensive experimental results, we show the effectiveness of the proposed TTO method.

II. Proposed Method

A. Motivation

Assume that we have a post-processing CNN trained on uncompressed images and their corresponding compressed ones as the GT labels and input, respectively. Since uncompressed images do not exist on the decoder side, we need surrogate pairs of input and target images for TTO of the post-processing CNN. Fortunately, the VVC standard, as well as many other off-the-shelf video codecs, perform several in-loop filtering operations during decoding; thus, the input and output images of the in-loop filters are still available on the decoder side. Among the three adopted in-loop filters of the VVC, *i.e.*, DBF [3], SAO [4], and ALF [5], the difference between the input and output images of ALF best mimics the one between the uncompressed and compressed images, as illustrated in Fig. 1. We thus propose to use this valuable input and output image pair of ALF as the input and target images for TTO, respectively, to adapt the post-processing CNN on test data.

B. Test-Time Optimization

Fig. 2(a) shows the training procedure. Let f_θ denote a post-processing CNN parameterized by θ, which is trained to minimize the following:

$$\theta^* = \arg\min_\theta \sum_i \left\| f_\theta\left(\hat{I}_i\right) - I_i \right\|, \tag{1}$$

where $\hat{I}_i$ and I_i are the i-th pair of decoded and uncompressed images, and $\|\cdot\|$ measures the distance between the network output and target uncompressed images. Given the training dataset $\mathcal{I} = \left\{\hat{I}_i, I_i\right\}_{i=1}^{N_T}$, where N_T is the number of images, we can obtain a set of optimized parameters θ^* that work well on average for the training images. To make the post-processing CNN more effective on test data with diverse distributions, we propose an algorithm that can adapt network parameters to test data.

Fig. 2(b) illustrates the proposed TTO procedure. Let $\hat{J}_i$ denote the i-th decoded video frame on the decoder side. Since its corresponding GT image J_i is not available, we instead extract an image before ALF, denoted as $\tilde{J}_i$, and use it as input for TTO of the post-processing CNN. Specifically, the network parameter is first initialized as θ^* in (1) and then finetuned on

test data as follows:

$$\theta^{**} = \arg\min_{\theta} \sum_{i \in \Upsilon_i} \left\| f_\theta \left(\tilde{J}_i \right) - \hat{J}_i \right\|, \tag{2}$$

where Υ_i is a set of frame indices used for TTO. Since video characteristics change over time, TTO has to be applied continuously during decoding. In our implementation, we apply TTO at the group of pictures (GOP)-level, a set of 32 consecutive frames in VVC [1]. In each frame, ALF is selectively applied according to the rate-distortion cost [1]. Therefore, we define Υ_i as a set of indices of frames belonging to the same GOP with $\hat{J}_i$ that ALF processes. Once a GOP-level TTO is completed, all frames in the GOP are post-processed by $f_{\theta^{**}}$ to have artifacts-reduced video frames, as illustrated in Fig. 2(b). If no frame is processed by ALF in a GOP, TTO is not applied.

C. Implementation Details

Considering that the residual block (ResBlock) is one of the most common building blocks of recent image restoration models [10], [30], [31], [32], we design a ResBlock-based baseline network (denoted as M1), as shown in Fig. 3(a) for the performance verification of the proposed TTO. M1 contains 24 residual blocks, each composed of two 3×3 convolutional layers and a PReLU activation layer. The global skip connection is also included to ease the training. The network complexity of M1 is 58,600 multiply-accumulate operations per pixel (MAC/pixel). In addition, the UNet architecture is widely used in image restoration tasks, such as deblurring [33], [34] and denoising [35], [36], [37]; thus, we design a baseline network based on the UNet architecture (denoted as M2), as shown in Fig. 3(b). M2 consists of three levels of convolution blocks, down-blocks, and up-blocks. Each of these blocks is composed of 3×3 convolutional layers and PReLU activation functions. Additionally, each down-block and up-block incorporates a 2×2 convolution and transposed convolution layer, respectively. Skip connections are included between each level's down-block and up-block. The complexity of M2 is 104.3 kMAC/pixel, and the numbers of network parameters of M1 and M2 are 1.78M and 4.92M, respectively.

Since the proposed TTO is designed to alleviate the performance degradation of the post-processing CNN caused by the domain gap between training and test data, the required number of parameter update steps of (2) can differ for each video sequence or even for each GOP of the same video sequence. To this end, we devise a simple but effective strategy that terminates the parameter update if the following condition is satisfied:

$$\frac{1}{|\Upsilon_i|} \sum_{i \in \Upsilon_i} \left\| f_{\theta_{t+1}} \left(\tilde{J}_i \right) - f_{\theta_t} \left(\tilde{J}_i \right) \right\| \; < \; T_h \tag{3}$$

where $|\cdot|$ represents the cardinality of a set, θ_t represents a set of network parameters after the t-th update, and T_h is a threshold, empirically chosen as 0.0003. In other words, TTO is terminated if it leads to a marginal change after a certain update step. We also terminate TTO if the number of updates reaches the threshold N_h, empirically chosen as 6.

Inspired by [27], [38], [39], [40], we aim to preserve domain-agnostic information of the pre-trained model while updating only domain-specific features. To achieve this, instead of updating the parameters of all layers in the post-processing CNN, we opt to freeze earlier layers to retain low-level features as well as the last-stage image reconstruction layers. Consequently, the proposed TTO aims to learn high-level content-specific features by finetuning only the second half of the ResBlocks in the M1 model, as well as the Up-block section in the M2 model. Fig. 3 indicates the trainable layers by TTO. Last, since ALF obtains filter coefficients by minimizing the L2 loss between the original and restored images [41], we use the L2 loss in (2)-(3).

TABLE I
PERFORMANCE EVALUATION IN BD-RATE (%) UNDER RA CONFIGURATION FOR THE JVET-CTC TEST SEQUENCES.

Class	Sequence	M1	M1+TTO	M2	M2+TTO
B	MarketPlace	-3.09	-3.25	-1.13	-1.23
	RitualDance	-3.93	-4.13	-1.43	-1.49
	Cactus	-2.50	-2.83	-0.28	-0.66
	BasketballDrive	-3.66	-3.88	-0.47	-0.95
	BQTerrace	-1.05	-2.04	+0.62	-0.07
C	BasketballDrill	-3.88	-4.40	-0.70	-1.01
	BQMall	-4.45	-5.00	-1.01	-1.35
	PartyScene	-4.20	-4.48	-0.57	-0.73
	RaceHorsesC	-2.41	-2.75	-0.50	-0.53
D	BasketballPass	-5.89	-6.02	-1.66	-1.83
	BQSquare	-9.44	-9.91	-1.98	-2.33
	BlowingBubbles	-4.47	-4.62	-0.83	-0.99
	RaceHorses	-4.79	-4.87	-1.40	-1.57
Total	B Class	-2.85	-3.23	-0.54	-0.88
	C Class	-3.76	-4.16	-0.69	-0.90
	D Class	-6.15	-6.35	-1.47	-1.68
	Average	**-4.14**	**-4.48**	**-0.82**	**-1.13**

III. EXPERIMENTAL RESULTS

For performance evaluation, we used videos compressed by the VVC reference software VTM-11.0 using the random access (RA) configuration and five QPs = {22, 27, 32, 37, 42}. The BVI-DVC dataset [42] and the class B-D JVET-CTC test sequences [43] were used as training and test data, respectively. The BD-rate [44] of the Y channel was used as a performance metric, which represents the percentage of the average bit-saving over the VTM-11.0 baseline.

During the training stage, the baseline network was trained for a total of 100 epochs via (1) using the Adam optimizer with an initial learning rate of 0.0001. The learning rate decayed at every 20 epochs after 30 epochs with a decaying rate of 0.1. We trained five networks separately for the five chosen QPs. In the test stage, TTO was applied at the GOP-level using the ALF-applied frames. From the input and output image pairs of ALF, we randomly cropped three patches with a size of

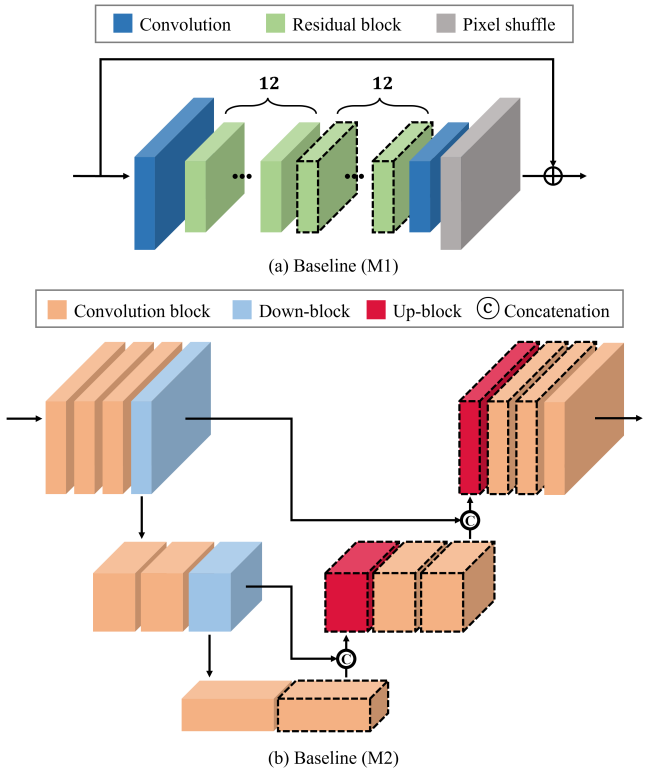


Fig. 3. Post-processing network architectures used in our experiments: (a) ResBlock-based baseline (M1) and (b) UNet-based baseline (M2). Only the dotted blocks are updated during TTO.

TABLE II
EXPERIMENTAL RESULTS ON THE NUMBER OF NETWORK UPDATE STEPS. THE PERFORMANCE WAS EVALUATED AS THE AVERAGE BD-RATE AND THE AVERAGE PROCESSING TIME PER GOP FOR THE CLASS B-D TEST SEQUENCES.

M1	N_h					
	0	2	4	6	8	10
BD-rate (%)	-4.14	-4.30	-4.41	-4.48	-4.49	-4.49
Time (s)	4.3	6.0	8.5	10.8	13.2	15.7

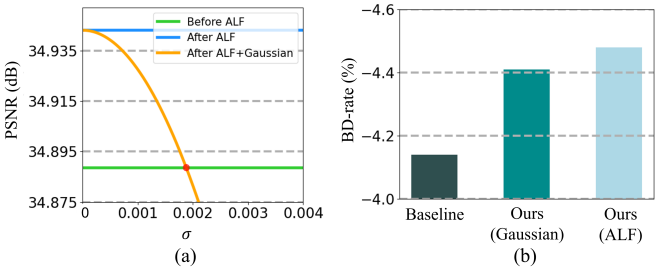


Fig. 4. Comparison of the TTO results obtained using different training datasets: (a) The PSNRs between the degraded and uncompressed (original) frames, where we add the Gaussian noise to the decoded (after ALF) frames of the JVET-CTC test sequences and obtain the noise standard deviation that results in the similar average PSNR of the frames before ALF, which is 0.0018 at QP=32; (b) Experimental results in BD-rates, where the Gaussian noise-added frames (Ours, Gaussian) and the frames before ALF (Ours, ALF) are used as input for TTO, respectively. The noise standard deviations are chosen to match the average PSNRs of the frames before ALF for all five tested QPs. Both TTO results show better performance than M1 while using the input and output image pairs of ALF for TTO yields a higher BD-rate gain.

256×256 and trained the network using (2) with the Adam optimizer and a fixed learning rate of 10e-6.

Table I shows the BD-rate results. Compared to the VTM-11.0 anchor, our post-processing networks, *i.e.*, M1 and M2, led to an improvement in the BD-rate of 4.14% and 0.82%, respectively. The proposed TTO resulted in 4.48% (M1+TTO) and 1.13% (M2+TTO), indicating its effectiveness. Note that TTO decreased the BD-rates for all test sequences, with a maximum gain of 0.99% for BQTerrace and a minimum gain of 0.08% for RaceHorses in M1 and a maximum gain of 0.69% for BQTerrace and a minimum gain of 0.03% for RaceHorsesC in M2.

Next, we obtained the experimental results using different values of N_h, as shown in Table II. As N_h increased, the average processing time of TTO per GOP increased approximately linearly. When $N_h = 0$, the pretrained post-processing model was applied without TTO. When $N_h \neq 0$, the processing time for TTO and the inference time after TTO were added together. Based on our experiments, with $N_h = 6$, the processing time ranged from a minimum of 4.3 seconds to a maximum of 10.8 seconds, depending on the test sequence characteristics. Note that the BD-rates did not continuously increase as the number of network updates increased. Since not a real image pair was used for TTO, but a surrogate image pair, we found performance saturation after certain update steps and thus chose $N_h = 6$ as the default value for both baselines, considering the increased processing time by TTO.

Lastly, to verify the effectiveness of using the image pairs before and after ALF as training images for TTO, we added Gaussian noise to the decoded image $\hat{J}_i$ and obtained its noisy version, denoted as $\overline{J}_i$. For a fair comparison, the standard deviation of the Gaussian noise was determined to make the average PSNR between $\hat{J}_i$ and $\overline{J}_i$ and that between $\hat{J}_i$ and $\tilde{J}_i$ similar, as shown in Fig. 4(a). For this experiment, the proposed TTO was applied while replacing $\tilde{J}_i$ by $\overline{J}_i$ in (2). As shown in Fig. 4(b), the proposed TTO enables improved performance in both cases compared to M1, indicating the effectiveness of using test-time data for finetuning. In addition, the use of the frames before and after ALF was found to be better than the use of noisier-to-noisy image pairs obtained by Gaussian noise addition [45], supporting the effectiveness of using compression-specific distortion for TTO purposes. More results can be found on the project page[1].

IV. CONCLUSION

This paper introduced a new use case of TTO for compression artifacts reduction. The proposed TTO method uses the signals before and after the in-loop filter of VVC, which are also available on the decoder side. By finetuning the post-processing model at the test stage using these signals as training data, we showed that compression artifacts could be

[1]https://yourhong1.github.io/PP-TTO/

reduced for test video sequences with different characteristics. We believe that our method provides an effective way of optimizing pre-trained models during the test stage, paving the way for further developments in artifacts reduction and quality improvement of compressed videos. Moreover, while our approach has focused on post-processing filters, the potential to extend TTO to other stages within the video codec, such as the in-loop filter or scaling techniques, presents a valuable direction for future research. By applying TTO to these components, we could further enhance overall BD-rate performance. This flexibility suggests that TTO could serve as a versatile tool for solving various challenges in video compression.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2017-0-00072, Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Teramedia

REFERENCES

[1] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, 2021.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] P. List, A. Joch, J. Lainema, G. Bjontegaard, and M. Karczewicz, "Adaptive deblocking filter," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 614–619, 2003.

[4] C.-M. Fu, E. Alshina, A. Alshin, Y.-W. Huang, C.-Y. Chen, C.-Y. Tsai, C.-W. Hsu, S.-M. Lei, J.-H. Park, and W.-J. Han, "Sample adaptive offset in the hevc standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1755–1764, 2012.

[5] C.-Y. Tsai, C.-Y. Chen, T. Yamakage, I. S. Chong, Y.-W. Huang, C.-M. Fu, T. Itoh, T. Watanabe, T. Chujoh, M. Karczewicz, and S.-M. Lei, "Adaptive loop filtering for video coding," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 6, pp. 934–945, 2013.

[6] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 38, no. 2, pp. 295–307, 2016.

[7] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, July 2017.

[8] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, June 2018, pp. 2472–2481.

[9] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2018, pp. 286–301.

[10] D.-W. Kim, J. R. Chung, and S.-W. Jung, "GRDN: Grouped residual dense network for real image denoising and GAN-based real-world noise modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit. Workshops*, 2019, pp. 2086–2094.

[11] M. Zhao, G. Cao, X. Huang, and L. Yang, "Hybrid transformer-cnn for real image denoising," *IEEE Signal Process. Lett.*, vol. 29, pp. 1252–1256, 2022.

[12] J. Wu, Y. Wang, and X. Zhang, "Lightweight asymmetric convolutional distillation network for single image super-resolution," *IEEE Signal Process. Lett.*, vol. 30, pp. 733–737, 2023.

[13] F. Zhang, C. Feng, and D. R. Bull, "Enhancing VVC through CNN-based post-processing," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2020, pp. 1–6.

[14] F. Zhang, D. Ma, C. Feng, and D. R. Bull, "Video compression with CNN-based postprocessing," *IEEE MultiMedia*, vol. 28, no. 4, pp. 74–83, 2021.

[15] W.-S. Park and M. Kim, "CNN-based in-loop filtering for coding efficiency improvement," in *Proc. IEEE Image, Video, and Multidimensional Signal Processing Workshop*, 2016, pp. 1–5.

[16] Y. Zhang, T. Shen, X. Ji, Y. Zhang, R. Xiong, and Q. Dai, "Residual highway convolutional neural networks for in-loop filtering in HEVC," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3827–3841, 2018.

[17] D. Luo, M. Ye, S. Li, and X. Li, "Coarse-to-fine spatio-temporal information fusion for compressed video quality enhancement," *IEEE Signal Process. Lett.*, vol. 29, pp. 543–547, 2022.

[18] Z. Wang, M. Ye, S. Li, and X. Li, "Multi-frame compressed video quality enhancement by spatio-temporal information balance," *IEEE Signal Process. Lett.*, vol. 30, pp. 105–109, 2023.

[19] C. Liu, H. Sun, J. Katto, X. Zeng, and Y. Fan, "A convolutional neural network-based low complexity filter," *arXiv preprint arXiv:2009.02733*, 2020.

[20] J. Deng, L. Wang, S. Pu, and C. Zhuo, "Spatio-temporal deformable convolution for compressed video quality enhancement," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 10 696–10 703.

[21] P. Svoboda, M. Hradis, D. Barina, and P. Zemcik, "Compression artifacts removal using convolutional neural networks," *arXiv preprint arXiv:1605.00366*, 2016.

[22] S.-J. Cho, J. R. Chung, S.-W. Kim, S.-W. Jung, and S.-J. Ko, "Compression artifacts reduction using fusion of multiple restoration networks," *IEEE Access*, vol. 9, pp. 66 176–66 187, 2021.

[23] W. Ma, C. Chen, S. Zheng, J. Qin, H. Zhang, and Q. Dou, "Test-time adaptation with calibration of medical image classification nets for label distribution shift," in *Proc. Int. Conf. Med. image Comput. Comput.-Assist. Intervention*, 2022, pp. 313–323.

[24] J. Zhang, X. Nie, and J. Feng, "Inference stage optimization for cross-scenario 3d human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 2408–2419.

[25] M. S. Rad, T. Yu, B. Bozorgtabar, and J.-P. Thiran, "Test-time adaptation for super-resolution: You only need to overfit on a few more images," in *Proc. Int. Conf. Comput. Vision Workshops*, October 2021, pp. 1845–1854.

[26] M. Choi, J. Choi, S. Baik, T. H. Kim, and K. M. Lee, "Test-time adaptation for video frame interpolation via meta-learning," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 44, no. 12, pp. 9615–9628, 2021.

[27] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2022, pp. 7201–7211.

[28] F. Azimi, S. Palacio, F. Raue, J. Hees, L. Bertinetto, and A. Dengel, "Self-supervised test-time adaptation on video data," in *Proc. Winter Conf. Appl. Comput. Vision*, 2022, pp. 3439–3448.

[29] D. Chen, D. Wang, T. Darrell, and S. Ebrahimi, "Contrastive test-time adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 295–305.

[30] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 43, no. 7, pp. 2480–2495, 2020.

[31] Y. Zhao, Y. Xu, Q. Yan, D. Yang, X. Wang, and L.-M. Po, "D2hnet: Joint denoising and deblurring with hierarchical network for robust night image restoration," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2022, pp. 91–110.

[32] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for fast image restoration and enhancement," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 45, no. 2, pp. 1934–1948, 2022.

[33] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. Int. Conf. Comput. Vision*, 2021, pp. 4641–4650.

[34] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2022, pp. 17–33.

[35] J. Gurrola-Ramos, O. Dalmau, and T. E. Alarcón, "A residual dense u-net neural network for image denoising," *IEEE Access*, vol. 9, pp. 31 742–31 754, 2021.

[36] F. Jia, W. H. Wong, and T. Zeng, "DDUNet: Dense dense u-net with applications in image denoising," in *Proc. Int. Conf. Comput. Vision*, 2021, pp. 354–364.

[37] C.-M. Fan, T.-J. Liu, and K.-H. Liu, "SUNet: Swin transformer UNet for image denoising," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2022, pp. 2333–2337.
[38] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 512–523.
[39] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vision*. Springer, 2014, pp. 818–833.
[40] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.
[41] F. Jin, P. Fieguth, L. Winger, and E. Jernigan, "Adaptive wiener filtering of noisy images and image sequences," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3, 2003, pp. 349–352.
[42] D. Ma, F. Zhang, and D. R. Bull, "Bvi-dvc: A training database for deep video compression," *IEEE Trans. Multimedia*, vol. 24, pp. 3847–3858, 2021.
[43] F. Bossen, J. Boyce, X. Li, V. Seregin, and K. Sühring, "VTM common test conditions and software reference configurations for sdr video (JVET-T2010)," *Joint Video Experts Team*, 2020.
[44] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *ITU SG16 Doc. VCEG-M33*, 2001.
[45] N. Moran, D. Schmidt, Y. Zhong, and P. Coady, "Noisier2noise: Learning to denoise from unpaired noisy data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12 064–12 072.

Lifelong Person Re-Identification with Backward-Compatibility

Minyoung Oh[1] and Jae-Young Sim[1,2,†]
[1]Graduate School of Artificial Intelligence, [2]Department of Electrical Engineering,
Ulsan National Institute of Science and Technology, Republic of Korea
E-mail: {mmyy2513, jysim}@unist.ac.kr

***Abstract*—Lifelong person re-identification (LReID) assumes a practical scenario where the model is sequentially trained on continuously incoming datasets while alleviating the catastrophic forgetting in the old datasets. In such a case, not only the training datasets but the gallery images are also incrementally accumulated, that requires a huge amount storage space to store the gallery images as well as computational complexity to extract the features at the inference phase. In this paper, we address the above mentioned problem by incorporating the backward-compatibility to LReID based on the replay scheme. We attempt to train the model using the continuously incoming datasets while maintaining the model's compatibility toward the previously trained old models without re-computing the features of the old gallery images. Specifically, we design the cross-model compatibility loss based on the contrastive learning with respect to the replay features across all the old datasets. Experimental results demonstrate that the proposed method achieves a significantly higher performance of the backward-compatibility compared with the existing LReID methods.**

I. Introduction

Person re-identification (ReID) aims to retrieve specific individuals, corresponding to a given query, from the galleries of person images. It has drawn much attention in various practical applications such as surveillance and security. Note that real-world scenarios of ReID are subject to changing datasets of different domains which are collected at different places and time instances. Most of the conventional ReID methods, developed in the supervised learning manner, perform model retraining whenever a new dataset is given [1], [2]. Therefore, retraining the model by using the whole accumulated datasets together requires a huge amount of memory space and dramatically increases the computational complexity. To reduce the burden of entire retraining while maintaining the performance across different domains, the ReID models should learn the knowledge of the changing datasets incrementally.

For this purpose, lifelong person ReID (LReID) has been introduced [3]–[7], which sequentially trains the model on continuously incoming datasets across different domains. Specifically, it attempts to alleviate the catastrophic forgetting, where the model prefers the knowledge learned from the recent dataset while forgetting the old knowledge learned from the previously trained datasets. However, in practice, the challenges lie not only on the training but also on the inference. Whenever the model is updated with a new training dataset based on the LReID scenario, the feature space is also changed. Thus, at the inference stage, the updated model should re-compute the features from all the accumulated gallery images again to ensure that the features are compatible with that of the query images extracted by the new model. We call this process as *backfilling*. Since the gallery images are also continuously collected, the backfilling process takes a substantial amount of time especially for long-term deployment of ReID systems.

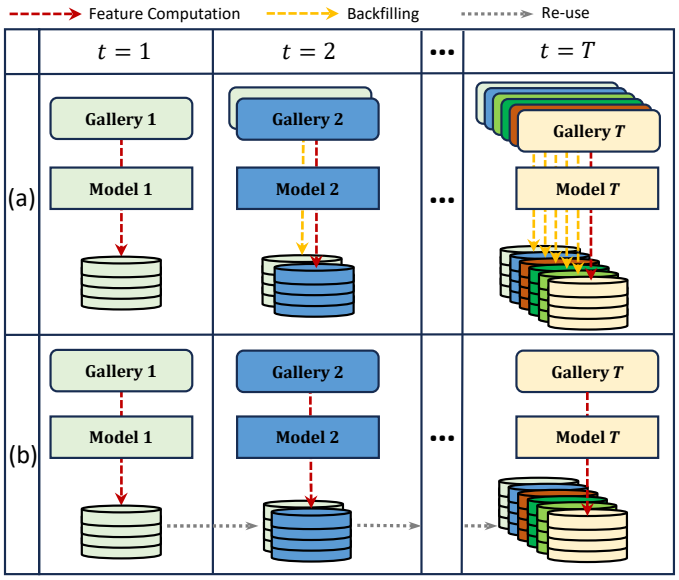


Fig. 1. Gallery feature computation at the inference phase. (a) The conventional LReID. (b) The proposed LReID with the backward-compatibility.

On the other hand, the backward-compatible training (BCT) has been investigated to address the time-consuming task of backfilling [8]–[10]. The BCT methods train the model by forcing the features computed by the new model to be close to the features obtained by the previously trained old models. Hence, based on BCT, the gallery features extracted by the old model can be re-used for re-identification with the new query features obtained by the new model, alleviating the burden of backfilling. However, the existing BCT methods jointly train all the accumulated datasets to update the model and do not consider the domain shift between the training datasets.

Based on the investigation of the complementary relationship between LReID and BCT, we devise a novel framework for LReID which ensures the backward-compatibility based on the replay dataset. Fig 1 illustrates the difference between (a) the conventional LReID and (b) the proposed LReID with the backward-compatibility in terms of the gallery feature

computation at the inference phase. Whereas the conventional LReID conducts the backfilling process repeatedly whenever the model is updated with new datasets, the proposed framework considering the backward-compatibility uses the stored old gallery features, and hence is free from the time-consuming backfilling process. To ensure that the model has compatibility with the old models, we utilize the replay datasets and the corresponding features to design a cross-model compatibility loss inspired by the supervised contrastive loss. Experimental results show that the proposed method yields much better performance of LReID compared with the existing methods while achieving the backward-compatibility for inference.

Our main contributions are summarized as follows:

- We introduced a novel framework of LReID based on the replay dataset that avoids the time-consuming backfilling process and ensures the model's backward-compatibility at the inference stage.
- We proposed the cross-model compatibility loss by utilizing the replay images and the corresponding features.
- Experimental results showed that the proposed method provides better performance of LReID compared with the existing methods in terms of the backward-compatibility.

II. Related Works

A. Person Re-Identification

Most of the existing ReID methods have been developed in the supervised learning manner, that attempt to obtain ID-discriminative representation dealing with small inter-class variations and large intra-class variations. Sun et al. [11] horizontally divided the person image to extract part-aware features. Zeng et al. [12] learned the illumination-invariant features by disentangling the illumination and identity features. Luo et al. [13] enhanced the benefits of using the cross-entropy loss and the triplet loss. He et al. [14] introduced a transformer-based network leveraging the token-wise knowledge. Zhang et al. [1] employed a transformer-based network to augment the high-frequency patches and obtain fine-grained features. Ren et al. [15] utilized the attention mechanism to deal with the occlusion problem.

Recently, LReID has been studied to overcome the retraining burden of the supervised methods assuming more practical scenarios where the training datasets in different domains are continuously collected and provided. Pu et al. [4] adaptively accumulated and tranferred the knowledge by using the learnable knowledge graph. Wu and Gong [5] mitigated the catastrophic forgetting by ensuring knowledge coherence between the old and new models in terms of the classification, feature distribution and representation. Huang et al. [16] proposed an unsupervised domain-adaptation method for LReID that uses a meta-learning scheme to strike a balance between the new and old knowledge. Ge et al. [3] considered the LReID as a domain-adaptation problem where the old datasets are regarded as the source domain and the new dataset as the target domain, respectively. Yu et al. [7] made the old model trainable to refresh the old knowledge, which helps the model to consolidate the knowledge between the old and new models. Xu et al. [6] developed a replay-free LReID method by harmonically leveraging the long and short-term knowledge. However, the existing LReID methods consider the computational burden and memory consumption at the training stage only, while overlooking the burden of iterative feature computation on the accumulated gallery images at the inference stage.

B. Backward-Compatible Training

BCT was introduced to alleviate the backfilling burden of retrieval and search systems, such as the face recognition, whenever the model is updated with newly collected training datasets. The purpose of BCT is to train the new model to be compatible with the old models in terms of the feature extraction. Shen et al. [9] first introduced the concept of BCT where the features obtained by the new model are forced to be compatible with the frozen old classifier. Meng et al [17] proposed the transformation module which maps the features of one model into the other model's feature spaces by using the alignment and boundary losses. Wu et al. [10] enhanced the backward-compatibility by using the contrastive loss with respect to the features and logits of the new and old models. Pan et al. [8] performed the adversarial learning which makes the features, obtained by using different models, indistinguishable. Recently, Wan et al. [18] maintained the backward consistency in the retrieval networks based on the continual learning. However, it does not consider the changing datasets with huge distribution-shift across different domains, and the performance is only evaluated on the classes seen in the training which lacks the applicability to practical scenarios.

III. Proposed Method

Fig. 2 illustrates the training and inference procedures of the proposed LReID framework with backward-compatibility. Following the conventional LReID formulation with the replay scheme, we assume that T different training datasets are given in order as $\mathcal{D}_i\,(i = 1, 2, \cdots, T)$, and the model ϕ is sequentially trained by using the training datasets, respectively. Specifically, we first train the model ϕ_1 by using $\mathcal{D}_1$, and sample the replay data $\mathcal{R}_1$ from $\mathcal{D}_1$. We compute the features of $\mathcal{R}_1$ by using ϕ_1 and store them into $\boldsymbol{\mathcal{F}}_{\mathrm{R},1}$, which is then used to train the next models. At the T-th time instance, we train the model ϕ_T by using $\mathcal{D}_T$ and the previously sampled replay data $\{\mathcal{R}_i\}_{i=1}^{T-1}$. The pre-computed replay features $\{\boldsymbol{\mathcal{F}}_{\mathrm{R},i}\}_{i=1}^{T-1}$ are also used for backward compatibility training. Then we compute the features of the newly sampled replay data $\mathcal{R}_T$ by using the new model ϕ_T and store them into $\boldsymbol{\mathcal{F}}_{\mathrm{R},T}$.

At the inference stage, we test the model on the galleries, $\mathcal{G}_i\,(i = 1, 2, \cdots, T)$ with respect to the query set $\mathcal{Q}$. In the conventional LReID scenario, the features of all the accumulated galleries are computed again for inference whenever the model is newly updated. On the contrary, we store the features of the gallery $\mathcal{G}_i$ computed by using ϕ_i to $\boldsymbol{\mathcal{F}}_{\mathrm{G},i}$ which is used for inference at the next time instances. Consequently, after the training of the last (T-th) model, we use the new model ϕ_T

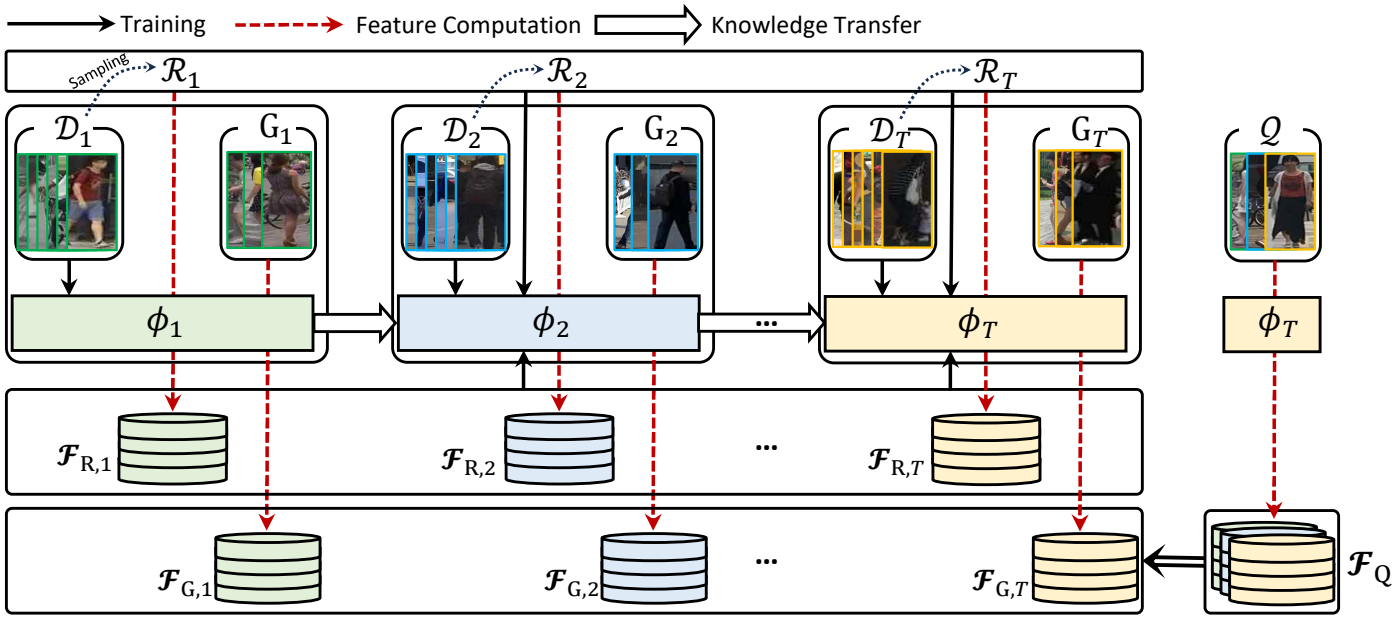


Fig. 2. Overview of the training and inference procedures in the proposed framework. We represent the feature extractors only for simplicity.

to compute the features of the gallery $\mathcal{G}_T$ only, and evaluate the performance of LReID on $\boldsymbol{\mathcal{F}}_{\mathrm{G},T}$ and the set of all the pre-computed gallery features $\{\boldsymbol{\mathcal{F}}_{\mathrm{G},i}\}_{i=1}^{T-1}$ together. Since we do not need to re-compute the features of previous galleries, it is a more practical setup without requiring the time-consuming backfilling process.

Whenever the model is sequentially trained on new datasets in the conventional LReID framework, the resulting feature spaces are changed to exhibit unique distributions according to the training domains. In such cases, the last model ϕ_T becomes incompatible with the previous models $\{\phi_1, \phi_2, ..., \phi_{T-1}\}$, and we should extract the features of all the galleries again by using ϕ_T. To ensure the compatibility of the new model with the previously trained models, we design the cross-model compatibility loss $\mathcal{L}_{\mathrm{cmc}}$ inspired by the concept of the supervised contrastive loss [19], given by

$$\mathcal{L}_{\mathrm{cmc}} = -\frac{1}{|B|}\sum_{k\in B}\log\left(\frac{\sum_{i=1}^{T-1}\sum_{j=1}^{|\boldsymbol{\mathcal{F}}_{\mathrm{R},i}^{k+}|}\exp\left\{\frac{\phi_T(\boldsymbol{r}_k)\cdot\boldsymbol{f}_{i,j}^{k+}}{\tau}\right\}}{\sum_{i=1}^{T-1}\sum_{j=1}^{|\boldsymbol{\mathcal{F}}_{\mathrm{R},i}|}\exp\left\{\frac{\phi_T(\boldsymbol{r}_k)\cdot\boldsymbol{f}_{i,j}}{\tau}\right\}+\mathcal{N}}\right), \tag{1}$$

where $\mathcal{N}$ is formulated as

$$\mathcal{N} = \sum_{j\in B_T}\exp\left\{\frac{\phi_T(\boldsymbol{r}_k)\cdot\phi_T(\boldsymbol{x}_j)}{\tau}\right\}. \tag{2}$$

B and B_T denote the mini-batches sampled from all the accumulated replay images and from the T-th dataset, respectively. $\boldsymbol{r}_k$ is the k-th image in B, and $\boldsymbol{x}_j$ is the j-th image in B_T. $\boldsymbol{\mathcal{F}}_{\mathrm{R},i}^{k+}$ is the subset of the replay features in $\boldsymbol{\mathcal{F}}_{\mathrm{R},i}$ positive to the identity of $\boldsymbol{r}_k$. $\boldsymbol{f}_{i,j}$ and $\boldsymbol{f}_{i,j}^{k+}$ denote the j-th replay features in $\boldsymbol{\mathcal{F}}_{\mathrm{R},i}$ and $\boldsymbol{\mathcal{F}}_{\mathrm{R},i}^{k+}$, respectively. τ is a temperature parameter. Note that the conventional supervised contrastive loss [19] takes the positive and negative samples within a mini-batch only without considering the model's backward-compatibility. On the contrary, we exploit the replay features across all the old datasets with different domains via $\{\boldsymbol{\mathcal{F}}_{\mathrm{R},i}\}_{i=1}^{T-1}$ based on the lifelong learning framework, and hence preserve the backward-compatibility more reliably by using the cross-model compatibility loss $\mathcal{L}_{\mathrm{cmc}}$ in (1).

Furthermore, the negative samples with respect to the each identity can help the model to improve the discriminative capability via the contrastive learning [21]. Based on the assumption that different datasets do not include the same identity simultaneously, we regard the images in the T-th dataset are the negative samples to the replay images, and minimize the similarity between the features of the old replay image $\boldsymbol{r}_k$ and the features of the new training images via the additional negative term in (2).

There has been an attempt [10] to achieve the backward-compatibility by using the supervised contrastive loss, which uses $\mathcal{R}$ and the old model ϕ_{T-1} instead of $\boldsymbol{\mathcal{F}}_{\mathrm{R}}$ where $\phi_{T-1}(\boldsymbol{r}_i)$ is utilized to guide the new model to represent the feature space obtained by ϕ_{T-1}. However, in the proposed lifelong learning formulation, the training datasets are dynamically changing across long duration of time instances, and thus it is hard for ϕ_{T-1} to keep the knowledge of the feature space at very old time instances suffering from the backward-compatibility of ϕ_T towards the old models. It is well-known that the LReID performance on the old datasets gets worse as T increases with more and more incoming datasets due to the catastrophic forgetting [22]. Note that the features of $\boldsymbol{\mathcal{F}}_{\mathrm{R}}$, extracted by the old models, preserve the representation of the old domains and thus guide the new model ϕ_T to ensure the backward-compatibility.

TABLE I
QUANITATIVE PERFORMANCE COMPARISON. WE RE-IMPLEMENTED THE EXISTING METHODS TO EVALUATE THE PERFORMANCE OF THE BACKWARD-COMPATIBILITY.

Train Order: MA → DU → CU → MS										
Method	MA		DU		CU		MS		Average	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
Joint-train	81.1	93.4	71.1	86.0	91.8	93.2	41.1	69.8	71.3	85.6
Fine-tuning	59.3	80.3	54.2	74.7	82.3	84.7	38.6	67.6	58.6	76.8
LwF [20]	62.9	82.0	56.8	74.8	78.8	81.2	13.5	33.7	53.0	67.9
BCT [9]	75.1	89.7	60.2	78.0	84.6	86.3	29.8	56.6	62.4	77.7
CVS [18]	70.7	87.0	56.3	72.0	81.5	83.8	27.2	53.0	58.9	74.0
PTKP [3]	73.4	86.4	**62.7**	77.8	83.8	85.3	36.3	62.0	64.1	77.9
KRKC [7]	59.5	80.3	49.3	69.7	80.7	82.4	36.7	65.3	56.6	74.4
LSTKC [6]	48.0	61.9	49.3	63.9	77.5	78.3	**42.7**	**68.7**	54.4	68.2
Proposed	**78.4**	**91.3**	62.5	**78.0**	**86.8**	**88.6**	33.1	60.4	**65.2**	**79.5**

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

We used four benchmark datasets with unique domains for training: Market1501[23] (MA), DukeMTMC[24] (DU), CUHK-SYSU[25] (CU), and MSMT17[26] (MS), which are widely used in the existing literatures of LReID [3], [5], [27]. We pre-process the scene images in CUHK-SYSU for the purpose of ReID by using the ground-truth annotations following [5]. To evaluate the performance, we adopt the mean average precision (mAP) score and the Rank-1 (R-1) accuracy.

To train the network, we construct our baseline with $\mathcal{L}_{\text{base}}$ that is formulated as

$$\mathcal{L}_{\text{base}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{tri}}\mathcal{L}_{\text{tri}}, \tag{3}$$

where $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{tri}}$ are the cross-entropy loss and the hard triplet loss [28] calculated by using the feature extractor ϕ, respectively. λ_{ce} and λ_{tri} are hyper-parameters. We adopted the ImageNet pre-trained ResNet50 [29] as the feature extractor. Our total loss function to train ϕ_t is formulated as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{cmc}}\mathcal{L}_{\text{cmc}}, \tag{4}$$

where λ_{cmc} is a hyper-parameter. Hyper-parameters are empirically set as follows: $\lambda_{\text{ce}} = 1$ and $\lambda_{\text{tri}} = 1$ in (3), $\tau = 0.1$ in (1) and (2), and $\lambda_{\text{cmc}} = 0.05$ in (4). We followed [3] for other training details such as the replay memory update strategy, learning rate, etc.

B. Performance Comparison

To compare the performance of LReID with backward-compatibility, we re-implemented several existing methods: LwF [20], BCT [9], CVS [18], PTKP [3], KRKC [7], and LSTKC [6]. Table I shows the results of the proposed method compared with that of the existing methods. Joint-train means that we trained the model on all the datasets together, and Fine-tuning means that the model is independently trained on each new dataset in order. Both methods are trained with $\mathcal{L}_{\text{base}}$ only. The evaluation protocol is as follows. Each set of gallery features in $\{\boldsymbol{\mathcal{F}}_{\text{G,t}}\}_{t=1}^{T-1}$ is pre-computed by using the corresponding model of each task $\{\phi_t\}_{t=1}^{T-1}$, respectively, and we compute all the query features $\{\boldsymbol{\mathcal{F}}_{\text{Q,t}}\}_{t=1}^{T}$ and the features of the T-th gallery $\boldsymbol{\mathcal{F}}_{\text{G},T}$ by using ϕ_T after training on the last dataset.

We observe that the performance of almost methods on MSMT17, the last dataset, degrades compared to that of Fine-tuning due to the stability-plasticity dilemma [30] that the bias on the old model's knowledge disrupts the adaptation capability of the new knowledge. For instance, LwF suffers from the biasing toward the old knowledge and yields poor performance on MSMT17. On the one hand, PTKP achieves noticeable performance although it does not explicitly consider the backward-compatibility in training. Since PTKP casts the LReID problem into the domain adaption problem regarding the old dataset as source dataset, the model maps the features from the new dataset into the feature space of the old model that is trained with the source dataset. Thus, it could enhance backward-compatibility of the new model.

On the other hand, KRKC also addresses the LReID problem but it trains the old model as well unlike the other methods that freeze the parameters of the old model while preserving its feature space. Hence it exhibits poor backward-compatibility while adapting the new knowledge with the impressive performance on MSMT17 compared to the other methods. As the old model is trained, the feature space of old model is gradually forgotten, and the old model becomes no longer suitable for guiding the new model to ensure the backward-compatibility in KRKC. It is noteworthy that the existing LReID methods, e.g., PTKP, can enhance the backward-compatibility of the model, however it does not always guarantee the backward-compatibility as in the case of KRKC.

V. CONCLUSION

In this paper, we proposed a replay-based framework of LReID that ensures the backward-compatibility. Whereas the conventional LReID methods re-compute the features of the old gallery images whenever new datasets are given, the proposed method alleviates such time-consuming backfilling process by training the new model to be compatible toward the old domains. To this end, we designed the cross-model compatibility loss that utilizes the features of the replay images to guide the new model to represent the old feature space. Experimental results showed that the proposed method

improves the performance of backward-compatibility in LReID by a large margin compared to the existing LReID methods.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00392536), and by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2020-II201336, Artificial Intelligence Graduate School Program(UNIST)).

REFERENCES

[1] G. Zhang, Y. Zhang, T. Zhang, B. Li, and S. Pu, "Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 14 133–14 142.

[2] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019.

[3] W. Ge, J. Du, A. Wu, *et al.*, "Lifelong person re-identification by pseudo task knowledge preservation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 688–696.

[4] N. Pu, W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew, "Lifelong person re-identification via adaptive knowledge accumulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 7901–7910.

[5] G. Wu and S. Gong, "Generalising without forgetting for lifelong person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2889–2897.

[6] K. Xu, X. Zou, and J. Zhou, "Lstkc: Long short-term knowledge consolidation for lifelong person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 16 202–16 210.

[7] C. Yu, Y. Shi, Z. Liu, S. Gao, and J. Wang, "Lifelong person re-identification via knowledge refreshing and consolidation," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 3295–3303.

[8] T. Pan, F. Xu, X. Yang, *et al.*, "Boundary-aware backward-compatible representation via adversarial learning in image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 15 201–15 210.

[9] Y. Shen, Y. Xiong, W. Xia, and S. Soatto, "Towards backward-compatible representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 6368–6377.

[10] S. Wu, L. Chen, Y. Lou, *et al.*, "Neighborhood consensus contrastive learning for backward-compatible representation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2722–2730.

[11] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.

[12] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3064–3074, 2020.

[13] H. Luo, W. Jiang, Y. Gu, *et al.*, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.

[14] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 15 013–15 022.

[15] X. Ren, D. Zhang, X. Bao, and Y. Zhang, "S^2-net: Semantic and saliency attention network for person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 4387–4399, 2023.

[16] Z. Huang, Z. Zhang, C. Lan, *et al.*, "Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 14 288–14 297.

[17] Q. Meng, C. Zhang, X. Xu, and F. Zhou, "Learning compatible embeddings," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9939–9948.

[18] T. S. Wan, J.-C. Chen, T.-Y. Wu, and C.-S. Chen, "Continual learning for visual search with backward consistent feature embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 16 702–16 711.

[19] P. Khosla, P. Teterwak, C. Wang, *et al.*, in *Proc. Adv. Neural Inform. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020, pp. 18 661–18 673.

[20] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, 2017.

[21] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[22] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci.*, vol. 114, no. 13, pp. 3521–3526, 2017.

[23] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[24] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 17–35.

[25] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3415–3424.

[26] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 79–88.

[27] J. Huang, X. Yu, D. An, *et al.*, "Learning consistent region features for lifelong person re-identification," *Pattern Recog.*, vol. 144, p. 109 837, 2023.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 815–823.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[30] M. Mermillod, A. Bugaiska, and P. Bonin, "The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects," *Frontiers Psychol.*, vol. 4, p. 504, 2013.

Enhancing Semiconductor X-RAY Images: A Framework Combining Denoising and Super-Resolution Modules With a Novel Dataset

Jae Hoon Shim*‡, Min Woo Kim*‡, Tae Gyu Lim†, Byung Seok Min†, Sang Hwa Lee*§, Nam Ik Cho*
* Department of ECE, INMC, Seoul National University, Seoul, Republic of Korea
† Central R&D Center, XAVIS Co., Ltd., Seongnam, Kyeonggi-do, Republic of Korea
‡ indicates equal contribution
§ corresponding author

***Abstract*—Semiconductors are essential components of modern technology, and inspecting their defects has become increasingly important. One popular approach is to inspect them using X-ray images, but the release of semiconductor images to the public is restricted due to security concerns. This limited dataset poses a challenge, particularly when developing learning-based inspection methods. Another problem with X-ray images is the noise and low resolution, which lowers the inspection accuracy of machine and human inspectors. To overcome this challenge, we introduce a new public semiconductor dataset and propose a new X-ray image enhancement framework to help inspectors increase inspection accuracy. Precisely, we have collected a new dataset consisting of three types of X-ray images obtained from actual industrial settings. Additionally, our enhancement method can improve all types of images without the need to train a separate model for each type. This is achieved by analyzing the noise within the dataset and building on previous studies, and using objective functions defined in the frequency domain. Various experiments and ablation studies have shown that our method can significantly improve X-ray semiconductor images, both qualitatively and quantitatively. Our proposed semiconductor dataset is publicly available at this URL: https://drive.google.com/drive/folders/1_YSuxR_YGNah1mnPGNd-pgrRULNZP5gn?usp=sharing.**

***Index Terms*—Image Enhancement, Semiconductor, X-ray Imaging, Image Denoising, Image Super Resolution**

I. Introduction

Semiconductors are important components used in various products and manufacturing machines today. They are also critical parts of machines where safety is a top priority, such as unmanned vehicles. Therefore, accurately detecting defects in semiconductors is becoming increasingly important, and improving the speed and accuracy of defect inspection is a critical objective in the semiconductor industry.

One popular semiconductor inspection method is to identify defects through X-ray imaging [1]–[4]. In general, the precision of semiconductor inspection is largely dependent on the quality of X-ray images, where two key factors that affect the image quality are the *energy level* and *exposure time* [5]. Although higher energy levels and longer exposure times can enhance the image quality, they can also have negative effects.

Precisely, it is required to have a delicate balance between image clarity and potential damage to the semiconductor circuits and other structures. While higher energy levels result in clearer images, excessive energy can harm the semiconductors. Therefore, it is crucial to limit the energy level to avoid damage. In addition, the exposure time should be long enough to produce clean images with minimal noise from the equipment. However, a long exposure time can also reduce semiconductor yields. Hence, to obtain high-quality images without harming the semiconductors, short exposure time with limited X-ray energy is necessary. But this can negatively impact the accuracy of both human and machine inspectors.

Therefore, in this paper, we present a new semiconductor image dataset and a deep-learning approach to improve the quality of X-ray images captured with various energy levels and exposure times. Since most semiconductor images are intended for industrial purposes, their public release is strictly prohibited. This restriction has resulted in a scarcity of datasets containing semiconductor X-ray images, which in turn has hindered the advancement of image processing technologies related to semiconductors. As demonstrated in Figure 1, our dataset includes three types of semiconductor images captured under various equipment and conditions, and we expect that the dataset can help the academic community overcome the research limitation.

Furthermore, based on prior research [6] and our observation of commonalities in the noise distribution across them, we suggest an integrated framework to enhance input images, instead of training separate networks for each type. Our framework consists of two modules: a denoising module and a super-resolution module. Initially, to reduce noise, we employ downsampling in the spatial domain using an encoder. Subsequently, due to the importance of preserving essential details like circuit and bump boundaries on semiconductor wafers, we adopt the architecture proposed in Soh *et al.* [7]. In addition to the spatial domain loss, we incorporate frequency-domain functions, Focal Frequency Loss [8] and Wavelet-domain High-Frequency Loss [9], which are shown to keep high-frequency components. As a result, our method shows significant improvements in both qualitative and quantitative ways.

Our contributions can be summarized as follows:

- We have proposed a new semiconductor X-ray image

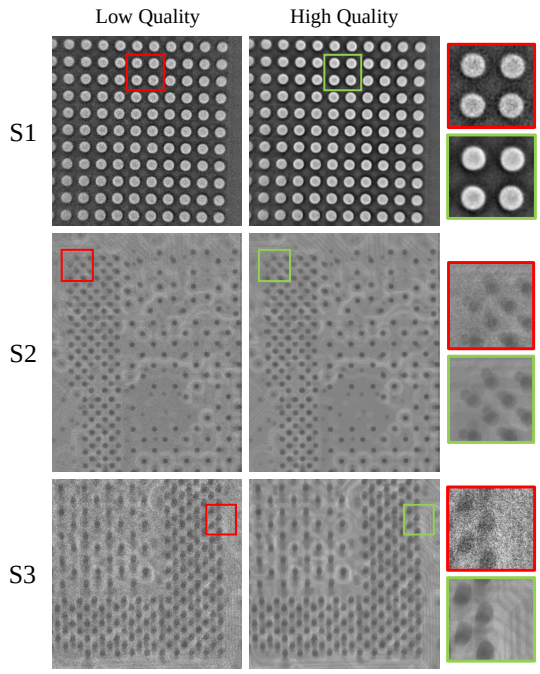


Fig. 1. Overview of our proposed dataset. The dataset consists of three types of data, each having a different resolution ($S_1 \in \mathbb{R}^{1024\times1024}, S_2 \in \mathbb{R}^{1936\times1536}, S_3 \in \mathbb{R}^{1900\times1900}$). Low-Quality images have significant noise due to short-duration X-ray exposure, while High-Quality ones have less noise as they are projected over a comparatively longer duration.

dataset that comprises three types of images captured under different equipment and conditions. The aim of this dataset is to facilitate research in the academic community, particularly learning-based X-ray image processing research.

• We have demonstrated that there are common statistical characteristics among the types of data. Based on the analysis, we designed an integrated framework capable of improving all three types at once.

• Through experiments and ablation studies, we have validated that our approach makes a significant improvement in both qualitative and quantitative results.

II. Method

A. Novel Semiconductor Dataset

We propose a novel semiconductor X-ray image dataset. As depicted in Figure 1, to provide diversities, our dataset includes three distinct types of images, depending on the filming equipment and environment. These are denoted as S1, S2, and S3. Each type comprises pairs of Low Quality (LQ) and High Quality (HQ) images. LQ images are captured by subjecting semiconductor wafers to a brief X-ray exposure (5 seconds), resulting in images with significant noise. Meanwhile, HQ images are produced by a longer X-ray exposure (90 seconds), yielding images with considerably less noise. There are 2135 images in S1, 2000 in S2 and 1040 in S3 dataset, and we have partitioned each data type into train and test datasets. For S1 and S3, the proportion between training and test datasets is 80:20, while it is 90:10 for S2.

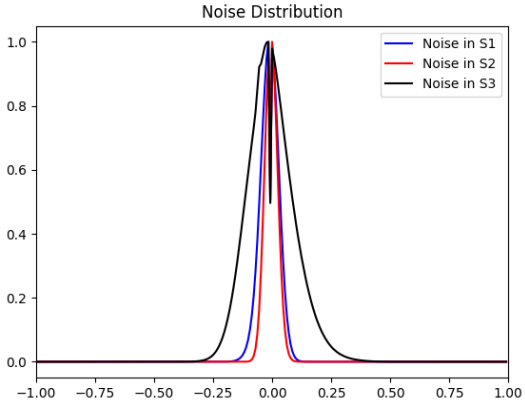


Fig. 2. To investigate the noise characteristics added to Low Quality (LQ) images compared to High Quality (HQ) ones, we plot histograms of the pixel values obtained by subtracting LQ from HQ, using 100 training pairs from each data type. Across all three data types, the shapes of histograms are similar to Gaussian.

B. Theoretical Background for the Framework

1) Noise Analysis in Dataset: In order to examine the nature of noise in LQ images relative to HQ ones, we compute the differences between corresponding pairs and analyze the pixel histogram for each type of data. As shown in Figure 2, we find that the additional noise present in LQ images across all types (S1, S2, and S3) can be approximated by a Gaussian distribution and exhibits a mathematically similar form. There have been many studies [6], [10], [11] for handling Gaussian noise of various standard deviations using a single model. This research provides the theoretical foundation for our integrated model.

2) Noise2Noise: Noise2Noise [6] suggested that it is possible to train the network to produce clean data solely with the use of noisy image pairs. They demonstrated that if the noise expectation is zero, the model can learn to approximate the clean ground truth despite each gradient aligning with the noisy data. The optimization task can be formulated as:

$$argmin_\theta \sum_i L(f_\theta(\hat{X}_i), \hat{Y}_i) \; where \; \hat{Y}_i = Y_i + n_i. \quad (1)$$

In this context, θ represents the trainable parameters of the model, and the pair $(\hat{X}_i, \hat{Y}_i)$ denotes a set of noisy training images. Note that Y_i and n_i indicate the clean data and noise, respectively. Under the assumption that the noise has a zero mean, such as Gaussian noise, the network's parameters are optimized to generate an unobserved clean target as:

$$E[\hat{Y}_i|\hat{X}_i] = Y_i \quad (2)$$

especially when the $L2$ loss is utilized. Furthermore, employing the $L1$ loss eliminates outliers, which similarly guides the network's training process, resulting in less blur in the output images.

Our goal is to convert data with noise caused by brief X-ray exposure times into clean data. However, obtaining clean

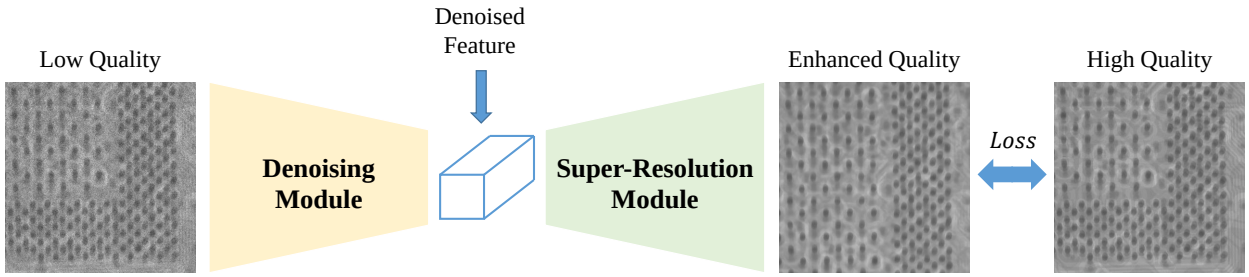


Fig. 3. Overall structure of the proposed framework. The LQ image, targeted for enhancement, initially goes through the Denoising Module. This process results in the denoised feature where unnecessary information like noise is eliminated. However, it retains sufficient information for the Super-Resolution Module in the feature ($z \gg c$ in Equation 3). Then, the feature undergoes the Super-Resolution Module to produce an Enhanced-Quality image. Utilizing Enhanced Quality and HQ images, the final objective function (Equation 4) is calculated to train our framework.

semiconductor X-ray images theoretically needs infinite X-ray exposure time, which is physically unfeasible. Thus, inspired by Noise2Noise research, we intend to project input data onto the manifold of clean data using pairs of noisy images, involving more noisy images (Low Quality, LQ) and less noisy ones (High Quality, HQ).

C. Framework

The primary objective of our framework is to enhance LQ images to achieve clarity comparable to HQ ones while eliminating noise. Hence, our framework consists of two key modules: the denoising module and the super-resolution module, as illustrated in Figure 3.

Traditional techniques for noise reduction involve low-pass filters and bicubic downsampling. However, with the advancement of deep learning, it has been proven that an end-to-end trained network outperforms these traditional methods in denoising task [12]–[16]. Especially, the downsampling process within such networks is critical for removing high-frequency bands that contain noise components [17]. Consequently, we incorporate an encoder network with downsampling in the denoising module to effectively filter out high-frequency noise from LQ images.

To ensure semiconductor images perform optimally in post-processing stages, such as anomaly detection, it is important not only to eliminate noises but also to improve image clarity. However, in the spatially compressed feature obtained through downsampling via an encoder, there is a lack of high-frequency information, which is related to the image detail. Hence, when enlarging the spatial size of the embedding feature to match the original image, we utilize a network designed for the Super-Resolution (SR) task. Within our framework, we adopt the fractal residual learning (FRL) architecture proposed in NatSR [7]. Although NatSR uses a discriminator to determine if the input patch lies on the natural manifold, we only utilize the SR network for processing semiconductor X-ray images because the discriminator's role is not proper in our task.

The above series of processes can be written as the following equation:

$$\hat{y} = D(f) = D(E(x)), \tag{3}$$

where $x \in \mathbb{R}^{h \times w \times c}$ and $\hat{y} \in \mathbb{R}^{h \times w \times c}$ indicate LQ and Enhanced Quality (EQ) image. Additionally, $f \in \mathbb{R}^{h/4 \times w/4 \times z}$ represents the denoised feature, which is the most spatially compressed element within our framework. E and D denote the denoising and the SR module, respectively. Note that in contrast to Singh et al. [18], we assign a larger value to z in f than to c in x. This assignment allows us to incorporate more information into f, which is essential for the SR module to effectively restore image details.

D. Objective Function

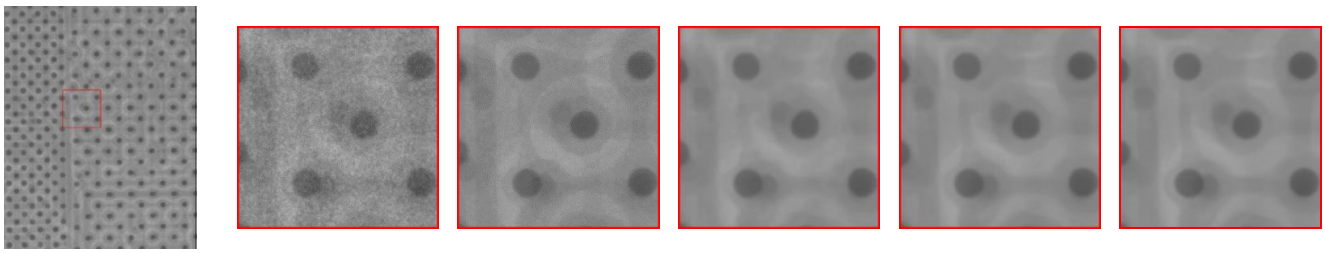
To train the network in the spatial domain, we utilize a distortion-oriented loss function. We refrain from adopting losses associated with the discriminator, as our goal is to produce the unobserved clean data manifold. Furthermore, since this field does not involve human visual perception and aesthetic appreciation, the use of any perceptual loss is deemed less relevant.

To improve the clarity of semiconductor lines and bump boundaries in images, we incorporate two losses defined in the frequency domain. First, we utilize Focal Frequency Loss (FFL) [8], suggested to narrow the frequency gap between EQ and HQ images. This method converts each image to the frequency domain via the Discrete Fourier Transform (DFT) and then calculates the Euclidean distance and associated weight at each spectral position. Nevertheless, these weights are proportional to the scale of the frequency band, assigning the low-frequency band with larger weights. To emphasize the high-frequency band, we also employ Wavelet-domain High-Frequency Loss (WHFL) [9]. This approach ensures that weights increase with frequency, effectively highlighting higher frequencies.

As a result, our final objective function for training the network can be written as:

$$L(y, \hat{y}) = |y - \hat{y}|_1^1 + \lambda_1 \cdot L_{FFL}(y, \hat{y}) + \lambda_2 \cdot L_{WHFL}(y, \hat{y}), \tag{4}$$

where y and $\hat{y}$ denote HQ and EQ images, respectively. Also, λ_1, λ_2 are hyper-parameters balancing among each loss term.



(a) HQ whole image (b) LQ (c) HQ (d) $DS_{1\times}$ (e) $DS_{2\times}$ (f) $DS_{4\times}$

Fig. 4. Comparison of results among original Low Quality (LQ), High Quality (HQ) images and super-resolved images according to downsampling scale ($DS_{n\times}$). $DS_{1\times}$ means that the input is not downsampled at all. $DS_{2\times}$ and $DS_{4\times}$ indicate that the input goes through $2\times$ and $4\times$ downsampling, respectively.

TABLE I

EXPERIMENTAL RESULTS ON DIFFERENT SETTINGS FOR THREE DATA TYPES: S1, S2, AND S3. THE EXPERIMENTS COVER VARIOUS SETTINGS, INCLUDING "BICUBIC" (APPLYING BICUBIC DOWNSAMPLING TO THE INPUT), "DM" (DENOISING MODULE), "FFL", AND "WHFL". PERFORMANCE RANKINGS ARE INDICATED WITH RED DENOTING THE HIGHEST, AND BLUE INDICATING THE SECOND-HIGHEST PERFORMANCE FOR EACH DATA TYPE. NUMBERS IN PARENTHESES REPRESENT THE PERFORMANCE INCREASE COMPARED TO LQ IMAGES.

Setting	DM	FFL	WHFL	Data type	PSNR	Average PSNR	SSIM	Average SSIM
bicubic	✗	✗	✗	S1 S2 S3	34.91(+3.22) 33.76(+5.63) 33.94(+13.62)	34.39(+6.39)	0.9264(+0.2218) 0.8603(+0.1895) 0.8824(+0.7158)	0.8996(+0.3373)
bicubic + FFL	✗	✓	✗	S1 S2 S3	37.20(+5.52) 33.79(+5.67) 34.04(+13.71)	35.60(+7.60)	0.9296(+0.2250) 0.8600(+0.1893) 0.8819(+0.7154)	0.9010(+0.3387)
DM + FFL	✓	✓	✗	S1 S2 S3	35.29(+3.60) 33.97(+5.84) 34.62(+14.29)	34.80(+6.80)	0.9310(+0.2263) 0.8659(+0.1952) 0.8838(+0.7172)	0.9036(+0.3413)
DM + FFL + WHFL	✓	✓	✓	S1 S2 S3	35.34(+3.65) 33.91(+5.78) 34.68(+14.36)	34.83(+6.83)	0.9324(+0.2278) 0.8662(+0.1955) 0.8851(+0.7186)	0.9048(+0.3424)

III. EXPERIMENT

A. Experimental Settings

We conduct experiments to compare four different settings across three data types: S1, S2, and S3, using PSNR and SSIM as evaluation metrics. In the "bicubic" setting, we apply bicubic downsampling to the input image and perform super-resolution to obtain the resulting image, employing the L1 loss as the loss function. In the "bicubic + FFL" setting, the network architecture remains the same, but we adopt FFL in addition to the L1 loss. In the "DM + FFL" setting, we replace bicubic downsampling with the denoising module. Lastly, in the "DM + FFL + WHFL" setting, we combine both FFL and WHFL with the L1 loss.

B. Experimental Results

Table I presents the PSNR and SSIM results for the four settings. Our proposed framework, "DM + FFL + WHFL," exhibits the best performance across all data types in terms of SSIM. It achieves the second-highest PSNR results in S1 and S2 data types and the highest PSNR in S3. While "bicubic + FFL" demonstrates the highest average PSNR value, this is primarily attributed to the high PSNR value in S1. However, a detailed analysis reveals that each component we introduced (denoising module, FFL, and WHFL) effectively enhances both PSNR and SSIM. Note that we set the SR module as the learning-based method [7] because the module is an essential component for upsampling the downsized intermediate results back to the original image size, excluding the component from this analysis.

Firstly, comparing "bicubic" and "bicubic + FFL," we observe improvements across all aspects, especially a significant increase in PSNR for S1. This suggests that FFL can compensate for the L1 loss for improving the quality of result images.

Secondly, "DM + FFL" enhances most metrics across nearly all data types when compared to "bicubic + FFL." Notably, "DM + FFL" employs a deeper feature, rather than a downsampled image, as an intermediate feature. This approach more effectively preserves the information necessary for restoring image details. Although increasing the channel dimension of the feature may adversely affect the performance in PSNR of S1, it improves overall performance including SSIM in S1. Hence, replacing "bicubic" with "DM" enhances the results.

As mentioned, the "DM + FFL + WHFL" setting, which incorporates WHFL into "DM + FFL", demonstrates superior performance. While maintaining PSNR values in S2, WHFL enhances all other metrics across every data type in comparison

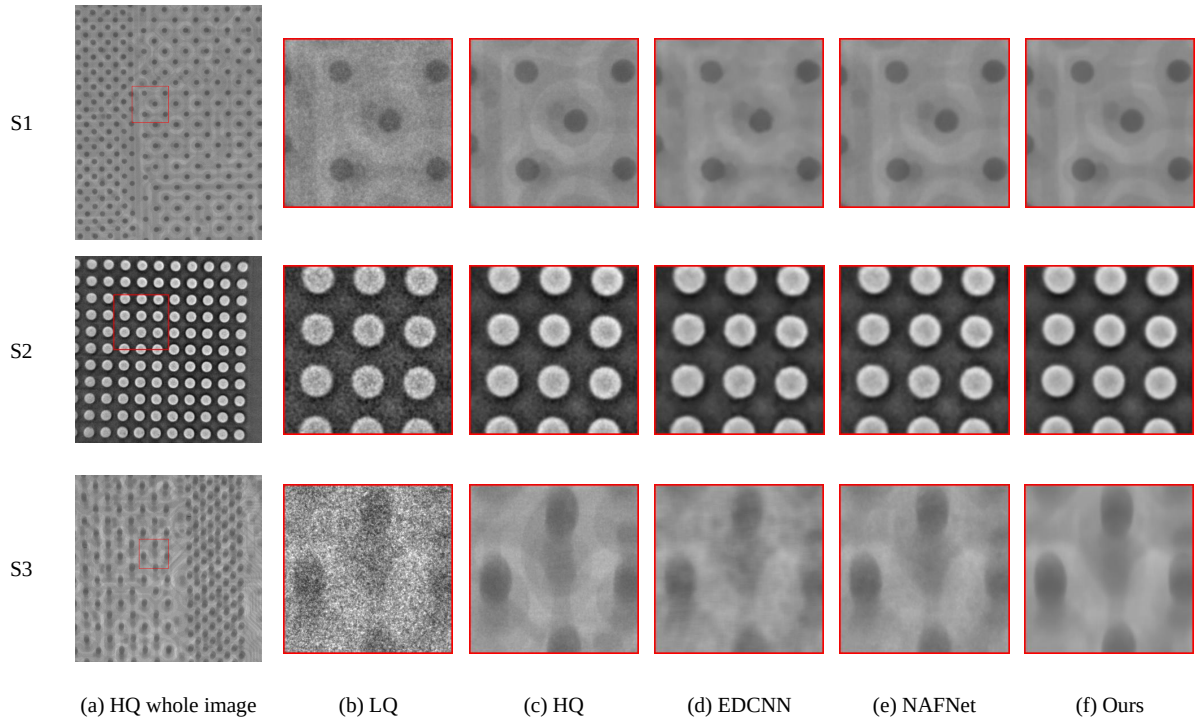


Fig. 5. Comparison among the original LQ, HQ images, and results from existing methods and ours for every data type (S1, S2, S3). Each row corresponds to each data type, and each column indicates an image category. Columns (d) and (e) show the results from EDCNN [19] and NAFNet [20], respectively. Column (f) displays the results of our full framework.

to "DM + FFL". As a result, given that SSIM is based on structural information, we can conclude that both FFL and WHFL significantly contribute to enhancing the quality of X-ray images for semiconductors.

C. Effects of Downsampling

We adopt downsampling as a preprocessing step for input images to eliminate noise and other unnecessary elements before proceeding to the SR module. Therefore, to investigate the effect of downsampling, we conduct an ablation study. In this section, we employ bicubic downsampling to the Low-Quality (LQ) input and subsequently super-resolve it using our deep network.

Figure 4 illustrates the results of applying $1\times$ (indicating no downsampling), $2\times$, and $4\times$ downsampling to the S1-type image. The image downsampled at $4\times$ shows better line preservation and more efficient noise removal compared to the $1\times$ and $2\times$ downsampled images, leading to clearer patterns. Despite the noise in the HQ image, downsampling contributes to obtaining an image cleaner than the HQ image.

Based on the above observation, we conclude that the downsampling process effectively eliminates unnecessary information, such as noise and stain-like artifacts, while preserving essential features like lines and curves. Thus, we design our framework by combining a denoising module for the downsampling process and the SR model to improve image quality, akin to an autoencoder.

D. Comparison to Existing Denoising Methods

We conduct a comparative analysis of our method against established denoising techniques. NAFNet [20] demonstrates exceptional performance in real-world image denoising, while EDCNN [19] works effectively in the context of CT image denoising, particularly in the medical image domain. We have re-trained both NAFNet and EDCNN on our dataset.

As illustrated in the qualitative results of Figure 5, the images denoised with EDCNN exhibit both blurriness and noise. Meanwhile, NAFNet achieves clearer results than EDCNN, but the residual noise artifacts still remain. Notably, more pronounced noise patterns are seen in the results of S2 and S3 data types. In contrast, our method produces clean denoised images with significantly reduced noise artifacts compared to other denoising methods. With our framework, images can be structurally reinforced, enhancing defect inspection capabilities.

IV. Conclusion

We have introduced a new dataset and a deep learning framework to improve the quality of semiconductor X-ray images. The dataset includes X-ray images taken at various energy levels and exposure times. In our deep learning framework, the input image is first passed through a denoising network that downsamples the image while preserving important information by expanding the feature's dimensionality. The feature is then processed by a super-resolution network. To

train the network, we combined pixel and frequency-domain functions to form a loss function. Our proposed architecture and these functions work together to enhance the quality of the output image, outperforming one of the leading deep networks in the field of real image denoising. We believe that our dataset and framework will contribute to the advancement of X-ray image processing for semiconductors and other areas.

ACKNOWLEDGMENT

This work was supported by the Technology Innovation Program (ATC+ program, 20014131, 25nm X-ray Inspection System for Semiconductor Backend Process) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea), in part by BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023, and partially by Samsung Electronics Co., Ltd.

REFERENCES

[1] Q. Zhang, M. Zhang, C. Gamanayake, *et al.*, "Deep learning based defect detection for solder joints on industrial x-ray circuit board images," in *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)*, IEEE, vol. 1, 2020, pp. 74–79.

[2] J. O'Leary, K. Sawlani, and A. Mesbah, "Deep learning for classification of the chemical composition of particle defects on semiconductor wafers," *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 1, pp. 72–85, 2020.

[3] K. Kim, H. Kim, J. Chun, M. Kang, M. Hong, and B. Min, "Real-time anomaly detection in packaged food x-ray images using supervised learning," *Comput. Mater. Contin*, vol. 67, no. 2, pp. 2547–2568, 2021.

[4] J. Ding, C. Ye, H. Wang, J. Huyan, M. Yang, and W. Li, "Foreign bodies detector based on detr for high-resolution x-ray images of textiles," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2023.

[5] I. O. Romero, Y. Fang, and C. Li, "Correlation between x-ray tube current exposure time and x-ray photon number in gate," *Journal of X-ray science and technology*, vol. 30, no. 4, pp. 667–675, 2022.

[6] J. Lehtinen, J. Munkberg, J. Hasselgren, *et al.*, "Noise2noise: Learning image restoration without clean data," *arXiv preprint arXiv:1803.04189*, 2018.

[7] J. W. Soh, G. Y. Park, J. Jo, and N. I. Cho, "Natural and realistic single image super-resolution with explicit natural manifold discrimination," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8122–8131.

[8] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 919–13 929.

[9] M. W. Kim and N. I. Cho, "Whfl: Wavelet-domain high frequency loss for sketch-to-image translation," in *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, 2023, pp. 744–754.

[10] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[11] M. El Helou, R. Zhou, and S. Süsstrunk, "Stochastic frequency masking to improve super-resolution and denoising networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, Springer, 2020, pp. 749–766.

[12] Z. Tu, H. Talebi, H. Zhang, *et al.*, "Maxim: Multi-axis mlp for image processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5769–5780.

[13] M. Chang, Q. Li, H. Feng, and Z. Xu, "Spatial-adaptive network for single image denoising," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, Springer, 2020, pp. 171–187.

[14] S. Lefkimmiatis, "Universal denoising networks: A novel cnn architecture for image denoising," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3204–3213.

[15] Z. Hong, X. Fan, T. Jiang, and J. Feng, "End-to-end unpaired image denoising with conditional adversarial networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34(04), 2020, pp. 4140–4149.

[16] C. Ren, X. He, C. Wang, and Z. Zhao, "Adaptive consistency prior based deep network for image denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8596–8606.

[17] A. H. Ribeiro and T. B. Schön, "How convolutional neural networks deal with aliasing," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 2755–2759.

[18] P. Singh, M. Diwakar, R. Gupta, *et al.*, "A method noise-based convolutional neural network technique for ct image denoising," *Electronics*, vol. 11, no. 21, p. 3535, 2022.

[19] T. Liang, Y. Jin, Y. Li, and T. Wang, "Edcnn: Edge enhancement-based densely connected network with compound loss for low-dose ct denoising," in *2020 15th IEEE International Conference on Signal Processing (ICSP)*, IEEE, Dec. 2020. DOI: 10.1109/icsp48669.2020.9320928. [Online]. Available: http://dx.doi.org/10.1109/ICSP48669.2020.9320928.

[20] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European Conference on Computer Vision*, Springer, 2022, pp. 17–33.

Monocular Depth Estimation for Autonomous Driving Based on Instance Clustering Guidance

Dahyun Kim[1], Dongkwon Jin[2], and Chang-Su Kim[1]
Korea University[1], Samsung Advanced Institute of Technology[2]
dhkim@mcl.korea.ac.kr, dongkwon.jin@samsung.com, changsukim@korea.ac.kr

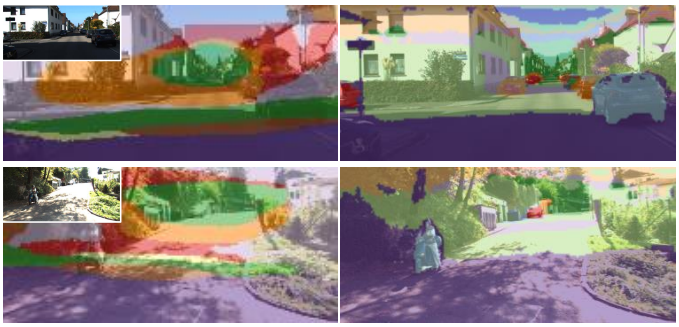
***Abstract*—A novel monocular depth estimator for autonomous driving, which produces reliable instance depths via instance clustering guidance, is proposed in this work. First, we extract multi-scale feature maps from a road scene and initialize depth clusters. Second, we update the depth clusters using the feature maps through transposed cross-attention. To guide the update process, we develop the instance clustering membership (ICM) loss, which employs an instance segmentation map. Third, we transfer the updated depth clusters to the feature map at the finest resolution, from which we produce the final depth map. Extensive experimental results show that the proposed algorithm yields competitive results to state-of-the-art techniques on the KITTI, Cityscapes-DVPS, and SemKITTI-DVPS datasets.**

I. Introduction

Monocular depth estimation (MDE) aims to predict the depth value for each pixel from a single RGB image. Since MDE can offer a 3D geometrical understanding, it is useful in various applications, such as 3D CAD model generation [1], augmented reality [2], and autonomous driving [3], [4]. In particular, in autonomous driving systems, it is crucial to precisely estimate the depths of moving obstacles, such as pedestrians and cars, while it is less critical to deal with background, such as roads and buildings.

Various MDE techniques have been proposed for road environments. Eigen *et al.* [5] introduced a deep-learning-based MDE method, and its variants [6], [7] have been designed using convolutional neural networks. Also, some attempts [8], [9] have been made to increase the depth quality by exploiting semantic segmentation data. In [8], a constraint was imposed on a depth map to maintain consistency between depth boundaries and segmentation results. In [9], a loss function was used to encourage depth discontinuity around semantic boundaries. In [2], an edge detector was employed to reduce errors near depth boundaries. Meanwhile, multi-task networks [10]–[13] have been developed to predict segmentation and depth results jointly. However, all these techniques do not focus on the depth accuracies of object instances, which are critically important for safe autonomous driving.

Recently, several MDE techniques [14]–[16] have been developed based on vision transformer [17]. Especially, Piccinelli *et al.* [16] proposed the iDisc algorithm to estimate depth maps via transformer-based clustering. They initialized depth clusters and then updated them via transposed cross-attention [18], [19] with contextual features. Then, they exploited the updated depth clusters to produce depth maps reliably. However, as in Fig. 1(a), iDisc yields clusters composed of semantically uncorrelated pixels, making the clustering results difficult to interpret. Also, iDisc does not consider the depth quality of individual instances.



(a) iDisc (b) Proposed

Fig. 1. It is crucial to precisely estimate the depths of moving obstacles in road environments. We aim to improve the depth estimation accuracies of instances through instance clustering guidance. Whereas uncorrelated pixels are clustered in iDisc [16] in (a), the clustering results roughly match individual instances in the proposed algorithm in (b).

In this work, we propose a novel monocular depth estimator for autonomous driving, which focuses on improving instance depth accuracies through instance clustering guidance, as illustrated in Fig. 1(b). First, we extract multi-scale feature maps from an image and initialize depth clusters. Then, we update the depth clusters through transposed cross-attention. To guide the update process, we develop a novel loss, called the instance clustering membership (ICM) loss, using ground-truth (GT) instance segmentation masks. Lastly, we transfer the updated depth clusters to the feature map at the finest resolution, from which we produce the final depth map. It is demonstrated through extensive experiments that the proposed algorithm yields better results than existing techniques on the Cityscapes-DVPS [10], SemKITTI-DVPS [10], and KITTI [20] datasets.

This paper has the following main contributions.

- The proposed algorithm estimates instance depths reliably via the instance clustering approach, which improves depth accuracies by exploiting semantic segmentation results during its training.
- We develop the ICM loss, which helps to align depth clusters to object instances.
- The proposed algorithm provides competitive results to existing state-of-the-art methods on the Cityscapes-DVPS, SemKITTI-DVPS, and KITTI datasets.

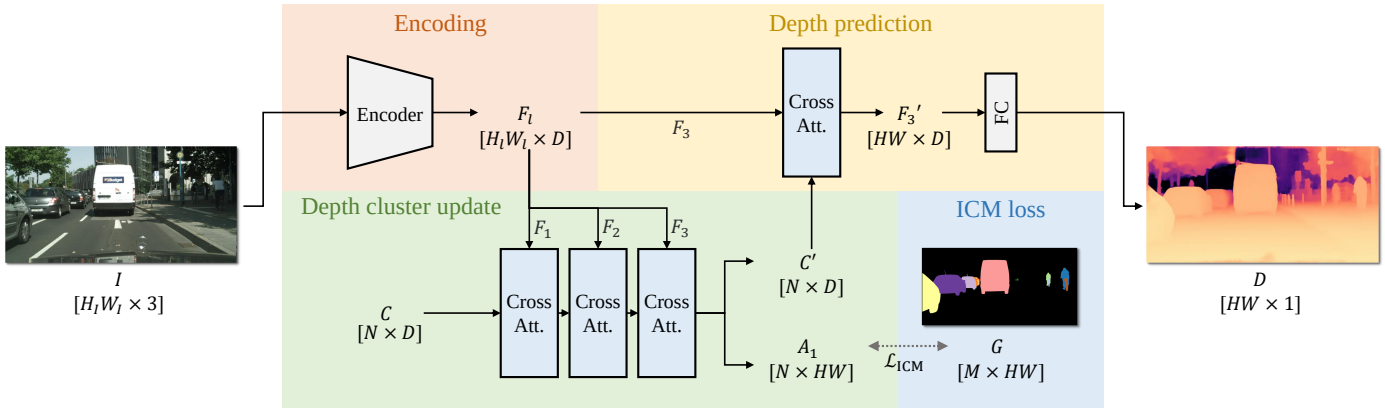


Fig. 2. The network architecture of the proposed algorithm. Given an input image I, the proposed network estimates a depth map D. First, multi-scale feature maps $\{F_l\}_{l=1}^3$ are extracted from I. Then, the depth cluster matrix C is updated to C' using the extracted feature maps through cross-attention. The update process is guided by the ICM loss. Lastly, the finest feature map F_3 is updated to F_3' using the depth cluster matrix C' and then projected to obtain the final depth map D.

II. Proposed Algorithm

The proposed algorithm has four components: encoding, depth cluster update, ICM loss, and depth prediction, as in Fig. 2. Note that we perform the depth cluster update and depth prediction steps differently from [16]. Moreover, the proposed ICM loss assists in the cluster update process.

A. Encoding

Given an image $I \in \mathbb{R}^{H_I W_I \times D}$, multi-scale feature maps $\{F_l\}_{l=1}^3$ are extracted using the Swin-L encoder [21], where $F_1 \in \mathbb{R}^{\frac{H}{4}\cdot\frac{W}{4}\times D}$, $F_2 \in \mathbb{R}^{\frac{H}{2}\cdot\frac{W}{2}\times D}$ and $F_3 \in \mathbb{R}^{HW\times D}$. These feature maps are used in both depth cluster update and depth prediction steps. We set $H = \frac{H_I}{4}$, $W = \frac{W_I}{4}$, and $D = 256$.

B. Depth Cluster Update

Let $C \in \mathbb{R}^{N\times D}$ be a learnable matrix representing N depth clusters, each with dimension D, which is randomly initialized. Then, using the multi-scale feature maps $\{F_l\}_{l=1}^3$, the depth cluster matrix C is updated to C' through transposed cross-attention [16], [18], [19] by

$$A_1 = \underset{N}{\text{softmax}}(C_q F_k^T), \quad C' = A_1 F_v + C. \tag{1}$$

Here, $C_q \in \mathbb{R}^{N\times D}$, $F_k \in \mathbb{R}^{H_lW_l\times D}$, and $F_v \in \mathbb{R}^{H_lW_l\times D}$ are linearly projected query, key, and value features from C and F_l. Also, $A_1 \in \mathbb{R}^{N\times H_lW_l}$ is the attention matrix, and the underscript N means that the softmax operation is done in the column direction, *i.e.*, over the N columns, as in [16], [19]. Thus, each element a_{ij} in A_1 indicates the probability that jth pixel belongs to cluster i. Using the matrix A_1, we transfer the contextual information of each pixel to the corresponding depth cluster.

C. Depth Prediction

With the updated depth cluster matrix C', the finest feature map F_3 is updated to F_3' through cross-attention by

$$A_2 = \underset{N}{\text{softmax}}(F_q C_k^T), \quad F_3' = A_2 C_v + F_3, \tag{2}$$

where $F_q \in \mathbb{R}^{HW\times D}$, $C_k \in \mathbb{R}^{N\times D}$, and $C_v \in \mathbb{R}^{N\times D}$ are linearly projected query, key, and value features from F_3 and C'. Also, $A_2 \in \mathbb{R}^{HW\times N}$ is the attention matrix, which is used to deliver the information of each depth cluster to the relevant pixels in F_3'. Note that, contrary to (1), the softmax operation is applied in the row direction, *i.e.*, over the N rows.

Finally, the output depth map is obtained by

$$D = f(F_3') \tag{3}$$

where f is a fully-connected layer.

D. ICM loss

Although depth features are produced through the information exchange between pixels and clusters in (1) and (2), those processes, in themselves, do not process object instances and background regions separately. Hence, without an extra mechanism, each depth cluster may comprise semantically uncorrelated pixels as in Fig. 1(a). In this work, we attempt to distinguish instance pixels from background pixels to enhance the depth features of foreground objects, which is crucial for autonomous driving systems to operate safely. To this end, we propose the ICM loss, employing GT instance segmentation masks, which is used to guide the attention matrix A_1 in (1).

First, we define the GT membership matrix $G \in \mathbb{R}^{M\times HW}$, where M is the number of instances in an image. Specifically, as illustrated in Fig. 3, we compute the centroid c_i of instance i as its median depth value using the GT depth map and instance segmentation mask. From the centroids $\{c_1, \ldots, c_M\}$, we then construct the GT membership matrix G based on the Fuzzy C-means (FCM) [22] as follows.

$$g_{ij} = \frac{|x_j - c_i|^{-\frac{2}{m-1}}}{\sum_{k=1}^{M} |x_j - c_k|^{-\frac{2}{m-1}}}, \quad g_{ij} \in [0, 1]. \tag{4}$$

Here, x_j is the depth value of pixel j, and m is the degree of fuzziness. The value of g_{ij} ranges from 0 to 1. When pixel j

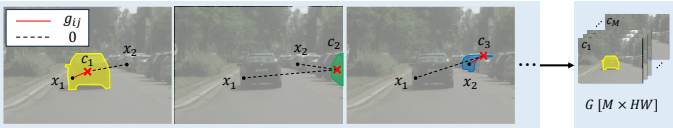


Fig. 3. Illustration of the GT membership computation. We define the GT membership matrix $G \in \mathbb{R}^{M \times HW}$, where M is the number of instances in an image. Given the GT depth map and instance segmentation mask, we compute the centroids $\{c_1, \ldots, c_M\}$ of M instances as their median depth values. If pixel j with a depth value x_j belongs to instance i, it is assigned a membership value g_{ij} via (4); otherwise, $g_{ij} = 0$.

is more associated with instance i, g_{ij} gets closer to 1. Also, we set $g_{ij} = 0$ for pixel j that does not belong to instance i.

Next, we compute the matching costs between A_1 and G by

$$\ell_{ii'} = -\hat{g}_{i'}^T \log \hat{a}_i - (1 - \hat{g}_{i'}^T) \log(1 - \hat{a}_i) \tag{5}$$

where $\hat{a}_i^T = a_i^T / \|a_i^T\|$ and $\hat{g}_{i'}^T = g_{i'}^T / \|g_{i'}^T\|$ are the normalized row vectors of A_1 and G, respectively. These vectors are normalized because the matching costs are computed by the cross-entropy of corresponding elements. After applying the Hungarian algorithm to the cost matrix, we define the ICM loss $\mathcal{L}_{\text{ICM}}$ as

$$\mathcal{L}_{\text{ICM}} = \mathcal{L}(A_1, G) = \frac{1}{N} \sum_{i=1}^{N} \ell_{i\sigma(i)} \tag{6}$$

where a_i and $g_{\sigma(i)}$ denote the pair with the overall minimum matching cost. By employing the ICM loss $\mathcal{L}_{\text{ICM}}$, each cluster aligns roughly to an instance, as shown in Fig. 4(c). Thus, we make the depth features of each instance more discriminative.

III. Experiments

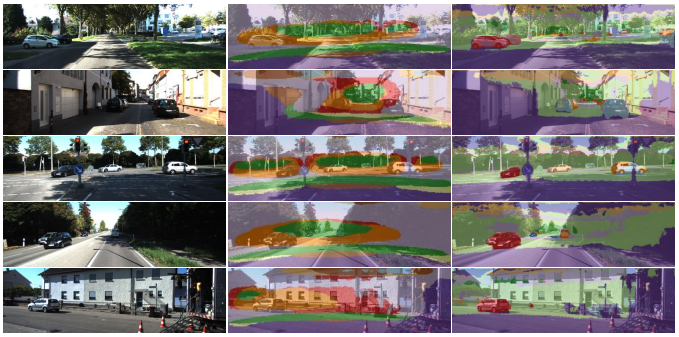
A. Datasets

We compare the proposed algorithm with existing techniques on the Cityscapes-DVPS [10], SemKITTI-DVPS [10], and KITTI [20] datasets.

Cityscapes-DVPS: This dataset extends the Cityscapes-VPS dataset [24] by incorporating depth labels obtained using stereo images. It consists of 3,000 frames, divided into 2,400 frames for training, 300 for validation, and 300 for testing. Both images and depth maps have a resolution of 1024×2048. Furthermore, it provides panoptic segmentation masks, including instance segmentation ones, which are utilized to compute the GT membership matrix G in (4).

SemKITTI-DVPS: It is based on KITTI [20]. Sparse annotations are derived from projecting panoptic-labeled 3D point clouds in SemanticKITTI [25] onto the image plane. It consists of 19,130 training, 4,071 validation, and 4,342 testing frames. Similar to Cityscapes-DVPS, it also provides instance segmentation labels.

KITTI: It is a well-known depth dataset containing stereo images and corresponding depth labels, with an average resolution of 1241×376. It consists of 23,158 frames for training and 652 frames for testing, as specified in [5]. In contrast to the other datasets, KITTI does not provide instance segmentation labels. Therefore, we generate pseudo segmentation labels using an off-the-shelf instance segmentation network [26].



(a) (b) (c)

Fig. 4. Visualization of membership matrices A_1 during the inference stage for input images in (a). When the ICM loss is not applied in (b), the clustering results are not related to object instances. In contrast, with the ICM loss in (c), clusters align roughly with individual instances, making the depth features of each instance more discriminative.

B. Implementation Details

We employ Swin-L [21] as the encoder backbone. The number of depth clusters N is set to 32, and the degree of fuzziness m in (4) is set to 2. During training, we combine a general depth loss and the ICM losses at three resolutions $l = 1, 2, 3$ by

$$\mathcal{L} = \mathcal{L}_{\text{SI}_{\log}} + \mathcal{L}_{\text{ICM1}} + \mathcal{L}_{\text{ICM2}} + \mathcal{L}_{\text{ICM3}} \tag{7}$$

where $\text{SI}_{\log}$ denotes the depth loss in [5]. Note that the GT membership matrix G in (4) is downsampled for each resolution l to a smaller size than HW.

C. Comparative Assessment

Comparison on Cityscapes-DVPS: Table I compares the proposed algorithm with conventional algorithms [11]–[13], [16], [23] on the Cityscapes-DVPS dataset. The proposed algorithm outperforms all conventional algorithms in every metric. Especially, the proposed algorithm is better than DeepDPS [13], which is the state-of-the-art technique, by margins of about 0.003, 0.004, and 0.003 in A. Rel, $\text{RMSE}_{\log}$, and δ_1, respectively. Moreover, the proposed algorithm outperforms iDisc [16], which also adopts a transformer-based clustering process. This indicates that the proposed ICM loss guides the clustering process effectively to improve the depth estimation performances.

Fig. 5 presents some estimation results on Cityscapes-DVPS. The proposed algorithm predicts the depths of instances more precisely than PolyphonicFormer [12] and iDisc. Especially, in the first row, PolyphonicFormer and iDisc fail to estimate the depths of the bus reliably, whereas the proposed algorithm provides highly accurate results.

Comparison on SemKITTI-DVPS: Table II shows the comparison results with PolyphonicFormer and iDisc on SemKITTI-DVPS. Again, the proposed method outperforms

TABLE I
COMPARISON ON CITYSCAPES-DVPS.

	A. Rel ↓	$\text{RMSE}_{\log}$ ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
DPT-Hybrid [23]	0.0697	0.1106	0.9434	0.9914	0.9976
PanopticDepth [11]	0.0711	0.1125	0.9359	0.9919	0.9982
PolyphonicFormer [12]	0.0647	0.1013	0.9524	0.9950	0.9985
iDisc [16]	0.0689	0.1090	0.9462	0.9902	0.9975
DeepDPS [13]	0.0597	0.0940	0.9616	0.9953	0.9988
Proposed	**0.0570**	**0.0896**	**0.9650**	**0.9961**	**0.9991**

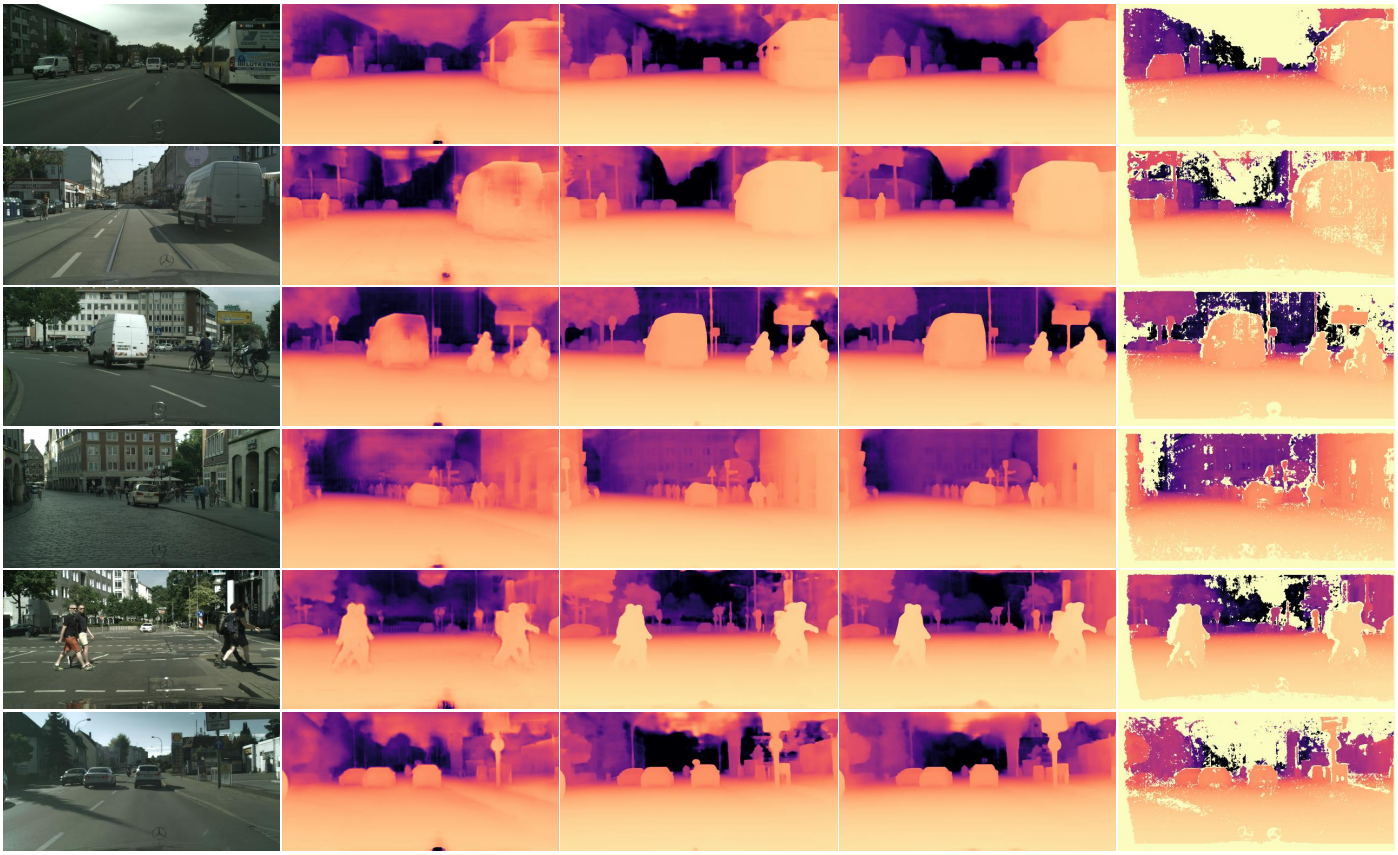


Fig. 5. Comparison of depth estimation results on Cityscapes-DVPS.

TABLE II
COMPARISON ON SEMKITTI-DVPS.

	A. Rel ↓	$\text{RMSE}_{\log}$ ↓	δ_1 ↑
PolyphonicFomer [12]	0.0900	0.1422	0.9084
iDisc [16]	0.0754	0.1248	0.9290
Proposed	**0.0697**	**0.1134**	**0.9398**

TABLE III
COMPARISON ON KITTI.

	A. Rel ↓	$\text{RMSE}_{\log}$ ↓	δ_1 ↑
BTS [7]	0.0563	0.090	0.964
AdaBins [14]	0.0585	0.088	0.964
NewCRF [15]	0.0520	0.079	0.974
iDisc [16]	0.0509	**0.077**	0.975
ZoeDepth	0.0576	0.089	0.965
Proposed	**0.0507**	0.079	**0.976**

the existing techniques in every metric. Particularly, the proposed algorithm yields better performance than iDisc with a significant margin of about 0.006, 0.011, and 0.011 in A. Rel, $\text{RMSE}_{\log}$, and δ_1, respectively. Fig. 6 shows estimation results. While iDisc yields inaccurate results around instance boundaries, the proposed algorithm yields clearer depth boundaries.

Comparison on KITTI: In Table III, we compare the proposed algorithm with the conventional MDE algorithms [7], [14]–[16] on KITTI. Note that we use pseudo segmentation masks since no GT segmentation masks are available in KITTI. Nonetheless, the proposed algorithm yields decent performances, ranking second in $\text{RMSE}_{\log}$ and achieving the best results in A. Rel. and δ_1. In Fig. 7, the proposed algorithm estimates instance depths more reliably than iDisc does.

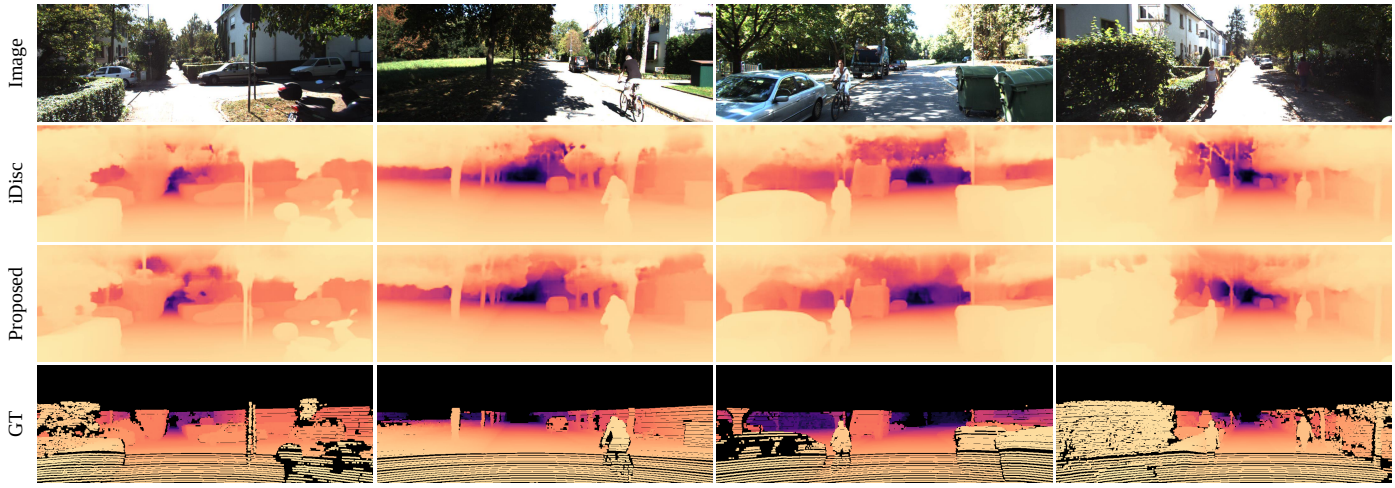


Fig. 6. Comparison of depth estimation results on SemKITTI-DVPS.

TABLE IV
COMPARISON OF δ_1 SCORES FOR INSTANCES AND BACKGROUND AREAS.

	Cityscapes-DVPS		SemKITTI-DVPS		KITTI	
	δ_1 (instance) ↑	δ_1 (background) ↑	δ_1 (instance) ↑	δ_1 (background) ↑	δ_1 (instance) ↑	δ_1 (background) ↑
iDisc [16]	0.9673	0.9374	0.9582	0.9245	0.9671	**0.9773**
Proposed	**0.9762**	**0.9541**	**0.9605**	**0.9253**	**0.9676**	0.9761

TABLE V
ABLATION STUDIES OF THE PROPOSED ALGORITHM ON CITYSCAPES-DVPS.

	Sequential update	ICM loss	A. Rel ↓	δ_1 ↑
I			0.0689	0.9462
II	✓		0.0625	0.9522
III	✓	✓	**0.0619**	**0.9531**

D. Analysis

Performance in instance and background areas: Table IV compares the proposed algorithm with iDisc in terms of δ_1 scores for instances and background areas separately. The proposed algorithm achieves higher δ_1 scores for instance areas in all three datasets. Especially, it provides a higher score for instances on KITTI, despite using pseudo segmentation masks. This indicates that the proposed algorithm produces more reliable depth results for instances by exploiting the proposed ICM loss.

Efficacy of key components: We conduct ablation studies to analyze the efficacy of the proposed algorithm and its components. Table V compares some ablated methods on Cityscapes-DVPS. Unlike the setting in Table I, we adopt ResNet50 [27] as the encoder backbone for faster training and comparison. Method I generates three depth cluster matrices C_l, C_2 and C_3 separately for each feature map. Then, three different depth maps are estimated from the cluster matrices, and the final depth map is obtained by averaging all these depth maps. Thus, method I is similar to iDisc [16]. In method II, a single depth cluster matrix is sequentially updated using multi-scale feature maps, as described in Section II-B. However, the proposed ICM loss is not employed for training. Compared to method III (the proposed algorithm), methods I and II underperform in terms of A. Rel and δ_1. This indicates that it is beneficial to update depth clusters sequentially and guide them with the proposed ICM loss.

IV. CONCLUSIONS

We proposed a monocular depth estimator for autonomous driving, which produces reliable instance depths via instance clustering guidance. The proposed algorithm consists of four main components: encoding, depth cluster update, ICM loss, and depth prediction. The ICM loss assists in the depth cluster update, aligning clusters to object instances. Experimental results showed that the proposed algorithms achieves competitive results to existing techniques. Moreover, the proposed algorithm yields more reliable depth estimation results on instances and offers meaningful clustering results.

ACKNOWLEDGEMENTS

This work was supported by the NRF grants funded by the Korea government (MSIT) (No. RS-2024-00397293 and No. NRF-2022R1A2B5B03002310).

REFERENCES

[1] H. Izadinia, Q. Shan, and S. M. Seitz, "IM2CAD," in *Proc. IEEE CVPR*, 2017.
[2] L. Talker, A. Cohen, E. Yosef, A. Dana, and M. Dinerstein, "Mind the edge: Refining depth edges in sparsely-supervised monocular depth estimation," in *Proc. IEEE CVPR*, 2024.
[3] X. Yang, Z. Ma, Z. Ji, and Z. Ren, "GEDepth: Ground embedding for monocular depth estimation," in *Proc. IEEE ICCV*, 2023.
[4] J. Moon, J. L. G. Bello, B. Kwon, and M. Kim, "From-Ground-To-Objects: Coarse-to-fine self-supervised monocular depth estimation of dynamic objects with ground contact prior," in *Proc. IEEE CVPR*, 2024.

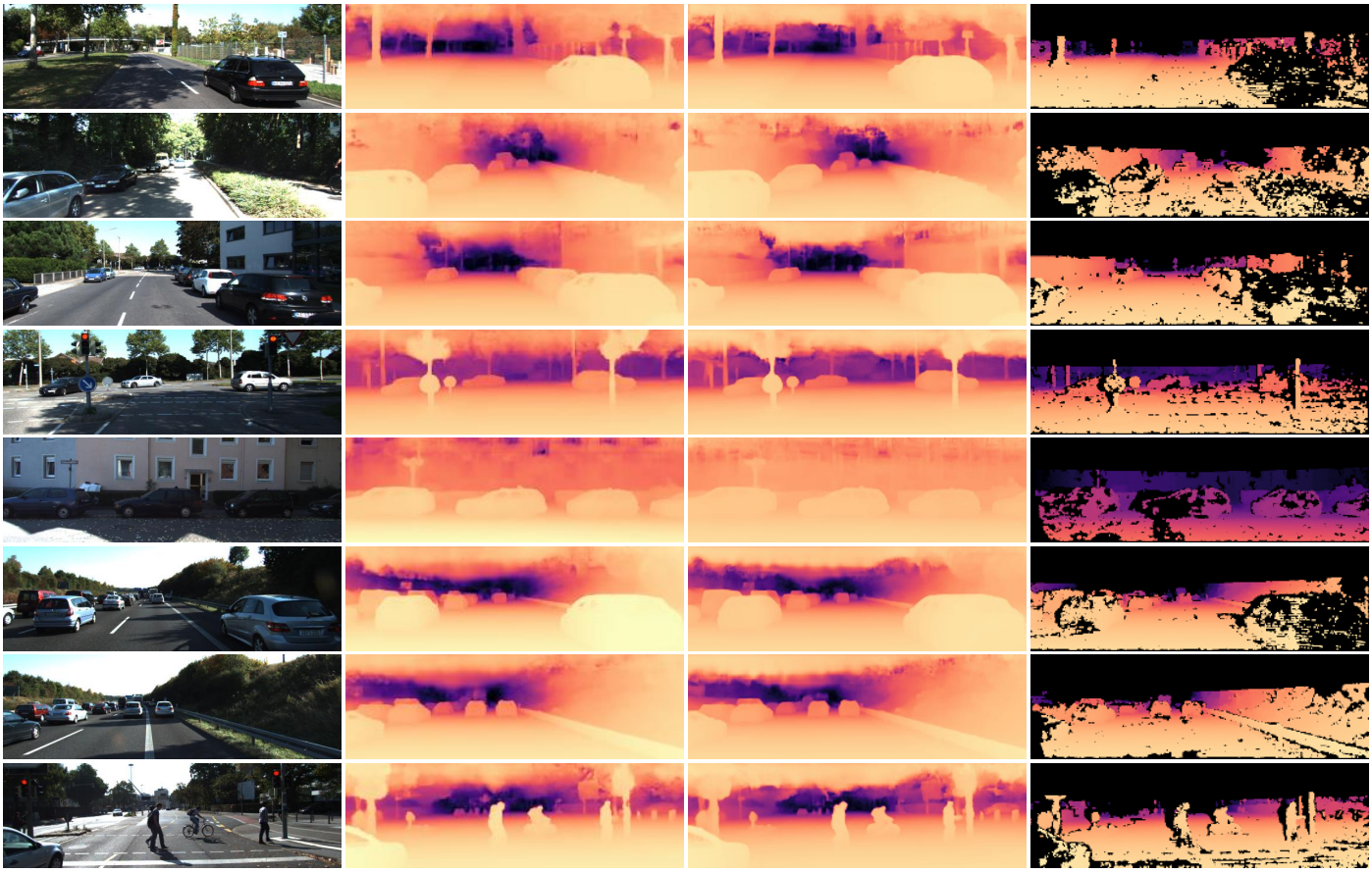


Fig. 7. Comparison of depth estimation results on KITTI.

[5] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. NIPS*, 2014.

[6] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE CVPR*, 2018.

[7] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," in *arXiv:1907.10326*, 2019.

[8] S. Zhu, G. Brazil, and X. Liu, "The edge of depth: Explicit constraints between segmentation and depth," in *Proc. IEEE CVPR*, 2020.

[9] F. Saeedan and S. Roth, "Boosting monocular depth with panoptic segmentation maps," in *Proc. IEEE WACV*, 2021.

[10] S. Qiao, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "ViP-DeepLab: Learning visual perception with depth-aware video panoptic sementation," in *Proc. IEEE CVPR*, 2021.

[11] N. Gao, F. He, J. Jia, Y. Shan, H. Zhang, X. Zhao, and K. Hunag, "PanopticDepth: A unified framework for depth-aware panoptic segmentation," in *Proc. IEEE CVPR*, 2022.

[12] H. Yuan, X. Li, Y. Yang, G. Cheng, J. Zhang, Y. Tong, L. Zhang, and D. Tao, "PolyphonicFormer: Unified query learning for depth-aware video panoptic segmentation," in *Proc. ECCV*, 2022.

[13] J. He, Y. Wang, L. Wang, H. Lu, BinLuo, J.-Y. He, J.-P. Lan, Y. Geng, and X. Xie, "Towards deeply unified depth-aware panoptic segmentation with bi-directional guidance learning," in *Proc. IEEE ICCV*, 2023.

[14] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE CVPR*, 2021.

[15] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, "Neural window fully-connected CRFs for monocular depth estimation," in *Proc. IEEE CVPR*, 2022.

[16] L. Piccinelli, C. Sakaridis, and F. Yu, "iDisc: Internal discretization for monocular depth estimation," in *Proc. IEEE CVPR*, 2023.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020.

[19] Q. Yu, H. Wang, S. Qiao, M. Collins, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "k-means mask transformer," in *Proc. ECCV*, 2022.

[20] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmar suite," in *Proc. IEEE CVPR*, 2012.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE ICCV*, 2021.

[22] J. C. Dunn, "A Fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[23] R. Ranftl, Z. Bochkovskiy, and V. Kotun, "Vision transformers for dense prediction," in *Proc. IEEE ICCV*, 2021.

[24] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Video panopcic segmentation," in *Proc. IEEE CVPR*, 2020.

[25] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE ICCV*, 2019.

[26] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Mask2Former: Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE CVPR*, 2022.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016.

Cloud Removal in Hyperspectral Satellite Images Using Low-rank Tensor Completion

Chuong Hoang Vo, Truong Thanh Nhat Mai, and Chul Lee
Department of Computer Science and Artificial Intelligence, Dongguk University, Seoul, Korea
E-mail: vhchuong@dgu.ac.kr, mtntruong@mme.dongguk.edu, and chullee@dongguk.edu

***Abstract*—We propose an unfolding-based low-rank tensor completion (LRTC) algorithm for cloud removal in hyperspectral satellite images. We first formulate cloud removal as an LRTC-based joint optimization problem, incorporating handcrafted priors for hyperspectral image acquisition and implicit regularization functions to compensate for modeling inaccuracies. We then solve the optimization problem iteratively and develop a multistage deep unfolded network. In this network, each stage corresponds to an iteration of the iterative algorithm in which the optimization variables and regularizers are updated using closed-form solutions and learned deep networks, respectively. Experimental results demonstrate that the proposed algorithm achieves better restoration performance than state-of-the-art algorithms in both quantitative and qualitative comparisons.**

I. Introduction

Advancements in remote sensing technology have enabled the acquisition of hyperspectral images (HSIs) with higher spatial, spectral, and temporal resolutions. Because of this higher-quality image data, HSIs are frequently used in Earth observation applications [1] including environmental monitoring [2], object detection [3], and land cover classification [4]. However, as clouds cover most of the Earth's land surface [5], spaceborne optical satellites cannot record entirely noise-free data of the Earth's surface. Therefore, cloud removal in HSIs has become a crucial issue in enhancing the usability of satellite imagery, for which various approaches have been developed.

Conventional cloud removal approaches for HSIs can be broadly divided into two categories of algorithms: model-driven and data-driven. Model-driven approaches formulate optimization problems by incorporating prior knowledge such as spatial information, spectral correlations among bands, and data from different sensors. Spatial approaches consider cloud removal as an image inpainting problem and restore occluded regions using spatial information from neighboring cloud-free regions within the same image [6]–[8]. While they are effective in filling small cloud-covered regions, they may struggle to restore larger occluded regions with complex structures. Spectral-based algorithms [9]–[11] exploit the correlations among spectral bands in HSIs. For example, in [9], [10], a low-rank tensor completion (LRTC) model was adopted to exploit spatial, spectral, and temporal information to recover cloud regions. However, because they adopted handcrafted priors such as smoothness and low-rank properties to regularize multispectral images, they may fail to capture the rich structure of RSIs, resulting in degraded performance.

More recently, data-driven approaches using deep neural networks have been actively developed, demonstrating significant performance improvements due to their superior learning capabilities. In particular, convolutional neural networks (CNNs) [12], [13], generative adversarial networks (GANs) [14], and transformers [15] have been employed for cloud removal. However, constructing large datasets for training networks is challenging, which also makes the training process computationally expensive. Furthermore, using predefined cloud masks in a training set limits adaptability to real data containing diverse characteristics of cloud regions, including different resolutions and spectral bands.

To address the aforementioned limitations of conventional approaches, we propose an unfolding-based algorithm based on LRTC for cloud removal in HSIs. First, we formulate the cloud removal task as a joint optimization problem, which comprises handcrafted priors for HSI observation model and implicit regularizers to compensate for modeling inaccuracies. In addition, a learnable attention block is employed to maintain multidimensional spatial dependencies and to focus on important regions. We then solve the optimization problem iteratively by dividing the optimization problem into subproblems that can be independently solved. In each iteration, the learnable regularizers are updated by CNNs, and optimization variables are updated by closed-form solutions. Experimental results show that the proposed algorithm outperforms state-of-the-art algorithms in quantitative and qualitative comparisons.

II. Proposed Algorithm

A. Problem Formulation

Let $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ both in $\mathbb{R}^{H\times W\times C}$ denote the cloudy and clean images, respectively, where H, W, and C are the height, width, and the number of spectral channels. Then, the cloud-contaminated HSI acquisition can be modeled as

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{M}} \odot \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{C}}, \tag{1}$$

where $\boldsymbol{\mathcal{M}}$ and $\boldsymbol{\mathcal{C}}$ denote the cloud mask and the cloud component, respectively, and $\odot$ represents element-wise multiplication.

The goal of cloud removal is to reconstruct a cloud-free image $\boldsymbol{\mathcal{X}}$ given a cloudy image $\boldsymbol{\mathcal{Y}}$, provided that the cloud-free regions in $\boldsymbol{\mathcal{X}}$ are the same as those in $\boldsymbol{\mathcal{Y}}$. Specifically, the

optimization problem for cloud removal can be formulated as

$$\begin{aligned} &\underset{\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{M}}}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{\mathcal{Y}}-\boldsymbol{\mathcal{M}}\odot\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{C}}\|_F^2 \\ &\text{subject to } \mathcal{P}_{\boldsymbol{\mathcal{M}}^{\mathsf{c}}}(\boldsymbol{\mathcal{X}})=\mathcal{P}_{\boldsymbol{\mathcal{M}}^{\mathsf{c}}}(\boldsymbol{\mathcal{Y}}), \end{aligned} \tag{2}$$

where $\boldsymbol{\mathcal{M}}^{\mathsf{c}}$ is the complement of $\boldsymbol{\mathcal{M}}$, and $\mathcal{P}_\Omega$ denotes a sampling operator in the observed region Ω as

$$[\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{X}})](i,j,k)=\begin{cases}\boldsymbol{\mathcal{X}}(i,j,k), & \text{if } (i,j,k)\in\boldsymbol{\Omega},\\ 0, & \text{otherwise.}\end{cases} \tag{3}$$

The optimization problem in (2) is ill-posed because it requires the recovery of the unknown cloud-free image $\boldsymbol{\mathcal{X}}$, cloud $\boldsymbol{\mathcal{C}}$, and cloud mask $\boldsymbol{\mathcal{M}}$ from a single cloudy image $\boldsymbol{\mathcal{Y}}$. Therefore, we introduce constraints for each optimization variable to obtain a solution. First, assuming that adjacent spectral bands in the HSIs are highly correlated and can be effectively captured using a low-rank representation, we model the cloud-free background image as a low-rank tensor. Second, since the cloud-covered regions are relatively small compared to the HSI, we model the cloud region as a sparse noise component. In this work, to incorporate these constraints, we employ a tensor factorization strategy [16]. Specifically, the low-rank tensor $\boldsymbol{\mathcal{X}}$ with tubal rank r can be factorized into the t-product of two tensors $\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{L}}*\boldsymbol{\mathcal{R}}$, where $\boldsymbol{\mathcal{L}}\in\mathbb{R}^{H\times r\times C}$ and $\boldsymbol{\mathcal{R}}\in\mathbb{R}^{r\times W\times C}$. Then, the optimization in (2) becomes

$$\begin{aligned} &\underset{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{M}},\boldsymbol{\mathcal{C}}}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{\mathcal{Y}}-\boldsymbol{\mathcal{M}}\odot\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{C}}\|_F^2+\frac{1}{2}\|\boldsymbol{\mathcal{L}}*\boldsymbol{\mathcal{R}}-\boldsymbol{\mathcal{X}}\|_F^2+\lambda\|\boldsymbol{\mathcal{M}}\|_1 \\ &\text{subject to } \mathcal{P}_{\boldsymbol{\mathcal{M}}^{\mathsf{c}}}(\boldsymbol{\mathcal{X}})=\mathcal{P}_{\boldsymbol{\mathcal{M}}^{\mathsf{c}}}(\boldsymbol{\mathcal{Y}}), \end{aligned} \tag{4}$$

where the parameter λ controls the relative importance among the three terms.

Note that the second term $\|\boldsymbol{\mathcal{L}}*\boldsymbol{\mathcal{R}}-\boldsymbol{\mathcal{X}}\|_F^2$ in (4) does not consider the spatial and channel dependencies in the tensor $\boldsymbol{\mathcal{X}}$, potentially degrading reconstruction performance. To address this limitation, we employ an attention mechanism from [17], which can improve the accuracy of the reconstructed tensor by focusing on important regions and preserving the multidimensional spatio-channel dependencies. In addition, because the optimization problem in (4) is based on certain assumptions, *i.e.*, the low-rankness of $\boldsymbol{\mathcal{X}}$ and the sparseness of $\boldsymbol{\mathcal{M}}$, it may fail to capture real-world scenarios precisely, thus degrading performance. Therefore, to compensate for modeling inaccuracies, we introduce two implicit regularization functions, $g:\mathbb{R}^{H\times W\times C}\to\mathbb{R}$ and $h:\mathbb{R}^{H\times W\times C}\to\mathbb{R}$, for $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{M}}$, respectively. Then, by integrating the attention and regulation functions, the optimization in (4) can be rewritten as

$$\begin{aligned} &\underset{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{M}},\boldsymbol{\mathcal{C}}}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{\mathcal{Y}}-\boldsymbol{\mathcal{M}}\odot\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{C}}\|_F^2 \\ &\qquad\qquad+\frac{1}{2}\|\sqrt{\mathcal{F}_{\text{att}}(\boldsymbol{\mathcal{Y}})}\odot(\boldsymbol{\mathcal{L}}*\boldsymbol{\mathcal{R}}-\boldsymbol{\mathcal{X}})\|_F^2 \\ &\qquad\qquad+\lambda\|\boldsymbol{\mathcal{M}}\|_1+g(\boldsymbol{\mathcal{X}})+h(\boldsymbol{\mathcal{M}}) \\ &\text{subject to } \ \mathcal{P}_{\boldsymbol{\mathcal{M}}^{\mathsf{c}}}(\boldsymbol{\mathcal{X}})=\mathcal{P}_{\boldsymbol{\mathcal{M}}^{\mathsf{c}}}(\boldsymbol{\mathcal{Y}}). \end{aligned} \tag{5}$$

B. Solution to the Optimization

We solve the optimization problem in (5) iteratively using the augmented Lagrange multiplier method [18]. To this end, auxiliary variables $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\mathcal{H}}$ are introduced for variable splitting, which reformulates the optimization in (5) as

$$\begin{aligned} &\underset{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{M}},\boldsymbol{\mathcal{C}}}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{\mathcal{Y}}-\boldsymbol{\mathcal{M}}\odot\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{C}}\|_F^2 \\ &\qquad\qquad+\frac{1}{2}\|\sqrt{\mathcal{F}_{\text{att}}(\boldsymbol{\mathcal{Y}})}\odot(\boldsymbol{\mathcal{L}}*\boldsymbol{\mathcal{R}}-\boldsymbol{\mathcal{X}})\|_F^2 \\ &\qquad\qquad+\lambda\|\boldsymbol{\mathcal{M}}\|_1+g(\boldsymbol{\mathcal{G}})+h(\boldsymbol{\mathcal{H}}) \\ &\text{subject to } \ \boldsymbol{\mathcal{G}}=\boldsymbol{\mathcal{X}}, \ \ \boldsymbol{\mathcal{H}}=\boldsymbol{\mathcal{M}} \\ &\qquad\qquad\mathcal{P}_{\boldsymbol{\mathcal{M}}^{\mathsf{c}}}(\boldsymbol{\mathcal{X}})=\mathcal{P}_{\boldsymbol{\mathcal{M}}^{\mathsf{c}}}(\boldsymbol{\mathcal{Y}}). \end{aligned} \tag{6}$$

Next, we define the augmented Lagrangian function for (6) as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{M}},\boldsymbol{\mathcal{G}},\boldsymbol{\mathcal{H}},\boldsymbol{\Lambda},\boldsymbol{\Gamma},\boldsymbol{\Phi}) \\ &=\frac{1}{2}\left\|\boldsymbol{\mathcal{Y}}-\boldsymbol{\mathcal{M}}\odot\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{C}}\right\|_F^2+\lambda\|\boldsymbol{\mathcal{M}}\|_1 \\ &\quad+\frac{1}{2}\left\|\sqrt{\mathcal{F}_{\text{att}}(\boldsymbol{\mathcal{Y}})}\odot(\boldsymbol{\mathcal{L}}*\boldsymbol{\mathcal{R}}-\boldsymbol{\mathcal{X}})\right\|_F^2 \\ &\quad+\frac{\mu}{2}\|\boldsymbol{\mathcal{M}}^{\mathsf{c}}\odot(\boldsymbol{\mathcal{Y}}-\boldsymbol{\mathcal{X}})\|_F^2+\langle\boldsymbol{\Lambda},\boldsymbol{\mathcal{M}}^{\mathsf{c}}\odot(\boldsymbol{\mathcal{Y}}-\boldsymbol{\mathcal{X}})\rangle \\ &\quad+g(\boldsymbol{\mathcal{G}})+\frac{\alpha}{2}\|\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{G}}\|_F^2+\langle\boldsymbol{\Gamma},\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{G}}\rangle \\ &\quad+h(\boldsymbol{\mathcal{H}})+\frac{\beta}{2}\|\boldsymbol{\mathcal{M}}-\boldsymbol{\mathcal{H}}\|_F^2+\langle\boldsymbol{\Phi},\boldsymbol{\mathcal{M}}-\boldsymbol{\mathcal{H}}\rangle, \end{aligned} \tag{7}$$

where $\boldsymbol{\Lambda}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Phi}\in\mathbb{R}^{H\times W\times C}$ are tensors of Lagrange multipliers; μ, α, and $\beta>0$ are penalty parameters; and $\langle\cdot,\cdot\rangle$ denotes the tensor inner product.

The optimal solutions to the problem in (6) can be obtained by minimizing the augmented Lagrangian function $\mathcal{L}$ as

$$\begin{aligned} &(\boldsymbol{\mathcal{L}}^*,\boldsymbol{\mathcal{R}}^*,\boldsymbol{\mathcal{X}}^*,\boldsymbol{\mathcal{M}}^*,\boldsymbol{\mathcal{C}}^*,\boldsymbol{\mathcal{G}}^*,\boldsymbol{\mathcal{H}}^*) \\ &=\underset{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{M}},\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{G}},\boldsymbol{\mathcal{H}}}{\arg\min}\ \mathcal{L}(\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{M}},\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{G}},\boldsymbol{\mathcal{H}},\boldsymbol{\Lambda},\boldsymbol{\Gamma},\boldsymbol{\Phi}). \end{aligned} \tag{8}$$

Since it is intractable to directly solve the joint optimization problem in (8), we adopt the alternating direction method of multipliers [19], which decomposes it into manageable subproblems for each variable and multiplier. These subproblems are then iteratively solved. We present how each subproblem is solved at the k-th iteration.

$\boldsymbol{\mathcal{X}}$-subproblem: Given variables from the previous iteration, $\boldsymbol{\mathcal{X}}$ is updated by

$$\boldsymbol{\mathcal{X}}_{k+1}=\frac{\mathcal{F}_{\text{att}}(\boldsymbol{\mathcal{Y}})\odot(\boldsymbol{\mathcal{L}}_k*\boldsymbol{\mathcal{R}}_k)+\boldsymbol{\Psi}_k}{\boldsymbol{\mathcal{M}}_k\odot\boldsymbol{\mathcal{M}}_k+\mathcal{F}_{\text{att}}(\boldsymbol{\mathcal{Y}})-\mu_k\boldsymbol{\mathcal{M}}_k^{\mathsf{c}}+\alpha_k\mathbf{1}}, \tag{9}$$

where $\boldsymbol{\Psi}_k=\boldsymbol{\mathcal{M}}_k\odot(\boldsymbol{\mathcal{Y}}-\boldsymbol{\mathcal{C}}_k)-\mu_k\boldsymbol{\mathcal{M}}_k^{\mathsf{c}}\odot\boldsymbol{\mathcal{Y}}+\boldsymbol{\Lambda}_k\boldsymbol{\mathcal{M}}_k^{\mathsf{c}}+\alpha_k\boldsymbol{\mathcal{G}}_k-\boldsymbol{\Gamma}_k$, and division is performed by elementwise.

$\boldsymbol{\mathcal{L}}$- and $\boldsymbol{\mathcal{R}}$-subproblems: Then, we update $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{R}}$ simultaneously by solving

$$(\boldsymbol{\mathcal{L}}_{k+1},\boldsymbol{\mathcal{R}}_{k+1})=\underset{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{R}}}{\arg\min}\ \frac{1}{2}\left\|\sqrt{\mathcal{F}_{\text{att}}(\boldsymbol{\mathcal{Y}})}\odot(\boldsymbol{\mathcal{L}}*\boldsymbol{\mathcal{R}}-\boldsymbol{\mathcal{X}}_{k+1})\right\|_F^2. \tag{10}$$

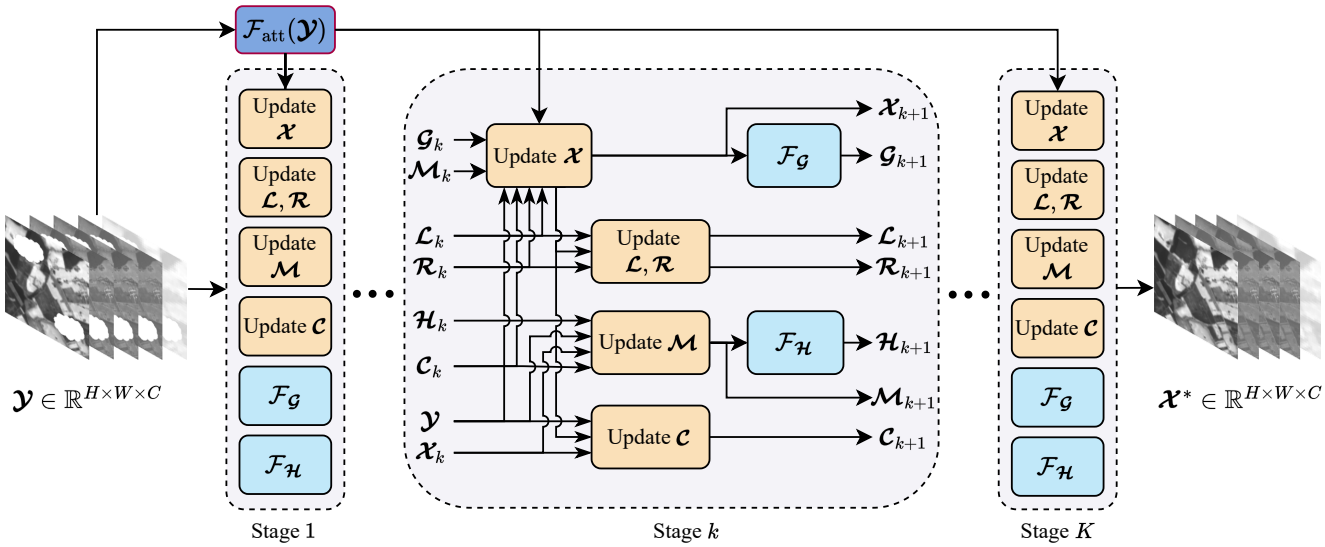


Fig. 1: Overview of the proposed algorithm. The network comprises an attention module and K stages.

We obtain the closed-form solutions for $\mathcal{L}$ and $\mathcal{R}$, respectively, in the Fourier domain [17] as

$$\mathcal{L}_{k+1}^{(i)} = \mathcal{F}_{\text{FFT}}^{-1}\left\{\widehat{\mathcal{X}}_{k+1}^{(i)}\left(\widehat{\mathcal{R}}_k^{(i)}\right)^H\left(\widehat{\mathcal{R}}_k^{(i)}\left(\widehat{\mathcal{R}}_k^{(i)}\right)^H\right)^\dagger\right\}, \quad (11)$$

$$\mathcal{R}_{k+1}^{(i)} = \mathcal{F}_{\text{FFT}}^{-1}\left\{\left(\left(\widehat{\mathcal{L}}_k^{(i)}\right)^H\widehat{\mathcal{L}}_k^{(i)}\right)^\dagger\left(\widehat{\mathcal{L}}_k^{(i)}\right)^H\widehat{\mathcal{X}}_{k+1}^{(i)}\right\}, \quad (12)$$

where $\mathbf{A}^H$ and $\mathbf{A}^\dagger$ represent the conjugate transpose and the pseudoinverse of matrix $\mathbf{A}$, respectively, and $\widehat{\mathcal{A}}(i,j,:) = \mathcal{F}_{\text{FFT}}\{\mathcal{A}(i,j,:)\}$.

$\mathcal{M}$-subproblem: The initial cloud mask $\widehat{\mathcal{M}}$ is obtained by

$$\widehat{\mathcal{M}} = \frac{\mathcal{X}_k \odot (\mathcal{Y} - \mathcal{C}_k) - \lambda_k \mathbf{1} + \zeta_k}{\mathcal{X}_k \odot \mathcal{X}_k + \mu_k(\mathcal{Y} - \mathcal{X}_k)\odot(\mathcal{Y} - \mathcal{X}_k) + \beta_k \mathbf{1}}, \quad (13)$$

where $\zeta_k = \mu_k(\mathcal{Y} - \mathcal{X}_k) \odot (\mathcal{Y} - \mathcal{X}_k) + \mathbf{\Lambda}_k(\mathcal{Y} - \mathcal{X}_k) + \beta_k\mathcal{H}_k - \mathbf{\Phi}_k$. Then, we perform thresholding to constrain the values in $\mathcal{M}$ to be binary, *i.e.*,

$$\mathcal{M}_{k+1}(i,j) = \begin{cases} 1, & \text{if } |\widehat{\mathcal{M}}(i,j)| > \epsilon, \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

$\mathcal{C}$-subproblem: The cloud component $\mathcal{C}$ is updated by

$$\mathcal{C}_{k+1} = \mathcal{Y} - \mathcal{M}_{k+1} \odot \mathcal{X}_{k+1}. \quad (15)$$

$\mathcal{G}$- and $\mathcal{H}$-subproblems: We employ CNNs, $\mathcal{F}_{\mathcal{G}_k}$ and $\mathcal{F}_{\mathcal{H}_k}$, for regulation functions $g(\cdot)$ and $h(\cdot)$, respectively, at the k-iteration to learn them from real-world data. Specifically, the auxiliary variables $\mathcal{G}$ and $\mathcal{H}$ are updated as

$$\begin{aligned}\mathcal{G}_{k+1} &= \arg\min_{\mathcal{G}} g(\mathcal{G}) + \frac{\alpha_k}{2}\|\mathcal{X}_{k+1} - \mathcal{G}\|_F^2 + \langle \mathbf{\Gamma}, \mathcal{X}_{k+1} - \mathcal{G}\rangle \\ &= \mathcal{F}_{\mathcal{G}_k}\left(\mathcal{X}_{k+1} + \alpha_k^{-1}\mathbf{\Gamma}_k\right), \quad (16)\end{aligned}$$

$$\begin{aligned}\mathcal{H}_{k+1} &= \arg\min_{\mathcal{H}} h(\mathcal{H}) + \frac{\beta_k}{2}\|\mathcal{M}_{k+1} - \mathcal{H}\|_F^2 + \langle \mathbf{\Phi}, \mathcal{M}_{k+1} - \mathcal{H}\rangle \\ &= \mathcal{F}_{\mathcal{H}_k}\left(\mathcal{M}_{k+1} + \beta_k^{-1}\mathbf{\Phi}_k\right). \quad (17)\end{aligned}$$

Multipliers: Finally, we update the Lagrange multipliers tensors as

$$\begin{aligned}\mathbf{\Lambda}_{k+1} &= \mathbf{\Lambda}_k + \mu_k \mathcal{M}_{k+1}^{\mathsf{c}} \odot (\mathcal{Y} - \mathcal{X}_{k+1}), \\ \mathbf{\Gamma}_{k+1} &= \mathbf{\Gamma}_k + \alpha_k(\mathcal{X}_{k+1} - \mathcal{G}_{k+1}), \\ \mathbf{\Phi}_{k+1} &= \mathbf{\Phi}_k + \beta_k(\mathcal{M}_{k+1} - \mathcal{H}_{k+1}). \end{aligned} \quad (18)$$

C. Unfolded Network

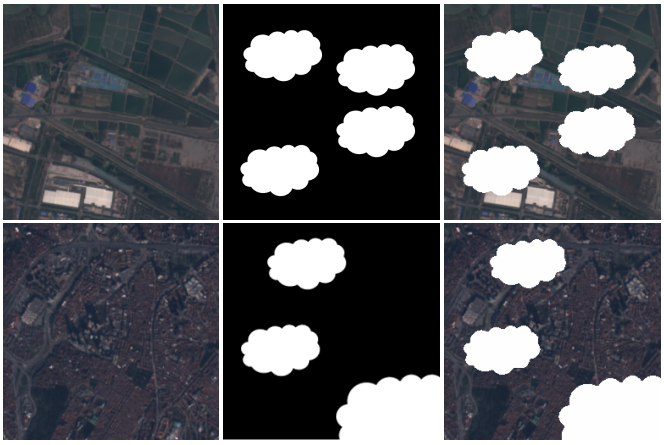
We construct a deep network for the proposed iterative optimization algorithm employing the algorithm unfolding strategy. Fig. 1 shows the architecture of the proposed deep unfolded network. In this work, we employ the attention module from AHDRNet [20]. In each stage k, the optimization variables $\mathcal{X}$, $\mathcal{L}$, $\mathcal{R}$, $\mathcal{M}$, and $\mathcal{C}$ are updated through closed-form solutions in (9), (11), (12), (14), and (15), respectively. Next, the auxiliary variables $\mathcal{G}$ and $\mathcal{H}$ are updated using two CNNs $\mathcal{F}_{\mathcal{G}}$ and $\mathcal{F}_{\mathcal{H}}$, respectively. We use the residual dense block (RDB) [21] with two convolution layers for both $\mathcal{F}_{\mathcal{G}}$ and $\mathcal{F}_{\mathcal{H}}$. The output $\mathcal{X}_K$ of the final stage K is the reconstructed HSI $\mathcal{X}^*$.

D. Implementation

We train the proposed network using a weighted l_1 loss, which is the weighted sum of the l_1-norm between the ground-truth $\mathcal{Y}$ and the output $\mathcal{X}_k$ at each stage with linearly increasing weights, given by

$$L = \sum_{k=1}^{K} \frac{2k}{K(K+1)}\|\mathcal{Y} - \mathcal{X}_k\|_1. \quad (19)$$

We use the Adam optimizer [22] with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 100 epochs. The initial learning rate is set to 10^{-5} and is reduced by a factor of 2 every 10 epochs. We fix the number of blocks K to 10 for all experiments and set the threshold value $\epsilon = 0.5$ in (14).



(a) Cloud-free images (b) Cloud masks (c) Synthetic images

Fig. 2: Examples of synthetic images in the synthetic dataset.

TABLE I: Quantitative comparison of cloud removal performance. For each metric, the **boldfaced** and underlined numbers denote the best and the second-best results, respectively.

	RMSE (↓)	PSNR (↑)	SAM (↓)	SSIM (↑)
SpA-GAN [14]	0.0246	32.60	0.0817	0.9139
DSen2-CR [12]	0.0169	36.14	0.0472	0.9334
GLF-CR [15]	0.0179	35.54	**0.0391**	0.9378
UnCRtainTS [13]	0.0176	35.70	3.5127	0.9559
Proposed	**0.0167**	**36.47**	0.0431	**0.9622**

Dataset: We generate a synthetic dataset to simulate real-world thick clouds using cloud-free HSIs from the SEN12MS-CR dataset [23] with 13-channel 256×256 images. We generate cloud masks with random shapes and coverage levels in the range of $[0.15, 0.45]$, as shown in Fig. 2(b).

We synthesize a cloudy image using the cloud mask as

$$\boldsymbol{\mathcal{Y}}(i,j,c) = \begin{cases} 1, & \text{if } \boldsymbol{\mathcal{M}}(i,j) = 1, \\ \boldsymbol{\mathcal{X}}(i,j,c), & \text{otherwise.} \end{cases} \tag{20}$$

The synthetic dataset has 5,000 images for training, 100 for validation, and 100 for testing. Fig. 2 shows example images in the synthetic dataset.

III. Experimental Results

A. Quantitative and Qualitative Comparison

We evaluate the performance of the proposed algorithm with those of state-of-the-art algorithms—SpA-GAN [14], DSen2-CR [12], GLF-CR [15], and UnCRtainTS [13]. We use four quality metrics: root mean squared error (RSME), peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [24], and spectral angle mapper (SAM) [25]. Higher PSNR and SSIM scores indicate better performance, while lower RSME and SAM scores indicate higher quality.

Table I provides a quantitative comparison of cloud removal performance. The proposed algorithm consistently outperforms the conventional algorithms in terms of RMSE, PSNR, and SSIM. In particular, the proposed algorithm yields 0.0063 higher SSIM scores than the second-best algorithm UnCRtainTS, indicating its effectiveness in preserving important structural details. This is due to the superior ability of the attention mechanism to focus on important regions in the scene with the incorporation of the low-rank prior. Note that although GLF-CR [15] yields the best SAM score, this comes at the cost of potential spectral inconsistencies, as demonstrated by poor PSNR performance.

Fig. 3 compares the cloud removal results obtained by each algorithm on the synthetic dataset. Conventional algorithms generate noticeable color and texture artifacts, significantly degrading the visual quality of the results. For example, SpA-GAN [14] in Fig. 3(b) generates textures in occluded regions, but they do not match those in adjacent regions. DSen2-CR [12], GLF-CR [15], and UnCRtainTS [13] in Figs. 3(c), (d), and (e), respectively, yield more plausible textures. However, the boundaries of the generated regions are noticeable. In contrast, the proposed algorithm in Fig. 3(f) reconstructs cloud regions most faithfully, with more accurate details and colors. These results also confirm that the proposed algorithm can capture the underlying spatial-spectral correlations, enabling a more accurate and contextually consistent reconstruction of the missing information.

TABLE II: Effects of weighted l_1 loss on cloud removal performance.

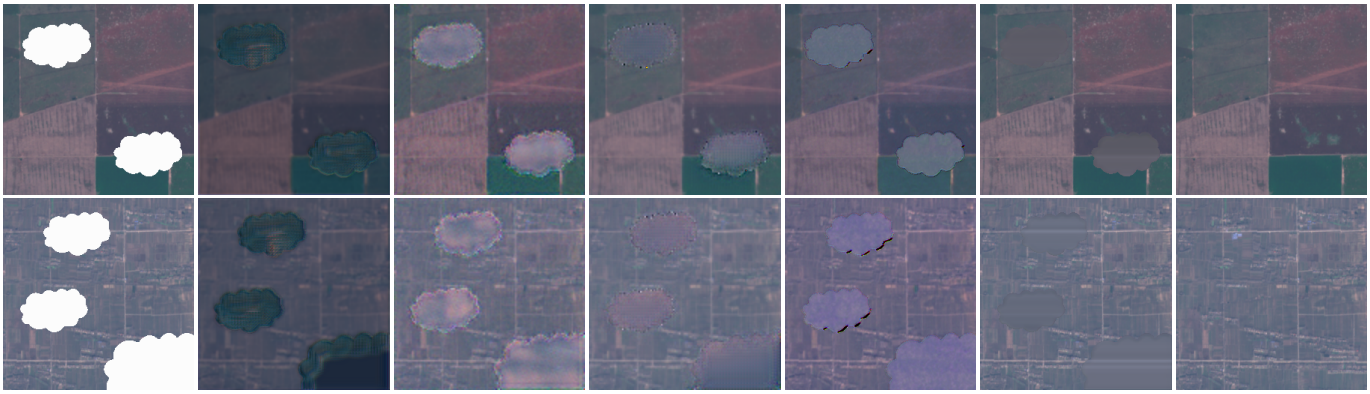
	RMSE (↓)	PSNR (↑)	SAM (↓)	SSIM (↑)
l_1 loss	0.0175	36.08	0.0453	0.9500
Weighted l_1 loss	**0.0167**	**36.47**	**0.0431**	**0.9622**

B. Effectiveness of Weighted l_1 Loss

We analyze the effectiveness of the weighted l_1 loss in (19) on the performance by training the proposed network using different settings. Table II compares the performance obtained using conventional and weighted l_1 losses. The use of the weighted l_1 loss increases scores across all metrics. This is because the weighted l_1 loss enables the network to be progressively supervised at each stage, balancing the amount of learned information throughout the stages.

IV. Conclusions

We proposed an unfolded LRTC network for cloud removal in HSIs. Specifically, we first formulated a cloud removal task as a joint optimization problem based on LRTC by incorporating handcrafted priors for HSI acquisition and implicit regularization functions to compensate for modeling inaccuracies. We then solved the optimization problem using a deep unfolded framework comprising a series of blocks, in which the optimization variables are updated through closed-form solutions and regularizers are updated through learned deep networks. Experimental results demonstrated that the proposed algorithm achieves better reconstruction performance than state-of-the-art algorithms.



(a) Input (b) SpA-GAN [14] (c) DSen2-CR [12] (d) GLF-CR [15] (e) UnCRtainTS [13] (f) Proposed (g) Ground truth

Fig. 3: Qualitative comparison of the cloud removal results on the synthetic dataset.

ACKNOWLEDGMENT

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1F1A1074402) and in part by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under the ICAN (ICT Challenge and Advanced Network of HRD) grant funded by MSIT (RS-2023-00260248).

REFERENCES

[1] D. Hong, B. Zhang, H. Li, *et al.*, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, p. 113 856, 2023.

[2] J. Peng, S. Chen, H. Lü, Y. Liu, and J. Wu, "Spatiotemporal patterns of remotely sensed PM2.5 concentration in china from 1999 to 2011," *Remote Sens. Environ.*, vol. 174, pp. 109–121, 2016.

[3] N. Yokoya and A. Iwasaki, "Object detection based on sparse representation and Hough voting for optical remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, pp. 2053–2062, 2015.

[4] D. Hong, L. Gao, N. Yokoya, *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, 2021.

[5] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the terra and aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, 2013.

[6] A. Maalouf, P. Carre, B. Augereau, and C. Fernandez-Maloigne, "A bandelet-based inpainting technique for clouds removal from remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2363–2371, 2009.

[7] L. Lorenzi, F. Melgani, and G. Mercier, "Inpainting strategies for reconstruction of missing data in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 8, no. 5, pp. 914–918, 2011.

[8] H. Shen, H. Li, Y. Qian, L. Zhang, and Q. Yuan, "An effective thin cloud removal procedure for visible remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, pp. 224–235, 2014.

[9] M. K.-P. Ng, Q. Yuan, L. Yan, and J. Sun, "An adaptive weighted tensor completion method for the recovery of remote sensing images with missing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3367–3381, 2017.

[10] W. He, N. Yokoya, L. Yuan, and Q. Zhao, "Remote sensing image reconstruction using tensor ring completion and total variation," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8998–9009, 2019.

[11] W.-J. Zheng, X.-L. Zhao, Y.-B. Zheng, J. Lin, L. Zhuang, and T.-Z. Huang, "Spatial-spectral-temporal connective tensor network decomposition for thick cloud removal," *ISPRS J. Photogramm. Remote Sens.*, vol. 199, pp. 182–194, 2023.

[12] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, 2020.

[13] P. Ebel, V. S. F. Garnot, M. Schmitt, J. Wegner, and X. X. Zhu, "UnCRtainTS: Uncertainty quantification for cloud removal in optical satellite time series," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2023, pp. 2086–2096.

[14] H. Pan, "Cloud removal for remote sensing imagery via spatial attention generative adversarial network," *arXiv preprint arXiv:2009.13015*, 2020.

[15] F. Xu, Y. Shi, P. Ebel, *et al.*, "GLF-CR: SAR-enhanced cloud removal with global-local fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 192, pp. 268–278, 2022.

[16] P. Zhou, C. Lu, Z. Lin, and C. Zhang, "Tensor factorization for low-rank tensor completion," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1152–1163, 2018.

[17] T. T. N. Mai, E. Y. Lam, and C. Lee, "Attention-guided low-rank tensor completion," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024, accepted for publication.

[18] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," University of Illinois, Urbana-Champaign, Tech. Rep. UILU-ENG-09-2215, 2009.

[19] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[20] Q. Yan, D. Gong, Q. Shi, *et al.*, "Attention-guided network for ghost-free high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019.

[21] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[23] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[25] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, *et al.*, "The spectral image processing system (SIPS)—Interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2–3, pp. 145–163, 1993.

Color Guided Disease Segmentation for Plant Images

So-Yeon Jang and Jong-Ok Kim
School of Electrical Engineering, Korea University, Korea
E-mail: {wkdjsy77, jokim}@korea.ac.kr

***Abstract*—This paper proposes a network architecture designed for semantic disease segmentation on plant images for smart agriculture. To effectively segment challenging small disease regions, the proposed method incorporates hierarchical fusions with color information into the conventional encoder-decoder network. The color information within input images can help the network to easily find regions associated with a target class. We extract color information by color layer decomposition, and it works as guidance for semantic features during the fusion process. The hierarchical fusion of encoder features and color features is highly informative, highlighting each class and enhancing disease segmentation performance. The proposed method exceeds existing semantic segmentation methods and handles small areas associated with diseases more effectively.**

I. Introduction

In agriculture crop management, it is crucial to discern regions affected by diseases from healthy ones. Relying on human judgment inherently leads to a time-consuming identification process, imposing a significant burden on human resources. Moreover, there is a fundamental limitation to require agricultural expertise in identifying plant disease areas.

Semantic segmentation could capture the portion of objects in an image through pixel-wise recognition. In the field of agriculture, it can help to determine the degree of disease damage automatically. Several studies have proposed deep learning methods for plant segmentation [1], [2] . In particular, some of them [3], [4], [5] concentrate on plant and disease segmentation in outdoor scenes with crop cultivation which we are considering in this paper. However, it is very challenging to accurately identify and segment disease areas on plant images. Plant images are frequently taken by drones in the sky, and due to long shooting distance, disease areas are captured much smaller on plant images. Thus, it is more difficult to segment small disease areas rather than relatively large plant ones. Moreover, as the network becomes deeper, boundary information is likely to be lost, and this makes segmentation results coarse. In particular, plant leaves are partially occluded, making it more difficult to learn the characteristics of each class in outdoor images.

In this paper, we propose a semantic segmentation network for outdoor plant images with three classes: plant, disease, and background (BG). The proposed network consists of two subnets, which are color layer decomposition (CLD) and color-guided segmentation (CGS). Our approach is inspired by the observation that objects with similar colors are likely to be in the same class. To decompose an image into the representative colors of a class, we first construct color palettes before training the network. Color palettes are obtained separately for each class using initial segmentation results by the pretrained network. K-means clustering is used for constructing a color palette for each class with its initial segmented image. In the CLD subnet, the input image is decomposed into color layers corresponding to palette colors, and they play a role of guidance for class-specific semantic features. In the CGS subnet, color guidance involves cross-attention between the color features and semantic features. By employing color features as a guide, we can highlight regions in the semantic information corresponding to disease areas where the segmentation tends to be coarse. Furthermore, the features are hierarchically fused, aiming at training the network which is applicable to various object sizes. Through experimental results, it was demonstrated that the proposed network outperforms the conventional methods in both quantitative and qualitative ways.

II. Related Work

Semantic segmentation plays an important role in the accurate identification of diseases. Some of the previous works such as UNet [6], ERFNet [7], ResNet [8], and Deeplabv3+ [9], have been proposed built upon the foundation of the Fully Convolutional Network (FCN) [10]. As these neural networks become deep, there is an inherent risk of feature information degradation, particularly in the context of small objects. To address this challenge, several studies have been done for small object segmentation. HRNet [11] is a network that maintains high-resolution representations to improve spatial accuracy. BiSeNetV2 [12] enhances performance by aggregating detailed and semantic information. Similarly, STDC-seg [13] employs a fusion approach of both detailed and semantic information by obtaining detailed ground truth.

As prior research on palette-based color decomposition, Zhang et al. [14] introduced a method using K-means to extract representative colors, creating a palette and obtaining linear combinations of color vectors per pixel. Tan et al. [15] decompose an RGB image into additive mixing layers based on a palette, introducing a geometric approach that is more efficient than previous methods. Akimoto et al. [16] presented Fast Soft Color Segmentation (FSCS), an efficient color segmentation methodology using a palette. These studies primarily focus on color segmentation for re-coloring, unlike object segmentation in this paper.

Several studies have introduced a multi-scale fusion method for integrating rich semantic informative high-level features and low-level features with high resolution. [17] and [18]

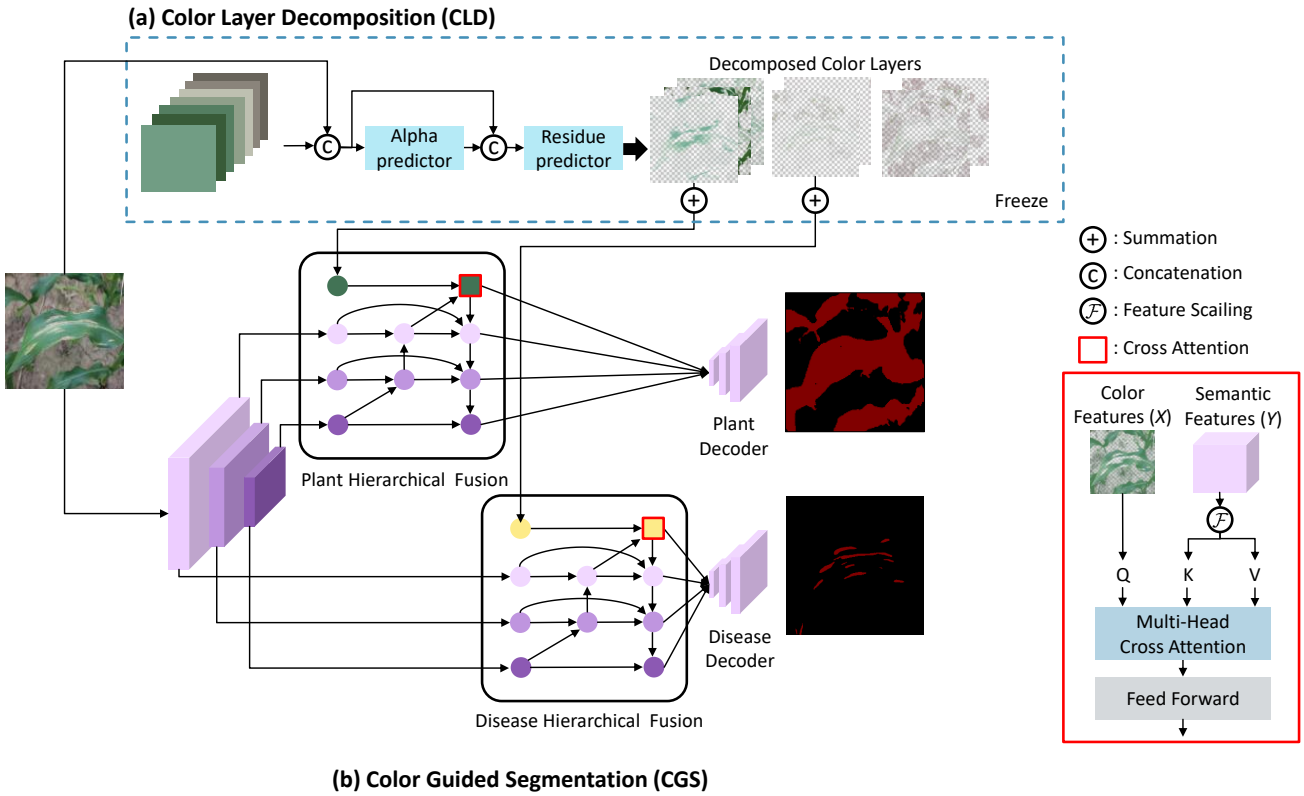


Fig. 1. The proposed network architecture which consists of (a) the color layer decomposition (CLD) subnet, and (b) the color guided segmentation (CGS) subnet.

employed a multi-scale feature fusion to reduce the semantic gap between feature maps in different scales and effectively learn objects of various sizes. However, applying these fusion techniques may not be effective due to differences between color features and encoder features. Based on the transformer, TransUNet [19] connects encoder features to the decoder, interpreting transformer results. SegFormer [20] employs a hierarchical transformer encoder to predict segmentation masks at various resolutions. UCTransNet [21] proposed a transformer module that takes inputs from four different scales.

Despite the progress of these studies, the segmentation of small objects remains an active and ongoing area of exploration. In this paper, we propose a new small object oriented segmentation network to include color guidance and hierarchical multi-scale fusion, in particular for small disease objects.

III. The Proposed Method

The proposed network comprises two subnets as illustrated in Fig. 1: Color Layer Decomposition (CLD) and Color-Guided Segmentation (CGS). In this section, we first introduce the CLD subnet, which is specifically designed for decomposing the segmented result of each class into color layers using a class-specific palette. Subsequently, we present the CGS architecture employed for the segmentation of plant and disease classes with hierarchical color fusion.

A. Color Layers Decomposition

The CLD subnet performs palette-based color layer decomposition of each class such as plant, disease, and BG as illustrated in Fig. 1. However, determining a color palette through clustering for the disease class proves to be challenging. It is difficult to obtain representative colors in disease regions due to their small size. Therefore, we adopt a pretrained segmentation network [7] which initially segments the input image into three classes. We can obtain class-specific RGB images by multiplying the pre-segmented results for each class, and a color palette is established independently every class. K-means clustering is applied to obtain a color palette for each class-specific RGB images. Fig. 2 shows an example of color extraction where the palette size is k for plant, l for disease, and m for BG. Once color palettes for all three classes are built, an input image is decomposed into multiple color layers. Then the palette color images and the input image are fed into the CLD subnet to generate color layers for each class.

In contrast to hard color segmentation that decomposes colors discretely, soft color segmentation probabilistically generates segmentation outputs. It allows for easy correction even in regions where color decomposition errors may occur. This approach ensures high confidence in areas with abundant color

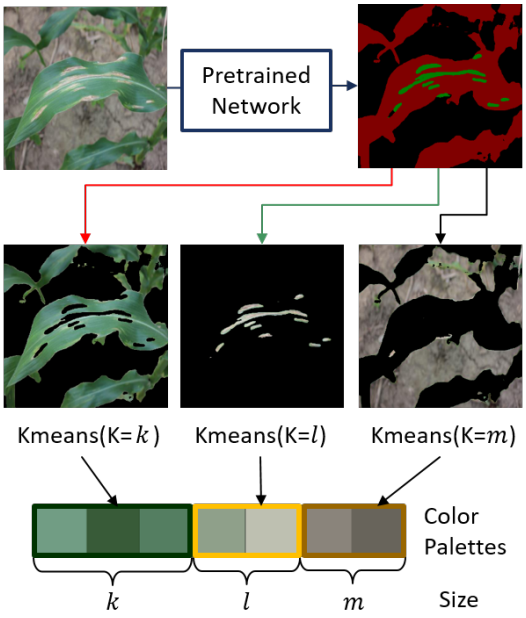


Fig. 2. Color palette is separately generated by applying K-means clustering to each segmented region (plant, disease, and BG).

information and allows more accurate prediction by considering false negative areas. Therefore, we conducted soft color segmentation method FSCS [16]. It allocates the probability to each color.

The CLD subnet is pretrained and frozen during the segmentation process. We first concatenate the input RGB image and the all-class color palettes which are broadcasted to match the shape of the input image, and the concatenated one is then fed into the alpha predictor. The alpha predictor in Fig. 1 decomposes the input image into the corresponding color layers in the palette. The residue predictor fills in the missing parts in the original image with the alpha predictor results and the palettes as inputs. Finally, decomposed color layers are obtained for each color palette. The color layers of the plant and disease classes are fused with the encoder of the CGS subnet.

B. Color-Guided Segmentation Network

Hierarchical feature fusion allows for effective integration of color information by combining semantic encoder features with high-resolution color features for small objects. Inspired by EfficientDet [17], we introduce a modified version of Bi-directional Feature Pyramid Network (BiFPN) to fuse features across different scales. Unlike traditional approaches, scaling in hierarchical feature fusion is not uniformly applied to all features. Instead, adjustments are made to the width, depth, and resolution of features, aligning them with each encoder feature scale of the backbone network through convolution and interpolation. This leads to slight adjustments in feature resolution, allowing minimal alteration of existing information.

However, due to the domain differences between the encoder and color features, summation fusion makes the fused features coarse. Cross-attention is applied to those features for fusion. We extend Restormer [22] to leverage its inherent computational efficiency. Cross-attention effectively addresses information disparities between the color and encoder features by dynamically adjusting the relative importance of each element, considering the semantic correlation between the two sequences. This approach can overcome the issue of inaccuracies in color decomposition by highlighting incorrect features. The color layer of each class is denoted by X for query (Q), and the low level feature is by Y for key (K) and value (V). $\mathcal{F}$ represents the feature scaling operation. The corresponding formula for feature scaling is expressed as:

$$Y' = \mathcal{F}(Y), \tag{1}$$

$$Q = XW_Q, K = Y'W_K, V = Y'W_V, \tag{2}$$

where W is a projection function and cross-attention is calculated as:

$$CrossAtten(Q, K, V) = V \cdot Softmax\left(\frac{K^T \cdot Q}{\alpha}\right), \tag{3}$$

where α is a learnable scaling parameter.

Two decoders are employed to interpret color-guided features for each class. To emphasize the color information specific to each class, hierarchical fusion is performed individually for the plant and disease classes. Although the encoder features are shared, the color layers X_p and X_d are utilized to obtain each query. In the case of two decoder outputs, the cross-entropy loss function enhances the training process by quantifying the disparity between the predicted output and the ground truth. The balancing parameter, λ, addresses class imbalance resulting from the extent of difference between plant and disease regions. In this context, the mathematical expression for the cross-entropy loss is defined as:

$$L = -(\lambda_p \sum_i y_i \log(\hat{y}_i) + \lambda_d \sum_j y_j \log(\hat{y}_j)) \tag{4}$$

where $i \in \{\text{plant, BG}\}$, $j \in \{\text{disease, BG}\}$.

IV. Experimental Results

A. Implementation Details

The CLD subnet is pretrained on the Places [23] dataset using a palette size of 7, extracted with an image by a traditional palette extract method. For semantic segmentation, we utilized the field images of maize annotated with Northern leaf blight (NLB) disease [24] for training, consisting of 81 images for training and 10 for testing. The images are resized to a resolution of 512 x 512 by bilinear interpolation. To the best of our knowledge, the NLB dataset is the only real-world open resource for disease detection. We constructed a new dataset with drone images of napa cabbages affected by Plasmodiophora brassicae in Korea. The experiments used 996 images for training and 50 images for test, each sized 512x512. ERFNet [7] is employed as the pretrained network in Fig. 2 and is trained with NLB. Furthermore, considering future deployment on drones, the lightweight ERFNet was used

TABLE I
IOU COMPARISONS FOR NLB DATASET AND NAPA CABBAGE DATASET

	NLB dataset			**Napa cabbage dataset**		
	Plant IoU	Disease IoU	mIoU	Plant IoU	Disease IoU	mIoU
UNet [6]	0.842	0.413	0.628	0.985	0.733	0.859
ERFNet [7]	0.838	0.366	0.602	0.986	0.734	0.860
Deeplabv3+ [9]	0.830	0.406	0.618	**0.988**	0.723	0.855
HRNet [11]	0.832	0.390	0.611	0.983	0.722	0.853
BiSeNetV2 [12]	0.837	0.416	0.627	0.982	0.719	0.851
TransUNet [19]	0.825	0.398	0.612	0.984	0.737	0.861
SegFormer [20]	0.834	0.400	0.617	0.981	0.661	0.821
STDC-Seg [13]	0.840	0.398	0.619	0.987	0.706	0.845
UCTransNet [21]	**0.843**	0.408	0.626	0.985	0.730	0.857
Proposed	0.842	**0.478**	**0.660**	0.986	**0.739**	**0.862**

as the backbone for the CGS subnet. The proposed network adopts the Adam optimizer for optimization, and the initial learning rate is set to 1e-3. In the loss function, we set the λ_p and λ_d values to 0.3 to mitigate class imbalance.

B. Comparison Results

When compared against conventional and state-of-the-art methods, our proposed approach demonstrates higher performance in the challenging task of small object segmentation (i.e., disease region). The proposed method is compared with commonly used techniques in semantic segmentation in Table I. and Fig. 3. The proposed network achieves the highest disease intersection over union (IoU) among small objects segmentation methods [6], [7], [9], [12], [13], [11] . Disease IoU and mIoU are even better compared with the transformer-based methods [19], [20], [21].

Fig. 3 shows that the proposed method achieves the highest detection behaviors. The proposed method exhibits the sharpest disease segmentation and effectively distinguishes between weeds and plants. This distinction is achieved not merely through color information but by fusing semantic information, demonstrating a complementary interaction between color features and semantic features. This hierarchical fusion enhances the accuracy of plant disease detection.

C. Ablation Study

The ablation studies aim to comprehensively analyze the impact of the proposed modules in our network. Without CLD, the proposed method actually relies on semantic features only without color guidance. As confirmed in Table II, color guidance has the largest performance gain for disease segmentation, significantly demonstrating its impact. Hierarchical fusion plays a crucial role in identifying small objects by learning from diverse scales of features, preventing the loss of information in small regions. This module demonstrates performance improvement, even in challenging edges, as depicted in Fig. 4. The use of summation without cross-attention led to a lower disease IoU compared to the proposed method. Because of the different domains of color and semantic features, simple summation tends to highlight incorrect information. Leveraging color features as queries within the cross-attention mechanism learn the mapping between color information and semantic features. This facilitates more accurate segmentation for small objects whose semantic features are hard to capture.

TABLE II
QUANTATIVE RESULTS OF ABLATION STUDIES ON NLB DATASET

	Plant IoU	Disease IoU	mIoU
w/o Color Guidance	0.849	0.428	0.639
w/o Hierarchical Fusion	**0.850**	0.443	0.647
w/o Cross Attention	0.846	0.448	0.647
Proposed	0.842	**0.478**	**0.660**

D. Effect of Color Palette Size

The performance of the proposed network was also evaluated under a variety of palette sizes. Table III lists the segmentation performance across different palette: sizes which are 3, 7, and 9 for plant, disease, and BG, respectively. The total palette size 3 is too small for representative color extraction, and it produces coarse color decomposition. Total palette size 9 has too many colors, which, on the contrary, hinders decomposition. For total palette size of 7, we achieved the best performance or disease segmentation. It seems like that the size of color palette depends on the extent of color diversity in the corresponding class. Thus, it would be better to configure its size adaptively, considering the color distribution of the class object.

V. CONCLUSIONS

The proposed method introduces a semantic segmentation framework with color guidance and hierarchical fusion. The CLD subnet focuses on color layer decomposition using palette colors, while the CGS subnet employs a hierarchical fusion of color features and encoder features, ensuring effective small object segmentation. Cross-attention addresses domain differences during fusion, dynamically adjusting the relative importance of elements based on semantic correlation. The

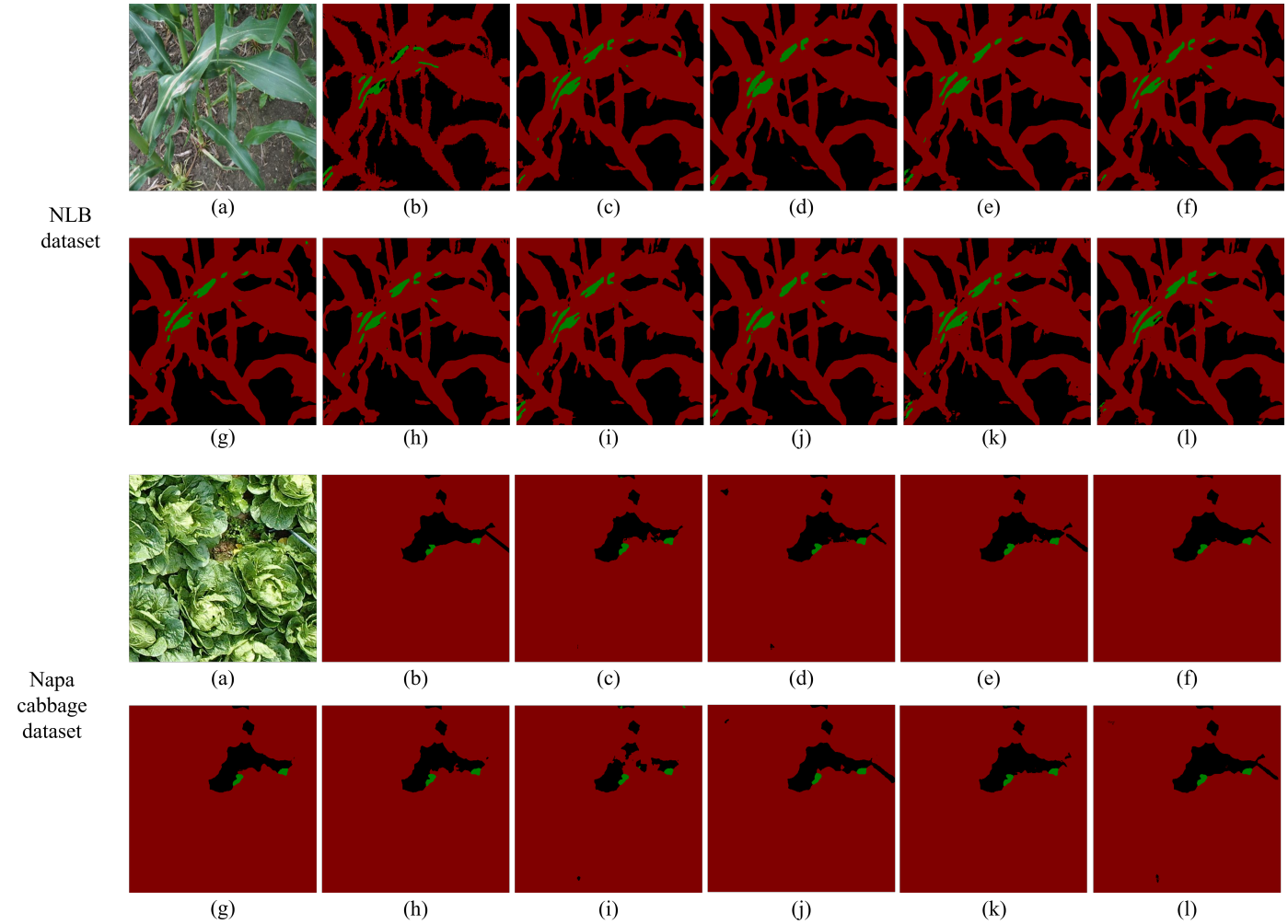


Fig. 3. Visualization comparisons on NLB dataset and Napa cabbage dataset. (a) RGB image, (b) ground truth, (c) UNet, (d) ERFNet, (e) DeepLabV3+, (f) HRNet, (g) BiSeNetV2, (h) TransUNet, (i) SegFormer, (j) STDC-Seg (k) UCTransNet, (l) proposed.

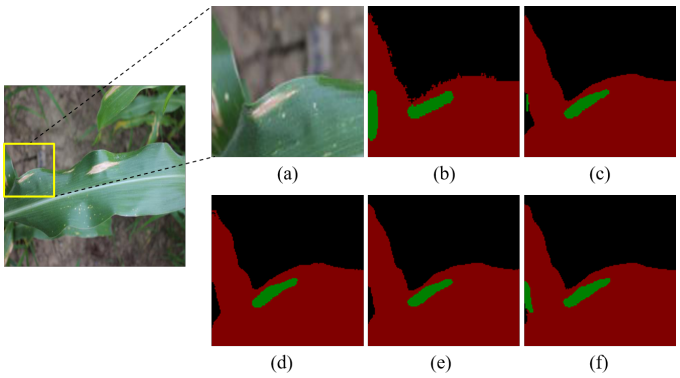


Fig. 4. Visualization of ablation studies results. (a) RGB image, (b) ground truth, (c) w/o color guidance, (d) w/o hierarchical fusion, (e) w/o cross attention, (f) proposed.

experimental results demonstrate that our approach can enhance the disease segmentation performance on plant images. The fusion of color and semantic features shows accurate and visually appealing results.

TABLE III
EXPERIMENTAL RESULTS VIA PALETTE SIZE

Palette Size (Plant, Disease, BG)	Plant IoU	Disease IoU	mIoU
3 (1, 1, 1)	0.847	0.455	0.651
7 (3, 2, 2)	0.842	0.478	0.660
9 (3, 3, 3)	0.837	0.442	0.640

ACKNOWLEDGMENT

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through Open Field Smart Agriculture Technology Short-term Advancement Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (322033-3).

References

[1] V. Singh and A. K. Misra, "Detection of plant leaf diseases using image segmentation and soft computing techniques," *Information processing in Agriculture*, vol. 4, no. 1, pp. 41–49, 2017.

[2] J.-Y. Jung, S.-H. Lee, and J.-O. Kim, "Knowledge transfer based spatial embedding network for plant leaf instance segmentation," *IEIE Transactions on Smart Processing & Computing*, vol. 12, no. 2, pp. 162–170, 2023.

[3] S.-E. Lee, S.-H. Lee, and J.-O. Kim, "Deep-clustering based plant disease segmentation network," in *2023 International Conference on Electronics, Information, and Communication (ICEIC)*, IEEE, 2023, pp. 1–3.

[4] G.-Y. Moon and J.-O. Kim, "Crop & match: Roi cropping and feature matching for segmentation of small objects," in *2023 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, IEEE, 2023, pp. 1–4.

[5] G.-Y. Moon and J.-O. Kim, "Roi-attention network for small disease segmentation in crop images," *IEEE Access*, 2024.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[7] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[11] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[12] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, pp. 3051–3068, 2021.

[13] M. Fan, S. Lai, J. Huang, *et al.*, "Rethinking bisenet for real-time semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9716–9725.

[14] Q. Zhang, C. Xiao, H. Sun, and F. Tang, "Palette-based image recoloring using color decomposition optimization," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1952–1964, 2017.

[15] J. Tan, J. Echevarria, and Y. Gingold, "Efficient palette-based decomposition and recoloring of images via rgbxy-space geometry," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–10, 2018.

[16] N. Akimoto, H. Zhu, Y. Jin, and Y. Aoki, "Fast soft color segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8277–8286.

[17] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[18] J. Fan, M. J. Bocus, B. Hosking, *et al.*, "Multi-scale feature fusion: Learning better semantic segmentation for road pothole detection," in *2021 IEEE International Conference on Autonomous Systems (ICAS)*, IEEE, 2021, pp. 1–5.

[19] J. Chen, Y. Lu, Q. Yu, *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[20] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[21] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, pp. 2441–2449.

[22] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[23] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018. DOI: 10.1109/TPAMI.2017.2723009.

[24] T. Wiesner-Hanks, E. L. Stewart, N. Kaczmar, *et al.*, "Image set for deep learning: Field images of maize annotated with disease symptoms," *BMC research notes*, vol. 11, no. 1, pp. 1–3, 2018.

MYMV: A Music Video Generation System with User-preferred Interaction

Kyungjune Lee, Mingyu Jang, Jungwoo Huh, Jeonghaeng Lee, Seokkeun Choi, and Sanghoon Lee*
Yonsei University, Seoul 03722, South Korea
E-mail: {naive2kj90, jmg1002, gjwjddn9, leedoright, csg, slee}@yonsei.ac.kr

***Abstract*—The advancements in computer vision and graphics have increased the demand for creating multi-modal content with dynamic 3D faces, such as in singing or talking face generation. However, due to the requirement for specialized knowledge in each modality, creating or reproducing such content is highly demanding in terms of manual effort. To fill this void, we present Make Your Music Video (MYMV), which enables users to easily produce multi-modal content. Our proposed system is featured for 1) song generation, 2) facial motion generation, and 3) virtual backgrounds generation. After all modalities are generated, they are combined into a 3D facial music video. Through our user-preferred interface, users can directly edit songs and backgrounds, facilitating easy participation in the music video production process. To evaluate our system, we construct a music video dataset from results of MYMV. The evaluation results show that our system achieves metrics indicating the generation of sufficiently natural 3D facial music videos. The demonstration video is available at https://github.com/naive2kj90/MYMV.**

I. Introduction

With the revolutionary advancements in computer vision and graphics techniques, the demand for creating multi-modal content using dynamic 3D faces has increased. For example, combining 3D faces with other modality components, which can be utilized for various types of content, has been proposed in forms such as singing [1] or talking [2], [3] face generation. However, creating or reproducing such multi-modal content is difficult because it requires manual labor with professional knowledge to combine each modality component.

Furthermore, while advancements in Artificial Intelligence across various modality components are being pursued with the objective of achieving peak performance within their respective fields, there is a lack of discourse on how to integrate these technologies into a unified system for user delivery.

To address the aforementioned issue, we propose a novel system called Make Your Music Video (MYMV) that allows users to easily produce a multi-modal content. Our proposed system divides the multi-modal content into three modality components: 1) *Song*, 2) *Dynamic 3D faces*, and 3) *Virtual backgrounds* that suit the atmosphere of the song. First, our system requests users to input song metadata, including the title, genre, and lyrics, to generate a song. Subsequently, it generates a 3D singing facial motion synchronized with the song. Lastly, our system generates a virtual background suitable for the atmosphere of the song. These generated modalities are combined into the 3D facial music video using the Physically Based Rendering (PBR) process, which renders 3D singing facial motion by considering illumination effects. Moreover, we provide users with a user-preferred interface, which enables users to show their preference for results. At each pipeline, the generated result is displayed on the interface. If the user is not satisfied with the result, such as generated songs and virtual backgrounds, they can request the system to regenerate new results. This offers users an intuitive system for direct engagement in 3D facial music video production.

To demonstrate the efficiency of the proposed system, we construct a music video dataset and conduct a system evaluation. The music video dataset consists of output videos generated from MYMV, including a total of 80 music videos featuring two songs from ten different genres, each represented by four different identities. We present 80 videos to 23 participants, who are asked to evaluate the music video dataset using a Likert 5-point scale (with '5' indicating satisfaction). The proposed system achieves a singing naturalness score of 4.02 and a visual harmony score of 3.83.

II. System

In this section, we describe the operation of MYMV as shown in Fig.1. First, our proposed system requests the user to provide song metadata, including the title, genre, and lyrics, along with the multi-view images. Then, the system generates each modality component of content through three pipelines: 1) Song generation, 2) Facial motion generation, and 3) Background environments generation. The system offers an interface for users to select each modality component according to their preferences. If the user is unsatisfied with the generated modality component, they can choose to regenerate it. Finally, the system renders a music video by combining all generated modality components, which are the song, 3D facial motion, and the background environments, using PBR. More details are described in the subsections.

A. Modality components generation

1) Song generation: To generate the song using the requested song metadata, we employ the transformer-based Suno API [4] which produces a song from the given text prompt. Then, our system separates vocals and accompaniment from the generated song through vocal separation, a U-net-based method proposed in [5]. The disentanglement of song components is essential because the potential noise from mixed

* Corresponding author

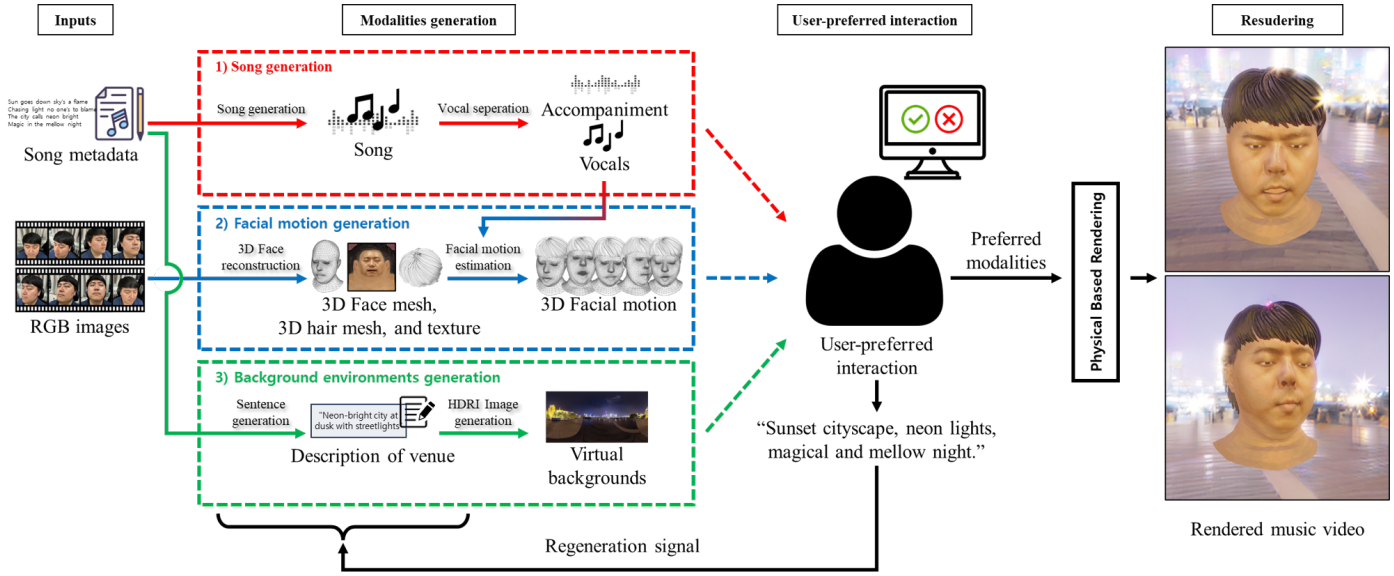


Fig. 1. The overview of our MYMV system.

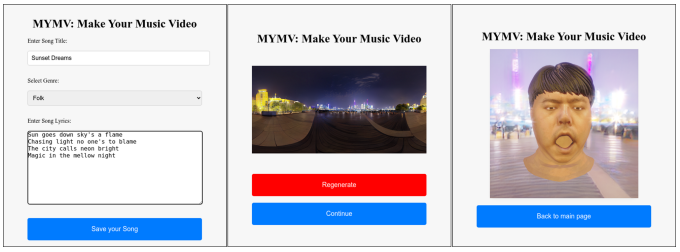


Fig. 2. The user-preferred interaction interface. The user can decide to use or regenerate the modality components.

vocals and accompaniment may become an artifact for facial motion generation in the forthcoming Sec. II-A2.

2) Facial motion generation: To easily animate a 3D face, we utilize a 3D morphable model (3DMM) [6], which enables us to obtain both 3D face geometry and textures through a linear combination of coefficients. To get an accurate initial 3D face mesh, we reconstruct the facial geometry and estimate camera poses based on input multi-view images using Agisoft Metashape [7]. Then, we fit the 3D face to the scanned geometry by optimizing the shape parameters and using an ICP-based rigid deformation method [8]. Most facial textures can be obtained by projecting multi-view images onto UV coordinates as similar as [9]. However, certain occluded regions, like the ears, may remain incomplete due to their intricate shapes. The effective way to represent hair involves generating each 3D hair strand individually, but this is generally impractical due to the extensive computational resources required for physics simulations. Additionally, hair material exhibits anisotropic reflection characteristics, complicating realistic rendering. To achieve plausible rendering of 3D hair meshes, we treat hair as a single mesh created from separate strands. For generating the 3D hair mesh, we first use a strand-based hair estimator [10]. Then, we generate a solid 3D hair mesh along each strand by determining the direction at each point on the strands and merging all strands into a unified mesh following [11]. As a result, it is possible to deliver realism induced by detailed geometry, while maintaining moderate computational burden.

Based on previous studies [12] indicating that human hair typically exhibits a uniform color, we opt to render all hair with a uniform color for the sake of rapid computation and moderate computational burden. We segment the hair region in the input image and apply the average color of this region uniformly to the hair mesh.

Finally, we need to place the 3D hair mesh and the 3D face mesh obtained from the input image into the same space. Since the two 3D meshes are generated based on their respective local coordinates, their world coordinates do not align, resulting in the face and hair being positioned in different spaces. We perform the realignment by matching the position of the hair root r. Using an optimization method, we can compute the scale s, rotation R, and translation t between the two 3D meshes. Therefore, the hair root displacement is optimized by minimizing the following energy function:

$$E = \min \sum_{i=1}^{N} ||sRr_i^H + t - r_i^F||^2, \tag{1}$$

where E is the distance difference between the hair root from 3D face mesh r_i^F and the hair root from 3D hair mesh r_i^F. Position of i-th hair root r_i can be written as,

$$r_i = \lambda_{i1}\mathbf{v}_{i1} + \lambda_{i2}\mathbf{v}_{i2} + \lambda_{i3}\mathbf{v}_{i3}, \tag{2}$$

where λ_i is the barycentric coordinate is the barycentric coordinate of the root with respect to the vertices $(\mathbf{v}_{i1}, \mathbf{v}_{i2}, \mathbf{v}_{i3})$ of the triangle in which the root is located and N is number of total hair root. Realignment of hair root can make 3D face mesh with 3D hair mesh on the scalp region.



Fig. 3. Rendered music video samples with the same RGB images and different song metadata.

To generate a 3D facial motion driven by a singing voice and 3D face, most recent algorithms estimate mouth movements from input songs effectively but fail to adequately address the entire facial region, such as eye movements. Thus, we address these challenges by training CodeTalker [13] on our own song dataset. We built a dataset of 4 hours of in-the-wild singing videos and then fitted 3DMM sequences to the videos using EMOCA [14].

3) Virtual backgrounds generation: To make a more immersive and impressive music video, it is essential to use virtual backgrounds that suit the atmosphere of the music [15]. To generate suited virtual backgrounds, we use Text2Light [16] which produces high-resolution 360-degree high dynamic range images (HDRI) from the given text prompt. Therefore, for the text prompt, a description of the venue considering the song's atmosphere is needed. We use Prompt Engineering [17], [18] to generate a description from the input lyrics. One of the recent advanced models in LLMs, ChatGPT [19] has shown great capability in performing this task effectively and is appropriate for analyzing lyrics and generating the description.

B. User-preferred interaction

For the user-participatory music video production process, the system checks the user's preference for the resulting modality components of each pipeline using a user-preferred interaction interface as shown in Fig. 2. If the users are not satisfied with the result, they can regenerate it by editing the song metadata.

C. Rendering

Finally, once the user is satisfied with the overall results, the system combines them into a music video using the PBR process. The rendered music video is shown in Fig. 1 and the user can see it through the interface. The results can be seen in our supplemental video.

III. Experiments

A. Music video dataset

To effectively and clearly test the visual performance of our algorithms, we collected datasets from different song genres. Fig. 3 is some samples of our results, from the same input images, and different song metadata. For example, the same input image results in the same 3D face mesh, and different song metadata results in different song, facial motion, and background environments. This results in the same 3D face (same people) singing different songs. We generated 80 music videos from various song metadata, with 2 different metadata on each genre, and 4 different input images.

TABLE I
Likert-scale score for Singing Naturalness on music video dataset

Method	Singing Naturalness
Cudeiro, Daniel, et al. [3]	1.96
Richard, Alexander, et al. [20]	2.89
Fan, Yingruo, et al. [21]	3.21
Xing, Jinbo, et al. [13]	3.88
Ours	**4.02**

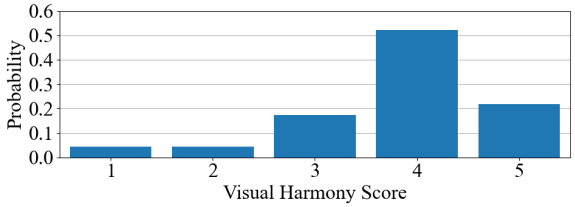


Fig. 4. Likert-scale score for visual harmony on music video dataset.

B. Applicability Performance

To compare whether our proposed algorithm is preferred by the users most, we perform a subjective test with generated music videos. The participants score a Likert-scale (1 to 5) for the singing naturalness and visual harmony of the rendered music video. The singing naturalness means how well the facial motion synchronizes with the song and visual harmony means the overall harmony between the song, facial movements, and background. The evaluation is conducted by 23 professionals in the field, ensuring a rigorous assessment of our algorithm's performance.

For the singing naturalness, we compared four widely used prior models with ours. Using the same generated input song (refer to Section. II-A1), we estimated facial motion using each model. The participants then viewed the resulting music videos and rated them based on the aforementioned criteria. The result is shown in Table. I. This evaluation provides valuable insights into the effectiveness of our proposed algorithm in generating realistic and harmonious music videos, as well as its comparative performance against existing models. This user-centric approach allows us to ensure that our algorithm not only performs well technically but also meets user expectations for visual and auditory coherence. For the visual harmony, participants were asked to score the overall harmony between the generated modality components such as song, facial movements, and background. We got 3.83 (standard deviation=0.96) as shown in Fig. 4, which implies that the participants well-received the combination of modality components. This result highlights the strength of our method in integrating various

elements to create a cohesive and visually pleasing music video.

Overall, our proposed algorithm demonstrates a significant improvement in singing naturalness when compared to existing models. In the first test, participants rated our algorithm higher than four widely used prior models for its ability to synchronize facial motion with the song effectively. In the second test, without direct comparison, the visual harmony was evaluated by asking participants to rate the overall harmony between the song, facial movements, and background. Participants gave favorable scores, indicating that the integration of modality components was well-received. The low standard deviation in visual harmony scores suggests consistent positive reception among participants, reinforcing the strength of our method in integrating various elements seamlessly. This comprehensive evaluation underscores the effectiveness of our approach in generating music videos that meet user expectations for both auditory and visual coherence, establishing its superiority in singing naturalness and its strong performance in visual harmony.

IV. Conclusion

We present MYMV, a music video generation system with user-preferred interaction. By utilizing multi-view images and song metadata as input, users can generate music videos preferred to their specifications. Our system can be utilized in the various user-generated content creation, offering an intuitive and accessible tool for music video production.

Acknowledgment

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C3011697). This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Global Talent for Generative AI Copyright Infringement and Copyright Theft, Project Number: RS-2024-00398413, Contribution Rate: 100%).

References

[1] P. Liu, W. Deng, H. Li, *et al.*, "Musicface: Music-driven expressive singing face synthesis," *Computational Visual Media*, vol. 10, no. 1, pp. 119–136, 2024.

[2] S. Lee, J. Lee, H. Song, and S. Lee, "Speech-driven emotional 3d talking face animation using emotional embeddings," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 7840–7844.

[3] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 101–10 111.

[4] Suno, *Create your own music video*, Accessed: 2024-05-21. [Online]. Available: https://suno.com/create.

[5] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, *Spleeter: A fast and state-of-the art music source separation tool with pre-trained models*, Late-Breaking/Demo ISMIR 2019, Deezer Research, Nov. 2019.

[6] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans.," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[7] Agisoft LLC, *Agisoft metashape professional*, https://www.agisoft.com/, Version 1.8.4, 2022.

[8] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, Spie, vol. 1611, 1992, pp. 586–606.

[9] K. Lee, J. Lee, H. Lee, M. Jang, S. Lee, and S. Lee, "Faceclone: Interactive facial shape and motion cloning system using multi-view images," in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, IEEE, 2023, pp. 512–513.

[10] Y. Zheng, Z. Jin, M. Li, *et al.*, "Hairstep: Transfer synthetic to real using strand and depth maps for single-view 3d hair modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 726–12 735.

[11] M. Jang, K. Lee, S. Lee, *et al.*, "Invite: Individual virtual transfer for personalized 3d face generation system," in *IJCAI*, 2024.

[12] J.-P. Ortonne and G. Prota, "Hair melanins and hair color: Ultrastructural and biochemical aspects," *Journal of investigative dermatology*, vol. 101, no. 1, S82–S89, 1993.

[13] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong, "Codetalker: Speech-driven 3d facial animation with discrete motion prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 780–12 790.

[14] R. Daněček, M. J. Black, and T. Bolkart, "Emoca: Emotion driven monocular face capture and animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 311–20 322.

[15] Y. Yan, Y. He, and L. Li, "Why time flies? the role of immersion in short video usage behavior," *Frontiers in Psychology*, vol. 14, p. 1 127 210, 2023.

[16] Z. Chen, G. Wang, and Z. Liu, "Text2light: Zero-shot text-driven hdr panorama generation," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–16, 2022.

[17] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–23.

[18] J. White, Q. Fu, S. Hays, *et al.*, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[19] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[20] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1173–1182.

[21] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 770–18 780.

Multi-Modal Video Summarization Based on Two-Stage Fusion of Audio, Visual, and Recognized Text Information

Zekun Yang * Jiajun He † and Tomoki Toda *
* Information Technology Center, Nagoya University, Japan
yang.zekun@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp
† Graduate School of Information Science, Nagoya University, Japan
jiajun.he@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—Multi-modal video summarization task provides a text summary by combining information from different inputs. Previous methods usually generate the summary from video and ground truth text inputs, but seldom study how audio, video, and ASR text influence the results. This work introduces a novel multi-modal video summarization approach that leverages video and audio inputs, along with speech-recognized text. Our method initiates by conducting speech recognition using a whisper ASR model, followed by representation of visual, textual, and audio information using video transformer, BART, and whisper, respectively. Subsequently, two-stage fusion between different modalities is applied. The fused features are fed into a BART decoder to generate text summaries. Experiments on the how2 dataset show that our method outperforms previous baselines. It is also revealed that incorporating audio information helps generate a better summary even if the texts contain certain errors.**

I. Introduction

E-learning has become a crucial way for people to acquire knowledge. With numerous educational videos being created and shared online every day, knowing their main contents and recommending videos based on the audience's interest become necessary. To address this issue, video summarization tasks have been proposed to retrieve essential information and generate a short and readable text summary [1]–[3].

In recent years, large-scale generative pre-trained language models [4]–[6] based on Transformers [7] have been proposed and applied in video summarization tasks, both in single modality and in multiple modalities. In single modality approaches, Sharma et al. [8], [9] utilizes audio input to transcribe speech into text and generate the summary. For multiple modalities, Yu et al. [3], Xu et al. [10], Qiao et al. [11], and Yuan et al. [12] leverage both visual and language features, employing information fusion techniques to produce summary texts.

Automatic speech recognition (ASR) techniques have made significant advancements by developing larger and more robust models [13]–[16] and making use of multi-modal information [17]–[21]. Moreover, ASR-oriented feature representation has been adopted in summarization [8], speech translation [22], emotion recognition [23] etc., showing its potential to provide necessary information for various applications. However, previous multi-modal summarization tasks tend to focus on video and ground truth text information, few researchers study how audio, video, and ASR text information jointly influence the quality of generated summary texts. Unfortunately, online videos usually do not contain subtitle texts. Since textual information is highly beneficial for video summarization [8], it is necessary to carry ASR to obtain audio transcripts and provide richer information beyond just audio and video inputs.

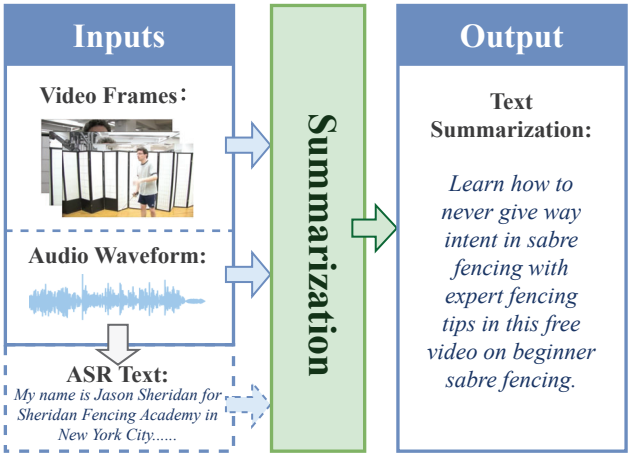


Fig. 1. The concept of the proposed video summarization method. Our method takes video and audio as inputs. We carry speech recognition and present two-stage fusion of visual, audio, and text to obtain text summaries.

In this work, we present a novel method that produces text summaries based on video and audio inputs, which is shown in Fig. 1. In this method, text is first generated by a whisper ASR model. Then, video transformer [7], BART [4] and frozen whisper [14] are used to extract features from visual, text, and audio modalities respectively. Two methods based on two-stage fusion are introduced to combine information from three different modalities (i.e. visual, text, and audio) together. A BART decoder is finally adopted to obtain the text summaries. We conduct experiments on the how2 dataset [24] to evaluate our method. The results show that our method outperforms previous baselines. We perform further experiments and prove

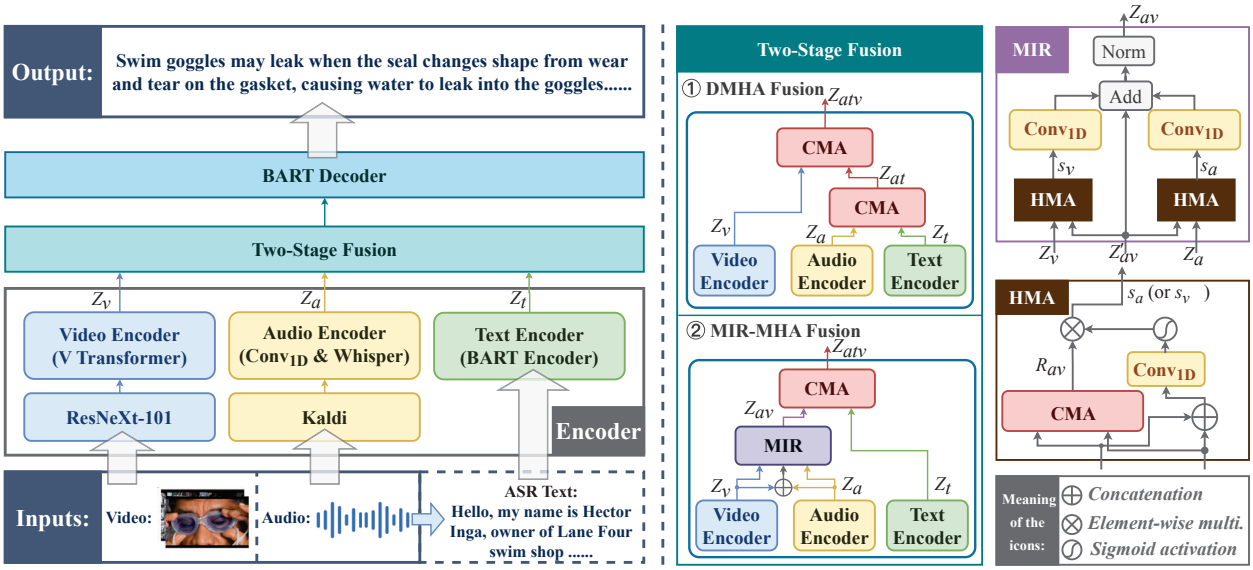


Fig. 2. The workflow of the proposed video summarization method. We propose two kinds of two-stage fusion methods, named DMHA fusion and MIR-MHA fusion, to combine information from different input modalities and generate the summary.

that audio input helps generate a better summary even if the input texts have certain errors.

II. Proposed Method

The workflow of our proposed method is shown in Fig. 2. This method takes video frames and audio waveforms as inputs and produces text summaries as the output. For the given video frames and audio waveform inputs, visual features, audio spectrograms, and ASR texts are first extracted from the corresponding input modalities. Then, video transformer [7], BART [4] and frozen whisper [14] are used to represent visual, text, and audio information respectively. Next, two-stage fusion strategies are introduced to combine information from video, audio, and text modalities together. The fused features are conveyed into a BART decoder to obtain the text summary. The detailed method will be introduced in the rest of this section.

A. Video and Audio Inputs

For video features, we follow previous work [2], [24] to retrieve 2048D representation by a 3D ResNeXt-101 model [25] pre-trained on the Kinetics dataset [26]. We further process this 2048D feature with a visual transformer [7] encoder, incorporating positional embeddings. The output from this process serves as our visual input.

For audio information, we follow previous work [24] to retrieve audio spectrogram from audio waveforms using Kaldi [27]. We extract 40D filter bank features from raw speech. Then, we obtain the audio feature representation using pre-trained whisper [14]. We set up a multi-layer convolution network to convert the input filter bank feature to 80D to fit the whisper inputs. We freeze the encoder and obtain the last hidden state from its output as the audio features.

B. ASR Text

Literature [8] has demonstrated that the text summary can be generated by fine-tuning an ASR model, which suggests that leveraging ASR texts contributes to the generation of text summaries. Motivated by this finding, we employ a whisper model [14] to generate ASR texts. We freeze the parameters in its encoder and decoder and just let whisper recognize speech and generate the corresponding text. Then, we use a BART encoder [4] to encode the generated texts.

C. Information Fusion

Information fusion at a deeper layer of the encoder tends to improve the quality of text summary [3]. Hence, we conduct two-stage fusion in the last layer of the encoder to combine information in video, audio, and text modalities together. We take ASR text as the main input flow to ensure the length is the same as that of the text. We employ two kinds of fusion methods called Dual Multi-Head-Attention (DMHA) fusion and Modality-Invariant-Representation Multi-Head-Attention (MIR-MHA) fusion respectively to combine the input information.

DMHA Fusion Yu et al [3] uses multi-head-attention to fuse text and visual information. We wonder how such a structure performs when audio information is also involved. Hence, we present DMHA fusion, which contains two multi-head-attention fusion. The fusion first carries A-T fusion which fuses audio and text information. Then it carries V-AT fusion which fuses the results of A-T fusion and the video information.

In A-T fusion, we use the last hidden state of BART encoder Z_t as the query and use the audio feature Z_a as both key and value. Linear projections W_Q, W_K and W_V are made to obtain Q_{AT}, K_{AT} and V_{AT}, respectively.

$$Q_{AT} = Z_t W_Q \tag{1}$$

TABLE I
EXPERIMENTAL RESULTS OF OUR METHOD AND OTHER METHODS, WHERE GT MEANS USING GROUND TRUTH TEXTS, ASR MEANS USING SPEECH-RECOGNIZED TEXTS.

Method	Input			Fusion		Rouge-1	Rouge-2	Rouge-L
	Modality	Text	Punctuation	Method	Order			
End-to-End [9]	A	–	–	–	–	60.9%	43.0%	55.9%
BASS [9]	A	–	–	–	–	64.0%	49.0%	60.1%
MCR [12]	V+T	GT	GT	MHA	(in parallel)	61.7%	45.2%	59.0%
VG-VPLM[*][3]	V+T	GT	GT	MHA	V-T	65.3%	49.3%	60.4%
	V+T	ASR	Fullstop	MHA	V-T	63.8%	47.2%	59.1%
	V+T	ASR	–	MHA	V-T	55.6%	36.7%	49.7%
Ours	V+A+T	GT	GT	DMHA	A-T+V-AT	64.7%	48.6%	59.5%
	V+A+T	ASR	Fullstop	DMHA	A-T+V-AT	66.0%	50.0%	61.3%
	V+A+T	ASR	Fullstop	DMHA	V-T+A-VT	64.9%	48.9%	60.8%
	V+A+T	ASR	–	DMHA	A-T+V-AT	59.6%	41.2%	53.7%
	V+A+T	GT	GT	MIR-MHA	A-V+AV-T	64.8%	48.7%	59.9%
	V+A+T	ASR	Fullstop	MIR-MHA	A-V+AV-T	**66.4%**	**50.2%**	**61.6%**
	V+A+T	ASR	–	MIR-MHA	A-V+AV-T	58.8%	40.4%	53.1%

*Based on our experiments.

$$K_{AT} = Z_a W_K \tag{2}$$

$$V_{AT} = Z_a W_V \tag{3}$$

We then set up cross-modal multi-head attention (CMA) to get the audio features R_{AT} based on the text query.

$$R_{AT} = \mathrm{CMA}(Q_{AT}, K_{AT}, V_{AT}) \tag{4}$$

To remove redundant information and noise, we use a forget gate on R_{AT} by setting up a forget gate mask M and doing element-wise multiplication (shown with $\otimes$) with R_{AT} to output the updated R'_{AT}:

$$M = \mathrm{sigmoid}(\mathrm{Concatenate}(R_{AT}, Z_t)W_m) \tag{5}$$

$$R'_{AT} = M \otimes R_{AT} \tag{6}$$

We concatenate Z_t and R'_{AT}, and linearly project it to specific dimensions with W_{AT} to obtain Z'_{at}.

$$Z'_{at} = \mathrm{Concatenate}(Z_t, R'_{AT})W_{AT} \tag{7}$$

We then obtain Z_{at} by adding Z'_{at} and Z_t.

$$Z_{at} = Z'_{at} + Z_t \tag{8}$$

We do a similar step for V-AT fusion. We use Z_{at} to obtain Q_{ATV}, and visual feature Z_v to obtain K_{ATV} and V_{ATV} respectively.

MIR-MHA Fusion Modality-Invariant Representation (MIR) [19], [23] captures the shared information in different modalities to ease the fusion process. In this work, we introduce MIR-MHA fusion based on MIR fusion. Our MIR-MHA fusion first carries A-V fusion for audio and video information using the MIR technique and then carries AV-T fusion to fuse the results of A-V fusion and the text information.

In A-V fusion, we first linearly project the audio feature Z_a and visual feature Z_v into the same dimension with W_a and W_v to obtain Z'_a and Z'_v, respectively.

$$Z'_a = Z_a W_a \tag{9}$$

$$Z'_v = Z_v W_v \tag{10}$$

We vertically concatenate Z'_a and Z'_v to obtain the shared information Z'_{av}.

$$Z'_{av} = \mathrm{Concatenate}(Z'_a; Z'_v) \tag{11}$$

Then, we convey Z'_a, Z'_v and Z'_{av} into the MIR-generator, where a hybrid-modal attention (HMA) is set up to extract information in each modality-specific representations.

$$s_m = \mathrm{HMA}(Z'_m, Z'_{av}), m \in \{a, v\} \tag{12}$$

where the details of the HMA module will be described later.

Next, the resulted s_m ($m \in \{a,v\}$) features are added to input sequence Z'_{av} to obtain modality-invariant representation Z^{inv}_{av}. 1×1 convolution with PReLU activation [28] and layer normalization [29] are used here.

$$Z^{inv}_{av} = \mathrm{Norm}(Z'_{av} + \sum_{m \in \{v,a\}} \mathrm{conv}(s_m)) \tag{13}$$

Finally, the modality-specific information and representations are concatenated to get the representations of MIR fusion Z_{av}.

$$Z_{av} = \mathrm{Concatenate}(s_v, s_a, Z^{inv}_{av}) \tag{14}$$

Where s_m contains information from both audio and visual modalities.

The HMA module in the MIR generator extracts representation for input audio and video features respectively using multi-head-attention. Here, Z'_{av} is used as query, audio (video) feature Z'_a (Z'_v) is used as both key and values.

$$R_{AV} = \mathrm{CMA}(Z'_{av}, Z'_m, Z'_m), m \in \{a, v\} \tag{15}$$

A parallel convolutional network is then used to learn the mask for the modality-specific information.

$$s_m = R_{AV} \otimes \sigma(\mathrm{conv}(\mathrm{Concatenate}(Z'_m, Z'_{av}))), m \in \{a, v\} \tag{16}$$

Where $\otimes$ means element-wise multiplication.

To fuse Z_{av} with text information on AV-T fusion, we set up another multi-head-attention fusion. We use the last hidden state of the original BART encoder Z_t as the query and use Z_{av} as both key and value. This multi-head-attention is similar to the one proposed in DMHA fusion.

D. Decoding and Generation

We use a BART decoder to decode the sequence. The features obtained by two-stage fusion are fed into the decoder, which generates potential hypotheses and their corresponding probabilities through beam search. The decoding proceeds in a left-to-right direction until the decoded token reaches an <end> mark or the maximum length of the generated summary. Finally, the hypothesis with the highest probability is chosen as the definitive summary.

III. Experimental Evaluations

A. Method Implementation

Dataset We conducted our experiments using the how2 dataset [24], which comprises 72,980 instructional videos spanning a total duration of 2,000 hours. This dataset provides video, audio, and text information. The audio information includes the sentence-level filter bank and the video-level filter bank, while the text information contains video transcripts and short text summaries. The audio filter bank in the dataset is derived from 16 kHz raw speech with a 25 ms time window and a 10 ms frame shift using Kaldi [27]. The videos in the dataset have a wide range of domains, with 68,333 for training, 2,520 for validation, and 2,127 for test. The corresponding sentence-level filter bank features are 1,013,715, distributed as 950,026 for training, 34,687 for validation, and 29,002 for test.

ASR Text Generation We used original filter bank features in the training set to generate ASR text. Since the input of whisper needs to be an 80D audio spectrogram, we establish a multi-layer convolution network preceding the pre-trained whisper model. We padded the filter bank features to fit the 3000 frame whisper input and trained the convolution layer before the whisper encoder, with the filter bank extracted from the top 2.5% (approximately 23,000) sentences. Subsequently, we employed this trained model to generate ASR text for the entire sentence-level dataset. To acquire ASR text at the video level, we concatenated the sentence-level ASR texts corresponding to each video.

Data Preprocessing We set the max length to 256 for video features, and 512 for text. The video/text sequences were padded or truncated to the corresponding max length before being conveyed into the model.

For audio features, video-level filter banks usually contain 6000-9000 frames [8], exceeding the capacity of the whisper input (3000 frames). To address this, we utilized whisper to generate audio features for every 3000 frames from the video-level filter bank. These generated frames were concatenated and down-sampled every 3 frames to fit the 3000-frame input of the whisper encoder.

Hyper Parameters We generated ASR text using pre-trained `whisper-tiny.en` model[1] with 4 layers in both encoder and decoder. We also generated the audio feature from the last hidden state of the encoder. We modeled the text feature using pre-trained `bart-base` model[2] with 6 layers in both encoder and decoder. Both models are not case-sensitive. For the video transformer, we used a 4-layer encoder with 8 attention heads and 2048 feed-forward dimensions. We set the max length of our generated summary to 64, the batch size to 8, and the learning rate to 3e-4. Adam optimizer [30] was used during optimization. The model was trained for 60 epochs.

Hardware and Software We performed our experiments on a computer with Intel Xeon Gold 6142 CPU, 128GB RAM, 3T HDD, and Nvidia Titan GPU. Our program was made using Python 3.9 and Pytorch 1.13.1 on Ubuntu 20.02.

[1]https://huggingface.co/openai/whisper-tiny.en

[2]https://huggingface.co/facebook/bart-base

B. Results and Analysis

We take End-to-End [8] and BASS [9] which use audio input only, VG-GPLM [3] and MCR [12] which use both video and ground truth text as our baseline. We adopt Rouge-1, Rouge-2, and Rouge-L [31] to evaluate the summarization results. We use a pre-trained fullstop model [32] to give punctuation to ASR text. We also test our method under the ground truth text with ground truth punctuation. Our results are in table I. Note that since whisper is not designed to be fine-tuned, we did not conduct experiments for video and audio inputs.

From the results, when we use V+A+T features with ASR text, our method has a better performance compared with VG-GPLM [3] that uses V+T features with ASR text, and End-to-End method [8]. Notably, our approach even outperforms models using ground truth text and punctuation. This means that information from different modalities (for V+T, it is A; for A, it is V+T) helps a lot to generate the summary.

We further compare results with and without punctuation, revealing better quality of the summary texts when punctuation is included. This is also reasonable since punctuation plays a crucial role in sentence segmentation and helps eliminate ambiguity in the input text.

IV. Discussion

Fusion Order From table I, we find that when we use DMHA fusion, the proposed method, which first fuses A-T features and then fuses V-AT features, generates a better text summary than the one that first fuses V-T features and then fuses A-VT features. The reason is considered as that, audio and text share similar information which will strengthen the representation of the fused feature when faced with video information. We also find that using MIR fusion with punctuation slightly improves the rouge score in summarization. This means that MIR gives a stronger feature representation when fusing video and text information.

The Contribution of Errors in Transcripts We find that our approach outperforms the model using ground truth text and punctuation. We wonder why this happens and conduct a deeper investigation on this problem using our model with DMHA fusion and VG-GPLM [3] model.

The first finding is that there is a space before each punctuation in the ground truth text, making the encoding of the same punctuation different between ground truth and ASR text. We remove space to form a GT-Fixed set and run experiments again. The results are in the middle of table II.

TABLE II
EXPERIMENTAL RESULTS OF OUR METHOD USING DIFFERENT KINDS OF TEXT, WHERE ASR MEANS USING SPEECH-RECOGNIZED TEXTS, GT MEANS USING GROUND TRUTH TEXTS, GT-FIXED MEANS USING GROUND TRUTH TEXTS WITH FIXED PUNCTUATION, REMOVE SHORT AND ADD EXTRA MEANS THE SIMULATION OF ASR SHORT DELETION AND INSERTION ERRORS RESPECTIVELY.

Text	Method	Rouge-1	Rouge-2	Rouge-L
ASR Text	Ours	66.0%	50.0%	61.3%
	VG-GPLM	63.8%	47.2%	59.1%
GT	Ours	64.7%	48.6%	59.5%
	VG-GPLM	65.3%	49.3%	60.4%
GT-Fixed	Ours	**66.2%**	50.2%	61.4%
	VG-GPLM	66.1%	50.1%	61.3%
Remove Short	Ours	66.1%	**50.3%**	**61.7%**
	VG-GPLM	66.0%	50.4%	61.6%
Add Extra	Ours	65.4%	49.4%	60.6%
	VG-GPLM	64.5%	48.5%	59.5%

We also wonder how ASR errors affect the results. We compare the ASR transcripts and ground truth transcripts and simulate ASR short deletion and insertion errors. Here, we define *short deletion* as no more than 10 tokens between ASR text and ground truth text in each transcript and we insert all the extra tokens, i.e., all ASR insertion errors to the ground truth text for the insertion error. We use `fullstop` model to rewrite the punctuation based on such simulations while keeping other transcripts the same as is. We show these results in the latter part in table II.

From the results, when we run the experiments on the GT-Fixed set, our model has comparable results with the one using ASR text. This means punctuation has a crucial influence when connecting audio and text information.

We find that when we use *Remove Short* text, the rouge score has no obvious difference between our method and VG-GPLM method. This means using audio or not has little influence on the quality of the generated summary.

We also find that when we use ASR text and *Add Extra* texts, the rouge score using our method is better than using VG-GPLM method. This means that the audio signal in such summarization task helps resist some errors caused by input audio in the texts and complements the information that cannot be captured in texts. This further implies that such summarization tasks can be finished with their original audio and video information.

Prediction Example We give an example of generating the summary using our method with MIR-MHA fusion and VG-GPLM [3] respectively in Fig. 3, where both methods use ASR texts. The texts in red are the matched texts between the ground truth summary and the summary generated from both methods. Note that we omitted some ASR texts in the figure.

From the example, our method using V+A+T input has more matched words compared with the one using V+T input. To analyze in detail, the proposed method correctly generates the word `expert` and `planting`, but fails to generate the word `spring`. The proposed method also continuously generates the words `get professional advice from an expert on` corresponding to the ground truth, while the method using V+T input generates `get advice from a professional on`. This means audio also provides effective information for a better summary.

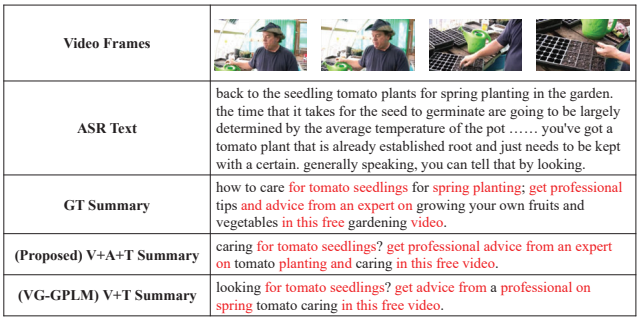


Fig. 3. An example showing the generated summary with ASR text using our method with MIR-MHA fusion and VG-GPLM. Texts in red are the matched texts between the ground truth and the summary generated from both methods.

V. CONCLUSION AND OUTLOOK

In this work, we proposed a new method that generates text summaries based on video and audio inputs. We used whisper to generate text from the input audio. Then, we employed the video transformer, BART, and frozen whisper to represent input information. We designed two kinds of two-stage fusion methods to combine information from different modalities and adopted a BART decoder to obtain the final summary texts. Experiments showed that our method outperforms the previous baselines. Further experiments proved that audio information helps generate a better summary even if the texts contain errors. Next, we want to apply a general-purpose sound feature extractor for audio modality instead of ASR. We are also interested in generating key video frames to make our summary contain both graphics and text.

Acknowledgments This work was supported in part by a project, JPNP20006, commissioned by NEDO, and JST CREST Grant Number JPMJCR22D1, Japan.

REFERENCES

[1] N. Liu, X. Sun, H. Yu, W. Zhang, and G. Xu, "Multistage fusion with forget gate for multimodal summarization in open-domain videos," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 1834–1845.

[2] S. Palaskar, J. Libovickỳ, S. Gella, and F. Metze, "Multimodal abstractive summarization for how2 videos," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6587–6596.

[3] T. Yu, W. Dai, Z. Liu, and P. Fung, "Vision guided generative pre-trained language models for multimodal abstractive summarization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3995–4007.

[4] M. Lewis, Y. Liu, N. Goyal, *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[5] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[6] W. Qi, Y. Yan, Y. Gong, *et al.*, "Prophetnet: Predicting future n-gram for sequence-to-sequencepre-training," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2401–2410.

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] R. Sharma, S. Palaskar, A. W. Black, and F. Metze, "End-to-end speech summarization using restricted self-attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 8072–8076.

[9] R. Sharma, K. Zheng, S. Arora, S. Watanabe, R. Singh, and B. Raj, "Bass: Block-wise adaptation for speech summarization," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2023, 2023, pp. 1454–1458.

[10] Z. Xu, X. Meng, Y. Wang, *et al.*, "Learning summary-worthy visual representation for abstractive summarization in video," *arXiv preprint arXiv:2305.04824*, 2023.

[11] L. Qiao, C. Wu, Y. Liu, H. Peng, D. Yin, and B. Ren, "Grafting pre-trained models for multimodal headline generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2022, pp. 244–253.

[12] J. Yuan, J. Yun, B. Zheng, L. Jiao, and L. Liu, "Mcr: Multilayer cross-fusion with reconstructor for multimodal abstractive summarisation," *IET Computer Vision*, 2023.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023, pp. 28 492–28 518.

[15] Y. Zhang, W. Han, J. Qin, *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[16] Z. Yao, L. Guo, X. Yang, *et al.*, "Zipformer: A faster and better encoder for automatic speech recognition," *arXiv preprint arXiv:2310.11230*, 2023.

[17] P. H. Seo, A. Nagrani, and C. Schmid, "Avformer: Injecting vision into frozen speech models for zero-shot av-asr," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 922–22 931.

[18] M. Han, F. Chen, Z. Ni, *et al.*, "Vilas: Integrating vision and language into automatic speech recognition," *arXiv preprint arXiv:2305.19972*, 2023.

[19] Y. Hu, C. Chen, R. Li, H. Zou, and E. S. Chng, "Mirgan: Refining frame-level modality-invariant representations with adversarial network for audio-visual speech recognition," *arXiv preprint arXiv:2306.10567*, 2023.

[20] J. Hong, M. Kim, J. Choi, and Y. M. Ro, "Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 783–18 794.

[21] J. He, Z. Yang, and T. Toda, "Ed-cec: Improving rare word recognition using asr postprocessing based on error detection and context-aware error correction," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–6.

[22] M. Anwar, B. Shi, V. Goswami, W.-N. Hsu, J. Pino, and C. Wang, "Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation," *arXiv preprint arXiv:2303.00628*, 2023.

[23] J. He, X. Shi, X. Li, and T. Toda, "Mf-aed-aec: Speech emotion recognition by leveraging multimodal fusion, asr error detection, and asr error correction," *arXiv preprint arXiv:2401.13260*, 2024.

[24] R. Sanabria, O. Caglayan, S. Palaskar, *et al.*, "How2: A large-scale dataset for multimodal language understanding," in *NeurIPS*, 2018.

[25] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[26] W. Kay, J. Carreira, K. Simonyan, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[27] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[29] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[32] O. Guhr, A.-K. Schumann, F. Bahrmann, and H. J. Böhme, "Fullstop: Multilingual deep models for punctuation prediction," Jun. 2021.

Prediction-error-based Adaptive SpecAugment for Fine-tuning the Masked Model on Audio Classification Tasks

Xiao Zhang* Haoran Xing* Mingxue Song* Daiki Takeuchi† Noboru Harada† Shoji Makino*
* Waseda University, Japan
E-mail: zhang_x07@toki.waseda.jp, haoranxing@suou.waseda.jp, smx_shue@toki.waseda.jp, s.makino@waseda.jp
† NTT Corporation
E-mail: daiki.takeuchi.ux@hco.ntt.co.jp, noboru.harada.pv@hco.ntt.co.jp

***Abstract*—Spectrogram augmentation (SpecAugment), a data processing method that enlarges datasets without adding new samples, has been widely employed in the fine-tuning process of the masked model for audio classification. In the conventional SpecAugment, the positions of time masking and frequency masking, which directly determine the available information that the model can learn from, are selected randomly. As a result, the random position masking in the conventional SpecAugment may prevent the model from fully utilizing the input information. To this end, we propose a Prediction-error-based Adaptive SpecAugment (PEAS), which incorporates two auxiliary tasks based on reconstruction and then introduces a mask position selector in the fine-tuning process for the masked model. Rather than masking at random positions, the proposed PEAS generates masks on the parts of the spectrograms that are hard to reconstruct for the model. Masking these positions forces the model to learn more generalized audio features, which can effectively prevent the model from learning classification labels by identifying individual special features. Besides, to accelerate the learning of audio features during the early training epochs, we progressively increase the proportion of adaptive masks. Experimental results demonstrate that our proposed PEAS can match or outperform both the conventional method and random masking strategy on the ESC-50 and Speech Commands V2 datasets.**

I. Introduction

The audio classification task aims to enable the machine to automatically recognize and distinguish between different kinds of sounds. Classifying audio segments into predefined categories [1] provides useful information for understanding audio content, which is crucial in several domains, including speech recognition [2], environmental sound classification [3], emotion recognition [4], etc.

Spectrogram augmentation (SpecAugment) [5] is a regularization technique used for audio data processing, aiming to improve the performance of large-scale deep neural network (DNN) models on small datasets [6] [7]. This paper focuses on two groups of SpecAugment strategies [8] [9]: time masking and frequency masking. These techniques simulate scenarios where some data is lost during actual data acquisition, thereby forcing the model to extract meaningful features from incomplete input data.

However, conventional time masking and frequency masking strategies still have limitations due to the predefined random strategy of masked position selection. These positions of masks directly determine the available audio information for feature extraction, which is highly correlated to the performance of audio classification. The random masking strategy prevents the model from fully utilizing the limited data to obtain generalized audio features. Thus, we propose a Prediction-error-based Adaptive SpecAugment (PEAS) that encourages the model to mask where it is most beneficial.

The related work Hard Patches Mining (HPM) [10] in Computer Vision (CV) employs a new training paradigm for Masked Visual Modeling (MVM), which enables the model to identify patches that are difficult to reconstruct, namely, hard patches. It introduces two processes: patch reconstruction and hard patch prediction. They use two models with the same structure: a student model mainly to reconstruct patches and a teacher model mainly to identify the hard patches, respectively. The teacher model can identify the parts of the images that are hard to reconstruct and mask these parts during the pretraining process, thereby bootstrapping the performance of masked image modeling across various downstream tasks. HPM method has proven that hard patches often contain more specific features, and masking these patches can compel the model to discern finer differences between various classes using more generalized image features, thereby enhancing the stability and performance of the model. Since HPM has shown superior performance and considering the similarity between visual images and spectrograms, the HPM idea can be extended to the fine-tuning process of the masked model for audio classification tasks.

Although the HPM method is expected to improve the performance of masked models during fine-tuning, its high training cost and the complex update mechanism for two models limit its application on low-resource terminals. To solve this issue, we employ aforementioned two processes as auxiliary tasks on one masked model during the fine-tuning process inspired by the concept of Multi-Task Learning (MTL) [11], which can decrease the number of parameters and save training time. In addition, block masking has proven effective in SpecAugment, thus we mask consecutive time frames and frequency bins as well. These two auxiliary tasks share the

same parameters, serving for spectrogram reconstruction and hard block prediction, respectively.

To determine which blocks to mask in the spectrograms, we propose a mask position selector. In PEAS strategy, each of the masks has a block shape along the time and frequency axes, respectively. The blocks with a larger mean prediction error are desired to be masked. The proposed PEAS enables the model to identify the hard blocks in the input spectrograms, which are believed to have specific information for audio classification. Thus, masking these parts is expected to force the model to learn more generalized features from the input spectrogram, thus improving the efficacy of the masked model for audio classification tasks.

We apply Self-Supervised Audio Spectrogram Transformer (SSAST) [12] as the backbone of proposed PEAS. SSAST is a masked model that significantly improves performance on various downstream tasks in the filed of audio classification without the need for large amounts of labeled data. Moreover, SSAST also includes generative masked spectrogram patch modeling as one of the joint training tasks, which is the same as one of our proposed auxiliary tasks.

The rest of this paper is organized as follows: Section II provides an overview of related work; Section III details the proposed methodology; Section IV presents the experimental setup and discusses the results; and concludes the paper in Section V.

II. Related Work

A. Time masking and Frequency Masking for SpecAugment

SpecAugment, which is commonly used in various fields [5] [13] [14], can enhance the generalization ability of models at a low cost. Specifically, time masking and frequency masking proposed in [5] can be regarded as an augmentation policy that acts on spectrograms directly. The positions of masks on the spectrogram are important, as they determine the available information for the model to learn. However, SpecAugment selects the starting point from a uniform distribution of the given parameters and generates masks with a random width. This random masking in both consecutive time frames and frequency bins is inefficient for obtaining the input information.

B. Hard Patches Mining for Masked Image Modeling

Hard Patches Mining (HPM) [10] is an approach for masked visual modeling initially deployed in CV pretraining. It has been demonstrated that patches with higher predicted losses are often more discriminative, and thus masking these patches brings a hard situation. HPM introduces two collaboratively trained models. The student model reconstructs the masked patches, while the teacher model predicts the positions of hard patches and generates masks. This iterative process drives the model to continuously generate and tackle difficult tasks, thereby enhancing its overall comprehension of the visual content. The findings in HPM have demonstrated superior performance based on the ViT model, which is a Transformer-based model used for image classification tasks. Therefore, HPM is supposed to be suitable for the masked model with Transformer structure in the audio classification task as well, which takes the spectrograms as input. It is proved that HPM can integrate with existing frameworks and improve performance consistently.

C. Multi-Task Learning

The Multi-Task Learning (MTL) [15] method simultaneously leverages the common information from multiple related tasks to enhance the generalization ability and learning efficiency of the model. By using a shared module for feature extraction, models can efficiently learn the common features, which are then fine-tuned through task-specific layers. The model proposed in [16] utilizes cross-task knowledge for feature extraction from spectrograms and adopts a convolutional neural network with a multi-task learning scheme for different subtasks, which effectively reduces the computational cost of multi-task.

D. Self-Supervised Audio Spectrogram Transformer

Self-Supervised Audio Spectrogram Transformer (SSAST) [12] is a self-supervised pretrained model designed for audio classification tasks. It not only matches but often exceeds the performance of previous supervised models. Also, it exhibits enhanced generalization capabilities. SSAST introduces two auxiliary tasks for joint pretraining, i.e., reconstruction and discrimination of masked patches. It uses both audio and speech datasets for pretraining, which allows it to achieve good performance on both audio classification and speech recognition tasks.

III. Proposed Method

To improve the classification ability of the masked model, we propose a Prediction-error-based Adaptive SpecAugment (PEAS) using a generative masking strategy to calculate the mask positions in the fine-tuning process. In our work, we utilize two auxiliary tasks to calculate the mean prediction errors of each time and frequency blocks, which indicate the positions that are hard to reconstruct, and then generate masks on hard positions via a mask position selector. The overall architecture of the proposed model is shown in Fig. 1.

A. Auxiliary Tasks for MTL Model

The two auxiliary tasks are designed as spectrogram reconstruction and hard block prediction, respectively. Unlike the HPM method, which employs auxiliary tasks on two different models with a same structure, the proposed adaptive masking strategy conducts both tasks on a single model to aid audio classification based on the MTL method.

The first auxiliary task is to reconstruct the original spectrogram. Its process is illustrated as step 1 in Fig. 1. The target of this auxiliary task involves reconstructing the input spectrogram and calculating the reconstruction error to evaluate the reconstruction quality. The value of the reconstruction error can be mapped to the concept of reconstruction difficulty. For each point in spectrograms, a larger reconstruction error

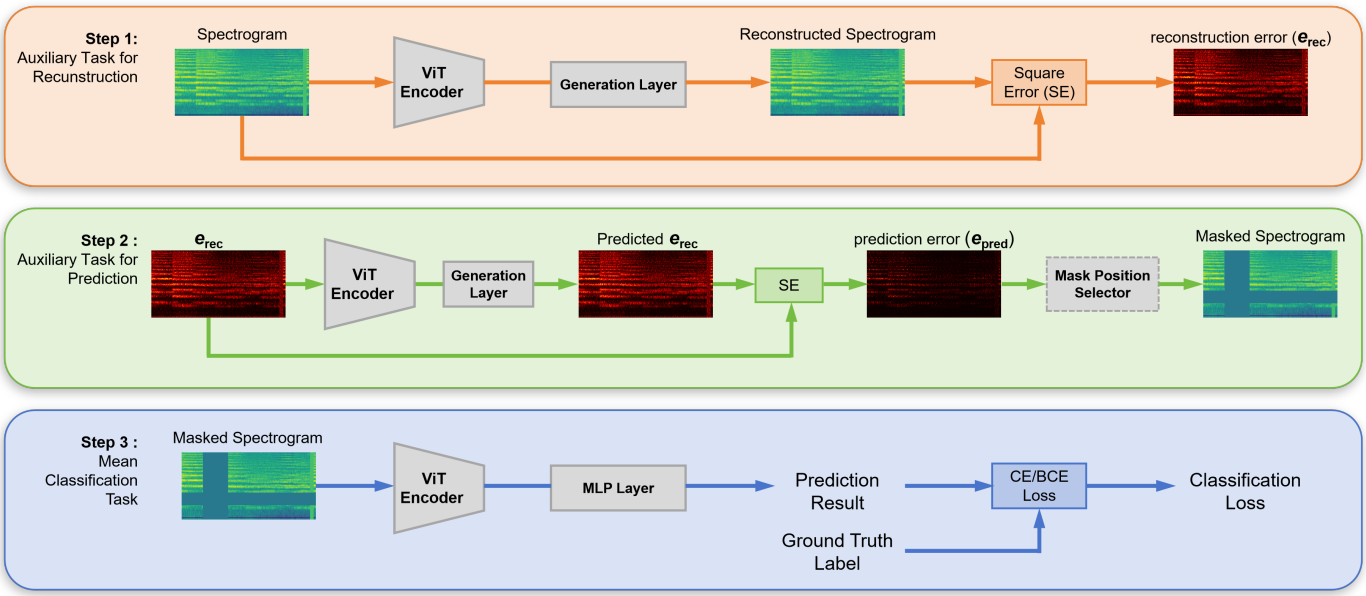


Fig. 1. Overview of the PEAS model process over one training epoch. In Step 1, the reconstruction error e_{rec} is calculated using the original spectrogram. In Step 2, the prediction error e_{pred} is derived from e_{rec}. Finally, Step 3 involves the audio classification task, utilizing a masked spectrogram as input. Each step's input is the output from the preceding step.

indicates a greater prediction deviation, implying increased difficulty in accurate reconstruction.

Assuming that the input spectrograms are 3D tensors with the shape $[B, H, W]$, where B denotes the batch size, H represents the input frame length, and W represents the number of input mel bins. In each iteration, we reconstruct the spectrograms and calculate the point-wise error using the Squared Error (SE) from the input spectrograms.

$$e_{\text{rec}} = (G(E(x)) - x)^2, \tag{1}$$

where x is the original spectrograms, E is the encoder and G is the generation layer. e_{rec} denotes the reconstruction error. As a result, the reconstruction error e_{rec} has the same shape as the input spectrogram x.

The second auxiliary task is to predict the reconstruction difficulty of blocks, as depicted in picture (ii) in Fig. 2 (b). The objective of this task is to enable the model to accurately indicate which positions in the spectrogram have the largest reconstruction error. Subsequently, the model is supposed to generate masks at these positions. We use the point-wise SE function between the true reconstruction error and predicted value of reconstruction error as the output named prediction error for the hard block prediction task:

$$e_{\text{pred}} = (G(E(e_{\text{rec}})) - e_{\text{rec}})^2, \tag{2}$$

where e_{pred} indicates prediction error.

The final classification task and two auxiliary tasks introduced above are employed on a pretrained SSAST-base model provided in [12]. The update of the model mainly relies on the loss of the main classification task. We use the masked spectrograms as input to predict the audio labels and update the whole model using the cross-entropy (CE) loss or binary cross-entropy (BCE) loss, while the auxiliary tasks indirectly bootstrap the training of the classification task by generating masks. This approach prevents the model from using special information to learn classification labels, thus improving the model's learning ability and generalization performance. Even if the model parameters are not directly updated with the loss from the auxiliary tasks, these tasks can still act as a regularizer through the masking strategy, preventing the model from overfitting.

B. Mask Position Selector

With the auxiliary tasks, the mask position selector is designed to identify hard blocks in the spectrograms and generate masks at those positions. As depicted in picture (iii) in Fig. 2 (b), we compute the average error for each block along the time frames and frequency bins separately, using the expected maximum mask width as a unit of calculation with a stride of 1, and then we rank the results. This allows the model to identify the predicted time and frequency blocks with higher average errors, i.e., the blocks that are hard to reconstruct. Generating masks with similar shape to those of the SpecAument method in these blocks can maximize the coverage of difficult information.

IV. Experiment

We validated the proposal for audio classification on the ESC-50 dataset and the Speech Commands V2 (SC-V2) dataset. The SSAST model was employed as the backbone of the audio classification model, along with the aforementioned

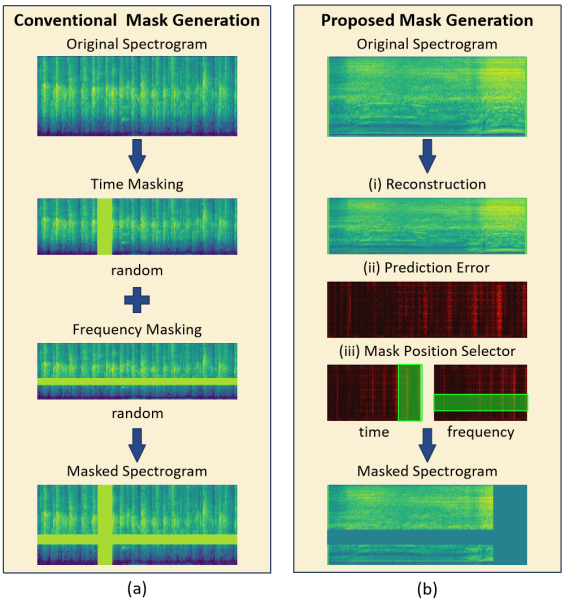


Fig. 2. The comparison between the conventional random masking method and the proposed PEAS masking strategy. In (a), the traditional SpecAugment masking method is shown. In (b): (i) Reconstruction of the original spectrogram, (ii) Prediction of difficulty, where brighter colors represent more challenging parts, and (iii) Illustration of hard block selection using a sliding window.

two auxiliary tasks. For better classification performance, we introduced the progressive strategy into the training process.

To verify the effectiveness of the proposed PEAS in the field of SpecAugment, we compared the performance of the SSAST model with the conventional SpecAugment method and the proposed adaptive SpecAugment method. To ensure the consistency of approaches, we adopted the same experimental settings as those used in the original SSAST paper for parameters as long as possible.

To further validate the superiority of the proposed method, we compared the performance of the proposed masking position selector and the random masking strategy. In addition, we explored the different numbers of masking blocks' impacts. We also generated multiple masks for evaluation to further illustrate the superiority of adaptive masking generation over the random masking strategy.

A. Datasets

We validated the proposed PEAS on the ESC-50 [17] and SC-V2 [18] datasets. The ESC-50 dataset contains 2,000 environmental sound clips (5 seconds each) across 50 classes, with 40 samples per class, and we used 5-fold cross-validation. The SC-V2 dataset includes over 105,000 one-second speech commands from diverse speakers, covering 35 classes, using separate validation and evaluation sets.

TABLE I
EXPERIMENTAL SETUP

Settings	ESC-50	SC-V2
Num of Classes	50	35
Spectrogram Size (T/F)	512 / 128	128 / 128
Initial Learning Rate	1e-4	2.5e-4
Batch Size	10	16
Epochs	25	30
Ground Truth Loss	Cross Entropy	Binary Cross Entropy
Pretraining	Librispeech/AudioSet	Librispeech/AudioSet
Time Masking	96	48
Frequency Masking	24	48
Mix up	0	0.6

B. Training Settings

Table I shows detailed training parameters employed in our experiments. We used batch sizes of 10 and 16 for the ESC-50 and SC-V2 datasets, respectively, and the remaining settings were consistent with those in SSAST. The backbone SSAST model was pretrained on the AudioSet [19] and Librispeech [20] datasets using 400 patch-level masks. All experiments were repeated at least three times, and the average accuracy was reported.

C. Adaptive Masks Settings

To investigate the superiority of proposed PEAS over random masks and explore the impact of different adaptive mask configurations, we set up control groups with different numbers of mask blocks, i.e., T mask blocks are generated in the time domain and F mask blocks are generated in the frequency domain, respectively, denoted as (T, F). The combination of (T, F) was set to (1, 1), (2, 2) and (3, 3) in our experiments, and the performance with different masking strategies were compared in Table III.

During the first 80% of the training epochs, we generated masks with uniformly distributed widths, and the maximum width is defined by predefined parameters. In the final 20% of the training epochs, we generated masks with a fixed-width which is the expected value of the uniformly distributed widths.

In addition, when the number of generated mask blocks was set to more than one, the total number of mask blocks was the sum of the number of random and adaptive mask blocks. We progressively increased the proportion of adaptive mask blocks according to the training period. The number of adaptive mask blocks N_{a} generated is given by

$$n_{\mathrm{a}} = \min\left(n_t, \left(0.1e + 0.9 \times n_t \times H\left(e - 0.8e_t\right)\right)\right), \qquad (3)$$

where n_{a} is the number of adaptive mask blocks, n_t is the total number of random and adaptive mask blocks generated in the time and frequency domains, respectively, e is the current training epoch, e_t is the total number of training epochs, and H is the Heaviside step function.

TABLE II
CLASSIFICATION ACCURACY OF CONVENTIONAL SPECAUGMENT AND THE PROPOSED PEAS

SpecAugment	Num of masks	ESC-50	SC-V2
Conventional	(1, 1)	88.80%	97.62%
PEAS (Proposed)	**(1, 1)**	**91.15%**	**97.61%**

Num of masks means T mask blocks are generated in the time domain and F mask blocks are generated in the frequency domain, respectively, denoted as (T, F).

TABLE III
PEAS ABLATION STUDY: DIFFERENT MASK BLOCK NUMBERS AND MASKING STRATEGIES

SpecAugment	Num of masks	ESC-50	SC-V2
Conventional	(1, 1)	88.80%	97.62%
Random Masking	(3, 3)	89.35%	96.13%
	(2, 2)	89.65%	97.22%
	(1, 1)	90.00%	97.38%
PEAS (Adaptive Masking)	(3, 3)	90.35%	97.20%
	(2, 2)	90.70%	97.23%
	(1, 1)	**91.15%**	**97.61%**

D. Results

Table II shows the comparison between the conventional SpecAugment method and the proposed PEAS under the same setting. The PEAS with the mask position selector showed better performance than the conventional SpecAugment with a random masking strategy on the small datasets. In particular, this method had a 2.35% improvement in accuracy on the ESC-50 dataset and had comparable performance on the SC-V2 dataset.

Table III shows results of the proposed PEAS with different numbers of mask blocks and different masking strategies. According to the results in Table III, the proposed method shows higher accuracy than the random masking strategy on the ESC-50 dataset and the SC-V2 dataset with identical conditions, regardless of the (T, F) configuration. The proposed strategy improved audio classification accuracy on the ESC-50 and SC-V2 datasets by 1.15% and 0.23%, respectively. In addition, several configurations of mask blocks on the time and frequency axes were compared. When (T, F) were set to (1, 1), the model performed best on the ESC-50 and SC-V2 datasets, with accuracy of 91.15% and 97.61%, respectively.

E. Discussion

The above results showed that the proposed PEAS improved the performance of the SSAST model compared to the conventional SpecAugment method, proving that masking hard blocks can help to improve the audio classification performance of the masked model. The accuracy does not improve on the SC-V2 dataset, which may be due to the fact that the model relies on more stable temporal information and semantic features in the learning of speech data. In addition, since speech is a structured signal, the model uses contextual and linguistic rules to infer ambiguous information, so the effect of the positions of the masks will not be as significant as in scene sounds.

As the number of mask blocks increases, the classification accuracy decreases. This may be because masking more blocks containing important information prevents the model from acquiring enough necessary learning information, making the learning process less effective. Additionally, introducing multiple masks can lead to excessive fragmentation of the original data, disrupting the continuity of audio signal features in the time-frequency domain. This makes it difficult for the model to extract effective features during training. However, even with multiple masks, the adaptive masking strategy still outperforms the random masking strategy. This indicates that selectively masking hard blocks allows for more effective guidance of the model in feature learning and generalization.

V. CONCLUSIONS

The conventional SpceAugment method for audio classification tasks generates masks for spectrograms in the time and frequency dimensions randomly. This randomness causes the model to acquire less effective information, resulting in an incomplete understanding of the generalized features of the audio. We propose a Prediction-error-based Adaptive SpecAugment (PEAS) that integrates a mask position selector with the progressive generation strategy for the masked model in the field of audio classification. This approach incorporates auxiliary tasks to identify hard blocks where masks are generated, thereby enforcing the model to learn more generalized features and enhancing the overall performance in the fine-tuning phase. We validated our method on the ESC-50 and SC-V2 datasets, which achieved superior or comparable results to the conventional method, demonstrating its robustness and adaptability in handling different audio data types. Furthermore, ablation studies confirm that PEAS bootstraps masked models for better feature learning and accuracy compared with random mask generation. These findings underscore the potential of adaptive masking to enhance the performance of masked models.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 23H03423.

REFERENCES

[1] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, Oct. 2002.

[2] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? Adversarial examples against automatic speech recognition," *arXiv preprint arXiv:1801.00554*, Jan. 2018.

[3] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. MLSP*, Nov. 2015, pp. 1–6.

[4] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, no. 101894, May 2020.

[5] D. S. Park *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, Dec. 2019.

[6] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proc. ICCV*, May 2021, pp. 9650–9660.

[7] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, Jan. 2017, pp. 131–135.

[8] J. Han, M. Matuszewski, O. Sikorski, H. Sung, and H. Cho, "Randmasking augment: A simple and randomized data augmentation for acoustic scene classification," in *Proc. ICASSP*, June 2023, pp. 1–5.

[9] G. Kim, D. K. Han, and H. Ko, "Specmix: A mixed sample data augmentation method for training with time-frequency domain features," *arXiv preprint arXiv:2108.03020*, Aug. 2021.

[10] H. Wang, K. Song, J. Fan, Y. Wang, J. Xie, and Z. Zhang, "Hard patches mining for masked image modeling," in *Proc. CVPR*, Apr. 2023, pp. 10 375–10 385.

[11] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," *arXiv preprint arXiv:1805.06334*, May 2018.

[12] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-supervised audio spectrogram transformer," in *Proc. AAAI*, vol. 36, Feb. 2022, pp. 10 699–10 709.

[13] D. S. Park *et al.*, "Specaugment on large scale datasets," in *Proc. ICASSP*, May 2020, pp. 6879–6883.

[14] P. Bahar, A. Zeyer, R. Schlüter, and H. Ney, "On using specaugment for end-to-end speech translation," *arXiv preprint arXiv:1911.08876*, Nov. 2019.

[15] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, Sep. 2020.

[16] T. L. Nwe, T. H. Dat, and B. Ma, "Convolutional neural network with multi-task learning scheme for acoustic scene classification," in *Proc. APSIPA*, Dec. 2017, pp. 1347–1350.

[17] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACMMM*, Oct. 2015, pp. 1015–1018.

[18] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, Apr. 2018.

[19] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, Mar. 2017, pp. 776–780.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, Apr. 2015, pp. 5206–5210.

Synchronization of Signals with Sampling Rate Offset and Missing Data Using Dynamic Programming Matching

Hayato Takeuchi*, Takao Kawamura*, Nobutaka Ono* and Shoko Araki†
* Tokyo Metropolitan University, Tokyo, Japan
† NTT Corporation, Japan

***Abstract*—In this paper, we propose a blind synchronization method for signals with sampling rate offset (SRO) and missing data, which occasionally occurs in distributed recording for acoustic scene classification. In our method, the correspondence between short-time frames is first estimated using cross-correlation and dynamic programming (DP) matching. Then, two methods for producing synchronized signals are compared. The first method is based on the overlap-add along the DP path, while the second method uses the DP path only to identify missing data positions and compensates for the SRO with a linear phase model. The performance of these methods is evaluated through experiments. The results are promising, and further applications to acoustic scene classification are expected.**

I. Introduction

Distributed microphone array consists of multiple independent recording devices such as smartphones, voice recorders, and notebook PCs, which do not require wired connections and allow for flexible placement of devices [1]. Furthermore, the use of devices with communication capabilities increases the convenience of the system [2]. In the distributed microphone array, we can use spatial information obtained from these devices for blind source separation [3], [4], source localization [5], speech activity detection [6], and acoustic scene classification [7]–[10].

However, when recording with different devices, recorded signals are not synchronized due to a sampling time offset (STO) and sampling rate offset (SRO). This may occur because the acoustic signal is converted to a digital signal by the AD converters in each device. In acoustic scene classification (ASC), this lack of synchronization degrades processing performance because time difference information plays an important role [7]–[9]. In addition, when there is also missing data in multi-channel signals due to microphone malfunctions, packet loss in network errors, and faulty connections of the microphone cable, ASC performance is significantly degraded because the multi-channel signals with missing data differ from those without missing data [11], [12].

Many blind synchronization methods have been proposed using only recorded signals [13]–[16]. On the other hand, a synchronization method that can cope with various types of missing data is desirable, although a study [17] that performs online SRO estimation based on a probability model of packet loss and a study [18] that detects the presence of missing data using CNNs have been proposed.

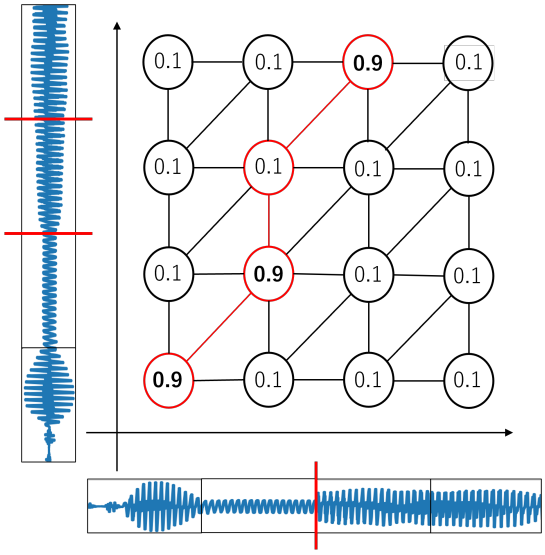


Fig. 1. Conceptual diagram of correspondence estimation based on DP matching. Each node indicates a cross-correlation of short-time frame signals, and the path represents the estimated correspondence. Note that the actual matching is performed on an overlapped-frame sequence.

Synchronization of asynchronous signals is generally performed using frame-by-frame correlations between signals, but under conditions where signals include missing data, it is not possible to simply add and accumulate the correlations between frames. For periodic signals, dynamic addition using their periodicity has been proposed [19], but this method cannot be used directly when the target is arbitrary.

In this study, we propose a new method to synthesis synchronous signals by blindly estimating the frame-by-frame correspondence of asynchronous signals including SROs and missing data by combining frame-by-frame correlation between signals and dynamic programming (DP) matching [20] (see Fig. 1). We confirm the effectiveness of the proposed method through evaluation experiments.

II. Related Works

Several blind SRO estimation methods using only recorded signals without prior information have been proposed [13]–

[16], [21], [22]. In these methods, SRO is estimated using the methods based on Linear phase drift (LPD) model. The methods based on coherence drift (CD) calculate the complex coherence between two consecutive time frames in frequency [21]. In the correlation maximization (CM)-based methods [14], the optimal SRO is searched to maximize the correlation between a pair of signals after resampling. There are also methods based on maximum likelihood (ML) [13]. Meanwhile, online SRO estimation methods assuming a statistical model for partial missing samples in recorded signals have been proposed [17].

III. Problem Settings

We consider the problem of synchronizing two discrete signals, $x[n]$ and $y[n]$. We assume that these signals capture the same audio signals but are not synchronized, specifically due to the time drift caused by sample rate offset (SRO) and partial data loss caused by packet loss or other reasons. The problem here is to produce $\hat{y}[n]$, which is synchronized to $x[n]$, from $y[n]$ by stretching and compressing the time axis.

Even if SRO or missing samples occur, there should be pairs of short-time frames in the two signals that have correlation. Then, we consider synchronizing two signals $x[n], y[n]$ by estimating the correspondence between short-time frame signals $x_i[m](i = 0, \ldots, I-1)$ and $y_j[m](j = 1, \ldots, J-1)$, which are split from $x[n], y[n]$ with frame length L and frame shift S. The variables i and j are short time frame indices, and $m = 0, \ldots, L-1$ is the local discrete-time index in the frame. The relation between the short-time frame signal and the original signal is expressed using an analysis window $w_a[m]$ of length L as follow:

$$x_i[m] = w_a[m]x[iS + m]. \tag{1}$$

The frame signal $y_j[m]$ is calculated in the same way as $x_i[m]$.

IV. Proposed Method

A. Estimation of correspondence between short-time frames

1) DP matching: DP matching [20] is a well-known method for finding the correspondence between two sequences that include insertions, deviations, stretches, and contractions. In this study, we consider the use of DP matching for synchronizing $x_i[m]$ and $y_j[m]$ in the presence of missing or SROs. Specifically, we seek the path (i_k, j_k) that maximizes the cumulative score D calculated using the node score $d_{i,j}$ and the edge score $e_{(i_{k-1},j_{k-1}),(i_k,j_k)}$. When the optimal path contains (i_k, j_k), $x_{i_k}[m]$ and $y_{j_k}[m]$ are in correspondence. The score D is given as

$$D = \sum_{k=1}^{K} d_{i_k,j_k} + e_{(i_{k-1},j_{k-1}),(i_k,j_k)}, \tag{2}$$

where K is the optimal path length, $(i_0, j_0) = (0, 0)$ and $(i_K, j_K) = (I-1, J-1)$, respectively. The k-th correspondence (i_k, j_k) satisfies the following constraints (see Fig. 2).

$$\begin{aligned}(i_k, j_k) \in \{&(i_{k-1}+1, j_{k-1}),\\ &(i_{k-1}, j_{k-1}+1),\\ &(i_{k-1}+1, j_{k-1}+1)\}.\end{aligned} \tag{3}$$

Algorithm 1 Algorithm of Dynamic Programming Matching

Input: d, e, I, J
Output: $(i_k, j_k)_{k=0}^{K}$

$D_{0,0} \leftarrow d_{0,0}$
for $i = 0, \ldots, I-1$ **do**
  for $j = 0, \ldots, J-1$ **do**
    if $(i, j) \neq (0, 0)$ **then**
      $\Lambda \leftarrow \{(i-1, j), (i, j-1), (i-1, j-1)\}$
      $(\hat{i}, \hat{j}) \leftarrow \underset{(i',j') \in \Lambda, i' \geq 0, j' \geq 0}{\arg\max} D_{i',j'} + e_{(i',j'),(i,j)}$
      $D_{i,j} \leftarrow D_{\hat{i},\hat{j}} + d_{i,j} + e_{(\hat{i},\hat{j}),(i,j)}$
      $\mathrm{prev}_{i,j} \leftarrow (\hat{i}, \hat{j})$
    end if
  end for
end for
$(i, j) \leftarrow (I-1, J-1)$
$K \leftarrow 0$
while $(i, j) \neq (0, 0)$ **do**
  $(i, j) \leftarrow \mathrm{prev}_{i,j}$
  $K \leftarrow K + 1$
end while
$(i_K, j_K) \leftarrow (I-1, J-1)$
for $k = K-1, \ldots, 0$ **do**
  $(i_k, j_k) \leftarrow \mathrm{prev}_{i_{k+1},j_{k+1}}$
end for

The algorithm for DP matching is shown in Algorithm 1. The inputs are all node scores d, all edge scores e, and the number of frames I, J for each signal. The output is a path $(i_k, j_k)_{k=0}^{K}$ that maximizes Eq. (2).

2) Node scores: The node score of $x_i[m]$ and $y_j[m]$ is the similarity between their frames. The variables $d_{i,j}$ and $u_{i,j}$ are the maximum value (node score) and the position of maximum value (time difference) of the normalized cross-correlation function $c_{i,j}[\tau]$, respectively.

$$d_{i,j} = \max_{0 \leq |\tau| \leq \frac{S}{2}} \{c_{i,j}[\tau]\}, \tag{4}$$

$$u_{i,j} = \operatorname*{argmax}_{0 \leq |\tau| \leq \frac{S}{2}} \{c_{i,j}[\tau]\}, \tag{5}$$

$$c_{i,j}[\tau] = \frac{\sum_{m=0}^{L-1} x_i[\tau + m] y_j[m]}{\sqrt{\sum_{m=0}^{L-1} x_i^2[m]} \sqrt{\sum_{m=0}^{L-1} y_j^2[m]}}. \tag{6}$$

The reason why the range of $|\tau|$ is restricted to $S/2$ is that if $|\tau|$ exceeds $S/2$, the target frame is considered to correspond to another frame. The frame signal $x_i[m]$ is defined for any integer a as

$$x_i[m] = x_i[m + aL]. \tag{7}$$

3) Edge scores: Edge scores are given as

$$e_{(i_{k-1},j_{k-1}),(i_k,j_k)} = \begin{cases} 0 & \text{if } (i_k, j_k) = (i_{k-1}+1, j_{k-1}), \\ 0 & \text{if } (i_k, j_k) = (i_{k-1}, j_{k-1}+1), \\ p & \text{if } (i_k, j_k) = (i_{k-1}+1, j_{k-1}+1). \end{cases} \tag{8}$$

The score of $(i_k, j_k) = (i_{k-1}+1, j_{k-1}+1)$ as p is based on the prior information that the time axis of the signals $x[n]$ and $y[n]$ proceed in the same way due to almost same sampling

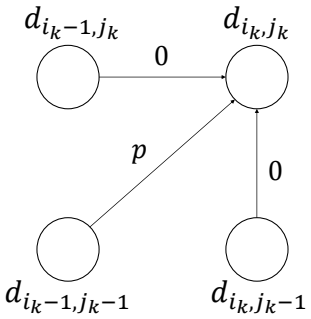


Fig. 2. Edge score (Eq. (8)). d and p indicate edge and node score, respectively.

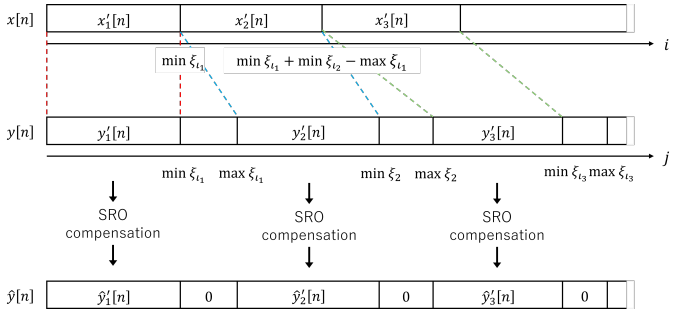


Fig. 3. Framework of missing intervals estimation and linear phase compensation. $x'[n]$ and $y'[n]$ are non-missing intervals of asynchronous signal and reference signal. $\hat{y}'[n]$ is a synchronized signal using $x'[n]$ after SRO compensation.

frequency. Here, a diagram of the edge scores is shown in Fig. 2.

B. Method 1 for synthesis of a synchronized signal: Overlap-add on DP path

We propose two methods to produce the synchronized signal in this paper. The first method is to apply the overlap-add technique along the DP path. In this method, using the correspondence (i_k, j_k) estimated by DP matching, time difference u_{i_k,j_k}, and $x[n]$, The synchronization signal $\hat{y}[n]$ is estimated by

$$\hat{y}[n] = \sum_{j=0}^{J} \tilde{y_j}[n - jS]. \tag{9}$$

Here, $\tilde{y_j}[m]$ is the frame obtained by using the correspondence and time difference, and it is defined as

$$\tilde{y_j}[m] = \begin{cases} 0 & \text{if } \mathcal{I}_j = \mathcal{I}_{j-1}, \\ w_s[m] x_{\mathcal{I}_j}[m + u_{\mathcal{I}_j,j}] & \text{else,} \end{cases} \tag{10}$$

where $w_s[m]$ is a synthesis window for $w_a[m]$. The variable $\mathcal{I}_j$ denotes the indice of the frame that has the largest node score among the frames that correspond to $y_j[m]$, and it is denoted as

$$\mathcal{I}_j = \underset{i_k, k \in \{k | j_k = j\}}{\operatorname{argmax}} \{d_{i_k,j}\}. \tag{11}$$

C. Method 2 for synthesis of a synchronized signal: Missing interval identification and linear phase compensation

The other method to synthesize a synchronization signal is to identify a missing interval using the DP correspondence and performing SRO compensation with linear phase compensation for the non-missing interval (see Fig. 3).

First, the set of indices of frames estimated as missing intervals is given as

$$\psi = \{j \mid \mathcal{I}_j = \mathcal{I}_{j-1}\}. \tag{12}$$

For any i, the set of j that corresponds to the missing intervals and is contained in ψ is given as

$$\xi_i = \{j \mid (\mathcal{I}_j = i) \wedge (j \in \psi)\}. \tag{13}$$

The set of i represented as the missing location in $x[n]$ is denoted as

$$\iota = \{i \mid \xi_i \neq \phi\}, \tag{14}$$

where ϕ indicates the empty set. The set ι is sorted in ascending order, with elements denoted by $\iota_1, \iota_2, \cdots, \iota_q, \cdots, \iota_Q$, where Q is the number of estimated missing intervals and $Q+1$ non-missing intervals. The partial signals $x'_q[n]$ and $y'_q[n]$ are the q-th $(1 \leq q \leq Q+1)$ non-missing interval in $x[n]$ and $y[n]$, respectively. The frame indices range of $x'_q[n]$ and $y'_q[n]$ are given as

$$\sum_{s=1}^{q-1} \min \xi_{\iota_s} - \sum_{s=1}^{q-2} \max \xi_{\iota_s} \leq i < \sum_{s=1}^{q} \min \xi_{\iota_s} - \sum_{s=1}^{q-1} \max \xi_{\iota_s}, \tag{15}$$

$$\max \xi_{\iota_{q-1}} \leq j < \min \xi_{\iota_q}. \tag{16}$$

We apply SRO compensation [13] for $x'_q[n]$ and $y'_q[n]$.

V. Experiments

A. Experimental condition

In this study, we chose speech signals with few silent intervals as a relatively straightforward example for our experiments. This choice allows us to evaluate the fundamental performance of the proposed method. Extending the experiments to more challenging scenarios, such as using environmental sounds for real-world acoustic scene classification (ASC), will be addressed in future work. Specifically, we used five speech signals from Japanese Newspaper Article Sentences (JNAS) [23] of Acoustical Society of Japan (ASJ). The sampling frequency was 16000 Hz. We conducted two cases: (a) using the same utterance, and (b) using convolutive mixtures simulated by [24].

In (a), we made ten mixtures of two speakers using five speech signals of three seconds. For each experiment, we added white noise to the mixture and treated it as a reference signal. For the asynchronous signal, we added white noise, which differed from the reference signal. The signal was resampled to a sampling frequency of 16001.6 Hz by cubic spline interpolation and then missed part of 22500 to 25500

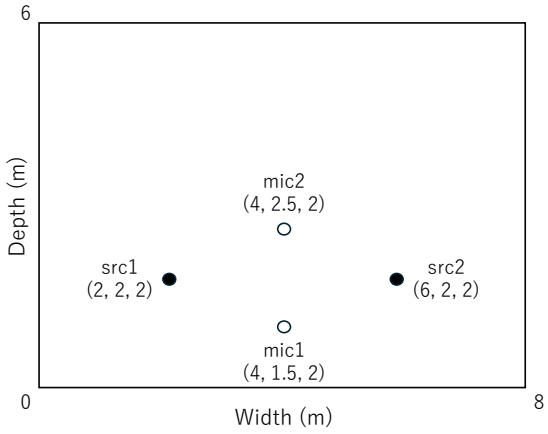


Fig. 4. Arrangement of sound sources and microphones in simulation experiments.The room size is $(8 \times 6 \times 4)$.

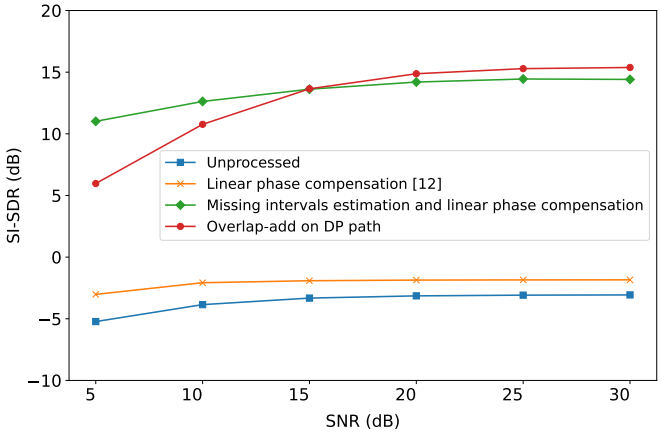


Fig. 5. Average SI-SDR of 10 utterances for each SNR

samples. Each white noise was added so that the signal-to-noise ratio (SNR) to mixtures was $5, 10, 15, 20, 25$, and 30 dB.

In (b), we made ten mixtures of two speakers where we convolved speech signals with an impulse response generated by simulation [24] (see Fig. 4). The room size is $(8 \times 6 \times 4)$, and there are two microphones and two sound sources. The reverberation time is 200 ms, and there is no noise. We regarded the recorded signal of mic1 as a reference signal. We resampled and missed the recorded signal of mic2 as in experiment (a) and regarded it as an asynchronous signal.

To verify the effectiveness of the proposed method, we compared it with linear phase compensation [13]. In the proposed method, frame length $L = 320$ samples, frame shift $S = 80$ samples, window function was Hamming window, and p in (8) was 1.5. In the linear phase compensation method, frame length, frame shift, and window function were the same as the parameters of the proposed method. For the evaluation, we used the scale-invariant signal-to-distortion ratio (SI-SDR) of the asynchronous signal and signal before resampling and missing.

B. Results

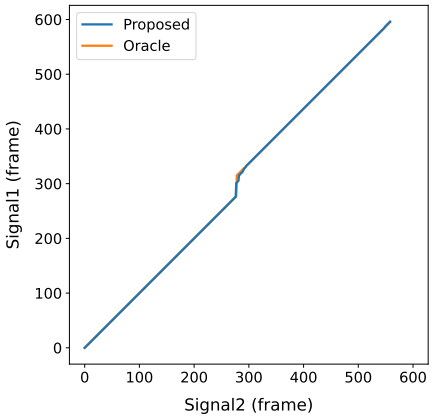


Fig. 6. Correspondence using the proposed method ("proposed") and oracle information ("oracle") for an utterance signal of SNR 5 dB. The signal shown on the horizontal axis include missing interval from 22500 to 25500 samples.

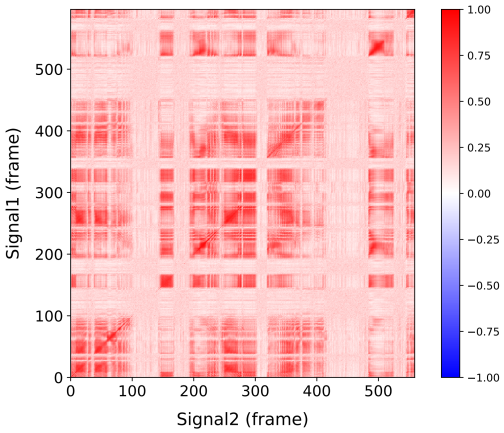


Fig. 7. Node score computed between signals. The signal shown on the horizontal axis include missing interval from 22500 to 25500 samples.

1) Results in (a): Figure 5 shows the average SI-SDR of 10 utterances for each SNR. "Overlap-add on DP path" and "Missing intervals estimation and linear phase compensation" represent the proposed methods described at Section IV-B and Section IV-C, respectively. "Linear phase compensation" is the conventional method [13] and "Unprocessed" is the Unprocessed case. The results show that "Linear phase compensation" performs better than "Unprocessed" due to SRO compensation, but both have a lower. It was considered that these methods could not compensate missing frame correspondence. On the other hand, the proposed methods ("Overlap-add on DP path" and "Missing intervals estimation and linear phase compensation") improved SI-SDR compared to "Unprocessed" and "Linear phase compensation" by more than 10 dB. Thus, we confirmed the effectiveness of the proposed methods. Furthermore, SI-SDR of "Missing intervals estimation and linear phase compensation" was higher than "Overlap-add on DP path" at low SNR. This is thought to be because linear phase compensation robustly worked in the presence of noise. However, in high SNR, "Missing intervals estimation

TABLE I
AVERAGE SI-SDR FOR EACH METHOD.

	Unproc.	Linear Phase comp. [13]	Missing intervals est. and linear phase comp.	Overlap-add on DP path
SI-SDR	-2.66	-2.83	13.41	9.52

and linear phase compensation" was lower than "Overlap-add on DP path". This is thought to be because the missing interval estimation was performed on a frame-by-frame, and the missing part of a sample-by-sample couldn't be considered.

We also showed the example of correspondence and node score for a more detailed analysis of the proposed method. Figures 6 and 7 show the alignment path and node score when the SNR was 5 dB. "Proposed" is the correspondence estimated by DP matching, and "Oracle" is the correspondence based on prior information. As for the estimation of correspondence, Figs. 6 and 7 show that "Proposed" has a path similar to that of "Oracle". It suggests that DP matching considered the local correspondences and allowed proper estimation even when low correlation was caused by noise in some areas.

2) Results in (b): Table I shows the SI-SDR for each method in the simulation experiment. "Unprocessed" corresponds to Unprocessed, "Linear phase comp." corresponds to Linear phase compensation, "Missing intervals est. and linear phase comp." corresponds to estimation and linear phase compensation, and "Overlap-add" corresponds to Overlap-add on DP path.

The table shows that SI-SDR of proposed methods improved more than 10 dB compared with "Unprocessed". We also confirmed that the SI-SDR difference between "Overlap-add on DP path" and "Missing intervals est. and linear phase comp." was about 4 dB. It was considered that the time difference estimation of "Overlap-add on DP path" was affected by the time difference of arrival due to reflected waves of speakers. On the other hand, "Missing intervals est. and linear phase comp." had to estimate a parameter SRO, thus robustly worked.

VI. CONCLUSIONS

In this paper, we proposed a new method for blindly synchronizing signals with SRO and missing data, and describe two methods for synthesizing synchronized signals based on the correspondence between frames estimated by DP matching. The first method involves overlap-adding the corresponding frames along the DP path, while the second method identifies missing intervals and compensates for the SRO using linear phase compensation. Evaluation experiments have demonstrated the effectiveness of these methods. Future work will focus on enhancing the performance of the proposed methods and exploring their application to array signal processing.

ACKNOWLEDGMENT

This work was supported by JST SICORP (JPMJSC2306).

REFERENCES

[1] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 161–164.

[2] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, 2011.

[3] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based mvdr beamformer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5694–5698.

[4] D. Wang, T. Yoshioka, Z. Chen, X. Wang, T. Zhou, and Z. Meng, "Continuous speech separation with ad hoc microphone arrays," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1100–1104.

[5] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1–24,

[6] P. Giannoulis, A. Brutti, M. Matassoni, *et al.*, "Multi-room speech activity detection using a distributed microphone network in domestic environments," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1271–1275.

[7] K. Imoto and N. Ono, "Spatial cepstrum as a spatial feature using a distributed microphone array for acoustic scene analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1335–1343, 2017.

[8] K. Imoto, "Graph cepstrum: Spatial feature extracted from partially connected microphones," *IEICE Transactions on Information and Systems*, vol. E103.D, pp. 631–638, Mar. 2020.

[9] Y. Shiroma, K. Imoto, S. Shiota, N. Ono, and H. Kiya, "Investigation on spatial and frequency-based features for asynchronous acoustic scene analysis," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1161–1166.

[10] T. Kawamura, Y. Kinoshita, N. Ono, and R. Scheibler, "Effectiveness of inter- and intra-subarray spatial features for acoustic scene classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[11] K. Imoto and N. Ono, "Acoustic topic model for scene analysis with intermittently missing observations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 367–382, 2019.

[12] Y. Shiroma, Y. Kinoshita, K. Imoto, S. Shiota, N. Ono, and H. Kiya, "Missing data completion of multi-channel signals using autoencoder for acoustic scene classification," *APSIPA Transactions on Signal and Information Processing*, 2023.

[13] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Processing*, vol. 107, pp. 185–196, 2015.

[14] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 571–582, 2016.

[15] A. Chinaev, P. Thüne, and G. Enzner, "Double-cross-correlation processing for blind sampling-rate and time-offset estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1881–1896, 2021.

[16] J. Schmalenstroeer, J. Heymann, L. Drude, C. Boeddecker, and R. Haeb-Umbach, "Multi-stage coherence drift based sampling rate synchronization for acoustic beamforming," in *Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2017.

[17] A. Chinaev, G. Enzner, T. Gburrek, and J. Schmalenstroeer, "Online estimation of sampling rate offsets in wireless acoustic sensor networks with packet loss," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1110–1114.

[18] T. Raissi, S. Pascual, and M. Omologo, "Sample drop detection for asynchronous devices distributed in space," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 815–819.

[19] K. Sumiyoshi, Y. Wakabayashi, and N. Ono, "Dynamic synchronous averaging for enhancement of periodic signal under sampling frequency variation," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 863–868.

[20] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[21] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[22] K. Yamaoka, N. Ono, and Y. Wakabayashi, "Sampling frequency mismatch estimation by auxiliary-function-based iterative maximization of double-cross-correlation," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1125–1129.

[23] K. Itou, M. Yamamoto, K. Takeda, *et al.*, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[24] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 351–355.

LEAD Dataset: How Can Labels for Sound Event Detection Vary Depending on Annotators?

Naoki Koga*, Yoshiaki Bando† and Keisuke Imoto*
*Doshisha University, Japan, †National Institute of Advanced Industrial Science and Technology, Japan
E-mail: keisuke.imoto@ieee.org

***Abstract*—In this paper, we introduce a LargE-scale Annotator's labels for sound event Detection (LEAD) dataset, which is the dataset used to gain a better understanding of the variation in strong labels in sound event detection (SED). In SED, it is very time-consuming to collect large-scale strong labels, and in most cases, multiple workers divide up the annotations to create a single dataset. In general, strong labels created by multiple annotators have large variations in the type of sound events and temporal onset/offset. Through the annotations of multiple workers, uniquely determining the strong label is quite difficult because the dataset contains sounds that can be mistaken for similar classes and sounds whose temporal onset/offset is difficult to distinguish. If the strong labels of SED vary greatly depending on the annotator, the SED model trained on a dataset created by multiple annotators will be biased. Moreover, if annotators differ between training and evaluation data, there is a risk that the model cannot be evaluated correctly. To investigate the variation in strong labels, we release the LEAD dataset, which provides distinct strong labels for each clip annotated by 20 different annotators. The LEAD dataset allows us to investigate how strong labels vary from annotator to annotator and consider SED models that are robust to the variation of strong labels. The LEAD dataset consists of strong labels assigned to sound clips from TUT Sound Events 2016/2017, TUT Acoustic Scenes 2016, and URBAN-SED. We also analyze variations in the strong labels in the LEAD dataset and provide insights into the variations.**

I. Introduction

Sound event detection (SED) [1], [2] is a fundamental task in environmental sound analysis in which types, onsets, and offsets of sound events included in a sound clip are estimated. SED has various real-world applications, such as residential street surveillance [3] and semantic video search based on audio and visual content [4]. The standard approach to SED is to utilize a deep neural network (DNN) in a strongly supervised manner [5]–[7].

Examples of datasets used in strongly supervised SED are TUT Sound Events (SE) 2016, 2017 [8], [9], URBAN-SED [10], [11], and AudioSet with temporally strong labels [12], [13]. TUT SE 2016 and 2017 contain 47 sound clips in total with 3–5 minutes long and strong labels with 18 and 6 sound event classes, respectively. TUT SE 2016 and 2017 show that annotating strong labels to sound events with relatively large signal lengths is very laborious. URBAN-SED contains 10,000 sound clips with 10 seconds long and 50,000 annotated sound events. To alleviate the labor involved in annotation, URBAN-SED has been constructed synthesizing independently recorded sound events.

In SED, it is very laborious to collect large-scale strong labels, and in most cases, multiple workers divide up the annotations to create a single dataset. Crowdsourcing is one of such methods used to assign a large number of strong labels to sound clips. However, previous works have shown that the strong labels created by multiple annotators tend to have a large variation in the type of sound events and temporal onset/offset [13], [14]. The strongly annotated dataset is used for not only the model training of strongly supervised methods but also the system evaluation of weakly supervised SED. Thus, the variation in the annotation of strong labels is still a serious problem that should be addressed for SED. Although many conventional studies have been conducted to investigate the variation in sound event labels, no previous studies have provided an open dataset for investigating the variation in strong labels.

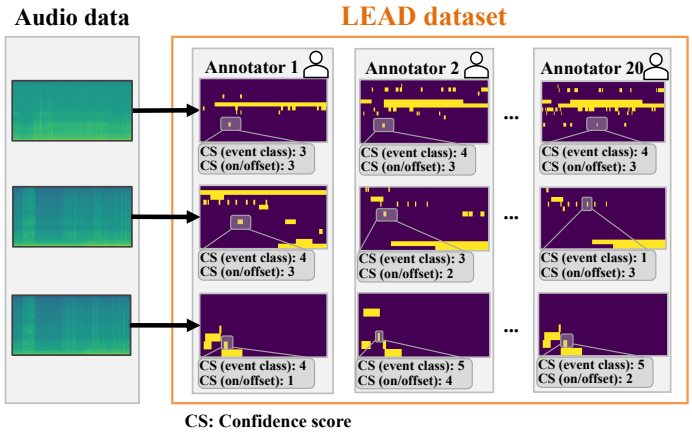


Fig. 1: Overview of LEAD dataset

To investigate the variation in strong labels and construct a robust method against the variation, we introduce the LEAD dataset as shown in Fig. 1. The LEAD dataset contains (i) strong labels that 20 annotators assigned to each sound clip in TUT SE 2016/2017, a part of TUT Acoustic Scenes (AS) 2016, and a part of URBAN-SED, (ii) a confidence score for selecting the sound event class to each sound event instance, and (iii) a confidence score for providing the onset/offset of each sound event instance. We describe the steps and conditions for assigning strong labels and discuss in detail the variations in strong labels and confidence scores in the LEAD dataset. We then evaluate the impact of the temporal variation in the strong labels on the SED performance.

II. Related Works

In some conventional works, the variation in sound event labels assigned to sound clips was investigated. Fonseca et al. [15] have released FSDnoisy18k, which is a dataset for sound event classification containing a small amount of clean data and large-scale real-world noisy data with weak sound event labels. They reported that about 60% of the data showed variation in sound event classes. Hershey et al. [13] have proposed a method in which sound event labels assigned by one annotator are modified by another annotator in sound event classification. The experimental results showed that no agreement was reached between annotators after five attempts at this modifying process. Several studies on human auditory features that were not limited to specific tasks such as sound event classification have also been conducted. Guastavino revealed that semantic features are more important than perceptual features when people perceive sounds [16].

Some works have been conducted on generating more reliable labels using crowdsourced labels. Cartwright et al. [17] have investigated the trade-off between reliability and redundancy in the assignment of sound event labels by crowdsourcing. When the number of annotators exceeds five, the reliability of additional annotators begins to decrease. Three findings were reported: first, when the number of annotators exceeds five, the value of additional annotators begins to decrease; second, the quality of labels improves by 90% when the number of annotators is increased to 16; and third, the degree in variation of labels differs across classes. Méndez Méndez et al. showed that when collecting sound event labels via a crowdsourcing platform, asking annotators to indicate their confidence level in each label assignment improves the quality of labels [18]. Martin-Morató et al. [14] proposed a method to substitute the difficult task of annotating strong labels with easier tasks of annotating multiple weak labels to create more reliable strong labels in SED. Li and Liu [19] proposed a method to select the optimal subset of annotators from a set of annotators when collecting labels by crowdsourcing, which utilizes a combinatorial optimization problem. Bilen et al. [20] proposed an evaluation metric called the polyphonic sound detection score (PSDS) in SED, which is robust to the variation in strong labels among the annotators.

III. LEAD Dataset

We have constructed the LEAD dataset, which can be used to investigate the variation in strong event labels for SED among annotators. The dataset consists of strong labels assigned to sounds in TUT Sound Events 2016/2017 [8], [9], a part of TUT Acoustic Scenes 2016 [8], and a part of URBAN-SED [10], [11]. For TUT Sound Events 2016/2017, we used all sound clips, which includes 2.9 h of sounds (47 files), whereas for TUT Acoustic Scenes 2016, we used a subset that includes all 15 acoustic scenes, which contains 2.6 h of sounds (314 files). For URBAN-SED, we used 0.17 h of sounds (60 files). As URBAN-SED is the sound event dataset synthesized by Scaper [10], the original dataset contains no variation among annotators. Moreover, in URBAN-SED, foreground sounds have a higher SNR than background sounds, and the annotators can more easily assign strong labels than in the case of using other datasets. The LEAD dataset thus enables us to compare annotations for datasets with different characteristics. For each sound clip, 20 annotators assigned sound event labels and gave their onset/offset. For TUT Sound Events 2016/2017 and TUT Acoustic Scenes 2016, we provided 20 candidate event labels for each acoustic scene, and the annotators selected the most suitable event labels from the candidates and then gave onsets/offsets. The annotators also assigned two confidence scores: one is the confidence score for selecting the type of sound event and the other is for providing the onset and offset of the sound event. These confidence scores are attributed to each sound event instance. We set the confidence score on a five-point scale ranging from 1 (very unconfident) to 5 (very confident). The annotators were asked to listen to an entire sound clip at least once during the annotation and to refer to a waveform plot and/or spectrogram if necessary. The annotators were instructed to consider sound events occurring at intervals of more than 1 s as different sound events.

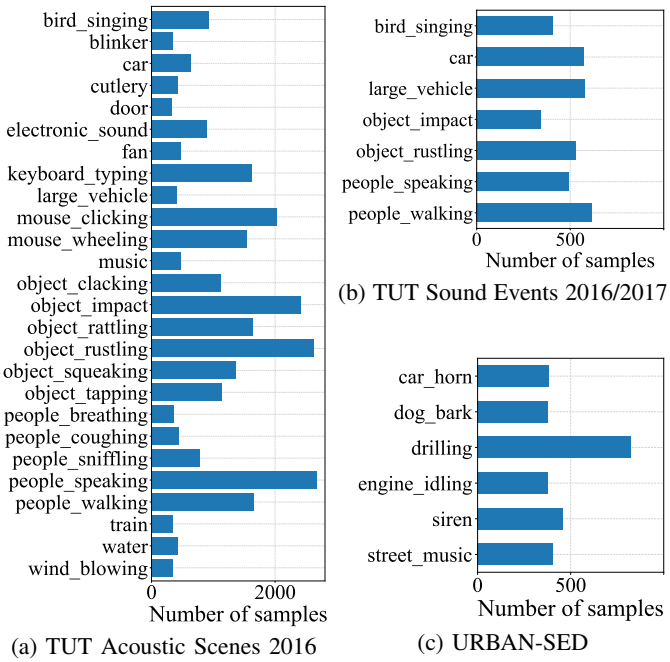


Fig. 2: Histograms of instances of sound events for excerpted classes in the LEAD dataset

The annotations of the LEAD dataset are organized in the same tsv format as those of TUT Sound Events 2016/2017 and TUT Acoustic Scenes 2016, except for the confidence scores (CSs) to the sound event class and the onset/offset of each sound event instance, as follows.

Onset	Offset	Event class	CS (event)	CS (on/offset)
0.619375	12.304331	car	4	3
1.671731	224.798167	bird singing	5	4
3.926606	5.889908	(object) impact	3	3
7.616967	44.334651	people walking	4	4

The LEAD dataset is freely available online[1].

Figure 2 shows the number of sound events for each event class collected through the annotation. Because of space limitations, only sound events with more than 300 instances are shown in Fig. 2. This indicates that the number of sound event instances and TUT AS 2016 is imbalanced, while URBAN-SED is well balanced since URBAN-SED is the synthesized dataset.

IV. Analysis of Variations In Strong Labels

In this section, we discuss possible variations in strong labels for SED by investigating the LEAD dataset. Variations in sound event labels are of two categories: variation of sound event classes and time variation of onset/offset of sound event labels. Then, we investigate confidence scores of annotation for the sound event category and onset/offset of sound event labels.

A. Variation in Sound Event Classes

Label deletion Label deletion indicates that an event label is not assigned to an expected sound event. If label deletion is present in the training data, a trained model might overlook sound events that should be detected. Moreover, if label deletion occurs in the evaluation data, the evaluation results may include improper false positive samples. Figure 3 shows an example of label deletion of the "bird_singing" event in the LEAD dataset. At around 50 seconds, we can see that annotators 008, 011, 015, 028, and 032 do not assign to an expected sound event. One countermeasure against label deletion is to use a majority vote for strong labels among annotators. Another reliable solution is to estimate the annotator's competence using the Multi-Annotator Competence Estimation (MACE) [21] and then take a weighted average of the annotators' opinions.

Label insertion Label insertion means assigning an event label to a sound event that does not exist. If label insertion occurs in the training data, the trained model may detect false positive samples. On the other hand, improper false negative samples are included in the evaluation results if label insertion occurs in the evaluating data. Figure 4 shows example of label insertion in the LEAD dataset, in which annotator 005 gives "water" and "washing_dishes" labels that probably do not exist. One way to address label insertion is to use a majority vote among the annotators.

Label substitution Label substitution refers to a situation in which a sound event class is confused with another similar category. One factor that causes label insertion is the existence of sound events that are auditorily similar but have different meanings. For example, the sound of a ventilation fan and the sound of a car passing by are auditorily similar sound events but have different meanings. Label substitution in the training dataset results in a mistrained model, which is likely to output cross-triggered results. Moreover, label substitution in the evaluation dataset causes a cross-trigger during system evaluation. A cross-trigger during system evaluation is better than a cross-trigger during model training but still disrupts system development in SED. Figures 4 and 5 show examples of label substitution, in which annotators 015 and 016 have confused sound events "water" and "washing_dishes". To avoid label substitution when constructing a dataset, it is important not to include label candidates that are easily confused in the label list.

Hierarchical label substitution Hierarchical label substitution represents label substitution for a sound event ontology [12], [22]. Since the LEAD dataset (including TUT Acoustic Scenes 2016/2017, TUT Sound Events 2016/2017, and URBAN-SED) do not use ontology-based label candidates, hierarchical label substitution does not occur explicitly. However, the label candidates may implicitly have the hierarchical label structure; thus, it is important to carefully select the candidates and determine the procedure for selecting which hierarchical level of sound events should be selected.

Label integration Label integration indicates that when multiple sound events with short duration occur in succession, only one integrated sound event label is annotated. Label integration is not necessarily a serious error. However, if it is inconsistent between the training and evaluation data, the evaluation result contains many inappropriate false positive or false negative samples, making it difficult to correctly evaluate the performance of the trained model. As an example of label integration, in Fig. 5, the sound event labels of annotators 012, 017, and 027 were integrated. As explained in Sec. III, the annotators were instructed to treat sound events that occurred at intervals longer than 1 second separately. Nevertheless, a few annotations include label integration, and it is important to have a common agreement in advance on how independently successive sound events are annotated.

B. Temporal Variation of Strong Labels

The temporal onset and offset of sound events also vary greatly depending on the annotator [13], [14]. For example, in Fig. 3, the unbiased standard deviation of the onset and offset are 2.87 and 1.39 s, respectively. Here, sound events where label integration has occurred are excluded from the calculation of standard deviation. The temporal variation in strong labels is inevitable, but sufficient solutions have not been found in previous studies [13], [14]. This not only affects the model training in SED, but also has a significant impact on the performance evaluation of the trained model. For example, in DCASE 2020–2024 Challenge task 4, a 0.2 s collar is used in event-based evaluation metrics [23], [24]. However, in the annotations shown in Fig. 3, the time variation in the onset and offset far exceeded 0.2 s. This indicates that time variation can affect the evaluation results and the collar setting can be improved. To mitigate the negative impact of time variation,

[1]https://github.com/KeisukeImoto/LEAD_dataset

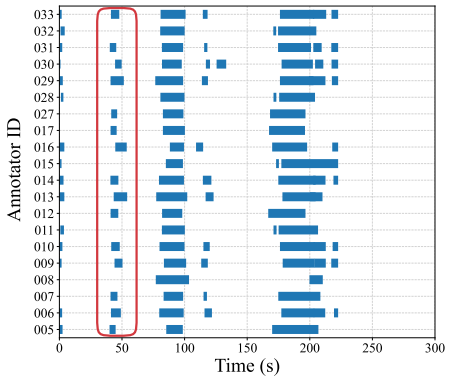


Fig. 3: "bird_singing" in b006.wav

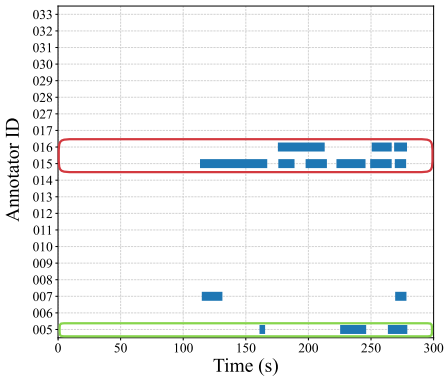


Fig. 4: "water" in b030.wav

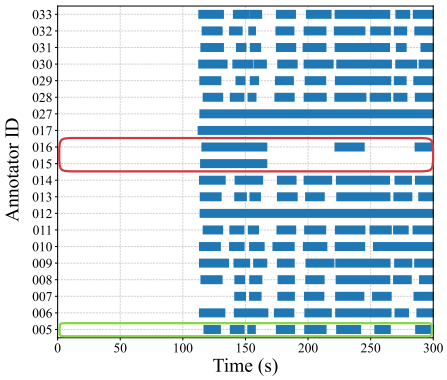


Fig. 5: "washing_dishes" in b030.wav

TABLE I: Average confidence scores on sound event class and onset/offset in LEAD dataset

Dataset	TUT SE 2016	TUT SE 2017	TUT AS 2016	URBAN-SED
Confidence score on sound event class	3.94 ± 0.16	4.00 ± 0.17	3.94 ± 0.12	4.03 ± 0.21
Confidence score on onset/offset	4.04 ± 0.12	4.00 ± 0.19	4.11 ± 0.11	4.13 ± 0.13

one approach is to use an evaluation metric that is robust to the time fluctuation in strong labels, such as PSDS [20].

C. Confidence score for sound event label and onset/offset of sound event

(ii) a confidence score for selecting the sound event class to each sound event instance, and (iii) a confidence score for providing the onset/offset of each sound event instance.

Table I shows the average confidence scores for selecting the sound event class to each sound event instance and for providing the onset/offset of each sound event instance with 95% confidence intervals. As URBAN-SED is the synthesized dataset, we expected that its confidence scores for both event labels and onset/offset would be high, however there is no significant difference from other datasets. We also found that the average length of sound events may not significantly affect the confidence scores. For example, for "people_breathing" (1.24 s), "cupboard" (3.99 s), and "people_shouting" (12.04 s), the confidence scores for the type and onset/offset of sound events were unaffected by the sound event length.

V. Experiments

A. Experimental Conditions

We conducted experiments using the LEAD dataset to evaluate how the variation in strong labels affects the SED performance. In this experiment, we evaluated a pseudo-SED system in which one of the 20 annotations was considered an output of the optimal SED system, and the remaining 19 annotations were considered the ground truth labels. We calculated $20 \times 19 = 380$ pseudo-detection results in the same way and evaluated the SED performance. As evaluation metrics, we calculated the segment-based, event-based [25], and intersection-based [20] micro-F-score. The detailed parameter settings are shown in Table II.

TABLE II: Experimental conditions

Segment length (segment-based F-score)	1.0 s
Collar (event-based F-score)	0.20 s
ρ_{GTC}, ρ_{DTC} (intersection-based F-score) [20]	0.1, 0.1

TABLE III: SED performance in terms of micro-F-score

Dataset	Segment-based micro-F-score	Event-based micro-F-score	Intersection-based micro-F-score
TUT SE 2016	68.21%	8.17%	53.87%
TUT SE 2017	61.79%	5.33%	32.59%
TUT AS 2016	68.57%	32.59%	54.25%
URBAN-SED	72.76%	50.51%	40.52%

TABLE IV: Example of SED performance

Data name	Segment-based micro-F-score	Event-based micro-F-score	Intersection-based micro-F-score
b006.wav	74.09%	7.03%	75.73%
a031.wav	56.69%	23.06%	65.58%

B. SED Performance

Table III shows the SED performance for TUT SE 2016/2017, TUT AS 2016, and URBAN-SED in terms of segment-based, event-based, and intersection-based micro-F-scores. Considering that the experimental result does not include the detection errors of the SED model, the fact that most SED performances are below 70% is an undesirable result. In particular, the event-based F-score is quite low. This result also supports that the idea the collar setting should be reconsidered.

Table IV shows an example of the SED performance for b006.wav and a031.wav in TUT Sound Events 2016. In

b006.wav, the sound of "bird_singing" is dominant, and the temporal variation in onset/offset is very large, leading to a decrease in the SED performance in terms of the event-based metric.

C. Investigation of SED Performance Dependence on Collar Setting

We then investigated the effects of temporal variation in sound event labels and the collar setting on SED performance. In this experiment, we evaluated the SED performance with various collar settings of the event-based metric using "water" sound in a031.wav. Figure 6 shows the annotation result of "water" in a031.wav. The collars were set at intervals of 0.1 s from 0.0 to 5.0 s. As shown in Fig. 7, the experimental results show that the event-based micro-F-score rapidly increases up to around 1.0 s. On the other hand, the unbiased standard deviations of onset and offset in water sounds in a031.wav are 3.14 and 3.73 s, respectively. These results indicate that if the collar is set too short, the error in the performance evaluation result caused by the time variation in the sound event labels will be large. For example, in DCASE 2022 Challenge task 4 [24], the collar is set to 0.2 s as in the experiment in Sec. V-B. These results may be affected by the temporal variation in sound event labels, and the collar setting in SED performance evaluation should be reconsidered.

VI. Conclusion

In this paper, we introduced the LEAD dataset, where 20 annotators assigned strong labels to each sound clip. We then investigated the variation in strong labels using the LEAD dataset, where the variation in strong labels is organized into the variation in sound event classes and the temporal variation in strong labels. We also conducted experiments using the LEAD dataset to evaluate the effect of the variation in strong labels on the SED performance. The experimental results show that the temporal variation in sound event labels seriously affects the SED performance: this could lower the SED performance by 30–40% points. Moreover, the experimental results showed that the short collar on the event-based F-score may affect the SED performance. We believe that the LEAD dataset will accelerate research into the variation in strong labels in SED.

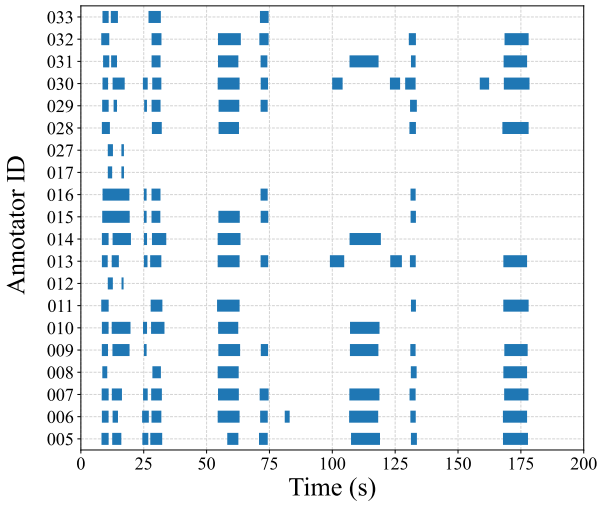


Fig. 6: "Water" labels in a031.wav

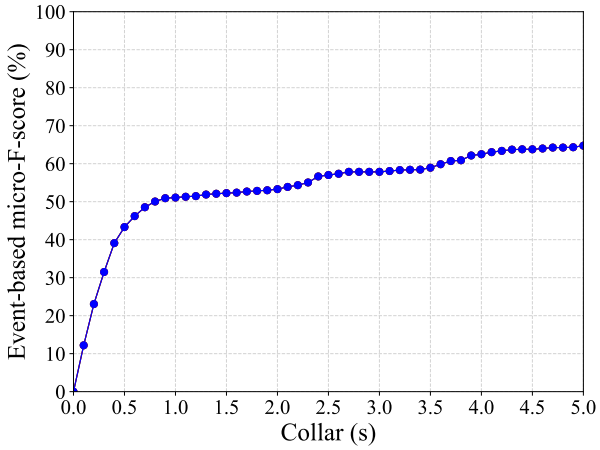


Fig. 7: Event-based micro-F-score performance for various collar ratios

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 22H03639 and 23K16908.

References

[1] A. Mesaros, A. Diment, B. Elizalde, *et al.*, "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.

[2] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[3] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1306–1309, 2005.

[4] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," *Proc. INTERSPEECH*, pp. 1151–1154, 2009.

[5] E. Çakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," *Proc. International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2015.

[6] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1–11, 2017.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 1128–1132, 2016.

[9] A. Mesaros, T. Heittola, A. Diment, *et al.*, "Challenge setup: Tasks, datasets and baseline system," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 85–92, 2017.

[10] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 344–348, 2017.

[11] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," *Proc. ACM International Conference on Multimedia (ACMMM)*, no. 4, pp. 1041–1044, 2014.

[12] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

[13] S. Hershey, D. P. W. Ellis, E. Fonseca, *et al.*, "The benefit of temporally-strong labels in audio event classification," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 366–370, 2021.

[14] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 902–914, 2023.

[15] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, 2019.

[16] C. Guastavino, "The ideal urban soundscape: Investigating the sound quality of french cities," *Acta Acustica united with Acustica*, vol. 92, pp. 945–951, 2006.

[17] M. Cartwright, A. Seals, J. Salamon, *et al.*, "Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations," *ACM Transactions on Computer-Human Interaction*, vol. 1, no. 29, pp. 1–21, 2017.

[18] A. E. Méndez Méndez, M. Cartwright, J. P. Bello, and O. Nov, "Eliciting confidence for improving crowdsourced audio annotations," *Proc. ACM Human-Computer Interaction*, vol. 6, no. 88, pp. 1–25, 2022.

[19] H. Li and Q. Liu, "Cheaper and better: Selecting good workers for crowdsourcing," *Proc. the Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 20–21, 2015.

[20] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61–65, 2020.

[21] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning whom to trust with MACE," *Proc. the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1120–1130, 2013.

[22] T. Nakatani and H. G. Okuno, "Sound ontology for computational auditory scence analysis," *Proc. the Association for the Advancement of Artificial Intelligence (AAAI*, pp. 1004–1010, 1998.

[23] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–6, 2019.

[24] F. Ronchini, R. Serizel, N. Turpault, and S. Cornell, "The impact of non-target events in synthetic soundscapes for sound event detection," *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, pp. 115–119, 2021.

[25] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

SSL-based Chewing and Swallowing Detection Using Multiple Skin-contact Microphones

Toshihiro Tsukagoshi*, Kazuhiro Koiwai*, Masafumi Nishida*, and Masafumi Nishimura*†
*Shizuoka University, Japan
†Aichi Sangyo University, Japan

***Abstract*—The recognition of eating behavior has a crucial role to play in the health management and monitoring of the elderly. Here, we propose an approach to the automatic detection of chewing and swallowing using skin-contact microphones. To improve the accuracy of the detection of chewing and swallowing, we explored the use of both self-supervised learning (SSL) and integrating information from multiple skin-contact microphones at different locations (2 channels below the ears and 2 channels on the throat). We evaluated our approach with training data collected from 27 subjects and test data from 5 subjects who did not contribute to the training data. We achieved a high-accuracy, with a character error rate of 2.88%, using an SSL-based model with a 4-channel summed signal. Furthermore, a frame-level F1-score evaluation showed a high degree of detection accuracy for both chewing and swallowing, achieving F1-scores of 0.954 for chewing and 0.97 for swallowing, with a 0.1 s allowance. These results demonstrate that SSL-based speech recognition models can be effectively applied to the detection of chewing and swallowing, even using a simple Early Fusion approach.**

I. Introduction

Eating behaviors such as chewing and swallowing provide important information for health management. Chewing activates extensive areas of the brain, producing improved cognitive function [1]. Conversely, insufficient chewing or the habit of eating quickly is a cause of lifestyle-related diseases and obesity [2]. In addition, age-related declines in swallowing function affect the nutrition and quality of life of the elderly, increasing the risk of aspiration pneumonia and choking. The importance of regular screening for swallowing function is recognized [3].

Non-invasive methods of recognizing eating behavior include the use of inertial measurement units (IMUs) and skin-contact microphones; further, Nguyen [4] proposed a method of estimating swallowing frequency using long short-term memory (LSTM) for data that are obtained from a neck-worn IMU. Eating behavior sounds have been used as a non-invasive and efficient method of recognizing eating behavior. Skin-contact microphones are robust against external noise, as they record from a microphone unit in direct contact with the skin, and they are able to record biological sounds such as chewing and swallowing more easily than close-talking microphones [5,6]. Nakamura et al. [7] proposed an approach combining recorded audio from multiple skin-contact microphones with an attention-based information integration method [8]. This model uses individual encoders and attention for each sound source, a common decoder, and, later, retrains using a hierarchical attention network, showing high-accuracy.

However, Nakamura et al.'s proposed method integrates information from multiple sources in the model, making it difficult to integrate with existing technologies for speech recognition. By contrast, if the effectiveness of an approach that integrates source data in advance, such as Early Fusion, can be verified, this would be easier to fuse with different types of speech processing technologies. For example, it would be possible to build an integrated system that simultaneously recognizes eating sounds and speech.

In recent years, models utilizing self-supervised learning (SSL) have been recognized as performing well in terms of speech recognition. Wav2vec2.0 [9] forms a type of SSL-based model reported to show high performance in various downstream tasks by pre-training using a large amount of unlabeled speech data and then fine-tuning it with a small amount of labeled data. It can be expected that the high-accuracy recognition of eating behavior is achievable through fine-tuning a pre-trained model that has been trained on a large amount of speech data, with labeled eating behavior data.

Conformer [10] is a model that integrates the global context-capturing ability of the Transformer architecture with the local feature-extraction capability of convolutional neural networks (CNN). Transformer excels in processing long-range dependencies, but has challenges in extracting local information. To overcome this, Conformer adopts a structure that effectively combines self-attention and CNN. Because Conformer also excels in processing complex time-series data as a speech recognition model, it may be suitable for the analysis of eating behavior sounds.

In this study, we propose effective methods for utilizing SSL and data obtained from multiple skin-contact microphones to improve detection accuracy for chewing and swallowing. As

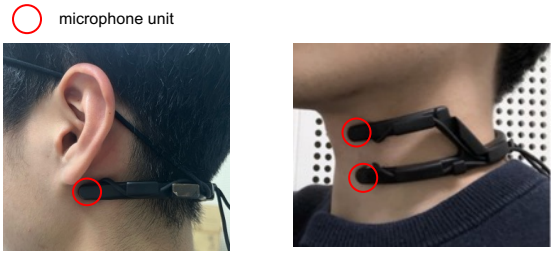


Fig. 1. Multiple skin-contact microphones

methods for utilizing multiple channels, we explored using input from individual channels as independent training data and also implemented a simple Early Fusion approach that sums multiple signals. Additionally, we consider data augmentation techniques. To evaluate the performance of this method, we conduct comparative experiments with the use of an open dataset respecting subjects, using Conformer as a comparison method. By means of these experiments, we seek to clarify the improvements in accuracy of the combination of SSL and multi-channel microphones in eating behavior detection and achieving a more accurate and robust detection system.

II. Approach

A. Data processing

To accurately record eating behavior, we placed skin-contact microphones below the ears on the left and right sides and upper and lower throat areas (Fig. 1). This arrangement was chosen to effectively capture the sounds of the main elements of eating behavior, namely, chewing and swallowing. As chewing involves the movement of both jaws, placing two-channel microphones on the left and right sides below the ears captures not only the differences in jaw movements but also the subtle changes in chewing sounds that are associated with position changes in food in the oral cavity. The microphones on the throat are primarily suitable for collecting swallowing sounds. Because, in swallowing, food flows from upper to lower, we placed two-channel microphones upper and lower throat areas. This arrangement allows us to capture a series of sounds going from the beginning to the end of swallowing, both temporally and spatially. In the recording environment, signals from the 4 channels are input into a single mixer, which produces parallel recorded audio data. In this study, we downsampled from the initial sampling rate of 22 to 16 kHz, which is suitable for inputting into speech recognition models. In addition, due to the characteristics of skin-contact microphones, low-frequency noise may occur due to blood flow in the neck area. To address this, we used a 100 Hz high-pass filter for all 4-channel signals to remove unwanted low-frequency components. We recorded eating behavior sounds using cabbage, crackers, gum, and water. For cabbage and crackers, we recorded both chewing and swallowing sounds.

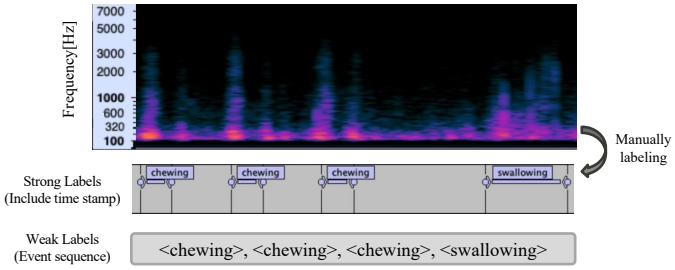


Fig. 2. Example of annotation for throat (upper) microphone

For gum, we recorded chewing sounds and saliva swallowing, and for water, we only recorded swallowing sounds.

We manually annotated strong labels that contained precise timing information for chewing and swallowing in eating behavior audio that was recorded with the 4-channel skin-contact microphones. Because speech recognition model training does not require exact timing information but only event sequences, we generated weak labels through omitting timing information from the strong labels. Fig. 2 presents an example of annotation for the throat (upper) microphone. We divided the audio data into segments of 10 seconds or less to maintain the time-series information of eating behavior and to ensure there was sufficient data for model training. When segmenting the audio, we referenced strong label data to avoid splitting swallowing or chewing events. This process ensures that each segment incorporates a complete sequence of eating behavior that is expected to improve model learning efficiency and recognition accuracy.

B. Models

For the creation of the SSL-based recognition model, we developed a model combining Wav2vec2.0, fully connected (FC) layers, and connectionist temporal classification (CTC). In this model, using Wav2vec2.0, we first extract 1024-dimensional features from the input audio data. The extracted features then pass through three FC layers (each layer's input and output dimensions are uniformly set to 1024), and finally connect to an FC layer with output nodes corresponding to the number of classification classes. The classification consists of three classes: <blank> for blank spaces, and <chewing> and <swallowing> for eating behaviors. We used the CTC algorithm for loss calculation.

For comparison's sake, we prepared a model combining Conformer with CTC and KL divergence loss. In this model, we extract STFT features every 25 ms and apply an 80-dimensional mel-filterbank. The extracted features pass through two CNN layers before being input into the Conformer. The Conformer consists of 12 encoder blocks and 6 decoder blocks, and it finally generates a 6-dimensional output through an FC layer. The tokenizer used is the same as that in Wav2vec2.0. The training employs a weighted combination of CTC loss and KL divergence loss, formulated as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{ctc} + (1 - \alpha) \cdot \mathcal{L}_{KL}$$

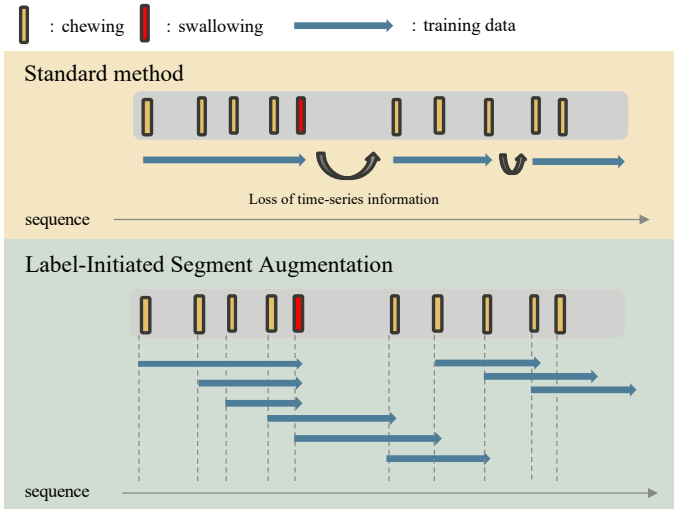


Fig. 3. Label-initiated segment augmentation

where the weighting factor α is set to 0.3. To compute the KL divergence loss, we use the loss between the model's output log likelihoods and a target generated with label smoothing. We utilized the speech recognition library SpeechBrain [11] for model training and evaluation.

C. *Data augmentation*

Unlike speech audio, eating behavior sounds feature a limited variety of labels, tending to have consecutive identical labels. Chewing labels make up the majority. The conventional method of dividing them into 10 s segments may result in a loss of time-series information for eating behavior at both ends of the segments. To address this, we propose a new data augmentation technique called Label-Initiated Segment Augmentation (LISA). LISA generates new training data through adopting each label as a starting point in addition to the existing dataset. Specifically, segments are allowed to extend continuously up to a maximum of 10 seconds. If adding the next label would cause the segment to exceed this 10-second limit, the next label is abandoned. As shown in Fig. 3, this method allows for a comprehensive incorporation of local time-series information included in the audio data to the training data. In addition, we introduced major audio augmentation techniques that are used in speech recognition model training. We applied Speed Perturbation [12], which brings about speed changes through resampling at a different sampling rate from the original waveform, Time Dropout [13], which zeros out random chunks of the original waveform, and Frequency Dropout [14], which zeros out signals in random frequency bands.

III. EXPERIMENTS

A. *Training data*

For training data, we used eating behavior audio drawn from 27 subjects (incorporating 17,539 instances of chewing and 1,329 instances of swallowing). To utilize 4-channel skin-contact microphone audio for learning, we considered two methods. The first is to sum audio signals. Taking the sum of these signals, we can integrate information from the microphones, and it is expected that the recognition model will produce more information. The second method is inputting each channel as independent training data for the model. Although information is not integrated at the input stage, it is expected that model accuracy can improve as data with different information used for learning. To verify the effects of the microphone placement, first, we used a single-channel method from the below-ear (left) microphone and single channel of the throat (upper) microphone as a baseline and then verified the combination using two channels of the below-ear microphone and two channels of the throat microphone. By means of these combinations of methods and placements, we comprehensively evaluated the effectiveness of each approach.

B. *Test data*

For the test data, we used eating behavior audio from five subjects who were not included in the training data (3,915 instances of chewing and 286 instances of swallowing). For the method using multi-channel summed audio for training data, we created and input-summed audio from the corresponding microphones for the test data.

C. *Training settings*

To create an SSL-based eating behavior recognition model, we used Wav2vec2-large-xlsr-53 as the pre-trained model. During the learning process, we updated not only the transformer blocks of Wav2vec2.0 but also the parameters for the feature extractor. For the Conformer model, we calculated the loss by combining CTC and KL divergence loss with a 3:7 weighting ratio. We used four Nvidia V100 SXM2 16 GB GPUs for model training. We used the correct weak labels from each dataset as targets to train each model. We allocated 10% of the training dataset for validation and continued training until the loss converged. We adopted the Adam optimizer in all models. For CTC decoding, we implemented beam search using a beam size of 10.

D. *Evaluation metrics*

The character error rate (CER) is widely used as an evaluation metric for speech recognition models. In this study, we also adopt CER for the evaluation criterion to measure the model's overall performance. In the context of the recognition of eating behavior, CER is suitable for evaluating classification accuracy in chewing and swallowing, in particular with regard to the accuracy of the chewing count. However, as chewing accounts for the majority of eating behavior, it is difficult to properly evaluate the recognition accuracy for swallowing using CER alone.

Thus, in this study, we used the F1-score for eating behavior detection with the use of strong labels as a second evaluation metric. This method adopts the frame-by-frame intermediate

Table 1. Comparison of recognition accuracy when using skin contact microphones (model: Wav2vec2.0)

	Training Data	*Test Data*	*LISA* *(Data Augmentation)*	*CER [%]*	*F1-score* *(Event: Chewing)*	*F1-score* *(Event: Swallowing)*
(1)	Below-ear: Left	Below-ear: Left	-	5.67	0.728	0.925
			○	3.27	0.754	0.933
(2)	Throat: Upper	Throat: Upper	-	13.7	0.661	0.948
			○	7.01	0.704	0.951
(3)	4ch-indipendent signal* (Below-ear, Throat)	Below-ear: Left	○	**2.69**	0.745	0.927
		Throat: Upper	○	3.76	0.720	0.935
(4)	4ch-summed signal (Below-ear, Throat)	4ch-summed signal (Below-ear, Throat)	○	2.88	**0.756**	**0.963**

* Each channel is input as independent training data for the model

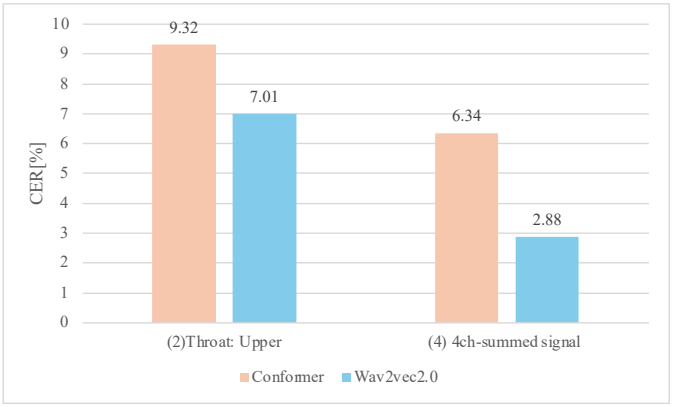


Fig. 4. Performance comparison between Conformer and Wav2vec2.0

output of CTC. Usually, CTC compresses blanks following a frame-by-frame output, but, by analyzing the output before compression, a more detailed frame-by-frame evaluation is possible.

For evaluation, we used manually annotated strong labels as the ground truth for the test data, and we compared them with the CTC frame output. In particular, we judged data to be as correct if the event frame for chewing or swallowing output by CTC was included in the range of the corresponding correct strong label. Furthermore, to explore the potential performance of the model, we set an allowance for the target labels and conducted evaluations at different tolerance widths.

E. Results

Table 1 presents the CER and F1-score for each training dataset with the Wav2vec2.0 model. The results of (1) show that LISA improved the accuracy for the Wav2vec2.0 model. This indicates that comprehensively capturing local time-series information improves the model's overall performance. A comparison of (1) and (2) shows that the detection accuracy of chewing was greater with the below-ear microphone and that of swallowing was higher using the throat microphone. The results of the F1-score showed that method (4) was the most accurate method for both chewing and swallowing. A comparison of the results of (3) and (4) showed that method (3) was the most accurate for CER values and method (4) was the most accurate for the F1-score. Because the CER value does not include time information, method (4) provides a more accurate output for time information. Further, the results

Table 2. F1-score (Wav2vec2.0 + 4ch-summed signal)

Event	*Allowance[s]*			
	0	**0.01**	**0.05**	**0.1**
Chewing	0.756	0.798	0.905	0.954
Swallowing	0.963	0.963	0.970	0.970

showed that accuracy improved with the use of summed signals across 4 channels rather than their use as independent training data for the model. These results indicate that the left and right below-ear and upper and lower throat microphones provide distinct information, interactively providing improved model performance. Fig. 4 compares the performance of Wav2vec2.0 and Conformer. Conformer also showed performance improvement by using a 4-ch summed signal, but it did not exceed the accuracy of Wav2vec2.0. The results demonstrate that SSL-based speech recognition models can be effectively applied to the detection of chewing and swallowing.

Table 2 presents the results of the F1-score for the learning method, applying the 4ch-summed signal for the Wav2vec2.0 model, which provides the highest-accuracy. These results indicate that even without setting a time allowance, the recognition accuracy for swallowing behavior is very high. Further, setting only a 0.01 s allowance brings the F1-score for chewing to about 0.8, and with a 0.1 s allowance, the F1-score exceeds 0.95.

These results indicate that the proposed method can detect both chewing and swallowing with high temporal accuracy. Moreover, it was confirmed that it is possible to achieve high-accuracy detection of swallowing and chewing behaviors, even with a relatively simple Early Fusion approach using summed signals from multiple channels.

IV. CONCLUSION

In this study, we proposed effective methods for the use of SSL and data obtained from multiple skin-contact microphones to improve detection accuracy for chewing and swallowing. Comparative experiments using Wav2vec2.0 showed that an SSL-based speech recognition model is effective in detecting chewing and swallowing and that the relatively simple Early Fusion approach is also capable of highly accurate detection.

We also showed that the newly proposed data augmentation technique LISA helps in the improvement of model performance.

Future works include developing further data augmentation techniques for swallowing-chewing sounds and the evaluation of the model in the context of more diverse eating patterns and environments. Based on the results of this study, we also plan to extend this study to applied work, such as building an integrated system that simultaneously recognizes speech and eating behavior sounds.

ACKNOWLEDGMENTS

These results were obtained through a consignment project (JPNP 20006) by the New Energy and Industrial Technology Development Organization.

This research was partially supported by JSPS Grants-in-Aid for Scientific Research 18H03260 and 21K18305.

REFERENCES

[1] MOMOSE, T., et al. Effect of mastication on regional cerebral blood flow in humans examined by positron-emission tomography with 15O-labelled water and magnetic resonance imaging. *Archives of oral biology*, 1997, 42.1: 57-61.

[2] OTSUKA, Rei, et al. Eating fast leads to obesity: findings based on self-administered questionnaires among middle-aged Japanese men and women. *Journal of epidemiology*, 2006, 16.3: 117-124.

[3] HORIGUCHI, Satoshi, et al. Screening tests in evaluating swallowing function. *JMAJ*, 2011, 54.1:31-34.

[4] NGUYEN, Dzung Tri, et al. SwallowNet: Recurrent neural network detects and characterizes eating patterns. *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2017, 401-406.

[5] DUPONT, Stéphane, et al. Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise. *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, 2004.

[6] HERACLEOUS, Panikos, et al. Fusion of standard and alternative acoustic sensors for robust automatic speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, 4837-4840.

[7] NAKAMURA, Akihiro, et al. Automatic Detection of Chewing and Swallowing Using Attention-Based Fusion. *IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021, 373-375.

[8] LI, Ruizhi, et al. Multi-stream end-to-end speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, 28: 646-655.

[9] BAEVSKI, Alexei, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 2020, 33: 12449-12460.

[10] GULATI, Anmol, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[11] RAVANELLI, Mirco, et al. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

[12] KO, Tom, et al. Audio augmentation for speech recognition. *Interspeech*, 2015, 3586.

[13] SRIVASTAVA, Nitish, et al. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014, 15.1: 1929-1958.

[14] ISLAM, Mobarakol; GLOCKER, Ben. Frequency dropout: Feature-level regularization via randomized filtering. *European Conference on Computer Vision. Cham: Springer Nature Switzerland*, 2022, 281-295.

JOSEPH: PHONETIC-AWARE SPEAKER EMBEDDING FOR FAR-FIELD SPEAKER VERIFICATION

Zezhong Jin*, Youzhi Tu*, and Man-Wai Mak†
* † The Hong Kong Polytechnic University
Department of Electrical and Electronic Engineering
Hong Kong SAR

***Abstract*—Performing speaker verification (SV) at a distance from the sound source is challenging because of the interference of noise and reverberation. In such a situation, incorporating phonetic information into speaker embeddings can help reduce the adverse effects of noise and reverberation. Inspired by this observation, we propose a *J*ointly *o*ptimized *s*peaker-*e*mbedding and *ph*onetic-matching (Joseph) framework to exploit phonetic content for far-field SV. The framework encourages the speaker embeddings to preserve phonetic information by matching the frame-based feature maps of a speaker embedding network with wav2vec's vectors. The intuition is that phonetic information can preserve low-level acoustic dynamics with speaker information and thus partly compensate for the degradation due to noise and reverberation. Results show that the proposed framework outperforms the standard speaker embedding on the VOiCES Challenge 2019 evaluation set and the VoxCeleb1 test set. This indicates that leveraging phonetic information under far-field conditions is effective for learning robust speaker representations.**

***Index Terms*—Far-field speaker verification, multi-task learning, phonetic content, wav2vec**

I. Introduction

Speaker verification (SV) plays an important role in various fields, such as biometric authentication, e-banking, and access control. Traditional SV models rely on statistical models like Gaussian Mixture Models (GMMs) [1] and i-vectors [2] to achieve good performance. With the advance in deep learning, deep neural networks, such as TDNNs [3], ResNets [4], and ECAPA-TDNNs [5], have been prevailing for speaker embedding. Notably, the ECAPA-TDNN has achieved state-of-the-art performance on various datasets, demonstrating its superiority in speaker verification tasks.

Conventional SV systems are usually trained on "clean" utterances and perform well on near-field speech signals. Under far-field conditions, however, due to uncontrollable noise and reverberation, a severe mismatch occurs between the near-field and far-field acoustics, and these systems suffer greatly [6]. Developing an SV system that can address the adverse conditions in the far field is essential.

Researchers attempted to address the far-field challenge by modifying the system architecture, exploring adversarial learning techniques, and leveraging advanced data augmentation strategies. For instance, the author in [7] introduced the channel-interdependence enhanced Res2Net (CE-Res2Net) to aggregate speaker information from multi-scale frame-level representations and achieved performance gains on VOiCES Challenge 2019 data. The authors in [8] used a domain separation network to disentangle and suppress the domain-specific information related to far-field noise and reverberation. In [9], a population-based searching strategy was proposed to optimize the augmentation parameters and greatly boosted far-field SV performance.

On the other hand, studies have shown that text-independent SV systems can be enhanced by incorporating phonetic information into speaker representation learning. In [10], the authors adopted a multi-task learning strategy by combining a phone classifier with a speaker classifier for speaker embedding and obtained superior performance. The authors of [11] investigated the usefulness of phonetic information at the segment and frame levels. They concluded that although phonetic content at the segment (embedding) level is detrimental to SV performance, using phonetic information at the frame level is beneficial. One possible explanation for the performance improvement in [10], [11] is that shared spectral dynamics exist at the lower (frame-level) layers, which are useful for speech and speaker recognition. Enriching content information at the frame-level layers also strengthens the information essential for speaker discrimination.

The aforementioned studies only focus on the contribution of phonetic information to near-field speaker verification tasks, and they rely on phonetic labels when incorporating phonetic information. However, obtaining phonetic labels for SV is challenging because most speech-to-text systems can only output word sequences instead of phone sequences, and the phones are not time-aligned after converting from words to phones.

Inspired by the above observations, we propose a framework that can jointly train a model to perform phonetic matching and speaker verification without any phonetic labels. The framework comprises a phonetic matching component for phonetic information extraction and a speaker identification component to enforce the segment-level layers to produce speaker discriminative vectors. We refer to the framework as **J**ointly **o**ptimized **s**peaker-**e**mbedding and **ph**onetic-matching (**Joseph**). Unlike

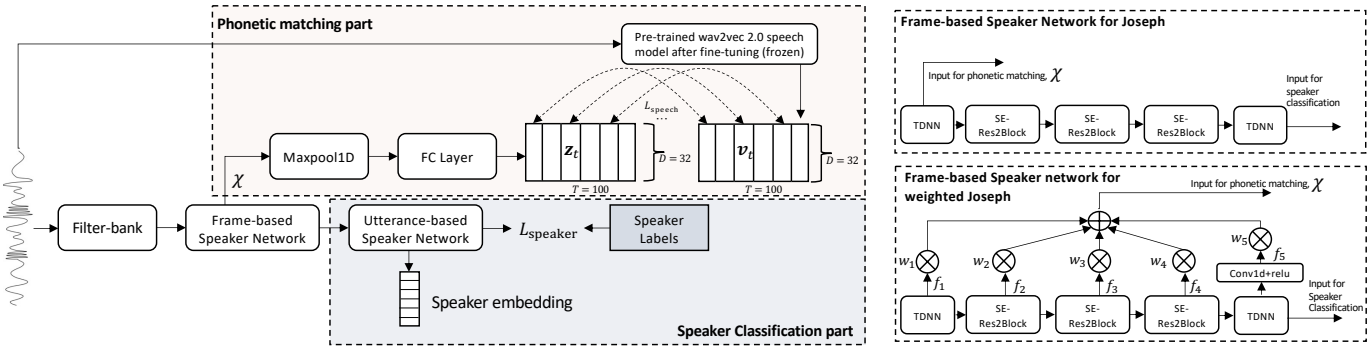


Fig. 1: *Left*: Framework of **J**ointly **o**ptimized **s**peaker-**e**mbedding and **ph**onetic-matching (Joseph) and Weighted Joseph. The utterance-based speaker network in the speaker classification part comprises a pooling layer and a fully connected layer. *Right*: The difference between Weighted Joseph and Joseph lies in the frame-based network. "⊗" represents multiplication and "⊕" represents element-wise addition.

[10], [11], phonetic labels are not required in Joseph. Instead, we use a pre-trained wav2vec 2.0 model to extract the phonetic content in an unsupervised manner. This strategy greatly saves the effort to transcribe the speech files in speaker recognition corpora. The rationale behind Joseph is that although noise and reverberation can blur speaker information in speech signals, the phonetic information extracted from wav2vec 2.0 assists in preserving the underlying acoustic dynamics shared by the speaker identity. Therefore, the degradation due to the far-field conditions can be compensated somewhat. As different frame-level layers contain distinct speaker and phonetic information, Joseph should be able to utilize the phonetic information from different frame-level layers rather than solely relying on the last layer. To address this issue, we propose Weighted Joseph, which utilizes learnable weights to aggregate the outputs of all frame-level layers as the input to the phonetic matching component of Joseph. In this way, the model can autonomously determine which frame-level layer's information is more important. Our main contributions are as follows:

1) We proposed a phonetic-aware framework called Joseph, which improves the robustness of far-field SV by exploiting phonetic information.
2) We incorporated a pre-trained wav2vec 2.0 model (after fine-tuning) in the phonetic matching part, eliminating the need for manually transcribing speaker verification datasets.
3) We improved the Joseph framework by introducing the Weighted Joseph framework, which significantly boosts the system's performance.

The rest of the paper is organized as follows. Section II introduces the wav2vec 2.0 and details the Joseph framework. Section III presents the experimental settings, and Section IV shows the results and analyses. We draw a conclusion in Section V.

II. Methodology

This section introduces the Joseph framework and its two components: phonetic matching and speaker classification.

A. Speech Recognition Model

In Fig. 1, we utilize a wav2vec 2.0 [12] network fine-tuned by CTC loss [13] as the speech model. Wav2vec 2.0 is a self-supervised learning framework that leverages a large amount of unlabeled data to learn speech representations. It takes in a waveform and produces context representations through a stack of CNN layers and Transformer layers. Through contrastive learning, the model is able to extract compact and meaningful speech representations that can be used for downstream speech tasks. Recently, pre-trained wav2vec 2.0 models have gained popularity as a front-end feature extractor in various speech applications.

B. Joseph Framework

As shown in Fig. 1, the phonetic matching part and the speaker classification part share the frame-level layers. The representations outputted from an intermediate frame-level layer are fed into the phonetic matching part. We denote these representations as $\mathcal{X} = \{\mathbf{x}_t \in \mathbb{R}^D; t = 1, \ldots, T\}$, where $\mathbf{x}_t$ is a D-dimensional vector at the t-th frame. For the speaker classification part, the feature maps produced from the last frame-level layer are processed by a pooling layer and a fully connected (FC) layer to derive an utterance-level embedding $\mathbf{e}$. The AAMSoftmax [14] loss is employed as the loss function (L_{speaker} in Fig 1).

For the phonetic matching part, the waveform is fed into the speech model and we obtain a sequence of T frames $\mathcal{V} = \{\mathbf{v}_t \in \mathbb{R}^{\widetilde{D}}; t = 1, \ldots, T\}$, where $\widetilde{D}$ is the dimension of speech vectors. A max-pooling layer is applied to $\mathcal{X}$ to ensure that the resulting $\mathcal{Z} = \{\mathbf{z}_t \in \mathbb{R}^{\widetilde{D}}; t = 1, \ldots, T\}$ has the same length as $\mathcal{V}$. We compute the speech loss as the cosine similarity between $\mathcal{Z}$ and $\mathcal{V}$:

$$L_{\text{speech}} = 1 - \frac{1}{T}\sum_{t=1}^{T}\cos(\mathbf{z}_t, \mathbf{v}_t). \quad (1)$$

Then, we average the L_{speech} across the utterances in a mini-batch. By making $\mathcal{Z}$ close to $\mathcal{V}$, we enable the frame-level lay-

ers of the speaker encoder to preserve useful phonetic information. Because phonetic information contains speaker-dependent acoustic dynamics, maintaining phonetic information at the frame level will also preserve speaker information in the embedding network. As will be demonstrated in Section IV-A, this speaker information preservation helps compensate for the performance degradation caused by far-field environments.

The total loss is defined as follows:

$$L_{\text{total}} = L_{\text{speaker}} + \lambda L_{\text{speech}}, \tag{2}$$

where L_{speaker} is the AAMSoftmax loss defined in [14] and λ is a hyperparameter that controls the contribution of phonetic information. During training, we freeze the parameters of the speech model.

C. Weighted-Joseph Framework

As shown in Fig. 1, the Joseph framework feeds a specific frame level layer's output into the phonetic matching part. However, each frame-level layer contains different phonetic and speaker information. This means that Joseph does not consider the information from other frame-level layers during the phonetic matching process. To address this issue, we have improved the Joseph framework and introduced the Weighted-Joseph (W-Joseph for short) framework.

As shown on the right side in Fig. 1, W-Joseph multiplies the output $\mathcal{F}_i = \{\mathbf{f}_{i,1}, \ldots, \mathbf{f}_{i,T}\}$ of each frame-level layer with the corresponding learnable weights w_i, where $i = \{1, \ldots, 5\}$ represents the index of the frame-level layers. To avoid dimension mismatch during element-wise addition, the frame-level output of each layer undergoes a one-dimensional convolution to unify the dimensions. After element-wise addition, we obtain $\mathcal{X}$, which is then fed into the phonetic matching part. The following equation is defined:

$$\mathcal{X} = \sum_{i=1}^{5} w_i \times \mathcal{F}_i. \tag{3}$$

In this way, Weighted Joseph considers the information from each frame-level layer in the phonetic matching part, and through a set of learnable weights w_i, the model can autonomously determine the significance of each frame-level layer's information.

III. Experimental Setup

A. Datasets and Data Preparation

The training data comprise the VoxCeleb1 development set and the VoxCeleb2 development set [15] [16], which consist of a total of 7,205 speakers. Voice activity detection (VAD) was not used. We followed the data augmentation strategy in Kaldi's recipes [17]. We added noise, music, and babble to the training data using MUSAN [18] and created reverberated speech data based on RIR [19]. For evaluation, we used the VOiCES Challenge 2019 evaluation (VOiCES19-eval) dataset [20]. The Voxceleb1 test Original (Vox-O), which comprises 40 speakers, was also used as the evaluation set.

B. Network Training

We used the standard x-vector [21] and ECAPA-TDNN [5] as our backbones. The channel size of ECAPA-TDNN is 512. The dimension of speaker embeddings is 192 for ECAPA-TDNN and 512 for x-vector, respectively. For the speech model, we used the wav2vec 2.0 model fine-tuned on the LibriSpeech dataset. The output of the wav2vec 2.0 was obtained from the projection layer of the fine-tuned model. For the unweighted Joseph, the frame-level representation from the second lowest-level TDNN of the x-vector network and the lowest-level ECAPA-TDNN were used as the input to the phonetic matching part (see the top-right of Fig. 1). For the Weighted-Joseph, the feature maps outputted by the five frame-level layers of these networks are linearly combined, as shown in Eq. 3 and the bottom-right of Fig. 1.

For ECAPA-TDNN, we extracted 80-dimensional filter-bank (Fbank) features from 16kHz audio signals using a 25ms window with a 10ms frameshift. For the x-vector network, we extracted 40-dimensional Fbank features. Each training segment in the mini-batch has a duration of 2 seconds. The batch size was set to 100 for ECAPA-TDNN and 50 for x-vector, respectively. We used an Adam optimizer with an initial learning rate of 0.001 and employed a step learning rate scheduler. The total number of epochs is 80. For the AAMSoftmax loss function, the margin is 0.2 and the scale is 30.

C. Performance Evaluation

We used a cosine scoring backend in all experiments. When performing evaluation on the Vox-O test set, we followed the setting in [5] to apply the AS-norm [22] on the scores. The performance metrics include equal error rate (EER) and minimum detection cost function (minDCF) with $P_{\text{target}} = 0.01$.

IV. Results and Analyses

We report the performance of Joseph and W-Joseph in this section. The comparison with conventional speaker embeddings is detailed.

A. Main Results

Table I presents the results of the baselines, Joseph, and W-Joseph on the VOiCES19-eval, Vox-O (clean), and Vox-O (noise) datasets. The baselines use the x-vector network or ECAPA-TDNN without phonetic matching. From Table I, it is evident that Joseph and W-Joseph outperform the baselines for both x-vector and ECAPA-TDNN on all datasets. In particular, on VOiCES19-eval, for ECAPA-TDNN, W-Joseph reduces the EER by 14.2% and minDCF by 18.69%. For x-vector, W-Joseph achieves a reduction of 22.99% and 33.44% on EER and minDCF, respectively. These reductions demonstrate the effectiveness of Joseph and W-Joseph in leveraging phonetic information for far-field SV.

To verify that Joseph and W-Joseph can partially compensate for the performance degradation due to adverse conditions under the far-field scenarios, we investigated their performance

TABLE I: Comparison of our methods (Joseph and W-Joseph) with the baselines (without phonetic matching) on the clean and noisy Vox-O datasets and the VOiCES19-eval sets.

Row	System	Speaker Embedding Network	VOiCES19-eval EER (%)	VOiCES19-eval minDCF	Vox-O (clean) EER (%)	Vox-O (clean) minDCF	Vox-O (noise) EER (%)	Vox-O (noise) minDCF
1	Baseline 1	x-vector	7.81	0.598	2.23	0.219	5.82	0.470
2	Joseph (Proposed)	x-vector	6.43	0.433	2.16	0.205	5.14	0.446
3	W-Joseph (Proposed)	x-vector	**6.01**	**0.398**	**2.01**	**0.174**	**4.88**	**0.173**
4	Baseline 2	ECAPA-TDNN	5.79	0.428	1.19	0.165	3.95	0.380
5	Joseph (Proposed)	ECAPA-TDNN	5.13	0.374	1.10	0.136	3.31	0.311
6	W-Joseph (Proposed)	ECAPA-TDNN	**4.97**	**0.349**	**1.05**	**0.143**	**3.05**	**0.252**

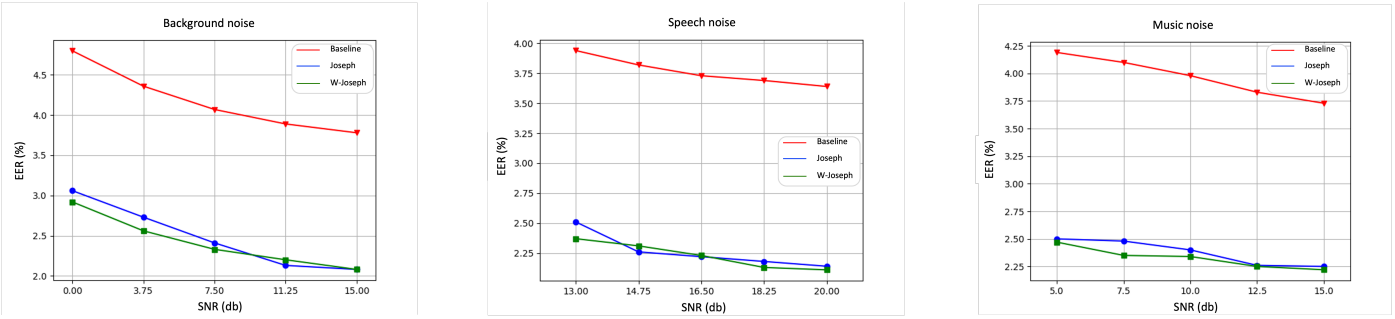


Fig. 2: The impact of different noise types in MUSAN and signal-to-noise (SNR) on the baseline, Joseph, and W-Joseph. The results were based on Vox-O (noise). The speaker embedding network is ECAPA-TDNN.

TABLE II: Comparison of our method (Joseph and W-Joseph) with other methods on the VOiCES19-eval dataset. The best results for each model are highlighted in bold.

System	Speaker Embedding	EER (%)	minDCF
[7]	CE-ResNet	5.72	0.423
[23]	x-vector	8.55	0.552
[24]	x-vector	6.85	0.536
[8]	x-vector	6.51	0.6309
[25]	x-vector	6.42	0.501
[26]	ECAPA-TDNN	5.90	_
Joseph (Proposed)	x-vector	6.43	0.433
W-Joseph (Proposed)	x-vector	**6.01**	**0.398**
Joseph (Proposed)	ECAPA-TDNN	5.13	0.374
W-Joseph (Proposed)	ECAPA-TDNN	**4.97**	**0.349**

TABLE III: Impact of preserving phonetic information at different frame-level layers of an ECAPA-TDNN on Joseph.

Frame-level Layer	Network Block	EER (%)	minDCF
Layer 4	TDNN	5.28	0.392
Layer 3	SE-Res2Block	5.54	0.379
Layer 2	SE-Res2Block	5.33	0.385
Layer 1	SE-Res2Block	5.27	0.390
Layer 0	TDNN	**5.13**	**0.374**

TABLE IV: Impact of preserving phonetic information at different frame-level layers of an x-vector network on Joseph.

Frame-level Layer	Network Block	EER (%)	minDCF
Layer 4	TDNN	6.85	0.483
Layer 3	TDNN	6.74	0.477
Layer 2	TDNN	6.54	0.485
Layer 1	TDNN	**6.43**	**0.433**
Layer 0	TDNN	6.44	0.459

on the clean and noisy Vox-O datasets. The "clean" set refers to the standard Vox-O test data, and the noisy Vox-O set was created by randomly adding noise and reverberation to the standard (clean) Vox-O data, following the data augmentation strategy in Section III-B. From Table I, we observe that Joseph and W-Joseph outperform the baselines on both clean and noisy Vox-O datasets. On the clean Vox-O, our methods achieve a slight improvement over the baselines. This improvement confirms the conclusion in [10], [11] that using phonetic content can benefit text-independent SV. On the noisy Vox-O dataset, all systems exhibit substantial performance degradation compared with the clean counterpart. Nevertheless, phonetic-aware systems (Joseph and W-Joseph) achieve remarkably greater performance gain over the baseline systems. This observation verifies our motivation that incorporating phonetic information into the speaker embedding system can improve SV performance, particularly in far-field

TABLE V: Impact of λ (in eq. 2) on the Joseph framework. The best results are highlighted in bold.

Speaker embedding	λ	EER (%)	minDCF
ECAPA-TDNN	0.001	5.69	0.410
	0.004	5.67	0.418
	0.01	5.59	0.401
	0.1	**5.13**	**0.374**
	0.4	5.85	0.448

TABLE VI: Impact of different speech models on the W-Joseph framework. The results based on the VOiCEs19-eval set.

Speaker embedding	Speech Model	EER (%)	minDCF
ECAPA-TDNN	Wav2vec2	4.97	0.349
	Hubert	5.04	0.343
	WavLM	**4.68**	**0.321**

environments with noise and reverberation.

To further demonstrate the effectiveness of our methods, we compared our method with several recent approaches on the VOiCES19-eval dataset. Table II shows the results. We observed that both Joseph and W-Joseph outperform other methods. W-Joseph achieves better performance compared to Joseph because it focuses on the phonetic information at every frame-level layer and allows the model to autonomously determine which layer's phonetic information is more important through learnable weights.

We have also conducted additional experiments on the Vox-O test set to evaluate the effects of three distinct noise categories in MUSAN: speech, noise, and music, at different SNR levels. Fig. 2 shows the results. From Fig. 2, it is evident that the phonetic aware systems demonstrate superior performance compared to the baseline across different types of noise and SNR levels. Furthermore, When the SNR decreases, the Joseph and W-Joseph system exhibits a relatively smaller decline in performance. This finding indicates the robustness of our methods to noise and reverberation, highlighting its ability to maintain stable performance even in challenging acoustic environments.

B. Ablation Study

We argue that the lower frame-level features contain richer shared spectral dynamic information related to speech and speakers. To support this argument, we conducted ablation studies to show the impact of preserving phonetic information at different frame-level layers of Joseph. Table III and Table IV show the results. In Table III, Layer 0 and Layer 4 correspond to the initial and final TDNN layers of the ECAPA-TDNN, respectively. The remaining three layers correspond to the three SE-Res2Blocks. In Table IV, Layer 0 to Layer 4 correspond to the five TDNN layers of the x-vector network.

Table III shows that the performance improvement of Joseph becomes more prominent when we feed features from lower layers into the phonetic matching part. Specifically, when we present features from the initial TDNN layer (Layer 0) to the phonetic matching part, we obtained the best result with an EER of 5.13%. However, performance gradually drops when we preserve phonetic information at the upper layers (with more abstract representations). This result is reasonable because the lower-level feature maps contain more speaker and content information that is entangled together. By contrast, the representations at the upper layers are more speaker-specific. Therefore, it is preferable to exploit phonetic information at lower layers. In Table IV, feeding the output of the bottom frame-level layers into the phonetic matching part yields better results. However, it is not the lowest layer. Our analysis suggests that the TDNN layers do not possess the same level of ability as the SE-Res2Blocks to filter out speaker-independent information. This is also the reason for using the bottom layer for phonetic information extraction in Section IV-B.

We also investigated the effect of λ in Eq. 2 on Joseph. The results are shown in Table V. We observe that the best performance is achieved when $\lambda = 0.1$, with an EER of 5.13%. As λ increases, the performance of Joseph gradually deteriorates. When λ was set to 0.4, the EER of the Joseph system is higher than that of Baseline 2 in Table I. The above observations suggest that excessive phonetic information can cause the speaker embedding network to focus on the content details and pay less attention to the speaker information, leading to performance degradation.

Table VI demonstrates the impact of different speech models on W-Joseph. It shows that W-Joseph performs the best when the speech model is WavLM.

V. Conclusions

In this paper, we propose two phonetic aware systems (Joseph and W-Joseph) to improve far-field SV performance. By using a pre-trained speech recognition model, we incorporate the phonetic information into the conventional speaker encoders. Also, we eliminate the reliance on transcriptions for the speech recognition task. Experimental results demonstrated that leveraging phonetic information can improve the performance of far-field speaker verification. In the future, we plan to replace the speech model with models suitable for other languages to test their performance on a wider variety of far-field datasets in different languages.

References

[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[2] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[3] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. In *Backpropagation*, pages 35–61. Psychology Press, 2013.

[4] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot. BUT system description to voxceleb speaker recognition challenge 2019. *arXiv preprint arXiv:1910.12592*, 2019.

[5] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. *in Proc. Interspeech*, pages 3830–3834, 2020.

[6] Qin Jin, Tanja Schultz, and Alex Waibel. Far-field speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2023–2032, 2007.

[7] Ling-jun Zhao and Man-Wai Mak. Channel interdependence enhanced speaker embeddings for far-field speaker verification. In *12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021.

[8] Lu Yi and Man-Wai Mak. Adversarial separation and adaptation network for far-field speaker verification. In *Proc. Interspeech*, pages 4298–4302, 2020.

[9] Weiwei Lin and Man-Wai Mak. Robust speaker verification using population-based data augmentation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7642–7646. IEEE, 2022.

[10] Yi Liu, Liang He, Jia Liu, and Michael T Johnson. Speaker embedding extraction with phonetic information. *arXiv preprint arXiv:1804.04862*, 2018.

[11] Shuai Wang, Johan Rohdin, Lukás Burget, Oldrich Plchot, Yanmin Qian, Kai Yu, and Jan Cernockỳ. On the usage of phonetic information for text-independent speaker embedding extraction. In *Proc. Interspeech*, pages 1148–1152, 2019.

[12] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[13] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International conference on Machine Learning*, pages 369–376, 2006.

[14] Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu. Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1652–1656, 2019.

[15] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[16] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[17] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The Kaldi speech recognition toolkit. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[18] David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[19] Marco Jeub, Magnus Schafer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proc. 16th International Conference on Digital Signal Processing*, pages 1–5, 2009.

[20] Mahesh Kumar Nandwana, Julien Van Hout, Mitchell McLaren, Colleen Richey, Aaron Lawson, and Maria Alejandra Barrios. The voices from a distance challenge 2019 evaluation plan. *arXiv preprint arXiv:1902.10828*, 2019.

[21] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018.

[22] Pavel Matejka, Ondrej Novotnỳ, Oldrich Plchot, Lukas Burget, Mireia Diez Sánchez, and Jan Cernockỳ. Analysis of score normalization in multilingual speaker recognition. In *Proc. Interspeech*, pages 1567–1571, 2017.

[23] Sergey Novoselov, Aleksei Gusev, Artem Ivanov, Timur Pekhovsky, Andrey Shulipa, Galina Lavrentyeva, Vladimir Volokhov, and Alexandr Kozlov. STC speaker recognition systems for the VOiCES from a distance challenge. *arXiv preprint arXiv:1904.06093*, 2019.

[24] Zhor Benhafid, Sid Ahmed Selouani, Mohammed Sidi Yakoub, and Abderrahmane Amrouche. Residual time-restricted self-attentive TDNN speaker embedding for noisy and far-field conditions. In *Proc. 2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 471–476, 2022.

[25] Weiwei Lin and Man-Wai Mak. Mixture representation learning for deep speaker embedding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:968–978, 2022.

[26] Sandipana Dowerah, Romain Serizel, Denis Jouvet, Mohammad Mohammadamini, and Driss Matrouf. Joint optimization of diffusion probabilistic-based multichannel speech enhancement with far-field speaker verification. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pages 428–435, 2023.

Vocal Tract Length Perturbation-based Pseudo-Speaker Augmentation Considering Speaker Variability for Speaker Verification

Hengyi Zou* Sayaka Shiota*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: zou-hengyi@ed.tmu.ac.jp

***Abstract*—In this paper, we propose an advanced method for pseudo-speaker augmentation using vocal tract length perturbation (VTLP) for automatic speaker verification (ASV) systems. The state-of-the-art ASV systems based on speaker embeddings require a substantial amount of training data to construct a reliable speaker embedding extractor. Traditional data augmentation methods for ASV typically focus on increasing the corpus size, while ensuring sufficient diversity of distinct speakers is also crucial for improving accuracy. A previous study has reported that VTLP is used as an effective pseudo-speaker generation method, and increasing the number of speakers through VTLP can enhance ASV performance. However, the previous method has demonstrated limitations in the number of pseudo-speakers that can be effectively used, indicating that these methods may not be sufficiently effective. Therefore, this paper proposes increasing the number of pseudo-speakers available for data augmentation by setting the VTLP parameters to ensure diversity for each speaker. The experimental results show that the proposed pseudo-speaker augmentation method can significantly improve the performance of ASV system based on ECAPA-TDNN.**

I. Introduction

In recent years, people's concern for security has been increasing as various systems have been digitized, and research on biometric authentication techniques has become increasingly important [1]. Biometric information used for authentication includes fingerprints, iris, veins, etc. Among them, biometric technology using voice is automatic speaker verification (ASV). ASV is expected to have practical value as a biometric authentication technology because of its low cost of implementation and the growing demand for online biometric authentication.

State-of-the-art (SoTA) ASV systems are based on speaker embeddings using deep learning, represented by x-vector [2] and ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network) [3]. These SoTA ASV systems employ deep learning techniques for extracting speaker embeddings, leveraging advanced methods to achieve efficient and accurate results. It is well known that in order to extract more expressive speaker embeddings, deep neural networks (DNNs) need to be trained on a large amount of training data. In addition to the existing training data, data augmentation methods are used to further enhance the training data and their effectiveness in speaker recognition has been proven [2]–[7].

In conventional ASV systems, the method widely used for data augmentation is to increase the number of utterances per speaker by adding noise. Futhermore, speed perturbation is also employed to further augment the data by varying the speaking speed of the utterances, thereby increasing the variability and robustness of the training dataset [8]. Additionally, the other research has reported that sufficient diversity among distinct speakers is crucial [9], [10]. To address this, vocal tract length perturbation (VTLP) [11], [12], which is derived from vocal tract length normalization (VTLN) [13], [14], has been proposed as a method to enhance speaker variability. By perturbing the frequency axis of voice, it is possible to generate a pseudo-speaker and increase the total number of speakers by adding it to the original training data, therefore it could increase the complexity of speaker embedding. The previous study [9] has reported that ASV performance is improved by selecting only pseudo-speakers with sufficiently large speaker variability, rather than simply adding pseudo-speakers. Nevertheless, this approach reduces the number of pseudo-speakers available for augmentation, thereby not fully achieving the original goal of increasing the number of speakers. To address this limitation, a refined method is proposed that balances the variability of speaker characteristics with the quantity of pseudo-speakers.

In this paper, we build on the prior work of [9] by proposing a method to further increase the number of pseudo-speakers by adjusting VTLP parameters to ensure sufficient variation in speaker characteristics. Our proposed method involves calculating the cosine similarity of speaker embeddings before and after applying VTLP for each speaker. For speakers whose characteristics do not exhibit significant variation, we modify the VTLP parameters to achieve the necessary changes. Speakers who exhibit sufficient variation are then added to the speaker pool as pseudo-speakers. This iterative adjustment process ensures that the generated pseudo-speakers introduce sufficient variability, thereby enhancing the effectiveness of our speaker verification system. In the experiments, an ASV system using ECAPA-TDNN is constructed, and evaluated it's performance when various data augmentation were applied to the training data. The experimental results confirm the significance of incorporating a diverse range of speaker characteristics into the training data. The system's performance

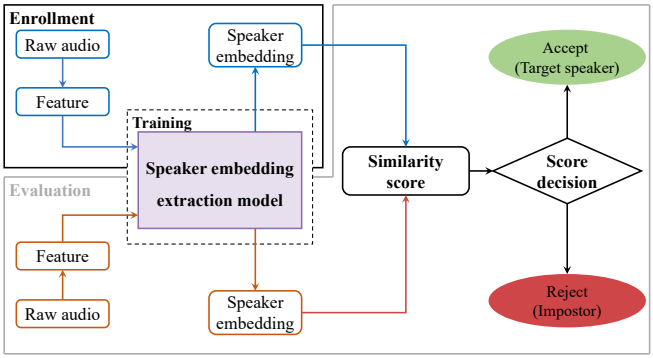


Fig. 1: Flow of speaker embedding-based ASV

improved most notably with the data augmentation method that maximized the dataset size while accounting for speaker variability. This was achieved by increasing the number of utterances per speaker through adding noise and expanding the number of speakers by VTLP. The importance of including a variety of speaker characteristic in the training data was confirmed.

The rest of the paper is organized as follows. Section 2 describes a genral ASV system based in speaker embedding. Section 3 introduced 2 kinds of augmentation. Section 4 will show how to select pseudo-speaker by variability of speaker characteristic. Section 5 gives details of experiment and results, and section 6 concludes this paper.

II. Speaker verification based on speaker embedding

Speaker recognition can be classified into the following two categories. The first is speaker identification in a multi-level classification task that identifies the most likely speaker from several registered speakers. Another is ASV in the binary classification task, which determines whether the input speech belongs to the registered speaker himself or not. Figure 1 illustrates the flow of the ASV system based on a recent speaker embedding technique. The ASV system comprises enrollment, evaluation, and training units for the speaker embedding extraction model, utilized in both enrollment and evaluation. Initially, the speaker embedding extraction model is trained using a large dataset to recognize a massive number of different speakers. The intermediate output of this model, which can accurately identify speakers, is used as the speaker embedding for both enrollment and evaluation. In the enrollment phase, the utterance of the enrolled speaker is converted into features, which are input to the model to extract the speaker embedding. The same procedure applies to the evaluation phase for the test speaker. The similarity between the extracted embeddings is calculated using cosine similarity or other metrics, and compared with a threshold to determine speaker identity.

III. Data augmentation for speaker verification

To enhance the performance of the speaker embedding extraction model, a substantial amount of training data is required. Besides using existing databases, data augmentation is widely employed to simulate additional training data [15]. In ASV, data augmentation typically involves increasing the number of utterances through methods such as adding noise and music. Moreover, augmenting the number of speakers using techniques like VTLP has been shown to improve accuracy [9], [16].

A. Speech augmentation using simulated noises

This method expands the number of utterances by adding noise and music to speech data. Meanwhile, it's often used in deep neural network training, helps stabilize training and improve robustness by simulating a noisy environment [17]. Augmenting the number of utterances in ASV can stabilize the training of embedding and extraction models for speaker identification.

B. Pseudo-speaker generation by VTLP

VTLN is a technique from speech identification to normalize the acoustic feature variability across different speakers. In this method, pseudo-speakers are generated by transforming the frequency axis of the logarithmic amplitude spectrum of speech. The normalization frequency of the original voice is ω, frequency after perturbation is ω', if the frequency expansion coefficient is α, it is represented by the formula (1).

$$\omega' = \omega + 2\arctan\frac{\alpha\sin(\omega)}{1-\alpha\cos(\omega)} \tag{1}$$

In addition, the level of pitch change could be adjusted by changing the parameter of the frequency expansion coefficient. In speech recognition, VTLN is used to normalize the vocal tract length to remove the effect of speaker characteristic, however in our research, this formula is used as VTLP to add variation to the vocal tract length. So far, it has been reported that VTLP can be used for data augmentation in ASV by adding vocal length perturbation [11], [12]. In these methods, the number of speakers is augmented by using the voice of a new speaker, which is processed to a different pitch from the original voice by applying VTLP to the training data, to increase the number of speakers in a pseudo form.

IV. Pseudo-Speaker generation considering speaker variability

A. Speaker verification by pseudo-speaker generation using VTLP

In ASV using data augmentation with VTLP described in the previous chapter, two methods were proposed: one consistently applies a frequency expansion coefficient to increase speakers, while the other selects pseudo-speakers with significant variability in speaker characteristics. The first method has the advantage of simply increasing the number of speakers. However, incorporating pseudo-speakers with small variability in speaker characteristics could lead to problem, as it can confuse the network and make it difficult to distinguish between the original speaker and the pseudo-speaker. On the other hand, the method that selects only pseudo-speakers with large speaker

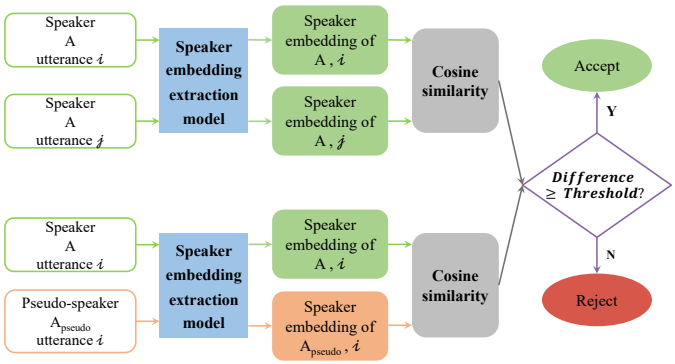


Fig. 2: Flow of calculating the amount of speaker variability

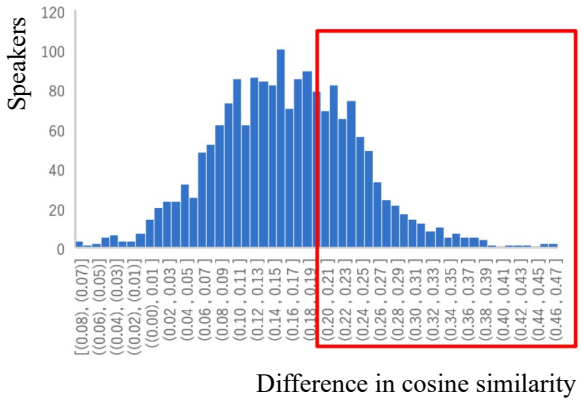


Fig. 3: Distribution of variation in speakability by VTLP

variability has better performance in ASV. This effect is likely due to the difficulty the network faces when distinguishing between the original speaker and a pseudo-speaker with small variability in characteristics. When the network assumes that a speaker with small variability is distinct from the original speaker, making accurate distinctions becomes challenging, which reduces the network's accuracy. Selecting only voices with substantial variability in speaker characteristics as pseudo-speaker voices can stabilize the training of the speaker embedding extraction network by excluding pseudo-speakers with minimal variability. Nevertheless, this approach also reduces the number of pseudo-speakers available for inclusion in the training data, which counteracts the primary objective of data augmentation. Consequently, the number of pseudo-speakers that can be added to the training data is reduced, and the cost is that the number of speakers cannot be increased sufficiently for the main purpose of data augmentation.

B. Data selection by variability of speaker characteristic

Pseudo-speaker generation involves applying VTLP to all speech data and assigning new speaker labels to the processed speech. The variability of speaker characteristics is calculated using cosine similarity between original and pseudo-speaker embeddings. Only those pseudo-speakers with variability above a certain threshold are added to the training data. Figure 2 outlines the process of calculating speaker variability. The cosine similarity is computed between the speaker embeddings of Speaker A's utterances i and j, and between the embeddings of Speaker A's utterance i and the pseudo-speaker A_{pseudo}'s utterance i. The difference in cosine similarity is then assessed by comparing the embeddings of Speaker A's utterances with each other and those of Speaker A's utterance with the pseudo-speaker A_{pseudo}'s utterance. This difference represents the variability of speaker characteristics. Since the cosine similarity between Speaker A's utterances can vary, multiple utterances from Speaker A will be used as reference utterances. Multiple cosine similarities will be calculated and averaged to determine the overall cosine similarity between Speaker A's utterances. In contrast, to obtain the similarity between the pseudo-speaker A_{pseudo}'s speech and Speaker A's speech, the cosine similarity is calculated between the embeddings of Speaker A's speech and all pseudo-speaker A_{pseudo}'s speech generated by applying VTLP to Speaker A's speech. If the calculated variability of speaker characteristics exceeds a predefined threshold, the pseudo-speaker A_{pseudo}'s speech is added to the training data. If it is below the threshold, it is not used as the pseudo-speaker. Figure 3 shows the distribution of the degree of changing speaker characteristics before and after adapting single VTLP factor. According to this figure, 0.2 is chosen as the threshold for the experiment.

C. Pseudo-speaker selection by variability of speaker characteristic

This paper proposes generating new pseudo-speakers with modified VTLP parameters for those initially deemed to have low variability. A previous study reported that the number of pseudo-speakers was reduced to about 40% through data selection to improve ASV performance. To address this limitation, we propose increasing the data for speaker augmentation by applying VTLP with modified parameters to the original speakers of excluded pseudo-speakers. Figure 4 provides an overview of the proposed method. Initially, the procedure in Fig. 2 is used to assess the variability of pseudo-speakers. For those with low variability, VTLP with modified parameters is applied to generate new pseudo-speakers with increased variability. This enhanced variability improves the effectiveness of data augmentation. The process in Fig. 2 is repeated for these new pseudo-speakers to determine their inclusion in the training data. By adjusting parameter variations and iteratively applying the method shown in Fig. 4, a greater number of pseudo-speakers can be automatically selected. However, as the absolute value of the VTLP parameter increases, speech distortion also increases. When this distortion reaches a threshold beyond which the speech can no longer be recognized as intelligible, the proposed data augmentation method will be halted.

V. EXPERIMENT

In order to verify the effectiveness of the proposed method, data augmentation for the number of speeches and speakers

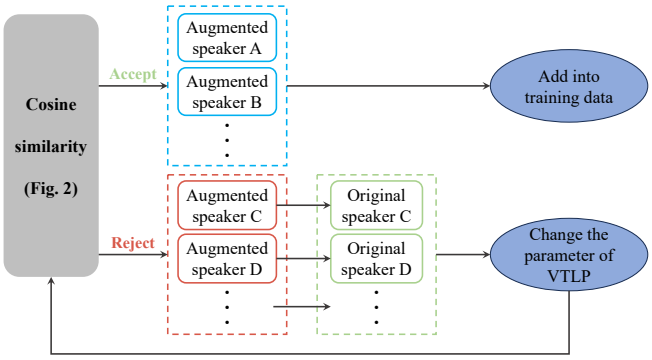


Fig. 4: Flow of Parameter adjustment considering speaker variability

was applied under various conditions, and ASV experiments were conducted.

A. Database

JTubeSpeech-ASV [18] was used to train and evaluate the ASV system, it is a speech corpus consisting of 900 hours of speech data automatically collected from YouTube videos. It consists of audio data of only one speaker appearing in one video, and one channel is considered as one speaker. This dataset mainly includes Japanese audio, and it also includes languages such as English, Chinese, and Korean. Among the subset for ASV of JTubeSpeech-ASV used for learning and evaluating the speaker embedding extraction model, the training dataset consists of 107,271 utterances from 1792 speakers. The test dataset for ASV is also from the JTubeSpeech-ASV dataset. The test dataset consists of 20,976 speech utterances from 92 speakers and amounting to 20,976 trials. Of the 20,976 sets, 228 are trials within the same speaker, and 20,748 are trials between different speakers. The sampling frequency is 16 kHz. The MUSAN database [19] is used for adding noise and music. The MUSAN database contains 42 hours of music of various genres, 60 hours of conversations in 12 languages, and over 900 types of noise. In the experiments, only the noise subset of the MUSAN database is used. The subset that this paper used contains a total of about 6 hours of noise, including mechanical and environmental noises.

B. Experimental setup

In the experiments, an ASV system based on speaker embedding is constructed. The pre-trained model ECAPA-TDNN is used as the speaker embedding extraction model, and it is known to have higher performance. A logarithmic mel filter bank was used as the input features of the model. The speaker embedding was extracted as a 512-dimensional vector using the output of the middle layer of the ECAPA-TDNN. More detailed hyper-parameters are shown in Table 1.

To verify the relationship between ASV performance and speaker characteristic variability based on different parameter settings, three methods were compared: the conventional method applying VTLP with a single parameter to all speakers (All), the conventional method selecting augmented speakers

TABLE I: Detailed settings of Hyper-parameter

	Parameter
Speaker embedding extraction model	ECAPA-TDNN
Input feature	Fbank (512 dimensions)
Iterations	Max: 500,000
Batch size	16
Optimization method	AdamW
Loss function	AAM-Softmax
Sampling frequency	16kHz

TABLE II: Number of speakers and total duration of speech for each data augmentation condition

Augmentation condition	**Num of speakers**	**Data duration (hrs)**	**Average data duration per speaker(hrs)**
(A) No augmentation	1792	498	0.278
(B) Noise	1792	1495	0.834
(C) VTLP (All)	5376	1494	0.278
(D) VTLP (Select)	2868	794	0.277
(E) VTLP (Proposed)	4086	1128	0.276
(F) VTLP (Proposed+Noise)	4086	5380	1.317

based on characteristic variability (Select), and the proposed method applying VTLP with various parameters to the speaker characteristics. In the experiments, three proposals have been prepared for augmenting speakers. The comparison conditions are shown below, and the number of data sets for training in each condition is summarized in Table 2.

(A) No augmentation

Only the training dataset of JtubeSpeech-ASV, no data augmentation was applied.

(B) Noise

For the training dataset of JtubeSpeech-ASV, the number of utterances is augmented by noise superimposed. The noise data was randomly selected from MUSAN's noise dataset, and the SNR was set to 0. Two types of noise (noise and music) data were superimposed for each voice data, and the number of utterances per speaker was tripled.

(C) VTLP (All)

Augmenting the number of speakers by VTLP on the JtubeSpeech-ASV training dataset. The VTLP (All) applies VTLP to 1,792 speakers in the training dataset. In the conventional method (All), VTLP is applied to the 1,792 speakers in the training dataset. The frequency expansion coefficients are set to 0.1 and -0.1, and VTLP is applied to each voice to generate two voices each, and 3,584 pseudo-speakers were added by assigning a new speaker label to the speech as if it were spoken by a different speaker from the original speaker.

(D) VTLP (Select)

Among the pseudo-speakers generated by the VTLP (All), by using ECAPA-TDNN as the extraction model for speaker embedding, only those pseudo-speakers with a large variability of speaker characteristics from the original speaker are selected for augmenting the number of speakers.

(E) VTLP (Proposed)

Among the pseudo-speakers generated by the VTLP (All), by using ECAPA-TDNN as the extraction model for speaker embedding, it is applied to the pseudo-speakers which are

excluded by condition D, therefore the amount of data could be increased. Though comparing the audio generated by different VTLP factors, after the factor reaches 0.17, almost no clear words could be heard compared with the original audio. So the frequency scaling factor will be two cases: increasing in increments of 0.01 from 0.1 to 0.17 or decreasing in increments of 0.01 from -0.1 to -0.17.

(F) VTLP (Proposed + Noise)

The experiment augmented the data for both the number of augmented speakers and the number of utterances by adding noise and music to the speech in condition (E). Noise data were randomly selected from the MUSAN noise data set, and SNR was set to 0. Two types of noise data were superimposed on each speech data, and the number of utterances per speaker was increased by a factor of three.

The experiment used the equal error rate (EER) as an evaluation index. The EER is calculated from the point where the acceptance rate of a person by another person is equivalent to his/her rejection rate, and the smaller the value is, the better the accuracy is evaluated.

C. Experimental Result

Table 3 shows the EERs of the ASV results for each of the conditions (A)-(F). First, comparing the two conditions (A) and (B), the EER for condition (B) is 0.781 points lower than that for condition (A), which confirms the effectiveness of speech augmentation by adding noise and music. Futhermore, comparing the conventional method (C) and (D), both EERs are almost the same as the EER of (A) without data augmentation. It is obvious that (C) has the highest EER overall, even though the amount of data is almost the same as in (B). From the results above, it suggests that the ECAPA-TDNN does not show much improvement in accuracy compared to the conventional method. About comparing the three conditions (C)-(E), the EERs of (D) and (E) are lower than those of (C) because the variability of speaker characteristic is taken into account while the number of speakers is augmented using VTLP. The fact that the EERs for both cases improved despite the smaller number of speakers and data volume compared to (C) confirms the necessity of taking the variability of speaker characteristic into account. Among these three conditions (C)-(E), the proposed method (E) has the lowest EER, which is 0.945 points lower than that of (A). Compared to both conventional methods, the proposed method (E) shows a significant improvement in accuracy, and its EER is 0.939 points lower than that of (D). In both (D) and (E), the number of speakers is augmented to account for the variability of speakerness, however (E) uses more pseudo-speakers for augmentation than (D). This confirms the effectiveness of using more pseudo-speakers by adjusting the parameters. In (F), the number of utterances per speaker is further augmented from (E), in which the number of speakers is augmented using the proposed method, and the EER is reduced by 0.651 points compared to (E). At the same time, the EER was 0.815 points lower than that of (B), which only expanded the number of utterances, resulting in the best performance among all conditions. From the above results, it is confirmed that, in data augmentation for ASV based on speaker embedding, the combined use of noise-superimposed utterance augmentation and speaker number augmentation with consideration of variability of speaker characteristic is very effective in improving the performance.

TABLE III: EER (%) of speaker verification for each condition

Augmentation condition	**ECAPA EER(%)**
(A) No augmentation	6.984
(B) Noise	6.203
(C) VTLP (All)	7.018
(D) VTLP (Select)	6.978
(E) VTLP (Proposed)	6.039
(F) VTLP (Proposed+Noise)	5.388

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a method for augmenting training data for ASV models by generating pseudo-speakers through VTLP and considering the variability of speaker characteristic for selection. Simultaneously, we augment the number of speakers by varying the VTLP parameters. In the experiment, data augmentation was applied to the JTubeSpeech-ASV training dataset by adding noise to increase the number of utterances, generating pseudo-speakers to increase the number of speakers, and using both methods. An ASV model based on speaker embeddings was trained and evaluated the performance of the ASV system. The experimental results demonstrated that the best performance was achieved when more pseudo-speakers were added to the training data, with a focus on selecting those with significant variations in speaker characteristics.

For future work, it is considered necessary to explore more quantitative methods for setting the most appropriate frequency scaling factors and thresholds. Additionally, since there can be differences in the variations of speaker characteristics even within the same speaker's speech, it is necessary to consider more suitable methods for selecting pseudo-speakers. Last but not least, we will analysis false acceptance and false rejection in the experiment, and evaluate the performance of the proposed method on the Voxceleb database.

ACKONWLEDGMENT

This work was supported in part by JSPS KAKENHI (Grant Number JP24K14993), and ROIS DS-JOINT (022RP2024) to S. Shiota. We would like to thank Ms. Ono for her valuable assistance with the statistical analysis and her insightful suggestion on the research methodology.

REFERENCES

[1] Z. Rui and Z. Yan, "A survey on biometric authentication: Toward secure and privacy-preserving identification," *IEEE Access*, vol. 7, pp. 5994–6009, 2019.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[4] S. Wang, J. Rohdin, O. Plchot, L. Burget, K. Yu, and J. Černocký, "Investigation of specaugment for deep speaker embedding learning," in *Proc. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[5] P. S. Nidadavolu, V. A. Iglesias, J. Villalba, and N. Dehak, "Investigation on neural bandwidth extension of telephone speech for improved speaker recognition," in *Proc. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[6] C.-L. Huang, "Exploring effective data augmentation with tdnn-lstm neural network embedding for speaker recognition," in *Proc. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[7] Y. Zhu, T. Ko, and B. K.-W. Mak, "Mixup learning strategies for text-independent speaker verification," in *Proc. Interspeech*, 2019.

[8] H. Yamamoto, K. A. LEE, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," in *Proc. Interspeech*, 2019.

[9] T. Wakamatsu, S. Shiota, and H. Kiya, "Vocal tract length perturbation-based pseudo-speaker augmentation for speaker embedding learning," in *Proc. 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023.

[10] K. Zhou, Q. Yang, X. Sun, and S. Liu, "A deep speaker embedding transfer method for speaker verification," in *Proc. Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Volume 1*, Springer, 2020, pp. 369–376.

[11] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML workshop on deep learning for audio, speech and language*, vol. 117, 2013, p. 21.

[12] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996.

[13] K. Johnson, "Vocal tract length normalization," *UC Berkeley PhonLab Annual Report*, vol. 14, no. 1, 2018.

[14] L. D. Lee and R. C. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 49–60, 1998.

[15] Z. Zhou, S. Xu, S. Yin, L. Li, and D. Wang, "A comprehensive investigation on speaker augmentation for speaker recognition," in *Proc. Interspeech 2024*, 2024, pp. 2160–2164.

[16] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Proceedings of INTERSPEECH*, 2015, pp. 3586–3589.

[17] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 1711–1723, 2007.

[18] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, "Jtubespeech: Corpus of japanese speech collected from youtube for speech recognition and speaker verification," *arXiv:2112.09323*, Dec. 2021.

[19] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *ArXiv*, vol. abs/1510.08484, 2015.

Differences Between Singer and Speaker Verification: Training Singer Feature Representation Extractor Utilizing Singing Voice Characteristics

Sayaka Toma*, Tomoki Ariga*, Yosuke Higuchi*, Ichiju Hayasaka†, Rie Shigyo† and Tetsuji Ogawa*
* Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan
† DAIICHIKOSHO CO., LTD., Tokyo, Japan
E-mail: toma@pcl.cs.waseda.ac.jp

***Abstract*—We aimed to construct a robust feature extractor for singer verification using a straightforward method that leverages the unique characteristics of singing voices. The speaker feature extractor based on ECAPA-TDNN needs to be trained to identify numerous speakers using voice data that vary within the same speaker and to distinguish voices that sound similar but belong to different speakers. However, collecting singing voice data is challenging, particularly when constructing a corpus with a large number of singing samples from the same singer. In this study, we address this problem by adopting a simple approach: segmenting singing voices before inputting them into ECAPA-TDNN during training. The validity of this approach arises not merely from increasing the amount of data, but from the inherent characteristics of singing voices. Specifically, compared to spoken voices, singing voices exhibit less acoustic variation over short periods but greater variation over long periods. By utilizing short segments of singing voices for training, we can develop a feature extractor that is robust to acoustic variations within the same singer. Our experiments demonstrate the effectiveness of the proposed approach of segmenting singing voices into three-second intervals for training and provide insights that this method does not yield the same benefits for spoken voices, highlighting its unique effectiveness for singer verification using singing voices.**

I. Introduction

Speaker verification technology, which determines whether different voice inputs belong to the same speaker, has been extensively studied for practical applications. However, most of these studies focus on spoken speech [1]–[5]. Conversely, achieving high accuracy in singer verification using singing voices could enable the creation and viewing of singing histories without requiring user registration. Additionally, singer verification can be seen as a technology that facilitates the comparative evaluation of acoustically similar singing, allowing for the recommendation of songs that are easier for users to sing (i.e., songs frequently sung by people with similar voices). Nevertheless, in contrast to the well-developed and extensively researched speaker verification technology for spoken speech, singer verification technology for singing voices has not been sufficiently explored [6]–[8].

In general, speaker verification involves using deep neural networks (DNNs) such as ECAPA-TDNN [9] or ResNet [10], [11], which are designed to identify numerous speakers from long-duration voice inputs. The similarity of the obtained speaker feature representations is then calculated using cosine distance [3] or probabilistic linear discriminant analysis (PLDA) [12], [13]. Specifically, speaker feature representations are obtained by accumulating and statistically processing local acoustic information extracted through convolutional layers over extended periods (attentive pooling) [14]. Ideally, the speaker feature extraction process should compact the acoustic data distribution of the same speaker (reducing false rejections) and separate the acoustic data distribution of different speakers (reducing false acceptances). To accurately achieve such functionality, it is crucial to train DNNs to identify numerous speakers using voice data that vary within the same speaker and voices that sound similar but belong to different speakers. However, for singing voices, there is no well-prepared corpus that meets these requirements, and it is particularly challenging to provide a large number of singing samples from the same singer.

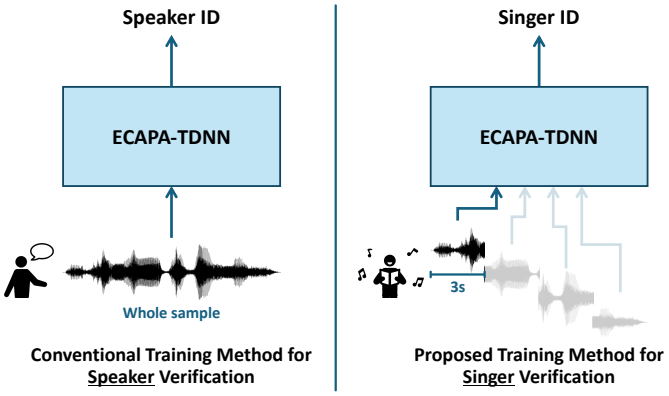


Fig. 1. Proposed strategy for building singer feature extractor robust to acoustic variability, compared to conventional strategy for training speaker feature extractor. In our approach, singing voice is first segmented into short, stationary signal intervals (e.g., three seconds) and then used to train ECAPA-TDNN to perform singer identification using these segments.

In contrast, this study seeks to address the problem using a straightforward approach that effectively leverages the unique characteristics of singing voices. Singing and spoken voices

have significant differences in their properties. Spoken voices exhibit substantial acoustic variation over short periods due to phonetic differences, but this variation seldom changes dramatically. Conversely, singing voices show relatively less acoustic variation over short periods but display greater acoustic variation over long periods due to singing techniques and song structures. To capitalize on these characteristics, this study proposes dividing singing voices into short, steady-state signals and using these as inputs to train an ECAPA-TDNN for singer identification (as illustrated in Figure 1). Given the nature of singing voices, it is expected that although the numerous short signals are individually steady-state, they will exhibit significant variation in the embedding space produced by the ECAPA-TDNN. Consequently, training the singer feature extractor in this manner is anticipated to yield robust singer feature representations that are resilient to acoustic variations within the same singer and sensitive to acoustically similar inputs from different singers.

By comparing the results of training ECAPA-TDNN using entire songs versus segmented singing voices, we aim to clarify the effectiveness of the proposed approach. Additionally, we will conduct experiments by segmenting spoken voice inputs and feeding them into ECAPA-TDNN to demonstrate that the proposed approach specifically addresses issues unique to singing voice verification. The contributions of this study are as follows:

- Provide insights into the differences between singing and spoken voices.
- Present a simple approach to significantly reduce false rejections in singer verification.
- Demonstrate the effectiveness of the proposed approach through experiments using a large-scale karaoke singing voice dataset and provide insights into the differences from speaker verification using spoken voices.

The findings from this study are expected to be valuable for researchers and engineers who wish to develop singer verification systems more easily.

The remainder of this paper is organized as follows. Section II explores the acoustic differences between spoken and singing voices. Section III details the singer verification system employed in this study. Section IV presents the singer verification experiments conducted using karaoke singing voices and evaluates the effectiveness of training the feature extractor with segmented voice inputs. Finally, Section V provides this study's conclusion.

II. Differences in Acoustic Variations Between Spoken and Singing Voice

In general, singing voices exhibit less acoustic variation over short periods compared to spoken voices. Figure 2 shows histograms of acoustic variation within short segments of spoken and singing voices. The acoustic variation for short segments was measured by calculating the variance of mel-frequency cepstral coefficients (MFCCs) using a 25-millisecond frame size and a 10-millisecond frame shift for each three-second segment. These variance values were then averaged for each speaker or singer, with the analysis performed on data from 100 individuals. The histograms indicate that acoustic variation within short segments is smaller for singing voices than for spoken voices.

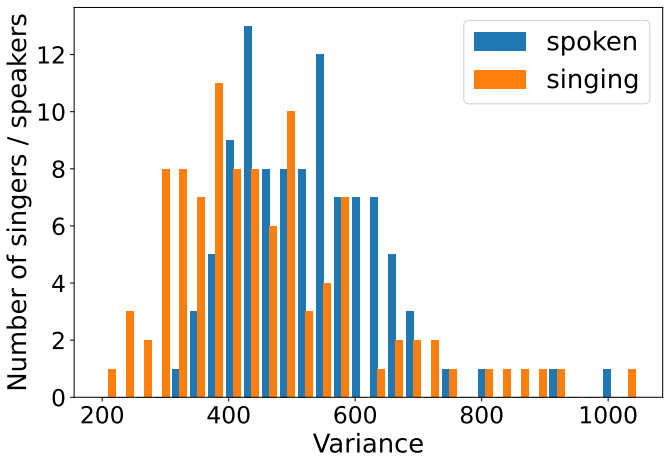


Fig. 2. Histogram of intra-singer and intra-speaker acoustic variations over short periods for singing and spoken voices.

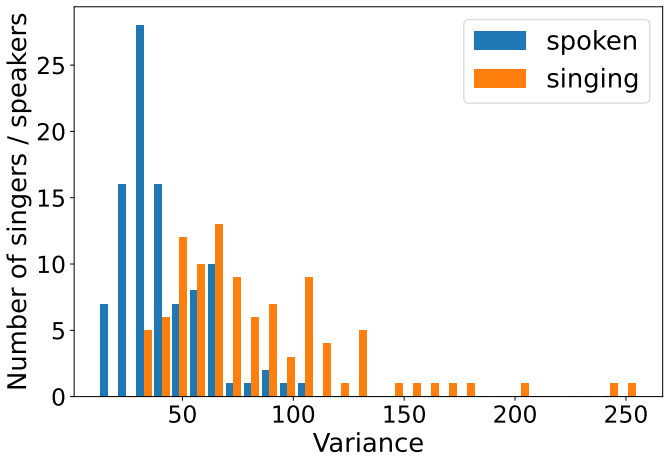


Fig. 3. Histogram of intra-singer and intra-speaker acoustic variations over long periods for singing and spoken voices.

Conversely, due to significant changes in melody and pitch throughout different parts of a song, acoustic variation over long periods is greater for singing voices compared to spoken voices. Figure 3 represents histograms of acoustic variation over long segments, which were computed as the variance within each speaker of the mean MFCCs calculated for each three-second frame. This graph demonstrates that acoustic variation over long periods is greater for singing voices than for spoken voices.

TABLE I
STRUCTURE OF ECAPA-TDNN.

Index	DNN Module	Input Dim.	Output Dim.
1	TDNNBlock	$80 \times T$	$1024 \times T$
2	SE-Res2Block	$1024 \times T$	$1024 \times T$
3	SE-Res2Block	$1024 \times T$	$1024 \times T$
4	SE-Res2Block	$1024 \times T$	$1024 \times T$
5	TDNNBlock	$3072 \times T$	$3072 \times T$
6	AttentiveStatisticsPooling	$3072 \times T$	256
7	Conv1d	256	192

III. SINGER VERIFICATION

A. ECAPA-TDNN

We focus on utilizing a widely adopted speaker verification system for singer verification, leveraging ECAPA-TDNN [9] for embedding singer information from singing voice data and employing cosine similarity scoring for verifying singers [3].

The ECAPA-TDNN-based feature extractor, configured with the parameters detailed in Table I, is first trained using our singing voice database (see Section IV-A). The intermediate layer outputs are then extracted as singer embeddings, which are subsequently used for cosine similarity scoring. ECAPA-TDNN incorporates attentive pooling and Squeeze-and-Excitation (SE) blocks [15] to enhance the learning of relevant features within the model. Additionally, by combining the outputs not only from the preceding layer but also from shallower layers during the pooling process, ECAPA-TDNN is expected to capture more robust singer embeddings.

Increasing the number of identification classes in this network is expected to capture finer distinctions among speakers, potentially reducing the false acceptance rate. However, if the data available for each singer is insufficient, it may hinder the model's generalization, potentially leading to an increased false rejection rate.

B. Proposed Training Strategy for Singer Feature Extractor

In speaker verification tasks involving spoken voice, shortening the input length for ECAPA-TDNN is generally not recommended. Spoken voice exhibits significant acoustic variations over short periods, as phonemes change within a few hundred milliseconds, but the overall acoustic variation across an entire utterance is relatively small. In contrast, this study on singer verification seeks to train the ECAPA-TDNN using singing voice samples segmented into short durations. As discussed in Section II, singing voice typically shows minimal acoustic variation over short periods but significant variation over the duration of an entire song. Because of this characteristic, short segments of the same song may display considerable variability in the embedding space. By intentionally increasing the acoustic variability within the same singer's samples during training, we aim to create an optimal feature embedding for singer verification: it suppresses intra-singer acoustic variations while effectively distinguishing between different singers.

In this approach, whether the entire song or segmented portions are used as input, the total amount of training data remains the same. However, capturing the precise acoustic variations across an entire song in the averaged embedding representation after TDNN pooling is likely to be limited. Segmenting the singing voice offers a simple and effective approach to overcoming this limitation.

IV. SINGER VERIFICATION EXPERIMENTS

To evaluate the effectiveness of the proposed approach, we conducted singer verification experiments using a custom-built singing voice corpus. In the first experiment, we compared the verification performance of ECAPA-TDNN when trained with segmented singing voice inputs versus entire song inputs. In the second experiment, to elucidate the differences between singing and spoken voices, we compared the verification performance for spoken voices using segmented versus unsegmented inputs during the training of ECAPA-TDNN.

A. Database

To conduct a singer verification experiment, we compiled a comprehensive database of Japanese singing voices. This dataset includes 4950 songs performed by 1666 amateur singers in karaoke settings. Of the 1666 participants, 1000 singers performed only one song, 225 sang two songs, and 136 sang three songs. The recordings were made using directional microphones installed in karaoke booths, capturing the singing voices with minor contamination from background music, other voices, and reverberation.

B. Experimental Setups

The singing voices used in the experiment were sampled at 16 kHz and quantized into 16-bit data. To mitigate the impact of noise, particularly from background music, we performed data augmentation for training singer embeddings using rirs-noises [16] and MUSAN [17], which include human voices, environmental sounds, and music as noise sources. The training of the ECAPA-TDNN model [9] utilized our singing voice dataset, which comprised 3500 songs performed by 441 singers (each singing three or more songs).

For the enrollment and verification phases, we used singing voice data from 225 singers, with each singer contributing two songs—one for enrollment and the other for verification. The total number of trials was determined by the combinations of enrollment and verification data for the 225 singers, resulting in 225×225 trials. Notably, the enrollment and verification data were sampled from different songs. We computed similarity scores for all possible pairs of data, and the evaluation was based on the Equal Error Rate (EER) and the minimum Decision Cost Function (minDCF). Both the training and inference processes were conducted using the SpeechBrain toolkit [18].

C. Experiment 1

1) Experimental Procedure: To assess the effectiveness of dividing singing voice data into segments of N seconds for training a singer feature extractor, we compared the verification performance for segment lengths of 3 seconds, 10 seconds, and

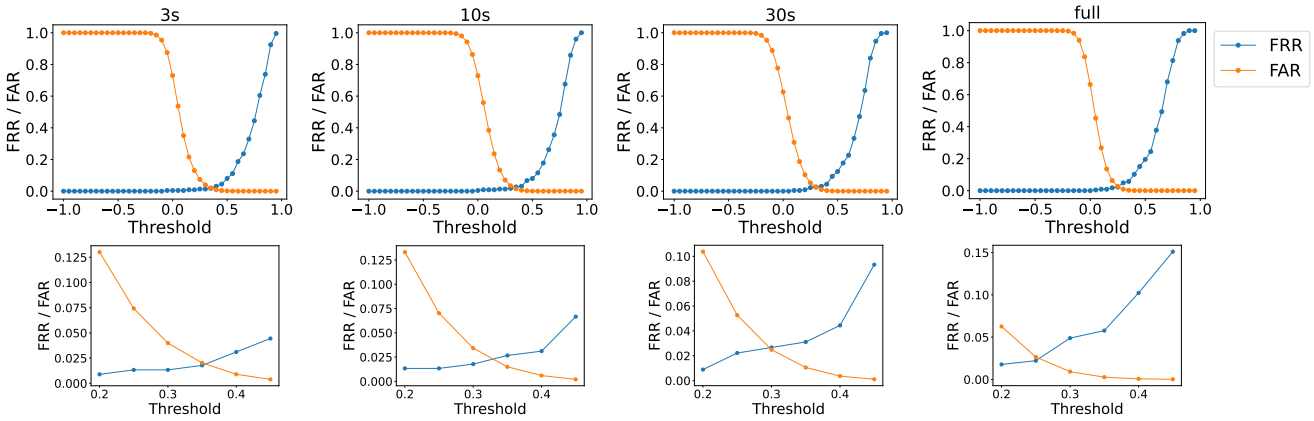


Fig. 4. Upper figure: False rejection rate (FRR) and false acceptance rate (FAR) as function of threshold for cosine similarity. Lower figure: An expanded figure near the EER.

TABLE II
EQUAL ERROR RATE FOR VARIOUS LENGTHS OF TRAINING AND VERIFICATION DATA. ENTIRE SONG (FULL) WAS USED FOR ENROLLMENT DATA.

Length of Training Data	**Length of Verification Data**		
	full	30s	10s
full	1.72	2.92	9.33
30s	2.22	2.85	7.53
10s	1.78	**2.67**	4.44
3s	**1.33**	2.76	**3.97**

TABLE III
NUMBER OF FALSE REJECTIONS AND MINDCF. C_{MISS} AND C_{FA} REPRESENT COSTS ASSOCIATED WITH MISSED DETECTION ERRORS AND FALSE ALARMS, RESPECTIVELY.

Length of Training Data	**Number of False Rejections**	**minDCF** ($C_{\text{miss}} = 2$, $C_{\text{fa}} = 1$)
full	29	0.15
30s	21	0.13
10s	22	**0.11**
3s	23	0.14

30 seconds with the performance obtained without segmentation. We ensured that the sample count of the training data remained constant regardless of N by overlapping the segmentation windows. Additionally, to examine the effects of varying input audio lengths between enrollment and verification data, we evaluated verification performance using the entire song, the first 10 seconds, and the first 30 seconds for the verification audio, while the entire song was used for the enrollment audio.

2) Experimental Results: Table II presents the verification performance for various lengths of training and verification data. The results demonstrate that training the feature extractor with short, segmented singing voice data (three seconds) directly enhances verification performance, regardless of the input length of the verification data. Additionally, using an entire song for both enrollment and verification data yielded the best performance, with an EER of 1.33. These findings suggest that while using longer-duration data for enrollment and verification enhances the reliability of the verification process, a feature extractor trained on short, segmented audio can effectively capture precise (capable of distinguishing between similar but different singers) and robust (suppressing acoustic variations of the same singer) singer features.

When applying singer verification to build and reference individual singing histories in karaoke, minimizing false rejections is essential. In particular, ECAPA-TDNN trained with short, segmented singing voice data effectively reduced false rejections. Figure 4 illustrates the false rejection rate (FRR) and false acceptance rate (FAR) as the threshold for cosine similarity in verification is varied in increments of 0.05 for each input length of the singing voice data used in training (entire song, 3 seconds, 10 seconds, 30 seconds). The results show that the integral of the FRR curve is smallest when the training data input length is 3 seconds, and it increases progressively as the input length extends to 10 seconds, 30 seconds, and an entire song.

Table III presents the minDCF values, adjusted to prioritize the reduction of false rejections (with the false rejection rate weighted twice as heavily as the false acceptance rate), along with the corresponding number of false rejections at the adjusted threshold settings. The results show a decrease in false rejections when using segmented inputs of 30 seconds, 10 seconds, and 3 seconds, compared to using the full-length song, with 10-second segments achieving the best minDCF performance. Although Table II indicates that the EER worsens with 30-second and 10-second segments compared to full-length inputs, it is important to recognize that these segmentations help reduce false rejections. This suggests that the observed deterioration in EER is likely due to an increase in false acceptances.

D. Experiment 2

1) Experimental Procedure: To evaluate whether similar effects can be achieved with spoken voice data, we conducted

TABLE IV
EER FOR TRAINING DATA LENGTH IN SPEAKER VERIFICATION.

Length of Training Data	EER
full	1.48
3s	**1.41**

an experiment similar to Experiment 1, where we segmented spoken voice data and input it into the ECAPA-TDNN model. Using audio segments longer than six seconds from the VoxCeleb dataset, we compared the performance of training with three-second segments to that of training with full-length audio.

2) Experimental Results: The results are presented in Table IV. For spoken voice, the verification performance remained consistent regardless of whether the audio was segmented. Due to the relatively stable acoustic characteristics of spoken voice over long durations, segmenting the audio and inputting it into the ECAPA-TDNN did not enhance the extraction of robust speaker features against acoustic variability, and thus did not impact verification performance. These findings suggest that the observed improvement in verification performance from training with short segments is specific to singing voice data, which exhibits substantial acoustic variation over long durations.

V. Conclusion

In this study, we proposed a straightforward method for extracting robust feature representations for singer verification by leveraging the distinct characteristics of singing voice, which shows minimal acoustic variation over short periods but significant variation over longer durations compared to spoken voice. We demonstrated that segmenting the input into shorter segments for training the feature extractor can effectively reduce false rejections in singer verification. Furthermore, we established that this effect is unique to singing voice and does not extend to spoken voice.

References

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, pp. 19–41, 2000.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[4] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Proceedings of Interspeech*, vol. 2017, 2017, pp. 999–1003.

[5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[6] L. Regnier and G. Peeters, "Singer verification: Singer model .vs. song model," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 437–440.

[7] L. Wang, B. Wang, G. Tan, *et al.*, "ATRemix: An autotune remix dataset for singer recognition," in *Proceedings of Chinese Conference on Biometric Recognition*, 2022, pp. 348–355.

[8] H. Yakura, K. Watanabe, and M. Goto, "Self-supervised contrastive learning for singing voices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1614–1623, 2022.

[9] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proceedings of Interspeech*, 2020, pp. 3830–3834.

[10] C. Li, X. Ma, B. Jiang, *et al.*, "Deep speaker: An end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[11] T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 301–307.

[12] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision–ECCV European Conference on Computer Vision*, 2006, pp. 531–542.

[13] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE international conference on computer vision*, 2007, pp. 1–8.

[14] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proceedings of Interspeech*, 2018, pp. 3573–3577.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[16] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[17] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[18] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

Accelerated Real-Time Local Maxima Detection in Video Streams Using FPGA Technology

Anindhita Nayazirly Sukarno*, Yahwista Salomo†, Trio Adiono, Infall Syafalni,
Nana Sutisna, and Rahmat Mulyawan
* Institut Teknologi Bandung, Indonesia
E-mail: zirlysukarno@gmail.com
† Institut Teknologi Bandung, Indonesia
E-mail: yahwistasalomo@gmail.com

***Abstract*—In this paper, we present a successful implementation of a local maxima filter on a Zybo Z7-20 and PYNQ Z1 FPGA using their two HDMI ports in real-time. The proposed system uses the HDMI ports to capture video frames with a resolution of 640x480 pixels. The local maxima filter is then applied to the captured frames in real-time, allowing for the detection of peaks in the image data. The filter uses a sliding window approach to determine the local maxima, and a threshold value is set to identify and retain only the most significant peaks in the image. The system was implemented using SystemVerilog Hardware Description Language (HDL). The system was developed in the Xilinx Vivado design suite and the results show that the proposed system is able to process video frames at a rate of 60 frames per second with high accuracy and low latency. The proposed implementation using SystemVerilog presents a more efficient and flexible solution for image processing applications on FPGA, making it a promising approach for real-time image processing.**

I. Introduction

Image processing has become an integral part of modern technological systems, with applications ranging from medical diagnostics to autonomous vehicles [1]. Real-time processing of high-resolution video streams poses significant challenges, particularly in resource-constrained environments [2].

The identification of local maxima in image frames is a fundamental operation in many computer vision tasks, including feature detection and segmentation [3]. However, performing this operation in real-time on high-resolution video streams (e.g., 640x480 at 60 fps) requires substantial computational resources [4].

Traditional software-based approaches often struggle to meet the stringent timing requirements of real-time video processing [5]. While GPU-accelerated solutions offer improved performance, they typically consume significant power and may not be suitable for embedded systems [6].

There is a need for a hardware-accelerated solution that can perform real-time local maxima detection on high-resolution video streams while maintaining low latency and power efficiency [7]. Existing FPGA-based solutions often lack the necessary throughput or flexibility to handle varying image complexities [8].

This paper presents a novel FPGA-based system for real-time local maxima detection in video streams. Our key contributions include:

- A highly parallelized architecture capable of processing 640x480 pixel frames at 60 fps [9].
- An adaptive processing algorithm that dynamically adjusts to image complexity [10].
- An efficient memory management scheme optimized for row-wise image processing [11].
- A pipelined design that maximizes hardware utilization and minimizes latency [12].
- Comprehensive performance analysis and comparison with existing solutions [13].

Our proposed system demonstrates significant improvements in processing speed and efficiency, paving the way for advanced real-time image processing in resource-constrained environments [14].

II. Local Maxima Algorithm

The process of identifying 3x3 local maxima in an image involves a siterative procedure that examines each pixel in relation to its surrounding neighbors. This algorithm, which shares similarities with techniques described by Gonzalez and Woods [1] in their seminal work on digital image processing, can be delineated as follows:

- Pixel Iteration: The algorithm systematically traverses each pixel within the image matrix, constructing a 3x3 window centered on the current pixel under evaluation. This approach is reminiscent of convolution operations used in various image processing tasks [3].
- Comparative Analysis of Neighboring Pixels: For each target pixel, a rigorous comparison is conducted against its eight adjacent pixels within the defined 3x3 window. A pixel is designated as a local maximum if and only if its value surpasses those of all its neighbors. In scenarios where the target pixel's value is equivalent to one or more of its neighbors, a more nuanced approach is adopted: the target pixel's status as a local maximum is determined by the output status of its neighboring pixels, introducing a level of contextual decision-making [15].

- Image Analysis: The algorithm rigapplies this evaluation process to every pixel within the image bounds, ensuring a thorough analysis of the entire visual field. This approach is critical for maintaining the integrity of the local maxima detection across the image [10].
- Output Generation and Representation: The culmination of this process yields a set of coordinates corresponding to all identified local maxima. This output can be represented in various forms, such as a binary image or a list of coordinate pairs, depending on the specific requirements of subsequent processing stages [6].

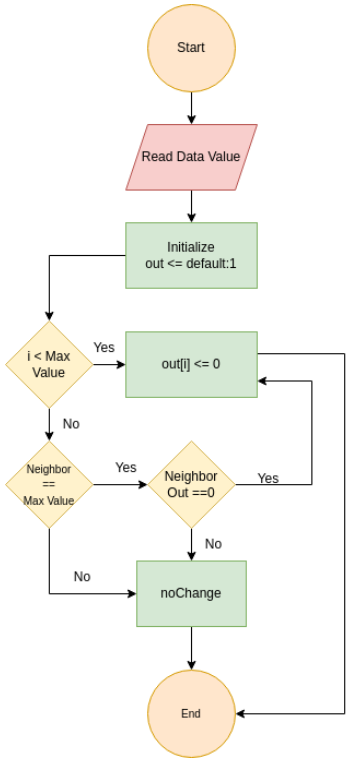


Fig. 1. Schematic representation of the local maxima algorithm (single iteration)

Fig. 10 elucidates the intricate timing management within the process block. The counter initiates its sequence upon the deactivation of the reset signal. The write_address and read_address signals are strategically delayed in relation to the counter, implementing a form of pipeline processing to optimize throughput. In this context, the notation 'd' denotes raw data input, 'o' represents the intermediate output values awaiting processing, and 'n' signifies the resultant values destined for storage in bram_out_curr.

A. Mathematical Representation of the algorithm

Below is the iterative algorithm for local maxima detection. We can represent this algorithm mathematically as follows: Let I be the input image of size $M \times N$, and O be the output image of the same size. Initialize O with all elements set to 1:

$$O_{i,j} = 1 \quad \forall i \in [1, M], j \in [1, N] \tag{1}$$

Define the neighborhood $\mathcal{N}(i, j)$ of pixel (i, j) as:

$$\mathcal{N}(i, j) = (i', j') : |i - i'| \leq 1, |j - j'| \leq 1, (i', j') \neq (i, j) \tag{2}$$

The algorithm then iteratively updates O until convergence: For each pixel (i, j):

Find the maximum value in the neighborhood:

$$\max_{local} = \max_{(i',j') \in \mathcal{N}(i,j)} I_{i',j'} \tag{3}$$

If $I_{i,j} < \max_{local}$:

$$O_{i,j} = 0 \tag{4}$$

If $I_{i,j} = \max_{local}$:

$$O_{i,j} = \min_{(i',j') \in \mathcal{N}(i,j), I_{i',j'} = I_{i,j}} O_{i',j'} \tag{5}$$

The algorithm continues until O no longer changes between iterations.

III. Proposed Design

A. System Architecture and Design

1) Image Processing Architecture: Both the input and output frames adhere to a 640x480 pixel resolution with a 24 bits per pixel (bpp) color depth, consistent with standard VGA specifications [4]. However, the system introduces an intermediate step where the received frame undergoes grayscale conversion to 6 bpp. This reduction in color depth not only optimizes memory usage but also simplifies subsequent processing steps, a common technique in image pre-processing [5]. Following the local maxima computation, the output is remapped to the full 24 bpp color space for transmission, ensuring compatibility with standard display devices. The system's memory architecture is built around three specialized block RAMs, each serving a distinct purpose in the processing pipeline: Input Array: Stores the incoming frame data from the DVI decoder. Output Array: Houses the intermediate results of the local maxima algorithm. Transmission Buffer: Contains a duplicate of the final output array, ready for transmission. This tripartite memory structure facilitates efficient data flow and minimizes latency, a crucial factor in real-time image processing systems [5].

2) Processing Unit: The processing flow incorporates a conditional branching mechanism: upon completion of the looping process, the final processed data is committed to a designated BRAM for subsequent transmission. However, if the process remains incomplete, the intermediate output is written to a buffer RAM and reprocessed. This iterative approach ensures the algorithm's convergence, a critical aspect in handling complex image structures [1]. The loop count increments with each iteration, with data read out sequentially until the process concludes. As illustrated in Fig. 3, each loop cycle necessitates (BRAM_Depth + 6) clock cycles to conclusively verify the finished signal, a design choice that balances processing thoroughness with system latency [2].

Fig. 4 provides an in-depth view of the processor's architectural scheme. The design incorporates sophisticated shift

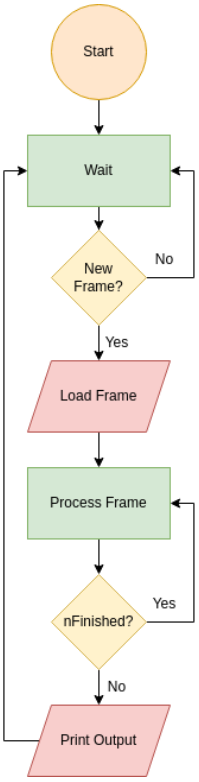


Fig. 2. Flowchart depicting the data processing scheme

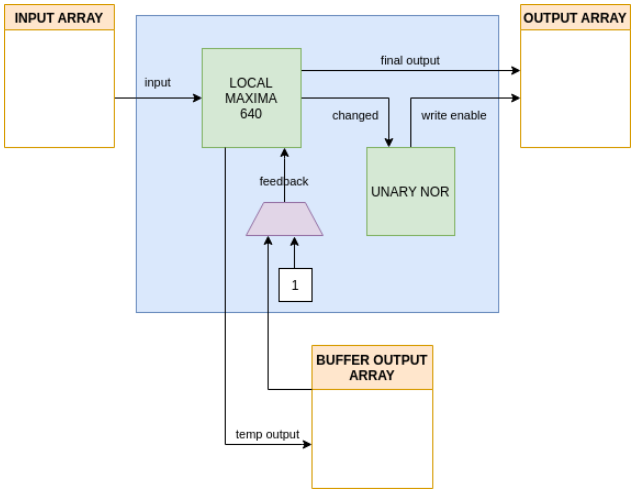


Fig. 3. Detailed block diagram of the advanced processor architecture

registers: at system initialization, out_buf_0 and out_buf_1 are preset to one, while data_buf_0 and data_buf_1 are initialized to zero. This initialization strategy ensures proper boundary handling, a crucial aspect in image processing algorithms [15]. In each clock cycle, the data in the shifters undergoes an upward shift. The local maxima combinational algorithm receives input from both the out_buf and data_buf shifters, with results subsequently buffered in the out_bram_curr memory. A notable feature of each of the 640 local_maxima combinational blocks is the inclusion of a specialized output port. This port is designed to detect variations from the previous value, triggering a reprocessing of the frame if changes are detected.

Fig. 5 offers a detailed view of the combinational logic within each local maxima processing block. Each module is designed to process 9 bits of image data from the BRAM alongside 9 bits of prior output data. The module then determines the output value, denoted as value_o. The 'changed' signal plays a crucial role in the system's adaptive behavior, determining whether additional processing iterations are necessary.

3) DVI Receiver: The DVI receiver subsystem is responsible for capturing and preprocessing the incoming video stream. Fig. 6 illustrates its complex architecture. The receiver integrates several key components:

4) DVI Transmitter: The DVI transmitter subsystem, depicted in Fig. 7, is responsible for converting the processed image data back into a format suitable for DVI transmission.

The DVI transmitter operates across multiple clock domains to meet the stringent timing requirements of the DVI standard:

A 25 MHz clock serves as the pixel clock, facilitating the output of 640x480 pixels at 60 frames per second, adhering to standard VGA timing specifications [4].

B. Performance Considerations and System Optimization

1) Pipelining and Parallelism: The system leverages both pipelining and parallelism to achieve high throughput. The 640 parallel "Local Maxima" combinational blocks represent a form of spatial parallelism, allowing the system to process an entire row of pixels simultaneously. This approach significantly reduces the time required to process each frame [12].

2) Memory Management: The use of three separate block RAMs (input array, output array, and transmission buffer) is a key feature that enables efficient data management. This approach minimizes memory conflicts and allows for simultaneous read and write operations, a technique that Ienne and Leupers [8] highlight as crucial for high-performance embedded systems. The row-based memory organization, with a depth of 480 and varying widths, is optimized for the specific requirements of image processing. This design choice facilitates burst reads and writes, reducing memory access latency and improving overall system performance [11].

3) Adaptive Processing: The inclusion of a 'changed' signal in the local maxima blocks, triggering reprocessing when necessary, represents an adaptive approach to image processing. This feature ensures that the system can handle varying levels of image complexity, dynamically adjusting its processing to maintain output quality. Such adaptive techniques are increasingly important in real-world image processing applications, as noted by Bovik [5].

IV. Hardware Implementation

As can be seen in Fig. 8 the system was implemented on Pynq Z1 (Digilent). The DVI source used for testing was Raspberry Pi 4 (Adafruit). The Raspberry Pi was used to play videos and stream camera. The DVI sink used for testing was

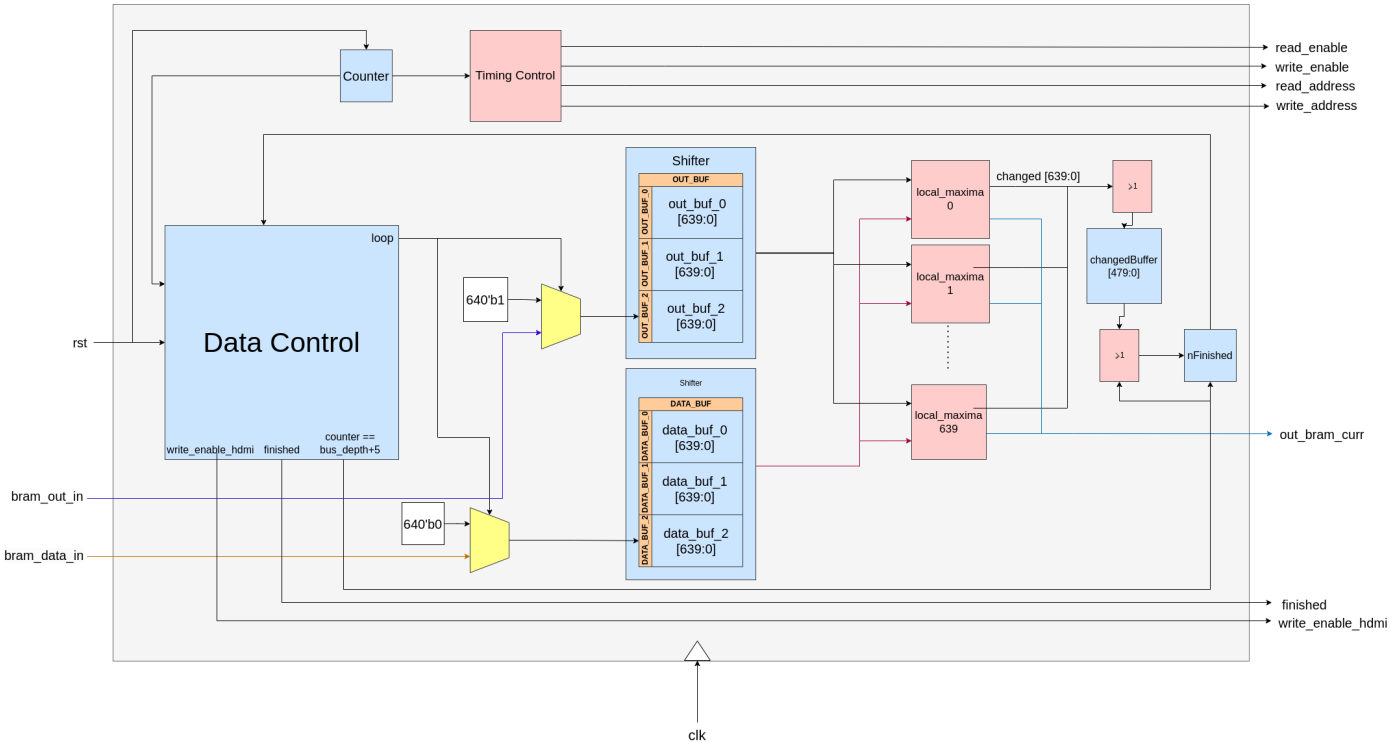


Fig. 4. Complete block diagram of the processor's internal structure

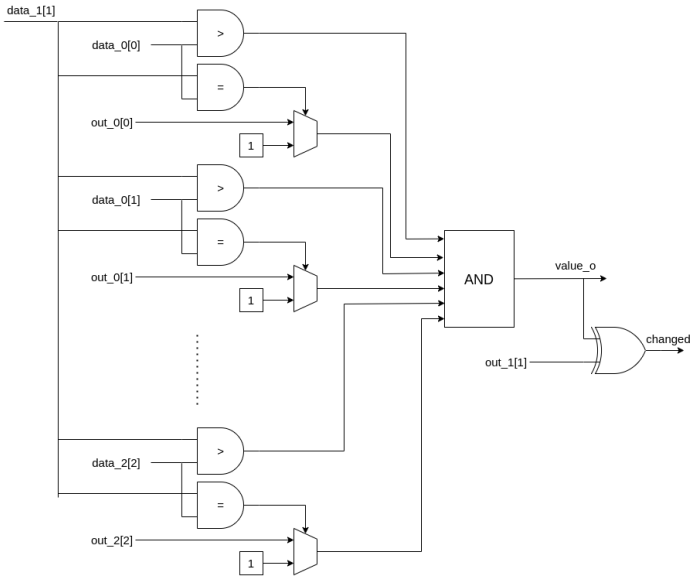


Fig. 5. Detailed combinational logic diagram of the local maxima processing unit

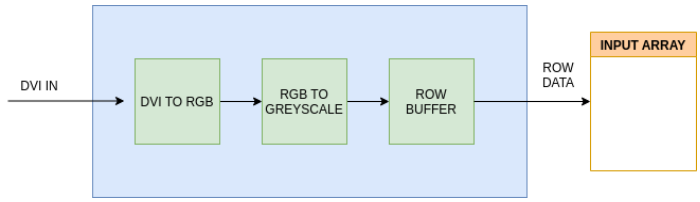


Fig. 6. Detailed block diagram of the advanced DVI receiver subsystem

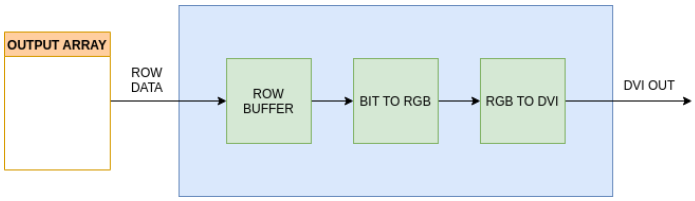


Fig. 7. Comprehensive block diagram of the advanced DVI transmitter subsystem

a generic display monitor. The Raspberry Pi unfiltered display was also mirrored to another display monitor for comparison. Fig. 9 shows a display of the two monitors. Video of the demo can be accessed at https://youtu.be/rrs4vBUoiV0.

Fig. 10 shows the expected behaviour of a single frame processed once in the loop. Note, Xilinx's bram requires two clock cycle to be registered.

V. Results and Discussion

A. Latency

The overall latency of the system is fixed as the time distance between two frames which are 16.67 milliseconds. The operating frequency of the datapath used is 25Mhz, which in its worst case scenario would require

$$Latency = 40ns * (BITWIDTH) * (BITDEPTH + 6)$$

or 12.4416 milliseconds to finish an operation.

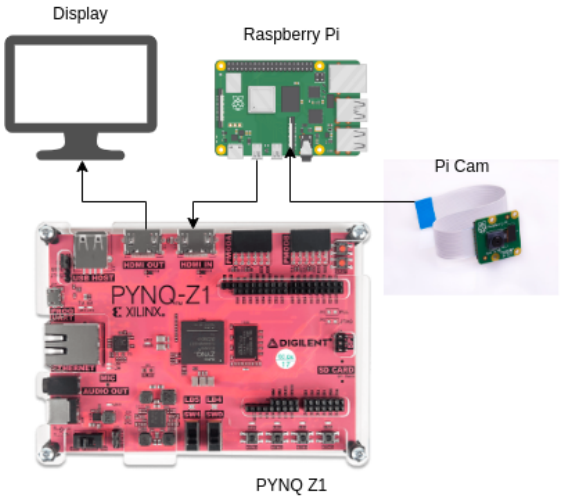


Fig. 8. Test setup

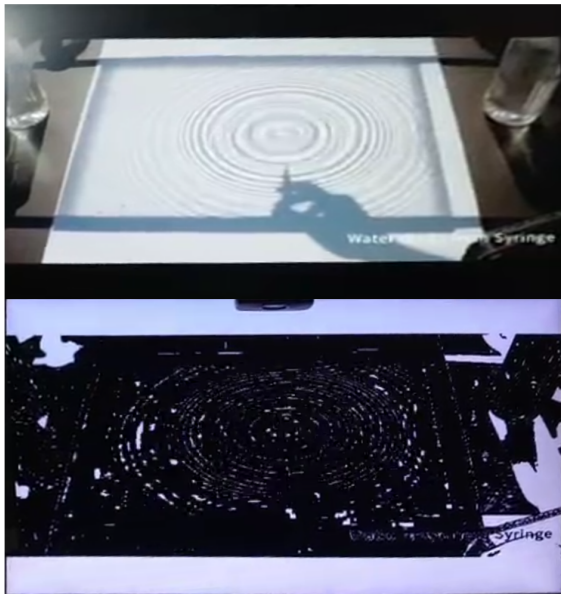


Fig. 9. Example of system input (top picture) and output (bottom picture)

TABLE I
RESOURCE UTILIZATION

Resource	Utilization
LUT	30581
LUTRAM	785
FF	24772
BRAM	88.50
IO	18
BUFG	6
MMCM	2

Note: LUT: Look-Up Table; FF: Flip-Flop;
BRAM: Block RAM; IO: Input/Output; BUFG: Global Clock Buffer;
MMCM: Mixed-Mode Clock Manager

Strictly speaking of the datapath delay path, ignoring other peripheral such as BRAM, ILA, or VIO, the critical path of the datapath of local maxima with 640 width, 480 depth, and 6 data width is 10.632ns. Which means the maximum internal operating frequency that can be achieved is around 94Mhz. with the worst possible scenario, could finish the operation in 3.3069 milliseconds and process 300 frames per second.

B. Scalability

The local maxima processing block are designed with scalability in mind, the relating hdl codes uses three parameter as its inputs, which are **DATA_WIDTH** to specify how many parallel data you want to be processed each clock, **DATA_DEPTH** to specify how many clock cycles per loop you want, and **BIT_WIDTH** to specify the number of bits used as the comparator.

C. Comparison of Equivalent Implementations

To contextualize the performance of our FPGA-based local maxima detection system, it's valuable to compare it with equivalent implementations on different platforms. This comparison highlights the advantages and trade-offs of various approaches.

1) CPU-based Implementation: Traditional CPU-based implementations, such as those using MATLAB or OpenCV, offer flexibility but often struggle with real-time performance for high-resolution video streams. For instance, Neubeck and Van Gool [16] reported processing times of 13.5 ms for a 512x512 image using an optimized algorithm on a 2.8 GHz Pentium 4. While this approaches real-time performance, it doesn't match our FPGA implementation's capability to process 640x480 frames at 60 fps (16.67 ms per frame).

VI. CONCLUSION AND FUTURE WORK/

The proposed system represents an approach to real-time local maxima detection in video streams. By leveraging advanced FPGA capabilities, including parallel processing, pipelined architectures, and efficient memory management, the system achieves high-performance image processing while maintaining the flexibility necessary for future enhancements. The design's modular architecture, with clearly defined input, processing, and output stages, provides a solid foundation for future research and development. As the field of real-time image processing continues to evolve, systems like this will play an increasingly important role in applications ranging from computer vision and robotics to medical imaging and autonomous vehicles.

Concluding, this paper has presented a successful design implementation on the Zybo Z7-20 and PYNQ Z1 using the HDMI output of Raspberry Pi. The proposed local maxima filter is a parallel 640 data blocks with a 10.632ns critical path, and is capable of much higher throughput than is currently achieved. Further works may include using it as a red blood cell counter or other image processing applications that require local maxima. With these improvements, the proposed design

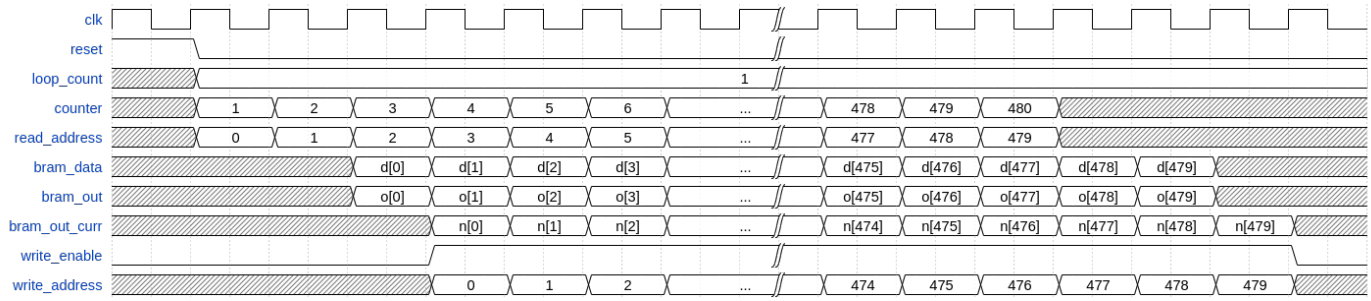


Fig. 10. Local maxima processor timing diagram for one loop

could provide a platform for more complex image processing tasks and applications.

While our current implementation demonstrates promising results in controlled environments using a Raspberry Pi as the video source, we acknowledge the need for more extensive testing in real-world conditions. Future work will focus on evaluating the system's performance across a wider range of scenarios, including varying lighting conditions, dynamic scenes, and different video resolutions. Additionally, we plan to conduct rigorous tests on the system's robustness in handling noisy or degraded video streams. This will involve introducing artificial noise, compression artifacts, and other forms of signal degradation to assess the algorithm's resilience. We also aim to explore adaptive thresholding techniques to enhance the system's ability to maintain accurate local maxima detection in suboptimal video conditions. These investigations will provide valuable insights into the system's applicability in diverse real-world applications, such as surveillance, autonomous vehicles, and medical imaging, where video quality can vary significantly.

References

[1] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Pearson, 2018.

[2] N. Kehtarnavaz, *Real-Time Digital Signal Processing: Based on the TMS320C6000*. Newnes, 2015.

[3] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[4] K. Jack, *Video demystified: a handbook for the digital engineer*. Newnes, 2007.

[5] A. C. Bovik, *Handbook of Image and Video Processing*, 2nd. Academic Press, 2010.

[6] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc., 2008.

[7] D. G. Bailey, *Design for embedded image processing on FPGAs*. John Wiley & Sons, 2011.

[8] P. Ienne and R. Leupers, *Customizable Embedded Processors: Design Technologies and Applications*. Morgan Kaufmann, 2006.

[9] K. Asanovic, R. Bodik, B. C. Catanzaro, *et al.*, "The landscape of parallel computing research: A view from berkeley," *Technical Report UCB/EECS-2006-183, EECS Department, University of California, Berkeley*, 2006.

[10] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[11] S. Hauck and A. DeHon, *Reconfigurable Computing: The Theory and Practice of FPGA-Based Computation*. Morgan Kaufmann, 2007.

[12] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, 6th. Morgan Kaufmann, 2019.

[13] L. Wanhammar, *DSP Integrated Circuits*. Academic Press, 1999.

[14] P. Pirsch, *Architectures for digital signal processing*. John Wiley & Sons, 1995.

[15] R. Jain, R. Kasturi, and B. G. Schunck, *Machine vision*. McGraw-Hill New York, 1995.

[16] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *18th International Conference on Pattern Recognition (ICPR'06)*, IEEE, vol. 3, 2006, pp. 850–855.

A Configurable OFDM Baseband Processor for RF-UOWC System-on-Chip

Trio Adiono*†, Erwin Setiawan†, Michael Jonathan†, Rahmat Mulyawan*†,
Nana Sutisna*†, Infall Syafalni*†, and Wasiu O. Popoola‡
* School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia
† University Center of Excellence on Microelectronics, Bandung Institute of Technology, Indonesia
‡ Institute for Digital Communications, University of Edinburgh, Edinburgh, EH9 3JL, UK
Email: {tadiono, erwin.s, michael.j, rahmat.mulyawan, nsutisna, infall}@itb.ac.id, w.popoola@ed.ac.uk

***Abstract*—Dual-hop radio-frequency underwater optical wireless communication (RF-UOWC) systems enable underwater communication as well as above-water communication. RF technology is used for communication above the water surface, and optical communication is used underwater. In this work, we present an OFDM baseband processor on a programmable system-on-chip (SoC) FPGA. Our register transfer level (RTL) design of the OFDM is configurable, so it can be implemented for RF as well as UOWC. We have tested our design in real-time data transmission with both RF and underwater channels. The OFDM processor for RF communication is tested on ADRV9361-Z7035 FPGAs. With a channel bandwidth of 3.2 MHz, the achievable real-time data rate is 5.4 Mbps. The OFDM processor for UOWC is tested on RFSoC 4x2 FPGAs. With a channel bandwidth of 32 MHz, the achievable real-time data rate is 54 Mbps.**

I. Introduction

The Earth's surface is 70% covered with oceans, and the demand for underwater communication systems for marine exploration has been increasing. However, the traditional acoustic communication systems for underwater communication are limited by data rates, while RF wireless communication is heavily attenuated by sea water [1]. Underwater optical wireless communication (UOWC) is an emerging technology for underwater wireless communication. UOWC provides higher data rates than the traditional acoustic communication systems [2]. UOWC is also less attenuated by seawater compared to RF wireless communications.

In recent years, various UOWC experiments have been introduced [3], [4] aiming at increasing throughput beyond 10 Gbps. However, those experiments are still carried out in laboratory conditions with waveform generators and signal analyzers that have a sampling rate of up to 100 GS/s. More practical works in UOWC are also demonstrated in [5], [6], which use FPGAs. However, those experiments are still conducted at a throughput far below 1 Gbps due to the limited sampling speed of FPGAs compared to laboratory instruments. In our recent works [7], [8], [9], we have demonstrated free-space optical transmission using FPGAs that are also integrated with the Linux TCP/IP stack. We have elaborated on one of the hardware impairments from clock offset that can affect the performance of real-time optical OFDM transmission.

Recently, researchers also combined RF technology with UOWC into a dual-hop RF-UOWC system [10]. The communication above the water surface (first hop) is carried out using RF, while the communication underwater (second hop) is carried out using UOWC as illustrated in Fig. 1. There has been a lot of research in RF-UOWC in terms of security layers [11], [12], [13]. However, to the best of our knowledge, there are no reported works yet dealing with configurable OFDM baseband processor for RF-UOWC system-on-chip.

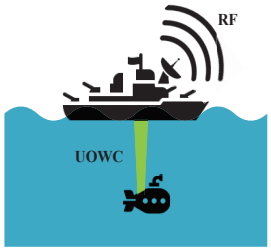


Fig. 1. An illustration of a hybrid RF-UOWC system. The above-water communication uses RF, while underwater communication uses UOWC.

Motivated by the above discussion, we consider designing a configurable OFDM baseband processor for the RF-UOWC system-on-chip. Our contributions are summarized as follows:

1) We designed a configurable OFDM baseband processor on FPGAs that can be used both for RF wireless communication and UOWC with minimal modification.
2) We evaluate the performance of the OFDM system in real-time TCP/IP transmission in the RF channel with the network and raw transmission in an underwater channel with a blue laser diode (LD) and photodiode (PD).

The remaining parts of this work are organized as follows. Section II describes the proposed system design. The experiment results are evaluated in Section III, followed by the conclusions in Section IV.

II. Proposed System Design

A. OFDM Processor Design

The most widely used OFDM in LED-based optical communication is intensity modulation/direct detection (IM/DD) that uses no carrier signal. In contrast, the OFDM for RF

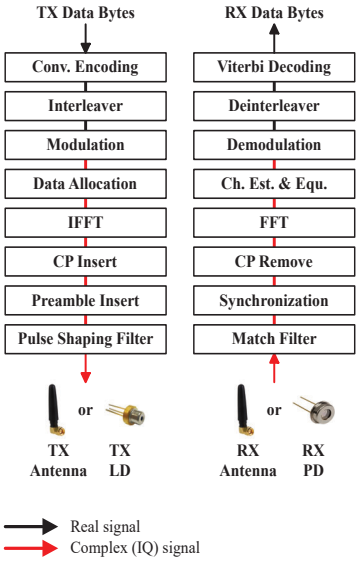


Fig. 2. Block diagram of the OFDM processor datapath.

TABLE I
LIST OF SUPPORTED MCS OF THE OFDM PROCESSOR.

MCS	Modulation	Rate
1	QPSK	1/2
2	16-QAM	1/2
3	QPSK	3/4
4	16-QAM	3/4

wireless systems is complex and bipolar. The Hermitian symmetry method is usually used to make LED-based OFDM produce real values only. However, the Hermitian method has a weakness, namely that the resulting channel bandwidth is reduced by half. Therefore, in order to achieve our goal of creating a configurable OFDM system for RF-UOWC, we consider using OFDM with a carrier signal for UOWC. As a result, the same OFDM design can be used both for RF and UOWC. This method is strandarized in the IEEE 802.11bb standard for light communication (LC) [14].

Fig. 2 shows our OFDM processor datapath block diagram. Our OFDM specification is based on the IEEE 802.16d standard. The transmitter (TX) datapath modulates TX data bytes into an OFDM frame. We implemented two types of encoding rates, namely 1/2 and 3/4, and we implemented two types of modulation, namely QPSK and 16-QAM. The modulation and coding (MCS) scheme that is used in our design is shown in Table I. Our OFDM symbol consists of 256 subcarriers consisting of 192 data points, 8 pilots, and 56 guards. On the receiver (RX) side, the datapath demodulates the OFDM frame to RX data bytes. We use synchronization to synchronize time and frequency. We use least squares channel estimation and the minimum mean squares (MMSE) equalizer. The input and output of the datapath are complex signals that will be further upconverted or downconverted before being connected to either an antenna or LD/PD.

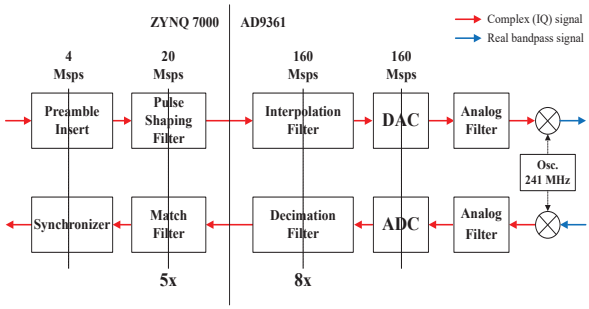


Fig. 3. ADRV9361-Z7035 digital-to-analog interface block diagram.

B. RF-UOWC System Configuration

There are two types of FPGA boards used in our work, namely ADRV9361-Z7035 for RF OFDM verification and RFSoC 4x2 for RF UOWC verification. The ADRV9361-Z7035 uses the Xilinx Zynq 7000 SoC chip, while the RFSoC 4x2 uses the Xilinx Zynq UltraScale+ RFSoC chip. The ADRV9361-Z7035 TX and RX paths support carrier frequencies from 70 MHz to 6.0 GHz, while the RFSoC 4x2 TX and RX paths support carrier frequencies from 0 to 6.0 GHz. We use a carrier frequency of 241 MHz for RF transmission and 49 MHz for UOWC transmission.

Fig. 3 shows the digital-to-analog system for the ADRV9361-Z7035 board. The signal bandwidth for the RF system is 3.2 MHz. We use a baseband clock frequency of 20 MHz. The pulse shaping filter on the TX interpolates the OFDM signal from 4 MSPS to 20 MSPS, and the match filter on the RX decimates the OFDM signal back from 20 MSPS to 4 MSPS. The interpolation filter on AD9361 interpolates the OFDM signal from 20 MSPS to 160 MSPS, and the decimation filter on AD9361 decimates the OFDM signal back from 160 MSPS to 20 MSPS. The sampling rate of the DAC and ADC is 160 MHz. The frequency cutoff of the AD9361's analog filter is set to 3.5 MHz. The carrier frequency is set to 241 MHz. On the RX path of the AD9361, there is an analog gain controller (AGC) with a maximum gain of 77 dB.

Fig. 4 shows the digital-to-analog system for the RFSoC 4x2 board. The signal bandwidth is 32 MHz. We use a baseband clock frequency of 200 MHz. The pulse shaping filter interpolates the OFDM signal to 200 MSPS, and the match filter decimates the OFDM signal back to 40 MSPS. The interpolation filter on RFDAC interpolates the OFDM signal to 3.2 GSPS, and the decimation filter on RFADC decimates the OFDM signal back to 200 MSPS. The sampling rate of the DAC and ADC is 3.2 GHz. The carrier frequency is set to 49 MHz. There are programmable numerically controlled oscillators (NCOs) and mixers inside the RFDAC and RFADC blocks that convert the baseband signal to the bandpass signal. However, on this FPGA board, there is no built-in AGC.

III. EXPERIMENT RESULTS

A. Hardware Resource

Table II shows the implementation results for the OFDM baseband processor. For the RF baseband processor, the syn-

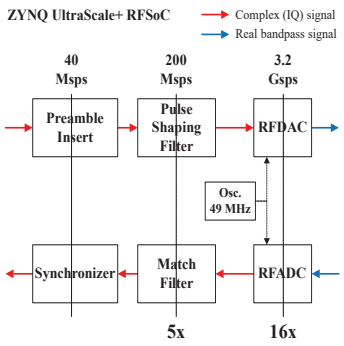


Fig. 4. RFSoC 4x2 digital-to-analog interface block diagram.

TABLE II
FPGA UTILIZATION OF THE OFDM PROCESSOR SOC.

Resource	RF System	UOWC System
LUT logic	28068	23894
LUT memory	1630	1652
FF	28416	25434
BRAM	93	60
DSP	51	51
Clock frequency	20 MHz	200 MHz
Power	0.873 W	0.776 W

thesis target is the Xilinx xc7z035ifbg676-2L device at an operating frequency of 20 MHz, while for the UOWC baseband processor, the synthesis target is the Xilinx xczu48dr-ffvg1517-2-e device at an operating frequency of 200 MHz. The result is that the RF system requires more resources compared to the UOWC system. This is justified due to the differences in the AXI interfaces and DAC/ADC interfaces used. The power consumption reported by the Xilinx tools for the RF system is 0.873 W, while for the UOWC system it is 0.776 W.

B. Experiment Setup

Fig. 5 shows our RF experiment setup. Two ADRV9361-Z7035 are used in the setup for TX and RX. We use 1/4 wave antennas that work at 241 MHz for downlink and 261 MHz for uplink. The distance between the boards is around 1.5 m. The host board is connected to a router that has an internet connection, while the client board is connected to the user's laptop via Ethernet. This RF setup is a full-duplex point-to-point system. The OFDM baseband processor is integrated into Linux's TCP/IP stack. Therefore, we can use real applications, such as `iperf`, to evaluate the system.

Fig. 6 shows our UOWC experiment setup. Two RFSoC 4x2 boards are used in the setup for TX and RX. For TX, we use Osram PL450B 450nm 80mW LD. This LD has a very high modulation bandwidth of approximately 1 GHz. We set the bias current for this LD to 30 mA by using a constant-current bench power supply. For RX, we use the Femto HSPR-X-I-1G4-SI ultra-high-speed Si-PIN PD module. This module has a bandwidth of 10 kHz to 1.4 GHz. The distance between TX and RX is 1.5 m. The boards are connected to the laptop via Ethernet and USB. This UOWC setup is a simplex system.

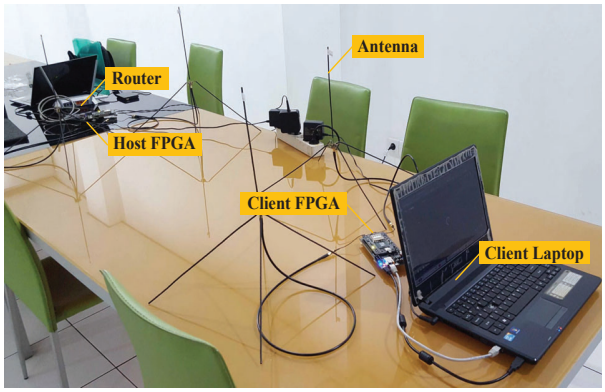


Fig. 5. Photograph of RF experiment setup with a 1/4 wave antenna at a distance of 1.5 m.

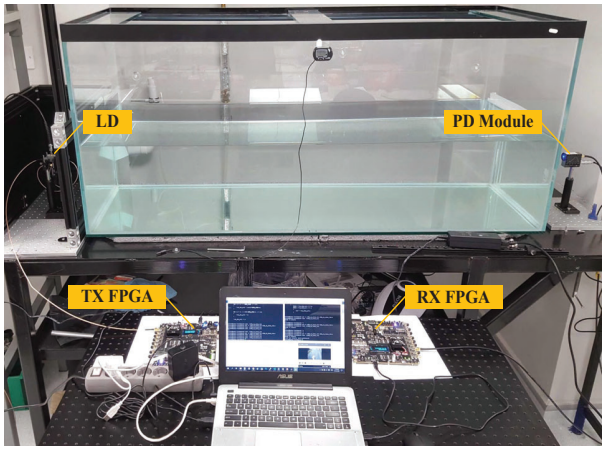


Fig. 6. Photograph of UOWC experiment setup with LD, PD, and a 1.5m water tank.

C. System Performance

We have tested the RF system without using any power amplifiers yet. The maximum output power from ADRV9361-Z7035 is 0 dBm. With this power, the maximum distance for error-free transmission is around 6 m for all MCS. The maximum real-time raw data rate is 5.4 Mbps, but after adding the Linux driver and TCP/IP stack, the maximum real-time data rate is 4.76 Mbps. Therefore, the overhead from the Linux driver and TCP/IP stack contributes to a 11.9% data rate penalty. The details of the RF system performance have been presented in [15].

We have tested the UOWC system by transmitting a video file. The video is RGB, and it has a size of 120x160 pixels. The video has a total of 120 frames. Therefore, the number of bits in the video file is 55296000. The maximum real-time data rate is 54 Mbps. The result is shown in Fig. 7. We also tested the UOWC system with turbulence. We added turbulence by adding a heater to the water tank. The scintillant index (SI) of the water is measured to be 0.188, while the normal condition without turbulence is 0.000322. The temperature of the water in the area of the heater is measured to be around 23°C, while the rest of the of the area is measured to be 21°C.



(a) (b)

Fig. 7. Screenshot of the video file after propagating through the underwater channel in (a) error-free condition and (b) non-error-free condition.

TABLE III
BER PERFORMANCE OF THE RF AND UOWC SYSTEMS.

MCS	RF BER	UOWC BER w/o turbulence	UOWC BER w/ turbulence
1	0	0	9.03e-03
2	0	3.52e-04	4.14e-02
3	0	0	3.01e-02
4	0	2.33e-03	1.57e-01

Table III shows the BER performance of the RF system, UOWC, without turbulence and with turbulence. The RF system produces a BER of 0. This can happen because the ADRV9361-Z7035 board has an AGC, which keeps the RX signal amplitude at the required limit. Meanwhile, in the UOWC system, which uses a RFSoC 4x2 board, there is no AGC. Therefore, this condition can affect the amplitude of the RX signal, which ultimately makes the BER worse. This can be seen in the UOWC system with the turbulence effect.

IV. CONCLUSIONS

In this work, we have designed a configurable OFDM baseband processor that can be used both for RF and UOWC systems. We have verified the system with the RF and UOWC channels. With a channel bandwidth of 3.2 MHz, the achievable real-time data rate is 5.4 Mbps. The OFDM design for UOWC communication is tested on RFSoC 4x2 FPGAs. With a channel bandwidth of 32 MHz, the achievable real-time data rate is 54 Mbps.

REFERENCES

[1] H. Kaushal and G. Kaddoum, "Underwater Optical Wireless Communication," *IEEE Access*, vol. 4, pp. 2169–3536, Apr. 2016.

[2] G. N. Arvanitakis *et al.*, "Gb/s Underwater Wireless Optical Communications Using Series-Connected GaN Micro-LED Arrays," *IEEE Photonics Journal*, vol. 12, no. 2, pp. 1–11, Apr. 2020.

[3] W. S. Tsai, H. H. Lu, H. W. Wu, C. W. Su, and Y. C. Huang, "A 30 Gb/s PAM4 underwater wireless laser transmission system with optical beam reducer/expander," *Scientific Reports*, vol. 9, no. 8605, pp. 1–8, Jun. 2019.

[4] C. T. Geldard, I. M. E. Butler, and W. O. Popoola, "Beyond 10 Gbps in Hostile Underwater Optical Wireless Communication Channel Conditions Using Polarisation Division Multiplexing," *Journal of Lightwave Technology*, vol. 42, no. 13, pp. 4444–4453, Jul. 2024.

[5] Y. X. Yao *et al.*, "Design of mQAM-OFDM Underwater Wireless Optical Communication System Based on LED Array," in *Proc. 2022 3rd Information Communication Technologies Conference (ICTC)*, China, May. 2022, pp. 45–50.

[6] I. B. Saksvik, V. Hassani, A. Pascoal, and A. Alcocer, "BlueLink: A Bidirectional Optical Modem for High-Speed Underwater Communication," in *Proc. 2023 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea)*, Malta, Oct. 2023, pp. 105–109.

[7] T. Adiono *et al.*, "Performance Evaluation of Sampling Clock Offset Compensation in OFDM VLC System," unpublished.

[8] T. Adiono *et al.*, "FPGA Implementation of SFO for OFDM-based Network Enabled Li-Fi System," in *Proc. 2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, Singapore, May. 2024.

[9] T. Adiono *et al.*, "A Scalable Design of a Full-Stack Real-Time OFDM Baseband Processor for Network-Enabled VLC Systems," *IEEE Access*, vol. 12, pp. 94 527–94 542, Jul. 2024.

[10] E. Illi, F. E. Bouanani, D. B. D. Costa, F. Ayoub, and U. S. Dias, "Dual-Hop Mixed RF-UOW Communication System: A PHY Security Analysis," *IEEE Access*, vol. 6, pp. 55 345–55 360, Sep. 2018.

[11] A. S. M. Badrudduza *et al.*, "Security at the Physical Layer Over GG Fading and mEGG Turbulence Induced RF-UOWC Mixed System," *IEEE Access*, vol. 9, pp. 18 123–18 136, Jan. 2021.

[12] T. Hossain, S. Shabab, A. S. M. Badrudduza, M. K. Kundu, and I. S. Ansari, "On the Physical Layer Security Performance Over RIS-Aided Dual-Hop RF-UOWC Mixed Network," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 2, pp. 2246–2257, Oct. 2022.

[13] M. Ibrahim, A. S. M. Badrudduza, M. S. Hossen, M. K. Kundu, and I. S. Ansari, "Enhancing Security of TAS/MRC-Based Mixed RF-UOWC System With Induced Underwater Turbulence Effect," *IEEE Systems Journal*, vol. 16, no. 4, pp. 5584–5595, Nov. 2021.

[14] E. Khorov and I. Levitsky, "Current Status and Challenges of Li-Fi: IEEE 802.11bb," *IEEE Communications Standards Magazine*, vol. 6, no. 2, pp. 35–41, Jun. 2022.

[15] T. Adiono *et al.*, "Performance Measurement of Real-time FPGA based OFDM System Implementation," in *Proc. 2021 International Conference on Electrical Engineering and Informatics (ICEEI)*, Malaysia, Oct. 2021.

Hammering Inspection System Using HPSS and Gradient Boosting with a Wall-Climbing Robot

Nichika Koyama* and Nari Tanabe† and Masaya Fujisawa‡
* Tokyo University of Science, Tokyo, Japan
E-mail: 4624517@ed.tus.ac.jp
† Suwa University of Science, Nagano, Japan
E-mail: nari@rs.sus.ac.jp
‡ Tokyo University of Science, Tokyo, Japan
E-mail: fujisawa@rs.tus.ac.jp

***Abstract*—This paper proposes an automatic judgment system that suppresses noise in the observed signal, which contains both hammering sound and noise, using Harmonic/Percussive Sound Separation (HPSS) to extract hammering sound, and then uses machine learning to judge whether the inspection area is healthy or damaged. The proposed method calculates hammering sounds from the observed signal using HPSS, and extracts features considering the timing of the hammering sound occurrence from the hammering sound. Subsequently, Gradient Boosting is used to automatically judge whether the area is healthy or damaged. The feature of the proposed method is that it suppresses noise while retaining the characteristic hammering sound necessary for judging healthy and damaged areas, and by extracting features considering the timing of the hammering sound occurrence, it allows for highly accurate classification.**

I. Introduction

One of the nondestructive inspection methods to determine the condition of invisible internal damage in bridges, highways and tunnels is hammering inspection. Hammering inspection involves lightly tapping the concrete structure's wall surface with a hammer and detecting internal abnormalities from the differences in the sounds produced. Recently, in Japan, there was an accident in which a ceiling panel of a tunnel fell.[1]

The cause was the detachment, loosening and corrosion of the anchor bolts, along with cracks in the protective concrete. However, the damage was overlooked because hammering inspection had not been conducted for 12 years. This accident has led to renewed recognition of the importance of hammering inspection and the appearance of three issues.

The first issue is that the noise overlaps the hammering sound, making them difficult to hear. Since the road cannot be closed during inspection, the inspection is conducted in an environment where vehicles are constantly passing. Therefore, the noise of vehicle traffic and surrounding environmental sounds is constantly heard as noise.

The second issue is that the accuracy of determining whether the inspection area is healthy or damaged varies depending on the level of skill of the operator. Experienced operators can distinguish whether the inspection area is healthy or damaged by listening to the frequency, amplitude, and timbre differences of the hammering sound. However, if the operator's ability to distinguish these sounds is not high, it is difficult to accurately determine whether the area is healthy or damaged.

The third issue is the danger to life due to hazardous inspection environments. On highways and bridges, work at high altitudes is necessary, and in tunnels, vehicles pass very close to the operators, putting them at risk.

Recently, research has been conducted using wall-climbing robots to perform hammering inspection to eliminate the third issue of life-threatening danger. However, the drive noise of the wall-climbing robots and the vacuum noise used to adhere to the wall are observed as noise, leaving the first issue unresolved. Additionally, since operators still need to listen to the hammering sound to determine whether the area is healthy or damaged, the second issue remains unresolved. Therefore, this paper proposes a system that suppresses noise suitable for the characteristics of hammering sound and noise identified through spectrogram analysis and extracts features considering the timing of the hammering sound occurrence. Gradient Boosting is then used to automatically judge whether the area is healthy or damaged.

II. Previous Method

In recent years, research has been conducted on the digitization of hammering inspection. Previous research [2][3] have focused on the features that represent the differences between healthy and damaged areas by conducting hammering inspection on bridges where deterioration was confirmed in real environments. In previous research [2], hammering inspection were performed using inspection hammers and rotary hammers, and the characteristics of acceleration and sound pressure obtained from the data were compared, demonstrating the effectiveness of the inspection method that evaluates sound pressure using a rotary hammer. In previous research [3], hammering inspection were performed on bridges with recognized deterioration and damage, specifically in areas with noticeable delamination and separation. The research verified changes in sound pressure characteristics between healthy and damaged areas, and a clear difference was observed.

Additionally, artificial damage areas were created to confirm the effectiveness of the frequency characteristics of sound pressure. As a result, it was confirmed that the frequency spectrum varies depending on the scale and condition of the damaged areas, but using a self-organizing map, judgment is possible. Moreover, research on hammering inspection using machine learning has also been reported in order to automate the judgment of healthy and damaged areas [4]. Previous research [4] compared multiple machine learning methods and neural networks for judgement. However, considering the use in the field, it is difficult to extract accurate features because the noise overlaps with the hammering sound. In previous research, noise was recorded separately and used as a reference or white noise was assumed, but noise suppression was not the main focus, leading to errors in judgment accuracy in real environments. In response, previous research [5] proposed an automatic judgment system using noise suppression with a Butterworth filter and a colored source-driven Kalman filter and a nonlinear Support Vector Machine, a type of machine learning, considering the time transition of the hammering sound for feature extraction. However, these studies could not solve the problems of life-threatening work at high altitudes, which is one of the problems with hammering inspection. Therefore, in recent years, research has been conducted on hammering inspection using drones and wall-climbing robots [6]. Hammering inspection using these robots can be conducted at high altitudes without life-threatening risks, but drive noise and vacuum noise overlap with hammering sound as noise.

Therefore, this paper proposes a hammering inspection support system with noise suppression suitable for the characteristics of hammering sound and noise identified through spectrogram analysis, a feature extraction method that extracts only hammering sound features, and Gradient Boosting.

III. Proposed Method

First, we define the observed signals obtained during hammering sound inspection. The observed signals obtained during hammering inspection are stereo audio signals, and the observed signals at time n can be expressed as $x_{L/R}(n)$, represented by the following equation:

$$x_{L/R}(n) = s_{L/R}(n) + v_{L/R}(n) \tag{1}$$

where $s_{L/R}(n)$ denotes the hammering sound signal, and $v_{L/R}(n)$ denotes the noise signal.

A. Noise suppression using HPSS

The observed spectrogram $X_{L/R}(\lambda, k)$, obtained by performing a Short-Time Fourier Transform (STFT) on the observed signal $x_{L/R}(n)$, is given by

$$X_{L/R}(\lambda, k) = \sum_{\mu=0}^{W-1} x_{L/R}(\lambda N + \mu) \cdot h(\mu) \cdot \exp\left(-j\frac{2\pi\mu k}{W}\right) \tag{2}$$

where λ is the frame number, k is the frequency bin number, μ is the bin number within the frame, W is the Discrete Fourier Transform (DFT) frame length, and N is the frame interval. A Hanning window function $h(\mu)$ is used.

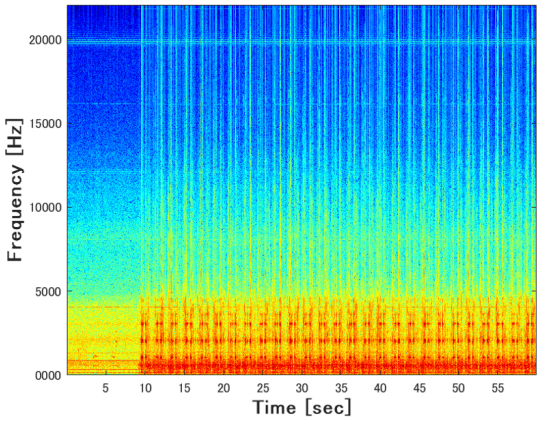


Fig. 1: Frequency analysis

From the spectrogram analysis (Fig. 1), it was found that hammering sound are continuous in the frequency direction, while noise is continuous in the time direction. Therefore, we applied Harmonic/Percussive Sound Separation (HPSS), an effective sound source separation method when such characteristics are present. However, while noise was effectively suppressed, the characteristic sounds (hereafter referred to as "characteristic hammering sound") disappeared, causing the feature distributions of sound and damaged areas to overlap, making judgment difficult.

Focusing on noise and characteristic hammering sound, it was found that noise is continuous in the time direction, while characteristic hammering sound occur intermittently in the time direction at the timing of hammering sound occurrences. This suggests that the disappearance of characteristic hammering sound is due to them being suppressed as noise since they occur intermittently in the time direction.

Additionally, the enhanced hammering spectrogram calculated in the next section uses a median filter applied to the frequency direction of the observed spectrogram. However, it was found that the relatively large characteristic hammering sound existing in a certain frequency band could converge to the median value within the range and disappear when the median filter is applied. Therefore, we determine the processing range in the frequency direction of the median filter for each frame based on the signal-to-noise (SN) ratio of the estimated hammering spectrum $\bar{S}_{L/R}(\lambda - 1, k)$ and the estimated noise spectrum $\bar{V}_{L/R}(\lambda - 1, k)$ from the previous frame. By narrowing the filter range when hammering sound are present and broadening it when they are not, we avoid applying the median filter to characteristic hammering sound.

The average previous frame SNR $\overline{\mathrm{SNR}}(\lambda, k)$ is given by

$$\overline{\mathrm{SNR}}(\lambda, k) = \frac{1}{K}\sum_{k=0}^{K-1} \mathrm{SNR}(\lambda, k) \tag{3}$$

where K is the number of samples in one frame, and

$\mathrm{SNR}(\lambda, k)$ is the length of the DFT frame:

$$\mathrm{SNR}(\lambda, k) = 20 \log_{10} \frac{\bar{S}_{L/R}(\lambda - 1, k)}{\bar{V}_{L/R}(\lambda - 1, k)} \tag{4}$$

Next, using the calculated average previous frame SNR, we compute the frequency direction processing range l_h for each frame. Since large noise often exists in the low-frequency range in the wall-climbing robot-based hammering sound inspection environment, we define the frequency direction processing range l_h as follows:

$$l_h\,[\mathrm{Hz}] = \begin{cases} f_n & (\overline{\mathrm{SNR}}(\lambda) < -5) \\ 1500 & (-5 \leq \overline{\mathrm{SNR}}(\lambda) < 5) \\ 1300 & (5 \leq \overline{\mathrm{SNR}}(\lambda) < 15) \\ 1000 & (15 \leq \overline{\mathrm{SNR}}(\lambda)) \end{cases} \tag{5}$$

where f_n is the Nyquist frequency.

By applying median filters in both the frequency and time directions, we calculate the enhanced hammering spectrogram and enhanced noise spectrogram. A median filter outputs the median of the input data, as expressed by the following equation:

$$y(n) = M\left\{x\left(n - \frac{d}{2} : n + \frac{d}{2}\right), d\right\} \tag{6}$$

where $x(n)$ is the input data and d is the filter length. the enhanced hammering spectrogram $\hat{S}_{L/R}(\lambda, k)$, which emphasizes the hammering components that are continuous in the frequency direction, is calculated by applying the median filter with the frequency direction processing range l_h obtained in Eq.(5). The enhanced hammering spectrogram $\hat{S}_{L/R}(\lambda, k)$ is given by

$$\hat{S}_{L/R}(\lambda, k) = M\left\{X_{L/R}\left(k - \frac{l_h}{2} : k + \frac{l_h}{2}\right), l_h\right\} \tag{7}$$

Next, the enhanced noise spectrogram $\hat{V}_{L/R}(\lambda, k)$, which emphasizes the noise components, is calculated by applying the median filter with an arbitrarily set filter length l_n in the time direction. The enhanced noise spectrogram $\hat{V}_{L/R}(\lambda, k)$ is given by

$$\hat{V}_{L/R}(\lambda, k) = M\left\{X_{L/R}\left(\lambda - \frac{l_n}{2} : \lambda + \frac{l_n}{2}\right), l_n\right\} \tag{8}$$

The values of the enhanced spectrograms in the current frame number λ_{now} for each enhanced spectrogram are denoted as the enhanced hammering spectrum $\hat{S}_{L/R}(\lambda_{now}, k)$ and the enhanced noise spectrum $\hat{V}_{L/R}(\lambda_{now}, k)$. Applying a Wiener filter to each enhanced spectrum, we calculate the estimated hammering spectrum $\bar{S}_{L/R}(\lambda_{now}, k)$. The estimated hammering spectrum $\bar{S}_{L/R}(\lambda_{now}, k)$ is given by

$$\bar{S}_{L/R}(\lambda_{now}, k) = X_{L/R}(\lambda_{now}, k) \cdot Q_{L/R}(\lambda_{now}, k) \tag{9}$$

where the observed spectrum is $X_{L/R}(\lambda_{now}, k)$, and the estimated hammering coefficient $Q_{L/R}(\lambda_{now}, k)$ is given by

$$Q_{L/R}(\lambda_{now}, k) = \frac{\hat{S}_{L/R}(\lambda_{now}, k)}{\hat{S}_{L/R}(\lambda_{now}, k) + \rho \cdot \hat{V}_{L/R}(\lambda_{now}, k)} \tag{10}$$

where ρ is the noise suppression parameter. Finally, by adding phase information to the estimated hammering spectrum $\bar{S}_{L/R}(\lambda_{now}, k)$ obtained in Eq.(9) and performing an inverse short-time Fourier transform (ISTFT), we calculate the estimated hammering signal $\overline{s}_{L/R}(n)$.

B. Feature extraction

Next, we extract features from the estimated hammering signal $\overline{s}_{L/R}(n)$. In this study, we used spectral peaks and their frequencies as features. However, since the rotary hammer used generates impacts in an arc, there is a pause when the hammer stops at both ends of the arc. Therefore, extracting one feature per frame (0.046 s) could lead to extracting features during periods when no impact is occurring. To address this, we consider selecting one feature from multiple frames.

First, we perform short-time fourier transform (STFT) on the estimated hammering signal $\overline{s}_{L/R}(n)$ to calculate the estimated hammering spectrum $\overline{S}_{L/R}(\lambda_{now}, k)$. Next, we search for peaks in the estimated hammering spectrum $\overline{S}L/R(\lambda now, k)$ and calculate the estimated hammering spectrum peak P_{λ_{now},k_p} and its frequency f_{λ_{now},k_p}. The estimated hammering spectrum peak is given by:

$$P_{\lambda_{now},k_p} = \max\left\{\hat{S}_{\lambda_{now},1}, \hat{S}_{\lambda_{now},2}, \cdots, \hat{S}_{\lambda_{now},K}\right\} \tag{11}$$

where k_p is the frequency bin number of the spectral peak.

Next, we prepare feature extraction frames of length l_T and select the maximum value $\overline{P}_\tau$ of the estimated hammering spectrum peaks and its frequency $\overline{f}_\tau$ from these frames to form the feature vector $\boldsymbol{\theta}_\tau$. Here, τ is an index that increments for each feature extraction frame l_T. The maximum value of the estimated hammering spectrum peaks for the feature extraction frame l_T is given by:

$$\overline{P}_\tau = \max\left\{P_{\lambda_i,k_{p_i}}, P_{\lambda_i,k_{p_{i+1}}}, \cdots, P_{\lambda_i,k_{p_{i+l_T}}}\right\} \tag{12}$$

C. Classification of healthy and damaged areas using Gradient Boosting

To classify healthy areas as 1 and damaged areas as 0, we define the class label t_τ and combine it with the extracted feature vector $\boldsymbol{\theta}_\tau$ to create the training data $\boldsymbol{r}_\tau$, as follows:

$$\boldsymbol{r}_\tau = \{t_\tau, \boldsymbol{\theta}_\tau\} = \{t_\tau, (f_\tau, \overline{P}_\tau)\} \tag{13}$$

We use Gradient Boosting to classify healthy and damaged areas using training data $\boldsymbol{r}_\tau$. First, we calculate the initial model $F_0(\boldsymbol{\theta})$ using:

$$F_0(\boldsymbol{\theta}) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{N} L(t_i, c) \tag{14}$$

where N is the number of samples and $L(t_i, c)$ is the cross-entropy loss function, given by:

$$L(t_i, c) = -[t_i \log(c) + (1 - t_i) \log(1 - c)] \tag{15}$$

Next, we calculate the residuals $r_i^{(m)}$ between the predicted labels and the true labels of the $m-1$-th model $F_{m-1}(\boldsymbol{\theta})$

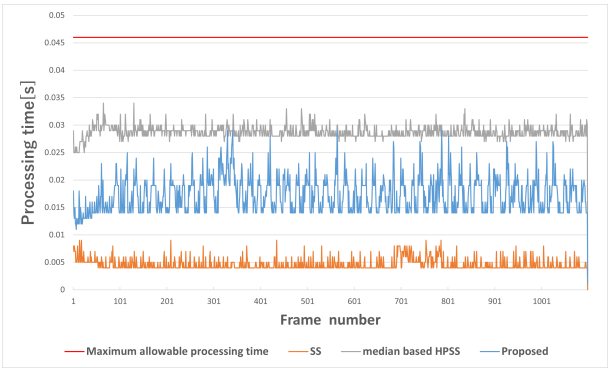


Fig. 2: Processing speeds

using:

$$r_i^{(m)} = -\left[\frac{\partial L(t_i, F(\boldsymbol{\theta}))}{\partial F(\boldsymbol{\theta})}\right]_{F(\boldsymbol{\theta})=F_{m-1}(\boldsymbol{\theta})} \tag{16}$$

Then, a decision tree h_m is fitted to the dataset $\{(\boldsymbol{\theta}_i, r_i^{(m)})_{i=1}^n\}$. We update the model by adding the new model to the $m-1$-th model:

$$F_m(\boldsymbol{\theta}) = F_{m-1}(\boldsymbol{\theta}) + \alpha h_m(\boldsymbol{\theta}) \tag{17}$$

where α represents the learning rate. By updating the model $M-1$ times in this manner, we obtain the final strong classifier $F_M(\boldsymbol{\theta})$. In this study, we use LightGBM[7], which offers fast processing speeds.

IV. Simulation

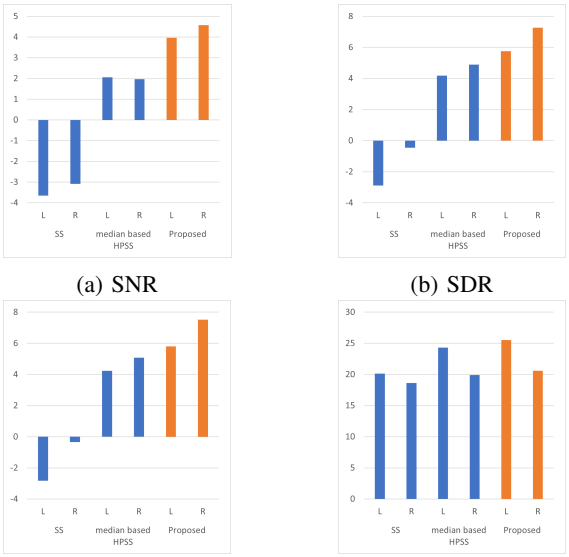
In the simulation, we compare the proposed method with previous methods in terms of noise suppression performance and classification performance. Noise suppression performance is evaluated by measuring processing speed and assessing noise suppression degree and hammering sound quality. Classification performance is evaluated by extracting features from each noise-suppressed hammering signal and comparing evaluation metrics.

A. Noise suppression performance evaluation

In the noise suppression performance evaluation experiment, we compare the proposed method with previous methods such as Spectral Subtraction (SS) [8][9] and Harmonic/Percussive Sound Separation (HPSS) [10][11], to demonstrate the effectiveness of the proposed method.

In the processing speed evaluation experiment, we measure the time required to estimate the hammering signal from the observed signal within one frame. If the proposed method and previous methods process within the maximum allowable processing time per frame (= 0.046 s), they are considered capable of real-time processing. Fig. 2 shows the processing speed evaluation. From Fig. 2, it can be confirmed that the processing time of the proposed method, indicated by the blue line, is below the maximum allowable processing time (= 0.046 s), demonstrating its capability for real-time processing.

Next, we evaluate noise suppression performance using SNR (Average Signal to Noise Ratio), SDR (Average Source to Distortion Ratio), SAR (Average Source to Artifacts Ratio), and SIR (Average Source to Interferences Ratio)[12]. The results of these evaluations are shown in Fig. 3. Here, SNR represents the ratio of the target sound to noise, SDR represents the quality of the target sound, SAR represents the minimal distortion caused by the separation process, and SIR represents the degree of separation. Higher values for each metric indicate better performance.



(a) SNR (b) SDR

(c) SAR (d) SIR

Fig. 3: Evaluation

TABLE I: Healthy area

	healthy area		
	Precision	Recall	F1-score
No processing	0.81	0.81	0.81
SS	0.79	0.73	0.76
Median based HPSS	0.76	0.87	0.81
proposed	0.86	0.93	0.89

TABLE II: Damaged area

	damaged area		
	Precision	Recall	F1-score
No processing	0.80	0.81	0.80
SS	0.74	0.80	0.77
Median based HPSS	0.85	0.72	0.78
proposed	0.92	0.85	0.89

From Fig. 3(a), 3(b), 3(c), and 3(d), it is evident that the proposed method outperforms previous methods in all evaluation metrics. This confirms that the proposed method provides superior noise suppression, hammering sound quality, minimal distortion, and degree of separation compared to previous methods.

B. Classification performance evaluation

For classification performance evaluation, we used cross-validation. The experimental data included 50-second WAV files collected using a wall-climbing robot, with three samples each for healthy and damaged areas. After noise suppression,

we compared precision, recall, F1-score, and accuracy for feature extraction and LightGBM application to verify the effectiveness of the proposed method. The results of the classification performance evaluation are shown in TABLE I and II. From TABLE I and II, the proposed method showed higher precision, recall, and F1-score compared to no processing, SS method, and median-based HPSS. Additionally, accuracy was 0.89 for the proposed method, 0.81 for no processing, 0.77 for the SS method, and 0.80 for median-based HPSS. Thus, the proposed method achieved the highest values for all classification evaluation metrics and was able to classify healthy and damaged areas with an average accuracy of 89%.

V. Conclusions

This paper proposed a Hammering Sound Inspection System Using HPSS and Gradient Boosting with a Wall-Climbing Robot and demonstrated its effectiveness through noise suppression and classification performance evaluations.

The proposed method has two main features. First, it determines the frequency direction median filter application range in the median-based HPSS from the SNR of previous frames, enabling noise suppression while preserving distinctive hammering sound present in the observed signal. Second, it performs feature extraction considering the timing of hammering sound generation, avoiding the extraction of features from periods when no hammering sound is occurring. The effectiveness of the proposed method was demonstrated by its superior performance in both noise suppression and classification metrics compared to previous methods.

References

[1] Research and R. Committee, "Report of the investigation and review committee on the falling tunnel ceiling plate accident," 2013.

[2] M. Omagari, Y. Sonoda, and O. Munemoto, "Fundamental study on rotational impact inspection of deteriorated concrete bridges," concrete engineering annual journal," *Concrete Engineering Annual Journal*, vol. 31, no. 1, 2009.

[3] Y. Sonoda and T. Watanabe, "Quantitative evaluation of defect detection capability in rotational impact inspection," *Journal of Structural Engineering A*, vol. 59, pp. 682–692, 2013.

[4] E. Hisao, B. Naoyuki, A. Hiromoto, and N. Yamato, "Evaluation of flaw detection in impact testing using ai techniques," *ScienAI and Dce Journal*, vol. 1, no. J1, pp. 514–521, 2020.

[5] K. Hori, N. Tanabe, and M. Fujisawa, "Nonlinear svm-type automatic dicision algorithm in noisy environment for hammering test system," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 275–281, 2021.

[6] K. Yamashita, S. Ogata, and T. Shimamura, "Improved spectral subtraction utilizing iterative processing," *Electronics and Communications in Japan*, vol. 90, no. 4, pp. 39–51, 2006.

[7] G. K. et al, "Lightgbm a highly efficient gradient boosting decision tree," *in Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] R. Martin, "Spectral subtraction based on minimum statistics," *power*, vol. 6, no. 8, pp. 1182–1185, 1994.

[9] S. K. Roy and K. K. Paliwal, "A noise psd estimation algorithm using derivative-based high-pass filter in non stationary noise conditions," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 1, p. 32, 2021.

[10] D. FitzGerald, "Harmonic/percussive separation using median filtering," *in Proc. 13th Int. Conf. Digital Audio Effects, Graz, Austria*, 2010.

[11] S. Oyabu, D. Kitamura, and K. Yatabe, "Linear multichannel blind source separation based on timefrequency mask obtained by harmonic/percussive sound separation," *in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[12] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

Implementation of Real-Time Oscillometric Based Algorithm for Blood Pressure Measurement in Patient Monitor

Trio Adiono*, Clarence Amadeus*, Sindy Novaria Cicilya Sinaga†, and Teuku Rafifsyah Thomi†
*School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia
†PT. Xirka Dama Persada, Indonesia

***Abstract*—As blood pressure is considered one of the most important metrics of human health, it is critical to have a device to measure blood pressure accurately and practically. Non Invasive Blood Pressure(NIBP) device is one of the solutions. To be able to perform well, NIBP devices need to be equipped with the right algorithm to measure blood pressure. One of the most popular blood pressure measurement algorithms is the oscillometric-based algorithm. This algorithm involves Oscillometric Waveform(OMW) extraction, envelope detection, and blood pressure estimation. In this paper, the authors want to develop and implement an oscillometric based algorithm for real-time NIBP measurement in a prototype patient monitor device. The algorithm will be tested in case-based and (Beat Per Minute) BPM-based data, and will be evaluated according to Indonesian Ministry of Health's standard. Our test yields promising result, with errors below 5mmHg**

***Keywords*—** NIBP, Oscillometric, Mean Arterial Pressure, Systolic Blood Pressure, Diastolic Blood Pressure

1. Introduction

Blood pressure is one of the most important metrics for human health. Humans can be very prone to disease if their blood pressure is too high or too low. Some of the most dangerous disease caused by blood pressure is Cardiac Arrest, Stroke, and many more. [1] To avoid this disease, periodical monitoring of blood pressure is vital. Traditionally, blood pressure is measured by medical staff using traditional medical equipment. This method requires good skill from the medical staff to be able to measure the patient's blood pressure. Another problem with this method is that this method requires a meeting between the medical staff and the patient. This is often impractical, especially in emergencies where the patient is unable to go anywhere.

To solve this problem, a simple and portable machine is developed to measure blood pressure automatically. This device is often called NIBP(Non Invasive Blood Pressure) Monitor. Such devices can be purchased easily in health equipment stores. Even some of the hospitals/clinics also use this kind of device because of its practicality [2]. Most NIBP monitor devices use an oscillometric-based algorithm to measure blood pressure. This algorithm uses a cuff that is inflated on the patient's arm to stop blood flow in the arm, then the pressure of the cuff is monitored using a high-precision pressure sensor. From the pressure data, the Oscillometric waveform can be separated and from this waveform, Mean Arterial Pressure(MAP), Systolic Blood Pressure (SBP), and Diastolic Blood Pressure (DBP) can be calculated. To be able to measure blood pressure accurately, the oscillometric algorithm needs to be tuned carefully. In this paper, the authors want to develop and implement an oscillometric based algorithm for real-time NIBP measurement in a prototype patient monitor device.

2. Previous Work

Basically, in NIBP devices, the Oscillometric based algorithm is widely used. It usually involves taking the maximum pressure in the Oscillometric Waveform as Mean Arterial Pressure (MAP), then deciding SBP and DBP based on the MAP. The algorithm to decide SBP and DBP is still an open problem, because different NIBP devices may have different approaches/algorithms. Some approaches use statistical-based method, some use DSP-based method, and some uses Neural Network Approach.

Some of the statistical-based methods involve statistical models to find the systolic and diastolic characteristic ratios. In [3] and [4], Bayesian Model is used to determine the SBP and DBP ratio, while in [5], Gaussian Mixture Model is used to estimate SBP and DBP ratio. These works try to find systolic and diastolic ratios by finding the posterior probability distribution function of these ratios. Although the ratios found using this approach is accurate, the process of calculating these statistical model can be hard and complex. This can make the blood pressure measurement time longer, which is not desired.

Another approach to estimate SBP and DBP is to use the derivative of the Oscillometric Waveform Envelope(OMWE). In [6] and [7], the maximum slope of OMW is used to estimate SBP and DBP ratios. This technique can also produce accurate SBP and DBP ratios compared to the empirically chosen ratios. The problem with this method is that the accuracy of the ratios is highly dependent on the quality of the OMWE. If the OMWE contains noise then the performance of this method can degrade significantly.

Besides using ratios, SBP and DBP can also be obtained from the oscillometric pulse itself. This can be done by evaluating the pulse shape in different cuff pressures [8]. In [9], Digital Signal Processing (DSP) based method is used to evaluate this pulse shape to obtain various pulse morphologies. This technique can lead to a better understanding of the relationship between blood pressure and OMW but requires a good understanding of the DSP concept, which is not practical. Another problem with this technique is that this technique is very noise prone because whenever the OMW is corrupted by noise, its pulse morphology can change drastically.

With the development of AI, various AI models have been proposed to estimate SBP and DBP. For example, Deep Neural Network is proposed to estimate SBP and DBP directly from OMWE in [10]. In [11], Adaptive Neuro-Fuzzy Inference System(ANFIS) is used to estimate SBP and DBP from OMWE. In [12], Principal Component Analysis (PCA) is combined with Neural Networks to estimate SBP and DBP from OMWE. These networks can estimate SBP and DBP accurately and are not sensitive to noise.

The downside of this approach is that some models may require large datasets and computational power.

Despite the success of implementing various concepts to estimate SBP and DBP, the most common and empirical way to determine SBP and DBP is to use the fixed ratio algorithm. In the fixed ratio algorithm, SBP and DBP are estimated using a predetermined fixed ratio to the MAP. This ratio is chosen usually by trial and error. This algorithm usually involves Oscillometric Waveform(OMW) extraction, envelope detection, and SBP-DBP estimation. Some papers discuss this algorithm, and their comparison can be seen in Table 1.

3. Proposed Method

In this paper, the authors used an Oscillometric based algorithm to measure blood pressure. The block diagram of this algorithm is shown in Figure 1. As seen in Figure 1, this algorithm takes the raw cuff pressure data read by the pressure sensor as an input. From this input, the following steps are performed:

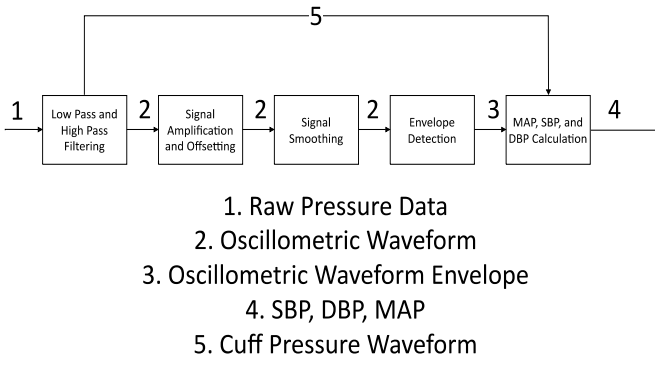


Figure 1. Block Diagram of The Algorithm Used in This Paper

1) Filtering
The main objective of this step is to separate the Oscillometric Waveform from the Cuff Pressure Waveform. To separate the Oscillometric Waveform from the Cuff Pressure Waveform, low pass, and high pass filtering is performed on the input signal. The signal is first low-pass filtered to get the cuff pressure waveform. Then the cuff pressure waveform is further filtered using the high pass filter to obtain the oscillometric waveform. The low pass filter and high pass filter cutoff frequency are tuneable parameters in this step. An example of the raw data and filtered data can be seen in Figure 2 and Figure 3, respectively.
As seen in Figure 3, the output of the low pass filter is the declining pressure of the deflating cuff. The MAP, SBP, and DBP will be determined from this cuff pressure waveform. The position of the MAP, SBP, and DBP will be obtained from the oscillometric waveform, which is the output of the high-pass filter.
2) Signal Amplification and offsetting
The oscillometric waveform obtained from the cuff pressure waveform is still very weak and it contains negative values. To make the oscillation clearer, the signal will be amplificated. After the signal is amplificated, a dc offset will be added to get rid of negative values.
3) Signal Smoothing

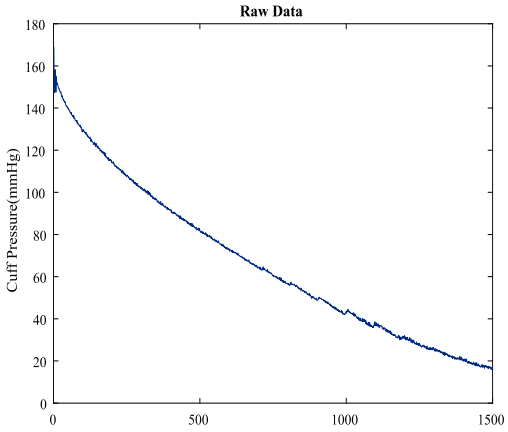


Figure 2. Raw Data of the Cuff Pressure Read by the Sensor

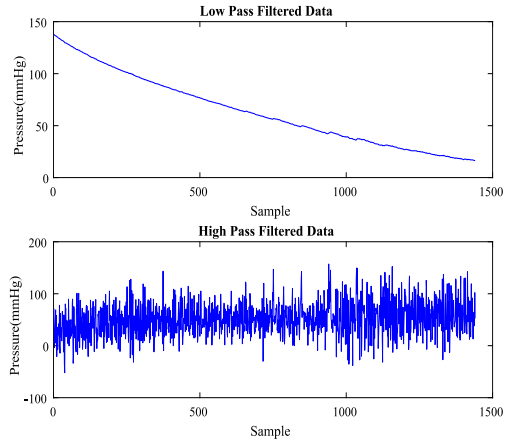


Figure 3. The Result of the Low Pass and High Pass Filter

The obtained oscillometric waveform is often still noisy, or it has too much oscillation. This will make detecting the envelope of the waveform hard. To avoid this, the signal will be smoothed using a moving average filter. The frame size and repetition of the filter are tuneable parameters in this step. The formula used in the moving average filter is shown in Equation 1.

$$y[n] = \frac{1}{N} \sum_{k=0}^{N-1} x[n-k] \tag{1}$$

where y[n] is the smoothed signal, N is the length of the signal, and k is the sumation iteration index
The comparison of the Oscillometric Waveform before and after signal smoothing can be seen in Figure 4. As seen in Figure 4, the oscillometric waveform obtained from the high pass filter is very noisy. This will cause the signal envelope hard to be obtained. Using a moving average filter, the oscillometric waveform is successfully smoothed and the signal is less noisy.
4) Envelope Detection
After the Oscillometric Waveform is obtained, the envelope of this waveform will be taken. This signal envelope is often called Oscillometric Waveform Envelope(OMWE). To obtain OMWE, an envelope detector

TABLE 1. COMPARISON OF FIXED-RATIO ALGORITHM IN DIFFERENT PAPERS

Paper	OMW extraction	Envelope Detector	SBP and DBP estimation
[13]	Filter 0.3 Hz-6 Hz	not mentioned	ratio of 0.55 and 0.85
[14]	Polinomial fitting	Cubic spline interpolation	ratio of 0.487 and 0.658
[15]	Filter 0.5Hz-1.5 Hz	Cubic spline interpolation	least square method

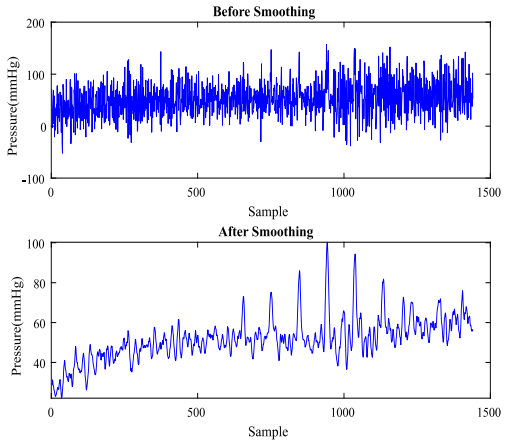


Figure 4. Oscillometric Waveform Before and After

algorithm is used. The flowchart of the algorithm is shown in Figure 5.

As seen in the flowchart, the algorithm takes the maximum point between i^{th} sample and $(i+frame)^{th}$ sample of the oscillometric waveform. After the maximum point is taken, the signal envelope is incremented until it reaches the maximum point. If the signal envelope is larger than the maximum point, the envelope signal will be decremented. After the signal envelope is obtained, it can be smoothed by the moving average filter. The frame length is the only tuneable parameter in this step.

5) MAP, Systolic, and Diastolic calculation

The last step of this algorithm is to determine MAP, SBP, and DBP, which are the value of Mean Arterial Pressure, Systolic Pressure, and Diastolic Pressure output of the measurement. The steps of the calculation are shown in Figure 6. As seen in Figure 6, the program accepts cuff pressure waveform(CPW) and Signal Envelope as input. The first step is to calculate yMAP, which is the maximum amplitude of the envelope signal. Then, the program will calculate ysys and ydia, which is the value of yMAP $\times$ rs and yMap $\times$ rd, respectively. The parameter rs and rd is a pre determined fixed ratio and will be the tuneable parameter of this step. After yMAP, ysys, and ydia is obtained, xMAP, xsys, an xdia is determined. The value of xMAP is taken as the index of yMAP. The value of xsys is taken as the index of min(yMAP-ysys), and the value of xdia is taken as the index of min(yMAP-ydia). The last step is to obtain MAP, SBP, and DBP. MAP is taken as CPW[xMAP], SBP is taken as CPW[xsys], and DBP is taken as CPW[xdia].

An example of MAP, SBP, and DBP calculation can be seen in Figure 7.

As seen in Figure 7, MAP is taken as the pressure in Cuff Pressure Waveform whose index corresponds to the

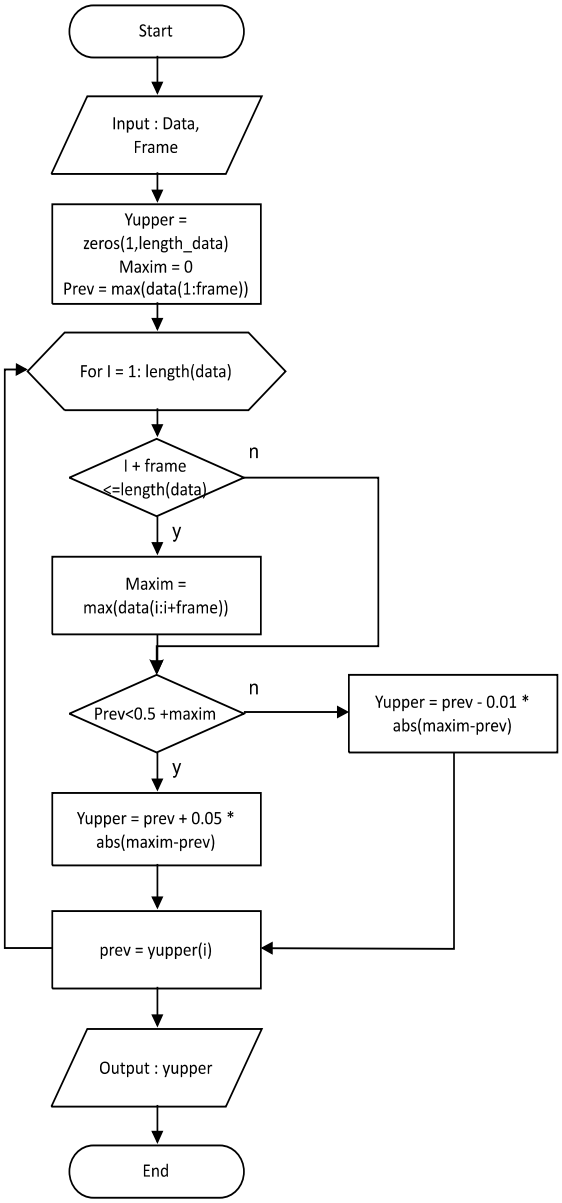


Figure 5. Envelope Detector Flowchart

maxpoint. Maxpoint is defined as the maximum pressure in the Oscillometric Waveform. SBP is taken as the pressure in Cuff Pressure Waveform which index corresponds to the pressure in the Oscillometric Waveform which is closest to sysratio*maxpoint. DBP is taken as the pressure in Cuff Pressure Waveform which index corresponds to the pressure in the Oscillometric Waveform which is

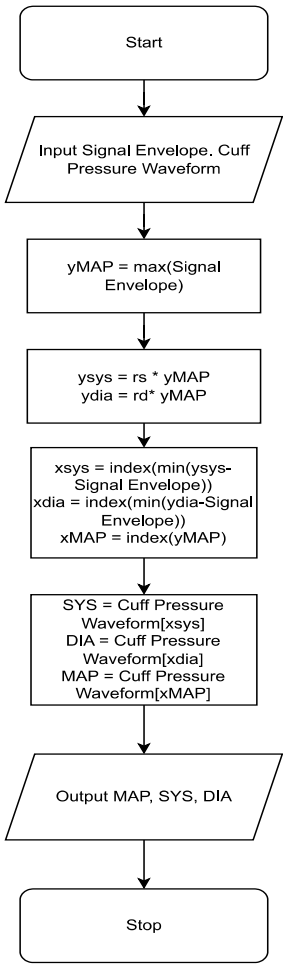


Figure 6. MAP,SYS,DIA Calculation Flowchart

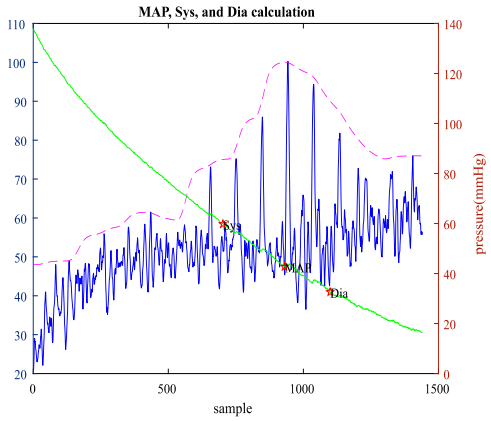


Figure 7. An Example of MAP, SBP, and DBP calculation

closest to diaratio*maxpoint. In this case, MAP is 42 mmHg, SBP is 59 mmHg, and DBP is 33 mmHg.
Although this algorithm is enough for most of the cases, sometimes there are cases such as shown in Figure 8. MAP is 211, SBP is 255, and DBP is 79. It can be seen that DBP is very wrongly estimated because MAP is 215, SBP is 255, and DBP should be 186. To minimize this kind of error, the determination of SBP and DBP can be limited to a certain part of the signal, centered at the maxpoint of the envelope signal. This introduces a new parameter p to determine how much portion to the left and right of the signal envelope is to be considered in SBP and DBP estimation. The example of SBP and DBP estimation of the case in Figure after using the p parameter is shown in Figure 9. Therefore, the tuneable parameters in this step are SysRatio, DiaRatio, and p.

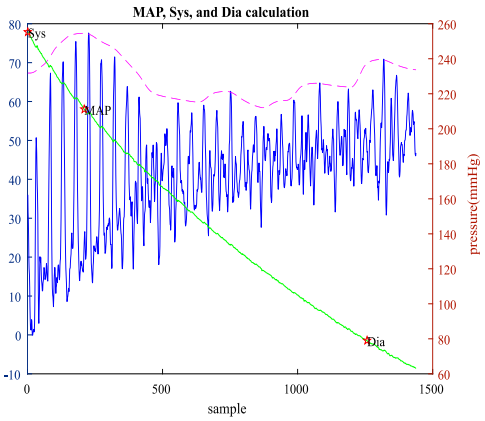


Figure 8. An Example of DBP Estimation Error

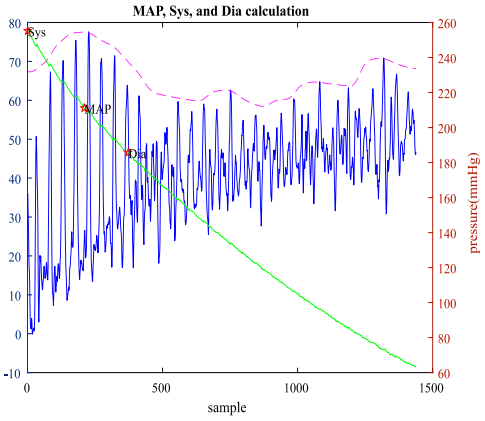


Figure 9. An Example of DBP Estimation After Using p Parameter

3.1. Parameter Selection

In the algorithm that is used in this paper, several tuneable algorithms are available. The value of each parameter that is used in this paper is shown in Table 2

3.2. Dataset Creation

The dataset used in this paper to evaluate the accuracy of the signal is made using CONTEC MS200 NIBP Simulator. The machine can be set to produce a cuff pressure waveform of a certain BPM, Systolic, and Diastolic blood pressure. The pressure produced by this calibrator is read using a pressure sensor, controlled by an ESP32 microcontroller. Finally, the data is printed into the serial monitor and saved to a txt file. The Systolic and Diastolic variation in this dataset can be seen in Table3. Each variation is

TABLE 2. PARAMETER OF THE ALGORITHM

parameters	value
low pass filter cutoff	20Hz
high pass filter cutoff	0.3Hz
moving average filter frame	10 , 5
envelope frame length	100
moving average filter frame	40
p	0.2*data length
r_s	0.75
r_d	0.9

taken for 30,60,120, and 180 BPM, and is repeated 6 times, so in total the dataset has 168 data.

TABLE 3. DATASET DISTRIBUTION

Case ID	Systolic(mmHg)	Diastolic(mmHg)	MAP(mmHg)
1	60	30	40
2	80	50	60
3	100	65	76
4	120	80	93
5	150	100	116
6	200	150	166
7	255	195	215

3.3. Algorithm Implementation

The program implementation is divided into two phases, which are algorithm development and real time implementation of the algorithm. The details of those phases will be explained in the next subsubsections.

3.3.1. Algorithm Development. The algorithm is developed using MATLAB R2023a. MATLAB is chosen because of its versatility in signal processing applications and to make it easier to plot and analyze the signal in each step of the algorithm. Although MATLAB has its library and toolbox to implement various DSP algorithms, in this paper the authors try to minimize the usage of such functions. This is done to make it easier to port the algorithm into other languages, for example, C language.

3.3.2. Real-Time Implementation. The algorithm is implemented in a prototype patient monitor device developed using ESP32. Block Diagram of the system can be seen in Figure 10. As seen in Figure 10, the system consists of a ESP32 microcontroller, a mprls0300yg00001b pressure sensor, 2 valves and valve drivers, a motor driver and a motor pump. The ESP32 controls the system by reading the pressure sensor, turning the pump on/off, and opening/closing the valve. This system is implemented for a prototype patient monitor that can be seen in Figure 11, while an example of the patient monitor display is shown in Figure 12.

4. Result and Discussion

In this section, the testing result will be displayed and discussed

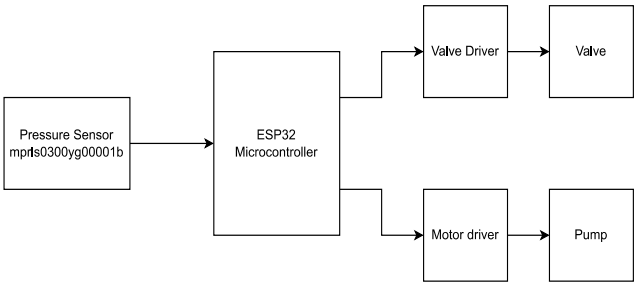


Figure 10. Block Diagram of the Patient Monitor



Figure 11. The Developed Patient Monitor

4.1. Algorithm Performance

The algorithm is tested in two ways. One way is to test the algorithm BPM based, and the other way is to test the algorithm Case-Based. These two methods are done to see if the algorithm is robust in terms of case and BPM. If the algorithm's performance is different in each case, then different parameters need to be adjusted in each case/BPM. The BPM and case-based performance of the algorithm can be seen in Table 4 and 5, respectively.

TABLE 4. BPM-BASED PERFORMANCE OF THE ALGORITHM

BPM	MAP Error(mmHg)	Sys Error	Dia Error
30	3.1667	4.0476	2.8333
60	3.0476	3.1905	2.3810
120	3.1429	4.0952	3.2143
180	3.4762	2.8571	2.5238

4.2. Discussion

According to the document issued by Indonesian Health Ministry on Bed Side Monitor Device testing, the maximum difference/error of the blood pressure reading is +-5 mmHg. This standard is taken to evaluate the performance of the algorithm.



Figure 12. The Developed Patient Monitor

TABLE 5. CASE-BASED PERFORMANCE OF THE ALGORITHM

Case	MAP Error(mmHg)	Sys Error	Dia Error
1	2.9583	3.8333	2.3750
2	3.1250	4.5	1.625
3	2.875	3.2917	2.7083
4	3.25	2.2083	2.3333
5	3.4167	4.8333	3.2083
6	3.5	4.375	3.5
7	3.4167	3.125	3.4167

Table 4 and 5 shows that in each BPM and case, no average error is more than 5 mmHg. This means that the algorithm works well in each BPM and each case. Hence, no further parameter tuning is needed.

5. Conclusion

In this paper, a algorithm for calculating blood pressure using NIBP method is proposed. The algorithm is suitable for real-time NIBP measurement. To evaluate the algorithm, it is then tested using a database created using the NIBP machine calibrator according to the Indonesian Ministry of Health standard. The algorithm manages to read Systolic and Diastolic blood pressure with mean error below +-5 mmHg, which satisfies the condition required by the Indonesian Health Ministry. This performance is valid for 7 different blood pressure cases and 4 different BPM, which shows the robustness of this algorithm. To further improve the work done in this paper, this algorithm can be implemented to measure a real patient's blood pressure. This will further test the algorithm's robustness, as well as the complexity of the algorithm.

Acknowledgment

This work was supported in part by Indonesia Endowment Fund for Education Agency (LPDP) under Grant PRJ-50/LPDP/2020, S-248/LPDP.4/2021, and S-446/LPDP.4/2023.

References

[1] *Clinical methods: The history, physical, and laboratory examinations*, 3rd ed. LexisNexis UK, 1990.

[2] P. S. Lewis, "Oscillometric measurement of blood pressure: A simplified explanation. a technical note on behalf of the british and irish hypertension society," *Journal of Human Hypertension*, vol. 33, no. 5, pp. 349–351, 2019.

[3] S. Lee, G. Jeon, and G. Lee, "On using maximum a posteriori probability based on a bayesian model for oscillometric blood pressure estimation," *Sensors (Basel, Switzerland)*, vol. 13, pp. 13 609–23, 10 2013.

[4] S. Lee, S. Rajan, H. R. Dajani, V. Z. Groza, and M. Bolic, "Determination of blood pressure using bayesian approach," in *2011 IEEE International Instrumentation and Measurement Technology Conference*, 2011, pp. 1–5.

[5] S. Lee, J.-H. Chang, S. W. Nam, C. Lim, S. Rajan, H. R. Dajani, and V. Z. Groza, "Oscillometric blood pressure estimation based on maximum amplitude algorithm employing gaussian mixture regression," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 12, pp. 3387–3389, 2013.

[6] M. Mafi, S. Rajan, M. Bolic, V. Z. Groza, and H. R. Dajani, "Blood pressure estimation using maximum slope of oscillometric pulses," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 3239–3242.

[7] A. H. Drzewiecki G, Hood R, "Theory of the oscillometric maximum and the systolic and diastolic detection ratios," *Annals of Biomedical Engineering*, vol. 22, 01 1994.

[8] M. Forouzanfar, H. R. Dajani, V. Z. Groza, M. Bolic, S. Rajan, and I. Batkin, "Oscillometric blood pressure estimation: Past, present, and future," *IEEE Reviews in Biomedical Engineering*, vol. 8, pp. 44–63, 2015.

[9] M. Mafi, S. Rajan, M. Bolic, V. Groza, and H. Dajani, "Blood pressure estimation using oscillometric pulse morphology," *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, vol. 2011, pp. 2492–6, 08 2011.

[10] Å. Ãolak and C. Isik, "Blood pressure estimation using neural networks," 08 2004, pp. 21 – 25.

[11] M. Forouzanfar, H. R. Dajani, V. Z. Groza, M. Bolic, and S. Rajan, "Adaptive neuro-fuzzy inference system for oscillometric blood pressure estimation," in *2010 IEEE International Workshop on Medical Measurements and Applications*, 2010, pp. 125–129.

[12] ——, "Oscillometric blood pressure estimation using principal component analysis and neural networks," in *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*, 2009, pp. 981–986.

[13] R. Jaafar, H. M. Desa, Z. Mahmoodin, M. R. Abdullah, and Z. Zaharudin, "Noninvasive blood pressure (nibp) measurement by oscillometric principle," in *2011 2nd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering*, 2011, pp. 265–269.

[14] S. Kumar, S. Yadav, and A. Kumar, "Oscillometric waveform evaluation for blood pressure devices," *Biomedical Engineering Advances*, vol. 4, p. 100046, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667099222000226

[15] P. Huang, J. Fu, and Y. Zhao, "A matlab demo software for nibp measurement based on hybrid oscillometric method," in *2011 4th International Congress on Image and Signal Processing*, vol. 5, 2011, pp. 2536–2539.

Exploration Robot Based On YOLOv8 Algorithm

Infall Syafalni[†‡], Angelica Winasta Sinisuka[‡], Dwi Kalam Amal Tauhid[‡], Farrel Ahmad[‡], Muhammad Alif Putra Yasa[‡], Steven Alexander Wen[‡], Erwin Setiawan[†], Nana Sutisna[†‡], and Trio Adiono[†‡]
[†]University Center of Excellence on Microelectronics, Bandung Institute of Technology, Indonesia
[‡]School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia

***Abstract*—Incidents or natural disasters, such as earthquakes, often create hazardous environments that are inaccessible to human rescuers. The extreme-condition exploration robot is the solution to it. The robot is designed to operate in these kinds of extreme conditions to assist in search and rescue operations. Equipped with sensors and utilizing machine learning capabilities, the robot can navigate through debris, detect gases, detect humans, and transmit real-time data to the rescue team. In this work, we propose a prototype extreme-condition exploration robot. The robot is equipped with an ESP32-CAM module. The video can be streamed to the server for object detection tasks. To get faster streaming times, JPEG compression is employed. Then, the YOLOv8 algorithm is employed for object detection. The YOLOv8 model is trained with 180 epochs to be able to detect three different classes. The model reached a precision of 0.951 for all classes and 0.928 mAP50. The model surpasses our target accuracy of 75% for our application.**

I. Introduction

Incidents or natural disasters, such as earthquakes, pose significant challenges for the rescue team due to the hazardous or extreme field conditions. Traditional methods of search and rescue can be dangerous and often fail to reach survivors in time. The extreme condition exploration robot is designed to address these challenges by exploring and providing information from the extreme environments to the rescue team [1]. Robots have been used a lot to respond to disasters [2], [3]. For instance, the Hitachi IRID robot [4] was deployed during the Fukushima nuclear incident to help assess the damaged reactors. There have been many other robots developed to navigate through hazardous environments. However, they often lack the integration of machine learning capabilities, which are necessary for efficiency.

This paper presents the design, implementation, and testing of the extreme condition exploration robot. The robot is equipped with IR sensors, gas sensors, and temperature/humidity sensors, all managed by the ESP32 development board. The robot also utilizes machine learning algorithms, specifically YOLOv8 [5] for object detection, and operates via teleoperation. The main objective of this research is to develop a robust robotic system capable of helping the search and rescue operation by providing real-time data from hazardous fields in order to save lives and reduce rescue risk.

The rest of this paper is organized as follows. Section II describes the related works. Section III describes the methodology. The results are evaluated in Section IV. Finally, the conclusion is in Section V.

II. Related Works

Search and rescue robots have been a topic of significant interest due to their potential to save lives in disaster situations. Various robotic systems have been developed and deployed to assist in these operations, each contributing valuable insights and advancements in the field.

A. Robots for Disaster Response

Several studies have been presented to show the use of robotics in disaster response [6], [7], [8]. One of the most notable applications of search and rescue robots was during the Fukushima nuclear disaster. The Hitachi IRID robot [4], for example, was deployed to explore and assess the damaged reactors. This robot was equipped with cameras and radiation sensors to navigate through the hazardous environment and provide critical data to human operators hitachi. However, its ability to navigate complex debris and perform detailed inspections was limited.

B. Advancements in Machine Learning

Recent advancements in machine learning have significantly enhanced the capabilities of search and rescue robots. Algorithms such as YOLO [9] have been utilized for real-time object detection and classification, enabling robots to better identify and respond to their surroundings. As an example, YOLOv8 has been used with ESP32-CAM for a cognitive assistance system to locate personal items for dementia patients [10]. However, integrating these algorithms with robust hardware systems that can withstand extreme conditions remains an ongoing challenge.

C. Gaps and Contributions

While these systems have made substantial progress, several gaps remain. Many existing robots lack the integration of advanced sensors for detecting hazardous gases and measuring environmental conditions. Additionally, real-time data transmission and remote control capabilities are often limited.

Our extreme condition exploration robot aims to address these gaps by incorporating a comprehensive suite of sensors, including gas detectors and temperature/humidity sensors, managed by an ESP32 development board. Moreover, the integration of YOLOv8 for object detection enhances the robot's ability to navigate and analyze extreme conditions effectively. This paper presents the design, implementation, and testing of this system, contributing to the advancement of search and rescue robotics in extreme environments.

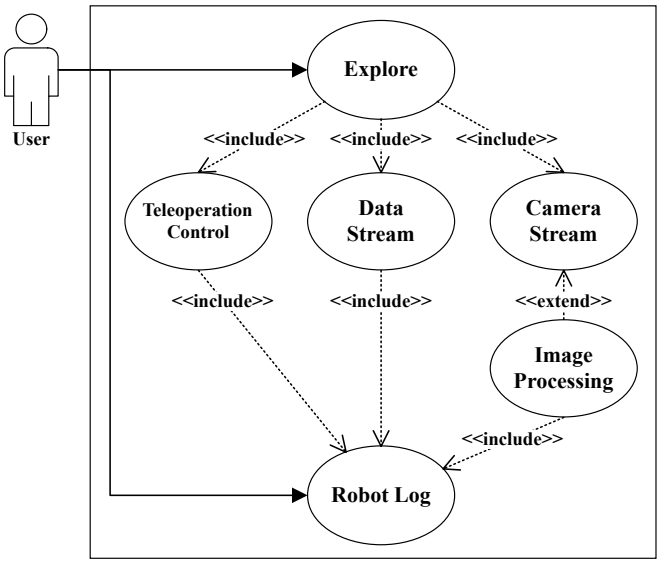


Fig. 1. Use case diagram for extreme condition exploration robot.

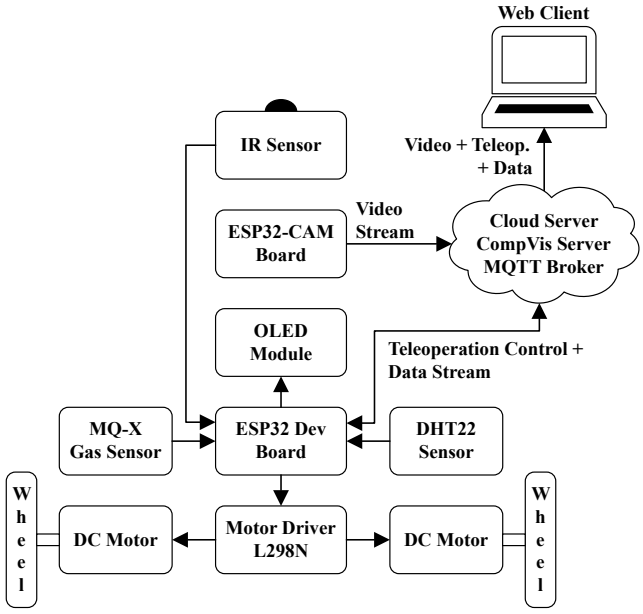


Fig. 2. The block diagram of extreme condition exploration robot.

III. Methodology

Our extreme condition exploration robot is equipped with a variety of sensors including IR sensors, gas sensors, and humidity sensors. The robot primary controls are based on the ESP32 development board.

A. Hardware

The robot hardware components include:

- **IR sensor**: for detecting obstacles and mapping the environment.
- **MQ-2 gas sensor**: to detect hazardous gases in the environment.
- **DHT22 sensor**: for measuring temperature and humidity.

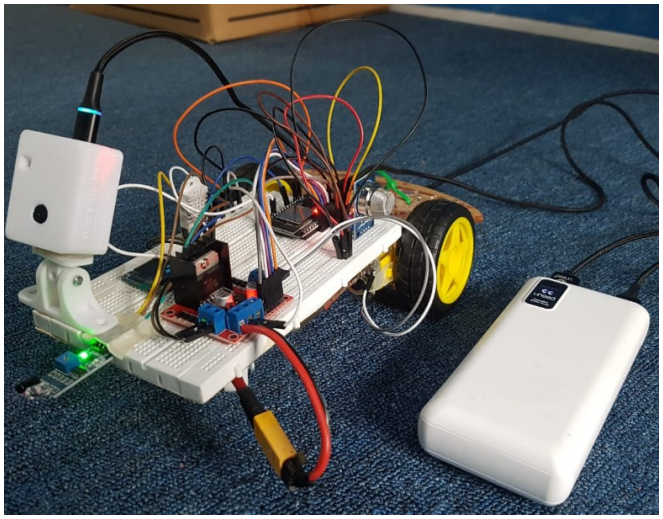


Fig. 3. Extreme condition exploration robot.

- **ESP32-CAM board**: for capturing video streams.
- **Motor driver L298N**: for controlling the DC motors.

B. Software

The software architecture includes:

- **ESP32 firmware**: manages sensor data collection, motor control, and communication with the server.
- **Machine learning**: utilizes YOLOv8 for object detection, implemented using TensorFlow.
- **Frontend**: serves as an interface for users to perform teleoperation control, view data streams visualization, and access exploration logs.
- **Backend**: provides an API for managing detected objects and exploration logs.
- **Cloud server**: hosts the data stream and provides teleoperation control via a web client. It employs Docker for containerization of the frontend, backend, and the MQTT broker, with Traffic routing the network traffic.

C. Data Collection and Analysis Method

Our research aims to identify the most effective object detection model for accurately locating three critical object categories: people, medical kits, and fire extinguishers, and to enable real-time detection. We prioritized four key performance metrics for this comparison: F1 score, precision, recall, and mean average precision (mAP). mAP was calculated at two distinct intersection over union (IoU) thresholds: 0.50, representing a moderate level of overlap between predicted and ground truth bounding boxes, and a stricter range of 0.50-0.95 to assess the model's ability to generate highly accurate detection. Given resource constraints that limit custom dataset training, we will prioritize model architectures with demonstrably fast inference times.

Each class is divided into two folders: training and validation. The training dataset comprises 80% of the images, while the validation set consists of the remaining 20%. Both

folders use the COCO dataset [11]. In order to test the model's performance on real-time data, images for validation are also captured using the ESP32-CAM. The NVIDIA T4 GPU is used as the computing resource for inference.

To determine the transmission time, the system employed the Network Time Protocol (NTP) for precise time synchronization. An NTP server provided timestamps at two crucial points: transmission initiation and data reception. By calculating the difference between these two NTP timestamps, the system accurately measured the total transmission time.

D. Use Case Diagram

The provided use-case diagram represents the functional interactions within the robot system as shown in Fig. 1. The primary actor in this system is the user, who initiates the main use case, explore. This exploratory use case has several functionalities, such as teleoperation control, data stream, camera stream, and robot log. The teleoperation control allows the user to remotely control the robot, ensuring precise navigation. The data stream manages the continuous flow of data from the robot, which is essential for real-time monitoring. The camera stream works similarly to the data stream and handles the streaming of visual data from the robot's camera. The robot log is responsible for logging all activities and data from the robot.

E. Block Diagram

Fig. 2 highlights the integration of various sensors and actuators along with the ESP32 development board to centralize the control and data flow. The ESP32-CAM Board streams video to the cloud server, where it is processed and served to the web client. The teleoperation control and data streams are managed through the cloud server, enabling the user to remotely operate and monitor the robot via the web client. This architecture ensures robust data handling and remote operability, which is crucial for extreme condition exploration.

F. Final Prototype

Fig. 3 shows the final prototype of the extreme condition exploration robot. The robot is equipped with various sensors and actuators. The robot is built on a compact chassis with two wheels driven by DC motors. The central processing unit of the robot is an ESP32 development board, which serves as the brain of the robot, managing all sensor inputs and motor controls. The ESP32-CAM is used to stream the video to the server.

IV. Results

Our tests demonstrate the robot's capabilities to navigate through simulated debris and detect objects using integrated sensors and machine learning models. The robot successfully transmitted data to the cloud server, and the web client provided an interface for remote control and data visualization.

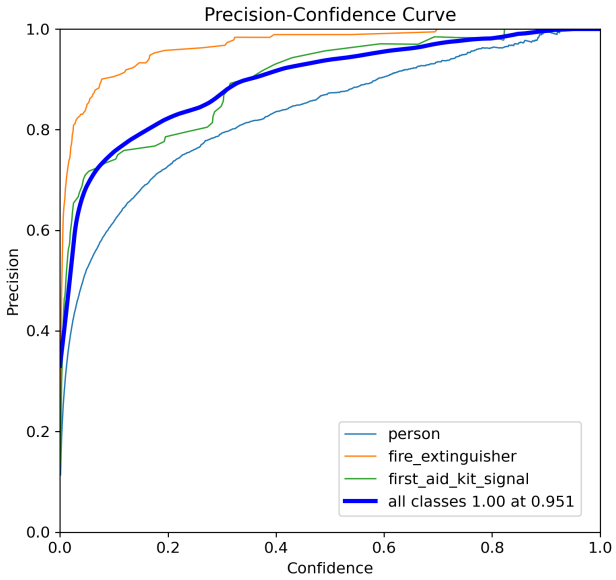


Fig. 4. Training results for a precision-confidence curve.

A. Performance Analysis

The YOLOv8 model is trained with 180 epochs. The model reached a precision of 0.951 for all classes and 0.928 mAP50 as shown in Fig. 4.

The graphs labeled train/box_loss, train/cls_loss, and train/dfl_loss in Fig. 5 all show a downward trend as the training epochs progress. This indicates that the model is learning to better predict bounding boxes, classify objects, and potentially filter out detections (dfl_loss).

In Fig. 7, the graphs labeled metrics/precision(B) and metrics/recall(B) show precision and recall for a bounding box class (B). Ideally, both precision and recall should be high. In the graphs, precision seems to be consistently higher than recall, although both metrics tend to increase slightly as training progresses.

The graphs in Fig. 6 (val/box_loss, val/cls_loss, val/dfl_loss) show similar trends to the training metrics, but with higher values.

In Fig. 8, while the validation metrics show a slight gap compared to the training metrics, indicating potential for further refinement, the model achieves a strong mAP50 exceeding 80%. This surpasses our target accuracy of 75% for our application. The detailed data for precision, recall, mAP50, and mAP50-95 for various classes is shown in Table I. These results demonstrate the model's capability for accurate object detection.

The experiment achieved low-latency communication, with a transmission time under one second. This encompasses image capture from the ESP32-CAM, inference processing, and data visualization on the user interface. Images with a resolution of 160x120 exhibited the fastest transmission times due to their smaller size, facilitated by JPEG compression.

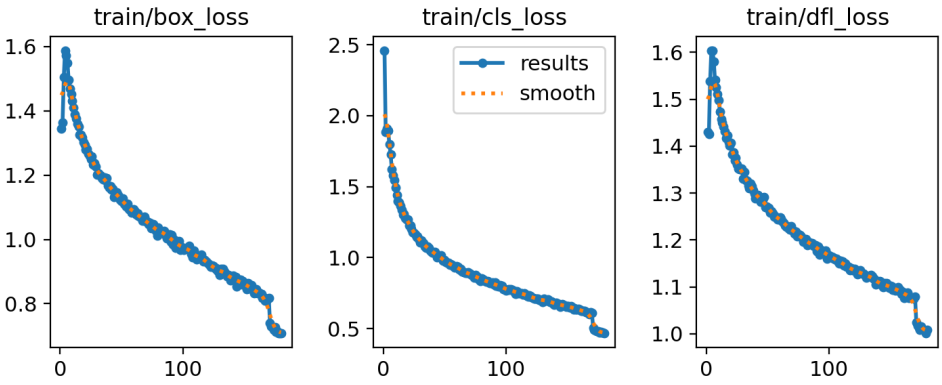


Fig. 5. Training results for train/box_loss, train/cls_loss, and train/dfl_loss.

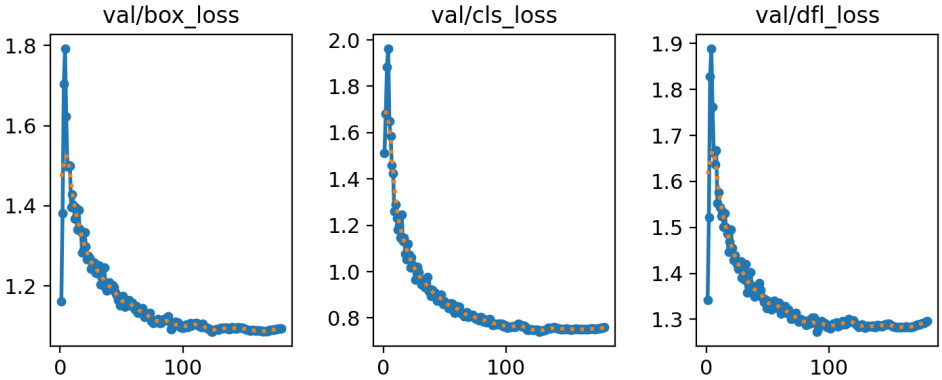


Fig. 6. Training results for val/box_loss, val/cls_loss, and val/dfl_loss.

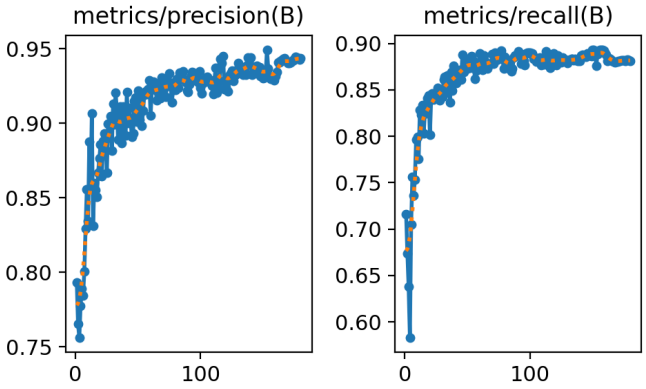


Fig. 7. Training results for metrics/precision(B) and metrics/recall(B).

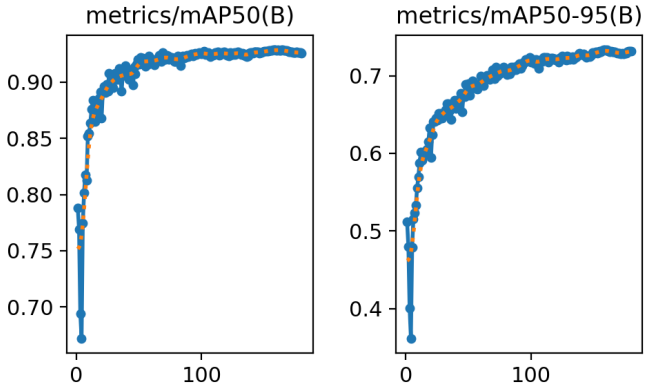


Fig. 8. Training results for metrics/mAP50(B) and metrics/mAP50-95(B).

B. Evaluation

To prioritize faster transmission times, the experiment employed JPEG compression, which introduces some image quality loss. Utilizing non-lossy image formats (e.g., PNG) could be explored to assess the impact of potential improvements in object detection accuracy. The experiment relied on a standard camera, which might be susceptible to lighting variations affecting detection performance. Employing an infrared camera could potentially improve object detection robustness in diverse lighting conditions. However, further research is needed to evaluate the trade-off between cost, power consumption, and the potential limitations of infrared cameras for this specific application. The current prototype of the extreme condition exploration robot shows promising results in terms

TABLE I
PRECISION, RECALL, MAP50, AND MAP50-95 FOR VARIOUS CLASSES.

Class	Images	Precision	Recall	mAP50/mAP50-95
all	1268	0.934	0.888	0.928/0.733
person	1268	0.860	0.699	0.806/0.536
fire extinguisher	1268	0.989	0.964	0.989/0.881
first aid kit	1268	0.953	1.000	0.990/0.783

of navigation, streaming, and data collection. However, further testing in more extreme or complex environments is necessary to validate the robot's performance in real-world scenarios.

V. CONCLUSION

The extreme condition exploration robot is a valuable tool for search and rescue operations in hazardous environments. By utilizing sensors and machine learning, the robot can provide critical information to rescue teams, which could potentially save more lives and improve the efficiency of disaster response. The YOLOv8 model is trained with 180 epochs to be The model reached a precision of 0.951 for all classes and 0.928 mAP50. The model surpasses our target accuracy of 75% for our application. Future works include improving the system robustness and implementing the system using more efficient platforms such as FPGAs and specific ICs.

ACKNOWLEDGEMENT

This work is supported by the 2024 ITB Young Researcher Program provided by LPPM, Bandung Institute of Technology under grant no. [STEI1.PN-6-04-2024].

REFERENCES

[1] R. R. Murphy *et al.*, "Search and Rescue Robotics," in *Springer Handbook of Robotics*, Berlin, Heidelberg: Springer, 2008, pp. 1151–1173.

[2] K. Nagatani *et al.*, "Emergency response to the nuclear accident at the Fukushima Daiichi Nuclear Power Plants using mobile rescue robots," *Journal of Field Robotics*, vol. 30, no. 1, pp. 44–63, Jan. 2013.

[3] X. Huang, F. Arvin, C. West, S. Watson, and B. Lennox, "Exploration in Extreme Environments with Swarm Robotic System," in *Proc. 2019 IEEE International Conference on Mechatronics (ICM)*, Germany, May. 2019, pp. 193–198.

[4] *"Robots help with decommissioning work for humans at Fukushima Daiichi Nuclear Power Station"*, IRID, https://irid.or.jp/wp-content/uploads/2022/04/IRID_ROBOTS_2022_ENG.pdf, (accessed Jul. 24, 2024).

[5] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-Time Flying Object Detection with YOLOv8," *arXiv:2305.09972*, May. 2023.

[6] S.-S. Oh, J.-H. Hahm, H.-J. Jang, S.-Y. Lee, and J.-H. Suh, "A study on the disaster response scenarios using robot technology," in *Proc. 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, South Korea, Jul. 2017, pp. 520–523.

[7] M. Chiou *et al.*, "Robot-Assisted Nuclear Disaster Response: Report and Insights from a Field Exercise," in *Proc. 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Japan, Oct. 2022, pp. 4545–4552.

[8] R. Mathew, P. Jagtap, and K. Mitra, "Ground Robots for Disaster Response: A Review," in *Proc. 2023 11th International Symposium on Electronic Systems Devices and Computing (ESDC)*, India, May. 2023.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, USA, Jun. 2016, pp. 779–788.

[10] D. K. Enna, P. Basuli, A. H. Mishra, and S. K. Singh, "Cognitive Assistance for Dementia Patients," in *Proc. 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, Bahrain, Jan. 2024, pp. 1741–1745.

[11] *"COCO Dataset"*, COCO, https://cocodataset.org/, (accessed Jul. 24, 2024).

Optimizing Deep Q-Network for Shortest Path Computation of Mobile Robot Agents

A Sumarudin*†, Nana Sutisna*†, Infall Syafalni*†, Bambang Riyanto Trilaksono†, and Trio Adiono*†
*University Center of Excellence on Microelectronics, Institut Teknologi Bandung, Bandung, Indonesia
†School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia
E-mail: 33222033@std.stei.itb.ac.id

***Abstract*—This paper proposes the application of Deep Networks (DQN) for shortest-path planning on mobile robots. Implementing DQN on mobile robots poses challenges due to the limited computational resources of embedded systems. The process consists of two modes: the training mode, where a model of DQN with weight and bias parameters is generated, and the inference mode, where the DQN model is loaded onto the robot to perform actions in a simulation maze environment with a gazebo. This study investigates optimal computational techniques for matrix operations, which are the primary operations in DQN. Computational methods performed on embedded system platforms utilizing GPUs (Jetson Xavier NX) involve studying matrix computation techniques using arrays, 2D arrays with GPUs, and tensors. Based on the experiment, the results show that for the DQN model used in shortest-path planning, computation based on an array matrix significantly speeds up the process.**

I. Introduction

Implementing Deep Q-Network (DQN) on mobile robots in real-time control applications requires several computational adjustments. The adjustments employ several embedded system-based approaches to perform matrix multiplications (MATMUL). The challenges are computational limitations on agents, especially on edge systems, where the devices are low resources CPUs, GPUs, and limited memory size [1]. This paper proposes optimizing matmul-based computations [2] to overcome the challenges. The application of DQN on robots [3] and USVs [4] are for path planning purposes. It had been ever applied in an embedded system for autonomous mobile yielded maximum delay and latency because of constraints by safety standards. The conclusion is leading to the challenges in the computational process [5]. The DQN, a mobile robot, like any deep reinforcement learning (DRL), is divided into two modes: training and inference. In the learning mode, the agent learns to achieve an optimized model in a simulation mode, while in the inference mode, the agent is tested in the real world [6]. In this research, the training mode is conducted by training the agent model within a specified environment. During the inference mode, the model yielded is loaded onto the agent. Firstly, the model is trained to achieve the model and also tested for the inference mode. Later, we transfer the calculation to a limited embedded system, and the results are evaluated by comparison analysis. Lastly, both are evaluated using a real mobile robot. The system's structure in this paper is shown in Fig. 1 and 2. Implementation DQN computation for training model DQN is shown in Fig. 1, and the implementation inference system of the system in this paper is demonstrated in Fig. 2, featuring a 10×10 maze environment. This environment is converted into a space state using one-hot encoding at the initial state where the DQN algorithm is applied; the algorithm then provides actions for the mobile robot agent to act upon.

II. DQN for mobile robot shortest path

A. Deep Q-Network design

The DQN in this paper is implemented in training and inference modes, both executed on an embedded system based on a single-board computer (SBC) refer Fig. 3. The training mode employs matrix multiplication (MATMUL) with computations performed on the embedded system; please refer to the paper [7] for the DQN algorithm and training computation shown in Fig. 1.

The developed DQN is described by equation 1 to determine the value of Q_{st+1}, which represents the temporal difference (TD) of the agent's next state action obtained from the policy generator. In this model, the policy generator uses epsilon-greedy. The loss function in equation 2 calculates the TD error between the Q value obtained from the next state and the Q value of the state when the agent takes action, serving as a reference for the learning mode. The reward for an agent is the value assigned when the agent selects an action based on the maximum value or a random action determined by the policy generator of the agent. The agent will receive a value of 10 if it reaches the goal, a value of -1 if it successfully avoids a wall, and a value of -10 if it collides with a wall.

$$Q(s_{t+1}, a_t) = R_t + \gamma \times \max_{a_t} Q(s_{t+1}, a_t) \tag{1}$$

The loss calculation uses the mean square error (MSE), with the reference value being the Q value at the next state of the agent and the reference value being the Q value at the agent's current state 2.

$$L(\theta) = \frac{1}{2} \times [Q(s_{t+1}, a_t) - Q(s_t, a_t : \theta)]^2 \tag{2}$$

During the training process, parameter settings are configured as shown in table I, with the fully connected layers consisting of input, hidden layers, and output. The output layer comprises four neurons representing the agent's actions, with the agent's actions selected based on the maximum value from the output neurons. The activation function used is sigmoid for

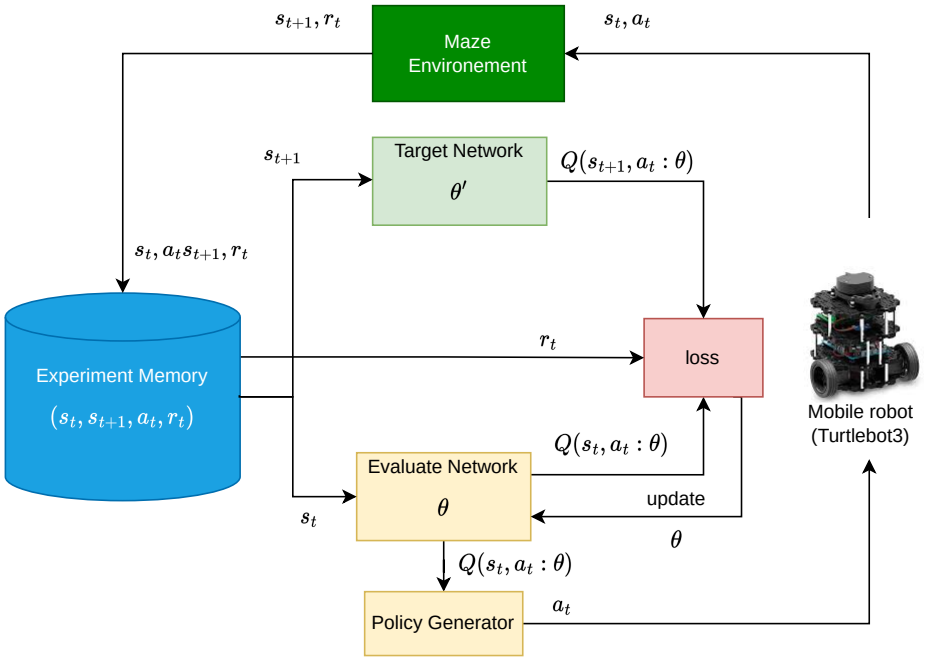


Fig. 1. DQN training mode for the shortest path.

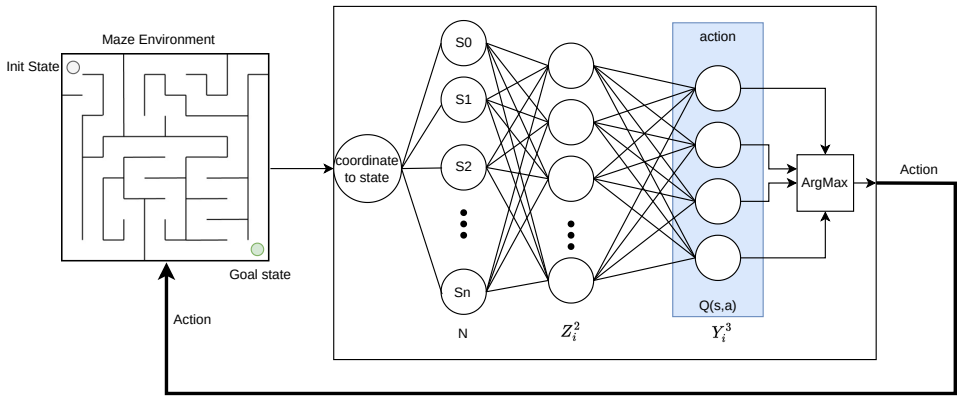


Fig. 2. DQN inference mode for the shortest path

the hidden layers and linear for the output layer. The optimizer employed is stochastic gradient descent with $\alpha = 0.1625$. The agent conducts exploration and exploitation based on an epsilon-greedy strategy with $\epsilon = 0.9$.

TABLE I
HYPER PARAMETER OF DQN

Parameters	Value
Number of layers	3
Number of neurons	100, 15, 4
Activation Functions	Sigmoid and Linear
Back Propagation	Stochastic Gradient Descent
α	0.1625
γ	0.85
ϵ	0.9

The computation process of the fully connected layer and the activation function using the sigmoid function is performed in inference mode on the robot agent. Here, j denotes the neuron in layer m, i represents the neuron in the layer, and σ refers to the activation function in the neural network layer.

$$y_j^m = \sigma\left(\sum_{i=1}^{i_n}(w_{i,j}^m \times y_i^{m-1}) + b_j^m\right) \tag{3}$$

The activation function process utilizes the exponential function of the output at layer 2. This process will be computed using both CPU and GPU on the embedded system.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{4}$$

Agent's actions are represented by specifying the agent's

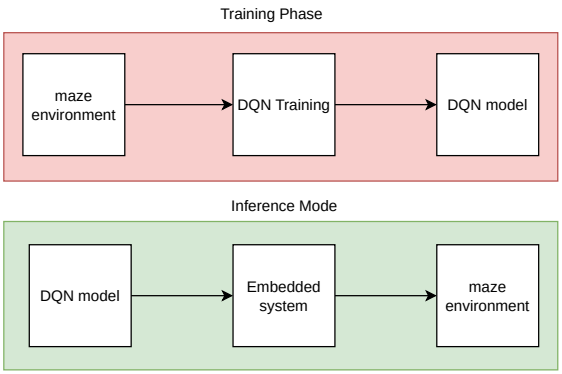


Fig. 3. Implementation mode DQN for mobile robot agent.

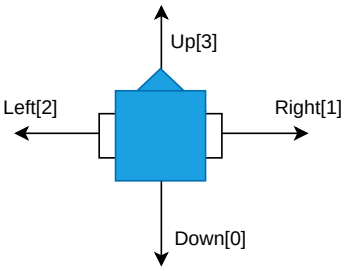


Fig. 4. Robot action with four actions.

direction within the maze, where x is an x-coordinate agent, and y is a y-coordinate agent. When the agent acts in the upward direction, the Y position is incremented by the length in meters (L) corresponding to the size of the maze cell, as well as for other robot actions as indicated in Table II and illustrated in Fig 4.

TABLE II
MOBILE ROBOT AGENT MOVEMENT.

Action	Robot movement
Down[0]	x,(y-L_{maze})
Right[1]	(x+L_{maze}),y
Left[2]	(x-L_{maze}),y
Up[3]	x,(y+L_{maze})

B. Mobile robot agent

The robot utilized in the implementation of this research is TurtleBot3. Kinematic modeling of mobile robot base on [6] and [8] as shown is given Equation 5 with x, y, and θ pose robot.

$$\begin{bmatrix} x \\ y \\ \theta \end{bmatrix} = \begin{bmatrix} -\sin\theta & 0 \\ \cos\theta & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v \\ \omega \end{bmatrix} \tag{5}$$

where v and ω are linear dan angular velocities, respectively.

The robot's pose is determined by the IMU (Inertial Measurement Unit) and the absolute rotary encoder on the mobile robot [9]. IMU utilizes a 3-axis Gyroscope and a 3-axis Accelerometer, while the absolute rotary encoder utilizes a contactless absolute encoder on the Dynamixel servo motor of the TurtleBot3.

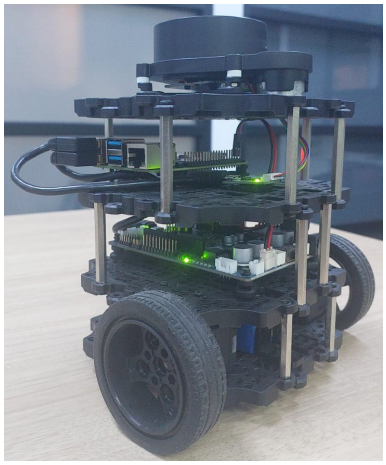


Fig. 5. A mobile robot agent uses Turtlebot3.

To convert the robot's coordinates (x and y) into states (S_t) ranging from state 0 to state n using Equation 6, where the robot's x-coordinate (x) and y-coordinate (y) are obtained from the IMU sensor, and M is the number of mazes in the robot's environment.

$$S_t = (y \times (M-1)) + x \tag{6}$$

Robot control involves three stages. In the first stage, the robot will face the target position using equation 7 with a constant angular velocity ω. The second stage, distance D in equation 8 with a constant linear velocity v, referenced by the distance and the third stage, θ from the robot [10].

$$\theta = \arctan(\frac{y_2 - y_1}{x_2 - x_1}) \tag{7}$$

$$D = \sqrt{((y_2 - y_1)^2 - (x_2 - x_1)^2} \tag{8}$$

The position control of the robot agent from the initial state (x_1, y_1) to the goal state (x_2, y_2) is illustrated in Fig. 6.

C. Computational design

Training and inference processes are conducted using an embedded system with the Jetson Xavier NX with a 20W mode and CPU with six cores. Matrix computation in Equation 3 involves matrix multiplication between the input neuron matrix and weights. For this computation, a 2D matrix multiplication in Fig. 7 approach is employed using NumPy array computation [11], executed with CUDA through the CuPy library [12], and tensor-based operations with PyTorch [13]. Optimization is performed by comparing different matrix multiplication models based on their computation time. Results from this computation are used as a reference for the robot to navigate toward the closest path to the goal within the maze.

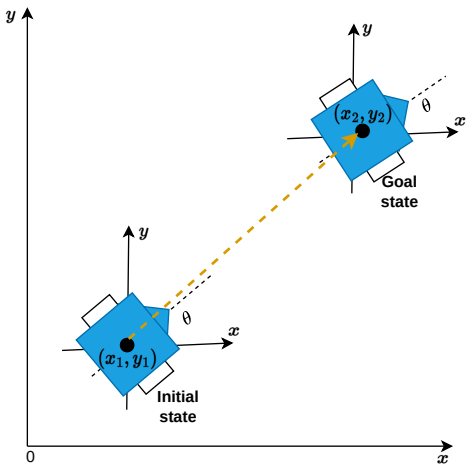


Fig. 6. Robot control position from the initial state to the goal state.

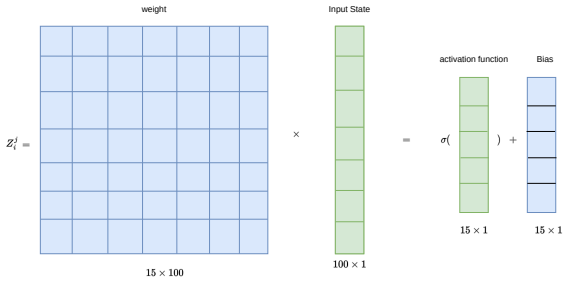


Fig. 7. Computation Matmul on DQN.

III. Experiment

A. Experiment Setup

The robot used is the TurtleBot3 in Fig. 5, which is simulated in a Gazebo with an environment consisting of a maze. The maze is a 10x10 grid, as shown in Fig. 10. Within the Gazebo maze, a rotation of $180°$ is applied to achieve the coordinates $x = 0$ and $y = 0$ at the Gazebo coordinate origin. Experimentation utilizes a simulation in Gazebo with a connection via Wi-Fi on the ROS2 Humble platform, as shown in Fig. 10. Path planning computations performed on the embedded system are transmitted via ROS2 to determine the actions of the mobile robot. The robot's position is transmitted to the embedded system through ROS topics, which define the robot's state. The robot's position serves as a reference for the Deep Q-Network (DQN) to plan the path it will follow to reach its goal by providing input to the fully connected layer of the DQN. The output from this fully connected layer is used to select the maximum value. Its argument is used to determine the action, as illustrated in Fig. 3. Actions represented by the arguments in the fully connected layer are encoded as 0 (down), 1 (left), 2 (right), and 3 (up) according to the coordinate robot in the environment, shown in Table II.

B. Experiment Results

The results of the training process are in Fig. 1 experiments for maze size environment (fig. 8) based on CPU at high-performance computing training with cumulative reward shown in Fig. 9, demonstrate that the model is capable of converging. The focus of this research is on the implementation of a robot in a 10×10 maze environment. The converged model for maze 10×10 on this study was saved with weight and bias parameters amounting to 1579 parameters refer to equation 9. Here, N represents the number of input layers, Z represents the number of neurons in the hidden layer, and Y represents the number of neurons in the output layer.

$$Param = ((N \times Z_i^2) + Z_i^2) + ((Z_i^2 \times Y_i^3) + Y_i^3) \quad (9)$$

Results of the path training with an initial state of 0 and a goal state of 99, as shown in Fig. 11, indicate that the model successfully achieved the goal with 36 actions from the robot. The matrix multiplication (MATMUL) computation in the experiment using $N \times N$ matrices was conducted, and the computation speed was measured for each computational tool on the GPU-based embedded system. The results of the matmul computation can be seen in Fig. 13. From the graph, it is observed that the use of CuPy is faster compared to other tools, with the maximum matrix size tested being 1000×1000. MATMUL operation is due to CuPy's utilization of CUDA on the GPU for parallel computation on 2D arrays, whereas PyTorch performs better when used with 3D arrays. However,

TABLE III
COMPUTATION INFERENCE PER TIME-STEP DQN FOR SHORTEST PATH MOBILE ROBOT.

Inference computing	Time(ms)
Cuppy	6.574
Torch	2.227
Numpy	1.048

when implemented in a fully connected Deep Q-Network (DQN) with a matrix size of $(100 \times 1) \times (15 \times 1)) + (15 \times 1)$ for the hidden layer and $(15 \times 1) \times (15 \times 4) + (4 \times 1)$ for the hidden layer with output, the computation performs better using NumPy as shown in Fig. 14. The inference results shown in table III on the embedded system for shortest path mobile robot in a maze environment indicate that CPU computation is more effective for implementing the DQN model in a 10x10 maze due to the small size of the neural network matrices. Numpy avoids the overhead of prepossessing between 2D arrays into CuPy arrays or PyTorch tensors.

IV. Conclusions

This paper proposes optimizing Deep Q-Network (DQN) implementation on an embedded system for shortest-path planning. The robot is implemented in two modes. First, training is conducted using a maze environment to obtain a DQN model with parameters consisting of weights and biases. In the inference mode, the model is loaded onto the

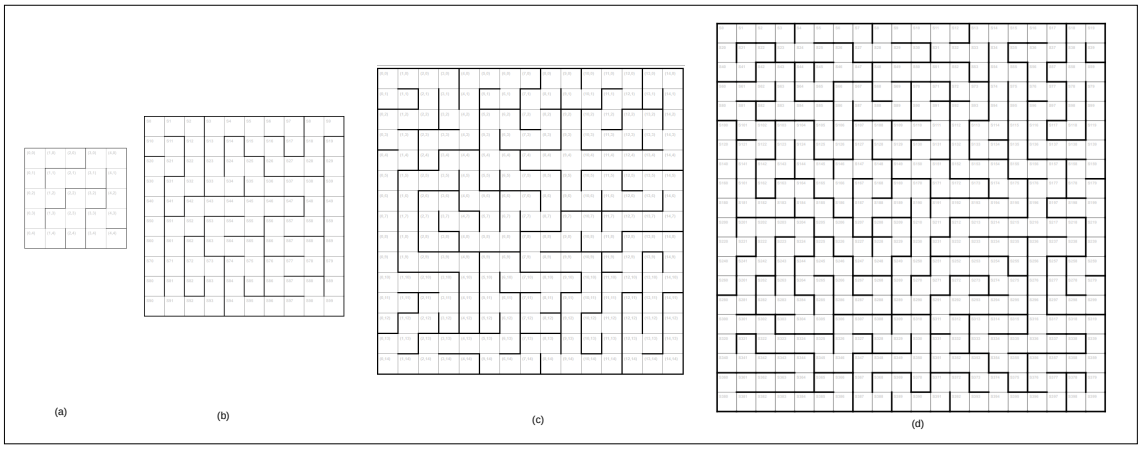


Fig. 8. Environment maze with size (a) 5×5, (b) 10×10, (c) 15×15, and (d) 20×20

.

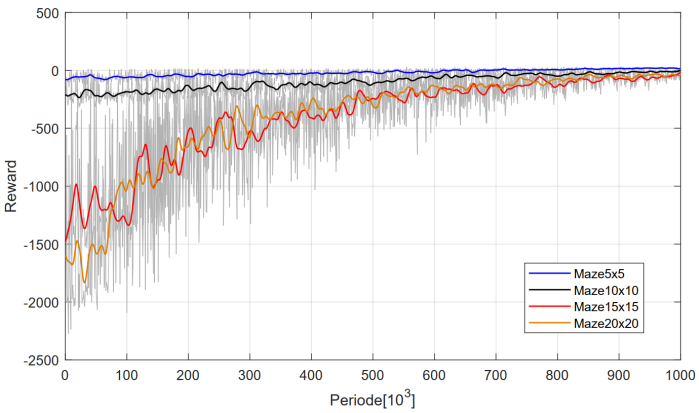


Fig. 9. Cumulative reward with varying maze sizes.

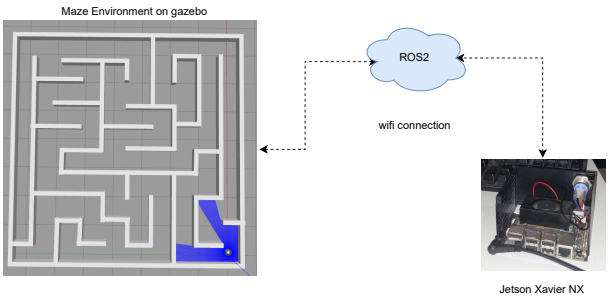


Fig. 10. DQN implementation for shortest path mobile robot agent on gazebo maze environment.

embedded system, enabling the robot to perform shortest-path planning within the maze environment. Based on the experiment, using an array matrix during the inference phase significantly improves execution time.

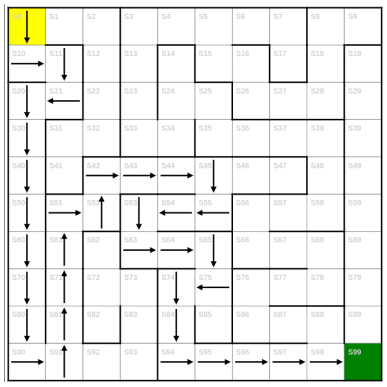


Fig. 11. Results training mode for maze size 10×10.

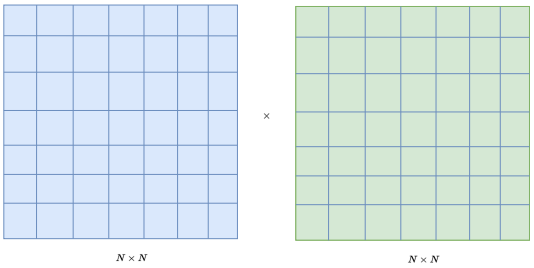


Fig. 12. Computational MATMUL $N \times N$ for computational DQN.

ACKNOWLEDGMENT

This work is sponsored in part by the Balai Pembiayaan Pendidikan Tinggi (BPPT), the Ministry of Education and Culture, Research, and Technology of the Republic of Indonesia, and the Indonesia Endowment Fund for Education Agency (LPDP) of the Republic of Indonesia, with a contract No. 00016/J5.2.3./BPI.06/9/2022.

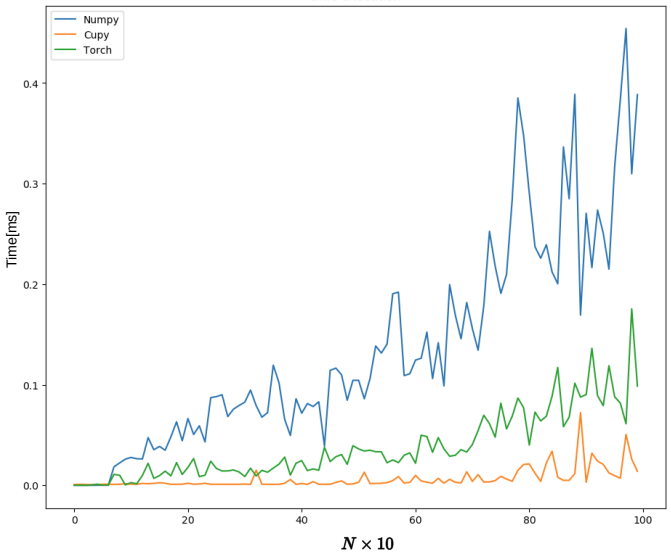


Fig. 13. Time execution MATMUL [NxN] on embedded systems.

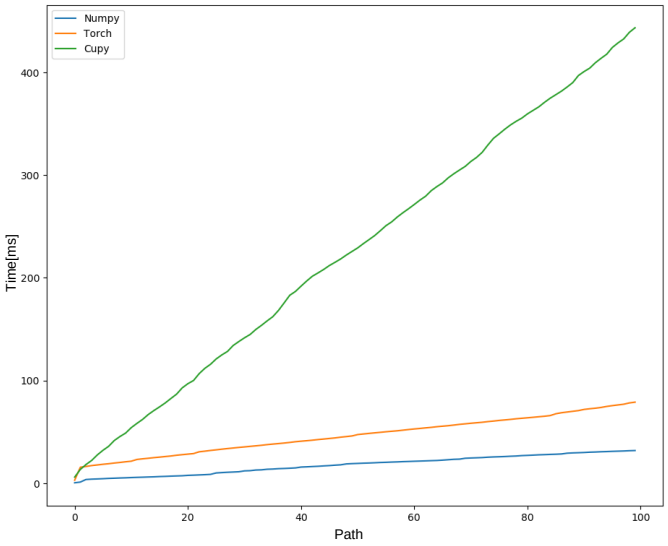


Fig. 14. Execution time of DQN inference with varying numbers of robot paths.

References

[1] M. Rothmann and M. Porrmann, "A survey of domain-specific architectures for reinforcement learning," *IEEE Access*, vol. 10, pp. 13 753–13 767, 2022. DOI: 10.1109/ACCESS.2022.3146518.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, ISSN: 00280836. [Online]. Available: http://dx.doi.org/10.1038/nature14236.

[3] S. Zhou, X. Liu, Y. Xu, and J. Guo, "A deep q-network (dqn) based path planning method for mobile robots," in *2018 IEEE International Conference on Information and Automation (ICIA)*, 2018, pp. 366–371. DOI: 10.1109/ICInfA.2018.8812452.

[4] Z. Huang, H. Lin, and G. Zhang, "The usv path planning based on an improved dqn algorithm," in *2021 International Conference on Networking, Communications and Information Technology (NetCIT)*, 2021, pp. 162–166. DOI: 10.1109/NetCIT54147.2021.00040.

[5] R. Hussain and S. Zeadally, "Autonomous cars: Research results, issues, and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1275–1313, 2019. DOI: 10.1109/COMST.2018.2869360.

[6] T. Nakamura, M. Kobayashi, and N. Motoi, "Path planning for mobile robot considering turnabouts on narrow road by deep q-network," *IEEE Access*, vol. 11, pp. 19 111–19 121, 2023. DOI: 10.1109/ACCESS.2023.3247730.

[7] A. Sumarudin, N. Sutisna, I. Syafalni, B. R. Trilaksono, and T. Adiono, "Dqn algorithm design for fast efficient shortest path system," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 254–260. DOI: 10.1109/APSIPAASC58517.2023.10317113.

[8] G. Campion, G. Bastin, and B. Dandrea-Novel, "Structural properties and classification of kinematic and dynamic models of wheeled mobile robots," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 1, pp. 47–62, 1996. DOI: 10.1109/70.481750.

[9] Y. Zhao, Y. Liang, Z. Ma, L. Guo, and H. Zhang, "Localization and mapping algorithm based on lidar-imu-camera fusion," *Journal of Intelligent and Connected Vehicles*, vol. 7, no. 2, pp. 97–107, 2024. DOI: 10.26599/JICV.2023.9210027.

[10] G. Farias, G. Garcia, G. Montenegro, E. Fabregas, S. Dormido-Canto, and S. Dormido, "Reinforcement learning for position control problem of a mobile robot," *IEEE Access*, vol. 8, pp. 152 941–152 951, 2020. DOI: 10.1109/ACCESS.2020.3018026.

[11] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[12] R. Okuta, Y. Unno, D. Nishino, S. Hido, and C. Loomis, "Cupy: A numpy-compatible library for nvidia gpu calculations," in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017. [Online]. Available: http://learningsys.org/nips17/assets/papers/paper_16.pdf.

[13] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

Leveraging IoT and Machine Learning for Efficient Rice Stock Monitoring and Prediction

Nana Sutisna[1,2†], Aditya Prawira Nugroho[1*], Christopher Jeffrey[1*], Patrick Amadeus Irawan[1*], Rizky Ramadhana[1*], Ronggur Mahendra[1*], Michael Jonathan[2†], Infall Syafalni[1,2†], and Trio Adiono[1,2†]
[1]*School of Electrical Engineering and Informatics*
[2]*The University Center of Excellence on Microelectronics*,
Institut Teknologi Bandung, Indonesia
E-mail: * {13520049, 13520055, 13520109, 13520151, 13519008}@std.stei.itb.ac.id,
† {nsutisna, michael.j, infall, tadiono}@itb.ac.id

***Abstract*—Recent technological advancements have made agricultural industry easier through the use of the Internet of Things (IoT). Many studies have been done to increase the adaptability of IoT in agricultural industries. IoT is mainly used as a tool to monitor and report data to a cloud server remotely. Those efforts are important to solve global challenges to produce 70% more foods, as the United Nation stated. To further solve those challenges, this paper tries to propose a comprehensive system to monitor the rice stock and forecast it for future use. This paper also shows that multivariable linear regression is an appropriate machine learning algorithm to forecast the rice stock. Furthermore, the forecast result is displayed in a web-based application built with Vue.js that exhibits a wonderful score when tested with Lighthouse.**

***Index Terms*—ESP32, IoT, machine learning, monitoring, rice stock, prediction**

I. Introduction

In recent years, new technologies have permeated a variety of industry sectors, including agriculture. These technologies allow individuals to effortlessly collect data and automate routine tasks. One such technology is the Internet of Things (IoT), which makes transmitting data and signals over networks possible. In the agricultural industry, it has the potential to automate traditional farming procedures and monitoring systems, increasing efficiency and effectiveness. The demand for food production increases exponentially as the global population grows. The United Nation (UN) Food and Agriculture Organization predicts a 70% increase in food production by 2050 compared to 2006 [1]. To meet this challenge, farmers must increase productivity and utilize their resources. One option to addressing these issues is to use IoT to track and forecast crop yields and stocks.

Many initiatives have been taken to address the difficulty of IoT. First, Jose [1] conducted studies that measured temperature, moisture, and some soil nutrients. This technology is applied in Kerala, India, because of the fall in rice crop yields despite having a vast quantity of arable land. Second, Sai's research [2] uses deep learning to identify diseases in rice agriculture. This study uses robots with several sensors, such as cameras and Wi-Fi modules, to transmit data on a regular basis.

In this paper, we have utilized IoT technology to monitor data, similar to the two studies mentioned before. However, this paper is focused more on the post-harvesting process, especially monitoring the rice stock. This paper also further analyzes the data and forecasts rice availability in the future using several machine learning methods, such as linear and polynomial regression.

II. Related Works

A. IoT for Food Monitoring

Prior research, as in [3], focuses on the developments and uses of multipurpose sensors in smart food packaging, emphasizing how well they can keep an eye on vital indicators including infections, gas concentrations, temperature, humidity, and storage time. These developments outperform conventional inspection techniques in terms of availability, efficacy, and efficiency. The technology uses a number of sensors to measure important aspects of the packaging, sending data via wireless connection through a network server to a central computer system. Customers may monitor the quality of vacuum-packed items in real-time from any place with this setup's user-friendly interface. Ensuring accurate and efficient quality and safety assessments throughout the storage duration provides a full view of storage status and promotes effective data sharing.

B. Food Demand Prediction using ML Models

In an effort to forecast accurate food orders for restaurants, [4] investigated the application of data mining, statistical, and machine learning models. The study was driven by the growing demand for food, climate change, and digitalization. Several models, such as lasso, ridge, gradient boosting approaches, support vector regression (SVR), decision tree, random forest, and Bayesian ridge regression, are evaluated in the study. Regression algorithm prediction performance has mostly been investigated in two ways: (1) choosing the best model to achieve certain objectives, and (2) assessing model performance by extracting pertinent characteristics from the dataset.

In this study, we will investigate both factors of the regression algorithm prediction performance. Given that our primary research objectives are to provide proof-of-concept by using various indicators to estimate present quality and future food stock demand, we will refer to existing methodologies that have shown promise using linear and polynomial regression models.

III. Methodology and System Design

This section describes the technique and system design for our IoT-powered rice stock monitoring system. We will look at individual hardware components, sensor selection, data gathering techniques, and overall system architecture. In this section, we'll go over how the selected sensors will interact with the rice storage unit and send out wireless data for real-time monitoring and stock-level analysis.

A. Hardware and Firmware

The firmware for our rice stock monitoring system must meet the following requirements: quantity measurement, quality measurement, and data transmission. Quantity measurement refers to the ability to determine how much rice is in a container using weight sensors. Quality measurement refers to the ability to use appropriate sensors to measure the quality of the rice in the container. Data transmission revolves around the ability to transfer collected data to a backend or database for further processing and monitoring.

To meet the requirement of measuring the quantity and quality of rice, the firmware integrates the following sensors:

- HC-SR04 [5] is an ultrasonic sensor that is used for measuring the level of rice in the container. By calculating the distance between the sensor and the surface of the rice, the firmware can determine the amount of rice left in the container.
- DHT22 [6] is a temperature and humidity sensor that is used to measure the temperature and humidity inside the container, which are critical indicators of rice quality. The firmware periodically reads the sensor data to monitor the environmental conditions affecting the rice.

To meet the need for transferring data to the database, we used a Wi-Fi-based connection with the HTTP protocol. The data is sent to the backend in JSON format to ensure compatibility and ease of processing. The payload is organized as shown in Table I.

TABLE I
JSON Payload Structure

Name	Data Type	Description
nodeID	String	uniqueID of the device
apiKey	String	apiKey for backend authentication
volume	Float	rice volume data
humidity	Float (0-100)	humidity data
temperature	Float	temperature data

To ensure proper assembly and connectivity of the IoT device, a detailed wiring diagram is provided. This diagram illustrates the connections between the ESP-32 microcontroller [7] and the sensors. The wiring diagram of the IoT system can be observed in Fig. 1.

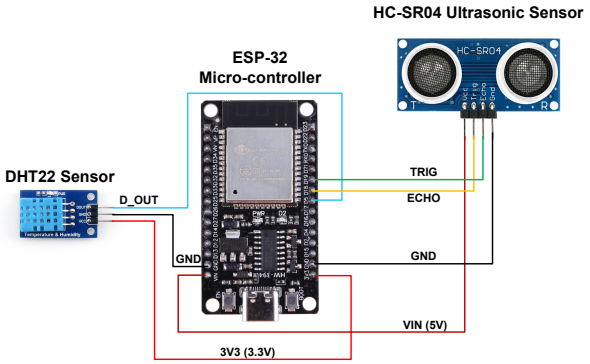


Fig. 1. Wiring diagram of the IoT system

The IoT system firmware is developed using the following tools and environments:

- Integrated Development Environment (IDE): Arduino IDE and Visual Studio Code with PlatformIO extension.
- Compiler: ARM GCC.
- Version Control: Git with GitHub for repository management.

During the designing phase, a number of configurations and design possibilities have been investigated in an effort to determine the best possible solution. Several alternatives were taken into account, but for various reasons they were not incorporated in the final design. These alternatives and considerations are as follows:

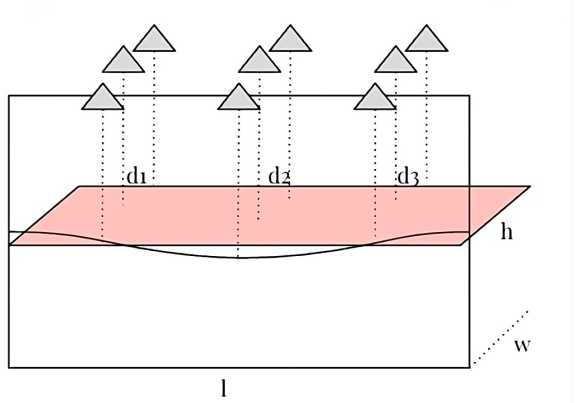
- **Photoresistor sensors**: we examined and tested the IoT system utilizing photoresistor sensors. The primary idea behind this alternative is to arrange a series of photoresistors inside the container, with a light source at the top. The volume measurement is based on determining whether the photoresistors are able to sense the light or not. If the photoresistors do not sense any light, it means that the rice level has reached the sensors. This design has several advantages, including lower costs and more precise measurements. However, we decided against using this configuration because it added complexity to both the setup and installation processes.
- **Infrared sensors**: in this alternative, we used infrared sensors to evaluate the IoT system. The basic idea is the same as with the previous option: instead of using photoresistors, an array of infrared sensors is placed within the container. The technique for measuring volume depends on determining whether or not the transmitted infrared (IR) signal reaches the sensors. When there is no IR signal detected, it means that the rice level has reached the sensors. This design has advantages of more precise measurements compared to the ultrasonic sensor utilization, and is also easier to install than the phororesistors design alternative, since the light source does not need to be placed on top of the container.

However, we chose not to employ this configuration due to the high cost of the infrared sensors, and according to the experiment, if the infrared sensors are not properly tuned, they will interfere with one another.

These alternative designs have been thoroughly reviewed to guarantee that the design solution is practical and efficient. The final design with the utilization of ultrasonic sensors was selected because it balances cost, performance, and requirements.

B. Data Processing and Analytics

The proposed feature extraction setup is depicted in Fig. 2. We use depth parameters acquired by numerous ultrasonic sensors at distinct plane coordinates to calculate the volume as our forecasting task's objective label. The average depth and volume are calculated using (1) and (2).



$$\bar{d} = \frac{d_1 + d_2 + d_3 + \cdots + d_n}{n} \tag{1}$$

$$V = w \times l \times (h - \bar{d}) \tag{2}$$

Fig. 2. A box with side lengths w and l, and height h, with d_i referring to depth for sensor$_i$ & V represents the volume of the box with $\bar{d}$ approximation

Our proposed feature extraction supports scalability and improved accuracy with an increased number of ultrasonic sensors, as it will better approximate the real volume, especially for uneven food stock surface conditions. For the stock forecasting task, we evaluate and select from linear regression models, with timestamp and volume as feature and label, respectively. We employ extensions of the base model to enhance prediction accuracy. The forecasting models are as follows:

- BASE Model: Simple linear regression is used as the BASE model to establish a relationship between the timestamp and the volume.
- POLYNOMIAL Model: The base model is extended by incorporating polynomial terms (POLY model), allowing it to fit non-linear relationships and improve accuracy.
- MOVING MULTILINEAR Model: The base model is further enhanced using a moving window approach, similar to the moving average principle. This method (MR) incorporates previous volume data within a window of size N, enabling the model to capture temporal dependencies and trends.

Algorithm 1 BASE LINEAR REGRESSION

Input: Timestamps t_i, Volume v_i
Output: Regression coefficients β_0, β_1

$$X \leftarrow \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}$$

$$Y \leftarrow \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

$$\beta \leftarrow (X^T X)^{-1} X^T Y$$

$$\beta_0, \beta_1 \leftarrow \beta$$

Algorithm 2 POLYNOMIAL REGRESSION

Input: Timestamps t_i, Volume v_i, Degree d
Output: Regression coefficients $\beta_0, \beta_1, \ldots, \beta_d$

$$X \leftarrow \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^d \\ 1 & t_2 & t_2^2 & \cdots & t_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^d \end{bmatrix}$$

$$Y \leftarrow \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

$$\beta \leftarrow (X^T X)^{-1} X^T Y$$

$$\beta_0, \beta_1, \ldots, \beta_d \leftarrow \beta$$

Algorithm 3 MOVING MULTILINEAR REGRESSION

Input: Timestamps t_i, Volume v_i, Window size N
Output: Regression coefficients for each window $\beta_0^{(w)}, \beta_1^{(w)}$
for each window $w = 1, 2, \ldots, \lfloor \frac{n}{N} \rfloor$ **do**

$$X_w \leftarrow \begin{bmatrix} 1 & t_{(w-1)N+1} \\ 1 & t_{(w-1)N+2} \\ \vdots & \vdots \\ 1 & t_{wN} \end{bmatrix}$$

$$Y_w \leftarrow \begin{bmatrix} v_{(w-1)N+1} \\ v_{(w-1)N+2} \\ \vdots \\ v_{wN} \end{bmatrix}$$

$$\beta_w \leftarrow (X_w^T X_w)^{-1} X_w^T Y_w$$

$$\beta_0^{(w)}, \beta_1^{(w)} \leftarrow \beta_w$$

end for

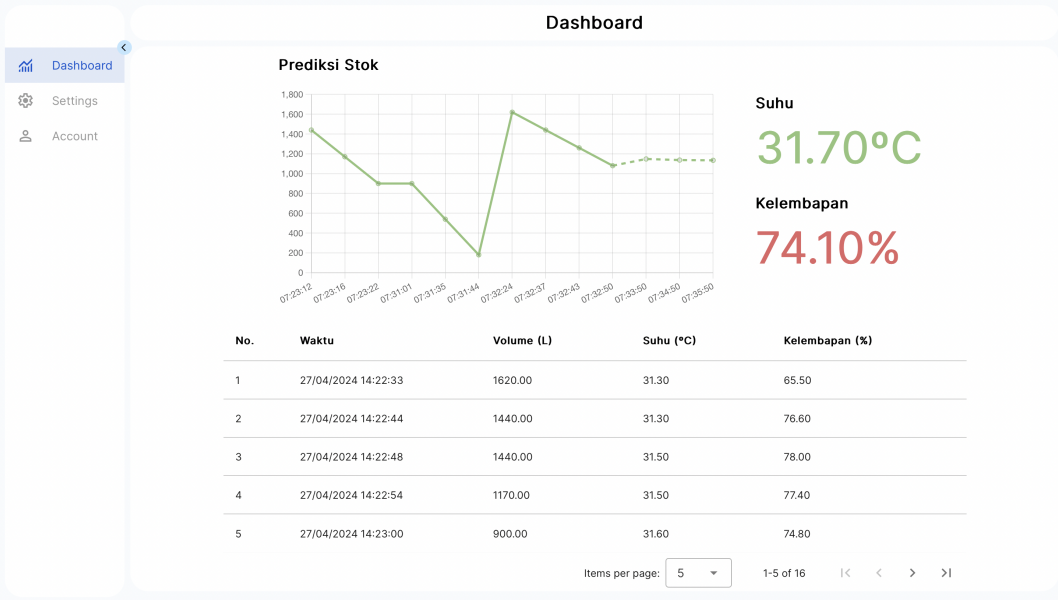


Fig. 3. Dashboard section to monitor data and forecast

C. End-to-end Software Development

A web-based software is developed to ease the user's ability to monitor the data and its forecast. In short, there are two components that build the software. First, the back-end component is composed of two main endpoints:

- **GET /node?startDate&secGap=&endForecast**: This endpoint is used to get all the node reading data. It accepts three query parameters. STARTDATE is used to specify the starting date of the node readings, SECGAP is used to specify the second gap used in the MOVING MUTLILINEAR model, and ENDFORECAST is used to specify the end date of the forecasted data generated by the model. To access this endpoint, the user is required to login as an authorization method. The response generated by this endpoint is an array of JSON consisting of values detailed in Table II.
- **POST /node**: This endpoint is used to post the sensor readings from a node. To use this endpoint, a node must be registered and provide a valid API key. This measurement takes place to prevent unauthorized nodes from posting data. The required data is the same as what is specified in Table II.

TABLE II
RESPONSE DATA

Field	Type	Description
nodeID	Integer	The id of the node posting the data
time	Datetime	The timestamp of the data
volume	Float	The measured volume of rice left in the container in cm^3
temperature	Float	The recorded temperature inside the rice container in Celcius
humidity	Float	The humidity of rice in percentage

The application is designed to handle multiple accounts, where each account can represent a different organization, corporation, or even a single individual. Each account has its own container, which is associated with a single node. A corporation might have multiple containers, each tracked separately, with a single node for each container. The interface must accommodate these needs. The front-end consists of several components built with Vue.js and Chart.js to visualize data. Every user must sign up and create an account to use this software. Users can have multiple nodes, each with its own API key, managed in the *Settings* section. The front-end includes three main pages:

- **Login and Registration:** This is where users are identified. By logging in or signing up, we can identify the correct node associated with the user.
- **Node Management:** Here, users can create, read, and delete nodes. Each node is identified by a unique API key, which is attached to the node and sent with each data transmission. Along with the API key, the node ID will be used as a claim, authenticated using the API key.
- **Dashboard:** The dashboard is used to monitor the quality of the rice stock. The node reads data from sensors and sends it to the backend by hitting the specified REST API endpoint. The data sent is a simple JSON object with three attributes: temperature, humidity, and volume. Users can view this data in the *Dashboard* section, as shown in Fig. 3. The forecasted data can also be seen on this page. The summary of the data is displayed in a bar chart, with a tabulated version also provided for users to see detailed data transmission.

The data is transmitted in small format, typically in kilobytes, and is sent at relatively long intervals, ranging from hours to days, depending on the settings. The JSON data

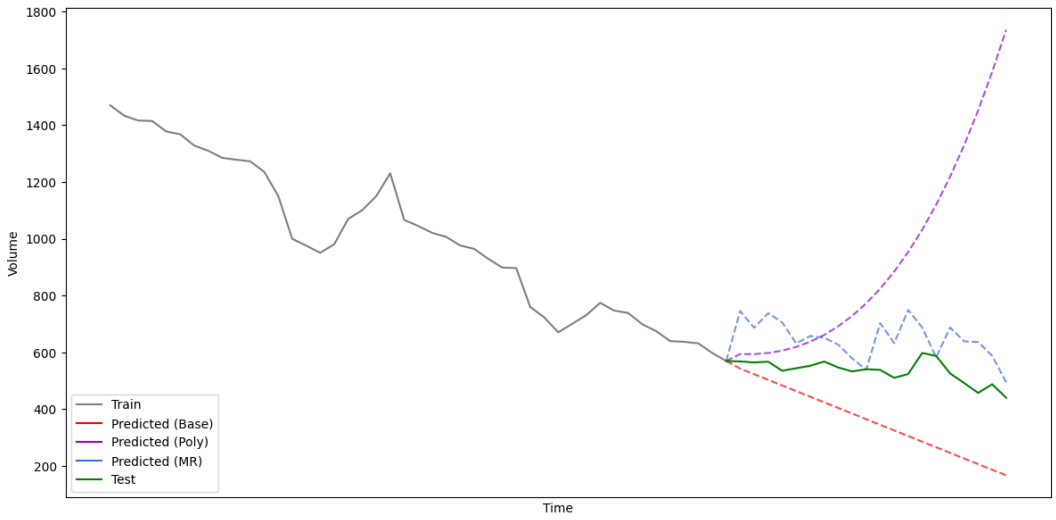


Fig. 4. Accuracy comparison between three methods (BASE, POLY, and MOVING MULTILINEAR) for a given time range

consists of three parts: **Humidity** (float), **Temperature** (float), and **Volume** (float). Humidity and temperature readings are sent directly from the sensor. However, the volume needs to be calculated first. The volume value depends on the container being used. The sensor is placed on top of the container, and two pieces of information must be manually input by the user: the height and the area of the container. We are assuming the containers have a consistent area, regardless of height. These parameters will be used for calculating the volume.

IV. RESULTS AND ANALYSIS

The results of the system's implementation are shown in this part, where they are interpreted using a variety of metrics. An evaluation of the system's performance and efficacy in predicting rice stocks will be conducted in this section.

A. Stock Forecasting

We use mean squared error (MSE) as our qualifying metric. By observing Fig. 4 and Table III, we can derive conclusive analysis from visual inspections and score analysis.

TABLE III
MSE RESULT AND COMPLEXITY FOR ALL METHODS

Method	**MSE**	**Time Complexity**	**Space Complexity**
BASE	38943.43	$O(n)$	$O(n)$
POLY	289600.88	$O(n \times d^2)$	$O(n \times d)$
MOVING MULTILINEAR	15838.23	$O(n \times N^2)$	$O(n \times N)$

BASE and MOVING MULTILINEAR appear to follow the general downtrend of the test data. However, POLYNOMIAL seems to deviate to extreme positive values, indicating there might be overfitting to the increased data points, hence the model can't capture the downtrend. BASE shows only stable forecasting, with no ability to capture the fluctuation of stock addition and extreme usage. MOVING MULTILINEAR appears to have the smoothest fit among the predicted methods and follows the test data most closely, with proper deviation as depicted in the MSE value from Table III.

MOVING MULTILINEAR regression emerges as the most suitable approach for this problem, anticipated due to its accounting for contextual considerations within a specified time window, enabling extrapolation of possible future values based on specifiable past trends. This algorithm is effective for analyzing data where relationships are expected to change over time or across different segments (as defined by the window size N). It offers flexibility in capturing varying trends but requires careful selection of N to balance computational efficiency and model accuracy. The quadratic time complexity relative to N suggests that larger N values may lead to increased computational demands. The analysis is based on a limited view of the graph. A wider timeframe might reveal different patterns. The performance of each method can vary depending on the specific characteristics of the time series data. Specific further fine-tuning might be needed to adjust performance.

B. System Performance

To measure the front-end component's performance, a tool called Lighthouse is used. Lighthouse is a Chrome's extension that allows users to measure the performance of a website. This approach is commonly used, including research that scores the Indonesian government's website [8]. As we can see in the Fig. 5 , the front-end component developed gives a satisfying score across many aspects. Two aspects, performance and best practice, even got a perfect score. While the other two aspects got a near-perfect score. This shows that the front-end component has served the functionality intended.

To measure the back-end component's performance, we simulate requests and make it as close as possible to a real-world scenario. We used Postman Runner [9], a feature in the

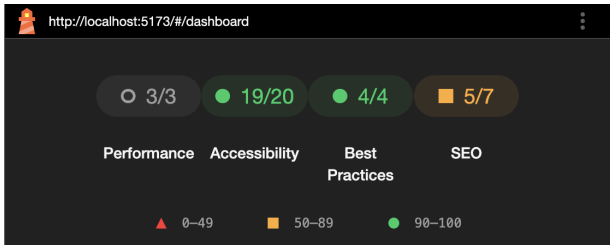


Fig. 5. Front-end performance measured by Lighthouse

Postman application, to simulate users making requests for a certain length of time. We tested two scenarios where the user can post the node readings and get the data back. To simulate those scenarios, we set the number of virtual users to 20 and set the time limit to five minutes. After running the test for five minutes, the result can be seen in Table IV - VI. The result shows that the POST request has a higher response time average than the GET request, meaning writing data to the back-end component is more consuming than reading data. Both of the endpoints have an error rate of 0.05%. In reading data, the forecasting task takes the majority of the time. While uploading data, authenticating node takes longer than writing to the database.

TABLE IV
BACK-END PERFORMANCE MEASURED USING THE POSTMAN RUNNER

Endpoint	Requests/second	Avg (ms)	Error %
Get Node Readings	6.67	831	0.05
Post Node Readings	6.84	1,154	0.05

TABLE V
AVERAGE TIME TAKEN TO COMPLETE EACH PROCESS WHILE READING DATA

Process	Average Time Elapsed (ms)
Reading data from database	133.79
Forecasting data	687.21

TABLE VI
AVERAGE TIME TAKEN TO COMPLETE EACH PROCESS WHILE UPLOADING DATA

Process	Average Time Elapsed (ms)
Authenticating node	692.4
Writing data to database	461.6

V. CONCLUSIONS

This study demonstrates the feasibility of employing Internet-of-Things (IoT) technology to aid and manage food stocks. Integrating IoT technology into agriculture has the potential to increase productivity. Monitoring and predicting food stocks can provide insight to the government and farmers, allowing them to make better decisions. Moving multi-linear regression demonstrates how to make predictions utilizing contextual data within a given time frame. The system's performance can manage up to twenty users with an average response time of roughly 1,000 ms, indicating that collecting data from twenty nodes of IoT devices is feasible.

Further study can be done on how to measure stock and rice condition in a different way, such as by weighing the rice or using another type of sensor. Another thing is to improve the data representation and provide filter capability to give the user more freedom in examining the data. Furthermore, a more efficient and strong back-end component framework is required to increase the system's scalability.

REFERENCES

[1] R. Jose and N. Thomas, "Smart farming via iot: A solution for monitoring rice productions," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 9, pp. 2320–2882, 2021.

[2] S. Kandula Kushal Sai, P. Satyanarayana, M. A. Hussain, and M. Suman, "A real time precision monitoring and detection of rice plant diseases by using internet of things (iot) based robotics approach," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, 2019.

[3] P. Alexandru, M. Hnatiuc, M. Paun, *et al.*, "An intelligent iot-based food quality monitoring approach using low-cost sensors," *Symmetry*, vol. 11, p. 374, Mar. 2019. DOI: 10.3390/sym11030374.

[4] S. Jayapal, "Food demand prediction using statistical and machine learning models," *International Journal of Engineering and Technology (IRJET)*, Jan. 2023.

[5] ElecFreaks, *Ultrasonic ranging module hc-sr04*. [Online]. Available: https://cdn.sparkfun.com/datasheets/Sensors/Proximity/HCSR04.pdf.

[6] T. Liu, *Digital-output relative humidity and temperature sensor*. [Online]. Available: https://www.sparkfun.com/datasheets/Sensors/Temperature/DHT22.pdf.

[7] E. Systems, *Esp32-wroom-32 (esp-wroom-32) datasheet*, 2018.

[8] M. A. Faizin, M. Nevin, and U. L. Yuhana, "Indonesia e-government website performance and accessibility evaluation using automated tool lighthouse," in *2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)*, 2024, pp. 210–215. DOI: 10.1109/ICoSEIT60086.2024.10497521.

[9] Postman, *Postman runner*, 2024. [Online]. Available: learning.postman.com.

Comparative Evaluation of Fine-Tuned Hybrid Transformer and Band-Split Recurrent Neural Networks for Music Source Separation

Ken Kalang Al Qalyubi*[1], Nur Ahmadi*†[2], Dessi Puji Lestari*†[3]
*School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, 40132, Indonesia
†Center for Artificial Intelligence (U-CoE AI-VLB), Bandung Institute of Technology, Bandung, 40132, Indonesia
Email: [1]alqalyubiken@gmail.com, [2]nahmadi@itb.ac.id, [3]dessipuji@itb.ac.id

***Abstract*—In recent advancements within the field of music source separation (MSS), state-of-the-art models such as Hybrid Transformer Demucs (HT Demucs) and Band-Split Recurrent Neural Networks (BSRNN) have been at the forefront. Although the pre-trained HT Demucs model is capable of separating six sources—drums, bass, guitar, piano, vocals, and others—it underperforms on guitar, piano, and other sources compared to its performance on bass, drums, and vocals, as measured by the utterance-level Signal-to-Distortion Ratio (uSDR) metric. To date, there has been no evaluation of the BSRNN model's ability to separate these six sources. This paper seeks to address this gap by investigating and comparing the performance of the BSRNN and HT Demucs models for six-source separation. Using the MoisesDB dataset, both models were developed and fine-tuned for this task. Their performance was then evaluated to identify the superior model for six-source separation. Experimental results reveal that the fine-tuned HT Demucs model surpasses the BSRNN model, achieving average uSDR and cSDR scores of 6.26 dB and 5.88 dB, respectively, compared to 5.52 dB and 5.38 dB for the BSRNN model. Moreover, the fine-tuned HT Demucs model outperforms its pre-trained counterpart on piano and other sources by 1 dB and 0.3 dB, respectively.**

I. Introduction

The task of music source separation (MSS) has drawn more and more attention in the community due to its wide application in music field [1]. MSS is a crucial technology in music information retrieval, aimed at isolating one or more target music sources from their mixture. Target music sources typically refer to various musical instruments such as bass, drums, and vocals, while the mixture refers to the combination of these source signals [2]. One of the popular models in music source separation (MSS) is Spleeter which is designed for ease and speed in separation tasks for its users [3]. However, this model has relatively low scores compared to current state-of-the-art models. In recent research, Demucs achieved state-of-the-art status with its latest development, Hybrid Transformer Demucs architecture [4]. In their latest study, Sparse HT Demucs achieved an average Signal-to-Distortion Ratio (SDR) score of 9.20 dB. The most competitive baseline model is BSRNN, which achieved better SDR scores for the other and vocal sources [4]. In the comparison of scores for each sound source, Demucs has an advantage in separating drum and bass sources, while the vocal and other sources are still better separated by the BSRNN model, with SDR scores of 10.47 and 7.08 compared to Demucs scores of 9.47 and 6.41. Currently, MSS models are largely limited to applications involving only four sources: bass, drums, vocals, and other. Other sound sources are grouped or merged into the other category. The lack of publicly available datasets in this domain also contributes to the limitation of these sound sources.

One of the datasets commonly used to compare the performance of MSS models is MUSDB18. This dataset only contains data for bass, drums, vocals, and other sources. Although MUSDB18 [5], [6] has significantly contributed to advancements in this task, its source grouping is still too coarse for many real-world remix applications [7]. A recent study developed a dataset specifically to address the challenges of MSS tasks, called MoisesDB. This dataset facilitates the creation and evaluation of detailed source separation systems, surpassing the limitations of using only 4 sources (drums, bass, other, and vocals) due to a lack of data [8]. It consists of 240 music tracks from different artists and genres with a total duration of over 14 hours. The study tested the pre-trained HT Demucs model on six source classes (drums, bass, other, vocals, guitar, and piano). The results showed that HT Demucs achieved an average SDR score of 6.24 dB across all sources. Specifically, for each source, HT Demucs obtained SDR scores of 9.55 dB for vocals, 11.93 dB for bass, 11.02 dB for drums, 0.28 dB for other, 1.60 dB for piano, and 3.07 dB for guitar. The SDR scores for guitar, piano, and other are still relatively lower compared to sources like bass, drums, and vocals.

Guitar and piano sources are derived from splitting the other stem where the previous stems group consists of drums, vocals, other, and bass. Although HT Demucs study shows that BSRNN model outperforms in the other class [4], there has been no research conducted on separating guitar and piano stems using this model. In this case, BSRNN might surpass HT Demucs model in separating guitar and piano sources. This highlights the need to investigate the BSRNN model in separating 6 stem sources. Additionally, it is necessary to rebuild the HT Demucs model for comparison to determine which model is the best for separating 6 stem sources.

II. Related Works

In recent years, various studies have explored advancements in music source separation. One notable contribution is Hybrid Transformer Demucs (HT Demucs), which improves music source separation by integrating Transformer layers into its architecture. This model achieved state-of-the-art results on the MUSDB dataset, reaching 9.20 dB SDR with additional training data [4]. By incorporating Transformer layers into the existing Hybrid Demucs framework, HT Demucs utilizes self-attention mechanisms within the temporal and spectral domains, alongside cross-attention between these domains. This architecture significantly enhances the model's ability to capture long-range dependencies in music signals, offering superior performance, particularly in the Music Demixing Challenge, compared to the original Hybrid Demucs.

Further improvements in music source separation have been demonstrated by the Band-Split RNN (BSRNN) architecture, which was one of the top-performing models in the Music Demixing Challenge 2021 [9]. BSRNN splits the spectrogram of a music mixture into different frequency bands, optimizing the processing of specific instrument characteristics. This band-wise processing allows for more precise separation of musical components, improving upon existing models. The study also introduces a semi-supervised finetuning pipeline that utilizes both labeled and unlabeled data, further boosting the performance of the BSRNN model. This approach represents a significant step forward in the field of music source separation by enhancing the isolation of individual instruments in complex musical mixtures.

Additionally, the development of new datasets has played a crucial role in advancing music source separation. One such dataset, MoisesDB, was created to address the limitations of existing datasets that focus predominantly on four sources (drums, bass, vocals, and others) [8]. MoisesDB contains 240 songs with separate sound sources, enabling more detailed source separation beyond the constraints of traditional datasets like MUSDB18 and MedleyDB [10], [11]. MUSDB18, although foundational for progress in music source separation, often groups sound sources too coarsely for complex remix applications [7]. Meanwhile, MedleyDB suffers from poor labeling, further highlighting the need for datasets like MoisesDB that can support more granular source separation tasks.

III. Methods

In this study, we aim to investigate the performance of BSRNN and HT Demucs models in separating six stem sources. We conduct several experiments to find the best model and configuration for each model.

A. Data

Data in MoisesDB will first be filtered to ensure that each song contains at least the following sources: guitar, piano, vocals, drums, bass, and one other sound source classified as other. Additionally, there will be a mixed data set, which is the combination of all sound sources. After filtering, the data for each sound source in each song will be checked to ensure they have the same duration. Any sound sources with differing durations will be trimmed to match the shortest sound source duration in the respective song. We then split the processed dataset accordingly with 70:20:10 proportions resulting in 62 train set, 18 test set, and 8 validation set. The dataset description is shown in Table I.

TABLE I
Dataset description

Item	Value
Number of data (songs)	88
Total duration (hours)	5
Average duration (minutes)	3

B. Band-Split Recurrent Neural Networks (BSRNN)

BSRNN architecture that we use for this experiment is the implementation found in this repository[1]. Therefore, the implementation script from that repository will be used and adjusted according to the goals of the research. The model will be built from scratch, without using pre-trained models. Six new models need to be constructed: vocal, drum, bass, guitar, piano, and other. For all experiments we train the model with an initial learning rate $1e^{-3}$ and a batch size of 2 [9]. We use the small model version because of the limitation of our computing resource, where we set the feature dimension N to be 64, hidden unit of BLSTM layers to be 2N = 128, the hidden size in the mask estimation MLP to be 4N = 256, and use 8 band and sequence modeling modules. BSRNN is a single-target model meaning that a model is only responsible for one stem [9]. Figure 1 depicts the inference process of the BSRNN model for separating different musical components from an input song. The input song is processed by separate BSRNN models, each specifically trained to isolate one stem. Each model outputs the corresponding isolated track for its respective stem.

For building the new sound source models, specifically guitar and piano, a bandsplit configuration is required. Thus, the first step in developing the BSRNN model is to determine the bandsplit configuration that provides the best performance for separating guitar and piano sources as the rest of the stems already have their own configurations. Subsequently, models for each sound source will be constructed as part of the final training process. Each model for guitar and piano will be tested with three bandsplit configurations [9]. These configurations include vocals configuration (v7), bass configuration, and drums configuration. For speed, we train this experiment with a maximum epoch of 50 for speed.

Based on these configurations, the best-performing bandsplit configuration for the guitar and piano models will be determined. Once the optimal bandsplit configuration for guitar and piano as new sound sources is identified, the next step is to train all stems. These models include vocals, drums, bass, guitar, piano, and other. We also use the V7 configuration

[1]https://github.com/magronp/bsrnn

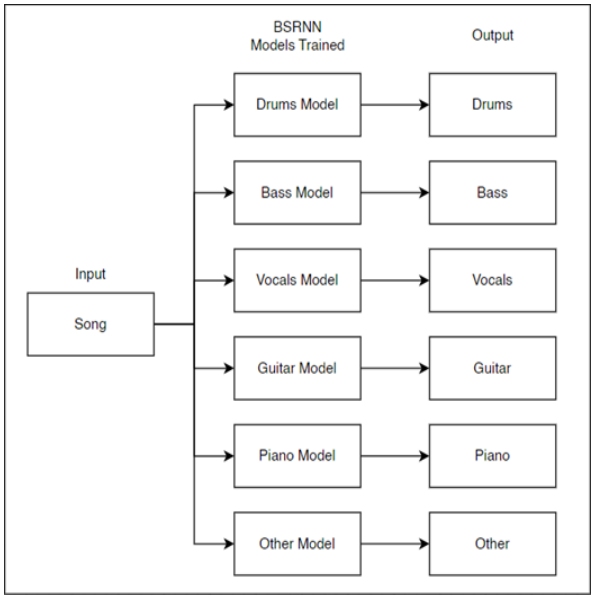


Fig. 1. BSRNN inference process for MSS.

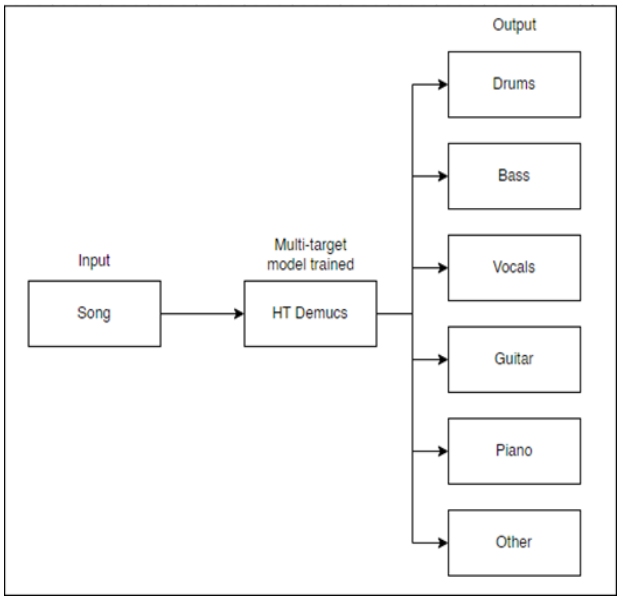


Fig. 2. HT Demucs multi-target inference process for MSS.

for the other stem, as previous research has done. For these experiments, we train the models with a maximum epoch of 100.

C. HT Demucs

HT Demucs model used is the official implementation from Meta AI, available on the GitHub repository[2] maintained by Facebook Research. Since Sparse HT Demucs is not publicly accessible, we have opted to use HT Demucs for this study. For all HT Demucs experiments, we train the models with a batch size of 2, optimized with Adam without weight decay. For HT Demucs we also use the small model where we set the channels to be 48 and the input/output dimension of the Transformer is 384. HT Demucs model can be train as a multi-target model or single-target model. Figure 2 shows how the inference process of the HT Demucs multi-target model for separating different musical components from an input song. The input song is processed by a single HT Demucs model trained for multi-target separation. The model outputs isolated tracks for each stem.

We first train HT Demucs as a multi-target model. We use learning rate of 3×10^{-4} as stated on [4]. We then tune the model to be single-target model. We follow the proposed procedure [4] where one copy of the multi-target model is fine-tuned on single target task with a learning rate of 10^{-4} and max epoch of 50. As a result, six separate models are developed, similar to the BSRNN approach. Subsequently, a comparison is made between the multi-target and single-target models.

D. Performance Metrics

We evaluate the performance of models with two following metrics:

[2]https://github.com/facebookresearch/demucs

1) uSDR: uSDR corresponds to the modified utterance-level signal-to-distortion ratio metric, and used as the default evaluation metric in the Music Demixing (MDX) Challenge 2021 and MDX Challenge 2023. This was proposed in [12] and used in [9]. We report the mean across the SDR scores of all test songs.

2) cSDR: cSDR corresponds to the chunk-level SDR calculated by the standard SDR metric in bss eval metrics [13] and served as the default evaluation metric in the Signal Separation Evaluation Campaign (SiSEC). It was also used on [9]. We use the official implementation which reports the median across the median SDR over all 1 second chunks in each song.

IV. Results and Discussion

A. Band-Split RNN (BSRNN)

We find that both guitar and piano achieve their optimal performance when utilizing the drums bandwidth configuration as can be seen in Table II. This phenomenon is attributed to the non-constant frequency ranges inherent in both instruments. Unlike vocals or bass, which typically maintain more constant frequency ranges, the guitar and piano produce a wide array of tones and harmonics. Moreover, drums configuration has more subband generated with its configuration compared to vocals and bass. This is why piano and guitar might be better when using drums configuration.

The lowest scoring stem in both uSDR and cSDR is the other stem. Most likely this is because BSRNN is unable to capture or separate inconsistent sources. The other stem is indeed a mix of rarely appearing sources such as flutes, tambourines, and even sound effects. Additionally, these sounds are often relatively non-dominant compared to main sounds like drums and vocals.

TABLE II
COMPARISON OF GUITAR AND PIANO MODEL USING BASS, VOCALS, AND DRUMS CONFIGURATIONS.

Instrument	Guitar		Piano	
	uSDR	cSDR	uSDR	cSDR
Bass	1.33	0.89	1.21	0.60
Vocals	1.83	1.38	1.51	0.78
Drums	**1.85**	**1.72**	**1.73**	**0.99**

B. HT Demucs

Empirically, we find that there is not much difference between training for 200 epochs compared to 100 epochs, indicating that further training is unnecessary. We achieve an overall score of 5.58 dB in uSDR and 5.26 dB in cSDR when the model is trained for 200 epochs, with only a 0.02 dB difference compared to training for 100 epochs.

We then fine-tune the 200 epochs model on a single-target task. We find that following this pipeline shows a significant improvement by differences of 0.3 dB – 1 dB. We compare this model with the open-source 6 stems HT Demucs and found that the model trained for piano stem achieves a score of 2.75 dB on uSDR and 2.03 dB on cSDR, compared to the pre-trained model, which only achieves 1.76 dB uSDR and 0.72 dB on cSDR. Additionally, this improvement is also observed on other stem. However, the trained HT Demucs model has not yet surpassed the average score of the pre-trained model, with an average uSDR score of 6.48 dB compared to 6.26 dB for the trained model, and a cSDR score of 6.03 dB compared to 5.88 dB for the trained model.

C. Comparison of BSRNN and HT Demucs

We find that HT Demucs single-target (fine-tuned) model has successfully surpassed the average score of the HT Demucs multi-target version model and even the BSRNN model as can be seen in Table III. In both uSDR and cSDR metrics, HT Demucs single-target model demonstrates that it produces relatively clean signals and is consistently clean both in cSDR and uSDR when compared to other models.

We carried out further analysis to see more clearly the quality of the signals produced by BSRNN and HT Demucs. The results on Figure 3 and Figure 4 show that piano stem produced by HT Demucs has more complete signal compared to BSRNN. We also found that both the left-channel and right-channel signals demonstrate that the HT Demucs produces a balanced signal in the guitar source. However, when the bass source is played, the sound produced by HT Demucs still contains noise from the guitar stem. On the other hand, the output from the BSRNN model is cleaner of noise, although the sound is not balanced. This also applies to the vocal stem, where noise from the drums is still present in the output of the HT Demucs. For other stem, no sound is produced by BSRNN, whereas HT Demucs model still produces audible sound for the other stem.

TABLE III
COMPARISON OF BSRNN AND HT DEMUCS MODELS.

Instrument	Metric	Model		
		BSRNN	HT Demucs	Fine-tuned HT Demucs
Guitar	uSDR	2.39	2.46	**3.01**
	cSDR	2.38	2.16	**2.70**
Piano	uSDR	2.22	1.73	**2.75**
	cSDR	1.58	0.89	**2.03**
Vocal	uSDR	9.12	8.39	**9.28**
	cSDR	8.15	7.63	**8.72**
Bass	uSDR	9.45	10.37	**11.25**
	cSDR	10.39	10.87	**11.48**
Drum	uSDR	9.93	9.85	**10.41**
	cSDR	9.80	9.90	**10.12**
Other	uSDR	7.04×10^{-6}	0.69	**0.85**
	cSDR	1.34×10^{-7}	0.13	**0.26**
Overall	uSDR	5.52	5.58	**6.26**
	cSDR	5.38	5.26	**5.88**

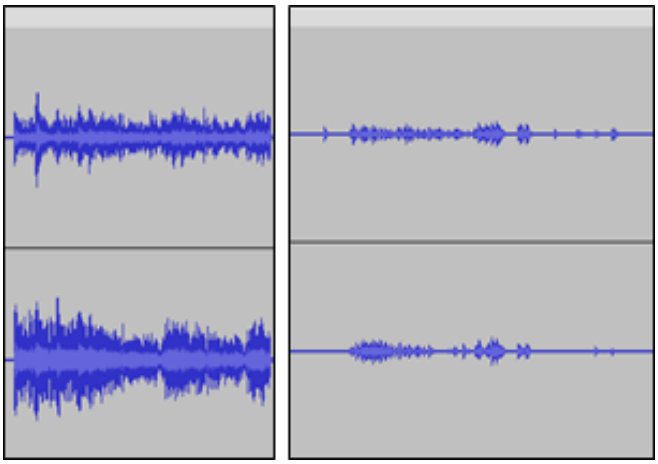


Fig. 3. Results of piano stem separation using the HT Demucs (left) and BSRNN (right).

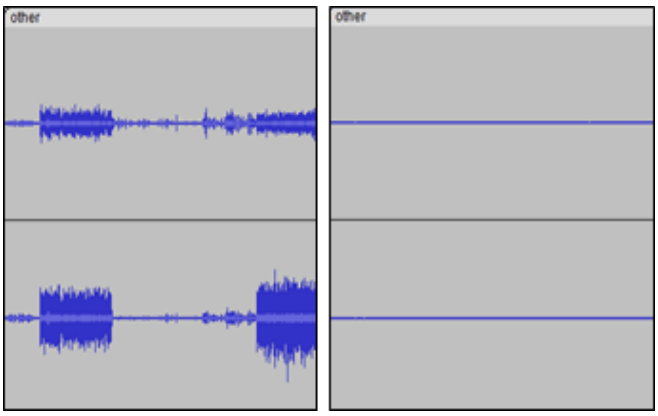


Fig. 4. Results of other stem separation using the HT Demucs (left) and BSRNN (right).

V. CONCLUSION

In this paper, we investigate the performance of the Band-Split RNN (BSRNN) and Hybrid Transformer Demucs (HT Demucs) models in separating six stem sources. We find that BSRNN model configuration for guitar and piano that

produced the best performance was when using the drum configuration. This is because the drum bandsplit configuration is more dispersed and results in more subbands compared to the vocal and bass configurations. It achieved an average uSDR score of 5.526 and a cSDR score of 5.387 across all sources. However, BSRNN model is not yet optimal for capturing sources that do not reside at constant frequencies, such as "other" as evidenced by evaluation scores that are close to zero for this category.

For HT Demucs, The single-target model achieved the best performance with a uSDR score of 6.264 and a cSDR score of 5.888. These results were achieved by training the multi-target model to focus on one specific audio source. The trained single-target HT Demucs model surpassed the performance of the pre-trained HT Demucs model on piano and other audio sources. For the piano source, the trained HT Demucs model outperformed with a difference of 1 dB in uSDR and 1.3 dB in cSDR metrics. For the "other" source, the trained model excelled with a difference of 0.3 dB in uSDR and 0.2 dB in cSDR.

Overall, the HT Demucs model outperformed BSRNN model based on the uSDR and cSDR evaluation metrics, with average scores of 6.26 dB and 5.88 dB, respectively. The analysis also showed that HT Demucs model excelled in producing balanced audio signals in both the left and right channels and was superior in signal completeness. On the other hand, the BSRNN model excelled in producing clearer audio with less noise, although the signals generated were less balanced and, in some cases, incomplete.

REFERENCES

[1] O. Gillet and G. Richard, "Extraction and remixing of drum tracks from polyphonic music signals," in *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2005, pp. 315–318.

[2] J. Qian, X. Liu, Y. Yu, and W. Li, "Stripe-transformer: deep stripe feature learning for music source separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 2, 2023.

[3] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020.

[4] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] Z. Rafii, A. Liutkus, and F.-R. Stöter, "The musdb18 corpus for music separation," Zenodo: https: //doi.org/10.5281/zenodo.1117372, 2018.

[6] ——, "Musdb18-hq - an uncompressed version of musdb18," Zenodo: https://doi.org/10.5281/zenodo.3338373, 2019.

[7] E. Manilow, G. Wichern, and J. Le Roux, "Hierarchical musical instrument separation," in *2020 International Society for Music Information Retrieval Conference (ISMIR)*, 2020, pp. 376–383.

[8] I. G. Pereira, F. Araujo, F. Korzeniowski, and R. Vogl, "Moisesdb: A dataset for source separation beyond 4 stems," in *2023 International Society for Music Information Retrieval Conference (ISMIR)*, 2023, pp. 619–626.

[9] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.

[10] R. M. Bittner, J. Wilkins, H. Yip, and J. P. Bello, "Medleydb 2.0: New data and a system for sustainable data collection," *ISMIR Late Breaking and Demo Papers*, vol. 36, 2016.

[11] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *2014 International Society for Music Information Retrieval Conference (ISMIR)*, vol. 14, 2014, pp. 155–160.

[12] Y. Mitsufuji, G. Fabbro, S. Uhlich, F.-R. Stöter, A. Défossez, M. Kim, W. Choi, C.-Y. Yu, and K.-W. Cheuk, "Music demixing challenge 2021," *Frontiers in Signal Processing*, vol. 1, p. 808395, 2022.

[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

A Real-Time Platform for Portable and Scalable Active Noise Mitigation for Construction Machinery

Woon-Seng Gan*, Santi Peksi†, Chung Kwan Lai‡, Yen Theng Lee§, Dongyuan Shi¶ and Bhan Lam‖
*†§¶‖ School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
E-mail: {ewsgan*, speksi†, dongyuan.shi¶, bhan lam‖} @ntu.edu,sg
§ e190091@e.ntu.edu.sg
‡ Institute of Sound and Vibration Research, University of Southampton, UK
C.K.Lai@soton.ac.uk‡

***Abstract*—This paper introduces a novel, portable, and scalable Active Noise Mitigation (PSANM) system designed to reduce low-frequency noise from construction machinery. The PSANM system consists of portable units with autonomous capabilities, optimized for stable performance within a specific power range. An adaptive control algorithm with a variable penalty factor prevents the adaptive filter from overdriving the anti-noise actuators, avoiding non-linear operation and instability. This feature ensures that the PSANM system can autonomously control noise at its source, allowing for continuous operation without human intervention. Additionally, the system includes a web server for remote management and is equipped with weather-resistant sensors and actuators, enhancing its usability in outdoor conditions. Laboratory and in-situ experiments demonstrate the PSANM system's effectiveness in reducing construction-related low-frequency noise on a global scale. To further expand the noise reduction zone, additional PSANM units can be strategically positioned in front of noise sources, enhancing the system's scalability. The PSANM system also provides a valuable prototyping platform for developing adaptive algorithms prior to deployment. Unlike many studies that rely solely on simulation results under ideal conditions, this paper offers a holistic evaluation of the effectiveness of applying active noise control techniques directly at the noise source, demonstrating realistic and perceptible noise reduction. This work supports sustainable urban development by offering innovative noise management solutions for the construction industry, contributing to a quieter and more livable urban environment.**

***Index Terms*—Active Noise Control(ANC), construction machinery, output power constraint, in-situ noise control, modified filtered reference least mean square algorithm**

I. Introduction

The rapid evolution of modern construction methodologies and machinery has significantly expedited urban development and infrastructure projects worldwide. However, this progress incurs an environmental cost, particularly in the form of noise pollution [1]. Construction machinery, including generators, exhaust systems, and transformers, is a primary contributor to urban noise, adversely affecting public health and the environment. The World Health Organization (WHO) identifies noise pollution as the second-largest environmental cause of health issues, following air quality concerns [2].

Active Noise Mitigation (ANM) technologies complement traditional noise control strategies, providing greater effectiveness for low-frequency noise, while passive methods, such as noise barriers, excel at higher frequencies. Unlike passive methods, ANM systems utilize electronic means to generate a sound wave that matches the original noise in amplitude but is opposite in phase, effectively canceling it out through destructive interference. This principle, known as active noise control (or mitigation), has been extensively researched and developed over the past few decades, finding applications across various industries [3]–[9]. One of the earliest uses of active noise control technology was in aerospace, where it helped reduce cabin noise in commercial aircraft [10]. The technology has since been adapted for the automotive industry, with manufacturers integrating ANM systems to enhance driving experiences by reducing engine and road noise [11].

In construction, ANM technologies offer promising solutions for managing the complex noise generated by machinery. These systems can be integrated into machinery or deployed as mobile and scalable units, enhancing flexibility in noise management strategies. A pilot study demonstrated significant noise level reductions at a construction site using ANM systems, indicating potential for broader application in the industry [12]. However, it is important to note that these experiments have primarily been conducted in controlled laboratory settings, using loudspeakers to replicate engine or machinery noise, and have yet to be tested under real on-site conditions.

The portable and scalable Active Noise Mitigation (PSANM) presents substantial noise reduction advantages over traditional passive noise control methods, especially for low-frequency noise, which is challenging to mitigate effectively using conventional means. A comprehensive review outlines the effectiveness of various passive noise control materials and designs, noting their widespread application in urban construction settings [13]. However, while passive methods effectively address high-frequency noise, they often fall short against low-frequency noise generated by construction machinery [14], [15]. Additionally, physical constraints and aesthetic considerations in urban environments limit the applicability of extensive passive noise mitigation structures. The European Environment Agency emphasizes that while passive measures are crucial for noise management, they cannot singularly resolve the multifaceted challenges of urban noise pollution, underscoring the need for innovative solutions [16].

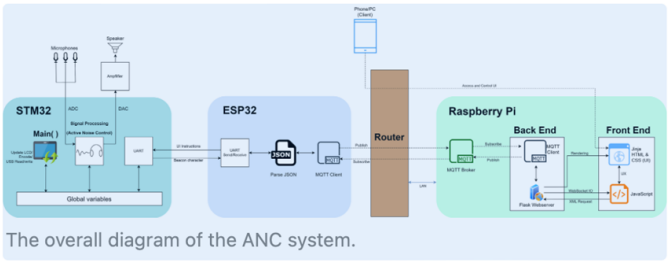


Fig. 1. he overall block diagram of the PSANM system consists of the active noise mitigation processing module, a communication module for remote operational mode control, and a module that performs data monitoring and performance evaluation.

This paper offers an overview of the current noise pollution landscape caused by construction activities and introduces the fundamental principles of PSANM technology. It examines PSANM's potential in controlling low-frequency construction noise without the need for a passive element. Furthermore, the paper delves into the technical aspects of the developed adaptive active noise control system, which has been tested in both laboratory and on-site environments with noise generators. It also explores future directions for innovation and wider adoption of PSANM systems.

By addressing the critical issue of noise pollution in construction, this paper contributes to the ongoing discussion surrounding sustainable urban development and the role of technology in mitigating environmental impacts. The paper is organized as follows: In Section II, a real-time PSANM system architecture is presented, followed by a detailed description of the hardware components, adaptive signal processing algorithms, software elements, and system integration with remote control of the PSANM. Section III and IV outline our controlled laboratory and in-situ experiments, respectively, highlighting the noise reduction limits and their spatial zones. We conclude our research work in Section V.

II. Proposed Real-Time PSANM

To develop a portable and scalable Active Noise Mitigation (PSANM) system aimed at reducing low-frequency noise emissions from construction machinery, we designed a series of portable units powered by embedded systems, including the STM32 microcontroller, ESP32 processing module, and Raspberry Pi 4 (RPi4) [17], as shown in Fig. 1. The STM32 microcontroller is programmed with an adaptive algorithm that is optimized to operate within a specified power output range, ensuring both distortion-free performance and system stability [18]. To achieve remote access and monitoring of the PSANM units, we have programmed the ESP32 module as a WiFi module to connect to a router, which is linked to the RPi4 via a LAN connection [19]. The RPi4 serves dual functions, hosting both the MQTT broker for efficient message handling and the web Server. The PSANM system can adapt to various types of machinery noise, effectively managing frequencies below $200\,\mathrm{Hz}$ with noise levels exceeding $90\,\mathrm{dB(C)}$.

A. Key Features and Innovations

One of the standout features of the PSANM system is its integrated web server, which enables streamlined remote management and monitoring. This web server provides significant operational flexibility, allowing users to access and control the system from any location with an internet connection. The remote access capability is crucial for managing PSANM units in large or hard-to-reach construction sites, ensuring continuous and efficient noise control operations.

To ensure durability and reliability in challenging outdoor environments, the PSANM units are housed in weather-resistant enclosures. These enclosures protect the sensitive electronic components from adverse weather conditions such as rain, dust, and extreme temperatures, ensuring the longevity and consistent performance of the system, as illustrated in Fig. 2. A photo of the dual PSANM units is shown in Fig. 3.

The system's active noise mitigation module uses reference and error microphones, which feed into an STM32 microcontroller equipped with a Nucleo ARM Cortex-M7 processor. The signals from these microphones are amplified and interfaced with the analog-to-digital converters (ADC) on the STM32 development board. The digitized signals are processed by a set of real-time adaptive algorithms, which will be further detailed in Section II.B. The processed anti-noise signals are then delivered to a secondary source, a 10-inch, 400-watt (RMS) subwoofer with a sensitivity of 85.7 dB 1W/1m.

To perform the initialization stage of secondary path calibration before the actual production of the anti-noise signal, secondary path estimation must be carried out in the noisy environment. Given the primarily tonal nature of the machinery noise, an uncorrelated and broadband training signal is used for this estimation. Experiments have shown that in-situ secondary path estimation is feasible and does not result in significant degradation of noise reduction performance. The following sections will further elaborate on the enhanced adaptive algorithm that can be applied in a real-world environment.

B. Adaptive Algorithms

The conventional Filtered reference Least Mean Square (FxLMS) Algorithm [13], [20]–[25] remains the most effective and practical adaptive algorithm for real-time implementation. The PSANM system adopts the baseline FxLMS algorithm for both feedforward and feedback modes to better address different noise characteristics.

In the feedforward mode, a shared reference microphone is required, and this reference information must be transmitted to all PSANM units either by wire or wirelessly. Additionally, cross-channel secondary path information is necessary to achieve centralized feedforward control among the PSANM units. In contrast, the feedback mode allows each PSANM unit to be self-sufficient with its own reference and error signals, enabling decentralized noise control. The feedback mode is particularly suited for handling narrow-band construction noise and offers better scalability and flexibility in different noise

regions compared to the feedforward mode, which can handle a broader frequency range but at a higher implementation cost.

Currently, there is a growing research interest in distributed active noise control to expand the spatial region of noise control. The PSANM system can play a crucial role in experimenting with the real-time performance of distributed noise control, contributing significantly to this emerging field.

C. Adaptive Algorithm with Variable Penalty Factor

In a PSANM system, the disturbance $d(n)$ can be suppressed by the anti-noise $y'(n)$ from the control signal $y(n)$ passed through the secondary path $s(n)$. The control signal is the result of processing the reference signal $x(n)$ by using the control filter $\mathbf{w}(n)$. To circumvent the output-saturation issue of practical PSANC systems [26]–[28], restricting the variance of the control signal has been proven to be an efficient and feasible solution. Hence, the objective of the PSANM algorithm can be represented as a quadratically constrained quadratic program (QCQP) that minimizes the mean squared error signal under the limited output power [29]–[33]:

$$\underset{\mathbf{w}}{\text{minimize}} \quad f(\mathbf{w}) = \mathbb{E}\left[e^2(n)\right], \tag{1a}$$

$$\text{subject to} \quad g(\mathbf{w}) = \mathbb{E}\left[y^2(n)\right] \leq \rho^2. \tag{1b}$$

Fig. 4 presents the block diagram for implementing a variant of the feedforward FxLMS algorithm, known as the Minimum Output Variance FxLMS (MOV-FxLMS) algorithm [34], which utilizes an optimal time-varying penalty factor to prevent output saturation in the PSANM system. Additionally, a feedback version of the FxLMS algorithm, based on internal model control, has been programmed as an alternative option for the PSANM system.

A real-time penalty factor, $\alpha(n)$, is computed for each time frame of M samples, based on the varying power gain of the secondary path, $G_s(n)$, and the power of the estimated disturbance, $d(n)$, as shown in the following:

$$\alpha(n) = \max\left\{G_s(n)\left(\sqrt{\frac{\sum_{m=0}^{M-1}\hat{d}^2(n-m)}{MG_s(n)\rho^2}} - 1\right), 0\right\}, \tag{2}$$

where 0 is the minimal value to avoid negative $\alpha(n)$. Therefore, the improved optimal time-varying penalty factor MOV-FxLMS algorithm can be rewritten as follow

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu[e(n)\mathbf{x}'(n) - \alpha(n)y(n)\mathbf{x}(n)]. \tag{3}$$

The advantage of using a real-time penalty factor is that it helps the MOV-FxLMS algorithm achieve optimal control under output constraints, even with abruptly changing noise. This capability is critical for the autonomous operation of the PSANM system.

D. System Integration and Remote Control

Remote control capabilities further enhance the system's flexibility and user convenience. We employ an ESP32 module as a Wi-Fi bridge, facilitating communication between the PSANM system and the web server via the MQTT (Message Queuing Telemetry Transport) protocol for network communications. This setup allows the Wi-Fi module to connect to a router, which is directly linked to the Raspberry Pi via LAN (Local Area Network). The ESP32 module updates the STM32 with any system status updates via serial communication through UART (Universal Asynchronous Receiver-Transmitter) ports once the updates are received from the MQTT broker. The Raspberry Pi hosts both the MQTT broker for efficient message handling and the web server responsible for maintaining the system's user interface, as shown in Fig. 1.

The web server executes JavaScript logic that enables user interaction, allowing users to adjust settings and receive updates on system status and noise control outcomes. This comprehensive digital infrastructure ensures seamless operation and interaction with the PSANM system, providing users with

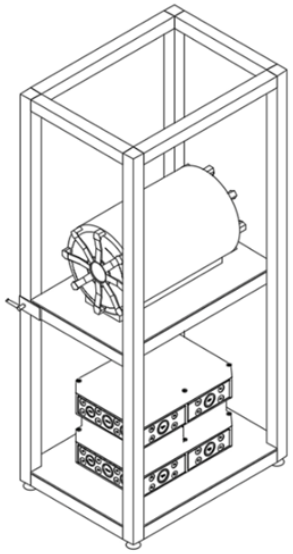


Fig. 2. Isometric view of a single PSANM device

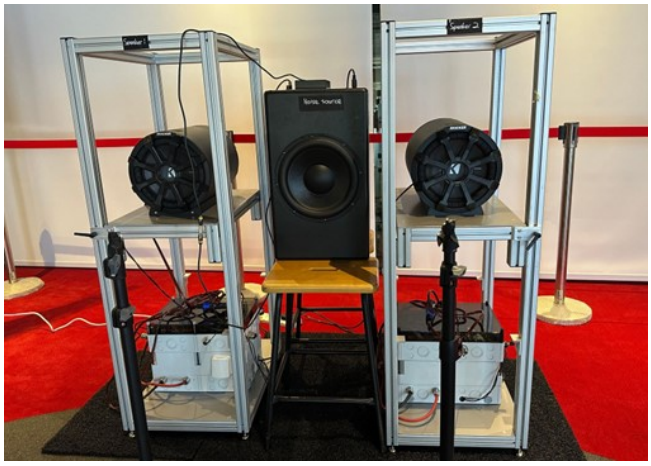


Fig. 3. The demo setup of dual-PSAMN device that place in front of the noise source (at center)

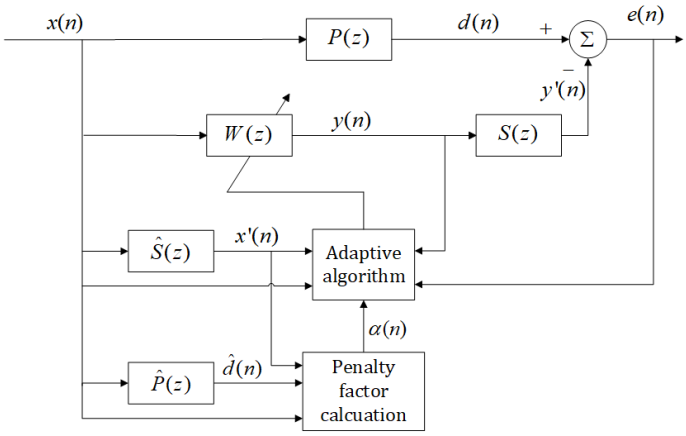


Fig. 4. The MOV-FXLMS (feedforward) algorithm that s programmed into the PSANM system to prevent output saturation and allow for autonomous operation

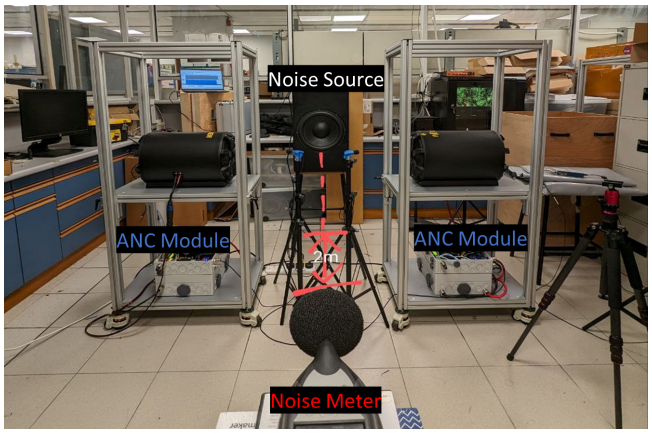


Fig. 5. Laboratory set-up to test and fine-tune the PSANM units

a powerful tool for managing noise pollution in construction environments.

III. Laboratory Experiment Set-up and Testing

To effectively deploy our PSANM system in real-world environments, we recognized the challenges posed by uncontrolled on-site conditions, such as traffic noise and other urban sounds. To address these factors, we conducted a series of controlled laboratory experiments to evaluate the feasibility and noise reduction performance of the PSANM units, optimizing their parameters.

In particular, to assess global noise control in free space, the distance d between the primary noise source and the secondary sources was identified as a critical factor in determining the amount of noise reduction, given a fixed number of secondary sources. Typically, achieving a 10 dB attenuation in total noise power requires a separation normalized distance (d/λ) of up to 0.1 for a single secondary source, where λ is the acoustic wavelength [4].

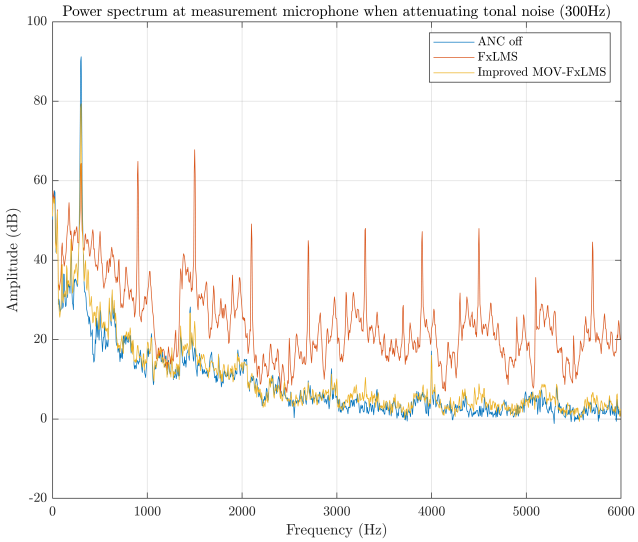


Fig. 6. The power spectrum of the measured tonal noise (150Hz), when FxLMS and the proposed MOV-FxLMS are tested under output saturation.

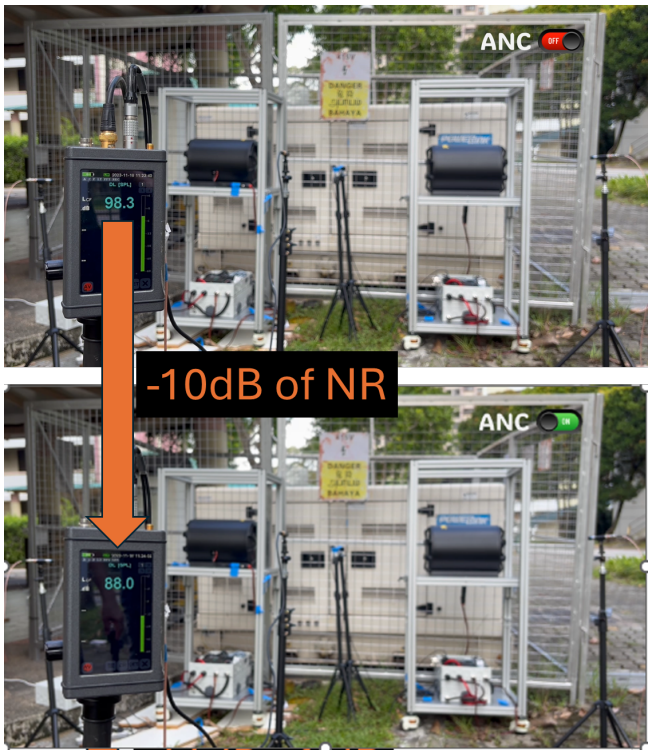


Fig. 7. The actual on-site setup of the PSANM units in front of the genset noise source achieved a global noise reduction of approximately 10 dB in the frontal working area of the genset.

Our laboratory experiments focused on replicating real-world noise conditions in a controlled setting. Fig. 5 illustrates the arrangement of two PSANM modules placed facing each other in front of a noise source played back by a subwoofer speaker. Various recorded construction machinery noises and simulated sounds were used to evaluate the performance of the

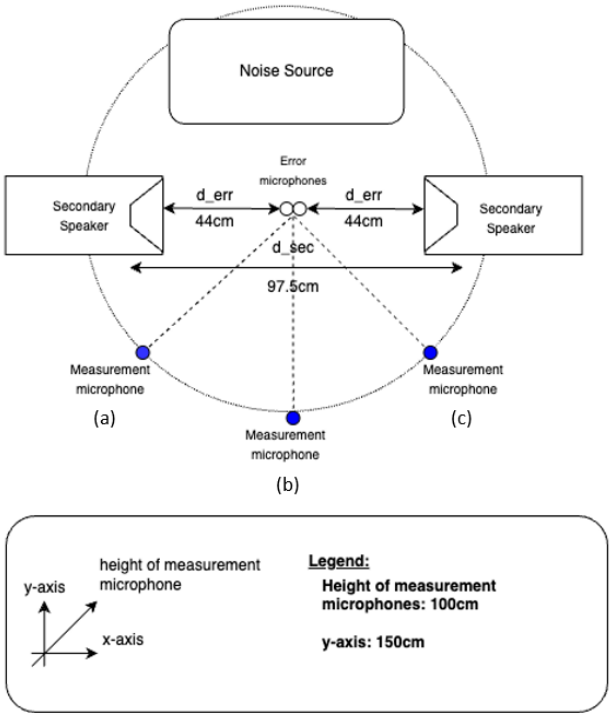


Fig. 8. The positions of the PSANM units, error microphones, and 3 different monitoring microphone positions (a), (b), and (c) are shown.

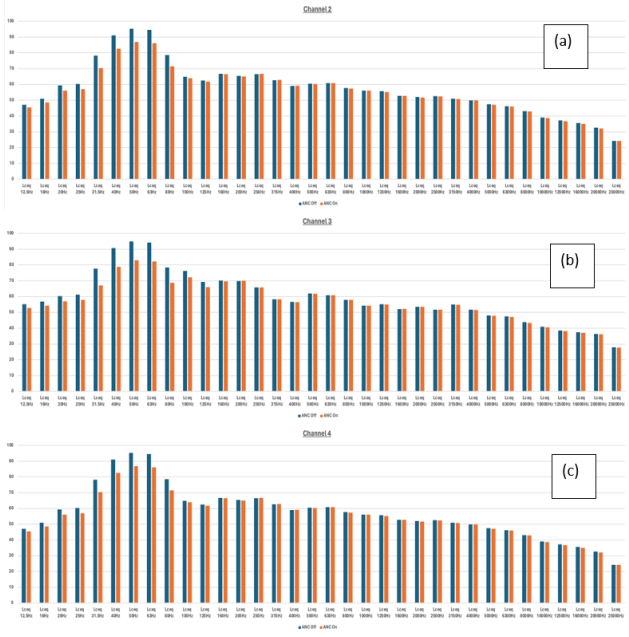


Fig. 9. The 1/3-octave noise reduction at 3 different positions (a),(b),(c) shows consistent and global noise reduction.

PSANM system under different algorithms, parameters, and different noise levels.

In an experiment, a simulated tonal noise at 150 Hz was used as the primary noise, with the constraint set near the peak power of the amplifier. Fig. 6 shows the frequency spectrum from the measurement microphone. It can be observed that the conventional FxLMS algorithm generates numerous harmonic signals at other frequencies due to its lack of output power limitation. This results in the control signal exceeding the output constraint when dealing with large amplitude disturbances, thus causing significant distortions in the anti-noise signal. In contrast, the proposed improved MOV-FxLMS algorithm constrains the output power, effectively reducing the disturbance and avoiding harmonics at other frequencies caused by output saturation.

IV. On-site Genset Noise Control

The predominant noise at construction sites is typically stationary and characterized by strong low-frequency components, particularly below 150 Hz, making it an ideal candidate for active noise control applications directly at the source. To test our PSANM system, we identified an operational generator set (genset) in an open field near a roadside to analyze real noise sources. This on-site deployment aimed to assess the system's performance in a typical construction environment, where noise levels can be unpredictable and influenced by various external factors.

The genset noise exhibits a high sound pressure level exceeding 90 dBC, often reaching around 100 dBC, with a significant resonance at 77 Hz and slight variations around this frequency. As shown in Fig. 7, we set up two PSANM units in front of the genset. Operating in feedback mode, the PSANM units achieved a global noise reduction of approximately 10 dB in the frontal area of the genset. Furthermore, we conducted comprehensive noise reduction measurements at various locations around the genset, as illustrated in Fig. 8. These measurements consistently demonstrated significant noise reduction, averaging close to 10 dB at low frequencies ranging from 31.9 Hz to 125 Hz. This performance was observed at distances of up to 1.5 meters from the genset and at a height of 1 meter above ground level, as shown in Fig. 9.

By demonstrating the PSANM system's efficacy in both controlled laboratory settings and on-site machinery noise control, we have laid the groundwork for broader adoption and further investigation into deploying the PSANM in a distributed manner around noise sources. Further minimization of PSANM hardware modules and optimization of the PSANM software architecture and algorithms will facilitate scaling the PSANM units for wide-scale deployment around larger noise sources.

V. Conclusions

In conclusion, the development and testing of our PSANM system in both laboratory and on-site deployments demonstrate its effectiveness in managing low-frequency noise emissions from construction machinery. In-situ tests showed a significant global noise reduction of 10 dB, effectively halving the perceived loudness. The use of the MOV-MFxLMS algorithm was crucial for maintaining power output constraints while

adapting to sudden noise disturbances. Both feedforward and feedback ANC have been implemented and can be extended to distributed active noise control.

The PSANM system can complement traditional passive noise mitigation approaches, offering broader and more effective noise control coverage. This patent-pending system has the potential for further enhancement and streamlining, enabling the creation of low-cost, mass-producible units for improved noise management and broader adoption in urban environments.

ACKNOWLEDGMENT

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (MOE-T2EP20221-0014)

REFERENCES

[1] E. Murphy and E. A. King, *Environmental Noise Pollution: Noise Mapping, Public Health, and Policy*, Elsevier, Amsterdam, 2014.

[2] World Health Organization (WHO), "Environmental noise guidelines for the european region," 2020, Retrieved from WHO.

[3] Yoshinobu Kajikawa, Woon-Seng Gan, and Sen M. Kuo, "Recent advances on active noise control: open issues and innovative applications," *APSIPA Transactions on Signal and Information Processing*, vol. 1, 2012.

[4] Bhan Lam, Woon-Seng Gan, DongYuan Shi, Masaharu Nishimura, and Stephen Elliott, "Ten questions concerning active noise control in the built environment," *Building and Environment*, vol. 200, pp. 107928, 2021.

[5] Xiaoyi Shen, Junwei Ji, Dongyuan Shi, Zhengding Luo, and Woon-Seng Gan, "The principle underlying the wireless reference microphone enhancing noise reduction performance in multi-channel active noise control windows," *Mechanical Systems and Signal Processing*, vol. 212, pp. 111284, 2024.

[6] Dongyuan Shi, Bhan Lam, Woon-Seng Gan, Jordan Cheer, and Stepehen J Elliott, "Active noise control in the new century: The role and prospect of signal processing," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Institute of Noise Control Engineering, 2023, vol. 268, pp. 5141–5151.

[7] Chuang Shi, Tatsuya Murao, Dongyuan Shi, Bhan Lam, and Woon-Seng Gan, "Open loop active control of noise through open windows," in *Proceedings of Meetings on Acoustics*. AIP Publishing, 2016, vol. 29.

[8] Cheng-Yuan Chang, Chun-Ning Chiou, and Sen M Kuo, "A complete design of smart pad that reduces snoring," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 3, pp. 649–656, 2023.

[9] Yi-Tsung Shen and Cheng-Yuan Chang, "A study of the microphone protection of active noise control for axial fan," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 1234–1239.

[10] D. Mylonas, A. Erspamer, C. Yiakopoulos, and I. Antoniadis, "Global control of propeller-induced aircraft cabin noise using active sound absorbers," *Journal of Sound and Vibration*, vol. 573, pp. 118213, 2024.

[11] C.-S. Oh, K.-D. Ih, J. Lee, and J.-K. Kim, "Development of a mass-producible anc system for road noise," *ATZ Worldwide*, vol. 120, pp. 58–63, 07 2018.

[12] A. Mostafavi and Y.-J. Cha, "Deep learning-based active noise control on construction sites," *Automation in Construction*, vol. 151, pp. 104885, 2023.

[13] C. H. Hansen and K. L. Hansen, *Noise Control: From Concept to Application*, CRC Press, 2021.

[14] M. Mostafa Mir, F. Nasirzadeh, S. Lee, D. Cabrera, and A. Mills, "Construction noise management: A systematic review and directions for future research," *Applied Acoustics*, vol. 197, pp. 108936, 2022.

[15] S. Sohrabi, T. Pàmies Gómez, and J. Romeu Garbí, "Suitability of active noise barriers for construction sites," *Applied Sciences*, vol. 10, no. 18, pp. 6160, 2020.

[16] European Environment Agency, "Noise," 2023, https://www.eea.europa.eu/en/topics/in-depth/noise.

[17] Raspberry Pi Foundation, "Raspberry pi 4 model b specifications," 2024, Accessed: 2024-08-05.

[18] STMicroelectronics, "Nucleo-h7s3l8," 2024, Accessed: 2024-08-05.

[19] Espressif Systems, "Espressif development kits," 2024, Accessed: 2024-08-05.

[20] Sen M. Kuo and D.R. Morgan, "Active noise control: a tutorial review," *Proceedings of the IEEE*, vol. 87, 1999.

[21] Dongyuan Shi, "Algorithms and implementations to overcome practical issues in active noise control systems," 2020.

[22] Zhengding Luo, Dongyuan Shi, and Woon-Seng Gan, "A hybrid sfanc-fxnlms algorithm for active noise control based on deep learning," *IEEE Signal Processing Letters*, vol. 29, pp. 1102–1106, 2022.

[23] Dongyuan Shi, Woon-Seng Gan, Jianjun He, and Bhan Lam, "Practical implementation of multichannel filtered-x least mean square algorithm based on the multiple-parallel-branch with folding architecture for large-scale active noise control," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 4, pp. 940–953, 2019.

[24] Feiran Yang, Jianfeng Guo, and Jun Yang, "Stochastic analysis of the filtered-x lms algorithm for active noise control," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2252–2266, 2020.

[25] Dongyuan Shi, Chuang Shi, and Woon-Seng Gan, "A systolic fxlms structure for implementation of feedforward active noise control on fpga," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[26] Dongyuan Shi, Chuang Shi, and Woon-Seng Gan, "Effect of the audio amplifier's distortion on feedforward active noise control," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 469–473.

[27] Dongyuan Shi, Woon-Seng Gan, Bhan Lam, and Shulin Wen, "Practical consideration and implementation for avoiding saturation of large amplitude active noise control," *Proc. 23rd Int. Congr. Acoust*, pp. 6905–6912, 2019.

[28] Yu Guo, Dongyuan Shi, Xiaoyi Shen, Junwei Ji, and Woon-Seng Gan, "A survey on adaptive active noise control algorithms overcoming the output saturation effect," *Signal Processing*, p. 109525, 2024.

[29] Dongyuan Shi, Bhan Lam, Woon-Seng Gan, and Shulin Wen, "Optimal leak factor selection for the output-constrained leaky filtered-input least mean square algorithm," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 670–674, 2019.

[30] Dongyuan Shi, Woon-Seng Gan, Bhan Lam, and Xiaoyi Shen, "Optimal penalty factor for the mov-fxlms algorithm in active noise control system," *IEEE Signal Processing Letters*, vol. 29, pp. 85–89, 2021.

[31] DongYuan Shi, Woon-Seng Gan, Bhan Lam, and Chuang Shi, "Two-gradient direction fxlms: An adaptive active noise control algorithm with output constraint," *Mechanical Systems and Signal Processing*, vol. 116, pp. 651–667, 2019.

[32] Dongyuan Shi, Woon-Seng Gan, Bhan Lam, Shulin Wen, and Xiaoyi Shen, "Optimal output-constrained active noise control based on inverse adaptive modeling leak factor estimate," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1256–1269, 2021.

[33] Woon-Seng Gan, Dongyuan Shi, and Xiaoyi Shen, "Practical active noise control: Restriction of maximum output power," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 1245–1249.

[34] C. K. Lai, D. Shi, B. Lam, and W.-S. Gan, "Mov-modified-fxlms algorithm with variable penalty factor in a practical power output constrained active control system," *IEEE Signal Processing Letters*, vol. 30, pp. 723–727, 2023.

PPHiFi-TTS: Phonetic Preserved High-Fidelity Text-to-Speech for Long-Term Speech Dependencies

Ravindrakumar M. Purohit, Dharmendra H. Vaghera, Arth J. Shah, and Hemant A. Patil
Speech Research Lab, DA-IICT, Gandhinagar, India
E-mail: {202321002, 202311034, 202101154, hemant_patil}@daiict.ac.in

***Abstract*—In recent years, TTS models have significantly improved the quality of synthesized speech, making it more natural-sounding and intelligible, particularly with the interaction of neural network-based models. Further, the neural vocoders-based on the Generative Adversarial Networks (GANs) have shown the potential to generate the raw speech waveform of the unseen speaker in the natural style. We introduce a novel architecture PPHiFiGAN, by combining the TTS model with the HiFi-GAN phoneme vocoder, where the Generator (G), and Discriminator (D) aim to enhance the synthesis quality and capture the in-depth phonetic nuances from the <word, phonemes> dictionary. This approach preserves the fine gradient details and captures the long-term speech characteristics. Our proposed method attained a Mean Opinion Score (MOS) of 4.23 with the LJSpeech recipe and 4.05 with the VCTK recipe, demonstrating the effectiveness of model in generating high-quality synthesized speech relative to proposed existing TTS architectures.**

***Index Terms*—Text-to-Speech (TTS), Generative Adversarial Network (GAN), Neural Vocoder, Mel Spectrogram.**

I. Introduction

Text-to-Speech (TTS) technology has undergone remarkable research and advances for the last several years. TTS enables the conversion of text into spoken natural speech, and helps to facilitate communication for individuals with better accessibility, and expressiveness with human interaction experience using digital devices. In recent years, TTS has been implemented in an autoregressive way of data distribution corresponding to the input, e.g., WaveNet [1], Tacotron [2]. However, autoregressive models could be more efficient in addressing the TTS quality. Later on, non-autoregressive TTS works on the motive of the teacher-student to increase the training-inference speed, and minimise latency issues of the autoregressive TTS, e.g., FastSpeech 2 [3] generates speech frames in parallel rather than sequentially, and transformer-based TTS [4] uses the multi-head attention mechanism to replace the Tacotron2 [5] to addresses a long-range dependency problem by directly connecting two input at different times using the self-attention mechanism. Due to the self-attention mechanism, it demonstrated the remarkable capability to generate high quality speech synthesis. Leveraging the ongoing research, TTS methods improved the naturalness and fluency for synthesized speech. However, due to the trade-off between speed *vs.* quality, non-autoregressive models frequently compromise on synthesis quality in favor of faster processing. Also, maintaining true parallelism by coherence, and fidelity in the synthesized speech degrades speech variability, such as different speaking rates, accents, or emotional expressions. To overcome these limitations, GAN-based TTS models can generate high quality natural sounding speech compared to the previous approaches. The GAN-based generalization of TTS shows state-of-the-art performance in terms of long input text, controllable utterances, and dynamic generative flow of speech synthesis.

As in the traditional methods [6], a preprocessor analyzes the input text, converting the sentence into words and *phonetic transcription*. Subsequently, the output of preprocessor fed to the Tacotron2 Mel generator (G). In this stage, the system transform the preprocessed text into a Mel spectrogram by predicting *temporal* and *pitch* (i.e., fundamental frequency, F_0), which generate the Mel spectrogram that involves using Fourier transform and Mel scale conversion. In context of adversarial training, G and D along with two additional adversarial losses, ensures the stability of the training process, and efficient data distribution in HiFi-GAN. During this phase, the synthesized Mel spectrogram and raw audio data are used to train the GANs architecture, where the G and D work together to train from the input Mel spectrogram into waveform data, which enable the model to produce a natural-sounding speech waveforms. To address this phonetic loss issue, we propose PPHiFi-TTS, which is an updated version of HiFi-GAN vocoder. In particular,

- We propose the PPHiFi-TTS using high-fidelity GAN vocoder. The proposed approach effectively works for a broader range of speech variations over the traditional TTS models, resulting in highly natural and intelligible speech with minimal distortion or loss of phonetic details.
- We combine Tacotron2 and HiFi-GAN vocoders for phonetically preserved waveform generation, where perceptual listening tests achives 4.16 MOS (Mean Opinion Score) on the scale of 5, which is 0.07 higher than the existing TTS models [3], [7], [8], [9].

The rest of the paper has following organization: Section II reviews related works in the recent TTS literature. Details of the proposed methodology are presented in Section III. Section IV outlines the experiments conducted in this study. We analyze and discuss the results obtained from these experiments in Section V. Finally, Section VI summarizes and concludes the paper, outlining potential future research directions.

II. Related Work

In 2016, Oord *et. al.* [1] proposed the *wavenet*, a deep neural network-based raw audio waveform generation. The

model architecture is auto-regressive, built on residuals and skip connections. Wavenet generated one sample from the previous samples, making it highly realistic, and natural prosody and intonation, significantly outperforming the traditional TTS systems. Later on, Arik *et. al.* [10] proposed the *Deep Voice*, which is built on five submodules: grapheme-to-phoneme model, segmentation model, phoneme duration model, fundamental frequency (F_0) model, and audio synthesis model. The audio synthesis submodule was derived from the Wavenet architecture, works autoregressively, and finds better performance in real-time inference, which runs 2x faster than in real-time (with MOS 3.94). Wang *et al.* [2] proposed the Tacotron with the end-to-end pipeline, where earlier WaveNet [1], and Deep Voice [10] were trained independently. Tacotron used the *seq2seq* architecture to map the text sequence to the Mel spectrogram frames; as an advantage, it reduces the complexity of the Wavenet and the Deep Voice, however, relies on the Griffin-Lim algorithm, which is less effective than the vocoders, such as WaveNet. Following these, Shen *et al.* [5] introduced Tacotron 2, combining Tacotron's *seq2seq* model with WaveNet as a vocoder, enhancing audio quality using WaveNet's waveform synthesis architecture. Arik *et al.* [10] then developed Deep Voice 2, an improved version of Deep Voice, incorporating multiple speakers and languages by utilizing separate modules for each task, achieving higher quality and versatility in speech synthesis, in context of Deep Voice 2, there is the phoneme block, which enhances the phonemes duration and separation in TTS. Ren *et al.* [3] presented FastSpeech 2, focusing on speed and efficiency using a non-autoregressive model, significantly reducing the inference time and producing high quality speech without relying on autoregressive sampling. At last, Kim *et al.* [6] proposed Glow-TTS, a flow-based generative model for TTS, which offers robust and expressive speech synthesis by directly modeling the distribution of speech features. FastSpeech 2 [3] builds upon FastSpeech with improved architecture and performance, enhancing controllability over speech synthesis parameters. However, it requires larger datasets and longer training times than the simpler models. Tacotron [2] can handle long sentences and capture prosody. However, careful training, hyperparameter tuning, and Prone are required to generate monotonic speech output. While Tacotron2 [5] improved the attention mechanism for better alignment between text and speech, it also has higher naturalness and quality in synthesized speech.

A. Speech Synthesis Model: Tacotron2

The Tacotron2 [5] architecture (as shown in Fig. 1) starts by entering textual content that needs to be converted into speech. This text is converted into person embeddings, converting the discrete textual symbols into continuous vectors that may be processed via the neural community. The person embeddings are then handled via a 2-layer fully-linked pre-Internet, which acts as a bottleneck layer to reduce the dimensionality and version complexity. The output of the pre-Internet is fed into

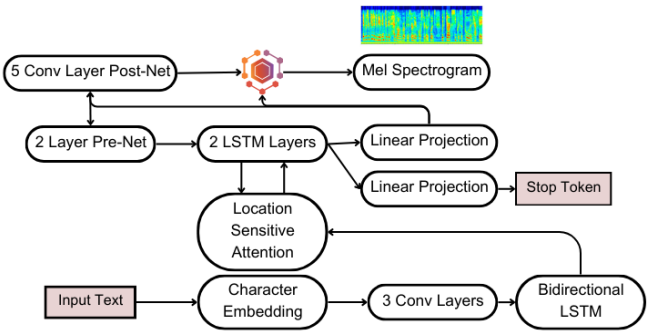


Fig. 1. Architecture of the Tacotron2 Model. After [5].

the encoder, consisting of three convolutional layers followed by a bidirectional Long Short-Term Memory (LSTM).

The convolutional layers extract nearby features from the statistics, while the LSTM captures sequential dependencies within the facts. A place-sensitive attention mechanism then processes the encoder output. This mechanism facilitates the model awareness of unique elements of the input series when producing every output time step, making the alignment between the input textual content and the output speech stronger and stronger. The attention mechanism is fed into LSTM layers for decoding. These layers process both the eye context and the outputs from the preceding steps to generate new frames of audio facts. A linear projection layer converts the output of the decoder LSTM layers into expected Mel spectrogram frames. Stop tokens, indicating the end of the collection, are predicted by another linear projection layer.

The five-layer convolutional post-net refines the expected Mel spectrograms to enhance the excellent output speech and speed up the convergence of the version. The last output is a Mel spectrogram, a visual representation of the spectrum of frequencies of the speech sign as it varies with time. A vocoder can then convert this into an audio waveform. This architecture enables Tacotron2 to generate high quality, natural-sounding speech from textual content. It finds extensive use in textual content-to-speech systems due to its effectiveness and efficiency [5].

B. Generating High-Quality Speech Waveforms: HiFiGAN

HiFiGANs is a GAN-based neural vocoder model [11]. It's training is done using a G that takes the random noise or input features and converts it through the upsampling structure. which captures patterns, and increases overall fidelity. Fidelity means the degree of exactness to which sound is reproduced into the speech waveform with the same high quality as natural human speech, whereas the D works as the evaluator to discern between the natural speech waves sourced from the training dataset or the G generates it. D delivers feedback to G through this classification process and guides its iterative improvement toward producing more realistic outputs for the two-stage TTS. It plays a significant role in the two-stage TTS architecture. It takes the Mel spectrogram [12] as input and produces the corresponding time-domain waveform. The trade-offs between

the G and D leads to the generation of high quality speech that resembles authentic human speech patterns.

1) Generator (G): The G is a critical component of its architecture, designed for efficient and high-fidelity speech generation. G uses fully-convolutional neural network that takes a Mel spectrogram as input and upsamples it through transposed convolutions until the output sequence's period suits the periodic resolution of raw waveforms. The G incorporates a Multi-Receptive area Fusion (MRF) module designed to study styles of numerous lengths in parallel [12]. G module captures numerous receptive field patterns by calculating the sum of outputs from a few residual blocks, each with distinct kernel sizes and dilation rates. The G is extraordinarily configurable, with adjustable parameters, which include the hidden measurement (h_u), kernel sizes (k_u) of the transposed convolutions, and kernel sizes (k_r) of the MRF modules. This flexibility allows for generating a sample with nearly equal long-term dependencies and consecutive input patterns.

2) Discriminator (D): The D in the HiFi-GAN architecture is essential in distinguishing between natural *vs.* synthesized speech waveforms. It contains multiple interconnected sub-components, each sub-component made to capture specific components of speech fidelity. The D is divided into two parts, namely, Multi-Period Discriminator (MPD) and Multi-Scale Discriminator (MSD), which interactively contribute to the discriminative capability of the version. The MPD is designed to analyze speech patterns through several temporal resolutions.

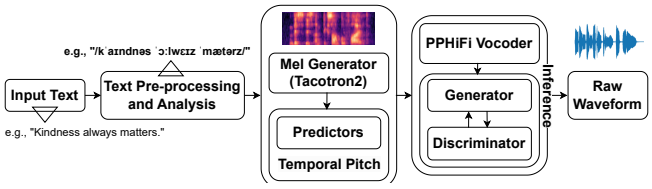


Fig. 2. Proposed Architecture: TTS Using Tacotron2 and PP-HiFi Vocoder.

III. Proposed Methodology

As shown in fig. 2, in proposed architecture, the text pre-processing phase employs phonetics embedding over the word level using tokenization, which is created based on two state-of-the-art models, namely, Tacotron2 [5], and HiFiGAN [11]. As shown in Fig.3, the proposed model starts with the input text undergoing the text-preprocessing and analysis phase, where text is tokenized and phonetic embedding is applied. Then, Tacotron2 converts the processed text into a Mel spectrogram. This spectrogram serves as input to the PPHiFi vocoder, which consist of predictors for various acoustic featuresm such as temporal and pitch (i.e., F_0) characteristics. The generator of vocoder, then synthesizes the raw speech waveform, and a discriminator helps to refine the output to ensure high-fidelity and naturalness; this architecture effectively captures and preserves the phonetic nuances of the input text, resulting in a natural-sounding speech output.

A. Phonetic Embedding on the Linguistic Features

The input text passes through pre-processing [13]. The proposed workflow is designed to increase the richness of the phonetics in the input textual content before going to Mel spectrogram generation. As shown in Fig.3, it begins with segmenting textual content into sentences, and tokenizing those sentences into words for further processing. Again, words break down into possible units as phonemes. However, identifying accents in the text is crucial for determining the language of the input. Thus, language identification is used to establish the text's language, essential for accurate phonetic transcription and converting textual content into phonemes. Then, last, phonemes are combined with words. e.g., words *W*, and phonemes *ph* as $<W \rightarrow ph>$. At last, the process is combined, where text is prepared as input for the Tacotron2 model, and goes into the subsequent stage of the Mel spectrogram prediction.

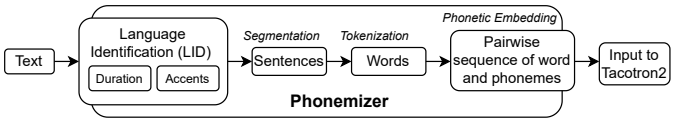


Fig. 3. Proposed Phonetic embedding on the linguistic features.

B. Algorithm

As shown in the *Algorithm* 1, synthesizing speech waveform from enter text the usage of the PPHiFi-TTS Vocoder and Tacotron2 TTS algorithm with Phonetic Preservation. Here, algorithm takes textual content as input and produces a synthesized speech waveform as output. The entered textual content is tokenized into individual devices, and those tokens are transformed into phonetic or linguistic capabilities using a phonetic renovation version denoted as Phi. This step aims to extract phonetic data from the text. Mel spectrogram generation: The phonetic features acquired from the previous step are then fed into Tacotron2, a textual content-to-spectrogram synthesis version, to generate a Mel spectrogram illustration of the speech.

IV. Experiments

To assess the efficacy of the proposed approach outlined in Algorithm 1, we conducted experiments using the LJ Speech dataset [14]. To train the PPHiFi-TTS, we employed Mel spectrogram feature extracted from the unique audio recordings. These Mel spectrograms served because of the entire representation for each the WaveNet (the in-constructed vocoder in Tacotron2) and Tacotron2 networks. During training, we utilized a batch size of 8 and set the values of $\beta 1$ and $\beta 2$ to 0.8 and 0.99, respectively. Following default configurations, the Fast Fourier Transform (FFT) size was maintained at 1024, even as the hop length was set to 256. At the final output stage, the sampling frequency and number of the channel changed to a 22050 Hz and 1, respectively, and the learning rate is 0.0001. These settings were chosen based on fast convergence to minimize loss.

Algorithm 1 Proposed PPHiFi-TTS Vocoder.

Input: Input Text
Output: Synthesized Speech Waveform
Preprocessing:
- Tokenize input text $T = \{t_1, t_2, ..., t_n\}$
- Convert tokens into phonetic or linguistic features using a phonetic preservation model Φ.
$P = \Phi(T)$
- Mel Spectrogram Generation:
- Use Tacotron to convert phonetic features into Mel Spectrogram,$M = \text{Tacotron}(P)$
Speech Synthesis:
- Phoneme Dictionary Lookup:
- Map phonetic features to phonemes using a phoneme dictionary, D
$Phonemes = D(P)$
- Waveform Synthesis:
- Use HiFi-GAN vocoders with phoneme information to convert Mel Spectrogram into waveform $W = \text{HiFi-GAN}(M, Phonemes)$
Post-processing:
- Waveform Enhancement:
- Apply post-processing techniques (denoising, equalization, etc.) on phonetic features
- Apply noise reduction techniques selectively based on phonetic context
Output:
- Synthesized Speech: Output the synthesized waveform,W

A. Dataset Used

In our proposed architecture, we used the LJSpeech recipe [14], which is widely used in speech synthesis and processing due to its high quality and small size. The dataset consists approximately 24 hours of speech recordings and converted text transcription. Speech data was recorded from a single female speaker at a sampling rate of 22050 Hz. For training data, we shuffled the data, and divided them into 3 sets, e.g., 12500, 800, and 300 samples, respectively, for training, validation, and testing purposes. To check the robustness of our proposed model for different accents of speech, we used the VCTK dataset [15], which contains the data uttered by 109 native speakers of English with multiple accents. The speech was selected randomly from the newspaper, and rainbow passage. Each speaker reads a specific set of sentences from newspapers. The dataset uses a 48 kHz sampling rate, and a single channel with 16-bit quantization.

B. Simulation Setup

We ran computer simulations using Python 3.8, Numpy 1.22.0, and Torch 2.2.1. We sourced the dataset from an open-source repository [14]. The entire process occurred on an MSI computing device with an Ubuntu 22.04 64-bit OS equipped with i7-12700, 32GB of RAM, and Duo NVIDIA 8 GB GTX 1080 armor graphics.

V. Experimental Results

In the comparative analysis illustrated in Fig. 4, spectrograms from three different sources are examined Fig. 4-a showcases the original LJSpeech dataset, while Fig. 4-b presents the spectrogram of speech synthesized using Tacotron 2 [5]. Fig. 4-a represents the original pattern from the LJSpeech dataset. Fig. 4-b, then again, shows the spectrogram of the Tacotron2 generated speech signal. It is found that the Real-

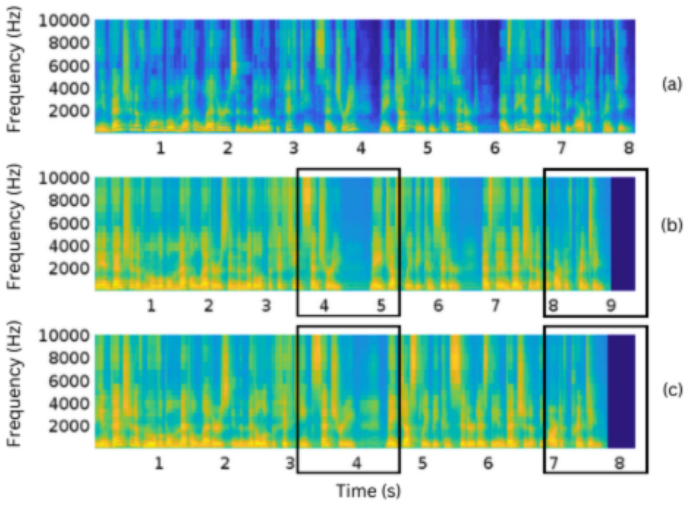


Fig. 4. Mel spectrogram comparison between the original speech sample, Parallel WaveGAN, and Proposed PPHiFi-TTS: (a) human speech sample from the LJSpeech corpus [14] *Vs.* (b) Parallel WaveGAN synthesised TTS sample [7] *Vs.* (c) proposed PPHiFi-TTS synthesized TTS sample.

Time Factor (RTF) of the synthesized TTS is about 12.5 % higher than the length of the speech itself, indicating that Tacotron2 takes more time to generate speech for a given quantity of text. In contrast, Fig. 4-c shows the spectrogram from the proposed PPHiFi-TTS. It demonstrates correct power, pitch (or fundamental frequency, F_0), and duration, indicating an extra green and specific speech synthesis. The highlighted region in Fig. 4-c indicates the spectrogram visualization, which would provide a more comprehensive insight into the frequency components of the signal over time, suggesting that PPHiFi-TTS can predict and reconstruct the waveform more accurately than Tacotron2. In Fig. 5(b), a notable divergence from the original speech sample (Fig.5-a) is observed as highlighted Fig.5-b, the parallel WaveGAN turns the speech around the 4-seconds into garbled noise, which is slightly less effective than our proposed PPHiFi-TTS, as shown in Fig.5-c. PPHiFi-TTS effectively reconstructs the wave signal near the original speech sample Fig. 5-a, making it more recognizable, and closely resembling the original human speech pattern. This demonstrates the effectiveness of PPHiFi-TTS in synthesizing more accurate and coherent speech samples compared to Parallel WaveGAN. Also, we perform the pitch assessment to analyze the variability, and dispersion of pitch within both the generated, and original speech signals. The variance in pitch value over time provides information on the overall stability

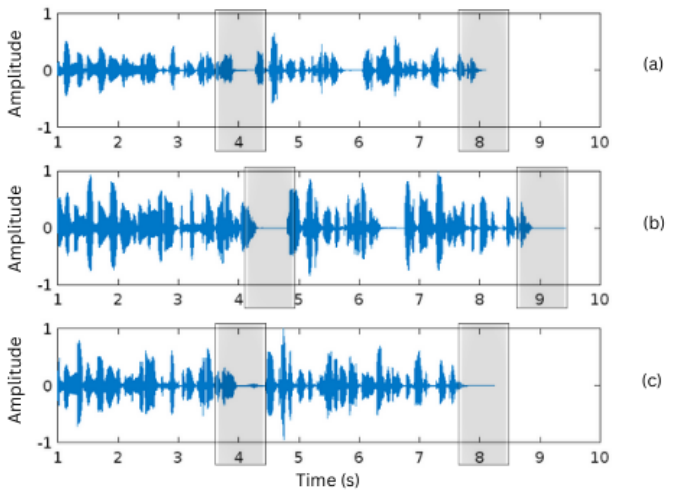


Fig. 5. Amplitude *vs.* time plot for original speech sample, speech samples obtained from Parallel WaveGAN, and from proposed PPHiFi-TTS: (a) Natural speech sample from the LJSpeech corpus [14] *Vs.* (b) Parallel WaveGAN synthesised TTS sample [7] *Vs.* (c) Proposed PPHiFi-TTS synthesized TTS sample.

of the pitch when the speaker conveys information about emotions, gender, and age. Standard deviation σ, skewness γ, kurtosis, and average dynamic time warping (DTW) for the comparison of speech quality between original speech *vs.* synthesized speech. As mentioned in Fig. 5, all moments of the synthesized speech are very close to the original speech, helping to identify differences in pitch usage, and its perceptual effects under various conditions [16].

A. Performance Metrics

We evaluated the performance of the generated samples using both objective and subjective metrics: Perceptual Evaluation of Speech Quality (PESQ), Mel Cepstral Distortion (MCD), and Signal-to-Noise Ratio (SNR) for objective assessment, and Mean Opinion Score (MOS) for subjective evaluation. Our model includes two TTS recipes namely, LJSpeech and VCTK. The results, shown in table I, demonstrate performance comparable to existing architectures such as Parallel WaveGAN [7], Deep Voice 2 [10], WaveGlow [8], FastSpeech 2 [3], and FastSpeech 2+HiFiGAN [9].

1) Subjective Measures:

We measure Five-point Mean Opinion Score (MOS) using TextVOX to compare the quality of synthesized speech from TTS [17], with R_n and N representing individual ratings and total ratings, respectively. Parallel WaveGAN received a MOS score of **4.16**, while PPHiFi-TTS achieved a higher score of **4.23** with the LJSpeech recipe and **4.05** with the VCTK recipe, indicating superior speech quality compared to Parallel WaveGAN. The MOS testing was conducted internally at the Speech Research Lab,DA-IICT in a controlled and quiet environment, with 18 subjects aged 18–24, having a 1:5 female-to-male ratio, and no subjects with listening disabilities.

$$\text{MOS} = \frac{\sum_{n=1}^{N} R_n}{N}. \tag{1}$$

2) Objective Measures:

1) ***Perceptual Evaluation of Speech Quality (PESQ):*** It is a standard used to measure the quality of speech as perceived by humans, especially in telecommunications and VoIP (Voice over Internet Protocol) systems. Developed by the ITU-T, where higher ratings indicate high quality [18]. In particular,

$$\text{PESQ} = 4.5 - 0.1 d_{\text{sym}} - 0.0309 d_{\text{asm}}, \tag{2}$$

where d_{sym} indicates symmetrically reduced speech quality; d_{asym} indicates asymmetrically reduced speech quality. As shown in Table I trajectories, PPHiFi-TTS achieves 1.11±0.11. In particular, PPHiFi-TTS exhibits a stronger resemblance to the original speech sample compared to existing TTS models.

2) ***Mel Cepstral Distortion (MCD):*** MCD measures the difference between two sets of Mel Frequency Cepstral Coefficients (MFCCs), often used in speech processing. It quantifies the spectral distortion between two signals or frames, providing a metric for evaluating speech quality or comparing different utterances, i.e.,

$$MCD = \frac{\sqrt{\sum_{t=1}^{N} \left(\sqrt{\sum_{i=1}^{D} (A(t,i) - B(t,i))^2} \right)^2}}{N}, \tag{3}$$

where N and D convey the total number of coefficients and dimension count of the coefficients, respectively, and $A(t,i)$ and $B(t,i)$ represent the Mel cepstral trajectories at varying time points and dimensions, respectively: As can be seen from Table I; MCD values of the 16.74±0.22, where as the Parallel WaveGAN have the MCD 19.36±0.50. Again, PPHiFi-TTS produces results similar to the original speech signal.

3) ***Signal-to-Noise Ratio (SNR):*** It is a measurement used to quantify the level of a speech signal. SNR is the ratio of the audio signal is compared to any unwanted background noise (dB),i.e.,

$$\text{SNR}_{\text{dB}} = 10 \cdot \log_{10} \left(\frac{\text{Speech Power}}{\text{Noise Power}} \right). \tag{4}$$

B. Diversity and Controllability

Diversity and controllability are crucial aspects of the production of the TTS system. It promotes realness and expressive speech via various accents, dialects, and speaking styles. Controllability helps to customize the TTS output to individual preferences per the needs of certain accents and pronunciation styles, e.g., internationalization and localization. As we do not make changes on the TTS encoder, but to make a stable enhancement on the TTS, the association with the rich phonetic dictionary, which contains the specific pitch, duration, and emphasis phonetic prosodic information, which is essential for the natural speech. However, the rhythm of the controllable prosody involves complex linguistic patterns, which is still a challenge.

TABLE I
COMPARISON OF PP-HIFIGAN TTS WITH OTHER EXISTING ARCHITECTURES FOR DEEPFAKE AUDIO. THE CONFIDENCE INTERVAL (CI) OF THE MOS IS THE 95%

Architecture	PESQ (↑)	MCD [dB] (↓)	SNR (↓)	MOS (↑)
Parallel WaveGAN [7]	2.85	**2.02**	2.67	4.16
Deep Voice 2 (80-layer WaveNet) [10]	0.98	5.67	3.29	3.53
WaveGlow [8]	1.90	5.13	3.02	3.9
FastSpeech 2 [3]	2.43	6.80	5.14	3.83
FastSpeech 2+HiFiGAN [9]	2.43	2.68	2.48	4.02
PP-HiFi TTS (LJ Recipe)	**3.58**	16.74	**2.08**	**4.23**
(Proposed)	**(±0.09)**	(±2.29)	**(±0.11)**	**(±0.50)**
PP-HiFi TTS (VCTK Recipe)	2.24	18.141	3.1548	4.05
(Proposed)	(±0.11)	(±2.29)	(±0.24)	(±0.67)

VI. SUMMARY AND CONCLUSIONS

In conclusion, our proposed architecture presents a promising avenue to advance TTS technology by integrating TTS models with HiFiGANs phoneme vocoders. Our model aims to utilize the power of GANs to enhance synthesis quality and capture intricate phonetic nuances, resulting in more natural and intelligible speech output. One of the strengths of the proposed work is the results shown with very low computational resources, in particular, 16 GB of GPU only. Some languages contain phonemes or phonotactic patterns that are very difficult to identify. To accurately ensure the TTS of this type of text, we need extensive training data and a rich cross-linguistic phonetic dictionary that pronounces them. Also, sounds often blend in natural speech due to coarticulation, where the primary sound influences another sound. Capturing the nuances of coarticulatory effects accurately is essential for smooth phonetics, and the preservation of TTS systems. However, accurately modeling and incorporating these nuances into TTS remains an open challenge within the research community. In the future, our goal is to create TTS models that preserve phonetic details while making the models more compact by enabling higher parameterization. Also, we would like to look at latency, which helps with accessibility and smooth conversations. Further, objectively evaluating the speech quality of TTS voice remains a technologically challenging open research problem.

REFERENCES

[1] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499,2016. {Last Accessed : 1st August, 2024}*.

[2] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017 {Last Accessed : 1st August, 2024}.

[3] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020 {Last Accessed : 1st August, 2024}.

[4] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Honolulu, USA.*, volume 33, pages 6706–6713, 2019.

[5] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada*, pages 4779–4783. IEEE, 2018.

[6] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada*, 33:8067–8077, 2020.

[7] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. pages 6199–6203, 2020.

[8] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom*, pages 3617–3621. IEEE, 2019.

[9] Dan Lim, Sunghee Jung, and Eesung Kim. JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech. *arXiv preprint arXiv:2203.16852*, 2022{Last Accessed : 1st August, 2024}.

[10] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. *Advances in Neural Information Processing Systems (NIPS)*, 30, Long Beach Convention Center, 2017.

[11] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems(NIPS), Vancouver, Canada*, 33:17022–17033, 2020.

[12] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. MelGAN: Generative adversarial networks for conditional waveform synthesis. *Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada,*, 32:2672–2680, 2019.

[13] Mathieu Bernard and Hadrien Titeux. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958, 2021.

[14] Keith Ito and Linda Johnson. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/, 2017 {Last Accessed : 1st August, 2024}.

[15] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), 2019. {Last Accessed : 1st August, 2024}.

[16] Oliver Niebuhr and Radek Skarnitzl. Measuring a speaker's acoustic correlates of pitch but which a contrastive analysis for perceived speaker charisma. In 19th *International Congress of Phonetic Sciences: Endangered Languages, and Major Language Varieties (ICPhS), Melbourne, Australia*, pages 1774–1778. Australasian Speech Science and Technology Association Inc, 2019.

[17] Robert Streijl, Stefan Winkler, and David Hands. Mean Opinion Score (MOS) revisited: methods and applications, limitations, and alternatives. *Multimedia Systems*, 22:213–227, 03 2016.

[18] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Salt Lake City, USA*, volume 2, pages 749–752. IEEE, 2001.

GPGAN-VC: Enhancing Voice Conversion using Gradient Penalty

Ravindrakumar M. Purohit, Dharmendra H. Vaghera, and Hemant A. Patil
Speech Research Lab, DAIICT, Gandhinagar, India
E-mail: {202321002, 202311034, hemant_patil}@daiict.ac.in

***Abstract*—This research introduces Gradient Penalty Generative Adversarial Network-Voice Conversion (GPGAN-VC), an upgraded version of the StarGAN-VC framework-a leading non-parallel VC method. Unlike traditional VC methods, which require parallel data, non-parallel multi-domain approaches, such as StarGAN-VC, have gained popularity due to their ability to function without this limitation. The main innovation in GPGAN-VC is including contrastive learning techniques in the GPGAN-VC Discriminator (D). Instead of using the typical structure, our approach introduces a gradient penalty mechanism. This modification not only increases the consistency of the training process but also effectively prevents over-fitting, resulting in a more robust and reliable VC model. We evaluated our generated sample using both subjective and objective metrics. Subjectively, our sample's Mean Opinion Score (MOS) was 3.71 on a 5-point scale, achieving a higher performance than existing VC architectures. Objectively, we assessed the genreted samples with Perceptual Evaluation of Speech Quality (PESQ), Mel Cepstral Distortion (MCD), Modulation Spectra Distance (MSD), Signal-to-Noise Ratio (SNR),Fréchet Audio Distance (FAD), and Root Mean Square Error (RMSE) metrics, which scores 1.13, 5.70, 0.08, -0.84, 1.78 and 0.04 respectively, which demonstrating that our model outperforms best-in-class non-parallel VC models.**

***Index Terms*—Voice Conversion (VC), Gradient Penalty (GP), Generative Adversarial Networks (GANs), Style Transfer, Non-Parallel VC.**

I. Introduction

Voice conversion (VC) has been attempted using recent Deep Learning (DL) methods [1], [2]. A DL-based technique is applied in VC and Text-To-Speech (TTS) as a speech synthesizer. The VC technique converts the voice of source speakers into the target speakers. The tonal quality of the output speech reflects the target speaker, but the prosody stays consistent with the speech features of source speakers. There is also TTS, which generates speech from the text. Unlike VC, TTS uses text data rather than speech features of source speakers.The patterns in the TTS depend on the training dataset, so it becomes difficult to regulate the tonal quality of speech by converting the characteristics of the speech features of source speakers. Before the advent of DL-based methods, traditional speech conversion relied on techniques, such as the Hidden Markov Model (HMM) [3], and the Gaussian Mixture Model (GMM) [4]. Later on, Recurrent Neural Networks (RNNs) were used to build VC-TTS systems. In 2014, Goodfellow [5] proposed Generative Adversarial Networks (GANs).One of the critical advancements encouraged by GANs is the evolution of end-to-end speech processing systems,e.g., WaveGANs, Parallel WaveGANs, and HiFi-GANs. Later on, Kaneko *et al.* [6] implemented the non-parallel VC using CycleGAN [7].

In the context of VC, the conversion effect of the target speaker is essential. Non-parallel VC techniques have a disadvantage in this regard compared to parallel VC. Achieving optimal stability between VC and speech synthesis quality is critical, particularly with respect to diverse conditions in non-parallel datasets for mapping source speakers to target speakers. To avoid this, the RNN used by Sun *et al.* [8] is a specialized Artificial Neural Network (ANN) designed to manage sequential data by preserving a memory function. Furthermore, an exemplar-based framework based on a notable approach grounded in non-negative matrix factorization (NMF) has likewise shown remarkable efficacy [9], [10]. Later on, a novel framework that uses conditional variational autoencoders (CVAEs) is used to integrate probabilistic modeling within encoder-decoder architectures seamlessly [11] . As discussed earlier, the VC approach stands out for its independence from parallel mapped data issues and compatibility with unaligned corpora. However, it is again a typical challenge with Variational Autoencoders (VAEs), including, in this context, the trade-offs and over smoothing in VC outputs. This can degrade converted speech quality, often leading to a fuzzy or unclear sound.

A GAN architecture trains two neural networks (NN) simultaneously to compete against each other and generate more real-like data for a specific application requirements. GANs have fueled the widespread adoption of speech processing applications, such as VC e.g., CycleGAN-VC [6], and StarGAN-VC [12]. StarGAN to enhance a technique that enables non-parallel M:M VC, the StarGAN-Picture Mapping [13] framework is an exploration that can potentially address the weaknesses of VAEs. This framework combines GANs architecture and offers a promising approach for overcoming the limitations of VAEs to CVAEs. StarGAN is a notable framework that can learn mappings among multiple domains, enabling non-parallel M:M VC. Next, we outline the major contributions in the development of the GPGAN-VC model.

1) We propose the gradient penalty mechanism in the Discriminator (D) network, which penalizes the norm of the gradient of the D's output w.r.t. its input; we minimized by adding the 1-Lipschitz and effectively mapping between the VC speaker features.
2) The additional GP is a significant step forward in addressing the issues of traditional GANs, such as *mode collapse* and oscillatory behavior during training. This

mechanism provides a stable training process, offering reassurance in the face of these challenges.

Remaining paper is structured as follows: Section II presents related work in the VC literature. Section III gives the details of proposed GPGAN-VC architecture and discusses the model details. Section IV presents standard experiments using the experiment setup. Finally, Section VI concludes the paper along with limitations and potential future research directions.

II. Related Work

In the domain of the VC, Hsu *et al.* [14] presented learning probabilistic models VAEs with latent variables using an unsupervised approach. VAEs generate similar data from the input, which helps generate diverse and realistic samples. Later on, CVAEs were introduced to transform speech characteristics [14], [15], [16]. From one speaker to the next, we maintain the linguistic content with the conditional variable, c. CVAEs are used to model the latent space distribution to acquire the essential features of the source speaker's voice.Later on, GANs [5] came with the semi-supervised potential to augment the quality of the synthetic data in a competitive manner, where two networks (the G and D) are working on a mini-max game for effective training as Eq. (1), where G tries to reduce the cost function $V(D,G)$, it encourages G to produce the samples that are indistinguishable from the real samples and make $D(G(z))$ close to 1. On the other hand, D is trying to maximize the cost function, which helps to assign a low probability to a fake sample compared to real samples, i.e.,

$$\mathrm{V(D,G)} = \mathrm{E}_{x\sim p_{\mathrm{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z))). \tag{1}$$

In particular, CycleGAN [7] enables the image transformation with non-parallel data. The generator (G), aims to create realistic images of the target-domain, while the D classifies these images as real or fake. Later on, CycleGAN-VC [6] presented the transfer of acoustic vocal characteristics from domain A to domain B, where dimension of $A \in \mathbb{R}^{M\times N}$ and dimension of $B \in \mathbb{R}^{M\times O}$. However, the motivation of CycleGAN-VC, Eq. (2) lies in understanding the mapping of G from $A \rightarrow B$, and F is the reverse mapping of the features, i.e, $A \leftarrow B$.

$$\begin{aligned} L_{\mathrm{cyc}}(G,F) = E_{a\sim p_A(a)}[\|F(G(a)) - a\|_1] \\ + E_{b\sim p_B(b)}[\|G(F(b)) - b\|_1]. \end{aligned} \tag{2}$$

CycleGAN-VC allows the production of the VC with sufficient training steps, but it's limited to learning from parallel One-to-One VC. Later on, StarGAN-VC [12], which allows multiple source and target speakers simultaneously with non-parallel, M:M VC scenarios.

As shown in Fig. 1, the generator G takes an acoustic feature sequence X and a target attribute label C as input, producing an acoustic feature sequence $G(X,C)$. The speech attribute is represented using a one-hot vector, where each vector has a single one in the position corresponding to the class of the attribute and zeros elsewhere. For example, when considering speaker identities, each one-hot vector component

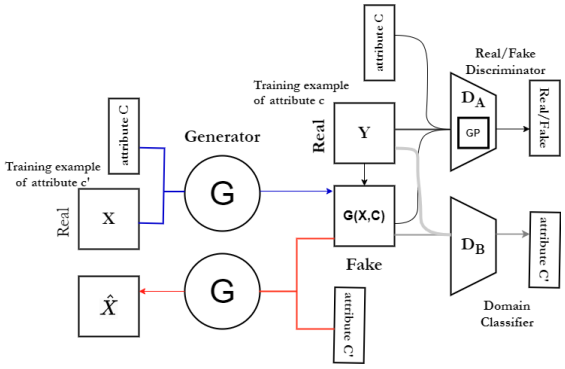


Fig. 1. StarGAN training Architecture. After [17].

corresponds to a distinct individual speaker. StarGAN-VC aims to create realistic speech features corresponding to attribute C. To achieve this, it uses the 0/1 discriminator D_A, similar to CycleGAN, and a domain classifier D_B, which predicts the input classes.

The discriminator D_A generates a probability $D_A(Y,C)$ indicating whether an input Y is a real speech feature. In contrast, the domain classifier D_B generates class probabilities $p_{D_B}(C|Y)$ for Y. In the context of GANs, the training loss for the discriminator D and the generator G is crucial. The goal is to minimize the adversarial loss for both the D and the generator. A small value of $L_D^{adv}(D_A)$ means that the D correctly classifies fake *vs.* real speech features. Similarly, a small value of $L_G^{adv}(G)$ means that the generator successfully deceives the discriminator by generating speech features that are misclassified as real. Dhar *et al.* [18] introduced a guided loss function that captures the different latent feature representations between the source and target domains. They also used feature mapping loss and improved reconstruction loss to preserve the linguistic content of source speech during the speech feature mapping (FM) and to maintain the contextual information of the input speech, respectively. partial least squares regression based totally on GMMs. This technique applied partial least squares regression to establish a mapping between supply and goal speech features, leveraging the statistical modeling talents of GMMs.

III. The Proposed GPGAN-VC

A. Dataset Used

The VCTK dataset [19] is widely used to gather speech data for research in the VC. It includes around 44 hours of speech recordings from 110 different speakers, where the speaker recites approximately 400 sentences from various sources, including newspapers and the *Rainbow Passage* from the Speech Accent Archive [20].

B. Generator (G)

G network aims to generate synthetic data, which faithfully preserves the underlying distribution of real data and its inherent features. G is supplied with two inputs. Real Data (A) constitutes authentic data samples that are extracted from

the domain, acting as the foundation for generating artificial data [12]. Attribute (C) offers contextual or auxiliary insights related to the data, potentially influencing the generation process or enhancing the synthesized output. Using the provided inputs, the generator produces synthetic data denoted as G(A, C), wherein G signifies the resultant output synthesized from the real data and the supplemental attribute.In the generator, the repeated number parameter determines the quantity of residual blocks. After this, two up-sampling layers have a transpose convolution layer, an instance normalizations layer, and a ReLU activation function [21]. The generator architecture starts and ends with the convolution for shaping and refining the input and output representations, increasing the model's effectiveness.

C. Discriminator (D)

The D evaluates the truthfulness of the model's predictions of generated data by classifying samples with 0 *vs.* 1 [22]. It determines the genuineness of the input data, e.g., domain classification. The D categorizes the inherent attributes of the input data, providing valuable insights into the domain-specific attributes of the generator for the upcoming training stage.In Discriminator, the initial layer has a convolution layer [23], which processes input data by a sequence of filters to acquire features, and the leaky ReLU activation function layer [24], which introduces non-linearity into the network, and helps to understand complex patterns and relationships in the data. Convolution layers are utilized to execute 2D convolution on the input. During each loop iteration, a convolutional operation is appended to the layers list. The leaky ReLU activation function layer incorporates a gradient penalty (GP) for negative input values, which regulates the extent of function leakage. Two convolution layers are added to the D network for a conditional *GAN*. The first D discriminates between real *vs.* fake samples by feature maps from the final layer of the D network as input. Another D is used for speaker classification. It utilizes identical input feature maps but produces multiple channels equivalent to the number of speakers.

1) Gradient Penalty (GP) Mechanism: We incorporate the gradient penalty (GP) within the D block, as illustrated in Algorithm 1. This component blends source and target samples to generate new ones, assesses them with the D, computes the gradients of the D's output, and applies a gradient penalty [25]. This penalty encourages the D's gradients to maintain a norm using $\|\text{gradients}\|_2$, which is close to 1(one) using $\|\text{gradients}\|_2 - 1$, which helps to stabilize the training of GANs [17]. Contrastive learning helps models differentiate speech attributes by clustering similar features and separating dissimilar ones, which is key for tasks such as verification and synthesis. Gradient penalty enhances this process by stabilizing the learning by ensuring well-behaved gradients, improving the quality of learned representations, and preventing issues such as gradient explosion or vanishing. This combination results in more robust and reliable models.

Algorithm 1 Proposed Gradient Penalty (GP) Algorithm in GP-GAN using PyTorch

Require:
D: Discriminator model
A_{real}: Real samples, A_{fake}: Fake samples
B_{real}: Converted real samples
B_{fake}: Converted fake samples
L_{real}: Loss function of real VC
L_{fake}: Loss function of fake VC (inverse)
β: Batch size, λ: Gradient penalty coefficient
Ensure:
Gradient Penalty (GP)

procedure GRADIENT PENALTY DISCRIMINATOR($D, real_audio, fake_audio, device, \lambda$)
  $\beta \leftarrow size(real_audio, 0)$
  $total_loss \leftarrow 0$
  $\gamma \leftarrow random_uniform(0, 1, \beta)$
  for $i \leftarrow 1$ to β **do**
    $A_{\text{real}} \leftarrow ExtractFeatures_{real}[i]$
    $A_{\text{fake}} \leftarrow ExtractFeatures_{fake}[i]$
    $D_real \leftarrow D(A_{\text{real}})$
    $D_fake \leftarrow D(A_{\text{fake}})$
    if $D_real \geq threshold$ **then**
      $L_{real} \leftarrow Calc_{Loss}(D_real, True_{Labels})$
    else
      $L_{real} \leftarrow Calc_{Loss}(D_real, False_{Labels})$
    end if
    if $D_fake < threshold$ **then**
      $L_{fake} \leftarrow Calc_{Loss}(D_fake, False_{Labels})$
    else
      $L_{fake} \leftarrow Calc_{Loss}(D_fake, True_{Labels})$
    end if
    $interpolated \leftarrow \gamma[i] * A_{\text{real}} + (1 - \gamma[i]) * A_{\text{fake}}$
    $interpolated \leftarrow interpolated.to(device)$
    $interpolated_score \leftarrow D(interpolated)$
    $gradients \leftarrow \nabla_{interpolated} interpolated_score$
    $gradient_penalty \leftarrow ((\|gradients\|_2 - 1)^2) \cdot \lambda$
    $loss \leftarrow L_{real} + L_{fake} + gradient_penalty$
    $total_loss \leftarrow total_loss + loss$
  end for
  return $total_loss/\beta$
end procedure

D. Training Methodology

GPGAN-VC operates based on the G-D architecture. We utilized the PyTorch library to implement multi-parallel training on the generator-D architecture of the training dataset. Consider $a \in \mathbb{R}^{Q \times N}$ and $b \in \mathbb{R}^{Q \times M}$, where a represents an acoustic feature sequence of speech from attribute domain A, and b represents an acoustic feature sequence of speech from attribute domain B. Here, Q denotes the Feature Dimension (FD), while N and M are the lengths of the sequences. The goal of generator is to deceive the discriminator and minimize the loss function associated with its adversarial training. Our

training model's batch size is 64 for faster convergence, leading to a more stable training, and improved generalized performance. The Learning Rate of 0.001 for both networks implies that the model undergoes relatively smaller updates during training, contributing to more gradual adjustments in the model parameters. Using two *adam* optimizers, along with the parameters β_1 and β_2, merges adaptive learning rates and momentum, enhancing the training efficiency of neural networks.

1) Model Details: The GPGAN-VC works on the *gradient penalty* to regulate the dependency of the training data. Indirectly, the gradient penalty increases the performance of the model. It enforces the D's *Lipschitz continuity* [26], it suggesting that minor variations in the input domain yield limited alterations in the output domain. This feature contributes to training stability and facilitates enhanced model generalization, particularly beneficial in adversarial training scenarios. The generator's goal is to transform the A to B.

2) Experimental Setup: All experiments were conducted using Python 3.8, PyTorch 1.10, and Librosa on Ubuntu 22.04 LTS. For subjective evaluations, we utilized the TestVox webapp.

IV. Experimental Results

The efficacy of the presented GPGAN-VC model is evaluated using objective *vs.* subjective metrics. We compared the outcomes of the GPGAN-VC model with those of the existing StarGAN-VC [12] and StarGAN-VC2 [27] for objective evaluation. The model is evaluated using human-level speech quality and the speaker parallel tests in the context of VC literature. We took the CSTR VCTK dataset as mentioned in Section III-A for training the model using the proposed GPGAN-VC. The result of Table I shows that the speech quality from the generated samples is rated using the Mean Opinion Score (MOS) [28], Mel Cepstral Distortion (MCD), Mel Spectral Distance (MSD), and Perceptual Evaluation of the Speech Quality (PESQ) score, Signal-to-Noise Ratio, Fréchet Audio Distance (FAD) helps to consider the generated VC quality, which shows that GPGAN-VC gave better results than others. The amplitude *vs.* time plot in Fig. 2 illustrates comparisons between generated samples of the StarGAN-VC and GPGAN-VC, where Fig.2(b,c) and 2(d,e) generated samples are mentioned, respectively. It can be observed that StarGAN-VC 2(b) has more fluctuations over time, and at approximately 2.5 seconds, it shows the unwanted noise in VC, wherever GPGAN-VC shows the smoothly formed frequency for all the regions of the plot in each time, which helps to determine the naturalness of the GPGAN-VC. Fig. 3(a) and Fig. 3(b) shows the Mel spectrogram of the original speech sample, which was taken from the VCTK dataset, and the generated VC sample, respectively. At 1 sec, we can see that in fig. 3(b) have a higher frequency than Fig.3(a). which is shown in the rectangular boxes in both sample Mel spectrograms. Also, we can find the speech-linguistic similarity via the given Mel scales. e.g., the similarity between the plots in Fig.3 (a) and Fig.3 (b) is that both spectrograms are similar structures of the frequency regions over each other (between timeframes of 1.4-1.8 sec. the pattern of the generated female speech sample's spectrogram is near-to the speech produced by the vocal folds of a male speaker).

A. Performance Metrics

We evaluated our architecture by subjective and objective measures. Subjective evaluations involve human listener rating the quality and naturalness of the VC, and objective measures include the quantitive analysis of the original and generated signal as shown in the Fig.4 and Table I. This improvement in performance is clearly visible when comparing GPGAN-VC with the existing VC architectures, such as StarGAN-VC [12], StarGAN-VC2 [27], and GPGAN-VC [18]. The GPGAN-VC model performs particularly well in capturing the target speaker's prosodic features compared to the other models. This dual approach provided a comprehensive understanding of our model's performance, demonstrating its distinction over existing models.

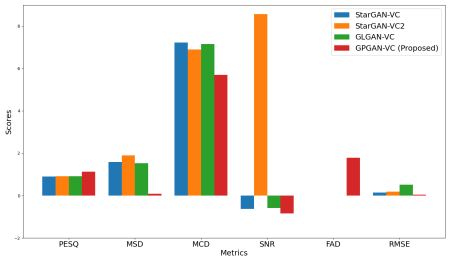


Fig. 4. Performance comparison for non-parallel VC.

1) Subjective Measures:

1) ***Mean Opinion Score (MOS):*** It is a subjective measure used to assess the quality of a product in the telecommunications domain [28], i.e.,

$$\text{MOS} = \frac{\sum_{n=1}^{N} R_n}{N}, \tag{3}$$

where R_n,N corresponds to individual and total ratings. Kaneko *et al.* [27]: StarGAN-VC2 received a MOS score of **3.1**. In comparison, GPGAN-VC, which was tested with a value of **3.7±0.95**, demonstrated the best speech quality among the VC models. Our MOS testing occurred in-house at the Speech Research Lab,@ " DA-IICT in a calm environment. We had 15 participants aged 18–24, with a female-to-male ratio of 1:2. Participants rated the samples on a scale from 1 to 5, where 1 indicated "Very poor or unacceptable quality", 2 indicated "Unsatisfactory and often unacceptable quality", 3 indicated "Acceptable but not good quality", 4 indicated "Good quality with minor issues". and 5 indicated "Enough high quality with no noticeable issues". None of the subjects had a listening disability, and each reviewed 20 shuffled samples. The majority of the listeners are non-native English speakers.

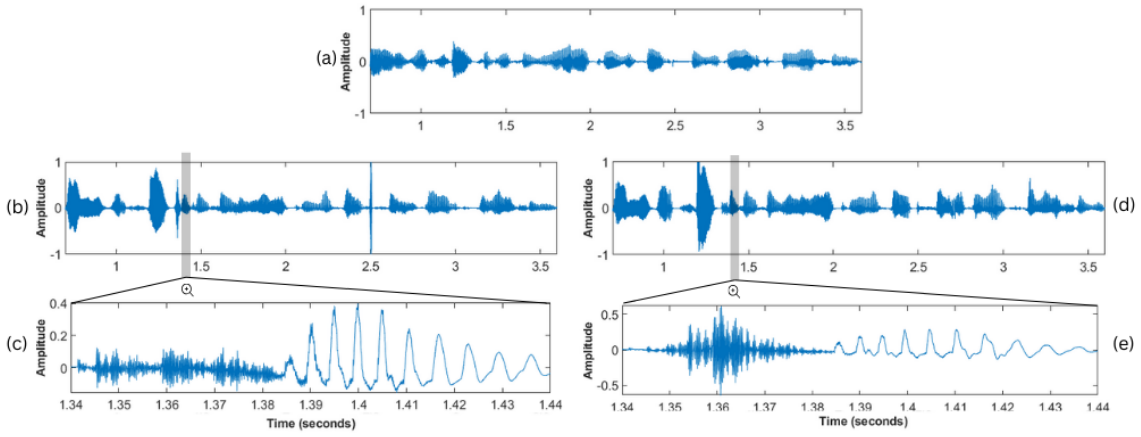


Fig. 2. Original speech signal and comparison of amp *vs*. time plot samples were acquired from the StarGAN-VC [12] and GPGAN-VC: (a) original speech samples, (b) StarGAN-VC generated speech waveform, and (c) magnified windows from the StarGAN-VC generated speech sample Fig.2(b) [12] (d) GPGAN-VC generated speech waveform, and (e) magnified windows from the GPGAN-VC generated speech sample Fig. 2(d).

TABLE I
COMPARISON OF THE PROPOSED ARCHITECTURE WITH OTHER NON-PARALLEL VC ARCHITECTURES.

Network	PESQ (↑)	MSD [dB] (↓)	MCD [dB] (↓)	SNR (↓)	FAD (↓)	RMSE(↓)	MOS(↑)
StarGAN-VC [12]	0.89	1.58	7.23	-0.63	0.75	0.14	3.50
StarGAN-VC2 [27]	0.43	1.89	6.90	8.57	0.68	0.18	3.10
GLGAN-VC [18]	0.91	1.52	7.15	-0.59	-	0.51	-
GPGAN-VC (Proposed)	**1.13 (±0.07)**	**0.08 (±0.02)**	**5.70 (±1.33)**	**-0.84 (±0.25)**	**1.78 (±0.46)**	**0.04 (±0.01)**	**3.71 (±0.95)**

A dash '-' indicates 'Not Applicable,' meaning the author either did not mention the result or the output sample is not available online.

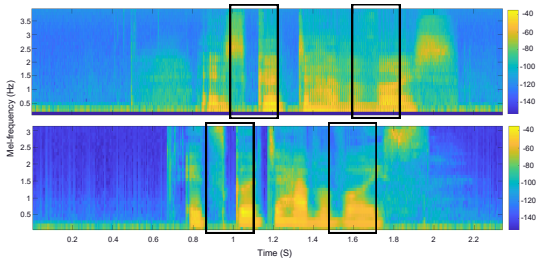


Fig. 3. Mel spectrogram comparison between original and voice converted samples: (a) original speech signal (taken from VCTK dataset [19], sample *p259_103.wav*, male speaker), and (b) converted sample (female speaker).

2) Objective Measures:

1) ***Perceptual Evaluation of Speech Quality (PESQ):*** The PESQ score measures speech quality. Higher PESQ ratings indicate high speech quality. In particular,

$$\text{PESQ} = 4.5 - 0.1d_{\text{sym}} - 0.0309d_{\text{asm}}, \tag{4}$$

where d_{sym} indicates symmetrically reduced speech quality, and d_{asym} indicates asymmetrically reduced speech quality. The PESQ values for the StarGAN-VC and the GLGAN-VC are 0.89 and 0.91, respectively however, the GPGAN-VC has a better quality VC sample than the existing models.

2) ***Mel Cepstral Distortion (MCD):*** It is used to quantify the variation among two sets of Mel cepstral coefficients, i.e.,

$$MCD = \frac{\sqrt{\sum_{t=1}^{N}\left(\sqrt{\sum_{i=1}^{D}(A(t,i)-B(t,i))^2}\right)^2}}{N} \tag{5}$$

where N and D convey the total number of coefficients and dimension count of the coefficients, respectively, and $A(t,i)$ and $B(t,i)$ represent the Mel Cepstral Coefficients at varying time points and dimensions, respectively.

3) ***Mel Spectra Distance (MSD):*** It calculates the likeness or disparity between two signals through modulation spectra. As given in Eq.(6), t and $\hat{t}$ represent target speech and converted speech, respectively. $s(y)^t$ and $s(y)^{\hat{t}}$ are converted logarithm modulation spectra, and N is the highest frame number. In particular,

$$\text{MSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(s(y)_i^t - s(y)_i^{\hat{t}}\right)^2}. \tag{6}$$

4) ***Signal-to-Noise Ratio (SNR):*** It is a metric to estimate the grade of the VC speech. SNR is the ratio of the signal power and the signal noise(dB), i.e.,

$$\text{SNR}_{\text{dB}} = 10 \cdot \log_{10}\left(\frac{\text{VC Speech Power}}{\text{VC Noise Power}}\right). \tag{7}$$

5) ***Fréchet Audio Distance (FAD)***: The FAD is used to evaluate the likeness between the two probability distributions. As per Eq. (8) in the VC context, it shows the *dissimilarity* between A and B. The FD involves the Gaussian distribution parameter to provide the dissimilarity between them.In particular,

$$F(A,B) = \inf_{a,b}\max_{t\in[0,1]}\{d(A(a(t)),B(b(t)))\}. \tag{8}$$

6) ***Root Mean Square Error (RMSE)***: RMSE helps to quantify the difference between the predicted and original speech samples by measuring the average magnitude

of the error. In particular,

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (x_{1,t} - x_{2,t})^2}{n}}, \tag{9}$$

where $x_{1,t}$ (ref eq. 9) represents the original speech signal at time t and $x_{2,t}$ represents the predicted or processed speech signal at the same time.

V. Summary and Conclusions

This study proposed GPGAN-VC, which integrates the gradient penalty into the architecture of GANs. D network of GP-GAN improved the learning of M:M VC, leading to the GPGAN-VC model showing favorable results in M:M speech mapping. The gradient penalty has helped to stabilize the training performance by improving both networks' trade-offs and avoiding the mode-collapse issue, which helps achieve equilibrium. The evaluation of the model is done by both subjective and objective assessment, which helps to provide a comprehensive insight into the speech quality and clarity, respectively. Although the proposed architecture has increased the optimization of VC and synthesis quality, maintaining the original prosody and emotional content from source to target speech remains challenging for researchers.Future work could focus on improving the model's preservation of prosody and emotional nuances by incorporating advanced emotional voice modeling techniques and exploring hybrid architectures that utilize various neural network strengths.

References

[1] Owais Nazir and Aruna Malik. Deep learning end to end speech synthesis: A review. In *2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, pages 66–71, 2021, Jalandhar, India.

[2] Quentin Langlois and Sébastien Jodogne. Practical study of deep learning models for speech synthesis. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*, pages 700–706, 2023, New York, USA.

[3] Eun-Kyoung Kim, Sangho Lee, and Yung-Hwan Oh. Hidden Markov Model based voice conversion using dynamic characteristics of speaker. In *European Conference On Speech Communication And Technology, EUROSPEECH, pp. 2519–2522, Rhodes, Greece, 1997*.

[4] Yannis Stylianou, Olivier Cappé, and Eric Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on speech and audio processing*, Vol. 6(2):pp.131–142, 1998.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS), Pro.* $27^{th}, 2014, pp.2672 - 2680, Montreal, Canada.$

[6] Takuhiro Kaneko and Hirokazu Kameoka. Cycle GAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. In 26th *European Signal Processing Conference (EUSIPCO), pp. 2100–2104, 2018, Rome, Italy*.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV), pp.2223–2232,*, 2017, Venice, Italy.

[8] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks, pp.4869-4873. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

[9] R. Takashima, T. Takiguchi, and Y. Ariki. Exemplar-based voice conversion using sparse representation in noisy environments. *IEICE Transactions on Information and Systems, pp.1946-1953*, E96-A(10), 2013.

[10] Z. Wu, T. Virtanen, E. S. Chng, and H. Li. Exemplar-based sparse representation with residual compensation for voice conversion. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 22(10):1506–1521, 2014.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013, {Last Accessed date:31 January, 2024}.

[12] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273, {Last Accessed date: 22nd February, 2024}, 2018, Athens, Greece.

[13] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multidomain image-to-image translation. *arXiv:1711.09020 [cs.CV]*, Nov 2017, {Last Accessed Date : 22 February, 2024}.

[14] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang. Voice conversion from non-parallel corpora using variational auto-encoder. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.1–6*, 2016, Jeju, Korea.

[15] D. P. Kingma, D. J. Rezendey, S. Mohamedy, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS), pp. 3581-3589*, 2014, Montreal, Canada.

[16] Gaku Kotani, Daisuke Saito, and Nobuaki Minematsu. Voice conversion based on deep neural networks for time-variant linear transformations. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:2981–2992, Sep 2022.

[17] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks, pp. 266-273. In *IEEE Spoken Language Technology Workshop (SLT), Athens, Greece*, 2018.

[18] Sandipan Dhar, Nanda Dulal Jana, and Swagatam Das. Glgan-vc: A guided loss-based generative adversarial network for many-to-many voice conversion, pp. 1-14. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[19] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019, {Last Accessed date: 22nd February, 2024}.

[20] Steven Weinberger. Speech accent archive. george mason university. *Online:¡ http://accent. gmu. edu*, 2015 {Last Accessed date: 22nd February, 2024}.

[21] Chao Zhang and Philip C Woodland. Parameterised sigmoid and relu hidden activation functions for dnn acoustic modelling, pp.3224–3228. In *INTERSPEECH*, 2015, Dresden, Germany.

[22] Yi Jiang, DeLiang Wang, RunSheng Liu, and ZhenMing Feng. Binaural classification for reverberant speech segregation using deep neural networks, vol. 22, pp.2112-2121. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014 {Last Accessed date: 22nd February, 2024}.

[23] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino. Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks. In *INTERSPEECH, pp. 1283-1287*, 2017, Stockholm, Sweden.

[24] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 503:92–108 , 2022.

[25] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems (NeurIPS)*, 30, 2017, Long Beach, United States.

[26] Khoat Than and Nghia Vu. Generalization of gans and overparameterized models under lipschitz continuity, 2022 {Last Accessed date: 22nd February, 2024}.

[27] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. StarGAN-vc2: Rethinking conditional methods for StarGAN-based voice conversion. *arXiv preprint arXiv:1907.12279*, 2019 {Last Accessed date: 22nd February, 2024}.

[28] Robert C Streijl, Stefan Winkler, and David S Hands. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems, Vol. 22, pp.213–227*, 2016.

Context-FFT: A Context Feed Forward Transformer Network for EEG-based Speech Envelope Decoding

Ximin Chen*‡, Yuting Ding*, Nan Yan†, Changsheng Chen‡, Fei Chen*
* Southern University of Science and Technology, Shenzhen, China
E-mail: 2020281106@email.szu.edu.cn, 12332166@mail.sustech.edu.cn, fchen@sustech.edu.cn
† Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
E-mail: nan.yan@siat.ac.cn
‡ Shenzhen University, Shenzhen, China
E-mail: cschen@szu.edu.cn

***Abstract*—Decoding speech envelope from electroencephalography (EEG) signals has been demonstrated to be useful for assessing speech intelligibility and boosting potential applications in neuroscience research as well as clinical diagnosis, which is also the focus of the ICASSP Auditory EEG 2023 Challenge. In order to further improve speech envelope decoding performance, this study proposes an end-to-end architecture based on multi-head attention mechanism called Context-FFT. Besides the transformer architecture, we also utilize a context layer to extract information and refine outputs based on given inputs. Notably, we decompose raw speech envelopes into envelopes in 12 frequency bands to model the relationship between 64-channel EEG signals and speech envelopes more precisely. Experiment results show that the proposed model achieves an average Pearson correlation value of 0.2148±0.1004 on held-out stories, outperforming the linear baseline by 51.65% and the VLAAI baseline by 23.17%, and 0.0701±0.0428 on held-out subjects. In terms of the final metric defined by the challenge, we obtain a final score of 0.1639, outperforming all submitted models of the ICASSP Auditory EEG 2023 Challenge. In the end, we explore the contributions of different brain regions to the process of continuous speech.**

I. Introduction

Electroencephalography (EEG) is a non-invasive neuroimaging technology that captures potential differences generated by electrical activities of the brain. Neurons exchange information between each other through synaptic connections in the form of electrical currents. When evoked, postsynaptic potentials generated by synchronized activities of a large number of neurons summate in the cortex and extend to the surface of the scalp, where they are recorded as EEG signals using a certain number of electrodes [1]. Due to its non-invasive and safe characteristics, EEG is widely used in fundamental research as well as clinical diagnosis [2]. EEG also has potential applications in smart hearing aids as researchers have found out that auditory attention can be decoded from EEG, which can improve outcomes of EEG-based smart hearing aids by determining the focus of a user's attention especially in noisy conditions [3].

To help understand how the brain processes continuous speech, recent studies have focused on neural tracking of speech features in EEG [4–7], which often refers to the time-locking effects of EEG to speech resources in a single-speaker scenario. One common paradigm to quantify the neural tracking of EEG is to reconstruct speech envelopes from EEG and calculate the similarity between the reconstructed envelope and the original envelope as the evaluation metric, which is also the focus of the ICASSP Auditory EEG 2023 Challenge [8] that calls for building the best model to relate speech to EEG. Linear models are commonly applied to reconstruct speech envelopes, but their reconstruction scores are low, ranging from 0.1 to 0.2 for subject-specific linear decoders [6, 9–12]. The limitation of linear models lies in assuming a linear relationship between speech and highly non-linear EEG [1].

Inspired by the success of deep learning architectures [13] in processing complex data such as text and audio signals [2], an increasing number of studies have been using deep learning models to relate speech to EEG. Accou et al. have proposed VLAAI, short for the Very Large Augmented Auditory Inference network, which stacks multiple convolutional blocks to enhance the model's non-linearity [6]. VLAAI has set the state-of-art performance by yielding an increase over the linear model by 52% and been chosen as the baseline of the challenge [8]. Notably, VLAAI utilizes an output context layer to take the output context into account, which contributes a 10% relative increase to the model performance, indicating that the context layer is capable to extract useful information from previous outputs. Transformer based on the multi-head attention mechanism is also widely used due to its significant success in fields such as natural language processing [14]. For instance, Yang et al. have proposed FastSpeech based on feed-forward Transformer (FFT) architectures to generate speech from text fast and robustly [15]. Drawing on its strong capability for fitting speech features, Piao et al. have proposed the HappyQuokka model consisting of FFT blocks and won the first place in the challenge [16], which indicates that incorporating the multi-head attention mechanism can effectively capture the dynamic variation characteristics of EEG signals. This motivates the present work to utilize the FFT architecture as well as the context layer to learn and fuse useful feature representations.

Besides model architectures, recent studies have also in-

corporated inherent characteristics of the auditory system to help model the complex relationship between EEG signals and speech stimuli. Thornton et al. have combined two decoders which exploit different EEG responses to speech: slow neural tracking of the speech envelope and high-frequency speech-related frequency following responses, achieving a significant improvement over the linear baseline [17]. Wu et al. have decomposed speech signals into envelopes in 12 frequency bands for more intelligible speech reconstruction [18]. This motivates the present work to utilize multi-band envelope information to guide the model towards more accurate envelope decoding.

To sum up, the aim of this work was to improve the decoding accuracy of reconstructing the speech envelope from the 64-channel EEG signal. Specifically, we propose a context feed forward Transformer network called Context-FFT, which combines multi-head attention mechanism and the context layer. Moreover, the raw speech is decomposed into envelopes of 12 frequency bands for model optimization. For model evaluation, we employ the SparrKULee dataset [19] as designated by the ICASSP Auditory 2023 Challenge [8].

II. METHODS

The proposed architecture, Context-FFT, is modified based on the system proposed by [16] as shown in Fig. 1. The input EEG signals are first fed into a convolutional layer for input embedding, then are passed through the modified FFT blocks to model the relationship between given EEG signals and speech features using the multi-head attention mechanism. The attention-weighted features are further refined through a context layer. Finally, the outputs of stacked blocks are fused by a linear layer to summarize extracted information into a single speech envelope. Key components and modifications are illustrated as follows.

A. Multi-head Attention Module

The multi-head attention module is based on Transformer's self attention mechanism [14], which can attend to specific tokens according to their attention scores. The normalization layer is placed inside the residual blocks instead of putting it between the residual block to help the model converge faster [20].

B. Context Layer

WaveNet [21] used a context module to expand the receptive field of the model. Similarly, VLAAI [6] employed a context layer to refine outputs by taking existing outputs into account. Inspired by these works, a context layer is introduced in the pre-LN FFT block, referred as the modified FFT block. Inside the context layer, attention-weighted features are passed through a convolutional layer with a kernel size of 9 and padding of 4, which is then followed by a nonlinear transformation using the ReLU activation function and a 1D convolutional layer. Finally, resulting features are normalized through layer normalization and adjusted using residual connections to obtain the refined features.

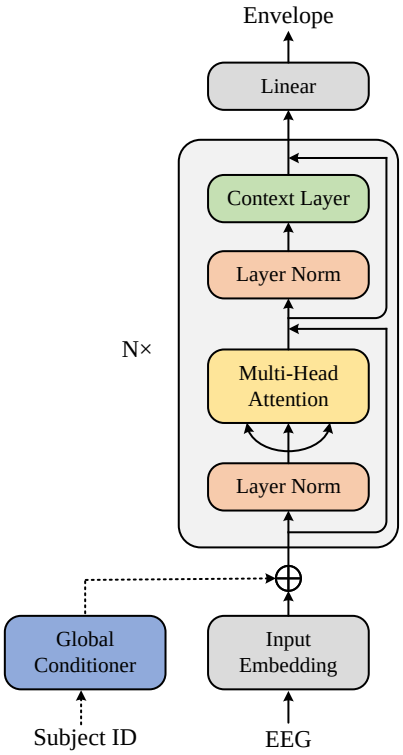


Fig. 1. Architecture of the proposed Context Feed Forward Transformer network, Context-FFT. Dash lines through the global conditioner block indicate that the block is only used in held-out stories test.

C. Auxiliary Global Conditioner

According to [16], the auxiliary global conditioner is only for within-subject generation, which provides the model with additional information of seen subjects. Specifically, it encodes a subject's identity into an embedding matching the dimension of the input embedding. This embedding is then added to the input embedding and fed into the modified FFT blocks.

III. EXPERIMENTS

For comparative studies, we perform experiments in SparrKULee dataset [19]. Details about the dataset and preprocessing procedure are provided as follows.

A. Dataset

The SparrKULee dataset [19] contains 85 participants with normal hearing. EEG data were recorded from 64 channels using a BioSemi ActiveTwo system at a sampling rate of 8192 Hz. All the stimuli are single-speaker stories spoken in Flemish by a native Flemish speaker, stored at a sampling rate of 48 kHz. Each subject listened to between 8 and 10 trials, each of approximately 15 minutes in length, totaling 168 hours of EEG data and 157 hours of stimuli.

For all our experiments, we use the preprocessed dataset provided by the challenge [8], which has already been normalized and split into train, validation and test set in a 1:1:1 ratio. Specifically, the training set includes EEG data from 71

subjects, numbered from sub-01 to sub-71. The test set consists of held-out stories (test set 1) and held-out subjects (test set 2). Test Set 1 includes EEG data from 71 subjects seen in training, while the corresponding stimuli are never used in the training set, thus referred as held-out stories, amounting to a total of 944 minutes. Test Set 2 includes EEG data from the remaining 14 subjects, numbered from sub-72 to sub-85 who are not in the training set, thus referred as held-out subjects, amounting to a total of 1260 minutes.

B. Data Preprocessing

The preprocessing procedure follows the setting of [8, 19]. For EEG data, firstly raw EEG recordings are downsampled from 8192 Hz to 1024 Hz and a multichannel Wiener filter is applied to remove eye blink artifacts. Then, the EEG signal is re-referenced to a common average. Finally, the EEG signal is downsampled to 64 Hz. As for stimuli, a sixth-order Butterworth filter is applied to band-pass filter raw speech signals to obtain 12 frequency band signals with center frequencies between 80 and 3000 Hz following the setting in [9, 18]. Next, each frequency band signal is filtered using a GammaTone filter bank with 28 subbands at equivalent bandwidth intervals with center frequencies ranging from 50 Hz to 5 kHz. Then, the absolute value of each sample in the filter is raised to the power of 0.6 and averaged to obtain a single speech envelope. Finally, the obtained envelope is downsampled to 64 Hz. During the training phase, we use the 12 subband estimated envelopes and the original signal's 12 subband envelopes as inputs of the loss function for optimization.

To improve computational efficiency, EEG signals and speech features are randomly cropped into segments of 5 seconds in the training phase. During testing, EEG signals are split into 5-second segments and are fed into the model, whose outputs are then concatenated together to make the entire envelope.

C. Channel Groups of Cortical Regions

To investigate the impact of different cortical regions on processing continuous speech, the cortical regions are divided into eight channel groups (CG) following the setting from [22], as shown in Table I. Broca's and Wernicke's areas are considered to be related to auditory production [23], and speech stimuli may also activate the auditory cortex. Moreover, the motor cortex, prefrontal cortex, and sensory cortex regions incorporate together in language production and processing [22].

D. Network Configuration

This work uses the TensorFlow framework to implement the linear decoder and VLAAI models for comparison, and the proposed model is implemented using the PyTorch framework. The number of modified FFT blocks is set to 8 and we use 2 heads for multi-head attention. The negative Pearson correlation is used as the loss function as required by the challenge [8]. We use the Adam optimizer with an initial learning rate of 0.0005 and a StepLR scheduler with a learning rate decay factor of 0.9.

IV. RESULTS AND DISCUSSION

In this section, we demonstrate the performance of Context-FFT on the SparrKULee dataset [19]. It is then compared with linear decoder [9] and VLAAI [6], which are both baselines provided by the challenge [8], as well as the HappyQuokka System [16], which placed the first in the challenge. Subsequently, an ablation study is conducted to evaluate effects of key components on model performance. In the end, we present the results of decoding performance of different brain regions and discuss its potential applications in smart hearing aids.

A. Model Performance

Table II shows the results of different models on the test set. Pearson correlation (Pearson *r*) is employed as evaluation metric between estimated envelopes and actual envelopes. Reconstruction scores of test set 1 (S_1) and test set 2 (S_2) are averaged across stimuli and subjects. Final score is a weighted sum of the average Pearson *r* of both test subsets defined by the challenge [8], computed as

$$\text{Final Score} = \frac{2}{3}S_1 + \frac{1}{3}S_2. \tag{1}$$

From Table II, we observe that our proposed model Context-FFT outperforms both baseline models and the state-of-art HappyQuokka model. Specifically, Context-FFT achieves an average Pearson *r* of 0.2180 ± 0.1004 for held-out stories and 0.0701 ± 0.0428 for held-out subjects, achieving the highest final score of 0.1639. Notably, our model has significantly improved in held-out stories by 51.65% and 23.17% compared with the linear decoder and VLAAI respectively. As for held-out subjects, the decoding performance slightly declines compared with other three models, possibly because the model's ability to learn unique characteristics of seen subjects has improved due to feature enhancement and context layer, however, feature patterns learned from seen subjects can't fit well to unseen subjects, thus leading to a decrease in decoding performance.

B. Ablation Study

To gain insight into what parts of the model are responsible for improvements on decoding performance, we have conducted ablation studies on the basis of the HappyQuokka model [16], which is referred as the benchmark model. Since our proposed model Context-FFT has outperformed baseline models by a large margin on held-out stories, while results of ablation studies on held-out subjects have minimal changes and are all below 0.1, we only present ablation studies on held-out stories. Specifically, key modules are added to the original stacked FFT block in the following steps:

1. Benchmark+FB: decomposing speech signals into 12 subbands and extracting their envelopes respectively on the basis of the benchmark model. FB is short for Filter Band.
2. Context-FFT: adding a context layer to the FFT blocks of the model in step 1.

Figure 2 shows the decoding performance of baseline models and each model variant in held-out stories. First of all,

TABLE I
CHANNEL GROUPS FOR CORTICAL REGIONS

Channel Groups	Cortical Regions	Channels
CG1	Broca's and Wernicke's areas	AF3, F3, F5, FC3, FC5, T7, C5, TP7, CP5, P5
CG2	Auditory cortex	FT7, FT8, T7, T8, TP7, TP8
CG3	Motor cortex	FZ, F1, F2, F3, F4, FC1, FC2, FC3, FC4, CZ
CG4	Prefrontal cortex	FPZ, FP1, FP2, AF3, AF4, F5, F6, F7, F8
CG5	Sensory cortex	CZ, C1, C2, C3, C4, CPZ, CP1, CP2, CP3, CP4
CG6	Left brain	FP1, AF3, F1, F3, F5, F7, FC1, FC3, FC5, FT7, C1, C3, C5, T7, CP1, CP3, CP5, TP7, P1, P3, P5, P7, PO3, PO5, PO7, CB1, O1
CG7	Right brain	FP2, AF4, F2, F4, F6, F8, FC2, FC4, FC6, FT8, C2, C4, C6, T8, CP2, CP4, CP6, TP8, P2, P4, P6, P8, PO4, PO6, PO8, CB2, O2
CG8	Whole brain	All 64 channels

TABLE II
PERFORMANCE OF MODELS ON TEST SETS OF SPARRKULEE DATASET

Model	Test Set 1	Test Set 2	Final Score
Linear Decoder [9]	0.1054±0.0538	0.0960±0.0387	0.1023
VLAAI [6]	0.1675±0.0732	**0.1139±0.0410**	0.1496
HappyQuokka [16]	0.1895±0.0869	0.0976±0.0444	0.1589
Context-FFT	**0.2180±0.1004**	0.0701±0.0428	**0.1639**

the median Pearson correlation values of all model variants have gradually increased. The decoding performance of Benchmark+FB has improved by 9.3% compared to the benchmark model, suggesting that the model can better find the correspondence between EEG signals and speech stimuli from the envelopes of different sub-bands than from the entire envelope. It also demonstrates that combining inherent characteristics of auditory-EEG processing can fit the relationship between EEG signals and speech stimuli more precisely, further enhancing the model's feature fusion capability. Furthermore, the performance of Context-FFT has also significantly improved by 13.1% compared with the benchmark model, indicating that expanding the model's receptive field can enhance its ability to extract useful context information.

C. Contributions of Cortical Regions

Figure 3 shows the comparison of decoding performance of different cortical regions in decoding speech envelopes from EEG signals.

First of all, we can observe that the Pearson r of CG8 is the highest among all groups, indicating the entire brain has involved in processing speech stimuli. Among the remaining groups, CG6 has the highest reconstruction score, followed by CG1 and CG2. Specifically, CG1 corresponds to Broca's

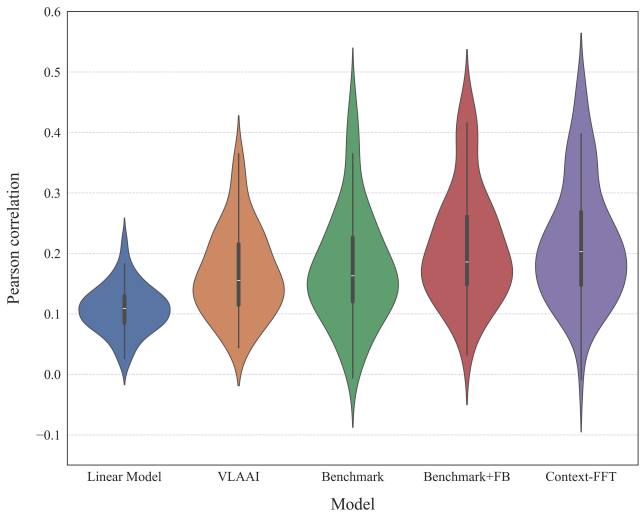


Fig. 2. Results of the ablation studies. Each point in the violin plot represents the average Pearson correlation for a subject averaged across stimuli.

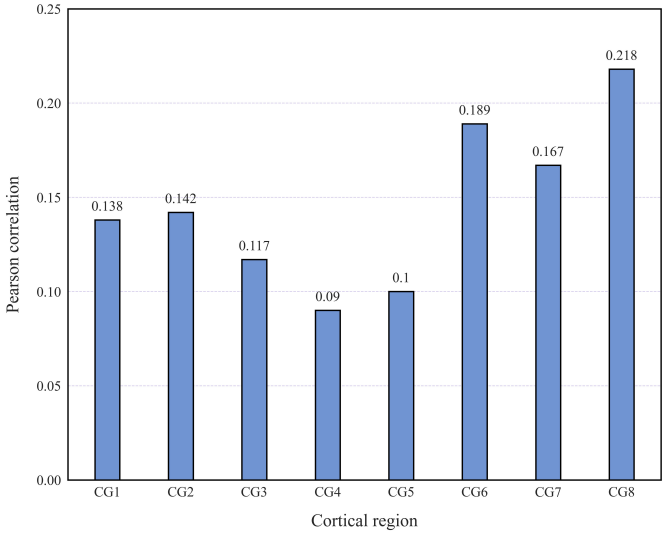


Fig. 3. Decoding performance of different cortical regions.

and Wernicke's areas which are located in the left brain region (CG6). Reconstruction scores of CG1 and CG6 are relatively close, suggesting that Broca's and Wernicke's areas play crucial roles in auditory processing of the left brain, which is consistent with conclusions obtained in [22, 23].

Moreover, we can observe from results of CG6 and CG7 that the involvement of the left brain is higher than that of the right brain in processing speech stimuli. Results of CG3, CG4, and CG5 are the lowest among all groups, indicating that when the brain processes speech stimuli, the corresponding motor cortex, prefrontal cortex and sensory cortex are also activated, but their contributions are relatively small compared to the auditory cortex and Broca's and Wernicke's areas.

Although CG2 covers only 6 channels, its Pearson r is relatively close to scores of CG1 and CG6, which cover 10 channels and 27 channels separately. Therefore, the above results have demonstrated that using fewer channels can achieve similar performance obtained by using full-channel EEG signals, which is essential for EEG-based smart hearing aids that require low computational costs and short computation times, allowing users to receive real-time feedback from the devices.

V. CONCLUSIONS

This work proposes a new model called Context-FFT for decoding speech envelopes from EEG signals and demonstrates that the performance of this model surpasses the SOTA model of the ICASSP Auditory 2023 Challenge. On average, Context-FFT achieves a Pearson r of 0.2180 on the held-out-stories and 0.0701 on the held-out subjects. Furthermore, through ablation experiments, it is determined that utilizing the context layer and incorporating the inherent characteristics of EEG signals play a crucial role in improving the model performance. Although our model achieves better results on seen subjects, the generalization performance on unseen subjects is not significantly improved, indicating that improving the generalization ability of the model remains challenging. Additionally, the decoding performance of the Broca's and Wernicke's areas containing 10 channels and the auditory cortex containing 6 channels are close to the performance obtained using 64-channel EEG signals, providing a theoretical basis for achieving a balance between computation speed and decoding performance in portable EEG-based cochlear implants and other hearing devices.

ACKNOWLEDGMENTS

This work was supported by Guangdong Basic and Applied Basic Research Foundation (Grant No. 2022B1515120056), and the Basic Research Foundation of Shenzhen (Grant No. JCYJ20220818101217037). Part of this study was the basis for the Bachelor's thesis of the first author (X.M.C.).

REFERENCES

[1] D. P. Subha, P. K. Joseph, R. Acharya U, and C. M. Lim, "EEG signal analysis: A survey," *Journal of Medical Systems*, vol. 34, no. 2, pp. 195–212, 2010.

[2] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, p. 051001, 2019.

[3] M. Thornton, D. Mandic, and T. Reichenbach, "Robust decoding of the speech envelope from EEG recordings through deep neural networks," *Journal of Neural Engineering*, vol. 19, no. 4, p. 046007, 2022.

[4] S. J. Aiken and T. W. Picton, "Human cortical responses to the speech envelope," *Ear and Hearing*, vol. 29, no. 2, p. 139, 2008.

[5] D. A. Abrams, T. Nicol, S. Zecker, and N. Kraus, "Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech," *Journal of Neuroscience*, vol. 28, no. 15, pp. 3958–3965, 2008.

[6] B. Accou, J. Vanthornhout, H. V. hamme, and T. Francart, "Decoding of the speech envelope from EEG using the vlaai deep neural network," *Scientific Reports*, vol. 13, no. 1, p. 812, 2023.

[7] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, H. V. hamme, and T. Francart, "Relating EEG to continuous speech using deep neural networks: A review," *Journal of Neural Engineering*, vol. 20, no. 4, p. 041003, 2023.

[8] M. J. Monesi, L. Bollens, B. Accou, J. Vanthornhout, H. Van Hamme, and T. Francart, "Auditory EEG decoding challenge for icassp 2023," *IEEE Open Journal of Signal Processing*, 2024.

[9] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European Journal of Neuroscience*, vol. 31, no. 1, pp. 189–193, 2010.

[10] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech intelligibility predicted from neural entrainment of the speech envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, pp. 181–191, 2018.

[11] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology: CB*, vol. 25, no. 19, pp. 2457–2465, 2015.

[12] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in Human Neuroscience*, vol. 10, 2016.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] Z. Piao, M. Kim, H. Yoon, and H.-G. Kang, "Happyquokka system for icassp 2023 auditory EEG challenge," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–2, IEEE, 2023.

[17] M. Thornton, D. Mandic, and T. Reichenbach, "Relating EEG recordings to speech using envelope tracking and the speech-FFR," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–2, IEEE, 2023.

[18] H. Wu and F. Chen, "A temporal envelope-based speech reconstruction approach with EEG signals during speech imagery," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 894–899, 2020.

[19] B. Accou, L. Bollens, M. Gillis, W. Verheijen, H. Van hamme, and T. Francart, "Sparrkulee: A speech-evoked auditory response repository of the ku leuven, containing EEG of 85 participants," *bioRxiv*, pp. 2023–07, 2023.

[20] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*, pp. 10524–10533, PMLR, 2020.

[21] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[22] M. Li, S. H. Pun, and F. Chen, "Impacts of cortical regions on EEG-based classification of lexical tones and vowels in spoken speech," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1–4, IEEE, 2023.

[23] M. N. I. Qureshi, B. Min, H.-j. Park, D. Cho, W. Choi, and B. Lee, "Multiclass classification of word imagination speech with hybrid connectivity features," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 10, pp. 2168–2177, 2018.

Effect of Dynamic Binaural Beats on Concentration Enhancement

Jun-Seok Lee[1, 2], Yun-Sung Lee[1, 2] and Han-Jeong Hwang[1, 2*]
[1] Dept. of Electronics and Information Engineering, Korea University, Sejong, Korea
[2] Interdisciplinary Graduate Program for Artificial Intelligence Smart Convergence Technology, Korea University, Sejong 30019, Republic of Korea.
E-mail: zck320@korea.ac.kr, ysung947@korea.ac.kr, *hwanghj@korea.ac.kr, Tel: +82-44-860-1762

***Abstract*— The objective of this study was to develop an auditory stimulation method that can improve concentration and to investigate its neurophysiological effects using electrocardiography (ECG). The auditory stimulus was developed by mixing the sound of white noise with dynamic binaural beats (DBB) that featured time-varying carrier frequencies. The experimental paradigm involved conducting an association memory (AM) task and a mental arithmetic (MA) task to assess cognitive performance. In the behavioral results, no significant differences were confirmed between the two experimental conditions (sound vs. no sound) for both cognitive tasks. In the AM task, no significant differences in HRV change between the baseline and task were observed between the two conditions. However, during the MA task, there was a significant decrease in HRV when participants were exposed to the auditory stimulation compared to when it was not.**

I. Introduction

Concentration or attention is a crucial factor that influences learning ability and can ultimately enhance learning efficiency. Recently, the use of electronic devices such as smartphones often results in loss of concentration, which can cause a decrease in learning ability. Therefore, various studies have proposed functional auditory stimuli such as white noise and binaural beats (BBs) to enhance concentration or attention.

BBs are functional auditory stimuli where tones with different carrier frequencies are independently presented to each ear. Studies focusing on concentration have primarily utilized BBs set at 40 Hz, with reported enhancements in concentration and cognitive function [1]. Traditional BBs are designed with a consistent frequency difference between the left and right ears. Recently, a new type of BB called dynamic BBs (DBB), featuring time-varying carrier frequencies, has been introduced [2], covering the entire range of delta frequency bands associated with sleep and its positive impact was demonstrated on enhancing sleep quality. However, further research is needed to determine the effects of DBB on concentration and cognitive functions.

Some studies suggest that the repetitive, unnatural sounds of BBs can make people uncomfortable and even cause dizziness. Therefore, researchers have addressed this issue by mixing BBs with other sounds, such as white noise and nature sounds, in a suitable ratio. They have confirmed the positive effects of this approach on enhancing cognitive ability and inducing sleep [3][4]. White noise contains all audible frequencies (20 to 20,000 Hz) without any specific auditory pattern. The presentation of white noise at an appropriate decibel (dB) level can enhance brain processing ability, thereby improving attention, concentration, and cognitive function [5].

In this study, we developed an auditory stimulus to enhance concentration by mixing DBB and white noise to cover a wider range of the gamma band, which is associated with cognitive functions centered around 40 Hz. Furthermore, we aimed to validate the effect of the developed auditory stimulus on concentration using electrocardiogram (ECG), which is associated with cognitive functions and attention [9][10].

II. Methods

A. Participants

Sixteen healthy subjects (9 males and 7 females, aged 22.1 ± 1.7 years) participated in this study. All participants had no neurological or psychiatric medical history that could have influenced the study's findings. Before the experiment, we provided the subjects with the information about the experimental procedures, and they signed a research consent form. The study protocol was approved by the Institutional Review Board (IRB) of Korea University [KUIRB-2023-0224-01].

B. Auditory stimuli

The auditory stimulus was developed by mixing the sound of white noise with DBB that featured time-varying carrier frequencies. The DBB was designed to oscillate at ± 3 Hz around 40 Hz to cover a wider range of the gamma band, based on previous studies showing that 40 Hz binaural beats are associated with cognitive functions and are known to enhance concentration [1][4][6]. As a result, the left auditory stimulus was modulated with 200 Hz, while the right one was modulated with 240 ± 3 Hz, covering a range of 37 – 43 Hz. White noise

contains all audible frequencies, but its high frequencies can be unpleasant to hear. Rain sounds have a similar effect on the brain because of their perceptual similarity to white noise [7] and are more comfortable to hear because of their relatively lower high-frequency components. Additionally, based on a previous study that showed the effectiveness of mixing binaural beats and ASMR sounds in a 1:2 ratio for inducing sleep [3], we developed the auditory stimulus by mixing DBB and white noise-based rain sounds in a 1:2 ratio. The generated auditory stimulus was presented at a 70 dB level through earphones, following previous studies indicating that appropriate noise levels can enhance cognitive functions [8].

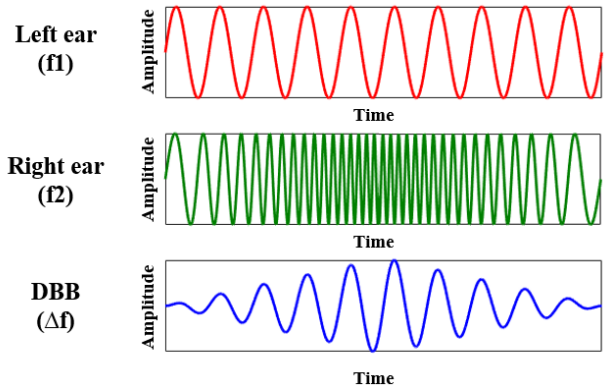


Fig. 1 Example of scheme of dynamic binaural beat (DBB)

C. *Experimental paradigm*

In this study, an association memory (AM) task and a mental arithmetic (MA) task were performed to evaluate concentration and cognitive function (Fig. 2). Both cognitive tasks were performed after a 1-minute resting state. The AM task involved 40 different pairs of pictures and word combinations, with 20 pairs where the meanings of words were associated with the given pictures and 20 pairs where the meanings were not associated. During the encoding phase, participants tried to memorize 40 pairs of pictures and words that were presented sequentially in a randomized order, with each pair shown for 2 seconds. To introduce a distraction between stimuli, a blank image was inserted after each pair presentation. After a 2-minute break, in the decoding phase, a randomly selected picture from the pairs memorized during the encoding phase was presented for 5 seconds, and participants were asked to provide the corresponding word. The auditory stimuli were presented only during the encoding phase, when memorization occurred. Behavioral performance was evaluated based on accuracy and reaction time (RT). The MA task involved repeatedly subtracting a one-digit number from a three-digit number for 10 seconds. The arithmetic expression was presented for 2 seconds, followed by a 10-second mental calculation period. After this, participants had 3 seconds to provide their final calculation result. It was repeated 30 times with different number combinations, while auditory stimuli were continuously presented during the execution of the task. Behavioral performance was evaluated based on accuracy and the count of calculations.

To compare the effects of auditory stimulation, two conditions were tested: one with auditory stimulation presented (Sound) and the other with silence (No Sound), each presented to subjects in a random order. Additionally, the AM and MA tasks were also presented randomly, which helped minimize any learning effects that could influence performance. Ultimately, all participants completed both the AM and MA tasks twice, once under each auditory condition (Sound and No Sound). During the experiment, ECG data were measured using the Lead-I configuration.

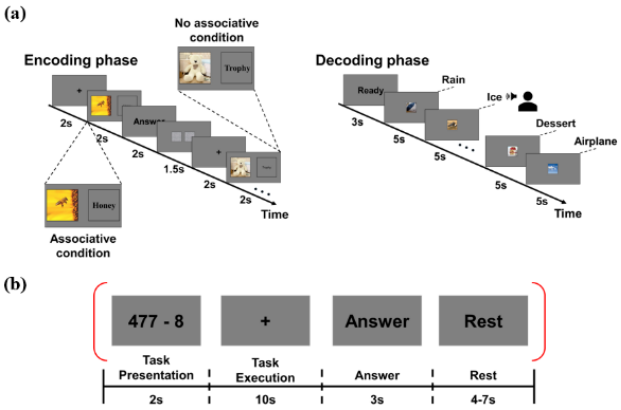


Fig. 2 Experimental paradigm (a) Association memory task, (b) Mental arithmetic task

D. *ECG data analysis*

In this study, ECG data were measured to investigate the neurophysiological effects of auditory stimulation on concentration and cognitive function. To remove artifacts from raw ECG data, we performed a band-pass filtering between 0.01 and 30 Hz. Subsequently, heart rate variability (HRV) was calculated during the resting state (baseline) and the cognitive tasks.

III. RESULTS

A. *Behavior results*

In the AM task, accuracy was about 2% lower, but the RT was approximately 150 ms faster in the Sound condition compared to the No Sound condition. In the MA task, accuracy was approximately 4% higher in the Sound condition compared to the No sound condition, and there was no difference in the count of calculations. No statistical significance was observed for any behavioral results (Wilcoxon signed-rank test, $p > 0.05$).

Table 1. Behavior results of both cognitive tasks

		Mean (std)	
		Sound	No sound
AM task	Acc (%)	60.2 (13.1)	62.0 (13.6)
	RT (ms)	2270.1 (482.6)	2415.1 (485.0)
MA task	Acc (%)	84.8 (11.9)	80.4 (11.3)
	Count (n)	4.6 (1.5)	4.6 (1.8)

B. HRV results

Fig. 3 shows the average HRV during the baseline and task for each Sound and No Sound condition, as well as the difference between the task and baseline HRV during the AM task. In both conditions, HRV significantly decreased during the task compared to the resting state (Wilcoxon signed-rank test, $p < 0.05$). However, no significant differences in HRV change between the baseline and task were observed between the two conditions.

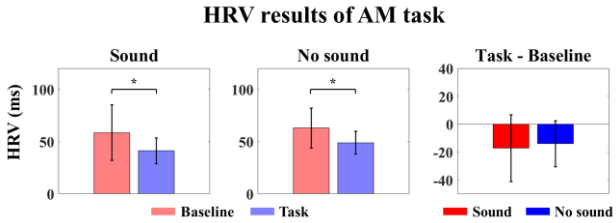


Fig. 3 HRV results of the AM task during baseline and task

Fig. 4 shows the HRV results of the MA task. Similar to the AM task, HRV significantly decreased during the task compared to baseline in both Sound and No Sound conditions. Furthermore, in the MA task, HRV decreased more significantly during the task in the Sound condition compared to the No Sound condition (Wilcoxon signed-rank test, $p < 0.05$).

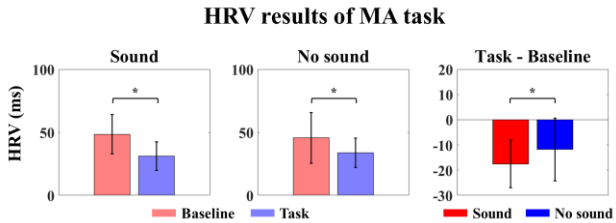


Fig. 4 HRV results of the MA task during baseline and task

IV. Discussion

In this study, we developed an auditory stimulus by mixing DBBs and rain sounds and investigated its effect on the performance of the memory (AM) and the mathematical calculation (MA). Previous studies have shown that a decrease in HRV is related to an increase in concentration or attention [9][10]. During the two cognitive tasks, HRV significantly decreased compared to the resting state, regardless of the auditory stimulation, indicating that participants were concentrated on the tasks. Specifically, during the MA task, HRV decreased more significantly when auditory stimulation was presented. This suggests that auditory stimulation based on the DBB may enhance concentration during MA tasks. However, during the AM task, the expected reduction in HRV associated with auditory stimulation could not be confirmed. This indicates that the auditory stimulation developed may not have had a significant effect on concentration during the AM task.

Despite the significant decrease in HRV, no significant differences were observed in behavioral results between the two conditions. A previous study has shown that the beneficial effect of background auditory stimulation is not found when performing easy arithmetic problems, and a ceiling effect is observed as the task difficulty increases [6]. Since this study also showed a high accuracy of over 80% in the MA task, further investigation is needed to verify the effect of the auditory stimulation at higher difficulty levels in the future. Moreover, 16 subjects are relatively small for ensuring statistical power, so it is necessary to conduct additional experiments with more subjects.

V. Conclusions

In this study, we investigated the effects of auditory stimulation mixing DBBs and white noise-based rain sounds on cognitive functions. The developed auditory stimulus induced a significant decrease in HRV during a MA task, indicating a potential for enhancing attention and concentration.

VI. Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2023- 00258971) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

[1] Ross, Bernhard, Lopez, Marc Danzell. “40-Hz Binaural beats enhance training to mitigate the attentional blink”, Scientific reports, Vol. 10, No. 1, p. 7002, 2020.

[2] LEE, Hwa-Ah-Ni, et al. “Effect of dynamic binaural beats on sleep quality: a proof-of-concept study with questionnaire and biosignals.”, Sleep, zsae097, 2024.

[3] LEE, Minji, et al. “Possible effect of binaural beat combined with autonomous sensory meridian response for inducing sleep.” Frontiers in human neuroscience, Vol. 13, p. 425, 2019.

[4] HOMMEL, Bernhard, et al. High-frequency binaural beats increase cognitive flexibility: evidence from dual-task crosstalk. Frontiers in Psychology, Vol. 7, p. 214422, 2016.

[5] Angwin, Anthony J., et al. “White noise facilitates new-word learning from context”, Brain and Language, Vol. 199: 104699, 2019.

[6] COLZATO, Lorenza S., et al. “More attentional focusing through binaural beats: Evidence from the global–local task.”, Psychological research, Vol. 81, pp. 271-277, 2017.

[7] PROVERBIO, Alice Mado, et al. “When listening to rain sounds boosts arithmetic ability.” PLoS One, Vol. 13, No. 2: e0192296, 2018.

[8] MEHTA, Ravi; ZHU, Rui; CHEEMA, Amar. “Is noise always bad? Exploring the effects of ambient noise on creative cognition.” Journal of Consumer Research, Vol. 39, No. 4, pp. 784-799, 2012.

[9] HANSEN, Anita Lill; JOHNSEN, Bjørn Helge; THAYER, Julian F. “Vagal influence on working memory and attention.”, International journal of psychophysiology, Vol. 48, No. 3, pp. 263-274, 2003.

[10] HOOVER, Adam, et al. “Real-time detection of workload changes using heart rate variability.”, Biomedical Signal Processing and Control, Vol. 7, No. 4, pp. 333-341, 2012.

EEG-based Evaluation of Enjoyment Emotion during Cognitive-Motor Task

Haruna Aoki[1], Sinan Zhang[2] and Yumie Ono[3]
[1] Graduate School of Science and Technology, Meiji University, Kanagawa, Japan
E-mail: ce241010@meiji.ac.jp
[2] Organization for the Strategic Coordination of Research and Intellectual Properties, Meiji University, Kanagawa, Japan
E-mail: zhang940211@gmail.com
[3] School of Science and Technology, Meiji University, Kanagawa, Japan
E-mail: yumie@meiji.ac.jp Tel: +81-44-934-7302

***Abstract*— Remote communication has become a widely used means of communication in modern society. However, the limited non-verbal information can make remote communication difficult, and there is a need to develop technologies that quantitatively evaluate and transfer users' emotional information for smoother communication. This study focuses on enjoyment, one of the positive emotions crucial for communication, and investigates EEG biomarkers for estimating enjoyment during cognitive-motor tasks. EEG was measured at six channels in 12 healthy young adults while they performed cognitive-motor tasks at five different difficulty levels. The task was presented as an immersive virtual reality game to simulate a virtual workspace with adjustable task difficulty. Participants answered a questionnaire to evaluate their "enjoyment" level for each gameplay. A stepwise regression analysis resulted in a linear model equation that could estimate enjoyment levels across all participants based on normalized EEG power values. This model was validated using leave-one-participant-out cross-validation, demonstrating its effectiveness in predicting enjoyment. Our findings suggest the importance of left parietal electrophysiological activity for estimating the positive emotion of enjoyment.**

I. INTRODUCTION

With the advancement of telecommunication and virtual reality (VR) technology, the opportunities for communication and collaboration in online environments have rapidly increased [1]. Compared to face-to-face communication, online meetings, remote classes, and avatar-mediated communication face challenges in efficient task assignment and smooth team communication [2] [3] due to the difficulty in obtaining non-verbal information [4]. Although research on face tracking, motion tracking, and emotion estimation from speech content has progressed [5], a method for estimating users' emotional states without relying on physical expressions or verbal information is needed to solve these problems. This is particularly relevant for online collaboration, lectures, and remote task instructions, where physical expressions or active conversations may not be present. The development of such methods could improve the ability to gauge emotional states in a wider range of contexts, potentially improving user experience and interaction in virtual and remote environments.

The effectiveness of quantitative emotion estimation using physiological information has been increasingly demonstrated in recent years. Previous studies [6] have shown that physiological information such as electroencephalography (EEG) and pulse waves can robustly estimate emotions with high accuracy, regardless of the individual, compared to facial expressions. Heart rate variability, another physiological measure of the autonomic nervous activity, is also gaining attention as an objective indicator of emotional responses and is expected to be a valuable emotional assessment metric [7]. In this study, we aimed to evaluate users' positive emotion of enjoyment during online tasks using EEG measurements, which are capable of monitoring the cognitive states of users with an excellent temporal resolution. We investigated multiple frequency components of EEG signals during the cognitive-motor task and their relationship with subjective enjoyment intensity. We also incorporated a stepwise regression analysis to determine the common biomarker signals that allow emotional evaluation of enjoyment during task engagement.

II. METHODS

Twelve healthy adults (5 males, 7 females, average age 20.8 ± 1.5 years) participated in the experiment. All experiments were approved by the institutional Ethics Committee and participants gave written informed consent before participation. Participants played five sessions of motor-cognitive tasks with different difficulty levels in a random order.

The task was an apple-catching game created by Unity engine as shown in Figure 1. An immersive VR game was selected because of its capability to adjust the task difficulty, simulate a virtual workspace, and evoke associated emotional responses. The goal of the game is to catch apples flying from the front center to either side of the player (red or blue area in Fig. 1) by moving the basket connected to the controller. The flying speed of the apple was constant throughout the game (4

m/s), and the game difficulty was adjusted by the five time intervals between the flying apples (2, 4, 6, 8, and 10 s). The shorter interval is assumed to have more enjoyment as a nature of the game. Each session lasted one minute, regardless of the difficulty levels.

EEG was measured at a sampling rate of 256 Hz under a 0.5–100 Hz band-pass filter using a bioamplifier (g.BCIsys32SUSB, g.tec medical engineering GmbH, Austria) during the VR games at six channels on the frontal (F3, F4), central (C3, C4), and parietal (P3, P4) regions. In the preprocessing procedure, I visually confirmed the absence of intense and large artifacts and removed segments exceeding ±100 μV. Participants first wore the electrode cap and a head-mounted display (VALVE INDEX, Valve Corporation, USA) over the cap. After each game play session, participants completed a questionnaire to provide subjective ratings of their enjoyment. The questionnaire used a 5-point Likert scale, where '1' indicated that participants found the game the most enjoyable and would like to play it again, '5' indicated that they found the game least enjoyable, boring, or would not play it again, and '3' indicated neutrality.

We investigated the relationship between the game difficulty (time intervals between the apples) and the questionnaire scores using a Spearman's rank correlation test. The EEG data were subjected to a fast Fourier transform to calculate the power value of each channel in four frequency bands (δ, θ, α, and β bands). The channel-wise power spectrum values were further normalized to those of a session in which the participant rated their enjoyment as neutral (scale of 3) and used as indices of cortical activity. If there were two or more sessions in which they rated their enjoyment as neutral, the averaged power value was used for normalization. We used the normalized power values to examine their correlation with subjective ratings of enjoyment.

We then conducted a stepwise regression analysis to create a linear model that best approximates the score of enjoyment out of 24 different normalized power values (6 channels x 4 frequency bands) as factors of the EEG index. We compared actual questionnaire scores with predicted score values obtained by substituting EEG data into the regression equation. Furthermore, leave-one-participant-out cross-validation (LOOCV) was conducted to validate the prediction model. In the LOOCV process, data from one participant were excluded as test data. Subsequently, a linear regression model was then built using only the remaining data of other participants. Finally, the test data were substituted into the model to estimate the enjoyment scores. This process was repeated for all participants. The error between the enjoyment scores estimated from the prediction model and the actual scores was evaluated. A pipeline summarizing the experimental flow is shown in Figure 2.

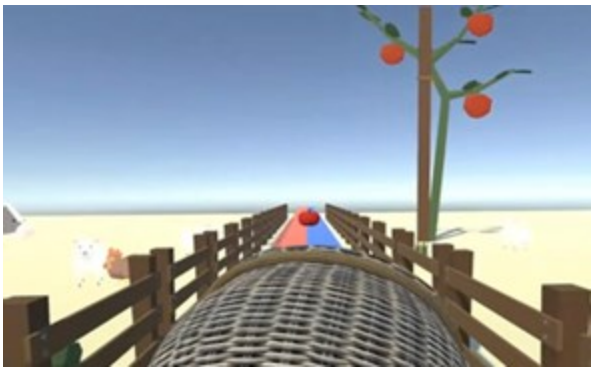


Fig.1 Motor-cognitive task: apple-catching VR game

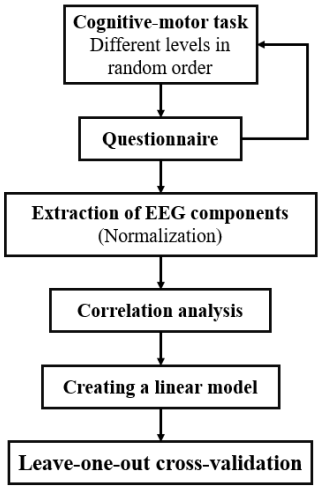


Fig.2 Experimental and analytical pipeline

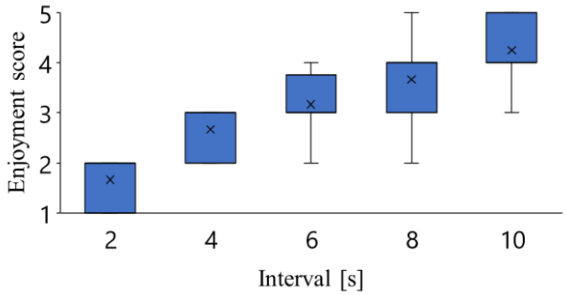


Fig.3 Relationship between task difficulty (interval) and questionnaire scores

III. RESULTS

Figure 3 shows the relationship between task difficulty (time intervals) and questionnaire scores (enjoyment). A significant strong positive correlation (r = 0.834, $p < 0.01$) was found between interval and enjoyment, indicating that the longer the interval, the less enjoyable and boring it is. This confirms that the enjoyment can be varied by changing the interval of the objects to interact.

Table 1 shows the correlation coefficients r between the subjective enjoyment obtained from the questionnaire and the normalized power values. The single asterisk indicates a

Table 1 Correlation coefficient *r* between questionnaire scores and normalized power values

r		Channels					
		F3	F4	C3	C4	P3	P4
Frequency	Delta band	-0.25*	0.050	-0.066	-0.088	0.056	0.19
	Theta band	-0.28*	-0.12	0.027	-0.034	0.20	0.28*
	Alpha band	0.051	-0.012	0.32*	0.52**	0.58**	0.51**
	Beta band	-0.28*	-0.16	0.13	0.080	0.50**	0.39**

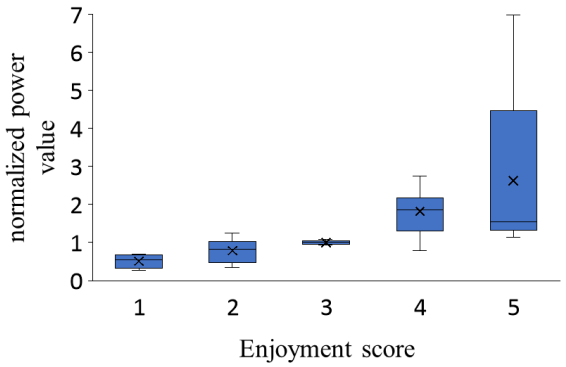


Fig.4 Relationship between questionnaire score and normalized power value

Table 2 Correlation coefficient *r* between task difficulty and normalized power values

r		Channels					
		F3	F4	C3	C4	P3	P4
Frequency	Delta band	-0.14	0.17	0.035	0.082	0.17	0.22
	Theta band	-0.17	0.013	0.11	0.13	0.24	0.31*
	Alpha band	0.13	0.038	0.38*	0.46**	0.49**	0.43**
	Beta band	-0.11	-0.10	0.23	0.21	0.48**	0.47**

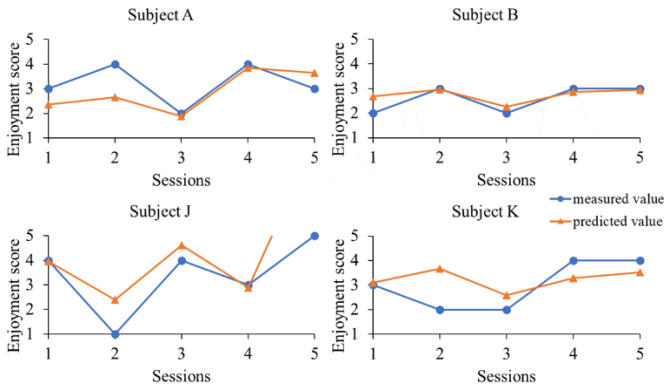


Fig.5 Comparison of measured and predicted values in representative participants.

statistically significant correlation at the 5% significance level, and the double asterisks denote a statistical significance at the 1% significance level. Significant positive correlations were observed in the central-parietal channels in the alpha and beta bands. Since the higher questionnaire score shows less enjoyment, these results indicate that the increased enjoyment is associated with decreased power in these channels and frequency bands. The strongest positive correlation ($r = 0.58$, $p < 0.01$) was found between enjoyment and normalized alpha band power values in channel P3 (Figure 4). On the other hand, a broadband signal from the left frontal channel (F3), except for alpha band power, showed a significant negative correlation with the enjoyment scores. The increase in frontal spectral power may be another signature of the neurophysiological response of enjoyment.

Table 2 shows the correlation coefficients *r* between the task difficulty (intervals) and the normalized power values. Significant and moderate correlations were also found in the central-parietal channels in the alpha and beta bands. However, there was no statistically significant correlation between task difficulty and left frontal spectral powers.

A stepwise regression analysis yielded the model equation as in Equation (1). Here, x_{ij} represents the normalized power values in the frequency band i at channel j, and y represents the estimated questionnaire scores, respectively.

$$y \approx -4.7 + 0.44x_{\delta F4} + 4.7x_{\theta P3} + 0.97x_{\alpha P3} + 8.1x_{\beta P3} - 6.4\,x_{\theta P3}x_{\beta P3}. \quad (1)$$

The correlation coefficient of the model equation was 0.814 ± 0.168 (mean ± standard deviation). The result indicates that P3 is important for estimating enjoyment [8]. Figure 5 shows a comparison of the measured and predicted values for representative four participants from the LOOCV analysis.

LOOCV results showed that the average error between the estimated enjoyment scores and the actual questionnaire scores was 0.37 ± 1.4. The mean and standard deviation of the participant-wise correlation coefficients between the measured and predicted scores of enjoyment were $r = 0.45 \pm 0.48$.

IV. DISCUSSION

Using a VR-based cognitive-motor task of varying difficulty, the current study investigated EEG biomarker signals for the positive emotion of enjoyment. Questionnaire scores confirmed that the five pre-defined levels of task difficulty successfully elicited different intensities of enjoyment, with shorter intervals between flying objects eliciting more enjoyment. Correlation analysis and stepwise regression both indicated the importance of left parietal activity in determining the intensity of enjoyment. The deviation of the estimated enjoyment intensity from the ground truth questionnaire score was small enough to classify the predicted value into the correct level of enjoyment

in most cases, suggesting the usefulness of the proposed approach.

The involvement of the left parietal lobe in positive emotional processing is consistent with the study by Dolan et al [8], who found specific activation of the left parietal lobe for happy face stimuli. Compared to the widely studied negative emotions, which are easier to discriminate into different classes, less attention has been paid to the recognition of positive emotions, which are not fully independent [9]. In recent years, studies have been conducted to classify positive emotions [10] [11], but few studies have focused on the degree of "enjoyment" within positive emotions, as in the present study.

Focusing on the large coefficients in the model equation, it is suggested that enjoyment is maximized when both the θ band and the β band are suppressed, and boredom is felt when the θ band and the β band are antagonistic. Table 1 shows that the suppression of the α band has the highest correlation with enjoyment, followed by the β band. In this model equation, the β band becomes a stronger feature by taking the product with θ band.

However, from Tables 1 and 2, significant correlations with normalized power values were found not only with the enjoyment scores from the questionnaire but also with the speed of the game, suggesting that the brain activity related to game speed may have confounded the results, rather than brain activity related to enjoyment alone. Because the parietal area is part of an association area that combines and processes multimodal sensory and motor information, it may have been activated by engagement with the task [12]. The task design needs to be improved in future studies to simulate conditions with different levels of enjoyment while ensuring that the required motor responses remain comparable. Another area for improvement is the use of the good temporal resolution of the EEG. While we derived the feature EEG values as average power over the whole minute of gameplay, there may have been changes in emotional states during a session. The time-frequency analysis of the EEG data can be combined with a more fine-grained online or post-hoc subjective rating [13] [14] to further improve the quality of the annotation and the accuracy of the model. In the future, we would also like to take advantage of the good temporal resolution of EEG to determine enjoyment in real time. For this reason, we used raw EEG data for this analysis without highly complex preprocessing such as noise reduction using independent component analysis [15]. However, the effect of artifact contamination on the emotion estimation should be quantified using the raw and cleaned datasets to confirm the reliability of the proposed method. In addition, further experiments need to be conducted to collect a larger and more diverse group of participants, which will provide more reliable data.

V. Conclusions

The current study used a VR-based cognitive-motor task of varying difficulty to investigate EEG biomarker signals associated with the positive emotion of enjoyment. The results indicated that left parietal activity plays a significant role in determining the intensity of enjoyment. The emotion of enjoyment may be represented by complex combination of electrophysiological responses. Suppression of the α band showed the highest correlation with enjoyment, followed by the β band, with the combination of θ and β bands. While caution should be taken regarding differences in game speed that may affect the results, the LOOCV results demonstrate the potential of using EEG to estimate the positive emotion of enjoyment. Future research should focus on refining task design to better isolate emotional responses and taking advantage of the high temporal resolution of EEG for more detailed analysis of emotional dynamics.

VI. Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number JP24K03244.

References

[1] Hart, J. D., Piumsomboon, T., Lee, G. A., Smith, R. T., & Billinghurst, M. (2021, May). Manipulating avatars for enhanced communication in extended reality. In *2021 IEEE International Conference on Intelligent Reality (ICIR)* (pp. 9-16). IEEE.

[2] Noah, J. A., Zhang, X., Dravida, S., Ono, Y., Naples, A., McPartland, J. C., & Hirsch, J. (2020). Real-time eye-to-eye contact is associated with cross-brain neural coupling in angular gyrus. *Frontiers in human neuroscience*, *14*, 19.

[3] Kelley, M. S., Noah, J. A., Zhang, X., Scassellati, B., & Hirsch, J. (2021). Comparison of human social brain activity during eye-contact with another human and a humanoid robot. *Frontiers in Robotics and AI*, *7*, 599581.

[4] Paradisi, P., Raglianti, M., & Sebastiani, L. (2021). Online communication and body language. *Frontiers in behavioral neuroscience*, *15*, 709365.

[5] Bahreini, K., Van der Vegt, W., & Westera, W. (2019). A fuzzy logic approach to reliable real-time recognition of facial emotions. *Multimedia Tools and Applications*, *78*, 18943-18966.

[6] Ikeda, Y., & Sugaya, M. (2016). Estimate emotion method to use biological, symbolic information preliminary experiment. In *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience: 10th International Conference, AC 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings, Part I 10* (pp. 332-340). Springer International Publishing.

[7] Appelhans, B. M., & Luecken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. *Review of general psychology*, *10*(3), 229-240.

[8] Dolan, R. J., Fletcher, P., Morris, J., Kapur, N., Deakin, J. F. W., & Frith, C. D. (1996). Neural activation during covert processing of positive emotional facial expressions. *Neuroimage*, *4*(3), 194-200.

[9] Zhao, G., Zhang, Y., Zhang, G., Zhang, D., & Liu, Y. J. (2020). Multi-target positive emotion recognition from EEG signals. *IEEE Transactions on Affective Computing*, *14*(1), 370-381.

[10] Hu, X., Yu, J., Song, M., Yu, C., Wang, F., Sun, P., ... & Zhang, D. (2017). EEG correlates of ten positive emotions. *Frontiers in human neuroscience*, *11*, 26.

[11] Liu, Y. J., Yu, M., Zhao, G., Song, J., Ge, Y., & Shi, Y. (2017). Real-time movie-induced discrete emotion recognition from EEG signals. *IEEE Transactions on Affective Computing*, *9*(4), 550-562.

[12] Tanzer, J. R., & Weyandt, L. (2020). Imaging happiness: Meta analysis and review. *Journal of Happiness Studies*, *21*, 2693-2734.

[13] Zhang, S., & Ono, Y. (2022). A study of autonomic nerve indexes related to subjective discomfort during a virtual reality experience. *The Autonomic Nervous System*, *59*(2), 246-254.

[14] Watanabe, S., Suzuki, T., Ono, Y. (2024). Predicting the Intensity of the Flow State using EEG and fNIRS biomarkers. Advanced Biomedical Engineering, in Press.

[15] Islam, M. K., Rastegarnia, A., & Yang, Z. (2016). Methods for artifact detection and removal from scalp EEG: A review. *Neurophysiologie Clinique/Clinical Neurophysiology*, *46*(4-5), 287-305.

Exploring Brain Connectivity Patterns and Cognitive Resilience in Aging: A Study with the LEMON Dataset

Kapeleshh KS†, Wei Chen†, Prince Aldrin Domer†, Hong Ji*
* The Shaanxii Key Laboratory of Clothing Intelligence, Xi'an Polytechnic University, China
E-mail: itsjihong@gmail.com Tel: +81-8057936828
† AI Research and Innovation Hub, MACNICA, Inc., Japan
E-mail: ks-k@macnica.co.jp Tel: +81-7088301120

***Abstract*—Investigating how brain connectivity and cognitive resilience evolve throughout life in healthy individuals is crucial to understanding the underpinnings of cognitive aging. This study uses the Leipzig Mind-Brain-Body (LEMON) dataset to explore the evolution of brain connectivity and cognitive resilience, focusing on differences in connectivity metrics among young (20–30 years) and old adults (70–80 years) and their correlation with cognitive performance measured by the California Verbal Learning Test (CVLT). Despite observable declines in connectivity with age, the data suggest a maintained network stability, potentially aiding in cognitive preservation. However, the weak correlations between connectivity metrics and cognitive performance suggest that cognitive resilience in aging could involve mechanisms beyond traditional connectivity measures, underscoring the importance of further exploring the complex interactions between brain networks and cognitive functions in healthy aging.**

I. Introduction

Aging is associated with numerous physiological, cognitive, and neural changes, often leading to declines in memory, attention, and executive function, along with structural and functional alterations in neural circuits and connectivity [1] Understanding these changes is crucial for developing effective strategies to promote healthy aging and mitigate cognitive decline. One promising avenue of research involves cognitive resilience, which refers to the ability of an individual to maintain cognitive function despite age-related neural changes. Cognitive resilience plays a key role in supporting healthy aging, allowing individuals to maintain independence and quality of life.

Research has shown that brain connectivity-the interactions and communication between brain regions is fundamental to understanding the processes that support cognitive function and resilience [2]. Connectivity within specific frequency bands, such as alpha, beta, and gamma, provides insight into how efficiently different regions of the brain communicate, which is crucial for maintaining cognitive abilities. Declines in connectivity efficiency, especially in higher-frequency bands like beta and gamma, are often associated with impairments in cognitive functions. On the other hand, alpha band connectivity is suggested to be more stable with age, which makes it a particularly interesting target for understanding resilience[3][4].

Existing studies on age-related changes in brain connectivity have often reported decreased connectivity and efficiency in older adults[5]. This reduction suggests a diminished capacity of the aging brain to integrate information across widespread networks, impacting overall cognitive performance. However, much remains unknown about the specific changes in connectivity measures and how they relate to cognitive resilience, especially when comparing young and old adults[6].

This research effort is crucial because early identification of connectivity patterns associated with cognitive decline can enable targeted interventions to maintain cognitive health [7][8]. By understanding how connectivity changes across the lifespan, we can better determine who might be at risk for cognitive impairments and intervene early to slow or even prevent further decline [9]. Research indicates that interventions such as cognitive training, physical exercise, and lifestyle modifications can positively impact brain connectivity and cognitive function, particularly if implemented during early or middle stages of aging [10][11]. Additionally, findings from this study can contribute to enhancing current practices in healthcare settings by providing objective markers for cognitive resilience, helping clinicians design personalized treatment plans aimed at improving quality of life in older adults [12].

This study aims to address these gaps by analyzing brain connectivity and cognitive resilience across the lifespan using the LEMON dataset. Specifically, we examine changes in brain connectivity metrics, such as connection counts, connection strengths, and global efficiency, to determine their association with cognitive resilience, measured by the California Verbal Learning Test (CVLT)[13]. In this study, we focus on two distinct age groups: younger adults (20–30 years) and older adults (70-80 years), comprising 124 and 22 subjects, respectively. By using Phase Locking Value (PLV) and Coherence (COH) matrices to quantify connectivity, we explore how differences in connectivity are linked to cognitive resilience across age groups.

Our study distinguishes itself from existing research by combining multiple connectivity measures (PLV, COH, global efficiency) with a cognitive performance assessment (CVLT), providing a more comprehensive understanding of cognitive

resilience and aging[14]. Additionally, we focus on specific frequency bands (alpha, beta, and gamma) to discern how different aspects of connectivity may change with age and impact cognitive function[15].

Identifying differences in brain connectivity and cognitive resilience between age groups is essential for understanding the underlying mechanisms of aging and for designing strategies that promote healthy aging[16][17]. Our study aims to enhance the understanding of how aging affects cognitive resilience, providing insights that may inform clinical practices and public health initiatives aimed at improving brain health across the lifespan.

II. Dataset

We used data from the publicly available Leipzig Study for Mind-Body-Emotion Interactions (LEMON) dataset[13], which consists of comprehensive neuroimaging and behavioural data from 228 healthy participants. A 16-min rs-EEG was recorded with a BrainAmp MR plus amplifier in an electrically shielded and sound-attenuated EEG booth using 62-channel (61 scalp electrodes plus 1 electrode recording the VEOG below the right eye) active ActiCAP electrodes (both Brain Products GmbH, Gilching, Germany) attached according to the international standard 10–20 extended localization system, also known as 10-10 system, and referenced to FCz. The dataset includes a young group (N=154, 25.1±3.1 years, range 20–35 years, 45 female) and an elderly group (N=74, 67.6±4.7 years, range 59–77 years, 37 female) acquired cross-sectionally in Leipzig, Germany, between 2013 and 2015.

During two-days assessment, participants completed a series of MRI scans at 3 Tesla and a 62-channel EEG experiment at rest. The resting-state EEG data was collected in two conditions: Eyes Open and Eyes Closed. Additionally, participants completed various cognitive tests, including the CVLT.

We did not use all the data from all the subjects in the LEMON dataset for this study. We selected only young participants aged between 20-30 years, totaling 124 subjects, and old participants aged between 70-80 years, totaling 22 subjects.

III. Methods

To analyze EEG signals, we employed PLV and COH matrices as alternatives to raw EEG data. These methodologies are favored due to their robustness against noise and their capability to reveal meaningful functional connectivity between brain regions. PLV quantifies phase synchrony, which assesses the consistency of phase relationships across trials, thereby illuminating synchronous neural activities. On the other hand, coherence measures the linear relationships between signals in the frequency domain, providing insights into the strength and stability of frequency specific interactions. Both methods are well-established in the study of brain network interactions and offer a biologically relevant, simplified representation of complex neural dynamics[1][18]. This representation facilitates the use of network analysis techniques, including connection count, connection strength, global efficiency, and resilience measures[19], which are crucial for understanding the efficiency and adaptability of the brain's functional network. By utilizing PLV and coherence, our analysis robustly focuses on dynamic interactions between brain regions, yielding more robust and interpretable results[20][21].

A. Preprocessing

The data were band-pass filtered to isolate alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-100 Hz) frequency bands for both the Eyes Closed, Eyes Open datasets. Phase Locking Value (PLV) and Coherence (COH) matrices were computed for each frequency band to quantify connectivity between brain regions. Cognitive resilience was assessed using the CVLT. Factor analysis was performed on CVLT scores to identify underlying cognitive factors, with the optimal number of factors determined using a scree plot, which resulted in 4 factors.

EEG data were recorded using a 62-channel system with participants at rest. The data were band-pass filtered to isolate alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-100 Hz) frequency bands for both the Eyes Closed, Eyes Open datasets. Phase Locking Value (PLV) and Coherence matrices were computed for each frequency band to quantify connectivity between brain regions. Cognitive resilience was assessed using the CVLT. Factor analysis was performed on CVLT scores to identify underlying cognitive factors, with the optimal number of factors determined using a scree plot, which resulted in 4 factors.

Connectivity patterns are essential for understanding brain activity, as they significantly affect regional activation during tasks. Connectivity is vital for shaping dynamic brain activity[20]. To explore these dynamic connectivity patterns, we employed PLV and COH methods, enabling a detailed analysis of temporal dynamics of brain function.

B. PLV Computation

The PLV is a measure of the phase synchrony between pairs of EEG signals, indicating the consistency of phase differences across trials [21]. Previous studies on age-related changes studied changes in phase synchrony using PLV [22]. The PLV for each electrode pair was computed as Eq 1:

$$PLV = \left| \frac{1}{N} \sum_{t=1}^{N} e^{i\Delta\phi(t)} \right| \qquad \text{(Eq 1)}$$

where N is the number of time points, j the imaginary unit, and $\Delta\phi_{ij}(t)$ is the phase between pairs of electrodes i and j calculated for each time point t. The PLV ranges from 0 to 1, with 1 indicating perfect phase synchrony and 0 indicating no synchrony.

C. COH Computation

COH is a measure of the linear relationship between the frequencies of two EEG signals, reflecting both amplitude and phase consistency. Similarly, previous studies have shown the utility of COH matrix in analysing age-related changes

Data Name	Optimal Threshold	AUC Score	t-test	p-val	Avg. Retain	Stat. Relevance
EC_alpha	0.4	0.71	3.32	0.00114	93%	True
EC_beta	0.9	0.63	1.49	0.13788	37%	False
EC_gamma	0.9	0.77	4.08	0.00007	29%	True
EO_alpha	0.4	0.73	3.70	0.00030	90%	True
EO_beta	0.9	0.74	2.83	0.00537	38%	True
EO_gamma	0.8	0.82	4.50	0.00001	41%	True

Table I: Optimal threshold values for each dataset, along with associated AUC scores, t-test, p-val, and average connection retention percentages. The table also indicates whether the results are statistically significant. The term 'EC' refers to the Eyes Closed condition, while 'EO' denotes the Eyes Open condition.

[23]. COH $C_{xy}(f)$ for each electrode pair and frequency was computed as Eq 2:

$$C_{xy}(f) = \frac{|\mathcal{S}_{xy}(f)|^2}{\mathcal{S}_{xx}(f)\mathcal{S}_{yy}(f)} \quad \text{(Eq 2)}$$

Where $S_{xy}(f)$, $S_{xx}(f)$ and $S_{yy}(f)$ are cross and power spectral densities respectively. COH values range from 0 to 1, with higher values indicating stronger linear relationships between the signals at a particular frequency.

For each frequency band, COH matrices were constructed. Each $n \times n$ matrix represents the COH between all pairs of n electrodes, where each element $C_{ij}(f)$ denotes the COH value between electrode i and j at a specific frequency.

For both PLV and COH Matrix, matrices were constructed for both Resting State Eyes Open and Resting State Eyes Closed conditions across all frequency bands.

D. Connection Counts and Connection Strengths

To quantify the connectivity patterns within brain networks, we computed two primary metrics: connection counts and connection strengths[24].

Connection counts indicate the number of connections between electrodes (nodes) and were obtained by summing the binary adjacency matrix for each electrode, reflecting the total number of significant connections[25].

Connection strengths represent the cumulative value of connections between electrodes, calculated by summing the original (non-binarized) connectivity matrix for each electrode, providing an overall measure of connection strength[26].

E. Global Efficiency

Global efficiency measures how efficiently information is exchanged across the entire network. Higher global efficiency indicates better integration of information across distant brain regions. The global efficiency E_{glob} for each subject's brain network was calculated using the following Eq 3:

$$E_{\text{glob}} = \frac{1}{N(N-1)} \sum_{i \neq j \in V} \frac{1}{d_{ij}} \quad \text{(Eq 3)}$$

where N is the number of nodes (electrodes) and d_{ij} is the shortest length between nodes i and j.

F. Network Resilience Analysis

Robustness to Node Removal: Random or targeted removal of nodes (electrodes) was simulated to evaluate the network's ability to maintain connectivity. Metrics such as giant component size, network diameter, and clustering coefficient were monitored to assess changes in network structure[27].

Resilience to Edge Removal: Similarly, edges (connections between electrodes) were selectively removed to analyze the impact on network connectivity. Measures like modularity and assortative were examined to understand network stability.

IV. Analysis results

We compared data from two groups:

- Young group: 124 participants (aged 20-30 years)
- Old group: 22 participants (aged 70-80 years)

Our goal was to analyze the differences between these age groups using various methods mentioned above to gain insights into cognitive abilities and brain network characteristics associated with aging.

A. Connection Counts and Connection Strengths

To calculate connection counts and strengths, determining the optimal threshold is essential. The Receiver operating characteristic (ROC) and Area Under the Curve (AUC) method was chosen for its comprehensive assessment of threshold performance across all possible thresholds, combining threshold agnosticism with statistical rigor for optimal selection. The optimal threshold was identified based on the maximum AUC score, representing the point where connectivity metrics effectively differentiated between young and old subjects. An independent t-test was conducted to compare connection counts between the age groups at the optimal threshold, assessing the statistical significance of the differences. Additionally, the average retain percent at this threshold was calculated to provide further insights into brain network connectivity, as shown in Table I. The analysis revealed that the average connection count and strength were consistently higher in young subjects than in old subjects across all frequency bands and both Eyes Open and Eyes Closed conditions, as shown in Table II and Table III.

B. Global Efficiency

Global efficiency, which measures how efficiently information is exchanged across the brain network, was computed for each subject. An independent t-test assessed the statistical

Metric	Age Group	No. Sub	EC alpha (thres=0.4)	EC beta (thres=0.9)	EC gamma (thres=0.9)	EO alpha (thres=0.4)	EO beta (thres=0.9)	EO gamma (thres=0.8)
PLV	20-30	124	**1720.51**	**713.48**	**563.59**	**1680.4**	**734.71**	**819.04**
	70-80	22	1649.27	661.04	446.0	1588.82	636.0	632.18

Metric	Age Group	No. Sub	EC alpha (thres=0.5)	EC beta (thres=0.5)	EC gamma (thres=0.6)	EO alpha (thres=0.1)	EO beta (thres=0.6)	EO gamma (thres=0.5)
COH	20-30	124	**944.23**	**937.46**	**529.99**	**1826.2**	**708.85**	**722.48**
	70-80	22	878.18	869.68	426.64	1812.5	618.4	557.27

Table II: This table described average connection retention count in PLV and COH matrices for each dataset at their optimal thresholds, comparing young and old subjects. The term 'EC' refers to the Eyes Closed condition, while 'EO' denotes the Eyes Open condition.

Metric	Age Group	No. Sub	EC alpha (thres=0.4)	EC beta (thres=0.9)	EC gamma (thres=0.9)	EO alpha (thres=0.4)	EO beta (thres=0.9)	EO gamma (thres=0.8)
PLV	20-30	124	**1361.5**	**684.00**	**540.35**	**1347.51**	**705.91**	**759.65**
	70-80	22	1296.99	635.75	427.77	1257.43	611.46	585.56

Metric	Age Group	No. Sub	EC alpha (thres=0.5)	EC beta (thres=0.5)	EC gamma (thres=0.6)	EO alpha (thres=0.1)	EO beta (thres=0.6)	EO gamma (thres=0.5)
PLV	20-30	124	**688.91**	**672.41**	**405.53**	**960.32**	**549.52**	**514.26**
	70-80	22	637.76	629.45	326.8	883.93	481.43	397.30

Table III: This table presents the average connection strength for Phase Locking Value (PLV) and Coherence (COH) matrices for each dataset at their optimal thresholds, comparing young and older subjects. The term 'EC' refers to the Eyes Closed condition, while 'EO' denotes the Eyes Open condition.

significance of differences between age groups. Significant differences in global efficiency were observed between young and old participants in the beta and gamma frequency bands, while differences in the alpha band were insignificant, shown as in Table IV. Correlation analysis between global efficiency and CVLT scores showed no significant relationship, indicating that global efficiency does not strongly correlate with cognitive ability in the participants.

		alpha		beta		gamma	
		t-test	**p-val**	**t-test**	**p-val**	**t-test**	**p-val**
Global Efficiency	PLV	1.64	0.10	3.16	0.002	2.53	0.012
	COH	1.85	0.07	2.63	0.009	2.37	0.018
Resilience Measures	PLV	-0.85	0.40	-0.91	0.37	-1.024	0.31
	COH	-0.75	0.45	-0.44	0.66	-1.85	0.065

Table IV: Global Efficiency and Resilience measure comparison scores between old and young subjects with t-test and p-val.

C. Resilience Measures

Resilience measures were calculated to evaluate the robustness and stability of brain networks in young and old subjects. These measures provide insights into the brain's ability to maintain functionality despite potential disruptions. Our findings suggest that overall network resilience, in terms of cognitive abilities and network integrity, did not significantly decline with age in our sample as shown in Table IV. Correlation analysis between resilience measures and CVLT scores revealed no significant relationship, indicating that network resilience may not strongly correlate with cognitive ability in the participants.

V. Discussion

The connection count and connection strength analysis revealed that the differences between young and old participants are not very significant. The constraint of our study, which involved a notably larger sample size of young participants (124 subjects) compared to old participants (22 subjects).

There was a significant statistically relevant difference in global efficiency between young and old participants only in the beta and gamma bands, whereas no statistically relevant difference was observed in the resilience measures. Previous studies have shown that global efficiency tends to decrease with age, although the extent and specific regions affected can vary. This decline indicates a reduced ability of the aging brain to integrate information across widespread networks, impacting overall cognitive performance[18].

Moreover, the decline in local efficiency is associated with impairments in cognitive functions, particularly those that rely on localized processing[28]. Some studies suggest that alpha band connectivity and power might not exhibit as pronounced age-related declines as beta and gamma bands. This empirical evidence supports the idea that alpha rhythms are more stable with age.

VI. Conclusion

This research systematically explored changes in brain connectivity and cognitive resilience with aging using the LEMON dataset. Our findings indicate minor declines in connectivity with age. however, the overall structural integrity and functionality of brain networks remain stable, demonstrating an intrinsic resilience that supports cognitive function in healthy old adults. Advancing knowledge of the aging brain empowers individuals to take proactive steps to protect their cognitive well-being, including adopting lifestyle changes and participating in targeted training programs[29][30].

Despite reductions in connectivity metrics, correlations between these metrics and cognitive performance, as assessed by

the CVLT, were weak. This suggests that cognitive resilience in the elderly may not be directly linked to conventional connectivity measures, potentially indicating other underlying factors or compensatory mechanisms that aid cognitive maintenance during aging. This highlights the complexity of brain connectivity and its relationship with cognitive function, which may not be fully captured by direct neural interconnections.

Our results enhance the understanding of healthy aging by emphasizing the need for further research into the interactions between brain networks and cognitive function. Future studies should integrate diverse metrics of brain function and cognitive performance to elucidate these relationships. Such efforts are essential for advancing our understanding of aging and developing targeted interventions to enhance cognitive resilience in the elderly.

This study underscores the significance of using EEG data and cognitive testing to explore brain-behavior relationships, supporting the development of strategies for healthy cognitive aging. As we continue to unravel these complex dynamics, the insights gained will enrich our understanding of the aging brain and improve the cognitive health and quality of life of old adults.

VII. Acknowledgment

We sincerely thank Prof. Toshihisa Tanaka from Tokyo University of Agriculture and Technology for his invaluable guidance.

VIII. Funding

This work was supported by National Natural Science Foundation Grant (Program No. 62106189) to HJ. The funding sources were not involved in the study design.

References

[1] E. J. Oosterhuis, K. Slade, P. J. C. May, and H. E. Nuttall, "Toward an understanding of healthy cognitive aging: The importance of lifestyle in cognitive reserve and the scaffolding theory of aging and cognition," *The Journals of Gerontology: Series B*, vol. 78, no. 5, pp. 777–788, 2023.

[2] E. Santarnecchi, G. Sprugnoli, E. Tatti, *et al.*, "Brain functional connectivity correlates of coping styles," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 18, pp. 495–508, 2018.

[3] A. Miljevic, N. W. Bailey, F. Vila-Rodriguez, S. E. Herring, and P. B. Fitzgerald, "Electroencephalographic connectivity: A fundamental guide and checklist for optimal study design and evaluation," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 7, no. 6, pp. 546–554, 2022.

[4] P. Babaeeghazvini, L. M. Rueda-Delgado, J. Gooijers, S. P. Swinnen, and A. Daffertshofer, "Brain structural and functional connectivity: A review of combined works of diffusion magnetic resonance imaging and electro-encephalography," *Frontiers in Human Neuroscience*, vol. 15, p. 721 206, 2021.

[5] L. Geerligs, R. J. Renken, E. Saliasi, N. M. Maurits, and M. M. Lorist, "A brain-wide study of age-related changes in functional connectivity," *Cerebral cortex*, vol. 25, no. 7, pp. 1987–1999, 2015.

[6] S. Baghernezhad and M. R. Daliri, "Age-related changes in human brain functional connectivity using graph theory and machine learning techniques in resting-state fmri data," *GeroScience*, pp. 1–18, 2024.

[7] F. Ribaldi, C. Chicherio, D. Altomare, *et al.*, "Brain connectivity and metacognition in persons with subjective cognitive decline (coscode): Rationale and study design," *Alzheimer's Research & Therapy*, vol. 13, no. 1, p. 105, 2021.

[8] S. Dautricourt, J. Gonneaud, B. Landeau, *et al.*, "Dynamic functional connectivity patterns associated with dementia risk," *Alzheimer's research & therapy*, vol. 14, no. 1, p. 72, 2022.

[9] V. La Corte, M. Sperduti, C. Malherbe, *et al.*, "Cognitive decline and reorganization of functional connectivity in healthy aging: The pivotal role of the salience network in the prediction of age and cognitive performances," *Frontiers in aging neuroscience*, vol. 8, p. 204, 2016.

[10] National Institute on Aging, *Lifestyle interventions can reduce dementia risk*, U.S. Department of Health and Human Services, 2021. [Online]. Available: https://www.nia.nih.gov/report-2020-2021-scientific-advances-prevention-treatment-and-care-dementia/lifestyle-interventions (visited on 10/04/2024).

[11] T. Salzman, Y. Sarquis-Adamson, S. Son, M. Montero-Odasso, and S. Fraser, "Associations of multidomain interventions with improvements in cognition in mild cognitive impairment: A systematic review and meta-analysis," *JAMA Network Open*, vol. 5, no. 5, e226744–e226744, 2022.

[12] M. Montero-Odasso, G. Zou, N. Kamkar, *et al.*, "Multidomain trials to prevent dementia: Addressing methodological challenges," *Alzheimer's research & therapy*, vol. 14, no. 1, p. 94, 2022.

[13] A. Babayan, M. Erbey, D. Kumral, *et al.*, "A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults," *Scientific data*, vol. 6, no. 1, pp. 1–21, 2019.

[14] R. Sanchez-Romero and M. W. Cole, "Combining multiple functional connectivity methods to improve causal inferences," *Journal of cognitive neuroscience*, vol. 33, no. 2, pp. 180–194, 2021.

[15] X. Li, Y. Wu, M. Wei, *et al.*, "A novel index of functional connectivity: Phase lag based on wilcoxon signed rank test," *Cognitive Neurodynamics*, vol. 15, pp. 621–636, 2021.

[16] F. Pourmotahari, H. Doosti, N. Borumandnia, S. M. Tabatabaei, and H. Alavi Majd, "Group-level comparison of brain connectivity networks," *BMC Medical Research Methodology*, vol. 22, no. 1, p. 273, 2022.

[17] M. Filippi, C. Cividini, S. Basaia, *et al.*, "Age-related vulnerability of the human brain connectome," *Molecular Psychiatry*, vol. 28, no. 12, pp. 5350–5358, 2023.

[18] L. K. Ferreira, A. C. B. Regina, N. Kovacevic, *et al.*, "Aging effects on whole-brain functional connectivity in adults free of cognitive and psychiatric disorders," *Cerebral cortex*, vol. 26, no. 9, pp. 3851–3865, 2016.

[19] P. N. Young, M. Estarellas, E. Coomans, *et al.*, "Imaging biomarkers in neurodegeneration: Current and future practices," *Alzheimer's research & therapy*, vol. 12, pp. 1–17, 2020.

[20] M. Bernstein-Eliav and I. Tavor, "The prediction of brain activity from connectivity: Advances and applications," *The Neuroscientist*, vol. 30, no. 3, pp. 367–377, 2024.

[21] M. J. Hülsemann, E. Naumann, and B. Rasch, "Quantification of phase-amplitude coupling in neuronal oscillations: Comparison of phase-locking value, mean vector length, modulation index, and generalized-linear-modeling-cross-frequency-coupling," *Frontiers in neuroscience*, vol. 13, p. 573, 2019.

[22] B. Moezzi, L. M. Pratti, B. Hordacre, *et al.*, "Characterization of young and old adult brains: An eeg functional connectivity analysis," *Neuroscience*, vol. 422, pp. 230–239, 2019.

[23] S. M. Bowyer, "Coherence a measure of the brain networks: Past and present," *Neuropsychiatric Electrophysiology*, vol. 2, pp. 1–12, 2016.

[24] Z.-Q. Liu, R. F. Betzel, and B. Misic, "Benchmarking functional connectivity by the structure and geometry of the human brain," *Network Neuroscience*, vol. 6, no. 4, pp. 937–949, 2022.

[25] F. V. Farahani, W. Karwowski, and N. R. Lighthall, "Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review," *frontiers in Neuroscience*, vol. 13, p. 585, 2019.

[26] S. H. Tompson, E. B. Falk, J. M. Vettel, and D. S. Bassett, "Network approaches to understand individual differences in brain connectivity: Opportunities for personality neuroscience," *Personality neuroscience*, vol. 1, e5, 2018.

[27] O. Artime, M. Grassia, M. De Domenico, *et al.*, "Robustness and resilience of complex networks," *Nature Reviews Physics*, vol. 6, no. 2, pp. 114–131, 2024.

[28] T. H. Grandy, M. Werkle-Bergner, C. Chicherio, F. Schmiedek, M. Lövdén, and U. Lindenberger, "Peak individual alpha frequency qualifies as a stable neurophysiological trait marker in healthy younger and older adults," *Psychophysiology*, vol. 50, no. 6, pp. 570–582, 2013.

[29] K. A. Tsvetanov, R. N. Henson, L. K. Tyler, *et al.*, "Extrinsic and intrinsic brain network connectivity maintains cognition across the lifespan despite accelerated decay of regional brain activation," *Journal of Neuroscience*, vol. 36, no. 11, pp. 3115–3126, 2016.

[30] J. I. Fleck, J. Kuti, J. Mercurio, S. Mullen, K. Austin, and O. Pereira, "The impact of age and cognitive reserve on resting-state brain connectivity," *Frontiers in aging neuroscience*, vol. 9, p. 392, 2017.

Effect of Phase-Locked Transcranial Alternating Current Stimulation on Vocal tremor

Junting Wang[1], Satoko Koganemaru[2], Atsushi Shima[3], Yedi Cao[4], Kana Hirakawa[5], Ken iwanaga [6], Atsushi Suehiro[7], Keiko Maekawa[8], Tatsuya Mima[9], Yumie Ono[10]

[1] Meiji University Graduate School of Science and Technology, Kanagawa, Japan

E-mail: wangjt0614@gmail.com

[2-4] Human Brain Research Center, Graduate School of Medicine Kyoto University, Kyoto, Japan

E-mail: [2] kogane@kuhp.kyoto-u.ac.jp, [3] shiima@kuhp.kyoto-u.ac.jp, [4] yedi0524@kuhp.kyoto-u.ac.jp

[5-7] Graduate School of Medicine and Faculty of Medicine, Kyoto University, Kyoto, Japan

E-mail: [5] k_nagahara@ent.kuhp.kyoto-u.ac.jp, [6] k_iwanaga@ent.kuhp.kyoto-u.ac.jp, [7] a_suehiro@ent.kuhp.kyoto-u.ac.jp

[8] Department of Otolaryngology, Kobe City Medical Center General Hospital, Kobe, Japan

E-mail: maekawarimura@kcho.jp

[9] The Graduate School of Core Ethics and Frontier Sciences, Ritsumeikan University, Kyoto, Japan

E-mail: t-mima@fc.ritsumei.ac.jp

[10] Meiji University School of Science and Technology, Kanagawa, Japan

E-mail: yumie@meiji.ac.jp Tel: +81-44-934-7302

Abstract— **Vocal tremor can be found in 30% of patients with essential tremor, causing significant communication difficulties. To develop more efficient and safer treatments with less side effects for vocal tremor, this study investigated the potential of transcranial alternating current stimulation (tACS) that is temporally synchronized with tremor envelope. Three female patients with vocal tremor received tACS over the cerebellum during their phonation. Patients received phase-locked stimulation of both in-phase and out-phase with respect to the tremor, and the changes in voice characteristics was evaluated. The results showed that out-phase stimulation significantly reduced tremor amplitude intensity (AtrI). After the treatment, the patients' subjective ease of phonation improved. These findings suggest that tACS, especially out-phase stimulation, can be an effective treatment for vocal tremor and warrants further study.**

I. INTRODUCTION

Vocal tremor refers to the tremor occurring in the muscles of the larynx, vocal tract, and other vocal organs, characterized by rhythmic movements causing pitch perturbation and amplitude perturbation of the voice in the range of 4-8 Hz. Among patients with essential tremor, a condition where tremors occur without underlying diseases such as stroke or Parkinson's disease[1], 30% of them experience this symptom [2]. The trembling of the voice leads to significant communication difficulties for patients, affecting their social life. Traditional treatments, such as voice therapy and pharmacotherapy, have certain symptom-relieving effects but are limited by insufficient efficacy or high risk of side effects, emphasizing the necessity for developing new treatment methods [3-5].

Recently, electrical stimulation therapy has gained attention as an innovative treatment for brain dysfunction and neurological disorders. Among these, transcranial alternating current stimulation (tACS) has shown potential as a non-invasive neural modulation technique for treating neurological diseases [6-10]. By applying weak alternating currents through electrodes placed on the scalp, tACS modulates neuronal activity and influences related neural network functions. Its unique advantage lies in the ability to adjust the frequency and intensity of stimulation in a non-invasive manner, allowing flexible intervention in neural activity [11]. Recent reports suggest that tACS, by stimulating the cerebellum, the central pattern generator, could potentially reduce upper limb tremor symptoms in the treatment of essential tremor[12]. Therefore, tACS holds great potential for a treatment option of vocal tremor.

Previous studies have used accelerometers attached to the affected limb to extract tremor frequency and phase. Sinusoidal current phase or at some phase difference with respect to the limb movement, was applied to the cerebellum [12-14]. When a sinusoidal tACS is applied at the same frequency as the patient's tremor, the tremor is entrained by the stimulus,

maintaining a constant phase difference. This phase difference varies from patient to patient and is referred to as the entrainment phase. Studies have also reported that phase-locked stimulation using the entrainment phase as the phase difference suppresses tremor amplitude [13]. Based on these previous studies on the treatment effect of tACS on motor tremor, we investigated the feasibility of tACS for the treatment of vocal tremor. analysis has been used to aid in the evaluation of vocal tremor focuses primarily on two aspects: frequency and amplitude. For example, studies on PD have found that the amplitude tremor intensity index (ATrI) is significantly higher in the PD group than in the control group [15]. Another study in essential tremor have shown that both ATrI and frequency tremor intensity index (FTrI) are significantly increased in patients [16]. Based on these studies, we used ATrI and FTrI from previous research as indices to assess tremor by comparing the intensity of vocal tremor before and after treatment. Using this approach, we investigated the potential of transcranial alternating current stimulation (tACS) for the treatment of vocal tremor.

II. METHOD

A. Participant

An experiment was conducted on three female patients with vocal tremor at Kobe City Medical Center General Hospital (Table 1). The study was approved by the Certified Review Board of Kyoto University, and Kobe City Medical Center General Hospital was registered as a collaborating research facility for this specified clinical research. Written informed consent was obtained from all participants, and the experimental procedures were conducted in accordance with the ethical guidelines approved by the Ethics Committee of the Certified Review Board of Kyoto University (approval number: jRCTs052210211). This study was also registered in the Japan Registry of Clinical Trials (jRCT ID: jRCTs052210211).

B. Vowel Selection

Previous research by Nieuwhof et al. discussed the amplitude uniformity of tremor patterns in the upper limbs and the efficacy of tACS stimulation, indicating that tACS treatment is more effective for non-jerky sinusoidal tremor patterns than for jerky tremor patterns [6]. Therefore, patients were first asked to pronounce the Japanese vowels "/a/", "/i/", "/u/", "/e/", and "/o/" to determine the vowel whose tremor sound envelope was closest to a sinusoidal wave. The selected vowel was then used consistently during the experiment. The vowels used by each patient are listed in Table 2.

An anode was placed on the cerebellum on the same side as the affected limb tremor, and a cathode was placed on the shoulder of the opposite side. The anode electrode was centered over the midpoint between the inion and the mastoid process [17-19]. The peak-to-peak amplitude of the tACS was -1.5 to +1.5 mA. The stimulation frequency used ranged between 4 and 6 Hz, matching the frequency of the patient's vocal tremor. Fig. 1 shows the electrode placement on the patient.

C. Electrode Arrangement and Stimulation

Fig. 2 shows a schematic representation of the experimental system. A microphone and the amplifier recorded the voice of the patient. The voice signal was captured and analyzed in real-time by an in-house developed MATLAB Simulink program to generate stimulation waveforms. The generated stimulation waveform was sent to the tACS stimulation device (DC-STIMULATOR PLUS, neuroConn, Ilmenau, Germany) as an analog voltage waveform via an I/O terminal board (MF634 and TB621, HUMUSOFT, Prague, Czech Republic). The stimulation voltage waveform was modulated to the current signal with the external input function of the tACS stimulation device, which was provided to the patient.

The analog input signal at the amplifier output and the analog input signal to the tACS device were recorded with an oscilloscope and used to investigate the relationship between the voice signal and the tACS signal. The sampling frequency in the oscilloscope and the MATLAB Simulink program was 10000 Hz and 11025 Hz, respectively. The data recorded by the oscilloscope were used to analyze the phase difference between the stimulus waveform and the tremor envelope. The voice data were also recorded as wave files in the Simulink program, which were used to determine ATrI and FTrI.

Table 1. Participant demographic information

Patient	Age (year)	Sex	Diagnosis	Affect-ed Side	VHI -10
1	80	F	Vocal tremor	Left	25
2	74	F	Vocal tremor	Left	66
3	73	F	Vocal tremor	Left	23

VHI-10: Voice Handicap Index-10

Table 2. Vowels used by each patient

Patient	Vowel
Patient 1	/u/
Patient 2	/a/
Patient 3	/o/

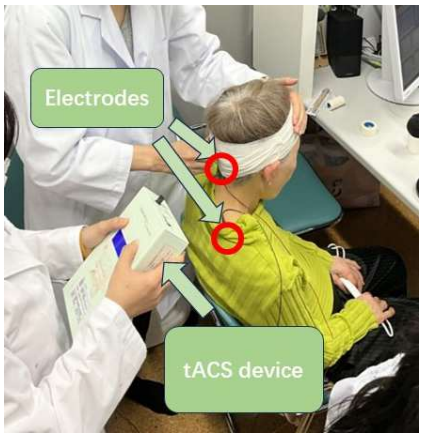


Fig. 1. Preparation of tACS electrodes with a patient

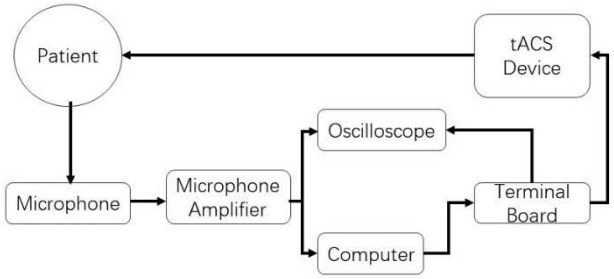


Fig. 2. Overview of the experimental system.

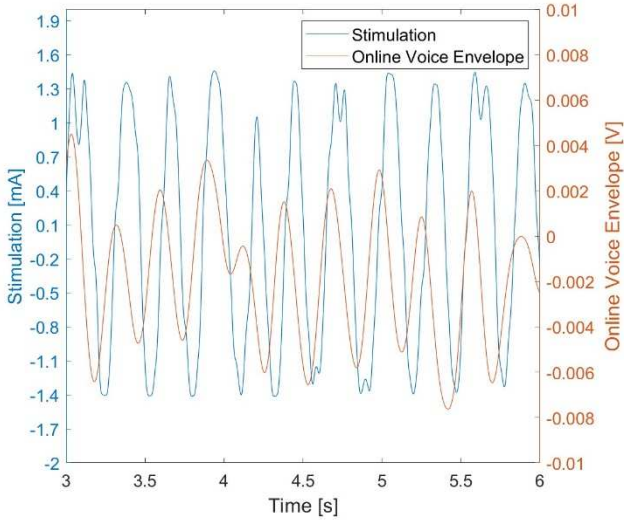
The oscilloscope simultaneously recorded the voice signal from the amplifier and the tACS waveform.

D. Experimental Protocol

The experiment consisted of a single session of an entrainment experiment and two sessions of a phase-locked stimulation experiment. Entrainment experiments were conducted to determine individual entrainment phase and phase-locked stimulation experiments were conducted to determine whether entrainment phase in-phase or entrainment out-phase stimulation was more effective. The entrainment experiment session was conducted with two pre-tACS non-stimulation voice measurements, three voice measurements under non-phase-synchronized tACS, and two post-tACS non-stimulation voice measurements. The inherent frequency of the patient's vocal tremor was calculated from the first two voice measurements. In the entrainment experiments, we applied a sinusoidal stimulus to the patients that was not phase-locked but had the same frequency as their voice amplitude tremor. The entrainment phase was determined by calculating the phase difference between the envelope of the sound waveform and the tACS stimulus waveform at its peak using the Hilbert transform. The phase differences were determined for each cycle of the tACS stimulus and presented in a histogram.

The phase-locked experiment consisted of in-phase and out-phase stimulation sessions. Each session consisted of three voice measurements under phase-synchronized alternating current stimulations, followed by two non-stimulation voice measurements. For the in-phase stimulation, the phase deviation between the stimulation waveform and the patient's voice envelope was maintained at the entrainment phase. For the out-phase stimulation, the phase deviation between the stimulation waveform and the patient's voice envelope was maintained at the antiphase of the entrainment phase (i.e. phase shift of π from the entrainment phase). The order of in-phase and out-phase stimulations was randomized. Fig. 3(a) shows an example of the entrainment phase in-phase stimulation waveform and the voice envelope. Fig. 3(b) shows a histogram of the phase difference between the envelope of the patient's voice and the stimulation waveform, calculated by the Hilbert transform. The horizontal axis represents the phase difference, and the vertical axis represents the frequency of occurrence of that phase difference. Fig. 3(b) indicates that in this task, the actual phase difference between the envelope and the stimulation waveform is close to the target phase difference.

(a)



(b)

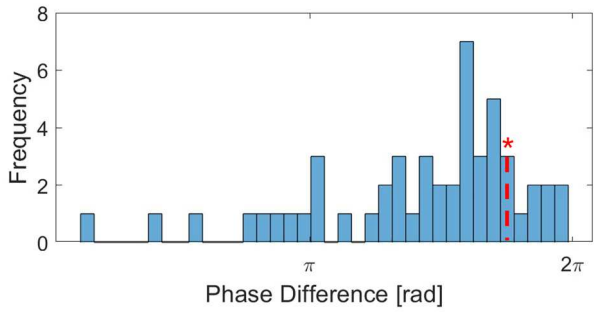


Fig. 3. Example of entrainment phase in-phase stimulation experiment.

(a) Voice envelope and the corresponding tACS waveform. In this example, the entrainment phase was set to 5.45 rad. (b) Histogram of the phase difference between the voice envelope and the stimulation waveform recorded in this experiment (results from single measurement). *:5.45 rad.

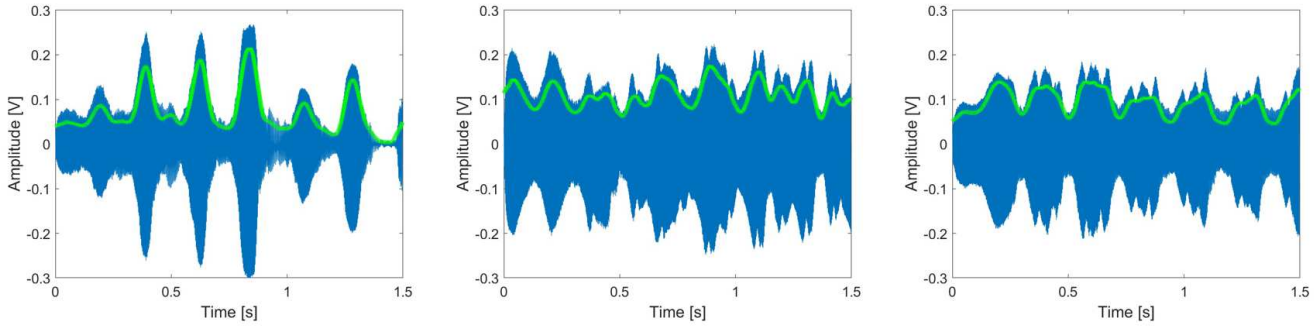


Fig. 4. tACS-related changes in patient voice characteristics. Figures show patient voice and its envelope waveforms (blue and green lines, respectively). Results from patient 1 is presented. (a) Before treatment. (b) After treatment with in-phase tACS. (c) After treatment with out-phase tACS.

E. Voice Signal Analysis

The initial 1.5 seconds of the voice waveform and its envelope were extracted to qualitatively observe the changes in voice amplitude. This duration was chosen based on the longest continuous voice period among all patients. To evaluate tremor changes, the FTrI and ATrI were calculated using the tremor2.06 algorithm of the voice processing program Prrat [20-21]. The definition of FTrI and ATrI is shown in Equation (1). The peaks and troughs of the voice envelope waveform were identified and set as maxima and minima, respectively, with the number of maxima and minima represented as m and n, respectively.

$$(\mathrm{F,A})\mathrm{TrI} = \left(\frac{\sum_{i=0}^{m}|max_i|}{m} + \frac{\sum_{j=0}^{n}|min_j|}{n}\right) \div 2 \qquad (1)$$

ATrI represents the vibration amplitude of the power value at the tremor frequency, indicating the intensity of the amplitude tremor. Thus, a higher ATrI indicates a greater amplitude tremor. FTrI indicates the intensity of changes in the tremor frequency, with a higher FTrI indicating greater variability in tremor frequency.

For patients from whom data could be obtained, subjective ratings of ease of voice production on a scale of 0-100 were obtained both before the intervention on arrival and after stimulation. On this scale, 0 represents the most difficult voice production, while 100 represents no difficulty at all.

III. Results

A. tACS-derived Changes in Voice Envelope Waveforms

Figure 4 shows the voice envelope of a representative patient (patient 1) measured on arrival and after in-phase and out-phase tACS treatment. In the pre-treatment state (Fig. 4a), a vocal tremor is evident, as indicated by the strong oscillations in the envelope waveform. In contrast, the post-treatment state shows a significant reduction in the amplitude of these oscillations (Fig. 4bc), indicating an improvement in the vocal tremor. Both in-phase and out-phase stimulation resulted in a comparative improvement in vocal tremor for this patient. While the responsiveness varied among patients, the amplitude variations of the voice envelope waveform were generally suppressed, and the envelope of the voice waveform became flatter after both in-phase and out-phase entrainment stimulation.

B. FTrI and ATrI

Figures 5 and 6 show the results of the voice signal analysis. The FTrI decreased in two out of three patients with both in-phase and anti-phase stimulation compared to the pre-stimulation state, although no significant changes were obtained across all patients. The ATrI decreased significantly after the out-phase stimulation compared to the pre-stimulation state (paired t-test, $p < 0.05$).

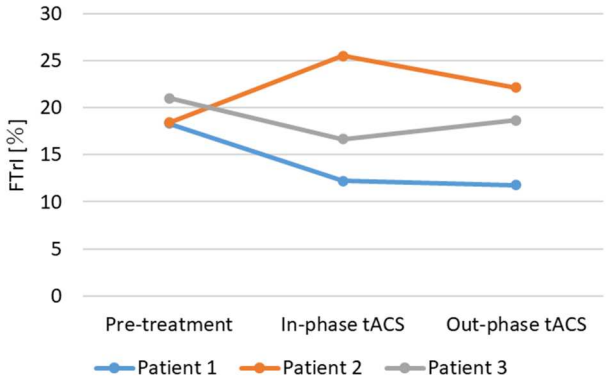


Fig. 5. FTrI changes before treatment, after in-phase stimulation, and after out-phase stimulation treatments.

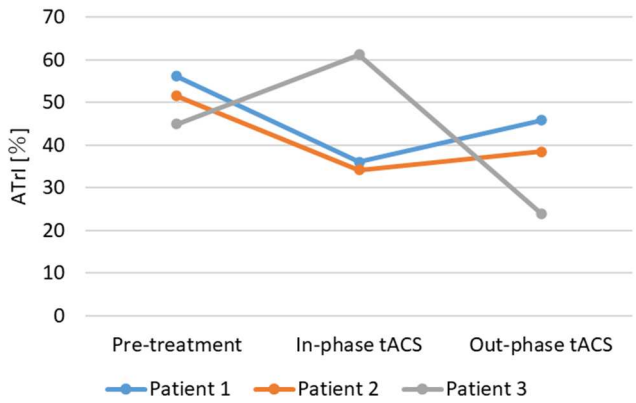


Fig. 6. ATrI changes before treatment, after in-phase stimulation, and after out-phase stimulation treatments.

C. *Questionnaire*

Subjective ratings for ease of voice production were obtained from patients 1 and 3 (Table 3). Both patients showed improvement in voice production after tACS treatment.

Table 3. Ease of voice production (scale of 100)

Patient	Before treatment	After in-phase stimulation	After out-phase stimulation
Patient1	25	60	Not available
Patient3	30	40	35-45

IV. Discussion

The current results demonstrated he potential of tACS as a treatment option for vocal tremor patients. We investigated the effect of both in-phase and out-phase stimulations to explore the effective treatment protocol. Although previous tACS research on patients with hand tremors showed clinical effectiveness of in-phase stimulation [12,14], statistically significant improvements in the vocal tremor amplitude were found in the out-phase stimulation. This difference may be related to the relationship between the site of the tremor and the cerebellum. For upper limb tremors, stimulation is applied to the cerebellum on the same side as the affected limb, but for vocal tremors, the relationship with the left and right cerebellum is not yet clear. Furthermore, the decrease in ATrI with in-phase stimulation is larger in patients 1 and 2, besides patient 3, indicating that the optimal phase might vary depending on the pathological condition of the patients. A similar tendency was observed in the FTrI, two out of three patients showed improved frequency tremor, but the other one showed its exaggeration regardless of the phase of stimulation. Further studies should measure both cerebellums and increase the number of patients to evaluate the therapeutic effect.

To date, the etiology of essential tremor is not clearly understood. However, a number of studies have shown that essential tremor is closely related to the central motor network that controls voluntary movements and the associated neural network, with the olivocerebellar-thalamic pathway playing an important role within this network [22]. Schnitzler et al. reported that in patients with essential upper limb tremor, upper limb tremor activity recorded by electromyography showed coherence with the electrophysiological activity of the primary motor cortex, possibly involving associated neural pathways of the primary motor cortex, thalamus, premotor cortex, cerebellum, and brainstem [23]. Therefore, the reason for the changes in tremor amplitude and frequency due to the application of tACS to the cerebellum may be due to the disruption of the coherence between cerebellar and skeletal muscle activities via the central motor network caused by the electrical stimulation.

V. Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number JP23H00459 and SECOM Science and Technology Foundation.

References

[1] K. P. Bhatia *et al.*, “Consensus Statement on the classification of tremors. from the task force on tremor of the International Parkinson and Movement Disorder Society,” *Mov. Disord.*, vol. 33, no. 1, pp. 75–87, 2018.

[2] W. C. Koller, K. Busenbark, and K. Miner, “The relationship of essential tremor to other movement disorders: report on 678 patients. Essential Tremor Study Group,” *Ann. Neurol.*, vol. 35, no. 6, pp. 717–723, 1994.

[3] S. Sharma and S. Pandey, “Treatment of essential tremor: current status,” *Postgrad. Med. J.*, vol. 96, no. 1132, pp. 84–93, 2020.

[4] B. Kundu, L. Schrock, T. Davis, and P. A. House, “Thalamic deep brain stimulation for essential tremor also reduces voice tremor,” *Neuromodulation*, vol. 21, no. 8, pp. 748–754, 2018.

[5] D. P. Newland, D. Novakovic, and A. L. Richards, “Voice tremor and botulinum neurotoxin therapy: A contemporary review,” *Toxins (Basel)*, vol. 14, no. 11, p. 773, 2022.

[6] A. V. Tavakoli and K. Yun, “Transcranial alternating current stimulation (tACS) mechanisms and protocols,” *Front. Cell. Neurosci.*, vol. 11, p. 214, 2017.

[7] S. Koganemaru *et al.*, “Cerebellar transcranial alternating current stimulation modulates human gait rhythm,” *Neurosci. Res.*, vol. 156, pp. 265–270, 2020.

[8] S. Koganemaru *et al.*, “Gait-synchronized rhythmic brain stimulation improves poststroke gait disturbance: A pilot study: A pilot study,” *Stroke*, vol. 50, no. 11, pp. 3205–3212, 2019.

[9] I. Nojima *et al.*, “Gait-combined closed-loop brain stimulation can improve walking dynamics in Parkinsonian gait disturbances: a randomised-control trial,” *J. Neurol. Neurosurg. Psychiatry*, vol. 94, no. 11, pp. 938–944, 2023.

[10] A. Shima *et al.*, "Case report: Backward gait training combined with gait-synchronized cerebellar transcranial alternating current stimulation in progressive supranuclear palsy," *Front. Hum. Neurosci.*, vol. 17, p. 1082555, 2023.

[11] J. Riddle and F. Frohlich, "Targeting neural oscillations with transcranial alternating current stimulation," *Brain Res.*, vol. 1765, no. 147491, p. 147491, 2021.

[12] S. R. Schreglmann *et al.*, "Non-invasive suppression of essential tremor via phase-locked disruption of its temporal coherence," *Nat. Commun.*, vol. 12, no. 1, p. 363, 2021.

[13] J.-S. Brittain, P. Probert-Smith, T. Z. Aziz, and P. Brown, "Tremor suppression by rhythmic transcranial current stimulation," *Curr. Biol.*, vol. 23, no. 5, pp. 436–440, 2013.

[14] F. Nieuwhof, I. Toni, A. W. G. Buijink, A.-F. van Rootselaar, B. P. C. van de Warrenburg, and R. C. Helmich, "Phase-locked transcranial electrical brain stimulation for tremor suppression in dystonic tremor syndromes," *Clin. Neurophysiol.*, vol. 140, pp. 239–250, 2022.

[15] E. A. Abraham and A. Geetha, "Acoustical and perceptual analysis of voice in individuals with Parkinson's disease," *Indian J. Otolaryngol. Head Neck Surg.*, vol. 75, no. 2, pp. 427–432, 2023.

[16] H. Silek and M. Dogan, "Voice analysis in patients with Essential Tremor," *J. Voice*, 2023.

[17] S. J. Groiss and Y. Ugawa, "Cerebellar stimulation in ataxia," *Cerebellum*, vol. 11, no. 2, pp. 440–442, 2012.

[18] Y. Shirota, R. Hanajima, T. Shimizu, Y. Terao, S. Tsuji, and Y. Ugawa, "Quantitative evaluation of cerebellar function in multiple system atrophy with transcranial magnetic stimulation," *Cerebellum*, vol. 21, no. 2, pp. 219–224, 2022.

[19] Y. Ugawa, Y. Uesaka, Y. Terao, R. Hanajima, and I. Kanazawa, "Magnetic stimulation over the cerebellum in humans: Magnetic Cerebellar Stimulation," *Ann. Neurol.*, vol. 37, no. 6, pp. 703–713, 1995.

[20] M. Brückl, E. Ibragimova, and S. Bögelein, "Acoustic Tremor Measurement: Comparing Two Systems," in *Proceedings of the International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA 2017)*, Firenze, Italy, 2017, pp. 13–15.

[21] M. Brückl and A. Ghio, "François Viallet, Measurement of Tremor in the Voices of Speakers with Parkinson's Disease," *Procedia Computer Science*, vol. 128, 2018.

[22] J. Raethjen and G. Deuschl, "The oscillating central network of Essential tremor," *Clin. Neurophysiol.*, vol. 123, no. 1, pp. 61–64, 2012.

[23] A. Schnitzler, C. Münks, M. Butz, L. Timmermann, and J. Gross, "Synchronized brain network associated with essential tremor as revealed by magnetoencephalography," *Mov. Disord.*, vol. 24, no. 11, pp. 1629–1635, 2009.

Complex CNN incorporating Hilbert transform for steady-state visual evoked potential BCI

Rintaro Takata* and Yoshikazu Washizawa†
* University of Electro-Communications, Tokyo
E-mail: t2331089@edu.cc.uec.ac.jp Tel: +81-0424435287
† University of Electro-Communications, Tokyo
E-mail: washizawa@uec.ac.jp Tel: +81-0424435976

***Abstract*—Brain computer interface (BCI) is a system that converts brain activity information into commands, allowing people to communicate with the outside world without the need for physical activity. Steady-state visual evoked potential (SSVEP) which is a kind of event related potentials (ERPs) is used for BCI. SSVEP shows higher performance and has been improved in various ways to improve the information transfer rate and expand the number of commands, such as phase-modulated SSVEP-BCI. To further improve the performance of SSVEP-BCI, the frequency and phase of SSVEP need to be detected in a short time.**

In this study, we apply the Hilbert transform to the complex-valued convolutional neural network (CVCNN) for phase-modulated SSVEP-BCI. The Hilbert transform and accompanying analytic signal enable us to determine the instantaneous frequency and phase. The proposed CVCNN incorporating the Hilbert transform not only detects instantaneous frequency and phase information from electroencephalogram (EEG), but also learns convolutoinal filter automatically to detect SSVEP frequency, phase, and their harmonics components. As the result of evaluation experiments on open SSVEP datasets, the proposed method showed higher accuracy than conventional methods.

I. Introduction

Brain-computer interface (BCI) is a system to operate computers or communicate with the outside world without the need for physical activity [1]. BCI is expected to be used as a communication tool for people who cannot move their muscles freely, such as patients with amyotrophic lateral sclerosis (ALS). Futhermore, BCI is applied for healthy people in areas such as gaming, entertainment, security, and education [2].

Electroencephalogram (EEG) is a scalp-based measurement of changes in electrical potentials due to cranial nerve activity. Event related potentials (ERPs) occurs in association with perceptual and cognitive processing. Visual evoked potential (VEP) is a kind of ERPs, that is evoked by visual stimuli. Steady-state visual evoked potential (SSVEP) is a kind of VEP, that is elicited by periodic flashes. SSVEP has higher signal-to-noise ratio than the other ERPs and less individual differences, thus, is applied to BCI. SSVEP-BCI has shown high accuracy [3], [4].

SSVEP is easily evoked around 10 Hz, and there is a limit to the frequency used for visual stimulation. The phase-modulated SSVEP was devised to increase the number of commands. The phase-modulated SSVEP attempts to modulate not only by frequency but also phase information. Extended canonical correlation analysis (CCA) method was proposed to classify the phase-modulated SSVEP, and showed high information transfer rate (ITR) of 117.75 bits/min and 105 bits/min [5], [6].

In recent years, deep learning has attracted much attention due to its success in pattern recognition such as image classification, object detection, and natural language processing. These are due to the easy availability of large datasets and increased computing power. Convolutional neural network (CNN) is one particularly successful model, which extracts features by convolution to provide position invariance to the model [7]–[9].

A complex-valued neural network (CVNN) is a neural network model that extends weights, inputs, outputs, activation functions, etc. to complex numbers. CVNN has been performed better than real-valued neural network (RVNN) in various fields, because complex numbers are suitable for representing amplitude and phase to process in the frequency domain via Fourier transforms, etc. [10]–[13].

This study proposes a complex valued convolutional neural network (CVCNN) model that incorporates the Hilbert transform. By the Hilbert transform, the proposed neural network model generates an complex analytic signal within the neural network model to obtain instantaneous amplitude and phase. We used this model for phase-modulated SSVEP classification. By extracting instantaneous amplitude and phase in the neural network, phase-modulated SSVEP is detected from short time frame accurately. We compare the performance of the proposed model with conventional methods and typical CNNs.

II. Method

A. Hilbert transform and analytic signal

The sinusoidal signal $x(t)$, whose amplitude and frequency vary with time, is given by

$$\begin{aligned} x(t) &= A(t)\cos(\omega(t)t+\theta) \\ &= A(t)\cos(\phi(t)), \quad \phi(t)=\omega(t)t+\theta, \end{aligned} \tag{1}$$

where $A(t)$ is the instantaneous amplitude, $\omega(t)$ is the instantaneous angular frequency, and θ is the initial phase. In order to determine the instantaneous amplitude $A(t)$ and frequency

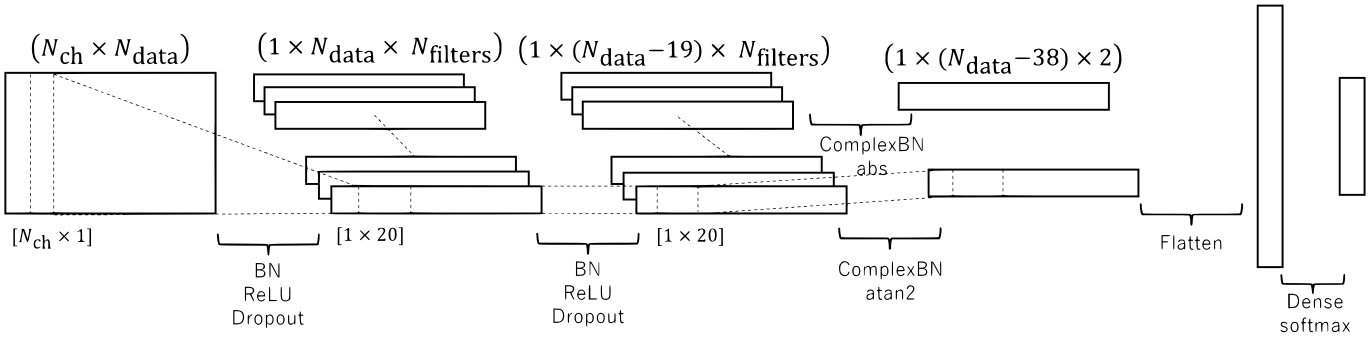


Fig. 1: Structure of the proposed network model

$\omega(t)$, the sinusoidal signal $x(t)$ is extended to complex analytic signal by the Hilbert transform.

The Hilbert transform shifts the phase of the sinusoidal signal $x(t)$ by 90 degrees. The Hilbert transform $y(t)$ of $x(t)$ is defined by the convolution of $x(t)$ and $1/\pi t$,

$$y(t) = x(t) * \frac{1}{\pi t} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau. \tag{2}$$

The analytic signal $z(t)$ is generated by putting the signal $x(t)$ in the real part and the Hilbert transform $y(t)$ in the imaginary part,

$$\begin{aligned} z(t) &= x(t) + jy(t) \\ &= A(t)\cos(\phi(t)) + jA(t)\sin(\phi(t)) \\ &= A(t)e^{j\phi(t)}. \end{aligned} \tag{3}$$

The instantaneous amplitude $A(t)$ and instantaneous phase $\phi(t)$ are respectively obtained as follows,

$$A(t) = \sqrt{x(t)^2 + y(t)^2} \tag{4}$$

$$\phi(t) = \mathrm{atan2}(y(t), x(t)), \tag{5}$$

where atan2 function calculates the declination angle in the complex plane. In general, the instantaneous angular frequency $\omega(t)$ is obtained by the derivative of $\phi(t)$.

The Hilbert transform filter is an FIR filter that performs the Hilbert transform (2). Its coefficients is generated by Parks-McClellan algorithm. The output signal is delayed in proportion to the size of the FIR filter. When applying an M-dimensional Hilbert transform filter, the signal is delayed by $\frac{M}{2}$ sampling points.

B. Proposed neural network

The proposed neural network achieves instantaneous frequency analysis through the Hilbert transform in the convolutional layer. The complex analytic signal is generated by the convolutional layer, and the latter complex valued activation function extracts frequency and amplitude information.

The activation function of the complex valued layer is not as simple as the real valued network, because it is a complex function. In the conventional CVNN, $\mathrm{ReLU}(x(t)) + i\mathrm{ReLU}(y(t))$ or $\tanh(|z(t)|)\exp(j\arg z(t))$ is used [14]–[16]. In this research, we use the absolute value function $|z(t)| = \sqrt{x(t)^2 + y(t)^2}$ and the declination function $\arg z(t) = \mathrm{atan2}(y(t), x(t))$ to obtain instantaneous amplitude and phase.

The structure of the proposed network model is shown in Fig. 1. BN and ComplexBN refer to batch normalization and complex batch normalization, respectively. This network model consists of three convolutional layers and one fully connected layer. The size of input matrix is the number of samplings N_{data} multiplied by the number of channels N_{ch}. The size of filters is N_{filters}.

The first convolutional layer reduces dimensionality with respect to the channel direction. After the convolution, batch normalization, activation function ReLU, and the dropout are applied.

In the second convolutional layer, the convolution is performed with respect to the time direction. This convolutional layer generates narrowband signals corresponding to the stimulus frequencies and harmonics of SSVEP.

The third convolution layer generates analytic signal. After the convolution, the absolute value function and the atan2 function are applied as the activation functions. Then the values are flattened, and discriminated by the fully connected layer. We set the initial value of the convolution filter in the third layer to the Hilbert transform filter. We considered two cases where the filter weights are updated and fixed during the learning process. We refer to them as "proposed train" and "proposed untrain" respectively.

III. Experiment

A. Datasets

We used two open datasets. The first dataset is phase-modulated SSVEP data of 35 subjects with 40 classes [5]. We used the same nine channels and eight experimental subjects as [5]. The sampling frequency was 250 Hz. The visual stimuli consist of 40 frequencies ranging from 8 to 15.8 Hz at 0.2 Hz intervals. The initial phase of each stimulus is also shifted by 0.5π. For each subject and each stimulus, six trials of data were collected. Similar to the reference study, the data were divided into training and test sets using six-fold cross-validation [5].

The second dataset is phase-modulated SSVEP data from ten subjects in twelve targets [6]. The following method is

used to create a series of visual stimuli at all frequencies and initial phases regardless of the refresh rate.

Let f be the frequency of the steady blinking stimulus, ϕ be the initial phase and N_f be the refresh rate. The visual stimulus is generated as follows,

$$s(f, \phi, n) = \text{square}[2\pi f(n/N_f) + \phi], \tag{6}$$

where square[] generates a 50% duty cycle square wave with levels 0 and 1. The stimulus sequence $s_i(n)$ of the i-th targets is defined as

$$s_i(n) = s(f_0 + (i-1)\Delta f, \phi_0 + (i-1)\Delta\phi, n), \\ i = 1, 2, \ldots, N_{\text{targets}}, \tag{7}$$

where f_0 is the minimum frequency used for stimulation, ϕ_0 is the initial phase at frequency f_0, Δf and $\Delta\phi$ are the frequency and phase intervals respectively, and $N_{\text{targets}} = 12$ is the number of frequencies used for stimulation [6]. The twelve visual stimuli were presented at a refresh rate of 60 Hz, and the frequency and phase values were $(f_0 = 9.25\text{Hz}, \Delta f = 0.5\text{Hz})$ and $(\phi_0 = 0, \Delta\phi = 0.5\pi)$. Sampling frequency was 256 Hz. For each subject and each stimulus, fifteen trials of data were collected. Similar to the reference study, the data were divided into training and test sets using fifteen-fold cross-validation [6].

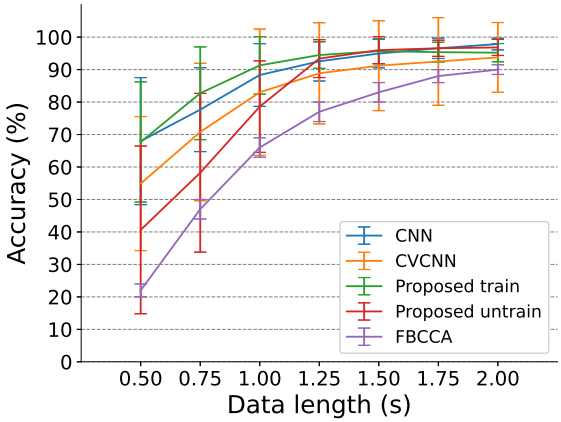
B. Methods

We compared three network models: real-valued convolutional neural network (RVCNN), CVCNN, and the proposed method. In the following, we will refer to RVCNN as CNN. For CNN and CVCNN, we used models similar to the structure of the proposed method. The both models have two convolutional layers and one fully connected layer. The number of filters and filter size were set to the same values as for the proposed method. For CNN, ReLU was used as the activation function after the convolutional layer, and the softmax function was placed after the fully connected layer. The activation functions of CVCNN are as follows; $\tanh(|z|)\exp(j\arg z)$ is used after the first convolutional layer, $\sqrt{x(t)^2 + y(t)^2}$ is used after the complex fully connected layer. In CVCNN, complex batch normalization was applied for the batch normalization part [15].

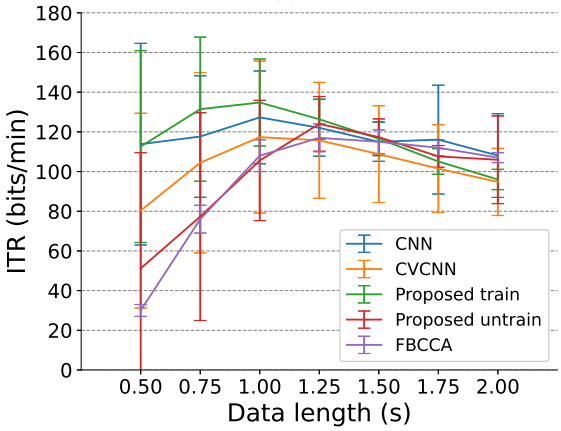
Grid searches were executed for the parameter selection of L2 regularization, dropout rate, learning rate, and number of epochs in $\{0.0001, 0.001, 0.01\}$, $\{0.0, 0.2, 0.4\}$, $\{0.0001, 0.001, 0.01\}$, $\{300, 400, 500\}$, respectively. Adam was used for optimizer. The loss function is mean squared error. All neural network models and experiments were implemented using Python 3.11.3, Numpy 1.24.3, Keras 2.13.1, and Tensorflow 2.13.0. We used remez function of the Python library Scipy for the Hilbert transform filter.

We conducted a comparison based on classification accuracy and information transfer rate (ITR). ITR represents the amount of information that can be transmitted per unit time. Let N be the number of targets, P be the classification accuracy, and T be the data length. ITR was calculated as follows,

$$\frac{1}{T}(\log_2 N + P\log_2 P + (1-P)\log_2(\frac{1-P}{N-1})) \text{ [bits/min]}. \tag{8}$$



(a)

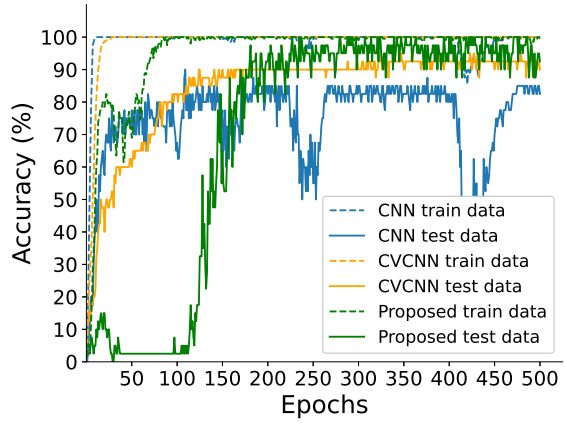


(b)

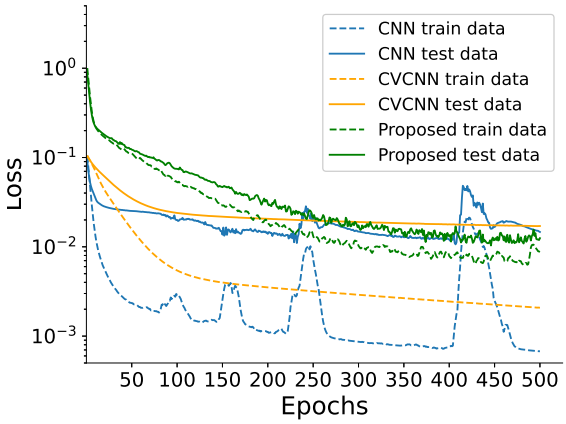
Fig. 2: (a) Classification accuracy for the first dataset with different data lengths from 0.25 s to 2.0 s with a step of 0.25 s. (b) Simulated ITRs across subject using different data lengths. The FBCCA method was quoted from [5]. The error bars indecate standard errors.

IV. Result

We compared classification accuracy and ITR. Fig. 2 shows the result of the first dataset. The proposed method showed the highest classification accuracy. The ITR of the proposed method also showed the highest value. It was demonstrated that the proposed method achieved higher accuracy compared to the FBCCA method from the reference study. Table. I shows the classification accuracy for each subject of the first dataset with 1s data length. In S1, S2, and S8, the proposed method showed higher accuracy than the conventional methods. Fig. 3 shows the learning curve of neural network models at one trial of S1. The proposed method shows that the difference in loss between training and testing is smaller than the other methods.

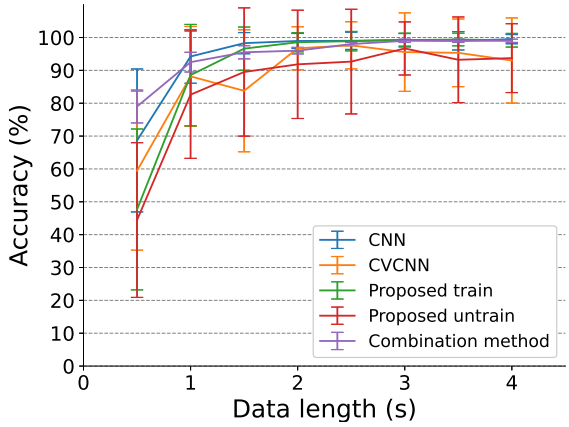


(a)

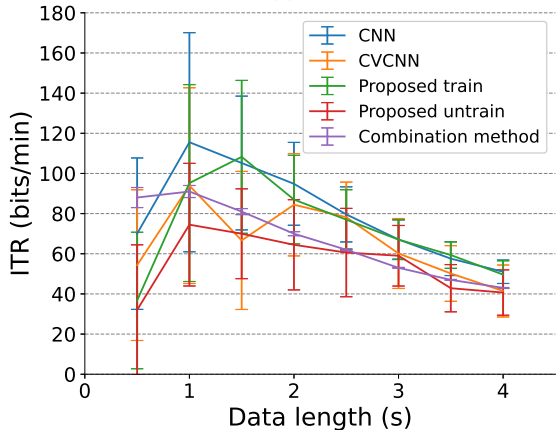


(b)

Fig. 3: (a) Learning curve accuracy and (b) learning curve loss for the first dataset with different data lengths from 0.25 s to 2.0 s with a step of 0.25 s. This is the result of one trial of subject1. "Proposed" model is "Proposed train" model.



(a)



(b)

Fig. 4: (a) Classification accuracy for the second dataset with different data lengths from 0.5 s to 4.0 s with a step of 0.5 s. (b) Simulated ITRs across subject using different data lengths. The combination method was quoted from [6]. The error bars indecate standard errors.

TABLE I: Classification accuracy (%) of the first dataset

	Proposed train	Proposed untrain	CNN	CVCNN
S1	95.00	76.67	87.50	85.83
S2	97.92	86.25	87.92	87.08
S3	99.17	95.00	98.75	98.33
S4	97.08	93.33	95.83	96.25
S5	95.00	87.50	96.67	88.75
S6	89.58	69.17	91.67	93.33
S7	73.75	57.92	75.83	37.92
S8	82.92	62.92	72.50	76.67
Avg.±STD	91.30±8.85	78.59±14.09	88.33±9.66	83.02±19.45

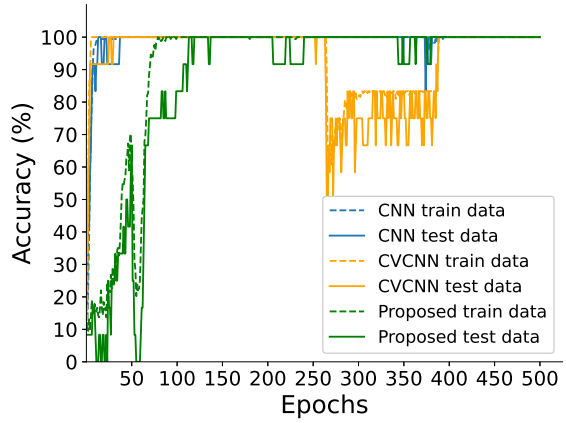
TABLE II: Classification accuracy (%) of the second dataset

	Proposed train	Proposed untrain	CNN	CVCNN
S1	79.44	65.00	95.56	92.78
S2	51.67	42.22	76.11	55.56
S3	89.44	76.67	100	100
S4	99.44	97.78	99.44	97.78
S5	100	98.33	100	96.67
S6	98.88	97.78	99.44	93.89
S7	99.44	95.56	98.89	97.22
S8	97.22	94.44	98.33	91.11
S9	92.78	92.78	95.56	93.33
S10	77.22	65.56	82.78	64.44
Avg.±STD	88.56±15.42	82.61±19.40	94.22±8.14	88.17±15.21

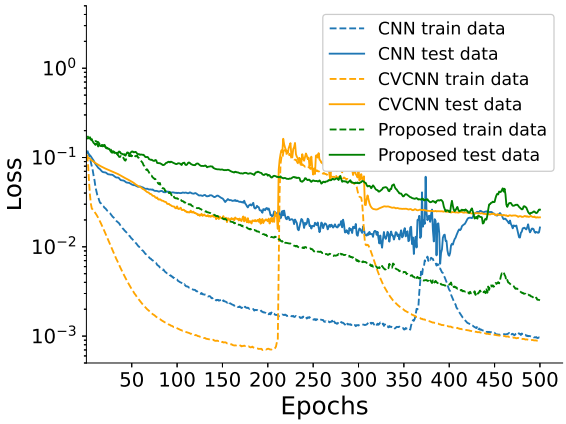
Fig. 4 shows the result of the second dataset. CNN has the highest classification accuracy. The proposed method was the second best classification accuracy and ITR. It was demonstrated that the proposed method achieved higher accuracy compared to the Combination method from the reference study [6]. Table. II shows the classification accuracy for each subject of the second dataset with 1s data length. In S7, the proposed method showed the highest accuracy than the conventional methods. Fig. 5 shows the learning curve of neural network models at one trial of S7. The proposed method shows that the difference in loss between training and testing is small and that the learning is stable.

V. Conclusion

The proposed method improved classification accuracy compared to the conventional method in the first dataset. ITR



(a)



(b)

Fig. 5: (a) Learning curve accuracy and (b) learning curve loss for the second dataset with different data lengths from 0.5 s to 4.0 s with a step of 0.5 s. This is the result of one trial of subject1. ”Proposed” model is ”Proposed train” model.

of the proposed method achieved 134.8 bits/min on average. The conventional method obtaines frequency information of the signal by the convolutional layer with respect to the time direction, but these models cannot obtain amplitude and phase information. The proposed method utilizes the initial phase information, thus that the proposed model showed higher classification accuracy for the dataset with a large number of frequencies used. In the proposed method, when the weights of the Hilbert transform filter were learned, an improvement in accuracy was observed. This is thought to be due to the optimization of the filter weights tailored to the data.

In the second dataset, there was one subject for whom the proposed method was the most accurate. The proposed method was more stable in learning the classification accuracy and loss than the conventional method. This indicates that overlearning is unlikely to occur and generalization performance is high. Because of the small number of frequencies used, we believe that the conventional method, which only performs frequency analysis, was sufficiently accurate.

The phase-modulated SSVEP data used in the experiment does not increase the number of stimuli based entirely on phase information while incorporating phase modulation. Future works will include considering phase-modulated SSVEP-BCI using stimuli with different initial phases at the same frequency. The proposed model may be even more accurate than the conventional method.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, “Brain-computer interfaces for communication and control,” *In Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1–2, 2007.

[2] V. Patelia and M. S. Patel, “Brain computer interface: Appications and P300 speller overview,” *In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–5, 2019.

[3] Z. Yue, X. Q. Shane, W. He, and Z. Zhiqiang, “Data analytics in steady-state visual evoked potential-based brain-computer interface: A review,” *IEEE Sensors Journal*, vol. 21, 2 2020.

[4] R. Aya and et al., “Braincomputer interface spellers: A review,” *Brain Sciences*, vol. 8, no. 4, p. 57, 2018.

[5] W. Yijun, C. Xiaogang, G. Xiaorong, and G. Shangkai, “A benchmark dataset for SSVEP-based brain-computer interfaces,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, 10 2017.

[6] M. Nakanishi, Y. Wang, Y. T. Wang, and T. P. Jung, “A comparison study of canonical correlation analysis based methods for detecting steady state visual evoked potentials,” *PLoS ONE*, vol. 10, no. 10, pp. 1–18, 2015.

[7] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” *International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015.

[8] L. Sergey and S. Christian, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *International Conference on Machine Learning*, vol. 37, pp. 448–456, 2015.

[9] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, “Deep residual learning for image recognition,” *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[10] A. Igor and K. Zain, “Analysis of EEG using multilayer neural network with multi-valued neurons,” *In 2018 IEEE Second International Conference on Data Stream Mining Processing*, pp. 392–396, 2018.

[11] P. Musa, S. Baha, and D. Dursun, “A novel method for automated diagnosis of epilepsy using complex-valued classifiers,” *IEEE Journal of Biomedical and Health Informatics*, pp. 108–118, 2016.

[12] Z. Junming and W. Yan, "A new method for automatic sleep stage classification," vol. 11, pp. 1097–1110, 5 2017.

[13] Z. Zhimian, W. Haipeng, X. Feng, and J. Ya-Qiu, "Complex-valued convolutional neural network and its application in polarimetric SAR image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, pp. 7177–7188, 12 2017.

[14] A. Ikeda and Y. Washizawa, "Spontaneous EEG classification using complex valued neural network," *International Conference on Neural Information Processing*, vol. 1142, pp. 495–503, 2019.

[15] T. Chiheb and et al., "Deep complex networks," *International Conference on Learning Representations*, 2018.

[16] K. Taehwan and A. Tulay, "Approximation by fully complex multilayer perceptrons," *Neural Computation*, vol. 15, pp. 1641–1666, 7 2003.

Electroencephalogram-Based Effective Features for Sustained Attention Assessment in Conversation

Masaya Togashi*, Ingon Chanpornpakdi†, and Toshihisa Tanaka†
* Department of Industrial Technology and Innovation,
Graduate School of Engineering, Tokyo University of Agriculture and Technology, Japan
E-mail: togashi22@sip.tuat.ac.jp
† Department of Electrical Engineering and Computer Science,
Graduate School of Engineering, Tokyo University of Agriculture and Technology, Japan
E-mail: mint@sip.tuat.ac.jp, tanakat@cc.tuat.ac.jp

***Abstract*—Sustained attention, or concentration, is the ability to focus on a task for a long time. Evaluation methods include subjective assessments using questionnaires and objective measures such as electroencephalogram (EEG). The EEG method offers promising concentration level insights, but it is a challenge to fully leverage the multi-channel nature of EEG at the same time. In this study, we conducted a button press task using the go/no-go paradigm to measure concentration during the simulated conversational scene while addressing the EEG multi-channel complexity. We measured response times and recorded EEG, simultaneously. The extracted features such as frequency band characteristics and information-theoretical features like entropy were used to evaluate the concentration level. We employed group Lasso and cross-correlation analysis to expose the explainable EEG feature for sustained attention relevant to the variance time course (VTC). As a result, we found that EEG power ratios were strongly related to concentration. Combining these several power ratios could open up the possibility of quantifying the concentration.**

***Index Terms*—sustained attention, electroencephalogram, sparse model, power, entropy**

I. Introduction

Sustained attention, also known as concentration, is the cognitive ability to maintain focus on a specific task over an extended period. In the workplace, this ability engages in work productivity and achievement. Besides the professional field, concentration enhances the performance of various activities from academic studies to hobbies. This ability is essential for our daily lives and closely related to life success. Consequently, demand is growing for precise and reliable methods to quantify concentration.

Both subjective reports and objective measures have been used to assess one's concentration levels. Subjective evaluations often rely on self-report questionnaires designed for various purposes. The cognitive failures questionnaire (CFQ) is a widely trusted subjective measure for assessing concentration [1], [2]. While questionnaire-based subjective evaluations are simple and less burdensome for individuals, they can lack consistency due to variations in interpretation and the possibility of false responses. Moreover, the responses can be biased by personal perception and social desirability.

In addition to subjective measures, behavioral experiments, and physiological signal-based indicators have been proposed. A typical example of behavioral experiments involves tasks where participants respond to image stimuli by pressing buttons and measuring response times and accuracy. The go/no-go task [3] is particularly notable, being used in research and clinical settings across a wide range of areas, from children to adults, including the assessment of higher brain dysfunction in brain injuries and epilepsy, and evaluating the effects of drug therapies [4]. During the task, the participant responds with a button press when the go stimuli are presented and refrains to press the button when other (no-go) stimuli are presented. The continuous performance test [5] and the sustained attention to response task [6] are examples of concentration measurements using the go/no-go paradigm.

Electroencephalogram (EEG) is a prominent physiological signal-based method for assessing concentration. It is closely linked to concentration and may capture potential neural activities of focus that are not apparent in behavior [7]. Furthermore, EEG has the potential to monitor temporal fluctuations in concentration due to its high temporal resolution [8]. Studies on EEG indicated that multiple oscillatory components of EEG are correlated with various states of concentration. During go/no-go tasks, theta and alpha band power in EEG at midline electrodes (Fz, Cz, Pz) are related to concentration levels, as compared with eyes-closed and eyes-open resting state[9]. Coelli et al. demonstrated a negative correlation between the beta-to-alpha power ratio and mean response time across electrode groups, suggesting that higher concentration is associated with greater beta power and lower alpha power [10]. Besides the spectral features, information-theoretic measures such as entropy have proven effective in characterizing EEG signals [11]. Despite advancements in the understanding of the relationship between concentration and EEG, most studies have focused on single-channel EEG analysis, neglecting the inherently multi-channel nature of EEG data. This gap in the literature limits our understanding of how different brain regions interact during sustained attention tasks. Therefore, this study aims to identify features that can represent concentration while accounting for the multi-channel characteristics of EEG.

These previous findings suggests that frequency power and information-theoretic (entropy) features [12] can represent concentration levels, we hypothesize that a small number of features and brain regions can effectively explain concentration. Our novel approach combines multi-channel EEG analysis with advanced machine learning techniques, specifically group Lasso [13]. This method aims to identify the most relevant EEG features across different brain regions that can explain behavioral indicators of concentration. While there are several feature selection techniques [14], we chose group Lasso because it can handle the two-dimensional nature (feature $\times$ location) of multi-channel EEG data, where we set up groups for each feature across all electrodes. Moreover, we designed a go/no-go task to collect our data on behavioral indicators and EEG. Frequency and entropy features were extracted from the recorded EEG, and their importance ($\ell_{2,1}$-norm) was evaluated using group Lasso. Additionally, we identified brain regions associated with concentration based on the spatial distribution of features selected by group Lasso.

II. Methods

A. Participants

Sixteen volunteers (mean age 24.0 ± 4.92 years, two females) participated in this study. Before the experiment, informed consent was obtained from all participants based on the approval of the Tokyo University of Agriculture and Technology Research Ethics Committee (231104-0555). None of the student participants were encouraged by their professors to participate, nor did they receive any academic credits for their involvement.

B. Experimental Task

The go/no-go task was designed using two facial stimuli as shown in Figure 1, considering the focus on concentration during human conversations. These faces were extracted from the publicly available racially diverse affective expression (RADIATE) face stimulus set [15]. One face had an open mouth (visible teeth), while the other had a closed mouth (no visible teeth). Participants were instructed to respond by clicking the left mouse button as quickly as possible to the go stimulus (the face with the open mouth) and to withhold response to the no-go stimulus (the face with the closed mouth). Stimuli were presented for 3000 milliseconds with an inter-stimulus interval of 2000 milliseconds. A total of 360 stimuli were presented throughout the task, including 288 go stimuli, which constituted 80% of the total. The task was created using Tobii Pro Lab version 1.232 (Tobii AB, Stockholm, Sweden).

C. Data Acquisition

EEG was recorded using 30 electrodes placed according to the international 10-10 system, as shown in Figure 2. The ground electrode was placed at AFz and the reference was the average of A1 and A2 electrodes placed on the earlobes. To monitor eye movements, an electrooculogram (EOG) was recorded using four electrodes placed as shown in Figure 3. These electrodes were connected to an AC amplifier, Polymate AP5148 (Miyuki Giken, Tokyo, Japan), and recorded using the



(a) Go stimulus with mouth opened.



(b) No-go stimulus with mouth closed.

Fig. 1: Two facial stimuli from the RADIATE Face Stimulus Set were used in the experiment [15].

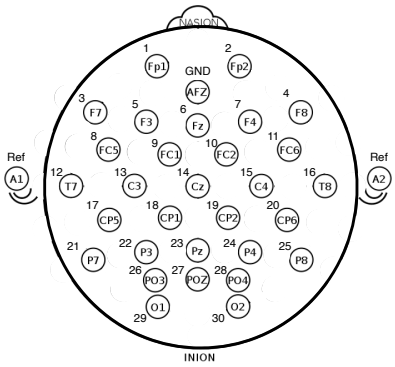


Fig. 2: Thirty EEG electrode placement based on the international 10-10 system.

accompanying software, AP Monitor. The sampling frequency was set to 1,000 Hz. In addition to EEG, we recorded the behavioral responses including button press (mouse clicking) and the response time.

D. Analysis Methods

The analysis pipeline consisted of several steps: defining a concentration index based on behavioral responses, preprocessing the EEG data, extracting relevant features, selecting features using sparse modeling, and performing cross-correlation analysis. Each stage is described in detail below.

1) Definition of Concentration Index Based on Response: Response time was defined as the duration from stimulus onset to the participant's button press, which was calculated for each trial. For trials without a response, linear interpolation using adjacent response times was applied. Based on response times, the variance time course (VTC) [16], a metric derived from the temporal dynamic of the response time, was used as an index of concentration. The VTC was computed by the moving average of standardized response times across all trials A Gaussian smoothing kernel with a full width at half maximum (FWHM) of 4 trials was used for the moving average. The window was applied to 7 trials before and after each point, resulting in the span of 15 trials in a window. Thus, we could achieve

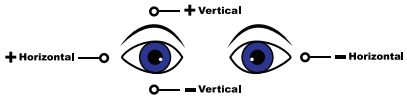


Fig. 3: EOG electrode placement.

the weight averaging that can balance between preserving the temporal dynamic of the central trial and reducing random variability in time responses. The obtained VTC then underwent the log transformation.

2) EEG Preprocessing: The preprocessing was performed using MNE-Python [17]. First, a 1–30 Hz FIR bandpass filter was applied to the raw EEG data. Next, the fast independent component analysis (FastICA) algorithm [18] was used to transform the EEG into 30 independent components, equal to the number of electrodes. The `find_bad_eog` function from MNE-Python was used to remove independent components highly correlated with EOG[1]. The data was then epoched from -200 to 300 milliseconds relative to the stimulus onset. Epochs with the peak-to-peak EEG amplitudes exceeding 100 μV were rejected. Two participants who had more than 20% of their epochs rejected were excluded from the further analysis in the next steps.

3) Feature Extraction:

a) Frequency Power Features: For each epoch, the power spectral density $P(f)$ of the EEG was calculated using Welch's method [19]. From $P(f)$, the relative power feature for the theta band (4–8 Hz) was defined as:

$$\theta = \frac{\int_4^8 P(f)df}{\int_4^{30} P(f)df} \tag{1}$$

The denominator 4–30 Hz represents the band extracted by the preprocessing bandpass filter. Similarly, relative power features α and β were defined for the alpha (8–13 Hz) and beta (13–30 Hz) bands, respectively. Power ratios $\frac{\theta}{\alpha}$, $\frac{\theta}{\beta}$, and $\frac{\alpha}{\beta}$ were also calculated. All relative powers and power ratios were expressed in decibels (dB).

b) Information Theoretic (Entropy) Features: The EEG was decomposed into theta, alpha, and beta bands using FIR bandpass filters for each epoch. Approximate entropy, sample entropy, permutation entropy, and spectral entropy were extracted from each EEG oscillation band. The Python library EntropyHub [20] was used for entropy calculations.

4) Feature Selection Using Sparse Modeling: Group Lasso [13], a sparse modeling technique, was used to select EEG features that best explain the VTC. It is a linear regression method that can shrink all parameters within a group to zero. This method was for feature selection by rejecting features with zero coefficients and selecting those that were left with non-zero coefficients. The regression coefficients (weight coefficients) $w \in \mathbb{R}^{p\times 1}$ was divided corresponding to different G features. Each group were represented as $w_g \in \mathbb{R}^{p_g\times 1}$, where p_g was the number of electrodes for feature g. Given the VTC $y^{(n)} \in \mathbb{R}^{l\times 1}$, where l represents the number or trials in the task for participant, and explanatory variables $X_g^{(n)} \in \mathbb{R}^{l\times p_g}$ for each feature of participant n, group Lasso estimates the regression coefficients as:

$$\widehat{w}_\lambda = \arg\min_w \left(\sum_{n=1}^{N} \left\| y^{(n)} - \sum_{g=1}^{G} X_g^{(n)} w_g \right\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g}\, \|w_g\|_2 \right) \tag{2}$$

where $\|\cdot\|_2$ is the ℓ_2-norm. The term $\sum_{g=1}^{G} \sqrt{p_g}\|w_g\|_2$ is the $\ell_{2,1}$-norm, which promotes sparsity across entire groups of coefficients. The regularization parameter λ was set to 0.001.

For each selected feature g, the $\ell_{2,1}$ norm term $\|w_g\|_2$ was compared, and the weights for each electrode of the top three features with the largest $\|w_g\|_2$ were examined.

5) Cross-correlation Analysis between Features and VTC: For each of these top three features, the electrode that provided the largest weight $\|w_g\|_2$, representing the most effective spatial location for that feature, was selected. The cross-correlation functions were calculated between the time course of the feature at the selected electrode and the VTC time course. Considering potential time lags between EEG features and VTC, t-tests were conducted on the maximum cross-correlation coefficients. The null hypothesis was that the cross-correlation coefficient is zero, with a significance level of 0.05.

III. Results

This section presents the outcomes of our analysis, focusing on two main aspects: the comparison of weights in selected features from the group Lasso analysis, and the cross-correlation between these features and the VTC. Our findings reveal patterns in EEG features that were associated with sustained attention during the go/no-go task.

A. Comparison of Weights in Selected Features

Application of group Lasso to EEG of all participants ($N = 14$) resulted in the selection of 10 features as shown in Figure 4. The top three features in terms of weight were the power ratios $\frac{\theta}{\alpha}$, $\frac{\alpha}{\beta}$, and $\frac{\theta}{\beta}$, where the top two features included α. Additionally, the spectral entropy of the beta band and the relative power of the alpha band were above-average weights. On the other hand, other selected features included sample entropy and approximate entropy of the beta band, which represent the temporal complexity of the beta band, had the weights below the average.

Figures 5a, 5b, and 5c show bar graphs of the weights for each electrode for the top three features. Electrodes were colored and plotted in blue for the left hemisphere, green for the midline, and orange for the right hemisphere. As shown in Figures 5a and 5c, $\frac{\theta}{\alpha}$ and $\frac{\theta}{\beta}$ both had positive weights for the midline electrodes Fz, Cz, and Pz. Figure 5b shows that for $\frac{\alpha}{\beta}$, electrodes in the frontal, temporal, and occipital regions had weights greater than 0.01. Moreover, we also observed a tendency for weights that were greater than 0.01 to be more distributed in the left hemisphere compared to the right hemisphere.

[1]Specifically, independent components with absolute z-scores greater than 3 for their Pearson correlation coefficients with EOG were removed.

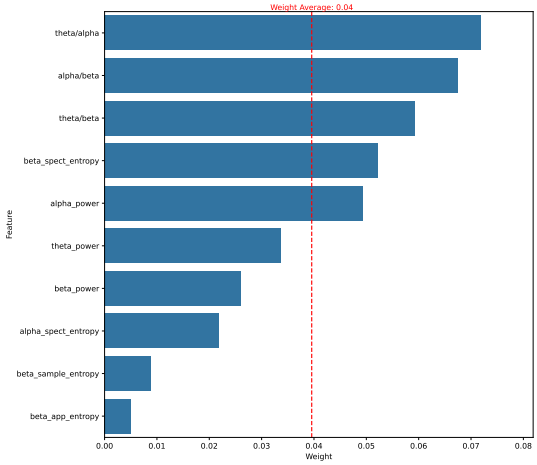


Fig. 4: $\|w_g\|_2$ values for the 10 selected features, shown in descending order. The vertical axis shows the features. The red dotted line indicates the mean $\|w_g\|_2$ value of the 10 features.

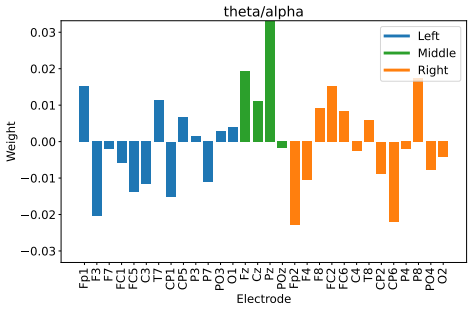
B. Comparison of Cross-correlation Coefficients between Features and VTC

Figures 6 shows examples of the line graphs overlaying the features and VTC for a participant for the features that have the largest (Figure 6a) and the lowest (Figure 6b) maximum cross-correlation at the electrodes with the largest weights. In the figures, blue lines represent the features and red lines represent the VTC. The maximum cross-correlation coefficient between the feature and VTC is shown in the top left of each figure. The highest maximum cross-correlation coefficient between the feature and VTC was 0.412, observed for $\frac{\theta}{\alpha}$ feature at the Pz electrode, while the lowest coefficient was 0.139 for $\frac{\alpha}{\beta}$ feature at the F3 location. Figure 7 shows the box plots of the maximum cross-correlation coefficients for each participant across the three features. T-tests were conducted under the null hypothesis that the cross-correlation coefficient is zero. As a result, the null hypothesis was rejected for all three features ($t(13) = 12.421, p < .001, d = 3.320$; $t(13) = 11.949, p < .001, d = 3.194$; $t(13) = 3.17, p < .001, d = 3.166$).

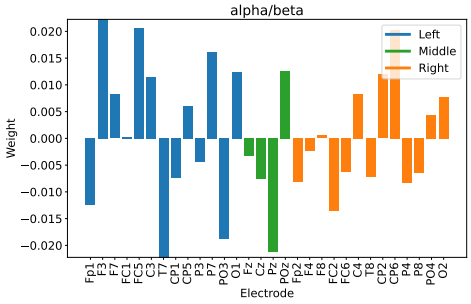
IV. Discussion

This study aimed to identify EEG features that effectively represent concentration levels during the go/no-go task, with a focus on the multi-channel nature of EEG data. Our analysis, using group Lasso for feature selection, revealed several key findings that provide insights into the neurophysiological correlates of sustained attention. In this section, we discuss the significance of our results, their relationship to existing literature, and their potential implications for understanding and measuring concentration.

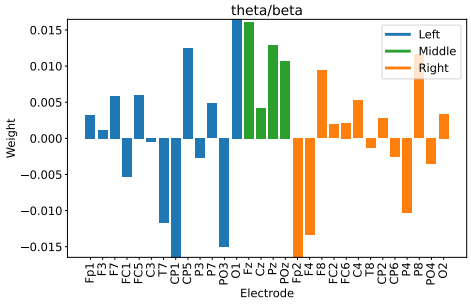
The feature selection results from group Lasso revealed that power ratios had the highest weights. This suggests that power ratios are more sensitive to fluctuations in concentration levels than features from single-frequency bands. As mentioned in Section I, it is known that as concentration decreases, power



(a) Electrode-wise weights for $\frac{\theta}{\alpha}$



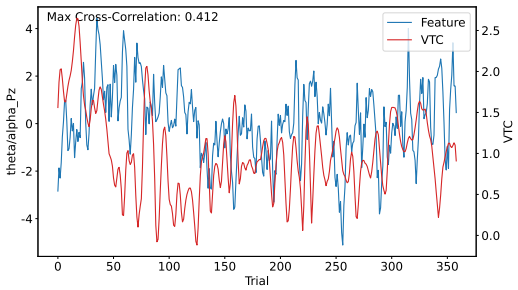
(b) Electrode-wise weights for $\frac{\alpha}{\beta}$



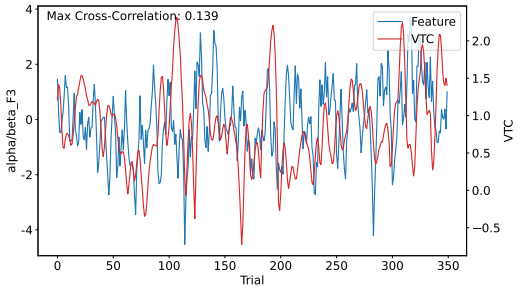
(c) Electrode-wise weights for $\frac{\theta}{\beta}$

Fig. 5: Electrode-wise weights for the top three features with the largest $\|w_g\|_2$ among the 10 selected features. The horizontal axis shows the 30 electrodes, and the vertical axis shows the weight of each electrode. Electrodes are color-coded: blue for the left hemisphere, green for the midline, and orange for the right hemisphere.

in the theta and alpha bands increases while power in the beta band decreases [10]. This inverse relationship between low and high frequencies may be strongly reflected in the power ratios. Regarding the electrode-specific weights of the power ratios, $\frac{\theta}{\alpha}$ and $\frac{\theta}{\beta}$ both showed positive weights for the midline electrode. This could be attributed to the influence of theta band power, which is common to both ratios. Theta waves are known to primarily appear during sleep [21]. In our experiment, which required sustained concentration over an extended period, participants may have temporarily entered a near-sleep state, resulting in increased theta band power. It is known that such temporary near-sleep states reduce concentration and activate theta band signals in the frontal and parietal regions along



(a) The highest maximum cross-correlation for $\frac{\theta}{\alpha}$ at Pz.



(b) The lowest maximum cross-correlation for $\frac{\alpha}{\beta}$ at F3.

Fig. 6: Example of the time course of feature and VTC at the most effective electrode for a participant. The horizontal axis shows the trial number, the left vertical axis shows the feature value and the right vertical axis shows VTC.

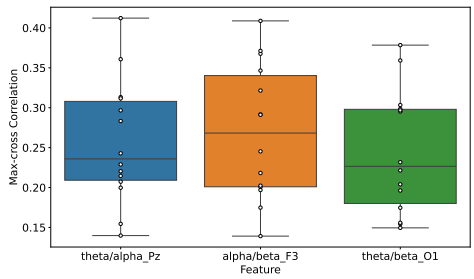


Fig. 7: Box plots of the maximum cross-correlation coefficients between VTC and the three features with the largest $\|\beta_g\|_2$ ($\frac{\theta}{\alpha}$, $\frac{\alpha}{\beta}$, $\frac{\theta}{\beta}$) at the electrodes with maximum weight coefficients. Circles indicate the maximum cross-correlation coefficients for each participant.

the midline [22]. Indeed, participants reported experiencing temporary drowsiness during the experiment. These findings suggest that power ratios including theta band activity from frontal and parietal midline regions may capture fluctuations in concentration levels. Moreover, Behzania et al. also observed that the theta power when the participant during a sustained attention task increases to a similar level as the closed-eye resting state [9]. This suggests that performing a long task, even without entering a near-sleep state, can increase theta power and potentially lower concentration. Therefore, the power ratio that included theta power can be promising for sustained attention assessment.

Focusing on $\frac{\beta}{\alpha}$, electrodes in the frontal, temporal, and occipital regions of the left hemisphere show large weights. The activation of frontal and occipital regions may be related to the default mode network (DMN) [23]. Activation of the frontal and occipital cortices and an active DMN are associated with states where individuals are relaxed and not focusing on external stimuli [24]. In our analysis, the relatively large weights for frontal, temporal, and occipital electrodes might imply that the DMN has been strengthened and resulted in enhancement of the alpha band activity in these regions [25]. The larger weights in the left hemisphere could be due to hemispheric dominance, as all except one of the analyzed participants were right-handed. It might also be noted that concentration has also been reported to be related to EEG asymmetry in the frontal region [26], suggesting that this hemispheric bias might reflect concentration levels as well.

When we compared the time course of the features and VTC, we found the high maximum cross-correlation coefficient between them was 0.412. Although the lowest coefficient was as low as 0.139, we could still observe a similar trend plotted for time course comparison, indicating that the features can capture the general trends of VTC. Furthermore, the cross-correlation coefficient tests demonstrated that these three features effectively evaluate individual concentration levels. Combining these features may enable the estimation of concentration levels from EEG characteristics.

V. Conclusion

This study aimed to identify EEG features that effectively represent concentration levels during a go/no-go task, addressing the multi-channel nature of EEG data. Our group Lasso analysis demonstrated that power ratios of theta, alpha, and beta bands in EEG are superior to the power itself in explaining the concentration levels relative to VTC. The ratios including the theta band showed relatively large weights in midline electrodes, while the $\frac{\beta}{\alpha}$ ratio showed prominence in frontal, temporal, and occipital electrodes. These findings align with established literature on attention networks and the default mode network, providing a neurophysiological basis for our results. We conclude that combining these EEG features could potentially estimate concentration from EEG data, addressing our initial research question.

While this study provides valuable insights into the EEG correlates of concentration, it is important to note limitations such as the small sample size and the specific task context. Moreover, the gap between static stimuli and conversational dynamics issues should be considered in the experimental design for the development of more accurate sustained attention assessment tools. Future research could explore the generalizability of these findings to other cognitive tasks and investigate how these EEG features relate to different measures of attention and cognitive performance.

Acknowledgment

This study was partly supported by the Moonshot Project, JPMJMS2031.

References

[1] D. E. Broadbent, P. F. Cooper, P. FitzGerald, and K. R. Parkes, "The cognitive failures questionnaire (cfq) and its correlates," *British Journal of Clinical Psychology*, vol. 21, no. 1, pp. 1–16, 1982.

[2] J. Smallwood, J. B. Davies, D. Heim, *et al.*, "Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention," *Consciousness and Cognition*, vol. 13, no. 4, pp. 657–690, 2004.

[3] B. L. Trommer, J.-A. B. Hoeppner, R. Lorber, and K. J. Armstrong, "The go-no-go paradigm in attention deficit disorder," *Annals of Neurology*, vol. 24, no. 5, pp. 610–614, 1988.

[4] Y. Kaga and H. Kanemura, "Event-Related Potentials for Cognitive Assessment of Patients with Epilepsy," *Pediatrics Therapeutics*, vol. 03, no. 03, 2013.

[5] H. E. Rosvold, A. F. Mirsky, I. Sarason, J. Bransome E. D., and L. H. Beck, "A continuous performance test of brain damage," *Journal of Consulting Psychology*, vol. 20, no. 5, 1956.

[6] I. H. Robertson, T. Manly, J. Andrade, B. T. Baddeley, and J. Yiend, "'Oops!': Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects," *Neuropsychologia*, vol. 35, no. 6, pp. 747–758, 1997.

[7] I. Pershin, G. Candrian, M. Münger, *et al.*, "Vigilance described by the time-on-task effect in EEG activity during a cued Go/NoGo task," *International Journal of Psychophysiology*, vol. 183, pp. 92–102, 2023.

[8] M. Torkamani-Azar, S. D. Kanik, S. Aydin, and M. Cetin, "Prediction of Reaction Time and Vigilance Variability from Spatio-Spectral Features of Resting-State EEG in a Long Sustained Attention Task," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2550–2558, 2020. eprint: 1910.10076.

[9] A. Behzadnia, M. Ghoshuni, and S. A. Chermahini, "EEG Activities and the Sustained Attention Performance," *Neurophysiology*, vol. 49, no. 3, pp. 226–233, 2017.

[10] S. Coelli, R. Barbieri, G. Reni, C. Zucca, and A. M. Bianchi, "EEG indices correlate with sustained attention performance in patients affected by diffuse axonal injury," *Medical and Biological Engineering and Computing*, vol. 56, no. 6, pp. 991–1001, Jun. 2018.

[11] W. Li, D. Ming, R. Xu, H. Ding, H. Qi, and B. Wan, "Research on Visual Attention Classification Based on EEG Entropy Parameters," *IFMBE Proceedings*, vol. 39 IFMBE, pp. 1553–1556, 2013.

[12] W. Wan, X. Cui, Z. Gao, and Z. Gu, "Frontal EEG-Based Multi-Level Attention States Recognition Using Dynamical Complexity and Extreme Gradient Boosting," *Frontiers in Human Neuroscience*, vol. 15, 2021.

[13] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[14] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Frontiers in Bioinformatics*, vol. 2, no. June, pp. 1–17, 2022.

[15] M. I. Conley, D. V. Dellarco, E. Rubien-Thomas, *et al.*, "The racially diverse affective expression (radiate) face stimulus set," *Psychiatry Research*, vol. 270, pp. 1059–1067, 2018.

[16] M. Esterman, S. K. Noonan, M. Rosenberg, and J. Degutis, "In the Zone or Zoning Out? Tracking Behavioral and Neural Fluctuations During Sustained Attention," *Cerebral Cortex*, vol. 23, no. 11, pp. 2712–2723, Nov. 2013, ISSN: 1047-3211. DOI: 10.1093/CERCOR/BHS261.

[17] A. Gramfort, M. Luessi, E. Larson, *et al.*, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, no. 267, pp. 1–13, 2013.

[18] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411–430, 2000.

[19] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[20] M. W. Flood and B. Grimm, "Entropyhub: An open-source toolkit for entropic time series analysis," *PLOS ONE*, vol. 16, no. 11, pp. 1–20, Nov. 2021.

[21] C. Cajochen, R. Foy, and D. Dijk, "Frontal predominance of a relative increase in sleep delta and theta eeg activity after sleep loss in humans," *Sleep Research Online : SRO*, vol. 2, no. 3, pp. 65–69, 1999.

[22] T. Andrillon, A. Burns, T. Mackay, J. Windt, and N. Tsuchiya, "Predicting lapses of attention with sleep-like slow waves," *Nature Communications 2021 12:1*, vol. 12, no. 1, pp. 1–12, Jun. 2021.

[23] M. E. Raichle, "The Brain's Default Mode Network," *Annual Review of Neuroscience*, vol. 38, no. Volume 38, 2015, pp. 433–447, Jul. 2015.

[24] R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter, "The brain's default network," *Annals of the New York Academy of Sciences*, vol. 1124, no. 1, pp. 1–38, 2008.

[25] M. S. Clayton, N. Yeung, and R. Cohen Kadosh, "The roles of cortical oscillations in sustained attention," *Trends in Cognitive Sciences*, vol. 19, no. 4, pp. 188–195, 2015.

[26] R. Liu and M. A. Bell, "Fearful Temperament in Middle Childhood Predicts Adolescent Attention Bias and Anxiety Symptoms: The Moderating Role of Frontal EEG Asymmetry," *Development and Psychopathology*, vol. 35, no. 3, p. 1335, 2023.

Forward Prediction-Guided Cross-Partition Targeted Pruning for VVenC

Jingyuan Tang* and Songlin Sun†
* Beijing University of Posts and Telecommunications, Beijing
E-mail: tangjingyuan2018@bupt.edu.cn
† Beijing University of Posts and Telecommunications, Beijing
E-mail: slsun@bupt.edu.cn

***Abstract*—With the rapid development of digital video technology, video coding plays a crucial role in multimedia communication, network streaming, and other fields. Efficient video encoding algorithms can significantly reduce data storage and transmission costs, and improve video quality. This article focuses on VVenC, an open-source encoding platform based on the latest video encoding standard H.266/VVC, and studies and optimizes its video encoding block partitioning strategy. In the encoding process, the division of encoding units and pattern prediction are interdependent, and the accuracy of division is transmitted to the accuracy of pattern prediction, and vice versa. In order to ensure encoding efficiency and reduce encoding time, we propose a pruning algorithm based forward prediction-guided and cross-partition (FPG-CTP). We perform targeted pruning on the horizontal and vertical block partition based on the prediction results of the forward prediction model. The experimental results show that the optimization strategy proposed in this paper significantly improves compression performance while keeping the encoding complexity basically unchanged. On various test sequences, compared to the original VVenC encoder, the average time savings reached 48.67%, and the average bit rate increased by 1.81%. In addition, the work of this article also provides reference for optimizing block partitioning in other video coding standards, and has a certain promoting effect on the further development of video coding technology.**

I. Introduction

The rapid development of the video industry has gradually made applications such as live streaming, video calls, panoramic videos, and short video socializing mainstream. These applications also drive the development of video coding with different levels of video quality requirements. The video encoding standard Versatile Video Coding (H.266/VVC) released by the Joint Video Exploration Team (JVET) in 2020 has improved encoding efficiency by 50% compared to the previous generation High Efficiency Video Coding (H.265/HEVC) [1]. The H.266/VVC testing model is mainly used as a technical testing platform for evaluating and validating proposals during the standardization process in the field of video coding. It is a universal reference implementation, but the official believes that it should not be used as an encoding and decoding platform in the industrial sector [2]. VVenC can provide a publicly available and efficient implementation of H.266/VVC encoder. VVenC has made further optimizations based on VTM, including assembly depth optimization, search algorithm improvement, and support for parallel encoding [3]. The VVenC encoder provides 5 preset modes: slower, slow, medium, fast, and faster, which represent different encoding speeds and compression quality offsets, respectively. The slowest preset has the slowest encoding speed but the best encoding quality, while the fastest preset has the fastest encoding speed but the worst encoding quality. In the fastest configuration, VVenC runs approximately 140 times faster than VTM, while still having a performance gain of about 10% compared to HM.

The VVenC encoder still adopts the mainstream hybrid encoding framework, which divides the input video frame into multiple encoding tree units (CTUs) using different partitioning modes to recursively divide into encoding units of different sizes (CU) for encoding. The method of block segmentation greatly determines the efficiency of subsequent tasks, and an appropriate partitioning method can quickly partition pixel groups with the same features into the same area. The latest generation of encoding H.266/VVC has made significant improvements in block partitioning compared to the previous generation, making partitioning more flexible. AVS3, independently developed in China, proposed a similar approach to extend Quad Tree (EQT) [4] partitioning. However, the increase in complexity of partitioning patterns means an increase in coding time, and achieving a balance between time and performance has always been a persistent issue in coding. In the H.266/VVC reference software VTM 3.0, in order to reduce the intra frame encoding complexity, some scholars have proposed a H.266/VVC encoding unit length pruning based on prospective prediction [5]. This method first determines the optimal intra prediction mode within a two partition region based on Satd's Mode Decision Making (SMD), and then calculates the rate distortion cost (RD Cost) [6] of the optimal prediction mode for each block to select a better partition direction.

Inspired by this, we propose a forward prediction-guided cross-partition targeted pruning algorithm. Our approach utilizes encoded features within the block for decision-making. Furthermore, in predicting partition directions, we incorporate the Hadamard transform coefficients of residual signals and block gradients, providing a more comprehensive and robust basis for partitioning decisions.

The remainder of this paper is organized as follows: Section 2 offers a thorough review of contemporary strategies for optimizing block partitioning algorithms, with a focus on the

H.266/VVC and VVenC encoding platforms. This section also examines related research efforts aimed at achieving a balance between block partitioning complexity and encoding efficiency. In Section 3, we provide a detailed exposition of the proposed partition pruning algorithm, highlighting the innovative use of forward and down-sampling predictive modes. Section 4 presents and analyzes the experimental results, demonstrating the efficacy and advantages of our approach through various performance metrics and comparative analyses. Finally, Section 5 concludes the paper with a summary of our findings, implications for future research, and potential applications of the proposed algorithm in practical encoding scenarios.

II. Related Work

From the development trend of intra frame coding, there are currently two mainstream approaches for optimizing intra frame video coding block partitioning: One is to consider the potential features of blocks that have an impact on block partitioning and subsequent prediction processes. Based on the probability distribution obtained from mathematical statistics, appropriate empirical values are found through experiments to terminate the partitioning in advance. Under certain conditions, the subsequent RDO process is truncated to achieve time savings.Texture information is an important consideration factor in reducing the complexity of block partitioning decisions[7][8][9]. Cui Jing et al. [10] used texture information from four directions to terminate early through numerical statistics and experimental thresholds. Zhang Qiuwen et al. [11] used Bayesian algorithm to calculate the rough mode decision evaluation cost of the current CU, the texture distribution position and complexity in the CU, and thus skipped bad candidate partitioning methods. Liu Shishi et al. proposed a scoring mechanism based on a bottom tracing split pattern to predict the likelihood of being selected as the optimal partition type for the time-consuming search process of CU partitioning.

The second approach is to use machine learning or neural networks for training. Based on the first approach, neural networks are introduced to better simulate the termination of partitions and seek termination algorithms that are closer to the final partition results. Quan He [12] and others used random forests to predict the partitioning of CU. Divide CU into three categories: simple, complex, and fuzzy. The first two categories of random forest classifiers directly predict the optimal partition mode; For fuzzy CU, train another random forest to predict whether the partitioning process will terminate. Li Tianyi et al. established a large-scale database and applied a multi-stage exit CNN (MSE-CNN) model with an early exit mechanism to determine CU partitioning. Wang Fengqin [13] et al. divided the CU partitioning process into two stages based on FSVM and dag-SVM, and symmetrically extracted some features of standard deviation, directional complexity, and content difference complexity of the same CU, and judged whether to terminate partitioning early based on these features. In addition, lightweight neural networks have also been applied by Maraoui et al. in fast partitioning decision-making [14].

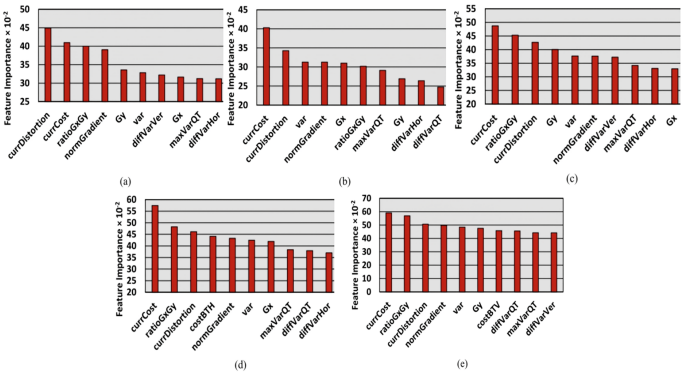


Fig. 1. Feature importance ranking of top ten features for (a) QT, (b) BTH, (c) BTV, (d) TTH, and (e) TTV classifiers [15].

The above two approaches are both aimed at the problem of increased encoding time caused by the complex RDO process that needs to be carried out in the video encoding process. The more accurate the simulation and early termination of the RDO process, the better the encoding performance can be provided. Overall, in the prediction process of CU partitioning, the information of CU itself and local regions play a crucial role, and gradients and textures play a decisive role in decision-making in most algorithms. In addition, there are also some local information that can potentially affect the partitioning method, including the current partitioning depth, current block QP, etc. Saldanha [15] and others conducted a large amount of statistical analysis on intra frame prediction and partitioning. By calculating the number of times different models use different features, they ranked the 10 features that are more important to the classifier, as shown in Fig.1.

Due to the high memory consumption in the calculation process of deep learning methods[16], in order to meet the common terminal encoding requirements, we optimize intra frame encoding based on traditional methods. At present, the optimization of encoders in VTM mainly focuses on the VTM platform, while there is less optimization for VVenC [17]. Although these solutions have excellent performance in H.266/VVC, they cannot be directly applied to VVenC. As a publicly available and efficient implementation of H.266/VVC encoder, further optimization of VVenC has practical significance in promoting the application of next-generation encoders.

III. Method

In most early termination algorithms for CU partitioning, the pattern prediction process and partitioning are often considered separately. However, in the actual encoding process, the division of encoding units and pattern prediction are interdependent, and the accuracy of division will be transmitted to the accuracy of pattern prediction. In order to ensure encoding efficiency and reduce encoding time, we propose a partition pruning algorithm based on forward prediction mode. By comparing the distortion values of horizontal and vertical

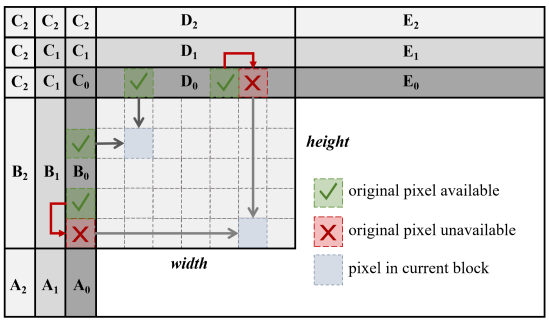


Fig. 2. Original pixel filling illustration.

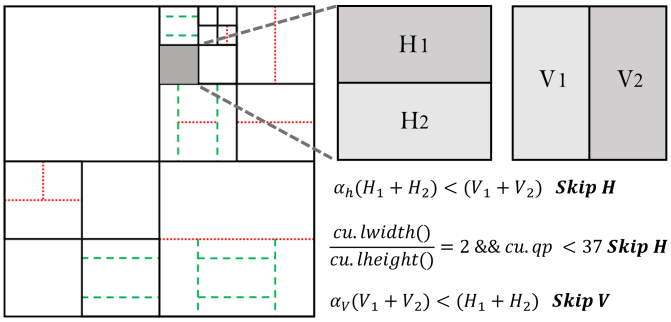


Fig. 3. Preliminary partitioning calculation illustration.

blocks, we determine whether certain prediction modes can be skipped.

The proposed methodology is encapsulated within a three-phase framework, encompassing forward prediction downsampling, original pixel filling, and preliminary partitioning computation. The following will provide a detailed introduction to the three steps.

- Forward Prediction Downsample
  Before performing intra frame mode prediction, downsample the predicted mode. Due to the expansion of angle prediction modes from 33 to 65 used in H.265/HEVC in the latest generation of video coding standards, complex mode selection is also increasing encoding time. We obtain the parent block prediction mode while retaining Planar, DC, and angle prediction modes 2, 4, 16, 32, 48, and 64. The above models have been retained from the earliest angle prediction models to the present day, which can represent as many directions as possible with the least number of angle models. The directions corresponding to these 6 angle modes include approximate directions of horizontal, vertical, diagonal, and transitional directions in the middle. Therefore, skipping the sub blocks in both horizontal and vertical directions in these modes is more representative. For the above prediction mode, its subsequent division is pruned through local features.
- Original Pixel Filling
  When the parent block operates in the specified mode, the reference pixels of the current prediction block are segmented and computed using either original pixel values or DC values to fill the reference pixels. If the reference pixel area is partially available, the algorithm first checks the availability of the reference pixel in the bottom-left corner. If available, all accessible reference pixels are filled with their original values, proceeding sequentially from bottom to top and left to right. Unavailable pixels are substituted with the nearest available pixel's original value. If the bottom-left reference pixel is unavailable, the algorithm traverses from bottom to top and left to right until the first available reference pixel is identified as shown in Fig.2. Since the reference pixels initially filled with reconstructed values are replaced with their original pixel values, streamlining the process of accessing reconstructed values.
- Preliminary Partitioning Calculation
  For each block that undergoes intra prediction, we calculate the SATD values for its four partitions. In video encoding, SATD (Sum of Absolute Transformed Differences) is a method of measuring the differences between two blocks. It is usually used in the block matching process to select the best matching block, thereby improving encoding efficiency. SATD calculates the difference between two blocks (usually the current block and the reference block). Firstly, the block difference is subjected to a transformation (such as Hadamard transformation), and then the absolute sum of the transformed coefficients is calculated. Assuming that $B1_{current}$ $and B1_{ref}$ represent the pixel values of the current block and the reference block, respectively, and T represents the Hadamard transform, the calculation steps for SATD are as follows:

$$SATD = \sum |T(B_{curr} - B_{ref})| \tag{1}$$

  SATD represents the sum of absolute values of transformed coefficients. During the partitioning process, the algorithm selects the more effective partitioning direction in both horizontal and vertical directions based on the predicted SATD value, and decides whether to skip, thereby effectively saving encoding time as shown in Fig.3. If the weighted distortion value in the vertical direction is less than the distortion value in the horizontal direction, horizontal segmentation is skipped, and vice versa. By comparing the distortion values in the horizontal and vertical directions, determine whether it is possible to skip the prediction in the corresponding direction.

In addition, we perform partition pruning based on forward prediction mode sampling based on the aspect ratio of important feature blocks and block level QP during the block partitioning process. We further prune small blocks with aspect ratio 2 and lower QP values. The specific algorithm steps are represented by pseudocode, as shown in algorithm 1.

This pseudocode demonstrates the main steps of intra prediction mode selection and segmentation pruning algorithms, and demonstrates how to optimize based on calculated SATD values and local features of blocks to save encoding time.

Algorithm 1 Intra Prediction Mode Selection and Partition Pruning Algorithm

Input: Current block B_{curr}, Reference block B_{ref}, Transform T, Weighting factors α_V, α_H
Output: Partition decision and SATD values
Initialization:
Define prediction modes: Planar, DC, Angular modes $\{2, 4, 16, 32, 48, 64\}$
SATD values: $minSadHad_{H1}$, $minSadHad_{H2}$, $minSadHad_{V1}$, $minSadHad_{V2}$
for each block B_{curr} **do**
  Downsample the prediction modes
  Use original pixel values instead of reconstructed values for reference pixels
  for each prediction mode in $\{Planar, DC, 2, 4, 16, 32, 48, 64\}$ **do**
    Calculate SATD for horizontal and vertical partitions:
    $SATD = \sum |T(B_{curr} - B_{ref})|$
    Compute SATD values: $minSATD_{H1}$, $minSATD_{H2}$, $minSATD_{V1}$, $minSadHad_{V2}$
  end for
  Evaluate partition efficiency:
  if $\alpha_V \times (minSATD_{V1} + minSATD_{V2}) < minSATD_{H1} + minSATD_{H2}$ **then**
    Skip horizontal partition
  else if $\alpha_H \times (minSATD_{H1} + minSATD_{H2}) < minSATD_{V1} + minSATD_{V2}$ **then**
    Skip vertical partition
  end if
  Prune partitions based on block characteristics:
  if block width/height ratio is 2 and QP is low **then**
    Skip horizontal partition
  end if
end for

IV. Experiments and Results

This experiment was conducted on a high-performance computer equipped with an Intel i9-4900K processor. The system is running a 64 bit Ubuntu 20.04 LTS operating system, equipped with 128GB memory and 1TB hard drive. All experiments were conducted on this unified platform to ensure consistency and comparability of results. We conducted experiments using the open-source video encoder VVenC (version 1.70). In order to study the performance of encoders under different configurations, we have made detailed settings and adjustments to the following aspects:

- Resolution
  We have chosen multiple different resolutions, including 720p, 1080p, and 4K, to evaluate the performance of the encoder under different video quality.
- Quantization parameter
  Four common quantization parameter, 22, 27, 32, and 37 were selected to test the performance of the encoder at different quantization levels.

The test sequences are provided by JVET, divided into six categories: A (3840x2160), B (1920x1080), C (832x480), D (416x240), E (1280x720), and F. Class A is further divided into two subcategories: A1 and A2. These videos use our proposed algorithm embed in VVenC v1.7.0 into their full function encoder (vencffapp) and encode them using slow presets in their All Intra encoding mode. In addition, we use VTM 12.0 reference software for corresponding decoding to ensure that the algorithm corresponds on the encoding and decoding end. To test the results and performance of our solution, we used Bjontegaard Delta PSNR (BDPSNR) and time savings to measure the effectiveness of the algorithm. BDPSNR is used to measure the quality of encoding, while time savings are used to measure the speed of encoding. The calculation Equation for saving time ΔET measured by encoding time is:

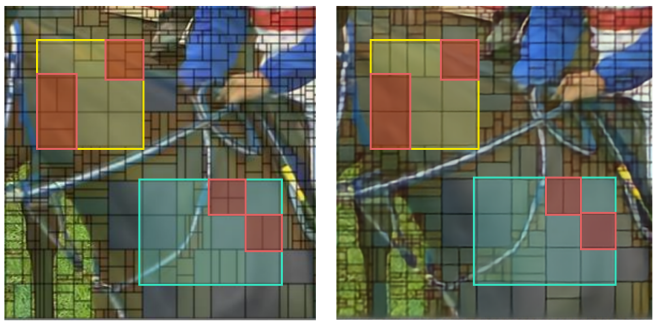
$$\Delta ET = \frac{1}{4} \sum_{QP_i \in \{22,27,32,37\}} \frac{T_R(QP_i) - T_P(QP_i)}{T_R(QP_i)} \quad (2)$$

T_R and T_P respectively represent the encoding time spent by the VVenC reference encoder and the VVenC recommendation encoder. We tested each video sequence using full frame mode for 2 seconds to ensure sufficient test frame rates. The Table I shows the test results. Through the above experimental

TABLE I
Algorithm performance test results

Class/Sequences		Y(%)	U(%)	V(%)	TS
A1	Campfire	2.11	1.77	1.65	0.38
	FoodMarket4	2.55	1.94	1.67	0.48
	Tango2	1.96	1.53	1.48	0.41
A2	CatRobot	2.23	1.61	1.44	0.47
	DaylightRoad2	2.17	1.69	1.56	0.43
	ParkRunning3	2.46	1.87	1.83	0.52
B	BasketballDrive	2.34	1.37	1.36	0.42
	BQTerrace	2.48	1.83	1.62	0.47
	MarketPlace	1.12	0.97	0.94	0.36
	RitualDance	2.57	1.84	1.91	0.46
	Cactus	2.27	1.48	1.46	0.45
C	BasketballDrill	2.78	2.11	2.12	0.56
	BQMall	2.45	1.60	1.48	0.55
	PartyScene	2.89	1.80	1.80	0.62
	RaceHorsesC	2.22	1.34	1.29	0.57
D	BasketballPass	2.97	2.43	2.25	0.59
	BQSquare	2.46	1.99	1.96	0.57
	BlowingBubbles	1.96	1.23	0.98	0.55
	RaceHorses	1.88	1.21	1.03	0.49
E	FourPeople	1.85	1.29	1.20	0.44
	Johnny	1.92	1.36	1.27	0.39
	KristenAndSara	2.05	1.82	1.80	0.47
F	BasketballDrilltext	2.34	1.92	1.77	0.50
	SlideEditing	1.79	1.25	1.17	0.42
	SlideShow	2.13	1.76	1.72	0.53

methods, we systematically evaluated the performance of the VVenC encoder. The experimental results demonstrate the



Original *Using FPG-CTP Algorithm*

Fig. 4. Comparison of the results of original partitioning and block partitioning using algorithms.

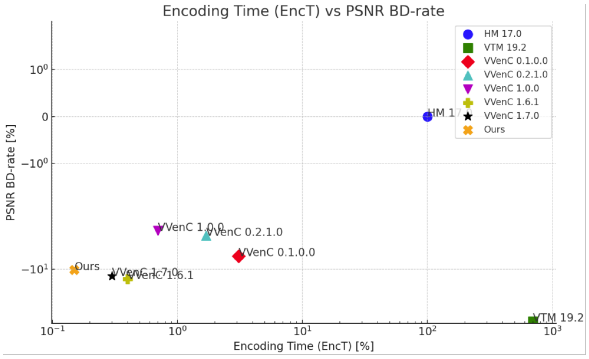


Fig. 5. Comparison of PSNR BD-rate performance corresponding to encoding time of different encoders.

encoding speed, compression efficiency, and video quality of the encoder at different QP value, sequence class, luminance and chromaticity components.

Table I presents a comprehensive comparison of the encoding performance between the algorithm proposed in this paper and the VVenC 1.7.0 anchor encoder, using the slow preset as a reference. The empirical evidence demonstrates that our solution markedly reduces the complexity associated with Coding Unit (CU) partitioning, with only a marginal loss in encoding efficiency. Specifically, the proposed method achieves an average time savings of 48.67%, while the Bjøntegaard Delta Bit Rate (BDBR) increases by a modest 1.81%.

The extent of complexity reduction varies across different video sequences. For instance, the complexity reduction for the Campfire video sequence is 38.12%, while for the PartyScene video sequence, it reaches a maximum of 62.03%. Correspondingly, the BDBR increases are 1.84% and 2.16%, respectively.

A. Performance Evaluation

Moreover, we have visually compared the partitioning results of the RaceHorseC sequence as shown in Fig.4. The left image represents the partitioning results of the original video encoding platform, while the right image shows the partitioning results after applying our algorithm.

The color blocks of identical colors in the Fig.4 illustrate the comparison of partitioning results within the same regions, exemplified by the yellow and green areas. The red area indicates the blocks that skip the binary tree partitioning process. From the comparative analysis, it is evident that our algorithm not only preserves a partitioning scheme similar to the original anchor framework but also minimizes the occurrence of multiple binary partitions. Consequently, this leads to a reduction in the overall complexity of video encoding.

B. Discussion

We conducted a comparative analysis of encoding time and Bjøntegaard Delta (BD) rate test results across multiple mainstream video encoding reference software. Using the reference software HM 17.0 of H.265/HEVC as the baseline, we evaluated the encoding performance of the reference software VTM 19.2 of H.266/VVC alongside various versions of VVenc, as illustrated in Fig.5. The scatter plot presented above delineates the relationship between Encoding Time (EncT) and Peak Signal-to-Noise Ratio (PSNR) BD-rate for the different reference software. The x-axis employs a logarithmic scale to more effectively represent disparities in encoding times. A shorter encoding time denotes higher encoding efficiency, while a higher BD-rate signifies superior encoding quality. Therefore, data points situated in the lower left corner of the plot indicate an optimal balance between encoding time and quality.

The baseline HM 17.0 is positioned at 100% encoding time (EncT) and 0% PSNR BD-rate, while VTM 19.2 shows a significantly higher EncT (700%) but offers the most substantial reduction in PSNR BD-rate (-41.50%), indicating a trade-off between encoding complexity and compression efficiency. Various versions of VVenC exhibit a range of trade-offs: VVenC 0.1.0.0 and VVenC 0.2.1.0 offer moderate encoding times (3.10% and 1.70%) with corresponding reductions in PSNR BD-rate (-7% and -4%), whereas later versions (VVenC 1.0.0, 1.6.1, and 1.7.0) further reduce encoding times but show varying impacts on PSNR BD-rate. Compare to HM 17.0, our proposed algorithm stands out with the lowest encoding time and a substantial reduction in PSNR BD-rate, highlighting its efficiency in achieving high compression performance with minimal encoding time.

These results underscore the efficiency of our algorithm in significantly decreasing the encoding time required for the CU partitioning process, thereby enhancing overall encoding efficiency. The slight trade-off in bit rate is justified by the substantial computational savings achieved, making our algorithm a valuable contribution to the field of video encoding.

V. Conclusion

This paper presents a novel partition pruning algorithm designed to enhance the efficiency of video encoding in the H.266/VVC standard, using the open-source VVenC platform. Our proposed algorithm, distinguished by its use of forward downsampling prediction mode to targeted pruning partition

process. Additionally, we integrates the block inner parameters during the encoding process for decision-making. Moreover, Hadamard transform coefficients of residual signals and block gradients are employed for predicting partition directions, thereby offering a more robust basis for partitioning decisions.

The experimental results indicate that our algorithm significantly improves compression performance while maintaining encoding complexity. Specifically, our optimization strategy achieves an average time savings of 48.67% and a 1.81% increase in bit rate compared to the original VVenC encoder. In this study, our proposed algorithm still has the potential for further optimization, but it can provide a reference for optimizing block partitioning in other video coding standards and promoting the development of video coding technology.

ACKNOWLEDGMENT

This work is funded by the National Key Research and Development Program of China for the research on the comprehensive plan for the extensive audiovisual environment of "Future TV", with Grant Number 2023YFF0904601.

REFERENCES

[1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012. DOI: 10.1109/TCSVT.2012.2221191.

[2] A. Wieckowski, J. Brandenburg, T. Hinz, *et al.*, "Vvenc: An open and optimized vvc encoder implementation," in *2021 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2021, pp. 1–2. DOI: 10.1109/ICMEW53276.2021.9455944.

[3] V. George, J. Brandenburg, G. Hege, *et al.*, "Efficient multi-threading strategies in vvenc, an open and optimized vvc encoder implementation," in *2022 IEEE International Symposium on Multimedia (ISM)*, 2022, pp. 18–25. DOI: 10.1109/ISM55400.2022.00009.

[4] J. Zhang, C. Jia, M. Lei, S. Wang, S. Ma, and W. Gao, "Recent development of avs video coding standard: Avs3," in *2019 Picture Coding Symposium (PCS)*, 2019, pp. 1–5. DOI: 10.1109/PCS48520.2019.8954503.

[5] M. Lei, F. Luo, X. Zhang, S. Wang, and S. Ma, "Look-ahead prediction based coding unit size pruning for vvc intra coding," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4120–4124. DOI: 10.1109/ICIP.2019.8803421.

[6] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998. DOI: 10.1109/79.733497.

[7] Z. Liu, M. Dong, X. H. Guan, M. Zhang, and R. Wang, "Fast isp coding mode optimization algorithm based on cu texture complexity for vvc," *EURASIP Journal on Image and Video Processing*, vol. 2021, pp. 1–14, 2021.

[8] Y. Li, G. Yang, Y. Song, H. Zhang, X. Ding, and D. Zhang, "Early intra cu size decision for versatile video coding based on a tunable decision model," *IEEE transactions on broadcasting*, vol. 67, no. 3, pp. 710–720, 2021.

[9] Q. Zhang, Y. Zhao, B. Jiang, L. Huang, and T. Wei, "Fast cu partition decision method based on texture characteristics for h. 266/vvc," *IEEE Access*, vol. 8, pp. 203 516–203 524, 2020.

[10] J. Cui, T. Zhang, C. Gu, X. Zhang, and S. Ma, "Gradient-based early termination of cu partition in vvc intra coding," in *2020 Data Compression Conference (DCC)*, 2020, pp. 103–112. DOI: 10.1109/DCC47342.2020.00018.

[11] Q. Zhang, Y. Zhao, B. Jiang, and Q. Wu, "Fast cu partition decision method based on bayes and improved de-blocking filter for h.266/vvc," *IEEE Access*, vol. 9, pp. 70 382–70 391, 2021. DOI: 10.1109/ACCESS.2021.3079350.

[12] Q. He, W. Wu, L. Luo, C. Zhu, and H. Guo, "Random forest based fast cu partition for vvc intra coding," in *2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2021, pp. 1–4. DOI: 10.1109/BMSB53066.2021.9547117.

[13] F. Wang, Z. Wang, and Q. Zhang, "Fsvm- and dag-svm-based fast cu-partitioning algorithm for vvc intra-coding," *Symmetry*, vol. 15, no. 5, 2023, ISSN: 2073-8994. DOI: 10.3390/sym15051078. [Online]. Available: https://www.mdpi.com/2073-8994/15/5/1078.

[14] M. Amna, W. Imen, and S. F. Ezahra, "Fast multi-type tree partitioning for versatile video coding using machine learning," *Signal, Image and Video Processing*, vol. 17, pp. 67–74, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248350588.

[15] M. Saldanha, G. Sanchez, C. Marcon, and L. Agostini, *Versatile Video Coding (VVC): Machine Learning and Heuristics*. Jan. 2022, ISBN: 978-3-031-11639-1. DOI: 10.1007/978-3-031-11640-7.

[16] K. Yang, D. Liu, Z. Chen, F. Wu, and W. Li, "Spatiotemporal generative adversarial network-based dynamic texture synthesis for surveillance video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 359–373, 2021.

[17] I. Taabane, D. Menard, A. Mansouri, and A. Ahaitouf, "A fast cu partition algorithm for vvenc encoder in intra configuration," in *2022 IEEE Workshop on Signal Processing Systems (SiPS)*, 2022, pp. 1–6. DOI: 10.1109/SiPS55645.2022.9919252.

Contrastive Learning Based Knowledge Distillation for Enhancing Defect Detection

Jing-Ming Guo*, Lun-Da Yuan, Cian Huang, and Yi-Chong Zeng
*National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C.
E-mail: jmguo@seed.net.tw Tel: +886-2-27303241

***Abstract*— Defect detection is critical for maintaining industrial production quality. This paper introduces pre-trained contrastive learning models to enhance defect detection. We integrate the simCLR model with two methods: fine-tuning and training from scratch. Fine-tuning adjusts pre-trained weights for defect detection tasks, yielding superior performance with reduced loss and higher Top-1 Accuracy. Training from scratch initializes model parameters specifically for anomaly detection, offering a tailored approach. The method employs the reverse knowledge distillation (RD) model with a simCLR-trained backbone, leveraging self-supervised and unsupervised learning exclusively on defect-free samples for practical implementation. Evaluation is conducted on the MVTec Anomaly Detection Dataset, encompassing 15 industrial product categories with high-resolution images. The experiment results demonstrate that fine-tuning achieves optimal performance, maintaining lower loss and higher Top-1 Accuracy.**

I. INTRODUCTION

Defect detection is critical in manufacturing to ensure product quality, reliability, and cost efficiency amidst automated production demands. Unsupervised and self-supervised learning methods have greatly improved defect detection by enhancing visual representation learning. Momentum Contrast (MoCo) improves self-supervised learning with a diverse set of negative samples [1]. SwAV learns visual features through cluster assignments without labeled data [2]. SimCLR shows that contrastive learning with simple designs and large batch sizes is effective [3].Siamese networks excel in one-shot learning and are useful for few-shot defect detection [4]. Knowledge distillation [5] helps smaller models learn from larger ones, improving performance. Bergmann et al. use a student-teacher framework for anomaly detection [6]. In the application of defect detection, recent advancements in deep learning technologies have introduced various approaches, such as Reverse Distillation (RD), CFA, and PatchCore. The RD improves anomaly detection by transferring knowledge from complex models to simpler ones. This technique uses one-class embeddings and distillation to enhance detection accuracy while making the process more efficient and less computationally demanding [7, 8]. CFA enhances anomaly localization by adjusting features with a coupled hypersphere method. This approach fine-tunes feature representations to better detect and localize specific anomalies, improving detection accuracy in complex scenarios [9]. PatchCore improves anomaly detection by using a memory bank to store features from normal samples. This approach helps in recognizing deviations from these patterns and effectively identifies subtle anomalies that might be overlooked otherwise [10].

This paper advocates using contrastive learning pre-trained models to enhance defect detection via knowledge distillation. Key contributions include:

- Recommend the contrastive learning pre-trained models to boost knowledge distillation model performance in defect detection.
- Explore and compare fine-tuning and training from scratch approaches for simCLR models, demonstrating that fine-tuning reduces loss and achieves high Top-1 accuracy.
- Enhance detection performance by employing pre-trained backbone architectures as auxiliary models in the RD model.

The rest of this paper is organized as follows: Section II introduces the methodology. Section III reveals the materials and experiment results. Sections IV and V are, respectively, the discussion and conclusion.

II. METHODOLOGY

The proposed method comprises the simCLR and the defect detection models. First, we introduce the processes of the training and the testing phases. Subsequently, the simCLR and defect detection models are presented in detail. Finally, anomaly scoring is defined for evaluating the performance.

A. Overall pipeline

First, we train the simCLR model [3], which serves as an auxiliary model for the Teacher Encoder (denoted as E^1、E^2、E^3) in anomaly detection via RD from the One-Class Embedding model. The proposed method enhances the performance of the Student Decoder (denoted as D^1、D^2、D^3) to normal images by training the teacher-student model. Fig.1 shows the architecture flow chart of the training phase.

During the testing phase in Fig.2, we remove the auxiliary model while retaining the trained mapping model named Projection Head. Then, defect images are inputted into the model sequentially. As the decoder learns the capability to reconstruct normal image features, encountering abnormal

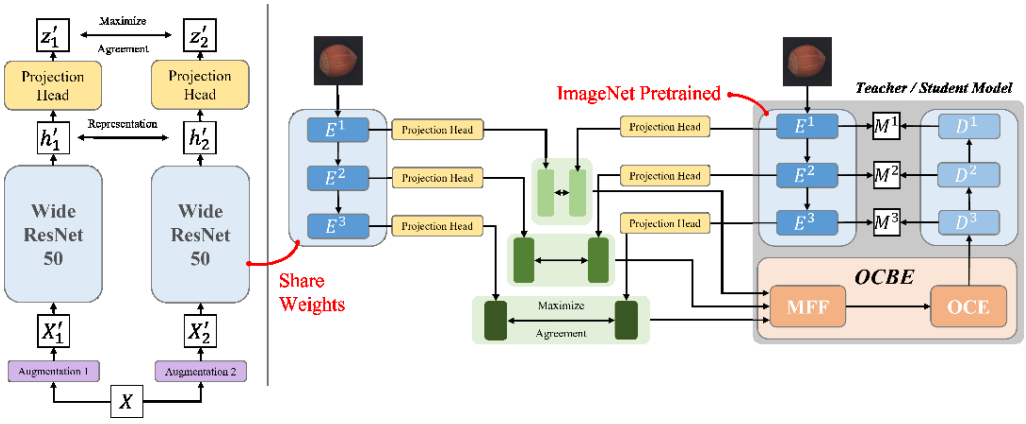


Fig. 1. Architecture flow chart of training phase

defects results in a significant difference between the feature representations outputted by the encoder and those reconstructed by the decoder. We present the feature extraction of each feature layer in the form of anomaly score maps (denoted as M^1、M^2、M^3) and finally fuse the differences of each layer to form the results of anomaly detection.

B. simCLR

The training methodology of simCLR uses contrastive learning. Given an input image x, data augmentation produces transformed images x'_1 and x'_2. These images are processed by a neural network, resulting in feature representations h'_1 and h'_2. These representations undergo non-linear transformation via a projection head, yielding outputs z'_1 and z'_2. Since both outputs originate from the same image x, the goal is to minimize the differences between z'_1 and z'_2 using a loss function, optimizing the model to reduce divergence.

The proposed method uses normal samples of the MVTec dataset as a benchmark. Data augmentation strategies include Random Resized Crop, Random Horizontal Flip, Random Color Jitter, Random Grayscale, and Gaussian Blur, which are applied to inputting RGB images. The backbone of the SimCLR model is Wide ResNet-50, whose projection head consists of two fully connected layers and a ReLU activation function shown in Fig.1.

C. Defect Detection Model

The Defect Detection Teacher-Student model has two main parts: an auxiliary model on the left and an RD model [8] on the right. The left side uses a pre-trained simCLR model [3] based on the Wide ResNet-50 architecture, with weights shared from simCLR. This model includes a trainable projection head that helps with defect detection during testing. The right side features the RD model, which functions as a teacher-student system. It includes a pre-trained Wide

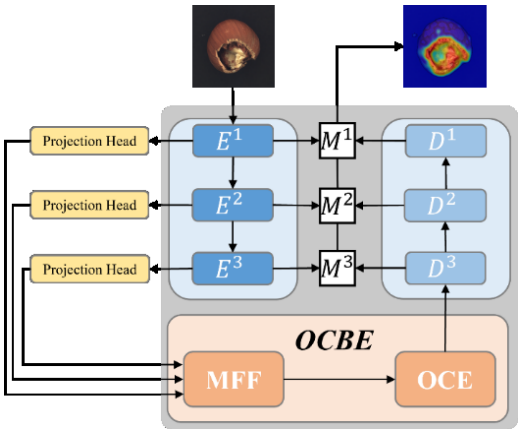


Fig. 2. Architecture flow chart of testing phase

ResNet-50 Teacher Encoder from ImageNet, an OCBE (One-Class Bottleneck Embedding) module, and a Student Decoder trained during the training phase. Both models process normal images, with the Wide ResNet-50 models pre-trained to enhance feature similarity, following principles similar to simCLR and contrastive learning.

Subsequently, feature representations flow into the OCBE module of the RD model for Multi-scale Feature Fusion (MFF) and dimensionality reduction through an OCE submodule. The Student Decoder reconstructs features from high to low levels, following reverse knowledge distillation principles. Cosine similarity continuously evaluates feature alignment between the Teacher Encoder and Student Decoder layers to ensure effective knowledge transfer.

Overall, the Defect Detection Knowledge Distillation model integrates two pre-trained Wide ResNet-50 backbones, three trainable projection heads, an OCBE module, and a Student Decoder, as shown in Fig.3. The left auxiliary model employs parameters derived from simCLR's Wide ResNet-50, while the RD model uses ImageNet pre-trained parameters.

D. Detection Mechanism

Before the testing phase, our method retains the trained projection head models and removes the pre-trained models

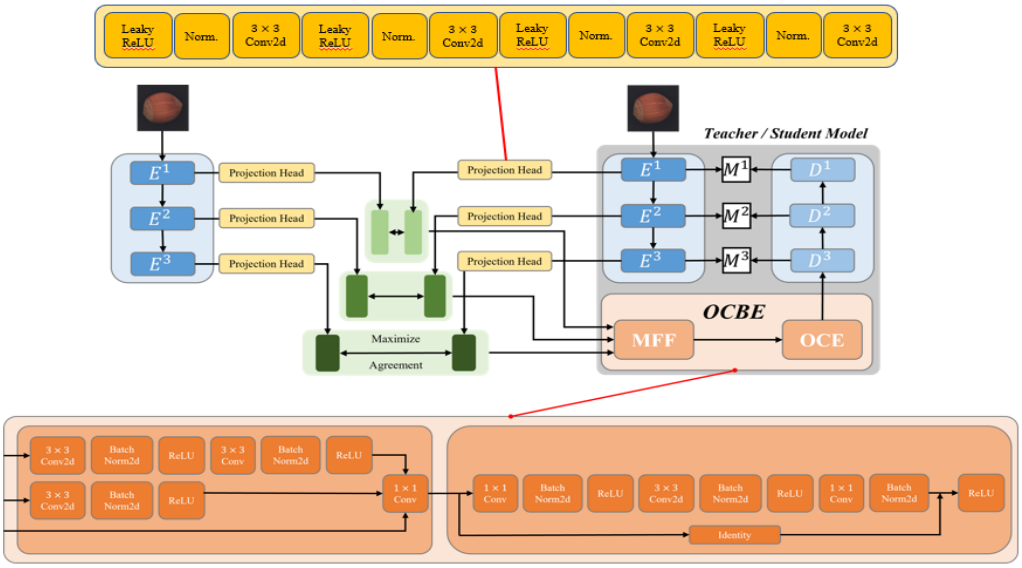


Fig. 3. Defect detection knowledge distillation model architecture – OCBE module and Projection head

used by the auxiliary model, as shown in Fig.2. During the testing phase, defect images replace normal ones. In the teacher-student model, defect images undergo initial processing by the Teacher Encoder by transforming each defect image into a high-dimensional embedding for precise cosine similarity comparison. Afterward, this embedding is passed to the Student Decoder, which reconstructs a normal image based on this information.

Feature extraction occurs layer by layer, capturing feature maps from each block. Those maps are inputted into the respective projection head models to fine-tune the auxiliary model through convolution operations with trained parameters. Fed feature maps from each layer into the OCBE module, and the MFF module adjusts these maps through zero or multiple convolution operations to standardize the size and dimensionality. A point-wise convolution adapts the overall channel count before the fused feature map proceeds to the OCE module for dimensionality reduction.

The Student Decoder restores the feature map, starting from high-level deep features and progressing to lower-level layers, capturing representations from each block. Next, cosine similarity between the restored feature maps from the Student Decoder and those extracted by the Teacher Encoder is calculated. This comparison generates anomaly score maps (Anomaly Maps) for each layer, indicating differences between encoding and decoding layers. Fig.4 depicts the scaled Anomaly Maps superposed on an original image highlighting defect-affected regions.

III. Materials and Experiment Results

This paper utilized the MVTec AD dataset, a benchmark for anomaly detection. MVTec AD simulates real-world industrial scenarios, a mix of normal and naturally occurring defect samples. Each category includes training and test sets

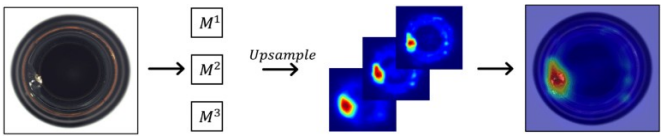


Fig. 4. Detection mechanism

with annotations. The training set contains only normal images, and the test set includes various defect types with annotations. Fig.5 shows normal images, defect images, and labeled data.

A. simCLR and RD model

In this experiment, we investigated two main aspects. First, we compared using an unsupervised learning (contrastive learning) model versus a supervised learning model as the teacher encoder in the RD model. Second, we evaluated the performance of the contrastive learning model when fine-tuned or trained from scratch within the defect detection teacher-student framework. Additionally, we assessed the training outcomes and detection performance of the defect detection model.

The results show that using the contrastive learning model to replace the RD model's teacher encoder was not as effective as a supervised learning approach with an ImageNet pre-trained backbone. However, 60% of the categories achieved AUC values above 0.9, indicating that the unsupervised pre-trained model is still effective for defect detection. Interestingly, training the backbone architecture from scratch produced better results. Table I lists the test results for both strategies, showing that the From Scratch method outperformed the Fine Tune approach.

B. Performance Evaluation

In the third experiment, we used the pre-trained backbone from the first experiment as an auxiliary model for fine-tuning

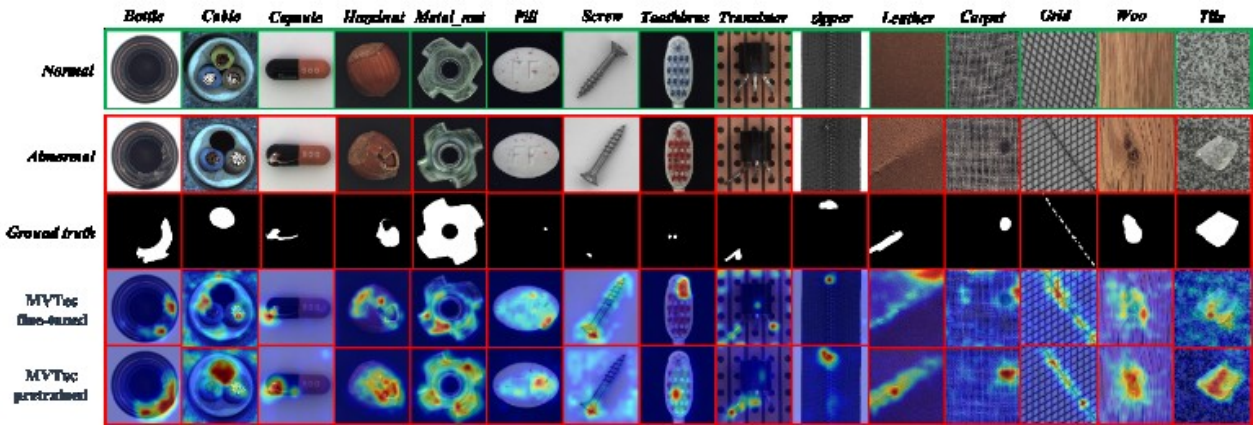


Fig. 5. MVTec dataset (1st and 2nd rows), labeled data (3rd rows), and visual results (4th and 5th rows)

Table I. Comparison of the AUC of each category and the original RD model in different pre-trained backbone architectures

	Bottle	Cable	Capsule	Carpet	Grid	Hazelnut	Leather	Metal nut
RD (Fine Tune)	0.8870	0.4300	0.9060	0.8480	0.9670	0.9100	0.8490	0.8360
RD (From Scratch)	**0.9030**	**0.8450**	**0.9200**	**0.8970**	**0.9780**	**0.9630**	**0.9860**	**0.8930**
	Pill	**Serew**	**Tile**	**Toothbrush**	**Transistor**	**Wood**	**Zipper**	
RD (Fine Tune)	0.8760	0.9570	**0.7800**	0.9230	0.5520	0.8450	0.8940	
RD (From Scratch)	**0.9080**	**0.9660**	0.7750	**0.9800**	**0.7370**	**0.8550**	**0.9700**	

Table II. Comparison of AUC of each category in different pre-trained backbone architectures

	Bottle	Cable	Capsule	Carpet	Grid	Hazelnut	Leather	Metal_nut
RD[8]	0.9870	0.9740	0.9870	0.9890	0.9930	0.9890	0.9940	0.9730
CFA[9]	**0.9884**	**0.9897**	**0.9911**	0.9928	0.9812	0.9885	0.9937	**0.9915**
PatchCore[10]	0.9860	0.9840	0.9880	0.9900	0.9870	0.9870	0.9930	0.9840
Proposed method (From Scratch)	0.9523	0.8651	0.9210	**0.9930**	**0.9951**	**0.9892**	**0.9949**	0.7588
	Pill	**Serew**	**Tile**	**Toothbrush**	**Transistor**	**Wood**	**Zipper**	
RD[8]	0.9820	**0.9960**	0.9560	**0.9910**	0.9250	0.9530	0.9820	
CFA[9]	**0.9893**	0.9891	0.9521	0.9896	**0.9806**	0.9153	**0.9902**	
PatchCore[10]	0.9740	0.9940	0.9540	0.9870	0.9630	0.9500	0.9880	
Proposed method (From Scratch)	0.9215	0.9278	**0.9586**	**0.9910**	0.7420	**0.9532**	0.9552	

the RD model's features. After fine-tuning, we removed the auxiliary model and kept only the trained mapping model. The results in Table II show improvements across categories. Specifically, training the backbone from scratch led to an average AUC increase from 83.07% to 90.51%. Further fine-tuning the backbone improved this to 91.86%, and training from scratch further increased it to 92.56%. Two key observations from the experiments are,

- Training the backbone architecture from scratch generally outperformed the fine-tuning approach in the knowledge distillation model.
- Using the pre-trained backbone architecture as an auxiliary model in subsequent knowledge distillation models was more effective than directly replacing the original teacher encode.

Additionally, the comparative analysis against the original RD model and existing defect detection methodologies, our

Table III. Comparison with other current modeling methods

	Carpet	Grid	Leather	Tile	Wood	Average
RD[8]	0.9890	0.9930	0.9940	0.9560	0.9530	0.9770
CFA[9]	0.9928	0.9812	0.9937	0.9521	0.9153	0.9670
PatchCore[10]	0.9900	0.9870	0.9930	0.9540	0.9500	0.9748
Proposed method (From Scratch)	**0.9930**	**0.9951**	**0.9949**	**0.9586**	**0.9532**	**0.9790**

proposed method achieved the highest AUC in 7 out of 15 categories, specifically Carpet, Grid, Leather, Tile, and Wood as listed in Table III. The fourth and fifth rows of Fig.5 display the visual results of defect detection using the fine-tuned and the pre-trained methods, respectively.

IV. Discussion

According to Table II, that the proposed method is weak in Cable, Metal nut, and Transistor due to issues with data augmentation techniques used during the backbone architecture training in the first experiment. As detailed in [3], simCLR employed augmentation methods Random Color Jitter reduces sensitivity to color variations, affecting the Cable category where color differences can lead to misclassifications as normal features.

Similarly, Random Horizontal Flip, which usually improves image recognition robustness, can cause misclassifications in the Metal nut category by treating rotated nuts as defects. In the Transistor category, variations like misplaced transistors (e.g., upside-down) are also seen as defects, which is problematic when no transistor is present.

It's important to explore alternative data augmentation techniques that enhance model performance without affecting defect detection accuracy. Such strategies can improve model stability and generalization, leading to more reliable defect detection across various datasets and scenarios.

V. Conclusion

This paper presents a method to enhance defect detection in knowledge distillation models using a contrastive learning pre-trained model. We evaluated the simCLR model with two strategies, namely Fine Tune and From Scratch. Our results showed that Fine Tune achieved lower loss and higher Top-1 Accuracy than From Scratch. We improved the Reverse Knowledge Distillation model by integrating the simCLR-trained backbone as an auxiliary model during training, which was removed for testing. The Fine Tune approach achieved an average AUC of 91.86%, while the From Scratch approach achieved 92.56%. This unsupervised method, which only requires unlabeled normal samples, is well-suited for industrial applications. Future research may explore multi-layer feature fusion and diverse data augmentation strategies for simCLR to further improve performance.

References

[1] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," in arXiv preprint arXiv:2003.04297, 2020.

[2] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," in arXiv preprint arXiv:2006.09882, 2020.

[3] T. Chen, S. Kornblith, M. Norouzi, & G. Hinton, "A simple framework for contrastive learning of visual representations," *International conference on machine learning*, pp. 1597-1607, 2020, PMLR.

[4] X. Chen, and K. He, "Exploring simple siamese representation learning," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.15750-15758, 2021.

[5] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," in arXiv preprint arXiv:1503.02531, 2015.

[6] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4183–4192, 2020.

[7] H. Deng and X. Li., "Anomaly detection via reverse distillation from one-class embedding," *in Proceedings of the IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 9737–9746, 2022.

[8] T. D. Tien, A. T. Nguyen, N. H. Tran, T. D. Huy, S. Duong, m, C. D. T. Nguyen, and S. Q. H. Truong, "Revisiting Reverse Distillation for Anomaly Detection, " *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24511–24520, 2023.

[9] S. Lee, S. Lee, B.-C. Song, "CFA: coupled-hypersphere based feature adaptation for target-oriented anomaly localization," *in IEEE Access*, no. 10, pp. 78446–78454, 2022.

[10] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14298–14308, 2022.

Screen Content Encoding Network Based on Deep Contextual Information

Tianyu Gong[1], Tao Zhang[2,*], Ye Zhong[1], Mengmeng Zhang[3] and Huihui Bai[1]
[1]Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[2]Beijing Polytechnic College, Beijing 100042, China
[3]The Faculty of Smart City, Beijing Union University, Beijing 102200, China
{24110103,21125295,hhbai}@bjtu.edu.cn, 2022001707@bgy.edu.cn, mengmeng@buu.edu.cn

***Abstract*—This paper proposes a deep contextual video coding network that integrates transformational jumps. The network takes a sequence of original in-screen content video frames as input and extracts motion features through motion estimation and a motion information encoder, with motion compensation occurring during decoding. It also extracts contextual information in the feature domain through a context encoder. This contextual information assists in both context encoding and decoding, as well as entropy encoding. By using contextual information as a condition for conditional encoding, the network transitions from predictive coding to conditional coding, aiding in the high-quality reconstruction of high-frequency content. The motion information encoder introduces a transformational jump branch into the analysis and synthesis process. This branch has the ability to extract coarse features and reconstruct them, thereby enhancing the encoding and decoding of visual signals. Finally, experiments confirm that this algorithm effectively improves the encoding performance of screen content videos.**

I. Introduction

In recent years, digital device advancements have made screen content images and videos central to communication. The growing data volume challenges image and video encoding technologies. Emerging needs in online education, virtual meetings, and cloud gaming highlight the urgency for efficient screen content encoding. This encoding aims to represent images and videos compactly with minimal visual quality loss, reducing storage and bandwidth use.

Camera-captured images often contain sensor noise and complex textures, whereas screen content images, typically computer-generated, are noise-free with discrete tones and fewer colors. They have finer lines, sharper edges, and more repetitive patterns. Traditional video encoding, designed for camera footage, uses a less effective hybrid prediction-transformation structure for screen content, risking detail loss and edge distortion. Current screen content encoding schemes, based on traditional methods, independently encode segmented blocks, ignoring overall content and causing high complexity. This paper contrasts video screen content encoding with traditional and deep video encoding algorithms H.264[1], H.265[2], DVC[3], Hu_ECCV[4], Lu_ECCV[5], FVC[6], DVCpro[7], EA_CVPR[8], and RLVC[9].

This work was supported in part by National Natural Science Foundation of China (No. 62331003, 62120106009), Beijing Natural Science Foundation (L223022), Natural Science Foundation of Hebei Province (F2024105029).

Existing end-to-end compression schemes, focused on natural images, overlook screen content image compression. The latter, being computer-generated, are noise-free with unique patterns and textures, affecting compression efficiency. Directly applying existing learning-based image compression approaches to screen content results in decreased rate and distortion performance. This paper aims to enhance the compression performance of screen content images. Inspired by deep contextual information and the concept of transform-skipping, we propose an end-to-end compression scheme for video screen content. Specifically, the proposed model thoroughly considers the characteristics of screen content, utilizing deep contextual learning to capture the high-frequency content in screen images, and applies deep contextual knowledge for the compression and reconstruction of video frames. Furthermore, a transform-skipping branch is integrated into the motion codec process. Transform-skipping refers to a technique used within the motion codec process, where specific transformations or layers in the feature extraction process are skipped to reduce computational complexity. This branch, capable of coarse feature extraction and reconstruction, effectively extracts multidimensional, high-dimensional features from screen content images. As a result, visual signals can be more simply interpreted at the encoder end and restored at the decoder end.

II. Overall Network Structure

To enhance screen content compression, this paper introduces a deep video compression algorithm tailored for screen content's high texture contrast and sharp edges. Utilizing deep contextual information and transform-skipping, the algorithm's structure, depicted in Figure 1, comprises a context codec and a motion information codec.

The algorithm processes video screen content frame sequences. Initially, motion estimation extracts motion information v_t from input and reference frames. The motion information codec, using a convolutional neural network with transform-skipping, compresses this information. The transform-skipping branch, adept at coarse feature extraction and reconstruction, simplifies screen content perception in its two branches and finely processes it in the main branch. This allows for straightforward encoding and precise reconstruction at the decoder end into $\hat{v}_t$. The second part utilizes the reconstructed motion information $\hat{v}_t$ and high-dimensional

features of the reconstructed reference frame $\hat{x}_{t-1}$, or $\tilde{x}_{t-1}$. These undergo warping and context mapping to produce deep contextual information $\overline{x}_t$. This context information then aids the context encoder and decoder in reconstructing high-frequency content, enhancing video quality. The context decoder outputs the reconstructed frame $\hat{x}_{t-1}$, with distortion calculated against the input frame x_t. The bitrate, derived from the high-dimensional features m_t and y_t quantized and entropy coded into a bitstream, is calculated alongside distortion. These metrics form the rate-distortion loss function, crucial for network optimization during training.

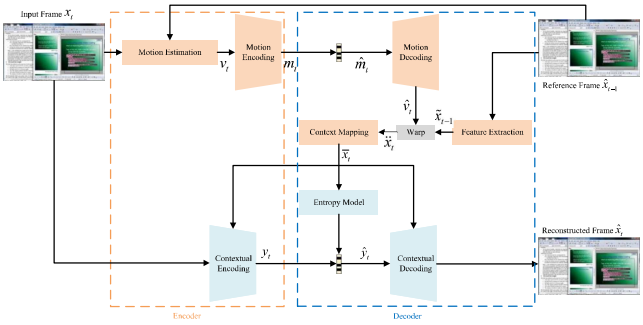


Fig. 1. Deep Context-Based Screen Content Coding Network.

The network framework of this paper can be represented as follows:

$$\hat{x}_t = f_{context}(\hat{x}_{t-1}) = f_{dec}(f_{enc}(x_t \mid \overline{x}_t) \mid \overline{x}_t) \otimes \overline{x}_t \quad (1)$$

In this representation, x_t denotes the input frame, $\hat{x}_t$ represents the predicted or reconstructed frame, $\overline{x}_t$ signifies the generated high-dimensional context, and $\hat{x}_{t-1}$ is the reference frame, i.e., the reconstructed previous frame. The function $f_{context}()$ is responsible for generating the context $\overline{x}_t$. $f_{enc}()$ and $f_{dec}()$ are the context encoder and decoder, distinct from the motion encoder and decoder. The deep contextual information is generated in a higher-dimensional feature domain, containing multidimensional, high-dimensional features. This provides a richer and more relevant set of conditions for encoding x_t, particularly for its high-frequency content.

III. Motion Encoding Network

A. Motion Encoding Network Based on Transform Skipping

This section details how integrating transform skipping into an end-to-end encoding framework can boost screen content image compression. Traditionally, transform skipping effectively handles sparse motion compensation residuals and intra prediction in screen content compression [10]. The motion encoding network's structure, based on transform skipping, is depicted in Figure 2. Our video screen content compression framework features an autoencoder-style encoder and decoder, with the multi-scale transform skipping encoder built upon the hyperprior model from Balle et al [11].

Figure 2 shows that the transform-skipping-based motion encoding network comprises convolutional, GDN, and ReLU

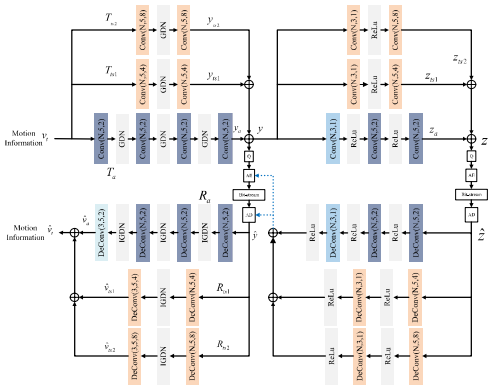


Fig. 2. Motion coding networks based on transform jumps.

layers. It features three branches in both the encoder and decoder, with most convolutional layers having a 5x5 kernel size and 128 channels (N=128). In the main branch, convolutional layers have a stride of 2, while the two side branches have strides of 4 and 8, respectively. This setup allows for local feature extraction while transform skipping captures larger-scale features, efficiently reducing spatial redundancy.

At the encoder, motion information v_t feeds into three scale analysis branches: a forward transform branch T_a and two transform skipping branches T_{ts1} and T_{ts2}. The forward transform branch T_a, resembling existing analysis transform structures, employs four sequential convolutional layers with a stride of 2. The transform skipping branches T_{ts1} and T_{ts2} use two convolutional layers each, with strides of 4 and 8, and a 5x5 kernel size with N channels. These branches collectively process the visual signal at different scales, extracting high-dimensional features across various scales. The T_a, T_{ts1}, and T_{ts2} branches generate the latent codes y_a, y_{ts1} and y_{ts2}, as follows:

$$\begin{aligned} y_a &= T_a(v_t, \theta_a) \\ y_{ts1} &= T_{ts1}(v_t, \theta_{ts1}) \\ y_{ts2} &= T_{ts2}(v_t, \theta_{ts2}) \end{aligned} \quad (2)$$

In this context, θ_a, θ_{ts1} and θ_{ts2} represent the model parameters of the main transform branch and the two transform skipping branches, respectively. The high-dimensional feature y is a combination of y_a, y_{ts1} and y_{ts2} as shown below.

$$y = y_a + y_{ts1} + y_{ts2} \quad (3)$$

Quantization is applied to the noise representation of the high-dimensional features, where $\widetilde{y}$ denotes the high-dimensional feature, and $\hat{y}$ is encoded into the bitstream as an entropy value.

$$\hat{y} = Q(\widetilde{y}) \quad (4)$$

This paper assumes that the distribution of the high-dimensional feature $\widetilde{y}$ follows a zero-mean Gaussian distribution. The motion encoding network based on transform

skipping predicts the scale of each distribution conditioned on the hyperprior z. In this section, two transform skipping branches are designed for the encoding and decoding processes of motion encoding. During the encoding process, the main branch and the two transform skipping branches separately generate z_a, z_{ts1} and z_{ts2}. The hyperprior feature represented by z can be expressed as follows:

$$z = z_a + z_{ts1} + z_{ts2} \tag{5}$$

After quantization, z is compressed using the entropy model [12]. At the decoder end, it is input into the hyperprior decoder. The main reconstruction branch and the two transform skipping reconstruction branches in the hyperprior decoder yield the standard deviation $\hat{\sigma}$ of the Gaussian distribution. Then, $\hat{\sigma}$ is sent to the arithmetic encoder and decoder to generate the probability distribution of $\hat{y}$ for encoding and decoding.

Similarly, reconstructing the motion information $\hat{v}_t$ from the high-dimensional feature $\hat{y}$ involves three branches. The main reconstruction path and the two transform skipping reconstruction paths are described as follows:

$$\begin{aligned} \hat{v}_a &= R_a(\hat{y}, \delta_a) \\ \hat{v}_{ts1} &= R_{ts1}(\hat{y}, \delta_{ts1}) \\ \hat{v}_{ts2} &= R_{ts2}(\hat{y}, \delta_{ts2}) \end{aligned} \tag{6}$$

Here, R_a, $\hat{v}_{ts1}$ and $\hat{v}_{ts2}$ respectively represent the main reconstruction branch and the two transform skipping branches in the decoder. δ_a, δ_{ts1}, and δ_{ts2} respectively denote the network parameters of the main reconstruction branch and the two transform skipping branches in the decoder. Finally, the outputs of the main reconstruction branch and the two transform skipping branches are summed to obtain the reconstructed motion information $\hat{v}_t$, as expressed in Equation 7:

$$\hat{v}_t = \hat{v}_a + \hat{v}_{ts1} + \hat{v}_{ts2} \tag{7}$$

Finally, the quantized high-dimensional feature $\hat{y}$ and the hyperprior feature $\hat{z}$, along with their formed bitstream, are used by the entropy model to calculate the bitrate. This is used in the training process's loss function to optimize the overall video screen content encoding network.

B. Feature Extraction and Contextual Refinement Networks

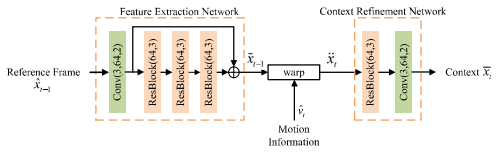


Fig. 3. Feature Extraction and Contextual Refinement Network.

This paper introduces a feature extraction network to generate contextual information in the feature domain. It consists of a convolutional layer with 64 channels, a 3x3 kernel size, and a stride of 2, followed by three 64-channel residual blocks with a 3x3 kernel size. The reference frame $\hat{x}_{t-1}$ inputs into this network, and through these layers, high-dimensional features $\widetilde{x}_{t-1}$ are obtained, shifting reference coordinates from the pixel to the feature domain. A $warp$ function, using motion information $\hat{v}_t$, distorts $\widetilde{x}_{t-1}$ to produce the preliminary context $\ddot{x}_t$. However, this initial context is coarse, as the warping may cause spatial misalignment. To enhance the preliminary context $\ddot{x}_t$, a context refinement network is employed. It comprises a 64-channel residual block with a 3x3 kernel size and a convolutional layer with 64 channels, a 3x3 kernel size, and a stride of 2. This network refines the initial context, addressing spatial discontinuities from warping, resulting in the final deep contextual information $\overline{x}_t$. Figure 3 illustrates the specific structures of both the feature extraction and context refinement networks. The process of feature extraction and context generation is summarized as a context generation function, detailed in Equation 8.

$$f_{context}(\hat{x}_{t-1}) = f_{cr}(warp(f_{fe}(\hat{x}_{t-1}), \hat{v}_t)) \tag{8}$$

$f_{context()}$ uses motion information $\hat{v}_t$ to guide context extraction and expand the learning area for the context. This generation process is akin to motion compensation in the feature domain. Without feature extraction, applying motion compensation directly in the pixel domain would mean that the learned context remains tied to the pixel domain. This could limit the relevance between the context's corresponding positions and the input frame x_t, reducing its efficacy in compressing x_t. Therefore, the feature extraction module is crucial. Deep contextual information in the feature domain holds extensive information, significantly enhancing the compression of high-frequency content. Furthermore, feature-domain context, as opposed to pixel-domain motion compensation, allows the network to utilize a larger reference area for extracting useful features.

IV. Context Coding Network

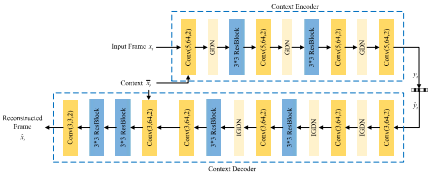


Fig. 4. Context coding network.

The context encoding network is mainly composed of convolutional layers, GDN layers, and residual blocks. In the context encoder, the key components are 64-channel convolutional layers with a 5x5 kernel size and a stride of 2. Conversely, the context decoder primarily utilizes 64-channel convolutional layers with a 3x3 kernel size and a stride of 2, as detailed in Figure 4. The context encoder's input includes the current

frame x_t and context $\overline{x}_t$, which it encodes into 96-channel, 16x downsampled high-dimensional latent features y_t. The context decoder begins by upsampling these latent features $\hat{y}_t$ to the original resolution, then concatenates them with the context $\overline{x}_t$ to produce the final reconstructed frame $\hat{x}_t$.

V. Entropy Model

According to [13], the cross-entropy between the estimated probability distribution and the actual distribution of the latent codes represents a tight lower bound of the actual bitrate, which can be expressed as:

$$R(\hat{y}_t) \geq E_{\hat{y}_t \sim q_{\hat{y}_t}}[-log_2 p_{\hat{y}_t}(\hat{y}_t)] \tag{9}$$

In this paper, $p_{\hat{y}_t}(\hat{y}_t)$ and $q_{\hat{y}_t}(\hat{y}_t)$ represent the estimated and actual probability density functions of the quantized $\hat{y}_t$, respectively. Arithmetic coding is employed to encode latent codes at a bitrate closely matching the cross-entropy, resulting in a negligible discrepancy between the actual bitrate $R(\hat{y}_t)$ and the cross-entropy bitrate. The primary goal is to develop an entropy model that precisely estimates the probability distribution $p_{\hat{y}_t}(\hat{y}_t)$.

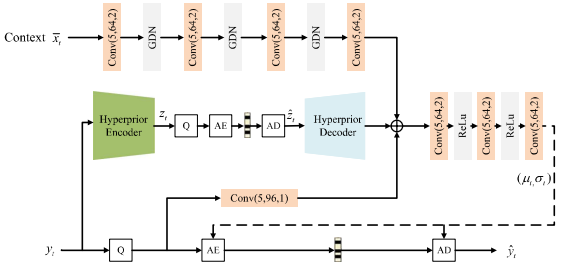


Fig. 5. Structure diagram of entropy model.

Figure 5 illustrates the entropy model's structure. Initially, a hyperprior model [11] learns hierarchical priors within y_t. Next, an autoregressive network with a $mask$ convolutional layer learns spatial priors in y_t's quantized bitstream. While hierarchical and spatial priors are standard in deep image compression, video features also exhibit temporal correlations, necessitating the removal of temporal redundancy. To address this, a temporal prior encoder analyzes the temporal correlations in $\overline{x}_t$, generating temporal priors. Subsequently, a prior fusion network combines these three priors: hierarchical, spatial, and temporal. It estimates the mean and variance of $\hat{y}_t$'s distribution. These estimates assist the arithmetic encoder and decoder in efficient entropy encoding and decoding of y_t. This paper assumes the mean and variance of $p_{\hat{y}_t}(\hat{y}_t)$ follow a Laplacian distribution. The expression for $p_{\hat{y}_t}(\hat{y}_t)$ is:

$$\begin{aligned} p_{\hat{y}_t}(\hat{y}_t \mid \overline{x}_z, \hat{z}_t) &= \prod_i (L(\mu_{t,i}, \sigma_{t,i}^2) * U(-\frac{1}{2}, \frac{1}{2}))(\hat{y}_{t,i}) \\ \mu_{t,i}, \sigma_{t,i} &= f_{pf}(f_{hpd}(\hat{z}_t), f_{ar}(\hat{y}_{t,<i}), f_{tpe}(\overline{x}_t)) \end{aligned} \tag{10}$$

In Equation 10, the index i represents the spatial position. $f_{hpd}()$ is the function form of the hyperprior decoding network, $f_{ar}()$ is the function form of the autoregressive network, $f_{tpe}()$ represents the function of the temporal prior encoder network used for learning the temporal correlations in the contextual information, and $f_{pf}()$ is the function form of the prior fusion network. $f_{ar}(\hat{y}_{t,<i})$ and $f_{tpe}(\overline{x}_t)$ respectively learn the spatial and temporal priors, while $f_{hpd}(\hat{z}_t)$ learns the supplementary information that cannot be derived from spatial or temporal correlations.

VI. Experimental Results and Analysis

A. Comparison and analysis of experimental results

To validate our proposed algorithm, we tested it on three videos from the HEVC Class F dataset, focusing on key foreground targets and overall content. We compared its performance against advanced video compression algorithms, including H.264[1], H.265[2], DVC[3], Hu_ECCV[4], Lu_ECCV[5], FVC[6], DVCpro[7], EA_CVPR[8], and RLVC[9], using PSNR and MS-SSIM metrics. Despite the limited availability of end-to-end video screen content encoding algorithms, we included notable natural image video encoding algorithms in our comparison.

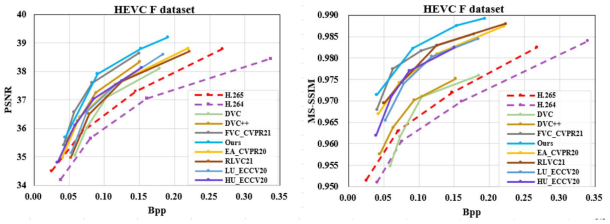


Fig. 6. Algorithm test results.

Our results, illustrated in Figure 6, show the proposed algorithm's superiority on the HEVC Class F dataset. The left side of the figure compares PSNR-bpp rate-distortion curves, while the right side focuses on MS-SSIM-bpp rate-distortion curves. The proposed algorithm outperforms others in both PSNR and MS-SSIM at higher bpp values, especially in MS-SSIM. It maintains this advantage even at lower bpp values, indicating its overall superior performance. Compared to natural image video encoding, our algorithm's PSNR and MS-SSIM values are higher, likely due to screen content characteristics like absence of noise, high contrast, sharp edges, and repetitive areas, which reduce spatial redundancy and complexity, enhancing performance in screen content encoding.

B. Ablation Experiment

To validate the effectiveness of the subnetworks in the deep video compression algorithm proposed in this paper, experimental tests were conducted on the HEVC Class F dataset, focusing on the motion codec network, feature extraction network, and context refinement network. The results of these experiments were then compared with the overall performance of the deep video compression algorithm proposed in this paper, as shown in Figure 7.

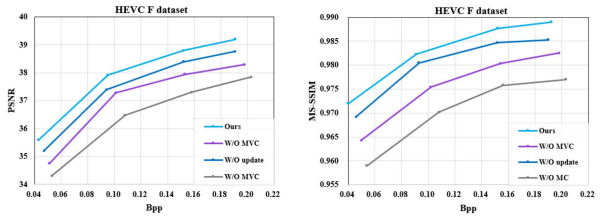


Fig. 7. Comparison of ablation results

Figure 7 displays, on the left, the PSNR-bpp rate-distortion curve comparisons between the proposed algorithm and several ablation algorithms. On the right, it presents the MS-SSIM-bpp rate-distortion curve comparisons. The curve labeled "Ours" represents the test results of the video compression algorithm proposed in this paper.

The curve "W/O CR," shows results without the context refinement network. Using distorted high-dimensional features directly for motion information, it forgoes context refinement, leading to spatial discontinuities. This impacts the context codec network and entropy model, reducing PSNR by about 0.4dB and MS-SSIM by 0.002. The curve "W/O FE," represents results lacking the feature extraction network. It directly employs motion information to distort the reference frame, leaving context in the pixel domain with limited information depth. This approach decreases PSNR by roughly 0.9dB and MS-SSIM by 0.007. The curve "W/O MVC," indicates results without the motion codec network. Foregoing our CNN-based motion codec network for direct motion encoding, this method fails to adequately reduce spatiotemporal redundancy, decreasing PSNR by approximately 1.5dB and MS-SSIM by 0.013. We also tested the efficacy of transform skipping branches in our motion encoding network on the HEVC Class F dataset. These tests, shown in Figure 8, compared configurations with only the main branch, one transform skipping branch, and the two original transform skipping branches from our method.

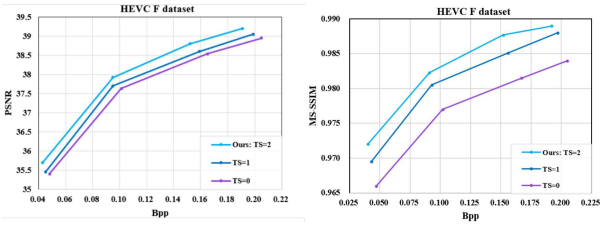


Fig. 8. Transform skip branch verification results

Figure 8 displays rate-distortion curve comparisons: PSNR-bpp on the left and MS-SSIM-bpp on the right. The "TS=1" curve shows results with one transform skipping branch, specifically the one with a convolution stride of 4. This setup slightly reduces performance, with an approximate 0.1dB decrease in PSNR and 0.002 in MS-SSIM. The minor impact suggests that the remaining branch effectively learns larger-scale features. Additionally, the algorithm's multi-layer convolutional networks contribute to learning larger-scale features. The "TS=0" curve, indicating results without transform skipping branches and relying solely on the main branch, shows a more significant reduction in performance: about 0.4dB in PSNR and 0.005 in MS-SSIM. This underlines the importance of transform skipping branches in enhancing screen content encoding.

VII. Conclusion

This paper focuses on the underexplored area of screen content images, distinct from natural video images, by developing a deep contextual screen content encoding network. It specifically addresses screen content characteristics like noiseless patterns, high repetitiveness, high texture contrast, and sharp edges. The network includes a motion encoding network featuring transform skipping, highly effective in compressing screen content motion information. Initially, the paper outlines the network's overall structure. It then describes the functions and principles of its various modules: motion encoding network, feature extraction network, context refinement network, context encoding network, and entropy model. Experimental tests on the HEVC Class F dataset and comparative analyses with other methods demonstrate the proposed algorithm's superior performance in objective evaluation metrics. Ablation studies further affirm the significance of each sub-module, validating the innovation and effectiveness of this approach.

References

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h. 264/avc video coding standard," *IEEE Transactions on circuits and systems for video technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 006–11 015.

[4] Z. Hu, Z. Chen, D. Xu, G. Lu, W. Ouyang, and S. Gu, "Improving deep video compression by resolution-adaptive flow coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 2020, pp. 193–209.

[5] G. Lu, C. Cai, X. Zhang, *et al.*, "Content adaptive and error propagation aware deep video compression," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 2020, pp. 456–472.

[6] Z. Hu, G. Lu, and D. Xu, "Fvc: A new framework towards deep video compression in feature space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1502–1511.

[7] G. Lu, X. Zhang, W. Ouyang, L. Chen, Z. Gao, and D. Xu, "An end-to-end learning framework for video compression," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3292–3308, 2020.

[8] E. Agustsson, D. Minnen, N. Johnston, J. Balle, S. J. Hwang, and G. Toderici, "Scale-space flow for end-to-end optimized video compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8503–8512.

[9] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, "Learning for video compression with recurrent auto-encoder and recurrent probability model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 388–401, 2020.

[10] M. Mrak and J.-Z. Xu, "Improving screen content coding in hevc by transform skipping," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, IEEE, 2012, pp. 1209–1213.

[11] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[12] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[13] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55, 2001.

A Coarse-to-Fine Change Detection Framework for Remote Sensing Sparse Cultivated Land

Yuan Hu, Yifan Zhang, Mingyang Ma and Shaohui Mei
Northwestern Polytechnical University, Xi'an 710072, P.R. China
E-mail: huyuan1999hy@mail.nwpu.edu.cn, yifanzhang@nwpu.edu.cn, mamingyang@mail.nwpu.edu.cn, meish@nwpu.edu.cn

***Abstract*—Remote sensing (RS) images contain rich geographic information. For specific application scenarios like cultivated land, it is necessary to select areas of interest to reduce data scale and focus on detailed features. In this article, an innovative coarse-to-fine change detection framework (CFCD) for sparse cultivated land is proposed to address these problems. Coarse screening module (CSM) first removes irrelevant low-difference image pairs, and then fine detection module (FDM) accurately locate change areas in remaining images. Experimental results show that two coarse screening methods can take out many disturbed images, and provide strong support for subsequent fine detection methods to achieve performance improvement.**

I. Introduction

In recent years, the rapid development of RS technology has driven the accumulation of relevant data, which provides strong support for the diversified application of Remote Sensing (RS) images. Taking change detection (CD) for example, its core idea is to accurately identify and analyze changes in the surface cover by jointly comparing two (or more) images acquired on the same geographical area at different times [1]. At present, it is widely utilized in various fields such as deforestation investigation [2], urban planning [3], disaster assessment [4] and other fields, demonstrating extremely high practical value and application prospects.

As the scarce land resource in the world, cultivated land not only serves as the lifeblood and main carrier for agricultural production activities, but also functions as the core factor in promoting food production and regional sustainable development [5]. The amount of adequate cultivated land is crucial for maintaining food security and fostering harmonious coexistence between humanity and nature. However, the ecological degradation caused by the change of cultivated land has become a prominent global issue [6]–[8]. On one hand, the distribution of cultivated land is deeply affected by natural conditions. On the other hand, the expansion of industrialization and urbanization has led to continuous compression and cutting of limited cultivated land, deepening the degree of sparsity and fragmentation. Therefore, it is urgent to implement protection policies for cultivated land.

How to quickly detect changes in cultivated land is a key step in effectively protecting the cultivated land. Under the background that the RS data acquired by a single satellite data center is dramatically increasing at a speed of several terabytes per day, traditional manual field investigation methods, as well as certain automated approaches, cannot meet the demands for direct real-time monitoring of large-scale images [9]. As the convolutional neural network (CNN) is introduced to CD, deep learning methods for cultivated land have attracted more attention from researchers. Liu *et al.* [10] designed a CNN-transformer network with multiscale context aggregation (MSCANet), which exploits context aggregate connections to fuse and aggregate features across different levels, fulfilling efficient and effective cultivated land CD. To address the problem of inadequate utilization of feature information, Miao *et al.* [11] proposed a Siamese network based on full-scale connected UNet (SNUNet3+), which combines the spatial and channel squeeze and excitation (scSE) attention mechanism and deep supervision modules to detect changes in cultivated land. In [12], A transformer-based multiscale feature fusion change detection network (M-Swin), which can capture the change information in small building through hierarchical windows and integrate the multiscale feature obtained from different windows to cope with the "scale gap" challenge.

Even though the existing mainstream methods have made better detection performance, they heavily rely on high-quality and well-produced standardized datasets, failing to fully consider the common problem of "non-focused change areas" in original large-sized RS images, which are far from practical application. In addition, the sparse characteristics of cultivated land also need to be further excavated.

In order to deal with the above-mentioned problems, a coarse-to-fine change detection (CFCD) framework is proposed in this paper for sparse cultivated land. The overall idea of this method is "coarse screening-fine detection". Firstly, differences between the two original bitemporal image pairs are analyzed and quantified through the method of coarse feature representation, and the image pairs with relatively small differences are discarded. Then the retained image pairs are finely detected by employing deep learning methods based on Siamese network. Finally, the patches of cultivated land change areas are located. The major contributions of this article are listed as follows.

- Two coarse feature representation methods are proposed for cultivated land CD, which aim to quickly identify areas with obvious changes and reduce the amount of data on subsequent processing.
- An efficient coarse screening-fine detection method is designed for cultivated land CD task based on large scale images. The proposed method can directly process

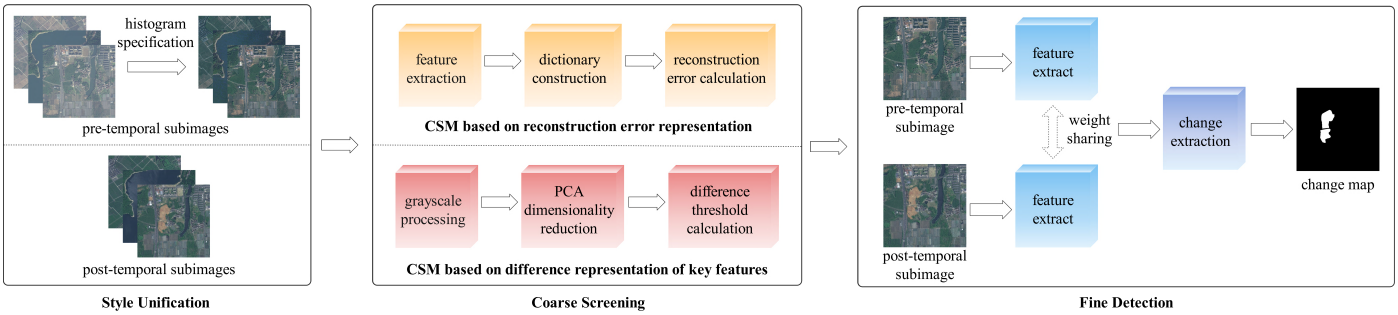


Fig. 1. Framework of the proposed CFCD

original images and significantly improve detection accuracy before being used in existing state-of-the-art (SOTA) methods.

- Extensive comparison and experiments with four SOTA CD models on ShangYu CLCD Datasets are conducted, which have proved the effectiveness of the proposed method.

II. Methodology

Except for the necessary preprocessing such as radiometric correction and geometric registration, which can truly reflect changes in land surface conditions, the original large-sized bitemporal images are directly cropped into unified small-sized subimages without any additional steps to form an image pair.

The proposed CFCD framework in this paper can rapidly remove low difference image pairs through coarse screening module (CSM), and extract the specific change areas from the remaining image pairs using fine detection module (FDM), so as to generate the final detection results. Framework of the proposed CFCD is shown in Fig 1.

A. Coarse Screening Module (CSM)

In large-scale RS images, the area of cultivated land changes usually occupies only a small part of them, and is sparsely distributed. The large presence of non-research areas not only consumes computing resources but also introduces noise, which greatly affects the accuracy of CD. To settle these problems, two coarse screening methods are designed in CSM modules of the proposed CFCD.

1) CSM based on reconstruction error representation (CSM-RER): Compressive Sensing (CS) theory manifests that even if only a small number of measured values are obtained, the original image can be effectively restored and reconstructed from sparse or sparsely represented images through specific measurement matrices and optimization algorithms [13], [14]. The proposed CSM based on reconstruction error representation adopts feature extraction algorithm to extract the same N-dimensional features from all image pairs. The features of the i-th post temporal subimage $\mathbf{y_i}$ is selected as the object to be reconstructed, and the corresponding features of pre-temporal subimage and neighboring subimages are selected to construct a dictionary $\mathbf{D_i}$ for sparse reconstruction, as shown in Fig 2.

The reconstruction error e_i is calculated using

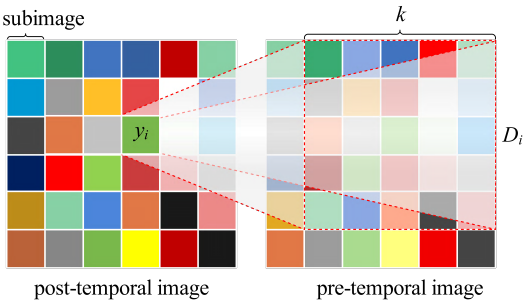


Fig. 2. CSM based on reconstruction error representation

$$e_i = \frac{\|\mathbf{y_i} - \mathbf{D_i} \cdot \mathbf{x_i}\|_2}{\|\mathbf{y_i}\|_2} < \varepsilon \tag{1}$$

where $\|\cdot\|_2$ represents the ℓ_2-norm, $\mathbf{y_i} \in \mathbb{R}^{N\times 1}$, $\mathbf{D_i} \in \mathbb{R}^{N\times M}$, ε is the error threshold. The number of words in the dictionary $M = k^2 \leq K$, with K the number of subimages cropped from each original image. Greater error indicates more likely changes, that is, the image pairs with smaller values can be discarded by sorting according to the reconstruction error.

2) CSM based on difference representation of key features (CSM-DRKF): Principal component analysis (PCA) is a technique that reduces the dimensionality of original data while maximizing the preservation of its intrinsic information [15]. As shown in Fig 3, grayscale processing is first performed on all subimages and two matrices (i.e. $\mathbf{A}$, $\mathbf{B}$) are composed of pre-temporal and post-temporal images. Then, to highlight the key information and eliminate the disturbing information, PCA is used to reduce the dimensionality of the two image matrices to N-dimension. Finally, the differences are calculated and sorted based on (2)-(4), and the image pairs with small differences are removed.

$$\mathbf{C} = \sum_{i=1}^{N}\sum_{j=1}^{K} c_{ij} = \mathbf{A} - \mathbf{B} \tag{2}$$

$$\mathbf{D} = \left(\sum_{i=1}^{N}|c_{i1}|, \sum_{i=1}^{N}|c_{i2}|, \cdots, \sum_{i=1}^{N}|c_{ij}|\right) \tag{3}$$

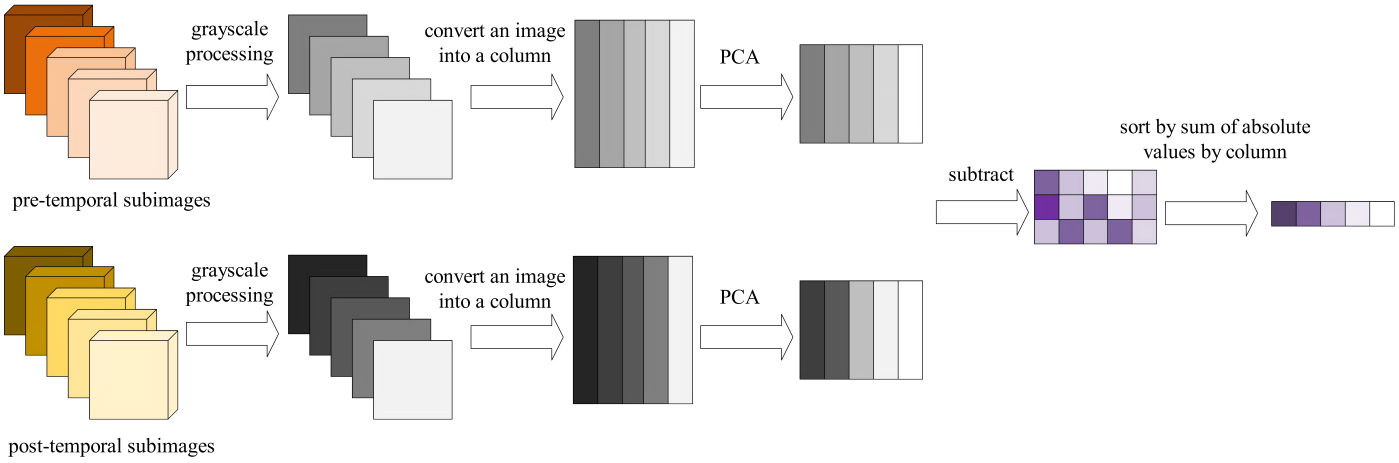


Fig. 3. CSM based on difference representation of key features

$$d_k = \sum_{i=1}^{N} |c_{ij}| < \delta \tag{4}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{N \times K}$, $\mathbf{D} = (d_1, d_2, \cdots, d_k) \in \mathbb{R}^{1 \times K}$, δ is the difference threshold.

B. Fine Detection Module (FDM)

Siamese network is a deep learning architecture used to solve the task based on similarity comparison, consisting of two or more identical subnetworks that share weights and parameters [16], as shown in Fig 4.

Each sub network receives an input sample and extracts features through a series of convolutional layers (or other types of layers). At present, many mainstream CD methods are proposed based on Siamese network structure, which have exhibited powerful feature extraction and comparison capabilities. In FDM of the proposed CFCD framework, the following four typical methods can be incorporated.

FC-Siam-conc [20] is a feature-concatenation method, which extracts multi-level features of bi-temporal images from shared Siam structures, and concatenate them to detect changes. The skipping feature connection can capture subtle changes with richer feature expressions, which is particularly important for CD in cultivated land.

Unlike FC-Siam-conc, FC-Siam-diff [20] is a feature-difference method, which extracts features by using the Siamese connection and continuously superimposes residuals during deconvolution to achieve favorable CD. This approach focuses on capturing different information between images, which can reduce the pseudo-changes caused by non-substantive factors such as light and seasonal changes to a certain extent, and is more sensitive to identification in change areas.

BIT [21] is a feature fusion method, which integrates Siamese tokenizer and transformer encoder-decoder structure to achieve more meaningful context-information to obtain the change map. By introducing transformer structure, it is possible to globally obtain the dependency relationships in RS images, which is helpful to identify the complex patterns of cultivated land changes.

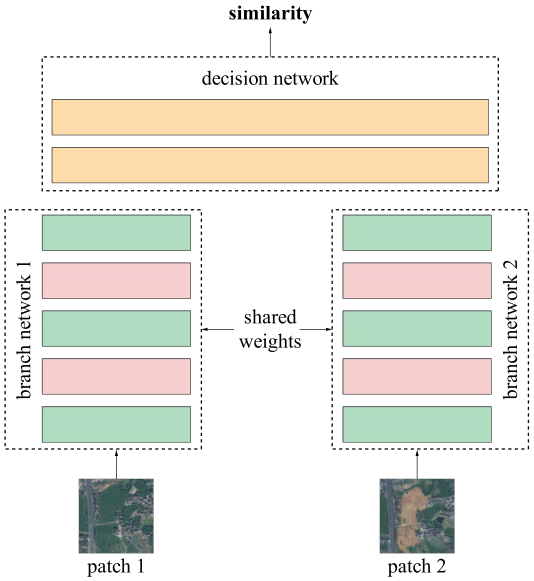


Fig. 4. Siamese network structure

DTCDSCN [22] is an attention-based method, which utilizes change information in spatial and channel information to extract more contextual features. It combines two tasks: CD and semantic segmentation. Through the shared feature extraction layer, the network can extract the features of cultivated land changes and the semantic information of cultivated land at the same time, which is obviously helpful for subtle change detection in cultivated land.

III. Experiments and Analysis

A. ShangYu CLCD Datasets

Many publicly datasets have been used for CD, typically including various typical scenarios such as urban areas, plains,

TABLE I
EXPERIMENTAL RESULTS ON CSM BASED ON RECONSTRUCTION ERROR REPRESENTATION (CSM-RER)

Dataset	Number of Images (K)	Number of Changes	Number of Words (k^2)	Error Threshold (ε)	Images Removal Rate
ShangYu-GF1b	5520	142	3×3	7.5264×10^{-2}	36.11%
			5×5	6.4046×10^{-2}	36.59%
			7×7	5.6378×10^{-2}	**37.07%**
			9×9	5.0041×10^{-2}	36.01%
ShangYu-BJ2	1596	168	3×3	5.5320×10^{-2}	15.60%
			5×5	5.2397×10^{-2}	15.66%
			7×7	4.8918×10^{-2}	15.73%
			9×9	4.6291×10^{-2}	**15.79%**

TABLE II
EXPERIMENTAL RESULTS ON CSM BASED ON DIFFERENCE REPRESENTATION OF KEY FEATURES (CSM-DRKF)

Dataset	Number of Images (K)	Number of Changes	Dimension (N)	Difference Threshold (δ)	Images Removal Rate
ShangYu-GF1b	5520	142	256	79.7307	45.65%
			512	123.2878	**46.59%**
			1024	187.5146	46.20%
ShangYu-BJ2	1596	168	256	74.2009	18.98%
			512	112.6452	19.49%
			1024	187.1340	**19.80%**

TABLE III
EXPERIMENTAL RESULTS ON FDM AND CFCD

Method	FDM					CFCD (FDM + CSM-DRKF)				
	Precision	Recall	F1	IoU	OA	Precision	Recall	F1	IoU	OA
FC-Siam-conc	40.71	46.00	58.58	37.39	63.69	76.88	63.98	71.92	56.14	82.10
FC-Siam-diff	52.07	50.24	42.80	44.21	63.77	69.51	44.99	56.32	49.20	75.28
BIT	52.62	52.98	53.66	46.24	64.02	72.93	43.33	54.36	47.33	77.13
DTCDSCN	32.15	43.29	41.78	38.27	52.03	64.06	49.05	55.34	43.03	76.73

cultivated land, and mountainous areas. However, the availability of datasets for cultivated land remains extremely restricted, greatly limiting the innovative application of CD methods. To address this issue, two large-scale optical remote sensing images of GF-1b and BJ-2 taken by Zhejiang Institute of Surveying and Mapping Science and Technology are obtained as the new dataset for detecting changes in cultivated land, with the sizes of 40598×35178 and 21243×19041 pixels, respectively, and spatial resolution ranged from 0.8 to 2 m. These images were uniformly cropped to 512×512 pixels, resulting in a total of 7116 image pairs (i.e. ShangYu-GF1b obtains 5520 pairs, ShangYu-BJ2 obtains 1596 pairs), forming ShangYu CLCD Datasets. The types of change targets mainly include buildings and structures. These samples are randomly separated for training, test and validation in the ratio of 6:2:2.

B. Implementation Details and Evaluation Metrics

Obvious stylistic differences exist between pre-temporal and post-temporal images of datasets, caused by different lighting and atmospheric conditions during images shooting. To reduce its impact on the accuracy of subsequent CD, histogram specification is used to unify the style [17]–[19].

The purpose of coarse screening is to exclude as many images as possible there are no changes, without affecting the precision of fine detection. Therefore, the effectiveness of two methods is measured with zero missed detection. This shows that error threshold (ε) and difference threshold (δ) are based on the image with the first missed detection, meaning it is less than this value and the change is negligible. As

for CSM based on reconstruction error representation (CSM-RER), feature extraction algorithm is applied to each subimage to extract 360 features (i.e. $N = 360$). And k is set to be 3, 5, 7, 9 respectively.

As for CSM based on difference representation of key features (CSM-DRKF), PCA is applied to each subimage and N is set to be 256, 512, 1024 respectively.

Five evaluation metrics are employed for quantitative assessment of the proposed CD method, including precision (Pre), recall (Rec), F1-score (F1), intersection over union (IoU), and overall accuracy (OA). They range from 0 to 1, and higher values indicate better performance.

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$F1 = \frac{2\mathrm{Pre} \times \mathrm{Rec}}{\mathrm{Pre} + \mathrm{Rec}} \tag{7}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{8}$$

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{9}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

C. Comparative Experiments

1) Comparison Experiments on Different CSMs: Table I and II lists evaluated metrics for different values of k and N on different CSMs. With the increasing of k or N, images removal rate shows an upward trend indicating that the better the quality of the processed dataset, the more obvious the removal effect of non-relevant changing images. In ShangYu-GF1b dataset, images removal rate reaches its maximum as 37.07% and 46.59% when k=7 or N=512, respectively. In ShangYu-BJ2 dataset, when k=9 or N=1024, images removal rate is 15.79% and 19.80%. Generally, both CSM-DRKF and CSM-RER can achieve better screening effects. In ShangYu CLCD Dataset, CSM-DRKF outperform CSM-RER as it can remove more invalid images.

2) Comparison Experiments on FDM and CFCD: Based on the results of CSM-DRKF, all images retained in ShangYu-GF1b and ShangYu-BJ2 datasets are combined to form new ShangYu CLCD Dataset, and comparative experiments of FDM are conducted. Most of the existing FDM methods directly detect images in the dataset, but the proposed CFCD architecture first coarsely screens images and then performs fine detection through any CD method, which reduces computational complexity and significantly improves accuracy. Table III illustrates quantitative evaluations of the proposed CFCD architecture employing four different SOTA CD methods. It can be observed that, compared to the results of ShangYu CLCD Datasets, the proposed CFCD framework can achieve better evaluation metrics in CD performance, with significant improvements in all five metrics. For example, the F1/IoU/OA of the proposed CFCD framework exceeds previous SOTA by 41.9/33.4/36.6% for ShangYu CLCD Datasets, respectively. It can also be observed from the Table III, the Precision, Recall, IoU, and OA values of traditional FC-Siam-conc method are lower than BIT. While the evaluation metrics of the proposed CFCD framework have been significantly improved because of the preprocessing procedure, exceeding those of BIT, which is 31.56/17.19/17.63/22.02%, respectively. In addition, although the ability of DTCDSCN to detect changes performed the worst on ShangYu CLCD Dataset, Precision, F1 and OA indicators have significantly improved after integration into the CFCD framework. These quantitative and qualitative comparisons indicate the superiority of the proposed CFCD framework in improving the detection performance of existing SOTA methods.

IV. Conclusions

In this paper, a CFCD framework based on dual modules (CSM and FDM) is proposed, which makes it possible to utilize the original RS images directly for detection. In CSM, a variety of non-cultivated land changes are quickly identified and screened, which greatly reduces the amount of data. In FDM, deep learning methods based on Siamese network are adopted which can specifically extract and locate the attention areas, to achieving the final CD results. Experimental results illustrate the effectiveness and competitiveness of the proposed CFCD framework. In future work, sparse feature extraction is also a direction worth further exploring.

Acknowledgment

This work was supported in part by the Zhejiang Province "Pioneering Soldier" and "Leading Goose" Research and Development Project under Grant 2023C01027.

References

[1] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 609–630, 2012.

[2] T. Hame, I. Heiler, and J. San Miguel-Ayanz, "An unsupervised change detection and recognition system for forestry," *International journal of remote sensing*, vol. 19, no. 6, pp. 1079–1099, 1998.

[3] L. Yang, G. Xian, J. M. Klaver, and B. Deal, "Urban land-cover change detection through sub-pixel imperviousness mapping using remotely sensed data," *Photogrammetric Engineering & Remote Sensing*, vol. 69, no. 9, pp. 1003–1010, 2003.

[4] D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using vhr optical and sar imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2403–2420, 2010.

[5] G. Jiang, M. Wang, Y. Qu, D. Zhou, and W. Ma, "Towards cultivated land multifunction assessment in china: Applying the "influencing factors-functions-products-demands" integrated framework," *Land Use Policy*, vol. 99, p. 104 982, 2020.

[6] W. B. Meyer and B. L. Turner, "Human population growth and global land-use/cover change," *Annual review of ecology and systematics*, vol. 23, no. 1, pp. 39–61, 1992.

[7] R. Lal, "Restoring soil quality to mitigate soil degradation," *Sustainability*, vol. 7, no. 5, pp. 5875–5895, 2015.

[8] H. Xie, Y. Zhang, Z. Wu, and T. Lv, "A bibliometric analysis on land degradation: Current status, development, and future directions," *Land*, vol. 9, no. 1, p. 28, 2020.

[9] Y. Ma, H. Wu, L. Wang, *et al.*, "Remote sensing big data computing: Challenges and opportunities," *Future Generation Computer Systems*, vol. 51, pp. 47–60, 2015.

[10] M. Liu, Z. Chai, H. Deng, and R. Liu, "A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4297–4306, 2022.

[11] L. Miao, X. Li, X. Zhou, *et al.*, "Snunet3+: A full-scale connected siamese network and a dataset for cultivated land change detection in high-resolution remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[12] J. Pan, Y. Bai, Q. Shu, Z. Zhang, J. Hu, and M. Wang, "M-swin: Transformer-based multi-scale feature fusion change detection network within cropland for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[13] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[14] Y. Tsaig and D. L. Donoho, "Extensions of compressed sensing," *Signal processing*, vol. 86, no. 3, pp. 549–571, 2006.

[15] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[16] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[17] A. K. Jain, *Fundamentals of digital image processing*. Prentice-Hall, Inc., 1989.

[18] K. R. Castleman, *Digital image processing*. Prentice Hall Press, 1996.

[19] D. Coltuc, P. Bolon, and J.-M. Chassery, "Exact histogram specification," *IEEE Transactions on Image processing*, vol. 15, no. 5, pp. 1143–1152, 2006.

[20] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, Ieee, 2018, pp. 2115–2118.

[21] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[22] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020.

Efficient Adaptation for Real-World Omnidirectional Image Super-Resolution

Cuixin Yang*, Rongkang Dong* and Kin-Man Lam*
* The Hong Kong Polytechnic University, Hong Kong
E-mail: cuixin.yang@connect.polyu.hk, rongkang97.dong@connect.polyu.hk, enkmlam@polyu.edu.hk

***Abstract*—With the increasing popularity of virtual techniques, such as virtual reality (VR) and augmented reality (AR), super-resolution (SR) of omnidirectional images has been crucial for more immersive and realistic experiences. This advancement also enhances the quality of images for various visual applications. Researchers have started exploring omnidirectional image super-resolution (ODISR). However, existing methods primarily address the problem using synthetic data pairs, where low-resolution (LR) images are generated using fixed, predefined kernels, such as bicubic downsampling. Consequently, the performance of these methods drops significantly when applied to real-world data. To address this issue, in this paper, we propose exploring the rich image priors from existing SR models designed for 2D planar images and adapting them for real-world ODISR. Specifically, we employ low-rank adaptation (LoRA) to adapt a large-scale model from the 2D planar image domain to the omnidirectional image domain by training only the decomposed matrices. This approach significantly reduces the number of parameters and computational resources required. Experimental results demonstrate that the proposed method outperforms other state-of-the-art methods both quantitatively and qualitatively.**

I. Introduction

With the development of virtual reality (VR) and augmented reality (AR), omnidirectional images (ODIs) have become increasingly prevalent in our daily lives, providing immersive virtual scenes. High-resolution (HR) ODIs are essential for these virtual scenarios, because they provide richer details and textures. However, it is costly to capture, store, and transmit HR ODIs [1]. Image super-resolution (SR) is a promising technique to address this challenge, aiming to reconstruct HR images from their low-resolution (LR) counterparts [2].

Recently, with the development of deep-learning techniques, significant progress has been made in many computer vision tasks [3], [4], including the field of omnidirectional image super-resolution (ODISR) [1], [5]–[10]. Despite the remarkable advancements in ODISR, existing methods still have several limitations. A major drawback is their focus on synthetic images [11], where LR images are generated from the HR images using fixed or predefined degradation methods, such as bicubic downsampling. However, real-world omnidirectional images suffer from complex and often unknown degradations, such as noise, blurriness, and downsampling [12]. Consequently, methods trained on synthetic images perform poorly in real-world scenarios, due to the substantial domain gap between synthetic training images and real-world test images. In the field of 2D planar SR, numerous works [13], [14] have addressed this problem by combining and mixing several degradations to better simulate real-world conditions. BSRGAN [13] and Real-ESRGAN [14] employ GAN losses to make the super-resolved output more realistic in the training process. However, GAN-based methods can be unstable, leading to undesirable artifacts in the generated images [15]. Recently, generative diffusion models, which utilize a Markov chain with hundreds of inference steps, have demonstrated strong representation abilities and have dominated many computer vision tasks, including real-world image SR. Some methods [16]–[19] have proposed leveraging the powerful capabilities of pretrained large-scale text-to-image (T2I) models to provide image priors. However, these diffusion-based methods have primarily been developed for 2D planar images. To date, limited research has been conducted in the field of ODISR to utilize large-scale pretrained generative models for the reconstruction of ODIs.

To address these issues, we propose leveraging a well-trained diffusion-based model, originally trained on a real-world 2D planar dataset, to exploit the rich image priors for the real-world ODISR problem. Specifically, to bridge the gap between 2D image SR and ODISR, we adopt an advanced parameter-efficient fine-tuning adapter, e.g., LoRA [20], to adapt the large-scale pretrained 2D planar image SR model to the ODISR model. LoRA is a low-rank decomposition method that freezes the weights of pretrained model and decomposes the weights of Transformer layers into trainable low-rank matrices, significantly reducing the number of trainable parameters. This allows us to efficiently transfer the well-trained knowledge from the 2D image domain to the ODI domain with a minimal number of trainable parameters.

The main contributions of this paper are as follows:

- Unlike previous works that focus on synthetic omnidirectional image super-resolution (ODISR), we investigate the challenge of real-world ODISR, where images are subject to multiple degradations, such as noise, blur, jpeg compression, and downsampling.
- We propose adopting the LoRA adapter to adapt a large-scale pretrained model from the 2D image domain to the ODI domain. By fine-tuning only the decomposed weight matrices of the Transformer layers, this approach significantly reduces the number of trainable parameters and training time compared to the original large-scale pretrained model.
- Experimental results demonstrate the effectiveness and superiority of the proposed method quantitatively and qualitatively.

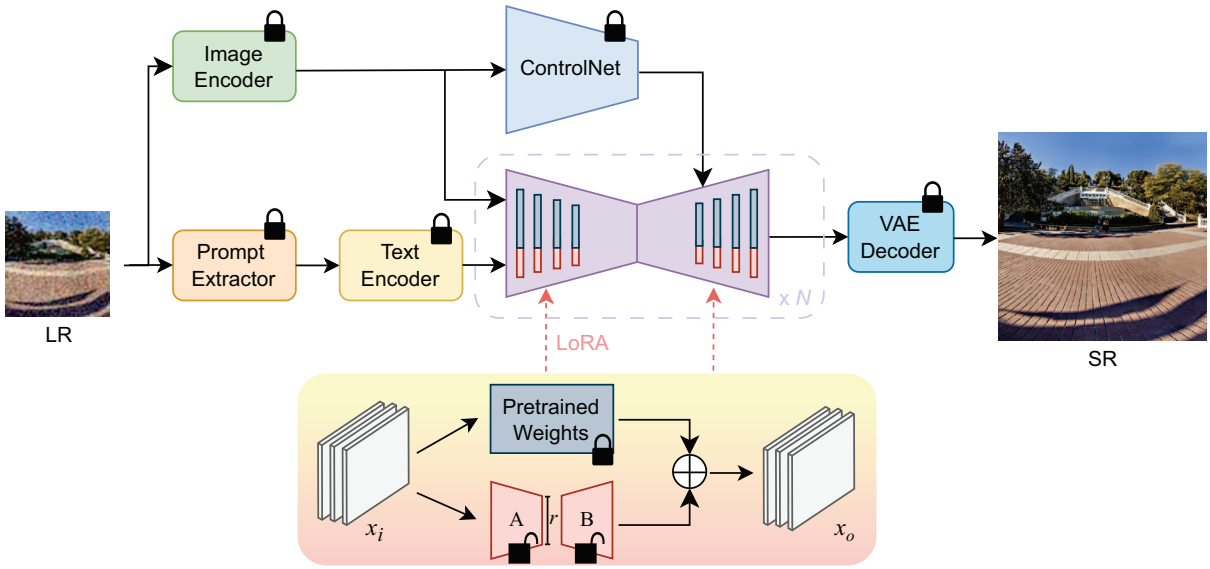


Fig. 1. The overview pipeline of the proposed method for real-world omnidirectional image super-resolution. Only A and B, which are low-rank decomposed matrices of the pretrained weight matrix, are trainable.

II. Related Works

A. Omnidirectional Image Super-Resolution (ODISR)

Omnidirectional images are projected as 2D equirectangular (ERP) images for image processing. LAU-Net [7] is a progressive pyramid network to highlight the non-uniform pixel density across latitudes in ERP images. However, training multiple network levels for different latitude bands is computationally expensive and can result in inconsistencies between the bands. Inspired by LIIF [21], SphereSR [9] creates a continuous spherical image representation to predict RGB values across various projection types. Although SphereSR can flexibly resolve ODIs with arbitrary projection types, it requires training multiple network branches for different projections. OSRT [47] utilizes Fisheye downsampling, applying uniform bicubic downsampling on the original ODIs. OSRT [1] was proposed to apply fisheye downsampling on the ODIs to perform uniform bicubic downsampling on the original ODIs. In addition, OSRT also proposes to adopt a deformable attention mechanism to make full use of the distortion map in the reconstruction. GDGT-OSR [10] proposes to utilize the rectangle-window-based transformer [22] for better adapting to the distortion of ERP images, and introduce the distortion-guided mechanism to modulate the attention area.

However, the abovementioned methods primarily focus on synthetic ODIs, which struggle to perform well on real-world data. In this paper, we explore real-world ODISR to address this problem.

B. Generative Models for Real-World Image Super-Resolution

In recent years, researchers have employed generative models, such as GAN [23] and diffusion networks [24], in the field of real-world image SR. SRGAN [25] was the first to adopt the GAN loss [23] in the SR training process to generate photo-realistic images. BSRGAN [13] and Real-ESRGAN [14] randomly shuffle the degradation of an image to generate realistic training pairs. However, training GAN models is unstable, and GAN models can easily generate unnatural visual artifacts. Recently, researchers have started exploring the potential of more powerful pretrained text-to-image models, such as Stable Diffusion [26], for solving real-world image SR problems. StableSR [16] employed a time-aware encoder and a feature warping module to balance quality and fidelity. PASD [17] feeds both low-level and high-level features into the pretrained Stable Diffusion model with a pixel-aware cross-attention module. SeeSR [19] takes both text prompts and image embedding as input to improve generation performance.

Despite this advancement, limited research has been conducted in real-world ODISR. In this paper, we aim to utilize the rich image priors provided by large-scale pretrained models to bridge the gap between 2D planar images and ODIs.

III. Methodology

A. Framework Overview

In this paper, we leverage a pretrained large-scale model that exploits both text and image representations, e.g., SeeSR [19], as our backbone. The overview pipeline, depicted in Fig. 1, consists of a prompt extractor, text encoder, image encoder, stable diffusion (SD) with ControlNet, and the LoRA adapter. During the training phase, all modules except the LoRA adapter are frozen. Specifically, we only train the decomposed matrices, e.g., A and B in the LoRA module, which are the low-rank matrices of the pretrained weight matrix. We applied LoRA to query, key, value and output projection matrices in the self-attention module, i.e., W_q, W_k, W_v, W_o.

B. LoRA for Real-World Omnidirectional Image Super-Resolution

Low-rank adaptation (LoRA) [20] is a parameter-efficient training technique that inserts a small number of trainable weights instead of fine-tuning all of the model's parameters. Based on the theory that pretrained models have a low "intrinsic rank" when adapting to a downstream task [27], LoRA posits that the updates to the model weights, i.e., ΔW, still have a low "intrinsic rank" and can be projected into a small space. The rank of a matrix represents the maximum number of linearly independent vectors (rows or columns) in it. When adapting a pretrained model to the downstream task, ΔW can be approximated by a small number of linearly independent vectors.

The bottom of Fig. 1 illustrates the mechanism of weight parametrization using LoRA. When fine-tuning the pretrained large-scale model, a large weight matrix $W \in \mathcal{R}^{m \times n}$ are decomposed into two trainable low-rank matrices, i.e., $A \in \mathcal{R}^{r \times n}$ and $B \in \mathcal{R}^{m \times r}$, where the rank $r \ll min(m, n)$. The pretrained weights are kept unchanged. Then, the dot product of the two low-rank matrices is added to the freezed pretrained weight matrix. The forward pass with LoRA can be expressed as follows:

$$x_o = Wx_i + \Delta W x_i = Wx_i + BAx_i, \tag{1}$$

where x_i and x_o are input and output features, respectively.



Fig. 2. An example of an equirectangular projection image.

In this paper, we address the problem of real-world ODISR, which is still a relatively unexplored area. In SR, omnidirectional images (ODIs) are typically projected as equirectangular projection (ERP) images for easier processing. Fig. 2 shows an example of an ERP image, which exhibits various geometric distortions due to the projection from a sphere surface (ODI) to a 2D plane (ERP image). These distortions vary across different latitudes, making ERP images distinct from common 2D planar images.

Significant advancements have been made in the field of real-world 2D image SR [30]–[33], particularly with algorithms utilizing recent large-scale models [16]–[19]. While these advanced real-world SR models provide rich image priors, directly applying 2D image SR models to ODIs is not advisable due to the domain gap between these two image types. One potential solution is to fine-tune large-scale real-world 2D image SR models. Considering the vast number of parameters in state-of-the-art (SOTA) large-scale models, fully fine-tuning these models from scratch is inefficient and requires substantial computational resources. To address these issues, we propose using the parameter-efficient adapter LoRA to adapt the 2D planar image SR model for ODIs. This approach not only maintains well-trained features for real-world image SR, but also facilitates adaptation for the ODI domain by training only a small number of parameters. Notably, the LoRA adapter is applied to the query, key, value, and output projection matrices in the self-attention module.

IV. Experiments

A. Datasets

We adopt the ODI-SR dataset [7], which consists of 800 HR images with a resolution of 1024×2048, for training. For testing, we use the ODI-SR testing dataset and the SUN 360 Panorama dataset [34]. Both testing datasets contain 100 HR images with a resolution of 1024×2048. Due to the lack of real-world ODI datasets, we follow the degradation pipeline of Real-ESRGAN [14] to synthesize LR-HR training pairs from these public datasets. The degradation involves Gaussian noise, Poisson noise, blur, jpeg compression, and downsampling. The scaling factor used in our experiments is 4. During training, the training pairs are cropped into patches, with the HR image patch size being 512×512.

B. Experimental Settings

We conducted all experiments using PyTorch [35]. For the LoRA adapter, we set the rank of the A and B matrices to 4. The learning rate was set to 5×10^{-5}, and the batch size was 4. We use the constant learning rate schedule as SeeSR [19] during the training process. The total number of training iterations 50,000. We used AdamW [36] as the optimizer with $\beta 1 = 0.9$ and $\beta 2 = 0.999$, and the weight decay is 0.01. We ran the experiments on an Nvidia GeForce RTX 3090 GPU, and the model was trained for approximately 2 days.

C. Experimental Results

1) Quantitative Comparison with State-of-the-Art Methods: Table I shows the quantitative results of different methods on the ODI-SR and SUN 360 Panorama test datasets. We compare our method with other SOTA SR methods, such as ESRGAN [28], SwinIR [29], PASD [17], and OSRT [1]. Notably, ESRGAN [28], SwinIR [29] and PASD [17] are SR methods for 2D planar images, while OSRT [1] is a SOTA method for ODISR. In addition, SwinIR [29] and OSRT [1] are transformer-based methods, while ESRGAN [28] and PASD [17] are photo-realistic SR methods based on generative models. The backbone of our method is the model proposed in SeeSR [19], which is also based on the generative diffusion model. We use multiple evaluation metrics to evaluate the performance of different methods. PSNR and SSIM are common metrics for evaluating SR methods. NIQE [37] is a no-reference image quality evaluation metric. FID [38] is

TABLE I

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE ODI-SR AND SUN 360 PANORAMA DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE SCALING FACTOR IS 4.

Methods	ODI-SR					SUN 360 Panorama				
	PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓	PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓
Bicubic	21.22	0.5128	9.4243	0.6264	-	21.2177	0.5128	9.4243	0.6218	-
ESRGAN [28]	20.92	0.4871	8.8908	0.6299	116.97	20.78	0.4744	8.8908	0.6262	143.78
SwinIR [29]	20.87	0.4852	8.7825	0.6307	-	20.74	0.4722	8.789	0.6268	-
PASD [17]	21.18	0.5727	**4.6861**	0.4602	77.43	21.27	0.5783	**4.6471**	0.4593	75.52
OSRT [1]	19.70	0.4389	7.9834	0.6345	-	19.12	0.4213	7.7438	0.8153	-
Ours	**21.51**	**0.5903**	5.4967	**0.4182**	**70.29**	**21.77**	**0.5989**	5.3447	**0.4307**	**66.02**

TABLE II

COMPARISON RESULTS ON DIFFERENT RANKS IN LORA. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE SCALING FACTOR IS 4.

Rank	Params	ODI-SR					SUN 360 Panorama				
		PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓	PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓
r=4	1.59M	**21.51**	**0.5903**	**5.4967**	**0.4185**	70.29	**21.77**	**0.5989**	**5.3447**	0.4307	66.02
r=8	3.18M	**21.51**	**0.5903**	**5.4967**	**0.4185**	**70.27**	21.69	0.5965	5.3657	**0.4181**	**65.4257**

TABLE III

QUANTITATIVE RESULTS WITH DIFFERENT TRAINING DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE SCALING FACTOR IS 4.

Finetune Dataset	ODI-SR					SUN 360 Panorama				
	PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓	PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓
DIV2K	21.49	**0.5903**	5.53	**0.4174**	**68.98**	21.74	**0.6009**	5.3904	0.4313	67.19
ODI-SR	**21.51**	**0.5903**	**5.4967**	0.4182	70.29	**21.77**	0.5989	**5.3447**	**0.4307**	**66.02**

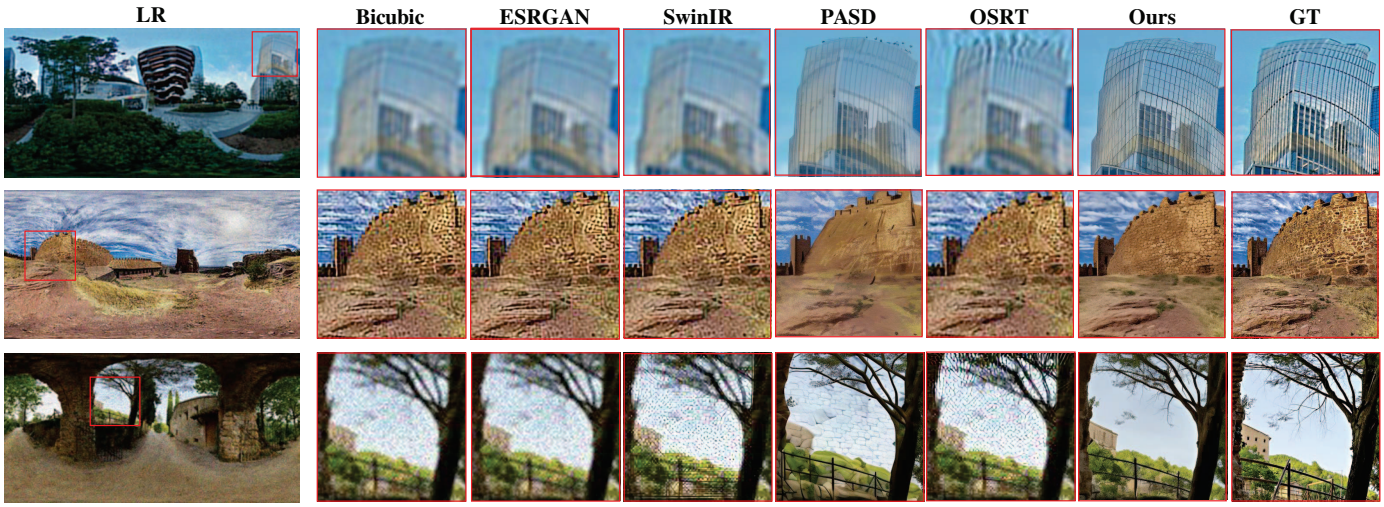


Fig. 3. Qualitative comparisons among different methods on SUN 360 Panorama testing dataset. The scaling factor is 4.

an evaluation metric that measures the distance between the distribution of images generated by a generative model and the distribution of real images (ground truth). LPIPS [39] is used to measure the perceptual similarity between two images. The higher PSNR and SSIM represent the better performance, while the lower NIQE, FID, and LPIPS mean the better results. As can be seen from Table I, our method outperforms other SR methods in most evaluation metrics on the ODI-SR and SUN 360 Panorama test datasets. In terms of the NIQE metric, the proposed method is the closest to PASD [17], which achieves the best result of NIQE. This demonstrates the effectiveness and superiority of our method.

2) Qualitative Comparison with State-of-the-Art Methods: Fig. 3 shows the visual results of different methods on the SUN 360 Panorama test dataset. As we can see, although SwinIR [29] and OSRT [1] perform well in restoring synthetic images, their performance degrades significantly when tested on real-world images which contain multiple and more complex degradation. ESRGAN [28] also struggles to reconstruct images with complex and severe degradation and introduces many unpleasant artifacts, which is a drawback of the GAN-based method. PASD [17] produces images that are too smooth and cannot restore details. Furthermore, from the third row in Fig. 3, we can see that PASD [17] restores the sky as the wall. The possible reason is that PASD is based on a generative

model, so it will generate some wrong content in some situations. Our method can produce more visually pleasing results with richer details and textures, which are closer to ground truths. Qualitative comparisons between different types of SR methods also demonstrate the superiority of the proposed methods.

3) The Rank in LoRA: The rank r determines the dimension of the trainable matrices, i.e., $A \in \mathcal{R}^{r\times n}$ and $B \in \mathcal{R}^{m\times r}$, in LoRA. A larger r represents more trainable parameters and more computational resources. We compare the impact and number of trainable parameters of different r on various evaluation metrics. Table II shows the comparisons between $r = 4$ and $r = 8$ on the two test datasets. Although the number of trainable parameters of the model with $r = 8$ is 3.18M, which is twice as large as that of the model with $r = 4$, i.e., 1.59M, the model with $r = 4$ performs equally well, or even better than the model with $r = 8$. One possible reason is that the domain gap between 2D planar images and omnidirectional images is not very large, so there is no need to fine-tune a large number of parameters. Therefore, we set $r = 4$ as the default value of the rank in our experiments. It is worth mentioning that the number of trainable parameters is 1.13B if the model is fully fine-tuned, which requires significantly more computational resources compared to using LoRA.

4) Fine-tuning Using Different Datasets: After applying LoRA, we fine-tune the model using different datasets, e.g., DIV2K and ODI-SR. DIV2K is a common SR dataset that involves 800 HR 2D planar images for training. DIV2K and ODI-SR represent two different image distributions. When fine-tuning using these two different datasets, we keep the architecture of the model and training configuration the same. As shown in Table III, the model trained using ODI-SR performs better than or on par with the model trained using DIV2K. This illustrates that it is important to keep the distributions of training data and test data consistent.

V. Conclusions

In this paper, we propose an efficient adaptation method for real-world omnidirectional image super-resolution (ODISR). Unlike previous real-world ODISR methods, we focus on SR of real-world ODIs with more complex degradations. Based on the advanced backbone for real-world 2D planar images, we propose to adopt the low-rank adaptation (LoRA) technique to adapt the SR domain from 2D planar images to omnidirectional images (ODIs). On one hand, training the low-rank decomposed matrices of the pretrained model is both time and resource efficient. On the other hand, the knowledge of pretrained model can be efficiently adapted to downstream taskS. Quantitative and qualitative results demonstrate the effectiveness and superiority of the proposed method.

References

[1] F. Yu, X. Wang, M. Cao, G. Li, Y. Shan, and C. Dong, "Osrt: Omnidirectional image super-resolution with distortion-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 283–13 292.

[2] C. Yang, J. Xiao, Y. Ju, G. Qiu, and K.-M. Lam, "Improving robustness of single image super-resolution models with monte carlo method," in *Proceedings of the IEEE International Conference on Image Processing*, 2023, pp. 2135–2139.

[3] R. Dong and K.-M. Lam, "Bi-center loss for compound facial expression recognition," *IEEE Signal Processing Letters*, vol. 31, pp. 641–645, 2024.

[4] Z. Lyu, J. Xiao, C. Zhang, and K.-M. Lam, "Ai-generated image detection with wasserstein distance compression and dynamic aggregation," in *2024 IEEE International Conference on Image Processing*, IEEE, 2024, pp. 3827–3833.

[5] H. Nagahara, Y. Yagi, and M. Yachida, "Super-resolution from an omnidirectional image sequence," in *Proceedings of the IEEE International Conference on Industrial Electronics, Control and Instrumentation*, vol. 4, 2000, pp. 2559–2564.

[6] Z. Arican and P. Frossard, "Joint registration and super-resolution with omnidirectional images," *IEEE Transactions on Image Processing*, vol. 20, no. 11, pp. 3151–3162, 2011.

[7] X. Deng, H. Wang, M. Xu, Y. Guo, Y. Song, and L. Yang, "Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9189–9198.

[8] A. Nishiyama, S. Ikehata, and K. Aizawa, "360 single image super resolution via distortion-aware network and distorted perspective images," in *Proceedings of the IEEE International Conference on Image Processing*, 2021, pp. 1829–1833.

[9] Y. Yoon, I. Chung, L. Wang, and K.-J. Yoon, "Spheresr: 360deg image super-resolution with arbitrary projection via continuous spherical image representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5677–5686.

[10] C. Yang, R. Dong, J. Xiao, *et al.*, "Geometric distortion guided transformer for omnidirectional image super-resolution," *arXiv preprint arXiv:2406.10869*, 2024.

[11] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1604–1613.

[12] K. Zhang, L. V. Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3217–3226.

[13] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4791–4800.

[14] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1905–1914.

[15] K. C. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "Glean: Generative latent bank for large-factor image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 245–14 254.

[16] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *International Journal of Computer Vision*, pp. 1–21, 2024.

[17] T. Yang, P. Ren, X. Xie, and L. Zhang, "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization," *arXiv preprint arXiv:2308.14469*, 2023.

[18] X. Lin, J. He, Z. Chen, *et al.*, "Diffbir: Towards blind image restoration with generative diffusion prior," *arXiv preprint arXiv:2308.15070*, 2023.

[19] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang, "Seesr: Towards semantics-aware real-world image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 456–25 467.

[20] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[21] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8628–8638.

[22] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yuan, *et al.*, "Cross aggregation transformer for image restoration," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 25 478–25 490.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[24] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 23 593–23 606.

[25] C. Ledig, L. Theis, F. Huszár, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[27] A. Aghajanyan, L. Zettlemoyer, and S. Gupta, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," *arXiv preprint arXiv:2012.13255*, 2020.

[28] X. Wang, K. Yu, S. Wu, *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision Workshops*, 2018, pp. 1–16.

[29] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

[30] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 32, 2019.

[31] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 701–710.

[32] Y. Huang, S. Li, L. Wang, T. Tan, *et al.*, "Unfolding the alternating optimization for blind super resolution," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5632–5643.

[33] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3118–3126.

[34] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, "Recognizing scene viewpoint using panoramic place representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2695–2702.

[35] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[37] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[38] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of Multivariate Analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

More Direct and stage-wise network for Face Super Resolution

Yohei Horiguchi*, Masaaki Ikehara* and Kei Shibasaki*
* University of Keio, Kanagawa, Japan
E-mail: horiguchi0722@keio.jp Tel: +81-45-566-1530

***Abstract*—In recent years, U-net-based models have demonstrated high performance in high-magnification face super resolution (FSR) tasks and are capable of outputting more vivid super resolution (SR) images. However, this model has some drawbacks, such as the possibility of unnecessary computations and uniform block selection at all stages of the decoder. Therefore, we propose a Direct and Stage-wise (DS) Net, which improves on CTCNet [1], a U-net based model with high qualitative results. This model outperforms previous networks by eliminating encoders to reduce computational waste, and by focusing on global feature extraction in the small tensor size stage.**

I. INTRODUCTION

Super-resolution (SR) is a traditional task in image processing that aims to reconstruct high resolution (HR) images from low resolution (LR) images. This task is an ill-posed problem since most LR images are affected by multiple degrading factors including aliasing, motion blur and out-of-focus and there are numerous possible SR results based on which degrading factors are dominant in each picture. To solve this task many machine learning-based methods are proposed.

Currently, various networks including RCAN [2] and SwinIR [3] have been proposed for low-magnification SR tasks such as 2×, and they can already output clear SR results that are not much different from ground truth (GT) images. However, research on SR tasks at high magnifications, such as 8×, has yet to output images like those of GT, and there is room for further development.

Face super resolution (FSR) is the major task in such high-magnification SR tasks. The reason why face images are often used in high-magnification SR is because face images have few high-frequency components. The skin area that occupies the majority of the face image can be represented with a certain degree of color, so there are fewer high-frequency components than in other types of images, such as building images. Since it is difficult to retrieve high-frequency components from LR images in high-magnification SR, face images with relatively few high-frequency components are considered suitable for this purpose. In addition, face image restoration can be utilized not only for entertainment applications but also for many tasks including verification, and analysis making it a task with high social demands.

In the field of FSR, many methods have been proposed. Recently, the mainstream approach involves up-sampling the LR image to HR size using bicubic interpolation and then passing it through a U-Net structured network, similar to image deblurring, to output the SR image, as seen in methods like SPARNet [4] and CTCNet [1].

However, there are two problems with these methods: First, the process of up-sampling the image size once and then down-sampling it again at the encoder part of the U-net structure is a two-step process, which increases the computational cost and may prevents meaningful feature extraction. The second problem is that in the decoder section, the same feature extraction block is used for all stages and feature extraction appropriate for each size stage is not performed.

To solve the above problems, we introduce Direct and Stage-wise (DS) Net which based on CTCNet [1], which is a Unet-based network that currently boasts high performance. This DSNet has 3 key aspects. First, we omitted the encoder part of the U-net and directly input LR images from the bottleneck part to reduce the computational cost. Second, we changed the Feature Refinement Module (FRM), the feature extraction block in the bottleneck part, to a Global-wise Feature Refinement Module (GFRM) that can extract more global information. Finally, the third improvement was achieved by removing the convolution neural networks (CNN) based Facial Structure Attention Unit (FSAU), which is good at extracting local information, from the Local-Global Feature Cooperation Module (LGCM) in the bottom layer of the decoder section, and replacing it with a transformer block only, thereby adjusting the number of blocks and improving accuracy. With these improvements, we were able to demonstrate superiority over CTCNet [1] and other major SR methods without a significant increase in computational load.

II. RELATED WORKS

Since Baker and Kanade [5] proposed the concept of FSR, many methods were proposed for FSR tasks. Some FSR

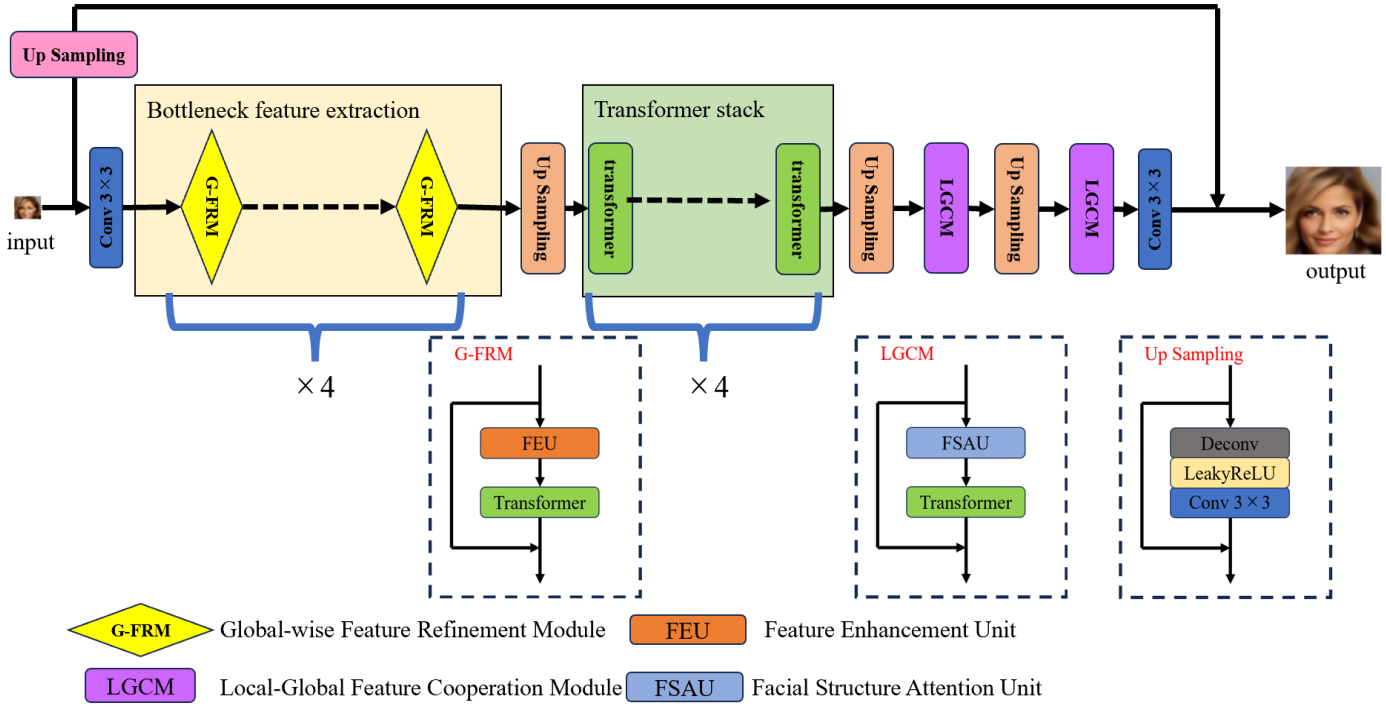


Fig. 1 The complete structure of the proposed Direct and Stage-wise (DS) network

methods utilize prior-guided training. These methods succeed in utilizing multiple information such as facial landmarks and action units to reconstruct SR images but it requires an additional labeled dataset [6]. Recently, the diffusion model [7] has contributed to the improvement of SR results however these models often require high computation capacity and therefore, implementation in end devices is still a long way off. CTCNet [1], which we will use as our base network model, can be trained on a dataset of HR-LR pairs and achieves SR with relatively low computational cost with high accuracy.

CTCNet [1] is composed of three parts: an encoder part, a bottleneck part, and a decoder part. Each part performs feature extraction using CNN-based blocks (FSAU and Feature Enhancement Unit (FEU)) and transformer blocks to enable both global and local feature extraction. The contents of each block will be lightly touched upon in the proposed method, but for details, please refer to this paper on CTCNet [1]. However, there are two problems with this network: The first is the wastefulness of the encoder part. In CTCNet [1], the LR image is up-sampled by bicubic sampling and then down-sampled while extracting features in the encoder part, but this requires a large computational load because of the feature extraction of the large-size image. In addition, it is questionable to what extent applying the high-dimensional information extraction characteristic of the U-net structure will contribute to SR results.

The second is that the same feature extraction block is used in each stage of the decoder section. The LGCM feature extraction block in the decoder section consists of a CNN-based FSAU block suitable for local information extraction and a transformer block suitable for global information extraction [1]. However, since the size of the tensor differs at each stage, it will be essential to adjust the size of the tensor at each stage to determine whether the emphasis should be placed on global or local information. Specifically, since information on fine details of a face image cannot be obtained when the tensor is small, we believe that the focus should be on global information extraction for the entire face and that the weight of local information extraction should be increased at the stage where a large tensor is processed.

III. PROPOSED METHOD

A. *Overall architecture*

DSNet has the structure shown in the Fig.1. In DSNet, we input LR images directly to the network and repeat feature extraction and up-sampling until we get images that are equivalent to the size of HR images. By this direct and simple network structure, we could drastically reduce unnecessary computation cost while preserving SR performance.

To better demonstrate the model, we define I_{LR}, I_{SR}, and I_{HR} as the LR input image, the recovered SR image, and the ground-truth HR image, respectively. Also, the size of the LR image is set to 16×16.

When an LR image is input to the network, shallow features are first extracted by 3 × 3 convolution, and then both global and local features are extracted using a G-FRM block in the Bottleneck feature extraction section. Next, the up-sampling block consisting of a 6 × 6 transposed convolutional layer with stride 2, a LeakyReLU activation function, and a 3 × 3

convolution with stride 1 is used for ×2 up-sampling to obtain 32 × 32 size feature map. After up-sampling to 32 × 32 size, features are extracted with emphasis on global information through the Transformer stack.

After that, the process of 2× up-sampling by the up-sampling block and feature extraction by the LGCM block is repeated twice, and finally, I_{out} is made using 3×3 convolution. Finally, the output image is given by $I_{SR} = I_{out} + I_{LR}^{8\uparrow}$. Meanwhile, $I_{LR}^{8\uparrow}$ is made by conducting ×8 bicubic up-sampling of I_{LR}. Given a training dataset $\{I_{LR}^{i}, I_{HR}^{i}\}_{i=1}^{N}$, DSNet utilize pixel-level loss function:

$$\mathcal{L}(\Theta) = 1/N \, \Sigma_{i=1}^{N} \left\| F_{DSNet}(I_{LR}^{i}, \Theta) - I_{HR}^{i} \right\|, \qquad (1)$$

Meanwhile, N, $F_{DSNet}(\cdot)$ and Θ denote the number of training datasets, DSNet and parameter set of DSNet, respectively.

B. Bottleneck feature extraction

In the bottleneck part that processes the 16×16 tensor, CTCNet [1] use the FRM block which consists of two feature extraction blocks, FSAU and FEU. Since both of these feature extraction blocks are CNN based ones, they are suited for local feature extraction. Especially, the FEU uses a double-branch structure to extract features from the original scale tensor and the down-sampled tensor, and finally fuse them to reduce the computational cost [1].

However, since feature extraction by FRM is performed on extremely small tensors in the bottleneck part, it is more important to extract global information than to extract local information.

Therefore, in DSNet we deleted FSAU blocks from FRM and inputted transformer blocks instead, naming it G-FRM block. The reason why we put transformer block instead of CNN based block is because transformer blocks can learn the relationships between elements of a sequence and thus suited for extracting global information [8]. In DSNet, we use transformer block that consist of Multi-Dconv head Transposed Attention (MDTA) and Gated-Dconv Feed-forward Network (GDFN) from [9]. MTDA is designed to address the computational challenges of conventional self-attention mechanisms. Instead of applying self-attention across the spatial dimensions, it operates across the channel dimensions, thus efficiently capturing global context with reduced computational cost. Additionally, it uses depth-wise convolutions to focus on local context before generating the overall attention map, ensuring a balance between local and global information processing. The GDFN modifies the traditional feed-forward network in two key ways. Firstly, it uses a gating mechanism, which involves two separate paths of linear transformations that are combined element-wise, with one path being activated by the GELU function. Secondly, it incorporates depth-wise convolutions to capture information from adjacent pixels, enhancing the network's ability to learn local image structures crucial for effective image restoration.

Thus, by combining the FEU, which can extract local information at low computational cost, and the transformer block, which can extract global information, it is possible to perform feature extraction suited to small tensors.

C. Transformer stack

In the bottom stage of encoding part, CTCNet use LGCM that use both CNN based FSAU block and transformer block in order to extract global and local features [1]. Similar to the changes made to the FRM in Bottleneck feature extraction, we decided to eliminate the FSAU and keep only the transformer, since this part also targets small 32 x 32 feature maps and we thought it would be better to put more emphasis on global information extraction. However, we thought that simply using a single transformer block would sacrifice pure feature extraction capability, so we decided to compensate for the lack of FSAU by connecting the transformers in series. After the experiment explained in ablation study, we identified that 4 that four transformers in series was optimal, and we named it the transformer stack.

IV. Experiment

A. Dataset

In this paper, we utilize 2 different datasets as training datasets in order to examine the efficiency of our method. The first dataset we use is the full dataset which consists of 18,000 samples of the CelebA [10] dataset for training, 200 samples for validating. This dataset is the same as the dataset which was used in the CTCNet paper [1], and used for comparing our method and other FSR methods. The second dataset we use is a small dataset which consists of 3,000 samples of the CelebA dataset for training, and 100 samples for validating. This small dataset is used for ablation study and since this small dataset required a shorter time to train, we used this dataset to mitigate the time we need for evaluation. Although the overall accuracy of the networks will be reduced due to the small number of samples that can be referenced in training, we believe that the relative advantage between networks can be confirmed even using a small dataset.

For the test data set, 1000 images from CelebA were used. In all data sets, the CelebA images were cropped to 128×128 size as HR, and LR images are made by conducting 8× bicubic down-sampling.

Table 1. Quantitative comparisons for ×8 SR on the CelebA test sets [1]. Best and second-best results are **bolded** and underlined.

Method	PSNR ↑	SSIM ↑
Bicubic	23.61	0.6779
SAN [12]	27.43	0.7826
RCAN [2]	27.45	0.7824
HAN [13]	27.47	0.7838
SwinIR [3]	27.88	0.7967
FSRNet [14]	27.05	0.7714
FACN [15]	27.22	0.7802
SPARNet [4]	27.73	0.7949
SISN [16]	27.91	0.7971
CTCNet [1]	28.37	0.8115
DSNet (ours)	**29.15**	**0.833**

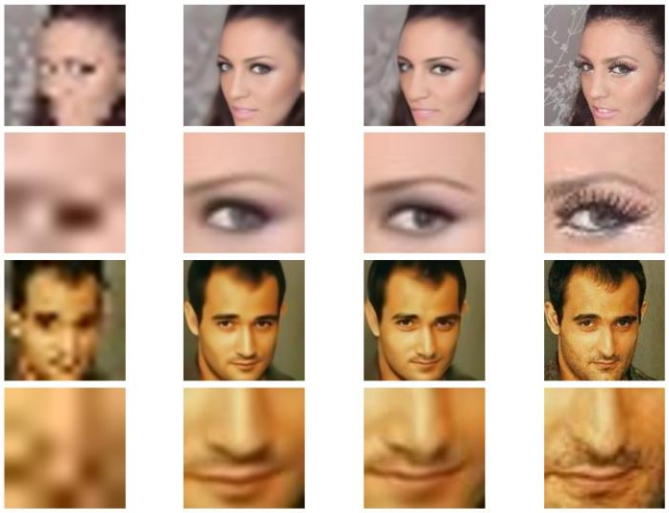


Fig. 2 Visual comparisons for ×8 SR on the CelebA test set.

B. Implementation details

Other training conditions are set as in CTCNet [1]. That is, implementation of our model is done by PyTorch framework, the optimizer is Adam and set $\beta1 = 0.9$ and $\beta2 = 0.99$, the initial learning rate is set to 2×10 -4, the batch size is 10 and the number of epochs is 100. Also, PSNR and SSIM [11] are used for qualitative comparisons.

C. Comparison with other methods

We compare DSNet with 9 major state-of-the-art methods. These methods include general SR methods SAN [12], RCAN [2], HAN [13], SwinIR [3], major FSR methods FSRNet [14], FACN [15], SPARNet [4], SISN [16], and CTCNet [1]. Since DSNet was experimented under the same conditions as CTCNet and When CTCNet was trained and tested under these conditions, the same results as reported in the CTCNet paper were obtained, for the qualitative evaluation values of the state-of-the-art methods, we referred to the result shown on CTCNet paper [1].

Table 2. Quantitative result of omitting encoder section. Better results are **bolded**.

Encoder	PSNR ↑	Params (M) ↓
✓	25.87	24.40
×	**26.04**	**17.67**

Table 3. Quantitative evaluation of contents in FRM block. Best and second-best results are **bolded** and underlined. Bottom contents are used in DSNet

FEU	FSAU	transformer	PSNR ↑	Params (M) ↓
✓	✓	×	26.17	19.87
×	×	✓	22.54	**11.43**
×	✓	✓	26.18	16.63
✓	✓	✓	26.26	22.62
✓	×	✓	**26.37**	19.34

1) Quantitative comparison:

When compared with other SOTA methods, we could see from Tab. 1 that results on DSNet outperforms other method in both PSNR and SSIM evaluation.

2) Qualitative comparison

In Fig.2 when we compare CTCNet [1] and DSNet, we could see, that DSNet can reconstruct SR images that are closer to GT images. For example, when comparing the results of the second male image, it can be seen that DSNet outputs results that are closer to GT in terms of nose shape, etc.

D. Ablation study

In this section, we conducted three ablation studies using the small dataset and examined the efficiency of omission of the encoder section, GFRM in Bottleneck feature extraction and transformer stack, respectively. Since we used small training dataset for Ablation study, quantitative results are decreased compared to results shown in tab. 1 however, we believe that even if the absolute amount of quantitative results has been reduced, the relative superiority relationship between different networks can still be confirmed.

1) Omission of encoder section

In order to assess efficiency of omitting encoder we compared CTCNet with/without encoder stage. Tab. 2 shows that the omission of encoder significantly reduces the number of parameters (reduces the parameters of the model to about 70% of the original) in the model, thus reduces the computational load, and also improves PSNR result.

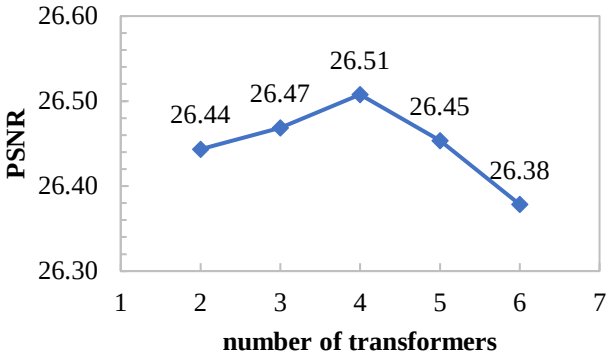


Fig. 3 Relationship between the number of transformers and PSNR in transformer stack.

2) GFRM in Bottleneck feature extraction

In order to verify the effectiveness of GFRM, we compared the qualitative results when a combination of FEU, FSAU, and transformer were introduced into the FRM for the feature extraction block at the bottleneck area. As shown in Tab. 3, it was found that the eliminating FSAU from the FRM and inserting transformer instead produced the best PSNR, and that the number of parameters could be kept at a reasonable value.

3) Number of transformer blocks in transformer stack

To determine the optimal number of transformers at the Transformer stack, we examined the PSNR values when varying the number of transformers. Fig. 4 shows that the best results were obtained when four transformers were used.

V. Conclusions

In this paper, we proposed DSNet, which is a significant improvement over CTCNet. In our study, we found that eliminating encoders in the U-net based SR models does not reduce the accuracy of SR and significantly reduces the computational cost. We also found that instead of using the same feature extraction block in all stages, it is better to use a global feature extraction-intensive network for feature extraction on small tensors. With these two modifications, DSNet was able to produce better quantitative results than other FSR methods.

There are two future prospects. Although the proposed DSNet is able to produce SR results that are somewhat closer to GT than existing methods, it still lacks the ability to reproduce the finest details. Therefore, we believe that a network that can extract more local information and restore details is required.

The second is the addition of more novel blocks. Although this study has succeeded in improving SR performance by making major modifications to CTCNet, the blocks themselves are not different from those used in CTCNet, so there is room to further improve SR performance by adding more novel blocks in the future.

References

[1] Gao, Guangwei, et al. "CTCNet: A CNN-transformer cooperation network for face image super-resolution." IEEE Transactions on Image Processing 32 (2023): 1978-1991.

[2] Zhang, Yulun, et al. "Image super-resolution using very deep residual channel attention networks." Proceedings of the European conference on computer vision (ECCV). 2018.

[3] Liang, Jingyun, et al. "Swinir: Image restoration using swin transformer." Proceedings of the IEEE/CVF international conference on computer vision. 2021

[4] Chen, Chaofeng, et al. "Learning spatial attention for face super-resolution." IEEE Transactions on Image Processing 30 (2020): 1219-1231.

[5] Baker, Simon, and Takeo Kanade. "Hallucinating faces." Proceedings Fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580). IEEE, 2000.

[6] Zhang, Chenggong, and Zhilei Liu. "Face super-resolution with progressive embedding of multi-scale face priors." 2022 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2022.

[7] Gao, Sicheng, et al. "Implicit diffusion models for continuous super-resolution." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[8] Khan, Salman, et al. "Transformers in vision: A survey." ACM computing surveys (CSUR) 54.10s (2022): 1-41.

[9] Zamir, Syed Waqas, et al. "Restormer: Efficient transformer for high-resolution image restoration." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[10] Liu, Ziwei, et al. "Deep learning face attributes in the wild." Proceedings of the IEEE international conference on computer vision. 2015.

[11] Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity." IEEE transactions on image processing 13.4 (2004): 600-612.

[12] Dai, Tao, et al. "Second-order attention network for single image super-resolution." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[13] Niu, Ben, et al. "Single image super-resolution via a holistic attention network." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer International Publishing, 2020.

[14] Chen, Yu, et al. "Fsrnet: End-to-end learning face super-resolution with facial priors." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[15] Xin, Jingwei, et al. "Facial attribute capsules for noise face super resolution." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.

[16] Lu, Tao, et al. "Face hallucination via split-attention in split-attention network." Proceedings of the 29th ACM international conference on multimedia. 2021.

Camera Focal Length Prediction for Neural Novel View Synthesis from Monocular Video

Dipanita Chakraborty*, Werapon Chiracharit*, Kosin Chamnongthai*, and Minoru Okada†
* King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand
† Nara Institute of Science and Technology, Nara 630-0192, Japan
* E-mail: kosin.cha@kmutt.ac.th

***Abstract*—Novel view synthesis is a challenging task that generates multi-view images from a single-view object by reconstructing the depth and spatial information between the camera and the object. This task specifically facilitates the rendering of 2D objects into 3D representations from monocular video scenes. Existing methods suffer from depth information loss between the camera and object when provided with limited single-view input images, resulting in poor reconstruction accuracy in 3D space. Moreover, they lack a high-level depth feature map representation of scene information. Therefore, we propose a multilayer encoder-decoder architecture-based network that efficiently predicts the focal length between the object and camera from a mono-view image. Additionally, we utilize a combined feature extraction strategy from the estimated depth feature map and RGB input image to synthesize novel views. While the encoder network extracts semantic high-level features at multiple scales, the decoder network refines these combined features for synthesis. Our method effectively improves depth information retention while achieving good reconstruction performance. Experimental results evaluated on a benchmark dataset demonstrate the efficacy of our proposed method.**

I. Introduction

The need for an accurate camera focal length estimation directly facilitates facial alignment analysis for biometric applications [1], subject-to-camera distance estimation [2], automatic camera calibration, 3D-object reconstruction, novel view synthesis [3], and several other crucial computer vision applications. Among all these applications, novel view synthesis is a significant computer vision task that can generate new viewpoints of an object from single-viewed images. Traditional novel view synthesis methods use fixed camera intrinsic parameters, such as camera focal lengths. However, these methods are unable to synthesize novel views of an object from in-the-wild images due to the unavailability of camera intrinsic parameters. Hence, this computer vision problem of camera focal length estimation from in-the-wild images has gained the attention of many researchers in the past few years [4].

Some depth-based novel view synthesis methods have been proposed from single-viewed images using disocclussion filling approach [1], [5]. Kim and Kim [6] proposed a machine learning-based novel view synthesis module using pixel generation and flow prediction. They investigated the performance variations based on the skip connections. Their method of using 3 skip connections for flow predictions returned SSIM loss scores of 0.9181 and 0.8903 for object categories "Car" and "Chair", respectively. The results indicate that there is still room for improvement since the higher the SSIM score, the better the performance. A self-supervised learning-based method was proposed in [7] that used a semi-parametric approach to integrate rich visual information. This method synthesized novel views of "car" and "chair" objects from 12 evenly-spaced azimuthal angles. The Neural Radiance Field (NeRF)-based approach using sparse RGB-D images was proposed in [8]. The method solved the main limitation of the large dataset requirement in the conventional view synthesis systems. The highest mean SSIM score achieved by this proposed model was 0.6100. However, the main limitation is that these methods can not generate novel views of in-the-wild images due to the unavailability of camera intrinsic metadata.

An effective way to resolve this problem is by predicting the camera focal length from in-the-wild images. A fast camera focal length extraction method was proposed in [9] that used parallel particle swarm optimization. This method returned good performance while reducing the time consumption. A novel camera focal length determination method was proposed in [10] using a calibration pattern of the checkerboard. However, these methods lack semantic feature information which is needed for estimating the focal length between the camera and the captured image of a complex scene. To address this issue, a deep learning-based approach was proposed in [11] that used the DLA-34 network to perform high-level visual feature extraction first, followed by a 3D box parameters regression task to predict the camera focal length. The achieved accuracy of 56.31% on the benchmark KITTI dataset indicates further need for improvement in their proposed method.

In our proposed method, we consider integrating the camera focal length estimation module into the encoder-decoder network, depth estimation module, and view transformation module to enhance the performance of novel view synthesis. The main contributions of our proposed method are as follows:

1) We propose a camera focal length prediction module using an encoder network and a linear activation function that extracts the semantic high-level features at multi-scale.
2) We utilize the combined feature extraction strategy from the input RGB image and corresponding depth map to improve the view transformation module.
3) We modify the decoder network with convolutional operations to synthesize the novel views.

This article is presented as follows: Section II reviews the existing state-of-the-art research, Section III describes the proposed methodology in detail, Section IV explains implementation and experimental evaluation, and lastly Section V concludes the achievements and future scope of this research study.

II. Related Work

We introduce an encoder-decoder novel view synthesis model architecture with an integrated camera focal length prediction model. Therefore, we summarize a brief of existing methodologies into two sections, camera focal length estimation model and novel view synthesis model.

A. Camera Focal Length Estimation Model

A robust camera pose and camera focal length estimation method were proposed in [12] to solve the perspective-n-point (PnP) problem for uncalibrated cameras, also referred to as unknown camera intrinsic parameters. This method utilized exhaustive linearization to solve the UPnP (Uncalibrated PnP) equation and achieved good accuracy. However, the time consumption was comparatively high. To address this issue, Zhou et al. [13] proposed an efficient PnP solution for cameras with unknown camera focal lengths using a polynomial system. Later, Yin et al. [14] proposed a robust PnP solution for cameras with unknown focal lengths using Gröbner basis minimal solver combined with convex optimization. Although, this proposed method achieved higher accuracy with lower time consumption, in some cases the matching algorithm failed to provide sufficient key points to localize the PnP problem. He et al. [15] proposed a deep learning-based approach for depth estimation using embedded camera focal lengths from single-viewed images. Their method could efficiently reconstruct depth feature maps, however, their system could not achieve the fastest of all state-of-the-art methods in terms of lowest run time. Ponimarkin et al. [16] proposed a focal length estimation method using a render and compare strategy from in-the-wild RGB images. This method significantly reduced the loss error than the existing methods. However, this method suffered from false retrieval of some 3D models.

B. Deep View Synthesis Model

The novel view of an object can be efficiently reconstructed by generative neural networks using depth feature extraction. Liu et al. [17] proposed a deep generative model, called SCGN based on encoder-decoder network architecture. The advantage of this network it can encode appropriate high-level features without the requirement of geometrical property rectifications, resulting in reduced geometrical distortion issues. However, this method did not investigate single-viewed source images. Lei et al. [18] proposed a powerful deep view synthesis network, called DGCC-Net by combining gradual and cycle synthesis. Whereas the gradual conversion technique learns the progressive rotation trend using intermediate transformation between the source and target images to synthesize a clearer target view, the cycle network maps the synthesized target view back to the source view to promote better feature learning of the target view. The proposed methodology performed well in terms of reducing the computation complexity. However, the performance can be improved by incorporating various features along with the RGB source image. Addressing this issue, Jiang et al. [19] proposed an encoder-decoder architecture-based network that uses the source image and its corresponding warp image as the input. Additionally, this method introduced to channel attention block technique to reduce the errors from depth estimation and target view generation process. However, due to the unsupervised training for depth estimation, some missing pixels were observed in the generated target view.

In this article, we aim to improve the novel view synthesis model by embedding predicted camera focal length information and semantic high-level visual features at multi-scale to generate a depth feature map. Additionally, we apply a combined feature extraction strategy from estimated depth features and the corresponding RGB input image to generate novel views.

III. Methodology

This section presents the proposed camera focal length prediction method to generate a novel view of a single-viewed object, as shown in Fig. 1. We describe the method in four steps. First, we use an encoder network to extract semantic high-level features from the RGB input image, as shown in Fig. 2. These features are then utilized to predict the focal length between the object and the camera. In the second step, we concatenate the extracted semantic high-level features with the predicted focal length features and feed them to a decoder network. The decoder network then estimates the depth feature map from the input image. In the third step, we use another encoder network to extract combined features from the input image and its corresponding depth feature map. Finally, we use a perspective transformation module on the combined features to transform into a new viewpoint. Then another decoder network is applied to synthesize a novel view image from the transformed features, upsampling it to the original input image resolution. The following subsections A, B, C, and D provide detailed material about the proposed methodology.

A. Camera Focal Length Prediction with Encoder

The main objective of this step is to predict the camera's focal length from an in-the-wild image, where the camera's intrinsic properties, including the focal length, are unknown. This step is crucial because it enhances our understanding of the geometrical properties, spatial location, and depth information of objects in the image. As shown in Fig. 3, the distance between the camera lens and the sensor is referred to as the focal length (F), while the angle of view (θ) represents the scene that the camera sensor can capture. The angle of view (θ) can be determined using the focal length and sensor width (W), as in (1).

$$\theta = 2 \times tan^{-1} \frac{W}{2 \times F}. \tag{1}$$

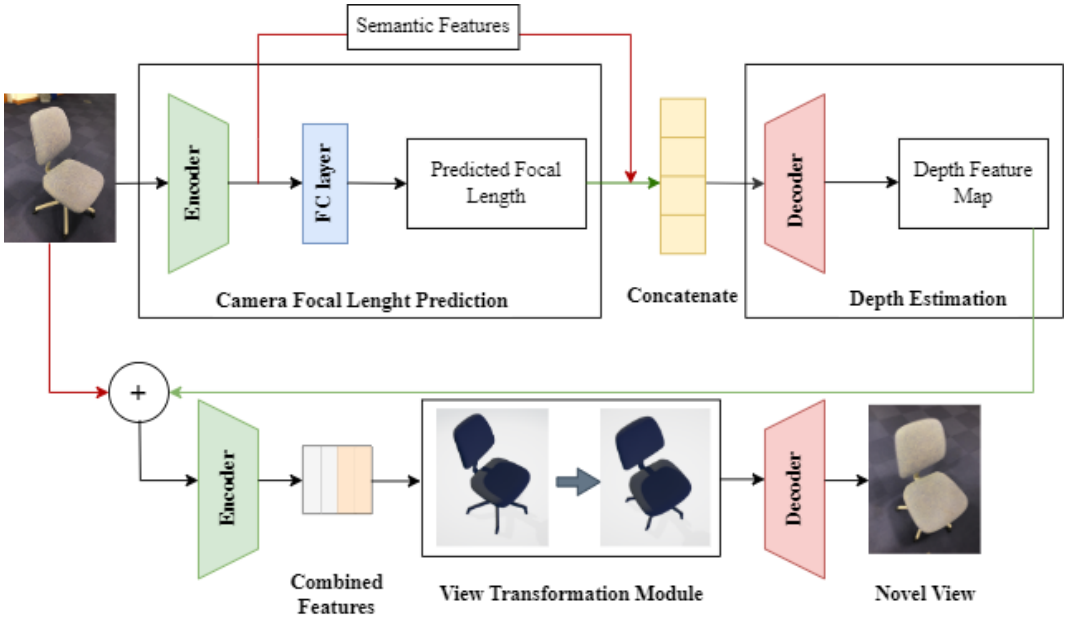


Fig. 1. Shows framework of the proposed methodology.

.

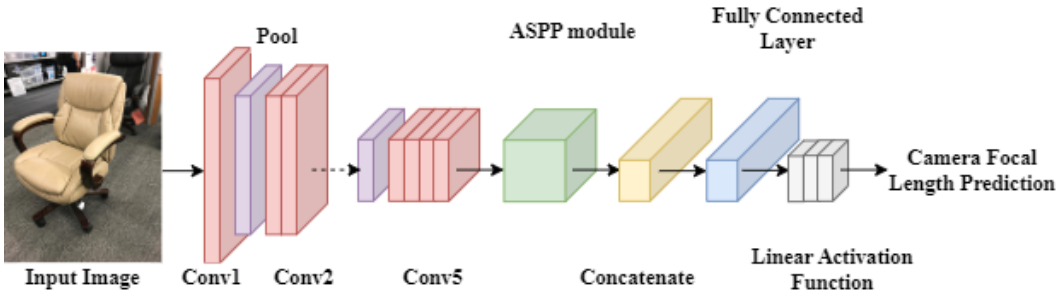


Fig. 2. Shows the proposed encoder network architecture based on DeepLabv3+ network, where Resnet-18 CNN is used as the backbone network.

.

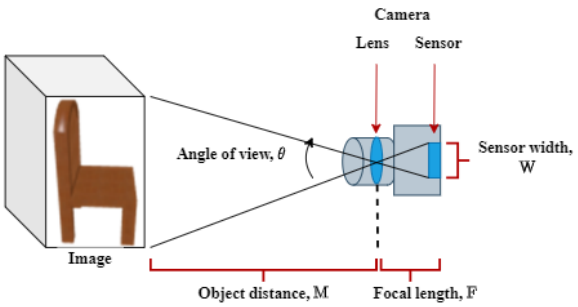


Fig. 3. Shows two of the camera intrinsic parameters, focal length and angle of view, which are used for novel view synthesis.

.

The camera focal length prediction approach directly facilitates depth estimation and 2D to 3D rendering of objects. We use an encoder network to extract the contextual features at a multi-scale. The encoder network is built upon a modified DeepLabv3+ network that encodes both low-level and high-level features from the source input image. We utilize the ResNet-18 convolutional neural network as the backbone of the encoder, as in [20]. As shown in Fig. 2, we extract the low-level feature from the Conv2 layer (64-channel feature map) and continue the backbone network flow until Conv5 layer. Then we feed the last feature map (512-channel feature map) to an **Atrous Spatial Pyramid Pooling (ASPP)** module. The ASPP module consists of one layer of point-wise convolution with the filter size 1×1 and 5 layers of depthwise convolution with the filter size 3×3. We employ different dilation rates of 2, 6, 12, 18, and 24 in the filters of depthwise convolutional layers which enables semantic high-level feature extraction at the multi-scale. The semantic high-level feature ($S_{i,j}$) with K numbers of Conv filters and $n-th$ dilated Conv layer at pixel position (i,j) can be defined as in (2).

$$S_{i,j} = \sum_{n=1}^{N} I_{i,j,nK} \times h_{i,j,nk}. \quad (2)$$

At the last layer, we use a fully connected (dense) layer, followed by a linear activation function to predict the focal length of the source image. In the next step, we combine this predicted focal length information with the semantic features for the depth estimation.

B. Depth Estimation with Decoder

The proposed decoder network generates the depth feature map using the extracted features from the encoder network. The depth map indicates the relative distances of objects in the image from the camera. After the prediction of the camera focal length, we concatenate the captured semantic high-level features with the corresponding focal length and feed this to the decoder network. The decoder refines the features, taking into account the focal length, to produce a depth map. We add 3×3 Conv. layer to refine the features. Finally, we add bilinear upsampling by factor 4 to the refined features, to make it compatible with the original image resolution. After that, we add a sigmoid activation function to estimate the depth feature map output of the corresponding input image.

C. Combined Feature Extraction

We combine the RGB input image and its corresponding depth feature map at this step. Similar to the previous encoder network architecture, we use another encoder network to extract combined features. This approach helps the network to learn semantic dense depth information of the input image. The combined feature extraction can be represented as in (3).

$$\beta_{combined} = Encoder_{combined}(I_{RGB} + S_{depth}). \quad (3)$$

Where, I_{RGB} and S_{depth} refer to RGB input image and estimated depth feature map generated using semantic high-level feature at multi-scale, respectively.

D. View Transformation Module

At this step, we transform the extracted combined features into a new viewpoint of the object based on the perspective transformation approach. The projection of the 3D point to the image plane (2D coordinates) can be defined as in (4).

$$p_{image} = T_l \rightarrow K \cdot [R|t] \cdot P_{world}. \quad (4)$$

where p_{image} and P_{world} are the homogeneous coordinates in the image plane and real-world coordinates, respectively. $[R|t]$ is combined rotation matrix R and a translation vector t. T_l represents the transformed features that belongs to p_{image}. The perspective projection matrix K includes the intrinsic parameters of the camera focal length F.

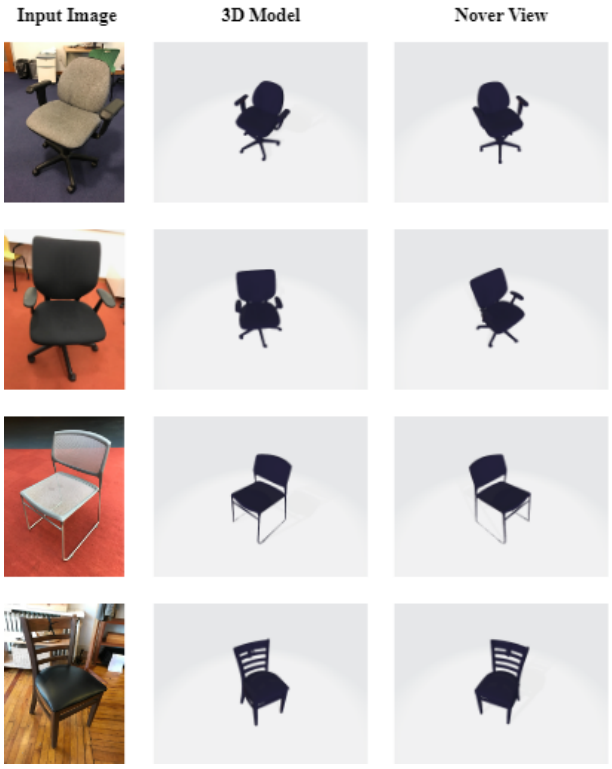


Fig. 4. Shows the RGB input image and novel views from Pix3D [22] dataset for object category "chair".

E. Generating Novel View

We use another decoder network that refines the transformed viewpoint features at this step. In this decoder network, we add two layers of Conv filters of size 3×3, followed by bilinear upsampling by factor 4 to generate the novel view of the same resolution as the input image. The synthesized image with a novel viewpoint using the transformed features can be represented as in (5).

$$\eta_{novel} = Decoder(T_l). \quad (5)$$

We discussed the implementation of the proposed methodology and evaluated results in the section IV.

IV. Experimental Results

Our proposed method is implemented in Python software on a personal computer with PyTorch library, NVIDIA Titan 1650i GPU of 8 GB memory, an Intel(R) Core(TM) i7-10750H CPU running at 2.60 GHz, a 64-bit operating system, and a 500 GB of solid-state drive (SSD). We use the Adam optimizer with momentum 0.9, LearningRate 0.001, and L2Regularization 0.005. We train and evaluate our proposed method on the benchmark ShapeNet [21] and Pix3D [22] datasets on the object category "chair". We divide the datasets for training, validation, and testing sets in a 70:20:10 ratio.

We use three loss functions to prevent overfitting: MSE for focal length loss, L1 loss for depth estimation, and SSIM loss

for novel view synthesis. Some examples of the RGB input image and generated novel views are represented in Fig. 4. The results show that our embedded camera focal length prediction model, and the combined feature extraction of RGB input and depth feature map technique performed well for novel view synthesis. However, our system encountered comparatively higher run time than the existing methods. This problem can be reduced by utilizing faster convolutional operations in the proposed network.

V. Conclusions

We propose a deep neural network model based on an encoder-decoder architecture, which includes an embedded camera focal length prediction component. This model can efficiently reconstruct new viewpoints from a single input image with unknown focal lengths. Our experimental results demonstrate good performance in depth feature map estimation by utilizing high-level semantic features at multiple scales. In the future, we plan to evaluate the proposed methodology on different object categories from benchmark datasets.

Acknowledgment

The authors highly appreciate the anonymous reviewers from APSIPA ASC 2024 for their valuable reviews and helpful comments. This research was supported by King Mongkut's University of Technology Thonburi's postdoctoral fellowship in the 2024 (2567) fiscal year under KIRIM project ID 27928 and agreement number 06/2567.

References

[1] X. Li, J. Liu, J. Baron, K. Luu, and E. Patterson, "Evaluating effects of focal length and viewing angle in a comparison of recent face landmark and alignment methods," *J. Image Video Process.*, vol. 2021, no. 9, Mar. 2021, doi:10.1186/s13640-021-00549-3.

[2] E. Bermejo, E. F.-Blanco, A. Valsecchi, P. Mesejo, O. Ibáñez, and K. Imaizumi, "FacialSCDnet: A deep learning approach for the estimation of subject-to-camera distance in facial photographs," *Expert Syst. Appl.*, vol. 210, Dec. 2022, doi:10.1016/j.eswa.2022.118457.

[3] I. Ahn and C. Kim, "A Novel Depth-Based Virtual View Synthesis Method for Free Viewpoint Video," *IEEE Trans. Broadcast.*, vol. 59, no. 4, pp. 614-626, Dec. 2013, doi: 10.1109/TBC.2013.2281658.

[4] G. Ponimatkin, Y. Labbé, B. Russell, M. Aubry, and J. Sivic, "Focal Length and Object Pose Estimation via Render and Compare," *in Proces. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, 2022, pp. 3815-3824, doi: 10.1109/CVPR52688.2022.00380.

[5] J. Thatte and B. Girod, "A Statistical Model for Disocclusions in Depth-based Novel View Synthesis," *in Proc. IEEE Visual Communications and Image Processing*, Sydney, Australia, 2019, pp. 1-4, doi: 10.1109/VCIP47243.2019.8966071.

[6] J. Kim and Y. M. Kim, "Novel View Synthesis With Skip Connections," *in Proc. International Conference on Image Processing*, Abu Dhabi, United Arab Emirates, 2020, pp. 1616-1620, doi: 10.1109/ICIP40778.2020.9191076.

[7] A. Palazzi, L. Bergamini, S. Calderara, and R. Cucchiara, "Warp and Learn: Novel Views Generation for Vehicles and Other Objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2216-2227, Apr. 2022, doi: 10.1109/TPAMI.2020.3030701.

[8] Y. -J. Yuan, Y. -K. Lai, Y. -H. Huang, L. Kobbelt, and L. Gao, "Neural Radiance Fields From Sparse RGB-D Images for High-Quality View Synthesis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8713-8728, Jul. 2023, doi: 10.1109/TPAMI.2022.3232502.

[9] C. Zheng, H. Qiu, C. Liu, X. Zheng, C. Zhou, Z. Liu, and J. Yang, "A Fast Method to Extract Focal Length of Camera Based on Parallel Particle Swarm Optimization," *in Proc. Chinese Control Conference*, Wuhan, China, 2018, pp. 9550-9555, doi: 10.23919/ChiCC.2018.8483981.

[10] C. T. Nguyen, R. D. Khlynov, A. A. Gorbachev, V. A. Ryzhova, S. N. Yarishev, I. A. Konyakhin, T. S. Djamiykov, M. B. Marinov, "Determining the Focal Length of a Video Camera Using a Calibration Pattern," *International Scientific Conference Electronics*, Sozopol, Bulgaria, 2022, pp. 1-4, doi: 10.1109/ET55967.2022.9920285.

[11] H. Xiao, K. Li, Y. Zhu, and J. Zhang, "3D Object Detection Based on Long and Short Focal Length Cameras," *International Conference on Computer Vision, Image and Deep Learning*, Zhuhai, China, 2024, pp. 545-548, doi: 10.1109/CVIDL62147.2024.10604155.

[12] A. P.-Sanchez, J. A.-Cetto, and F. M.-Noguer, "Exhaustive Linearization for Robust Camera Pose and Focal Length Estimation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2387-2400, Oct. 2013, doi: 10.1109/TPAMI.

[13] B. Zhou, Z. Chen, and Q. Liu, "An Efficient Solution to the Perspective-n-Point Problem for Camera With Unknown Focal Length," *IEEE Access*, vol. 8, pp. 162838-162846, 2020, doi: 10.1109/ACCESS.2020.3021313.

[14] X. Yin, L. Ma, X. Tan, and D. Qin, "A Robust Visual Localization Method With Unknown Focal Length Camera," *IEEE Access*, vol. 9, pp. 42896-42906, 2021, doi: 10.1109/ACCESS.2021.3065953.

[15] L. He, G. Wang, and Z. Hu, "Learning Depth From Single Images With Deep Neural Network Embedding Focal Length," *IEEE Trans. on Image Process.*, vol. 27, no. 9, pp. 4676-4689, Sept. 2018, doi: 10.1109/TIP.2018.2832296.

[16] G. Ponimatkin, Y. Labbé, B. Russell, M. Aubry, and J. Sivic, "Focal Length and Object Pose Estimation via Render and Compare", 2022, arXiv:2204.05145.

[17] Z. Liu, W. Jia, M. Yang, P. Luo, Y. Guo, and M. Tan, "Deep View Synthesis via Self-Consistent Generative Network," *IEEE Transactions on Multimedia*, vol. 24, pp. 451-465, 2022, doi: 10.1109/TMM.2021.3053401.

[18] J. Lei, B. Liu, B. Peng, X. Cao, Q. Huang, and N. Ling, "Deep Gradual-Conversion and Cycle Network for Single-View Synthesis," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 6, pp. 1665-1675, Dec. 2023, doi: 10.1109/TETCI.2023.3272003.

[19] L. Jiang, G. Schaefer, and Q. Meng, "An Improved Novel View Synthesis Approach Based on Feature Fusion and Channel Attention," *in Process. IEEE International Conference on Systems, Man, and Cybernetics*, Prague, Czech Republic, 2022, pp. 2459-2464, doi: 10.1109/SMC53654.2022.9945244.

[20] D. Chakraborty, W. Chiracharit, and K. Chamnongthai, "Semantic Scene Object-Camera Motion Recognition for Scene Transition Detection Using Dense Spatial Frame Segments and Temporal Trajectory Analysis," *IEEE Access*, vol. 12, pp. 21673-21697, 2024, doi: 10.1109/ACCESS.2024.3363233.

[21] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su et al., "ShapeNet: An information-rich 3D model repository," arXiv preprint arXiv:1512.03012, 2015.

[22] X.Sun, J.Wu, X.Zhang, Z.Zhang, C.Zhang, T.Xue,J.B.Tenenbaum, and W. T. Freeman, "Pix3D: Dataset and methods for single-image 3D shape modeling," *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 29742983.

Scene-Segmentation-Based Exposure Compensation for Tone Mapping of High Dynamic Range Scenes

Yuma Kinoshita* and Hitoshi Kiya†
* Tokai University, Japan
E-mail: ykinoshita@tokai.ac.jp
† Tokyo Metropolitan University, Japan
E-mail: kiya@tmu.ac.jp

***Abstract*—We propose a novel scene-segmentation-based exposure compensation method for multi-exposure image fusion (MEF) based tone mapping. The aim of MEF-based tone mapping is to display high dynamic range (HDR) images on devices with limited dynamic range. To achieve this, this method generates a stack of differently exposed images from an input HDR image and fuses them into a single image. Our approach addresses the limitations of MEF-based tone mapping with existing segmentation-based exposure compensation, which often result in visually unappealing outcomes due to inappropriate exposure value selection. The proposed exposure compensation method first segments the input HDR image into subregions based on luminance values of pixels. It then determines exposure values for multi-exposure images to maximize contrast between regions while preserving relative luminance relationships. This approach contrasts with conventional methods that may invert luminance relationships or compromise contrast between regions. Additionally, we present an improved technique for calculating fusion weights to better reflect the effects of exposure compensation in the final fused image. In a simulation experiment to evaluate the quality of tone-mapped images, the MEF-based tone mapping with the proposed method outperforms three typical tone mapping methods including conventional MEF-based one, in terms of the tone mapped image quality index (TMQI).**

I. Introduction

High dynamic range (HDR) imaging, also known as HDR image synthesis, is a technique that reconstructs the luminance information of real-world scenes with an extensive dynamic range by combining multiple images captured using imaging sensors with a limited dynamic range [1], [2]. The resultant images are termed HDR images, which are distinctly categorized from images obtained through conventional photography, namely low dynamic range (LDR) images. In contemporary times, smartphones enables us to capture HDR images without clipped whites and crushed shadows.

Dynamic range constraints are not limited to image sensors but also exist in display devices. The sRGB color space, which is widely used in current displays, assumes a relatively low peak luminance of $80\,\mathrm{cd\,m^{-2}}$. While HDR displays capable of reproducing a broader dynamic range with peak luminances exceeding $10^3\,\mathrm{cd\,m^{-2}}$ are becoming increasingly prevalent, the dynamic range of displays remains limited compared to real-world scenes. For instance, the luminance of direct sunlight can reach up to $10^9\,\mathrm{cd\,m^{-2}}$. To address this disparity and display HDR images on devices with restricted dynamic range, a process known as tone mapping is essential. This technique adjusts the image's tonal range to fit within the display's dynamic range capabilities [3]–[8].

Global tone mapping, which applies a single function such as a gamma curve or S-curve uniformly to all pixels, is a simple and widely used approach. However, this method tends to diminish the high local contrast between neighboring pixels inherent in HDR images. Consequently, research has been directed towards local tone mapping techniques, which adaptively apply tone mapping to individual pixels based on local information. Recently, there has been active research in multi-exposure image fusion (MEF) methods [9]–[16], which aim to represent the wide luminance range of real-world scenes within a limited dynamic range by fusing a stack of differently exposed images, called multi-exposure images. These fusion methods can be considered a variant of local tone mapping.

While research on fusion algorithms has progressed, Kinoshita et al. demonstrated that the quality of the final fused image is significantly influenced by exposure values of input multi-exposure images [17], [16]. For this reason, they also proposed an exposure compensation method that segments the HDR image into subregions based on luminance and determines an exposure value for each subregion to ensure well-exposure. Although this method enables to avoid information loss in fused images due to clipped whites or crushed shadows, it does not necessarily contribute to obtaining a visually appealing image because of the inversion of luminance relationships and the loss of contrast between regions.

Because of such a situation, in this paper, we propose an novel exposure compensation method for MEF-based tone mapping of HDR images. In this method, an input HDR image is first segmented into subregions based on luminance, similarly to the conventional method. Then, the exposure values of multi-exposure images are determined to maximize the contrast between regions. As a result, this method preserves the relative luminance relationships between regions, in contrast to the conventional method. Additionally, to better reflect the effects of exposure compensation in the final fused image, we also propose a method for calculating fusion weights.

We evaluated the efficacy of the tone mapping scheme with the proposed segmentation-based exposure compensation method in terms of the quality of tone-mapped images by a simulation, where tone mapped image quality index (TMQI) was utilized as quality metrics. In the simulation, the tone

mapping scheme with the proposed compensation method was compared with typical tone mapping methods, including a MEF based one. Experimental results showed that the proposed segmentation-based exposure compensation is effective for tone mapping. In addition, the segmentation-based exposure compensation outperformed three compared tone mapping methods in terms of TMQI.

II. Related work

In this paper, we propose a novel scene-segmentation-based exposure compensation method for tone mapping HDR images. Because the proposed method is related to conventional tone mapping and multi-exposure image fusion, these techniques including a scene-segmentation-based exposure compensation method are summarized in this section. The following notations are used throughout this paper.

- Bold italic letters, e.g., $\boldsymbol{x}$, denote vectors or vector-valued functions, and they are assumed to be column vectors.
- The notation $\{x_1, x_2, \cdots, x_N\}$ denotes a set with N elements. In situations where there is no ambiguity as to their elements, the simplified notation $\{x_n\}$ is used to denote the same set.
- The notation $p(x)$ denotes a probability density function of x.
- U and V are used to denote the width and the height of an input image, respectively.
- $\mathbb{P}$ denotes the set of all pixels, namely, $\mathbb{P} = \{(u, v)^\top | u \in \{1, 2, \cdots, U\} \wedge v \in \{1, 2, \cdots, V\}\}$.
- A pixel $\boldsymbol{p}$ is given as an element of $\mathbb{P}$.
- An image is denoted by a vector-valued function $\boldsymbol{x} : \mathbb{P} \to \mathbb{R}^3$, where its output means RGB pixel values.
- The luminance of an image is denoted by a function $\ell : \mathbb{P} \to \mathbb{R}$.

A. Tone mapping

Display devices have limitations in their dynamic range, like imaging sensors. To present HDR images on such a display device, it is necessary to employ a tone mapping operator (TMO), which adjusts the luminance range of the images to fit within the dynamic range capabilities of the display [3]–[8].

The display luminance $\ell_d(p)$ of a tone-mapped image $\boldsymbol{y}$ in the linear domain is generally calculated from the world luminance $\ell_w(p)$ of an input HDR image $\boldsymbol{E}$ as

$$\ell_d(p) = f(\ell_s(p)) = f(\ell_w(p)\Delta t), \tag{1}$$

where $\Delta t > 0$ is a parameter that determines the brightness of the resulting image, and f is a non-linear tone curve. This process relates to actual photography, where Δt corresponds to the shutter speed that determines the exposure, and f represents the camera response function that adjusts the tone.

TMOs are primarily categorized into two types: global operators, which apply a common function f to all pixels, and local operators, which apply different functions f to individual pixels based on local information. In the case of Reinhard's global operator [3], the parameter Δt is given as

$$\Delta t = \frac{\alpha}{G(\ell_w|\mathbb{P})}, \tag{2}$$

where $\alpha \in [0, 1]$ is the parameter called "key value", which indicates subjectively if the scene is light, normal, or dark and it is typically set as the middle gray $\ell_{\text{gray}} = 0.18$ of the display luminance on a scale from zero to one. The geometric mean $G(\ell_w|\mathbb{P})$ of the world luminance $\ell_w(p)$ is calculated as follows:

$$G(\ell_w|\mathbb{P}) = \exp\left(\frac{1}{|\mathbb{P}|}\sum_{p\in\mathbb{P}} \log \ell_w(p)\right). \tag{3}$$

The TMO f is given as

$$f(\ell) = \frac{\ell}{1+\ell}\left(1 + \frac{\ell}{\ell_{\text{white}}^2}\right) \text{ for } \ell \in [0, \infty), \tag{4}$$

where parameter $\ell_{\text{white}} > 0$ determines the white point. Specifically, for all ℓ where $\ell \geq \ell_{\text{white}}$, $f(\ell) \geq 1$. The equality holds when $\ell = \ell_{\text{white}}$.

Although global TMOs are simple and widely used, it often results in a loss of local contrast inherent in HDR images. To address this issue, local tone mapping has gained attention. In particular, multiple exposure image fusion has become a focal point in this field.

B. Multi-exposure image fusion

Multi-exposure image fusion (MEF) is a technique to represent the high dynamic range of a scene as an LDR image. MEF combines several images of the same scene captured at different exposure levels by varying shutter speeds [9]–[14], [18], [19]. A stack of differently exposed images is referred to as "multi-exposure images." Given a stack of M exposure images $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M$, a fused image $\boldsymbol{y}$ is produced:

$$\boldsymbol{y} = \mathscr{F}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M), \tag{5}$$

where $\mathscr{F}$ indicates a function for fusing M images into a single image.

MEF methods can be used to tone-map HDR images by performing exposure compensation for a given HDR image to generate multi-exposure images, which are then combined. This exposure compensation is done by substituting M distinct shutter speeds $\Delta t_1, \cdots, \Delta t_M$ into (1). For this reason, MEF methods can be considered as a type of local tone mapping methods.

To produce high-quality fused images from given differently exposed images while preserving local contrast of original scenes, many research works have been focused on the development of the fusion method $\mathscr{F}$. Mertens et al.[10] proposed a multi-scale fusion scheme in which contrast, color saturation, and well-exposedness measures are used for computing fusion weights. In the work by Nejati et al. [18], base-detail decomposition is applied to each input image, and base and detail layers are then combined individually. An exposure fusion method based on sparse representation, which combines sparse coefficients with fused images, was also proposed in [12].

While research on fusion algorithm has progressed, Kinoshita et al. demonstrated that the quality of the final fused image is significantly influenced by not only fusion algorithm but also exposure values or shutter speeds $\Delta t_1, \cdots, \Delta t_M$ used for generating multi-exposure images $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M$. Furthermore, they proposed a scene-segmentation-based exposure compensation method to automatically determine the exposure of input images, ensuring that the fused image does not contain over- or under-exposed regions.

The key idea of this method is to segment the scene into M small regions based on luminance values by clustering them using a Gaussian mixture model (GMM). In the segmentation, the distribution of luminance ℓ_w is modeled with M Gaussians as

$$p(\log \ell_w) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\log \ell_w \mid \mu_m, \sigma_m) \tag{6}$$

and a pixel $\boldsymbol{p} \in \mathbb{P}$ is assigned to a subset $\mathbb{P}_m$ of $\mathbb{P}$ when m satisfies

$$m = \arg\max_k \frac{\pi_k \mathcal{N}(\log \ell(\boldsymbol{p}) \mid \mu_k, \sigma_k)}{\sum_{j=1}^{M} \pi_j \mathcal{N}(\log \ell(\boldsymbol{p}) \mid \mu_j, \sigma_j)}. \tag{7}$$

Given $\mathbb{P}_m$, the exposure of the mth image $\boldsymbol{x}_m$ is determined so that the mth region $\mathbb{P}_m$ achieves proper exposure.

C. Scenario

In this paper, we focus on the exposure values of multi-exposure images for MEF, like the aforementioned literatures [17], [16]. In the literatures, Δt_m is calculated so that the geometric mean of luminance on mth region $\mathbb{P}_m$ equals the middle-gray ℓ_{gray}, as

$$\Delta t_m = \frac{\ell_{\text{gray}}}{G(\ell_w|\mathbb{P}_m)}. \tag{8}$$

By determining exposures in this manner, it is possible to avoid information loss due to highlights or shadows because many MEF methods assign greater weight to pixels having luminance values close to middle gray during the fusion process. However, the resulting fused image tends to have the average luminance of each region $\mathbb{P}_m$ converging towards the middle gray (see Fig. 1(a)). This can lead to issues such as the inversion of luminance relationships and the loss of contrast between regions.

Because of such a situation, in this paper, we propose an novel exposure compensation method that preserves the relative luminance relationships between regions, for MEF-based tone mapping. Additionally, we also propose a method for calculating fusion weights to better reflect the effects of exposure compensation in the final fused image.

III. Proposed Tone Mapping Scheme

The goal of the proposed method is to represent a HDR scene with a wide dynamic range as an LDR image while preserving the relative luminance relationships and local contrast. To achieve this, this study first divides the scene into multiple regions, similar to conventional approaches. Subsequently, the exposure of multi-exposure images is determined to maximize the contrast between regions (see Fig. 1(b)). HDR images are subjected to exposure compensation to generate multi-exposure images, which are then composited to obtain the final image. During the compositing process, to reflect the effects of exposure compensation, the weights used in the composition are determined based on the exposure value and the luminance value of each pixel.

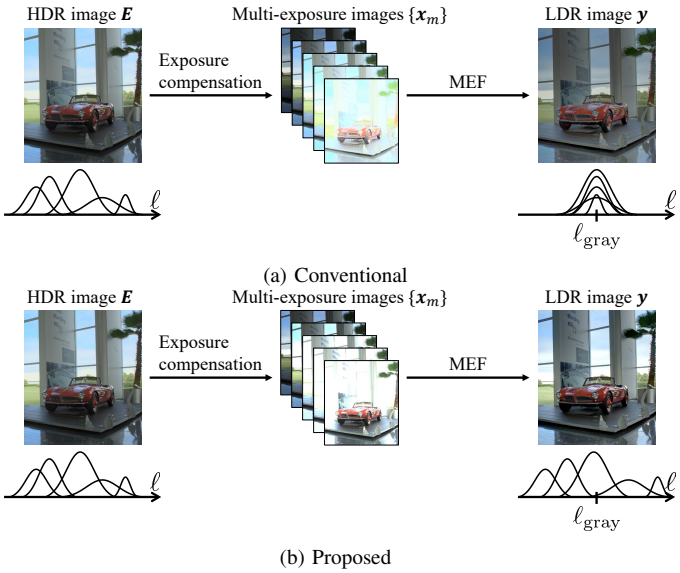


Fig. 1. Conventional and proposed exposure compensation based on scene segmentation

A. Exposure compensation

In this method, we consider a relative exposure value (EV) with respect to ℓ_{gray} as a reference point. In accordance with (1) and (2), we normalize the world luminance ℓ_w to 0 EV using the following equation:

$$\ell_s(\boldsymbol{p}) = \frac{\ell_{\text{gray}}}{G(\ell_w|\mathbb{P})} \ell_w(\boldsymbol{p}). \tag{9}$$

The distribution of the scaled luminance ℓ_s is modeled as a Gaussian mixture model (GMM) and $\mathbb{P}$ is divided into M regions $\{\mathbb{P}_m\}$, similar to (6) and (7):

$$p(\log \ell_s) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\log \ell_s \mid \mu_m, \sigma_m), \tag{10}$$

$$m = \arg\max_k \frac{\pi_k \mathcal{N}(\log \ell_s(\boldsymbol{p}) \mid \mu_k, \sigma_k)}{\sum_{j=1}^{M} \pi_j \mathcal{N}(\log \ell_s(\boldsymbol{p}) \mid \mu_j, \sigma_j)} \tag{11}$$

where we assume that mean values $\{\mu_m\}$ is sorted in ascending order and thus $\mathbb{P}_1$ and $\mathbb{P}_M$ indicates the darkest and the brightest regions, respectively. In the regions $\{\mathbb{P}_m\}$, a reference region $\mathbb{P}_{m_{\text{ref}}}$ is selected based on the mixture component that contributes most significantly to ℓ_{gray}. This selection is mathematically defined as:

$$m_{\text{ref}} = \arg\max_k \pi_k \mathcal{N}(\log \ell_{\text{gray}} \mid \mu_k, \sigma_k) \tag{12}$$

After the segmentation, we perform the exposure compensation considering the relative luminance relationships between regions. The exposure compensation factors $\{\Delta t_m\}$ are determined as the ratio of the target mean μ'_m to the mean μ_m of the m-th region:

$$\Delta t_m = \frac{\exp(\mu'_m)}{\exp(\mu_m)}. \tag{13}$$

Considering the dynamic range of the display luminance, we fix the target mean of the darkest region and the brightest region to $v_{\min}$ EV and to $v_{\max}$ EV, respectively:

$$\mu'_1 = \log 2^{v_{\min}} l_{\text{gray}}, \quad \mu'_M = \log 2^{v_{\max}} l_{\text{gray}}. \tag{14}$$

To maximize contrast between regions, we calculate the target mean evenly spaced in the log-luminance domain:

$$\mu'_m = \begin{cases} \dfrac{\mu_{m_{\text{ref}}} - \mu'_1}{m_{\text{ref}} - 1}(m-1) + \mu'_1 & 1 < m < m_{\text{ref}} \\ \mu_{m_{\text{ref}}} & m = m_{\text{ref}} \\ \dfrac{\mu'_M - \mu_{m_{\text{ref}}}}{M - m_{\text{ref}}}(m - m_{\text{ref}}) + \mu_{m_{\text{ref}}} & m_{\text{ref}} < m < M \end{cases}. \tag{15}$$

Finally, we apply the exposure compensation and tone mapping to obtain the luminance of the mth exposure image $\boldsymbol{x}_m$ as:

$$\ell_m(\boldsymbol{p}) = f(\ell_s(\boldsymbol{p})\Delta t_m), \tag{16}$$

where we use Reinhard's global TMO shown in (4) as the tone curve f with $\ell_{\text{white}} = 2^{v_{\text{white}}}\ell_{\text{gray}}$.

B. Weight calculation for multi-exposure image fusion

In our approach, we employ Mertens' well-knowm MEF method [10] as the fusion function $\mathscr{F}$. This fusion is performed by averaging the Laplacian pyramid decomposition of input multi-exposure images $\{\boldsymbol{x}_m\}$ as

$$\boldsymbol{L}_l[\boldsymbol{y}](\boldsymbol{p}) = \sum_{m=1}^{M} G_l[w_m](\boldsymbol{p})\boldsymbol{L}_l[\boldsymbol{x}_m](\boldsymbol{p}), \tag{17}$$

where $\boldsymbol{L}_l[\boldsymbol{x}_m](\boldsymbol{p})$ is the lth level of the Laplacian pyramid of the mth image $\boldsymbol{x}_m$ at pixel $\boldsymbol{p}$, and $G_l[w_m](\boldsymbol{p})$ is the lth level of the Gaussian pyramid of the fusion weight w_m for the mth exposure image at pixel $\boldsymbol{p}$.

To reflect the effect of the exposure compensation in the fused image, this method employs novel fusion weights. To calculate the weight w_m for the mth exposure image $\boldsymbol{x}_m$, we define the difference between the luminance ℓ_m and the target mean μ'_m as

$$d_m(\boldsymbol{p}) = g(\ell_m(\boldsymbol{p})) - g(\exp(\mu'_m)), \tag{18}$$

where $g(\ell) = (\gamma \circ f)(\ell)$ is a composite function of a gamma correction function γ and the tone curve f because the fusion in (17) is performed in the gamma-corrected sRGB domain. Using the difference $d_m(\boldsymbol{p})$, the weight w_m is calculated as

$$w_m(\boldsymbol{p}) = \frac{\exp(-(d_m(\boldsymbol{p}))^2)}{\sum_{k=1}^{M}\exp(-(d_k(\boldsymbol{p}))^2)}. \tag{19}$$

This weighting scheme emphasizes pixels whose luminance is close to the target mean μ'_m of their corresponding region $\mathbb{P}_m$.

C. Proposed procedure

The procedure of MEF-based tone mapping from an input HDR image $\boldsymbol{E}$ into $\boldsymbol{y}$, with the proposed exposure compensation and the fusion weight calculation, is summarized as follows:

1) Exposure compensation:
   a) Calculate the world luminance ℓ_w of the input image $\boldsymbol{E}$.
   b) Normalize the world luminance ℓ_w to 0 EV by using (9).
   c) Separate $\mathbb{P}$ into M areas $\{\mathbb{P}_m\}$ by using (10) and (11) to segment image x into M areas.
   d) Determine the target mean μ'_m for each region $\mathbb{P}_m$ by using (12), (14) and (15).
   e) Calculate the exposure compensation factor Δt_m for each region $\mathbb{P}_m$ by using (13).
   f) Tone map the input image $\boldsymbol{E}$ according to (16).
   g) Generate multi-exposure images $\{\boldsymbol{x}_m\}$ in the sRGB color space as

$$\boldsymbol{x}_m = \boldsymbol{\gamma}\left(\frac{\ell_m(\boldsymbol{p})}{\ell_w(\boldsymbol{p})}\boldsymbol{E}(\boldsymbol{p})\right), \tag{20}$$

   where $\boldsymbol{\gamma}$ is a function that applies gamma correction to each of the RGB values.

2) Multi-exposure fusion:
   a) Calculate fusion weights $\{w_m\}$ by using (18) and (19) from $\{\ell_m\}$ and $\{\mu'_m\}$.
   b) Fuse the Laplacian pyramid of the multi-exposure images $\{\boldsymbol{x}_m\}$ into that of the fused image $\boldsymbol{y}$ by using (17).
   c) Obtain an tone-mapped image $\boldsymbol{y}$ by reconstructing the fused Laplacian pyramid $\boldsymbol{L}_l[\boldsymbol{y}]$.

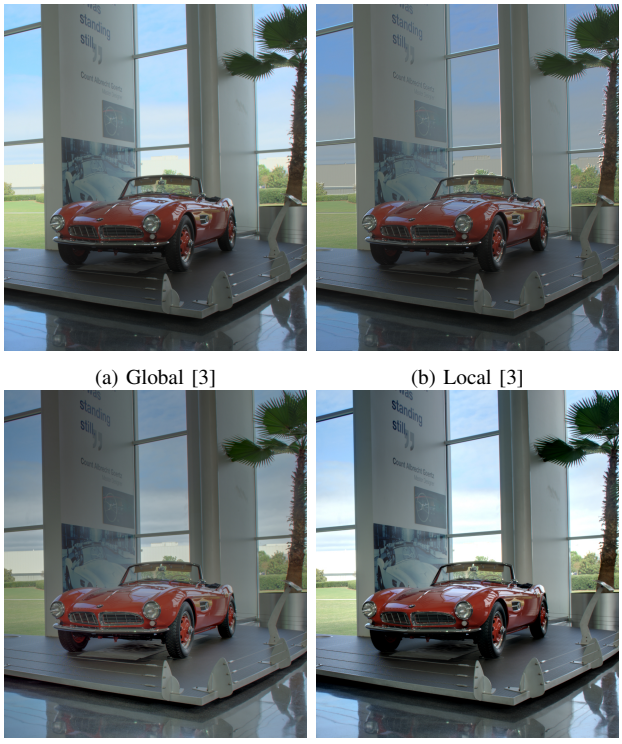
IV. Simulation

We evaluated the tone mapping scheme that uses the proposed segmentation-based exposure compensation and the proposed fusion weight, in terms of the quality of tone-mapped images $\boldsymbol{y}$. Hereinafter, the scheme is simply called the proposed method.

A. Simulation conditions

In the simulation, 20 HDR images in the literature [20] were used as input images $\boldsymbol{x}$. Each image $\boldsymbol{x}$ was tone-mapped by the proposed method and the quality of the resulting image $\boldsymbol{y}$ was evaluated.

The following four tone-mapping methods were compared in this paper: Reinhard's global tone mapping (Global) [3], Reinhard's local tone mapping (Local) [3], MEF with conventional exposure compensation (Conventional) [17], [16], and the proposed method (Proposed). For the two MEF based methods (Conventional and Proposed), Mertens' method [10] was used as a fusion function $\mathscr{F}$. In this simulation, we set the parameters in the proposed method as $v_{\min} = -3, v_{\max} = 1.5$, and $v_{\text{white}} = 2.5$.



(a) Global [3] (b) Local [3]

(c) Conventional [17], [16] (d) Proposed

Fig. 2. Tone-mapped images (“507”)



(a) Global [3] (b) Local [3]

(c) Conventional [17], [16] (d) Proposed

Fig. 3. Tone-mapped images (“LasVegasStore”)

To evaluate the quality of tone-mapped images by each of the four methods, objective quality assessments are needed. Typical quality assessments such as the peak signal to noise ratio (PSNR) and the structural similarity index (SSIM) are not suitable for tone-mapped images because they needs a reference tone-mapped image with the highest quality. We therefore used the tone mapped image quality index (TMQI) [21].

B. Results

Tone-mapped images from input images “507” and “LasVegasStore” are shown in Figs. 2 and 3, respectively. These figures show that the MEF-based conventional method produced grayish, low-contrast images have certain effects for enhancement. In contrast, the proposed method effectively tone mapped HDR images while preserving the contrasts.

Table I represents TMQI scores of tone mapped images generated by the compared methods for four representative scenes and their average and standard deviation for all 20 scenes. As shown in the table, we confirmed that the proposed method provided the highest average of TMQI scores. This means that the tone-mapped images produced by the proposed method has a higher quality than those produced by the other three methods.

TABLE I
TMQI SCORES FOR FOUR REPRESENTATIVE SCENES AND AVERAGE AND STANDARD DEVIATION FOR ALL 20 SCENES

Scene	Global [3]	Local [3]	Conv. [17], [16]	Prop.
507	0.8737	0.8185	0.8467	**0.9298**
LabTypewriter	0.8323	0.7881	0.7869	**0.8836**
LasVegasStore	0.9204	0.8707	0.8300	**0.9494**
WillyDesk	0.8564	0.8077	0.8677	**0.9326**
Average	0.8923	0.8388	0.8628	**0.9253**
Std.	0.0398	0.0386	0.0464	0.0395

V. Conclusion

In this paper, a scene-segmentation-based exposure compensation method was proposed for MEF based tone mapping of HDR images. For the exposure compensation, the proposed method segments an input HDR image into subregions and determines the exposure value for each subregion to maximize the contrast between regions while preserving the relative luminance relationships. Additionally, to better reflect the effects of exposure compensation in the final tone-mapped image, a novel fusion weight calculation method based on the difference between the luminance and its target was also proposed. In experiments, MEF-based tone mapping with the proposed compensation method and fusion weight calculation outperformed typical tone-mapping methods in terms of TMQI. Moreover, visual comparison results showed that the proposed segmentation-based exposure compensation and fusion weights are effective to produce images with high contrast.

References

[1] P. E. Debevec and J. Malik, “Recovering high dynamic range radiance maps from photographs,” in *Proceedings of ACM SIGGRAPH*, Los Angeles, CA, USA: ACM Press, 1997, pp. 369–378. DOI: 10.1145/258734.258884.

[2] C. Go, Y. Kinoshita, S. Shiota, and H. Kiya, "An Image Fusion Scheme for Single-Shot High Dynamic Range Imaging with Spatially Varying Exposures," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E102.A, no. 12, pp. 1856–1864, 2019. DOI: 10.1587/transfun.E102.A.1856.

[3] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 267–276, 2002. DOI: 10.1145/566654.566575.

[4] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 257–266, 2002. DOI: 10.1145/566654.566574.

[5] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 249–256, 2002. DOI: 10.1145/566654.566573.

[6] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive Logarithmic Mapping For Displaying High Contrast Scenes," *Computer Graphics Forum*, vol. 22, no. 3, pp. 419–426, 2003. DOI: 10.1111/1467-8659.00689.

[7] B. Gu, W. Li, M. Zhu, and M. Wang, "Local Edge-Preserving Multiscale Decomposition for High Dynamic Range Image Tone Mapping," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 70–79, 2013. DOI: 10.1109/TIP.2012.2214047.

[8] S. Paris, S. W. Hasinoff, and J. Kautz, "Local Laplacian filters: Edge-aware image processing with a Laplacian pyramid," *Communications of the ACM*, vol. 58, no. 3, pp. 81–91, 2015. DOI: 10.1145/2723694.

[9] S. Zokai and G. Wolberg, "Image registration using log-polar mappings for recovery of large-scale similarity and projective transformations," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1422–1434, 2005. DOI: 10.1109/TIP.2005.854501.

[10] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," *Computer Graphics Forum*, vol. 28, no. 1, pp. 161–171, 2009. DOI: 10.1111/j.1467-8659.2008.01171.x.

[11] A. Saleem, A. Beghdadi, and B. Boashash, "Image fusion-based contrast enhancement," *EURASIP Journal on Image and Video Processing*, vol. 2012, no. 10, pp. 1–17, 2012. DOI: 10.1186/1687-5281-2012-10.

[12] J. Wang, G. Xu, and H. Lou, "Exposure fusion based on sparse coding in pyramid transform domain," in *Proceedings of International Conference on Internet Multimedia Computing and Service*, ser. ICIMCS '15, Zhangjiajie city, Hunan, China: ACM Press, 2015, pp. 1–4. DOI: 10.1145/2808492.2808496.

[13] Z. Li, J. Zheng, Z. Zhu, and S. Wu, "Selectively detail-enhanced fusion of differently exposed images with moving objects," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4372–4382, 2014. DOI: 10.1109/TIP.2014.2349432.

[14] T. Sakai, D. Kimura, T. Yoshida, and M. Iwahashi, "Hybrid method for multi-exposure image fusion based on weighted mean and sparse representation," in *Proceedings of European Signal Processing Conference*, Nice: IEEE, 2015, pp. 809–813. DOI: 10.1109/EUSIPCO.2015.7362495.

[15] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proceedings of IEEE International Conference on Computer Vision*, Venice: IEEE, 2017, pp. 4724–4732. DOI: 10.1109/ICCV.2017.505.

[16] Y. Kinoshita and H. Kiya, "Scene segmentation-based luminance adjustment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 4101–4116, 2019. DOI: 10.1109/TIP.2019.2906501.

[17] Y. Kinoshita and H. Kiya, "Automatic exposure compensation using an image segmentation method for single-image-based multi-exposure fusion," *APSIPA Transactions on Signal and Information Processing*, vol. 7, no. e22, 2018. DOI: 10.1017/ATSIP.2018.26.

[18] M. Nejati, M. Karimi, S. R. Soroushmehr, N. Karimi, S. Samavi, and K. Najarian, "Fast exposure fusion using exposedness function," in *Proceedings of IEEE International Conference on Image Processing*, Beijing: IEEE, 2017, pp. 2234–2238. DOI: 10.1109/ICIP.2017.8296679.

[19] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust multi-exposure image fusion: A structural patch decomposition approach," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2519–2532, 2017. DOI: 10.1109/TIP.2017.2671921.

[20] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi, "Subjective quality assessment database of HDR images compressed with JPEG XT," in *Proceedings of International Workshop on Quality of Multimedia Experience*, Pylos-Nestoras: IEEE, 2015, pp. 1–6. DOI: 10.1109/QoMEX.2015.7148119.

[21] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 657–667, 2013. DOI: 10.1109/TIP.2012.2221725.

High-Quality Facial Pose Generation with Latent Space Processing

Wing-Ho Cheng*†, Wan-Chi Siu*†, *Life-FIEEE*, H. Anthony Chan*, *FIEEE*
* St. Francis University Hong Kong and
† The Hong Kong Polytechnic University

***Abstract*—Manipulating facial poses is challenging, especially when addressing significant pose variations. While extensive research has been dedicated to address large poses and manipulate various facial expressions, this frequently results in compromised image quality. The challenge may arise from non-linearity of the latent space. We must navigate a complex path along the high-quality image manifold and determine the optimal direction for the face rotation task, which may secure the most effective disentanglement. Moreover, the regularity of the latent space also affects directly the quality of the resulting image. In this paper, we have made a careful study of the latent space, and deliberately crafted our model to identify the complicated trajectory of rotating facial manipulation with exceptional disentanglement. Our facial pose generative model, aims at enhancing the quality of generated images while preserving the identity and fidelity and achieving better disentanglement. Data acquisition is another challenging aspect, requiring extensive preparation and meticulous setup. To address this, we suggest a flipping technique to mitigate dataset limitations. Ultimately, we strive to strike a balance between image quality and pose generation, ensuring that our results are both visually pleasing and accurately representing the desired facial pose.**

I. INTRODUCTION



Fig. 1 Illustration of our model to synthesize various poses from the input images (leftmost image in red).

Facial pose manipulation from a static image involves adjusting the orientation, rotation, and perspective of the human face while retaining its inherent features such as facial expression, hairstyle, and other attributes.

The proposed method entails altering the facial pose of a reference image without fundamentally changing the face's appearance, ensuring that the modified image appears as if captured from a different angle. Fig. 1 illustrate our model to synthesize various poses from an input image (i.e., leftmost image in red). Faces at the right-hand-side are the synthesized faces at different directions (i.e., different poses) ranging from -45, -30, 0, 30 and 45 degrees and their pose values are indicated above the faces.

Manipulating facial poses based on a static image is challenging. The intricacies inherent to the human face and the non-linear characteristics of pose alterations necessitate innovative approaches. Recent developments, particularly in generative models driven by deep learning, have paved the way for novel methodologies to address this difficult task. Unlike supervised tasks in machine learning, facial pose manipulation lacks a clear ground truth dataset for training and evaluation.

StyleGAN [1, 2] stands out as an innovative generative model that leverages the concept of the Generative Adversarial Network (GAN) [3]. This model can generate hyper-realistic human images. Beyond its ability to faithfully replicate facial features, StyleGAN exhibits considerable capabilities in various tasks. By manipulating the latent space, facial images can be transformed into different viewpoints or have their facial expressions altered.

Some models utilize StyleGAN as a generator to generate images [4, 5]. pSp [4] trains an encoder to identify the trajectory of semantic changes within the image domain and uses the StyleGAN generator to generate an image. InterfaceGAN [5] identifies a linear path in the latent space of StyleGAN for specific semantic attributes in the image domain. However, this approach produces artifacts, as the optimal path is typically not linear. Rotate and Render (RR) [6] obtains a 3D fit from the source image, rotates the 3D model to a different pose, and then reconstructs the image by applying texture using a trained model. CFR-GAN [7] employs several techniques to achieve face transformations for various tasks, including 3D face reconstruction and the Swap-R&R strategy. Although many methods are employed to handle pose manipulations; they frequently result in compromised image quality. BPRN [8] employs a Back-Projection method to enhance image quality, which is effective for super-resolution and various other applications. EFGPN [9] utilizes an edge-guided feature extraction encoder and achieved promising results.

A. Contributions

The primary contributions of this work can be outlined as follows. (1) We propose an unsupervised learning approach for pose manipulation through latent space manipulation. This approach eliminates the need to prepare ground truth images in

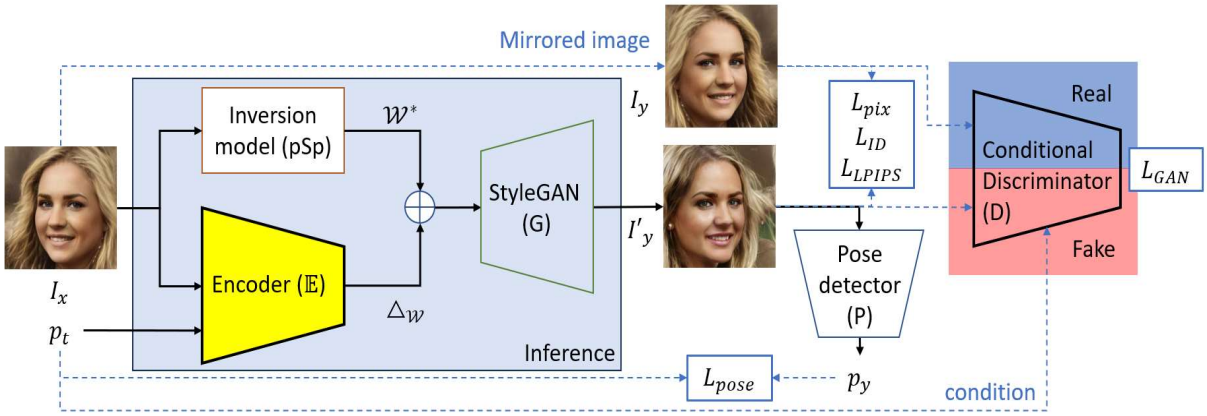


Fig. 2 Architecture of our system. The blue square at the right-hand side is the inference model. The whole model will be used during training.

training by utilizing a face-flipping strategy. (2) The model explicitly separates the learning of identity representation from the features of facial pose, enabling it to produce high-quality image of desired pose. (3) Our architecture utilizes an encoder and the StyleGAN [2] generator to synthesize the image. We also use a conditional discriminator [10] to restrict the generated images conforming it within the facial domain. In order to ensure that the generated face features align with the original image, the back projection technique [8] or the Back projection loss [8, 11] can be used. This architecture improves the quality of the generated image while preserving its identity. (4) The proposed model demonstrates outstanding image quality when generating facial poses on CelebA-HQ [12] and LFW [13] datasets. The high-quality images generated can effectively be utilized in various forms of creative entertainment and enhance the recognition rate of facial recognition systems. As shown below, the experimental results have verified that our approach give remarkable results compared with other approaches in the literature.

II. PROPOSED APPROACH

This paper is on pose generation of a face. We need face generation because we want to generate a face to have various pose appearances or want to make a real face photo to talk with various poses. The method focuses on manipulating the latent vector of an object in StyleGAN [2]. By incorporating a dedicatedly designed encoder to navigate pathways within the latent space, which includes the functionality of facial pose adjustments. The resultant images benefit from the high-quality synthesis capabilities of the StyleGAN model, ensuring realistic and detailed outcomes. Fig.2 shows the overall architecture of our approach to manipulating facial poses. In the diagram, the solid black lines illustrate the main data flow, while the dotted blue lines denote the path of data comparison with loss functions.

Inference: In Fig.2, I_x is the input face. Initially, the input image (I_x) undergoes inversion [4, 9] to produce a latent vector ($\mathcal{W}^*$) in the latent space. The Encoder (E) takes input from both the input image (I_x) and the target pose (p_t). The target pose (p_t) gives the angle to be turned. During the inference stage, we explicitly specify the desired output pose value, p_t. The pose layer is concatenated to the input image (3-layer of RGB) to form a 4-layer tensor and fed into the encoder to produce delta change ($\triangle_W$). This delta change will be added to the latent vector ($\mathcal{W}^*$) produced from the inversion model and fed into the StyleGAN generator (G) to generate the output image. It is worth noting that the pose detector and discriminator are not needed during the inference stage.

Training: Our objective is to train the system to change the pose of the input image I_x such that it matches closely the pose of the target image I_y. During training, the pose is determined by the target image, I_y using our novel image mirroring technique (to be explained shortly in the Training Strategy section) or a real number generated randomly. The input image undergoes the same processing as described in the inference stage to generate the output image. The generated image is compared with the target image to evaluate $\mathcal{L}_{pix}$, $\mathcal{L}_{ID}$ and $\mathcal{L}_{LPIPS}$ losses. Moreover, the pose of the generated image is compared with the target pose to calculate $\mathcal{L}_{pose}$ loss. Finally, the conditional discriminator (D) discriminates the difference between the mirrored image and the generated one, classifying them as real or fake under a condition based on the pose ($\mathcal{L}_{GAN}$). The solid-colored components in the model represent the modules to be trained (i.e., Encoder and Conditional Discriminator), while other components like the version model, StyleGAN (G), and the Pose Detector (P) remain fixed during training. Note that we have to train the encoder and the discriminator model such that they can be improved progressively. The process can be summarized as follows:

$$I'_y = \mathrm{G}(E(I_x, p_t) + \mathcal{W}^*) \quad (1)$$

where I_x is the input image, p_t is the target pose. E is the encoder. G represents the StyleGAN generator network that generates the output image I'_y (Fig. 2).

A. Architecture

The architecture of the encoder uses a Feature Pyramid Network (FPN) [14] built on top of a ResNet [15] backbone, which is similar to those in pSp [4, 9]. Initially, the encoder takes input from both the input image and the pose. The image and pose will be combined to create a 4-channel input. The FPN then extracts features from the input. These features are subsequently mapped to the 18 styles and inputted into StyleGAN generator.

The conditional discriminator used in our model base on the architecture of StyleGAN's discriminator while condition on the input pose (Conditional Discriminator (D) in Fig.2).

B. Training strategy

Novel image mirroring for training: Our methodology leverages the inherent symmetry of the human face to streamline the process of generating training pairs for pose manipulation. We intentionally mirror the image for training as shown in Fig. 2. To explain by an example, if the input image is facing +23 degrees, the target image will be laterally flipped image, and the pose of the target image will be -23 degrees. We use this input image and flipped input image as training pairs to train our model. This approach is grounded in the observation that facial features often exhibit bilateral symmetry, making it feasible to generate a close approximation of the ground truth image solely through mirroring techniques. This also simplifies the data collection process significantly, as well as eliminates the requirement for extensive manual annotation.

C. Loss function

Our encoder employs a variety of loss functions.

1) Pixelwise loss

We utilize the L2 norm distance to align the generated image with its corresponding counterpart, calculated through the following equations:

$$L_{pix} = \left\| I_y - I'_y \right\|_2 \quad (2)$$

where I_y and I'_y are the target image and the generated image respectively and $\|.\|_2$ is the L2 norm distance. Our proposed method has designed that the target image, I_y, will be a mirrored version of I_x if the target pose is facing the opposite direction to the input image or else it will be I_x itself. While other losses like ID loss, perceptual loss, or GAN loss prioritize capturing specific facial features from the source image, they may overlook finer skin details such as facial lines and pores. Using the L2 norm can improve the overall quality and fidelity of the generated face images.

2) ID loss

We employ the ID loss to steer the generation process, maintaining the identity of the source image. The identity of a human face encompasses various factors, such as the shape and size of facial components like the eyes, nose, ears, and mouth, as well as their positioning and coloration. The ID loss aids in capturing these intricate details. This loss quantifies the feature differences between the target and generated images, employing a feature extraction model built on Arcface [16] with a ResNet [15] backbone. It measures the cosine similarity between the output and target images.

$$L_{ID} = 1 - \left\| \mathrm{R}(I_y), \mathrm{R}(I'_y) \right\|_{cos} \quad (3)$$

where R is a pre-trained ArcFace network [16] and $\|.\|_{cos}$ is the cosine similarity.

3) Perceptual loss

In addition to maintaining identity consistency, we preserve the perceptual similarities in the generated images. The Perceptual loss shares similarities with the ID loss but emphasizes capturing details that enhance visual perception. To achieve this, we integrate the LPIPS loss [17] into our training framework. This loss metric helps the model learn and reinforce perceptual similarities between the generated face images and their corresponding target face images.

$$L_{LPIPS} = \left\| \mathrm{K}(I_y) - \mathrm{K}(I'_y) \right\|_2 \quad (4)$$

where K denotes the perceptual feature extractor [17] and $\|.\|_2$ is the L2 norm distance.

4) Conditional GAN loss [10]

To guarantee the high quality and photorealism of the generated image, we employ a GAN discriminator for validation purposes. The GAN loss contributes in capturing the overall distribution of facial images and ensures that the output closely aligns with facial properties. By adopting the idea of conditional GAN [10], we utilize pose conditioning to separate the pose feature from the facial feature, enhancing disentanglement from the generated face image. The adversarial loss can be calculated as:

$$L_{GAN} = -log\left(\mathrm{D}(I'_y, p_t)\right) \quad (5)$$

where I'_y and p_t are the generated image and target pose respectively. D represents the discriminator network conditioned on the pose of the image that is leaned to distinguish whether the generated face appears natural and realistic.

5) Pose loss

To evaluate the accuracy of the facial pose transformation, we employ a pre-trained facial pose predictor network, Hopenet [18]. This network analyzes a facial image and determines its pose. We then compute the L2 loss by measuring the difference between the facial poses of the generated and target images as follows:

$$L_{pose} = \left\| \mathrm{P}(I_y) - \mathrm{P}(I'_y) \right\|_2 \quad (6)$$

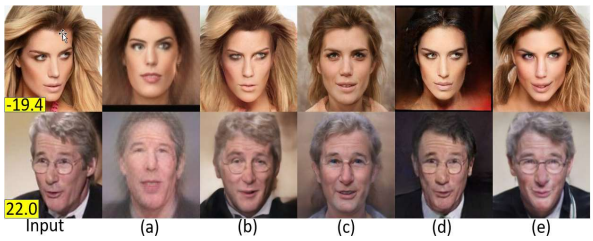


Fig. 3 Visual comparison with other SOTA models for the frontalization task using the CelebA-HQ (Upper row) and LFW (Lower row) dataset. (a) CFR-GAN [7], (b) InterfaceGAN [5], (c) pSp [4], (d) RR [6] and (e) ours.

where P is the facial pose prediction network, Hopenet [18] and $\|.\|_2$ is the L2 loss.

6) Back projection loss [8]

This loss function operates as a revert of the normal forward pass, where after generating the target pose, the generated image is then required to revert to the original pose. This method offers several advantages. Firstly, in the absence of a supervised or target facial image, we obtain a reversed target facial image that should ideally match the original input face image. This allows for precise comparison with the target image, facilitating accurate loss calculation.

Furthermore, this approach imposes constraints on the model to preserve the identity of the face image during the transformation process. If the model diverges from the correct transformation path, it becomes difficult to revert the image back to its original form. By enforcing this constraint through Back projection loss [8, 11], the model learns to produce high-quality face image with improved both fidelity and identity. We denote the Back projection loss as L_{cycle}. The same set of losses introduced in forward pass is applied to cycle pass.

7) Total loss

In summary, the loss for the forward pass is expressed as follow:

$$L_{forward} = w_{pix} \cdot L_{pix} + w_{ID} \cdot L_{ID} + w_{LPIPS} \cdot L_{LPIPS} + w_{GAN} \cdot L_{GAN} + w_{pose} \cdot L_{pose} \quad (7)$$

where w_{pix}, w_{ID}, w_{LPIPS}, w_{GAN}, w_{pose} are constants defining the loss weights.

To alleviate the impact of a lesser accurate approximation of the target image when the target pose is randomly generated, the weighting for pixel-wise loss, w_{pix}, will be reduced.

Combining both forward and cycle passes, the total loss is given by:

$$Total\ Loss = L_{forward} + w_{cycle} \cdot L_{cycle} \quad (8)$$

where w_{cycle} is the weight of the Back projection loss.

III. Experimental Works

A. Datasets

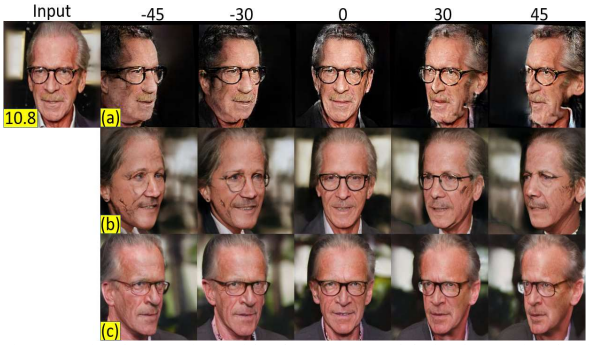


Fig. 4 Visual comparison with other SOTA models for the rotational task using the CelebA-HQ dataset. The input images are depicted in the leftmost column, while the synthesized images are presented on the right-hand side. (a) RR [6], (b) InterfaceGAN [5] and (c) ours.

We employed the unlabeled image dataset FFHQ [1] for training, consisting of 70,000 high-quality human face images sourced from the Flickr photo-sharing platform. This dataset is highly valued for its high resolution and extensive size. The images are already aligned with each image having a resolution of 1024x1024 pixels. The facial images are displayed in various poses. It proves particularly beneficial for applications requiring detailed facial features and realistic representations.

B. Implementation details

We implemented our model based on PyTorch and conducted our experimental work using training on a single GPU RTX3090 with 24G VRAM. We greatly appreciate the source code shared by Alaluf, et al. [19], which serves as a reference for our implementation. During the training phase, we adopted the Adam [20] optimizer with a batch size of 6. Throughout the training process, we maintained fixed parameters for the inversion model, StyleGAN generator [2], and the pose detection model, Hopenet [18]. The pre-trained models were obtained from official channels and loaded to these fixed models before training. We solely trained the encoder and the conditional discriminator network. We utilized FFHQ [1] as our training dataset, comprising 70,000 high-resolution faces with a resolution of 1024x1024.

In the testing phase, we employed CelebA-HQ [12] and LFW [13] to assess our trained network and compared methods. All experiments were conducted using the model output from unseen images during the training phase. Obviously, the discriminator will not be included in the testing phase.

C. Qualitative Comparison

We conducted a comparison with four state-of-the-art models with source code available. Two models, InterfaceGAN [5] and pSp [4] which utilize StyleGAN as a generator which share similarities with our architecture. The other two models: Rotate and Render (RR) [6], and CFR-GAN [7], employ different approaches for comparison. A short demonstration is included in the following webpage:

https://cis.sfu.edu.hk/2024.RotationalFace/index.html.

Our analysis in Fig. 3 indicates that the quality of our output surpasses that of the compared models in both datasets. Our model exhibits superior disentanglement across various aspects.

For instance, while the facial details and hairstyle preservation in the compared models (Column (a)-(d) in Fig.3) are inadequate, our model (Column (e) in Fig. 3) closely recovers the original input photo. Our model has produced the most photorealistic output compared to all other models. Artifacts rarely appear in the images generated by our model.

When outputs of Rotate and Render (RR)'s [6] (Column (d) in Fig.3) possess sufficient realism, they tend to lack the refinement and clarity found in higher-resolution images. This lower resolution may result in a loss of details and the overall visual quality.

InterfaceGAN [5] (Column (b) in Fig.3) adjusts the pose based solely on the original input and there is no specific input control mechanism to precisely regulate the output pose of the generated image. In order to compare this approach, we have to choose the frontal pose solely by human inspection. The output image quality by InterfaceGAN is significantly influenced by the amount of pose that deviates from the original photo. The synthesized image sometimes displays artifacts or image distortion due to diverges from the optimal image manifold (e.g. Fig.3. column (b) of 2nd row).

The pSp model [4] (Column (c) in Fig.3) can generate highly detailed images, sometimes these details may appear exaggerated, leading to images that deviate from reality. In their model, features such as shape, expression, and hairstyle were averaged and exhibited similarity across different instances. In contrast, our approach (Column (e) in Fig.3) better preserves the identity and achieves a higher level of disentanglement. Interestingly, pSp outputs often exhibit blurred and plain backgrounds. In contrast, our model offered vibrant color representation closer to the original and effectively maintained the overall appearance and atmosphere of the image, resulting in outputs that closely resemble the original photos.

Our synthesized output surpasses CFR-GAN [7] (Column (a) in Fig.3) in terms of realism and fidelity compared to the original image.

For the task of altering the pose of an image, models like pSp, CFR-GAN does not provide the feature to rotate a face to a specific view. Instead, we focus our comparison on RR [6] and InterfaceGAN [5], as they are capable of generating poses different from the frontal pose, as depicted in Fig.4. Let us showcase our model results across a range of pose angles ranging form -45, -30, 0, +30 to +45 degrees, and compare them with results from other models. Once more, our model (Fig.4 row (c)) demonstrates superior image quality compared to other models, and this distinction becomes apparent in larger pose generation.

All models exhibited varying degrees of artifacts or distortions as the magnitude of the generated pose increases. However, our model shows better disentanglement capabilities; for instance, glasses intermittently appeared or disappeared during the generation of different poses (Fig.4 row (b)). In order to reduce variations across different poses of the same image, such as stochastic fluctuations in hair generation, we have fixed the per-pixel noise input in the StyleGAN generation settings. Typically, distortions became bad at the larger poses on the left and right sides.

Although RR (Fig.4 row (a)) preserves many details and fidelity, it deviates considerably from the original image. Its image quality is realistic only around the frontal position, however, there were significant changes to the background, diverging from the original photos. Furthermore, the face undergoes more distortion at larger poses, and our model always does a better job for preserving the hairstyle and the prevention of artifact generation.

InterfaceGAN (Fig.4 row (b)) adjusts the pose based on the original input and the magnitude cannot be precisely controlled. Furthermore, generating poses with larger differences have proved to be more challenging and resulted in increased distortion and the generation of more artifacts. InterfaceGAN employs a linear interpolation of the original pose in the latent space. A larger pose difference resulted in larger discrepancies from the optimal facial latent representation, especially when the optimal path has deviated from the linear trajectory. This also suggested that the optimal path is non-linear in nature.

D. Quantitative Comparisons

It is challenging to perform a quantitative comparison of image outputs when the pose varies. The absence of an accurate ground truth for comparison also further complicates quantitative comparison. Generation of large poses is not the sole consideration; image quality is most important.

TABLE 1.
FID Scores (CelebA-HQ). The best model is highlighted in red while the 2nd best is highlighted in blue.

Method	Avg	0	-45	-30	-15	15	30	45
pSp [4]	-	67.8	-	-	-	-	-	-
CFR-GAN [7]	-	187.3	-	-	-	-	-	-
RR [6]	118.81	84.3	155.1	133.9	105.4	100.8	119.5	132.7
Ours	**23.3**	**23.2**	**30.6**	**20.7**	**20.2**	**21.0**	**21.2**	**26.2**

TABLE 2.
FID Scores (LFW). The best model is highlighted in red while the 2nd best is highlighted in blue.

Method	Avg	0	-45	-30	-15	15	30	45
pSp [4]	-	145.2	-	-	-	-	-	-
CFR-GAN [7]	-	106.2	-	-	-	-	-	-
RR [6]	58.6	50.9	75.1	51.3	49.1	50.1	54.4	79.1
Ours	**34.4**	**35.4**	**41.5**	**32.5**	**32.2**	**32.4**	**31.6**	**35.4**

Let us compare the frontalization output from different models using the Fréchet Inception Distance (FID) score [21] (lower value indicated better quality) and the Rank-1 recognition rate. To evaluate the Rank-1 recognition rate, we extracted features of the images in both datasets using LightCNN [22] for the original images and the generated pose

image. We then compared the original and generated images, selecting the pair with the top cosine similarity to compare their identities. Since CelebA-HQ lacks labeling in the datasets, each image is considered a unique identity, even if there are instances of the same identity present.

While InterfaceGAN can rotate images to different angles, it is hard to determine the output pose angle of the generated image. As a result, we did not use it for quantitative comparison. Note that only RR can rotate images into different poses, and there were not data available for other models.

Table 1 and Table 2 present the FID [21] scores of various models for different generated poses, our model gains significant superiority over other models, indicating that it achieves the highest similarity in terms of overall similarity and color distribution compared to the original datasets.

TABLE 3.
Rank-1 Accuracy (CelebA-HQ). The best model is highlighted in red while the 2nd best is highlighted in blue.

Method	Avg	0	-45	-30	-15	15	30	45
pSp [4]	-	**99.5**	-	-	-	-	-	-
CFR-GAN [7]	-	80.2	-	-	-	-	-	-
RR [6]	92.5	99.4	79.5	94.7	98.9	98.8	94.9	81.1
Ours	**97.0**	98.9	**90.5**	**97.9**	**99.3**	**99.1**	**98.6**	**94.9**

TABLE 4.
Rank-1 Accuracy (LFW). The best model is highlighted in red while the 2nd best is highlighted in blue.

Method	Avg	0	-45	-30	-15	15	30	45
pSp [4]	-	**98.3**	-	-	-	-	-	-
CFR-GAN [7]	-	53.8	-	-	-	-	-	-
RR [6]	83.4	97.4	66.5	87.4	94.9	93.7	84.3	59.5
Ours	**94.3**	97.9	**86.2**	**96.0**	**98.3**	**98.0**	**96.6**	**87.1**

Table 3 and Table 4 list the comparison of the Rank-1 Accuracy with other state-of-the-art models. For all models, the recognition rates were generally higher around frontal poses and the rate tended to decrease at larger poses. It may be because the datasets were dominated by frontal-facing images. In general, rotation to frontal image means fewer pose variations.

Our model has the highest average accuracy of 94.3% (CelebA-HQ) (Table 3. Row 4 Column 1) and 97.0 (LFW) (Table 4. Row 4 Column 1). Note that the result may be impacted by the presence of multiple images for the same identity existing within the CelebA-HQ dataset. Since the dataset lacks labeling, each image is treated as a distinct identity.

Interestingly, our model achieves its highest recognition rate for faces with small pose angle of -15 degrees, reaching 98.3% (Table 3. Row 4 Column 5) and 99.3% (Table 4. Row 4 Column 5) in CelebA-HQ datasets and LFW datasets respectively. Furthermore, our model outperforms other models notably for faces with large poses. In large poses, our model sustains exceptional image quality, demonstrating its ability to successfully disentangle facial features and maintain identity after pose rotation. Consequently, our model achieves a recognition rate significantly surpassing that of the other models.

IV. Conclusions

In conclusion, our research primarily focuses on manipulating the latent space of StyleGAN [2] to transform the original image into a different pose. By training the model in an unsupervised manner and incorporating image-flipping techniques, we aim to enhance its ability to learn diverse facial properties and enhance its image generation quality for various poses. Our model integrates Back projection loss [8, 11] to enhance output quality and leverages conditional GANs [10] to improve disentanglement during face rotation, thereby preserving identity. One future direction is the incorporation of facial expressions into the rotated faces. This not only would improve the visual fidelity of the generated images but also expand the potential use cases in fields such as virtual reality, animation, and human-computer interaction.

V. Acknowledgment

This work is partly supported by the St. Francis University (1SG200206) and UGC Grant (UGC/FDS11/E05/22) of the Hong Kong Special Administrative Region.

References

[1] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401-4410.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110-8119.

[3] I. Goodfellow et al., "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[4] E. Richardson et al., "Encoding in style: a stylegan encoder for image-to-image translation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2287-2296.

[5] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9243-9252.

[6] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5911-5920.

[7] Y.J. Ju, G.H. Lee, J.H. Hong, and S.W. Lee, "Complete face recovery gan: Unsupervised joint face rotation and de-occlusion from a single-view image," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 3711-3721.

[8] Z.S. Liu, W.C. Siu, and Y.L. Chan, "Joint back projection and residual networks for efficient image super-resolution," in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2018: IEEE, pp. 1054-1060.

[9] X. Cheng, W.C. Siu and J. Yang, "Large-Scale Blind Face Super-Resolution via Edge Guided Frequency Aware Generative Facial Prior Networks", Proceedings, pp. 1638-1643, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC'2022), Chiang Mai, Thailand, 2022, doi: 10. 23919/APSIPAASC55919.2022.9980270.

[10] P. Isola, J.Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125-1134.

[11] J.Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223-2232.

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2017.

[13] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition, 2008.

[14] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117-2125.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4690-4699.

[17] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586-595.

[18] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 2074-2083.

[19] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "Only a matter of style: Age transformation using a style-based regression model," ACM Transactions on Graphics (TOG), vol. 40, no. 4, pp. 1-12, 2021.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," Advances in neural information processing systems, vol. 30, 2017.

[22] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," IEEE transactions on information forensics and security, vol. 13, no. 11, pp. 2884-2896, 2018.

Hyperspectral Anomaly Detection Using Robust Principal Component Analysis with Autoencoding Adversarial Networks

Atsuya Emoto and Ryo Matsuoka
The University of Kitakyushu,
Faculty of Environmental Engineering,
1-1 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka, Japan.

***Abstract*—This paper introduces an unsupervised anomaly detection method for hyperspectral (HS) images, integrating robust principal component analysis (RPCA) and deep autoencoding adversarial Networks (AEAN). The proposed method is solved using a plug-and-play alternating direction method of multipliers (PnP-ADMM). This method uses AEAN as a proximity operator to promote low-rankness in reconstructed spectral data, effectively distinguishing anomalous pixels from the background. Moreover, anomaly detection performance is enhanced using the anomaly map estimated by the proposed method as weights for the WLRX detector. Experimental results show that our method outperforms some existing methods in detection accuracy.**

I. Introduction

Hyperspectral (HS) images, with several hundred spectral bands, are widely used in earth surface classification, crop health assessment, environmental monitoring, and mineral exploration. Detecting anomalies, which are small areas with unique spectral characteristics, poses a challenge and often relies on background modeling or statistical methods. Various approaches have been proposed, including statistical methods, subspace methods, linear mixture models, and deep neural networks [1]–[10].

A representative example is the Reed-Xiaoli (RX) anomaly detector [3], which detects anomalous pixels in HS data. The RX detector calculates the anomaly score of each pixel based on the Mahalanobis distance between the pixel and the background, with higher scores indicating a higher probability of being an anomaly. The spectral mean and covariance of the HS data characterize the background. Many advanced methods based on RX have been proposed, e.g., [1], [2].

Some deep learning-based methods have been proposed for HS anomaly detection [4]–[7], often using autoencoder (AE) models. These methods train AEs with pixels from roughly estimated background regions, resulting in low reconstruction errors for background regions and high errors for anomalous regions. Anomalous pixels are detected based on these errors. Adversarial learning between original and reconstructed spectra improves accuracy, termed deep autoencoding adversarial Networks (AEAN) [7]. Combining AEAN with the local weighted RX (WLRX) detector [2] further enhances detection accuracy. However, accurate background estimation is challenging, and training with inaccurate estimates can reduce detection accuracy.

In HS anomaly detection, the observed tensor is assumed to comprise a low-rank background and a sparse anomaly component and is decomposed using robust principal component analysis (RPCA) [8]–[10]. RPCA can be applied to the 2D matrix, utilizing a nuclear norm to enhance the spectral low-rankness. The 3D approach implements a tensor nuclear norm (TNN) to improve low-rank properties in both spatial and spectral domains. However, in complex scenes, TNN often leads to checkerboard artifacts that diminish the accuracy of anomaly detection.

In this paper, we propose a novel HS anomaly detection technique that integrates RPCA and AEAN. We employ the AEAN model to extract the spectral characteristics of the background regions from the observed data. Furthermore, to improve the separation accuracy of background and anomaly components and to enable robust anomaly detection, we incorporate the AEAN model as a regularization for the background component in the RPCA optimization problem. The proposed problem can be solved using the plug-and-play alternating direction method of multipliers (PnP-ADMM). Experiments show that the proposed method can more accurately detect anomaly pixels than some conventional methods.

II. Preliminaries

In this paper, bold-faced lowercase and uppercase letters indicate vectors and matrices, respectively. Real-valued N-dimensional vector spaces denotes by $\mathbb{R}^N$. We define the $N \times M$ real-valued matrix as $\mathbb{R}^{N\times M}$. Three-dimensional tensor is denoted by $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ and its $\{i,j,k\}$-th element is denoted by $[\boldsymbol{\mathcal{X}}]_{i,j,k}$.

A. Plug-and-Play Alternating Direction Method of Multipliers (PnP-ADMM)

Alternating Direction Method of Multipliers (ADMM) [11], [12] is a proximal splitting algorithm that can treat convex optimization problems of the form

$$\min_{\mathbf{x}\in\mathbb{R}^{N_1},\, \mathbf{z}\in\mathbb{R}^{N_2}} F(\mathbf{x}) + G(\mathbf{z}) \;\text{ s.t. }\; \mathbf{z} = \mathbf{L}\mathbf{x}, \tag{1}$$

where F and G are usually assumed to be a quadratic and proximable function, respectively, and $\mathbf{L} \in \mathbb{R}^{N_2\times N_1}$ is a

matrix with full-column rank. For any $\mathbf{x}^{(0)} \in \mathbb{R}^{N_1}, \mathbf{z}^{(0)} \in \mathbb{R}^{N_2}, \mathbf{b}^{(0)} \in \mathbb{R}^{N_2}$ and $\rho > 0$, the ADMM algorithm is given by

$$\begin{cases} \mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}} \left\{ F(\mathbf{x}) + \dfrac{\rho}{2} \|\mathbf{z}^{(t)} - \mathbf{L}\mathbf{x} - \mathbf{b}^{(t)}\|_2^2 \right\}, \\ \mathbf{z}^{(t+1)} = \arg\min_{\mathbf{z}} \left\{ G(\mathbf{z}) + \dfrac{\rho}{2} \|\mathbf{z} - \mathbf{L}\mathbf{x}^{(t+1)} - \mathbf{b}^{(t)}\|_2^2 \right\}, \\ \qquad = \mathrm{prox}_{1/\rho G}(\mathbf{L}\mathbf{x}^{(t+1)} + \mathbf{b}^{(t)}) \\ \mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} + \mathbf{L}\mathbf{x}^{(t+1)} - \mathbf{z}^{(t+1)}, \end{cases} \tag{2}$$

where the superscript (t) denotes the iteration number. The update of $\mathbf{z}$ reduces to the proximity operator of the function $G(\cdot)$[1]. The sequence generated by (2) quickly converges to an optimal solution of (1).

In PnP-ADMM [14], [15], the solution of the sub-problem w.r.t. $\mathbf{z}$ (assuming $\mathbf{L}$ is the identity matrix) is replaced by an off-the-shelf noise removal algorithm, to yield

$$\mathbf{z}^{(t+1)} = \mathcal{D}_\sigma \left(\mathbf{x}^{(t+1)} + \mathbf{b}^{(t)} \right), \tag{3}$$

where $\mathcal{D}_\sigma$ denotes the Gaussian denoiser with σ the standard deviation of the additive white Gaussian noise (AWGN). Auto-encoder networks are often used as denoiser [15].

B. Robust Principal Component Analysis

RPCA decomposes observational data, represented as a two- or three-dimensional tensor, into a low-rank tensor and a sparse tensor. When applied to HS images, RPCA enhances the low-rank structure of the tensor while preserving spatial correlation.

Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ (n_1 and n_2 are the height and width of the HS image, and n_3 is the number of the spectral bands) be an observed HS image. The standard RPCA optimization problem [8], [9] is defined by

$$\min_{\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{S}}} \|\boldsymbol{\mathcal{L}}\|_* + \lambda \|\boldsymbol{\mathcal{S}}\|_1 \quad \text{s.t.} \quad \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{S}}, \tag{4}$$

where $\boldsymbol{\mathcal{L}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ are the background and anomaly components, respectively, and $\lambda > 0$ is a balancing weight to control the two terms. The first term is the regularization term of $\boldsymbol{\mathcal{L}}$ based on the nuclear norm, which promotes the low-rankness of $\boldsymbol{\mathcal{L}}$. The second term is the regularization term of $\boldsymbol{\mathcal{S}}$ based on the ℓ_1-norm, which promotes sparsity of $\boldsymbol{\mathcal{S}}$.

III. Proposed Method

The proposed method aims to achieve high-precision and robust anomaly detection in HS images through three steps: training the AEAN model using denoised HS images by singular value decomposition (SVD), solving the optimization problem based on RPCA with AEAN using PnP-ADMM, and generating the anomaly map from the estimated anomaly component by combining the WLRX detector [1], [2].

[1]The proximity operator [13] is a key tool of proximal splitting techniques. Let $\mathbf{x} \in \mathbb{R}^N$ be an input vector. For any $\gamma > 0$, the proximity operator of f over $\mathbb{R}^N$ is defined by $\mathrm{prox}_{\gamma f}(\mathbf{x}) := \arg\min_{\mathbf{y} \in \mathbb{R}^N} f(\mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{y}\|^2$.

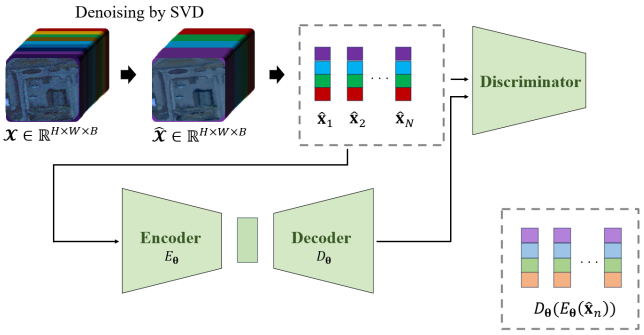


Fig. 1. The 1D-AEAN model and its training.

TABLE I
THE PROPOSED 1D-AEAN ARCHITECTURE. PARAMETER NOTATION: KERNEL SIZE × NUMBER OF INPUT CHANNELS × NUMBER OF OUTPUT CHANNELS

Nets	Layer	Parameters	BN	Activation
Encoder	Conv1D	11 × 1 × 5	YES	LReLU
	Conv1D	9 × 5 × 10	YES	LReLU
	Conv1D	7 × 10 × 15	YES	LReLU
	Conv1D	5 × 15 × 20	YES	LReLU
Decoder	Conv1DTranspose	5 × 20 × 15	YES	LReLU
	Conv1DTranspose	7 × 15 × 10	YES	LReLU
	Conv1DTranspose	9 × 10 × 5	YES	LReLU
	Conv1DTranspose	11 × 5 × 1	NO	Sigmoid
Discriminator	Conv1D	11 × 1 × 5	YES	LReLU
	Conv1D	9 × 5 × 10	YES	LReLU
	Conv1D	7 × 10 × 15	YES	LReLU
	Conv1D	5 × 15 × 20	YES	LReLU
	Dense	20 × 1	NO	Sigmoid

A. Training the AEAN Model

We train the 1D model of the AEAN (1D-AEAN) to learn an encoder-decoder structure for reconstructing background components from HS images. The conventional method [7] uses the modified kernel RX algorithm [3] to eliminate anomaly pixels and identify background pixels for training. However, RX is noise-sensitive and computationally expensive. To address this, we preprocess using SVD to remove noise and anomalous pixels, generating training data (Fig. 1).

Let an observed HS image $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{H \times W \times B}$ be rearranged into a 2D matrix $\mathbf{X} \in \mathbb{R}^{N \times B}$, where $N = HW$. The matrix $\mathbf{X}$ is denoised using SVD by retaining the top 90% of the singular components. The process involves centering the data, performing SVD, retaining the significant singular values, and reconstructing the matrix. This denoised HS data is then used to train the 1D-AEAN model.

Table I summarizes the details of the architecture of the proposed 1D-AEAN model.

B. Solving the proposed optimization problem

Next, we describe the proposed optimization problem based on RPCA with AEAN. In the proposed method, we enhance the accuracy of estimating the sparse tensor $\boldsymbol{\mathcal{S}}$, which represents the anomalous regions, by modeling the low-rank nature

of the background regions using a low-dimensional model (low-dimensional tensor $\boldsymbol{\mathcal{L}}$) based on the 1D-AEAN model.

To adopt the 1D-AEAN model as the proximity operator in the PnP-ADMM framework, we define $\mathcal{R}_{\text{AEAN}}$ as the regularization term in the minimization problem. We formulate the proposed optimization problem as follows.

$$\min_{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{S}}} \mathcal{R}_{\text{AEAN}}(\boldsymbol{\mathcal{L}}) + \lambda \|\boldsymbol{\mathcal{S}}\|_{2,1}^{(\mathcal{G})} \quad \text{s.t. } \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{S}}, \tag{5}$$

where $\lambda > 0$ is a balancing weight of two terms.

The second term is the group ℓ_1-norm of $\boldsymbol{\mathcal{S}}$, defined as

$$\|\boldsymbol{\mathcal{S}}\|_{2,1}^{(\mathcal{G})} := \left\| \left\{ \|\boldsymbol{\mathcal{S}}^{(\mathcal{G}_1)}\|_2, \ldots, \|\boldsymbol{\mathcal{S}}^{(\mathcal{G}_N)}\|_2 \right\} \right\|_1, \tag{6}$$

where $\{\mathcal{G}_1, \ldots, \mathcal{G}_N\} \in \mathcal{G}$ are the sets of indices of the one-dimensional vectors in the spectral direction, and $\mathcal{G}_n$ has the index sets of the spectral vector of the n-th pixel.

To find a solution of (5), we employ the PnP-ADMM algorithm described in Section II-A. The minimization problem (5) is not directly applicable to PnP-ADMM. Thus, we reformulate it as

$$\min_{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{S}}} \mathcal{R}_{\text{AEAN}}(\boldsymbol{\mathcal{L}}) + \lambda \|\boldsymbol{\mathcal{S}}\|_{2,1}^{(\mathcal{G})} + \iota_{\{\boldsymbol{\mathcal{X}}\}}(\boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{S}}), \tag{7}$$

where $\iota_{\{\boldsymbol{\mathcal{X}}\}}(\cdot)$ is the indicator function[2] of $\{\boldsymbol{\mathcal{X}}\}$. This function guarantees the equal constraint $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{S}}$ on optimization. Furthermore, by introducing auxiliary variables $\boldsymbol{\mathcal{Z}}_1, \boldsymbol{\mathcal{Z}}_2$, and $\boldsymbol{\mathcal{Z}}_3$, we rewrite the minimization problem into the following equivalent expression:

$$\begin{aligned} \min_{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{S}},\boldsymbol{\mathcal{Z}}_i (i=1,2,3)} & \mathcal{R}_{\text{AEAN}}(\boldsymbol{\mathcal{Z}}_1) + \lambda \|\boldsymbol{\mathcal{Z}}_2\|_{2,1}^{(\mathcal{G})} + \iota_{\{\boldsymbol{\mathcal{X}}\}}(\boldsymbol{\mathcal{Z}}_3), \\ & \text{s.t. } \boldsymbol{\mathcal{Z}}_1 = \boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{Z}}_2 = \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{Z}}_3 = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{S}}. \end{aligned} \tag{8}$$

The minimization problem (8) can be applied to PnP-ADMM[3].

C. Generating Anomaly Maps

In the WLRX-based detector [2], assuming that background pixels follow a multivariate Gaussian distribution, we use the estimated $\boldsymbol{\mathcal{S}}$ as weights in WLRX to calculate the final anomaly map. The proposed RPCA with AEAN separates the spectrum of anomalous regions into a sparse tensor, assigning high weights to anomalous pixels. To enhance robustness, we use the inverse of the result of the closing operation on $\sum_k |[\boldsymbol{\mathcal{S}}]_{i,j,k}|$ as weights $\omega_i (i = 1, \ldots, N)$.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^B$ denote the sub-vectors corresponding to each pixel of the input HS image $\boldsymbol{\mathcal{X}}$. We calculate the weighted mean vector $\mathbf{m}$ as $\mathbf{m} = \sum_n \hat{\omega}_n \mathbf{x}_n$ and the weighted covariance matrix $\mathbf{C}$ as $\mathbf{C} = \sum_{n=1}^{N} \hat{\omega}_n (\mathbf{x}_n - \mathbf{m})(\mathbf{x}_n - \mathbf{m})^\top$, where $\hat{\omega}_n$ is the n-th normalized weights. The anomaly score for $n = 1, \ldots, N$ is calculated by $AD(\mathbf{x}_n) = (\mathbf{x}_n - \mathbf{m})^\top \mathbf{C}^{-1} (\mathbf{x}_n - \mathbf{m})$.

[2]Let $\mathbf{x} \in \mathbb{R}^N$ be an input vector. For a given non-empty closed convex set $\mathcal{C}$, the indicator function of $\mathcal{C}$ is defined by $\iota_C(\mathbf{x})$ which returns 0 if $\mathbf{x} \in \mathcal{C}$, and $+\infty$ otherwise.

[3]We omit the details of the proposed PnP-ADMM algorithm due to space limitations.

TABLE II
AUCs OF EACH METHOD (THE HIGHEST AUC IS HIGHLIGHTED IN BOLD AND THE SECOND HIGHEST IN UNDERLINE.

Scene	LRX [1]	WLRX [2]	RPCA [8]	1D-AEAN [7]	Ours
Airport1	0.9563	0.9310	0.9575	<u>0.9598</u>	**0.9717**
Airport2	0.9623	0.9638	<u>0.9828</u>	0.9692	**0.9831**
Airport3	0.8280	0.9028	0.9147	<u>0.9252</u>	**0.9468**
Airport4	0.9821	0.9337	**0.9999**	0.9851	<u>0.9919</u>
Beach1	0.9751	<u>0.9853</u>	0.9760	0.9793	**0.9882**
Beach2	0.9443	0.9693	0.9457	<u>0.9804</u>	**0.9824**
Beach3	0.9984	0.9999	0.9957	**1.0000**	**1.0000**
Beach4	0.9708	**0.9910**	0.9612	0.9794	<u>0.9839</u>
Urban1	0.9912	0.9899	0.9771	<u>0.9929</u>	**0.9932**
Urban2	0.9548	**0.9971**	0.9649	0.9793	<u>0.9907</u>
Urban3	0.9791	0.9342	<u>0.9897</u>	0.9818	**0.9910**
Urban4	0.9822	0.9930	0.9943	<u>0.9960</u>	**0.9967**
Urban5	0.9573	0.9635	<u>0.9791</u>	0.9660	**0.9795**
Average	0.9601	0.9657	0.9722	<u>0.9765</u>	**0.9845**

IV. Experimental Results

The proposed HS anomaly detection method was evaluated using ROC and AUC metrics on the ABU (airport, beach, urban) dataset. Noisy short-wavelength bands were removed, image intensity was normalized to $[0, 1]$, and the spectral domain was resized to 80 bands using bicubic interpolation.

We compared the proposed method with local-RX (LRX) [1], WLRX [2], and RPCA [8]. In addition, to verify the effectiveness of RPCA with AEAN in our method, we also compared it with the 1D-AEAN model. For each input HS image, we trained the 1D-AEAN model as described in Section III-A, computed the reconstruction error, and generated the anomaly map from the error as described in Section III-C. This approach can be seen as a slightly modified version of the 1D-AEAN model presented in [7]. We termed it "1D-AEAN". Both the proposed method and the conventional methods were trained on each scene from the ABU dataset, learning parameters individually for each scene.

Table II shows the results of the AUC scores for the proposed and conventional methods. The proposed method consistently outperforms conventional methods in most scenes and secures high ranks in several specific scenes. While LRX shows a lower average performance, WLRX and RPCA perform competitively. Additionally, 1D-AEAN achieves notable results in specific scenes, including a perfect score in one and a high average performance across methods.

Figure 2 shows binary anomaly maps generated by our method and conventional methods to maximize the F-score against the ground truth. The proposed method results in fewer false detections and higher accuracy across all scenes. RX-based methods and RPCA show lower accuracy and higher false detections of non-anomalous pixels. 1D-AEAN has slightly more false detections and lower accuracy than the proposed method.

V. Conclusion

We introduced the unsupervised anomaly detection method for HS images. This method combines RPCA and AEAN

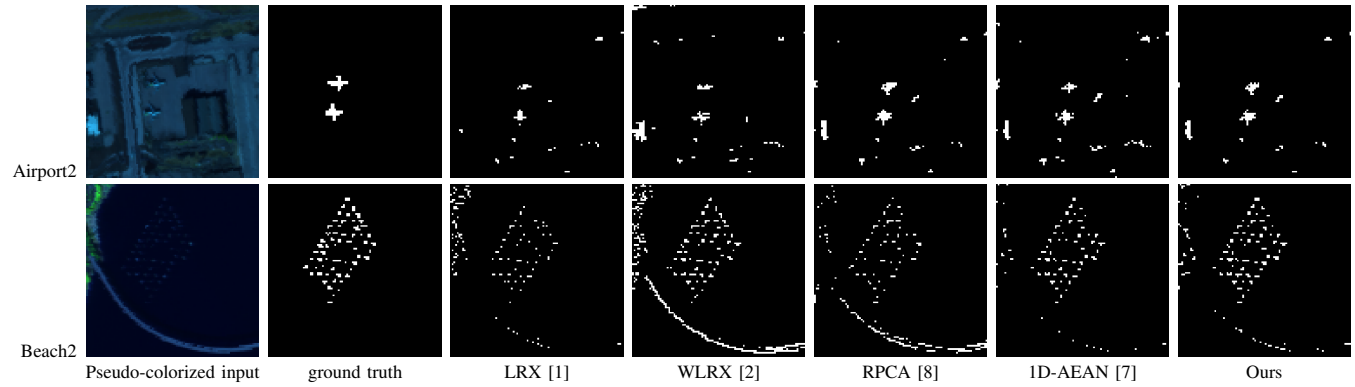


Fig. 2. Anomaly detection results obtained by binarizing the anomaly maps to maximize the F-score.

and is solved using PnP-ADMM. We utilize a pre-trained autoencoder as a proximity operator in RPCA to encourage low rankness in the reconstructed spectral data, effectively detecting anomalous pixels from the background. The experimental results confirmed the effectiveness of our method compared to some conventional approaches.

ACKNOWLEDGMENT

This work was partially supported by the FAIS Research and Development Project Support Program, "Satellite Data Utilization Demonstration and New Technology Development Project," and JSPS KAKENHI Grant Number 24K15075.

REFERENCES

[1] S. Küçük and S. E. Yüksel, "Comparison of RX-based anomaly detectors on synthetic and real hyperspectral data," in *Proc. Workshop Hyperspect. Image Sig. Process.: Evol. Remote Sens. (WHISPERS)*, IEEE, 2015, pp. 1–4.

[2] Q. Guo, B. Zhang, Q. Ran, L. Gao, J. Li, and A. Plaza, "Weighted-rxd and linear filter-based RXD: Improving background statistics estimation for anomaly detection in hyperspectral imagery," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 7, no. 6, pp. 2351–2366, 2014.

[3] C. Zhao, X.-F. Yao, and Y. Yan, "Modified kernel RX algorithm based on background purification and inverse-of-matrix-free calculation," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 544–548, 2017.

[4] W. Xie, J. Lei, B. Liu, Y. Li, and X. Jia, "Spectral constraint adversarial autoencoders approach to feature representation in hyperspectral anomaly detection," *Neural Networks*, vol. 119, pp. 222–234, 2019.

[5] W. Xie, B. Liu, Y. Li, J. Lei, and Q. Du, "Autoencoder and adversarial-learning-based semisupervised background estimation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5416–5427, 2020.

[6] Z. Zhao and B. Sun, "Hyperspectral anomaly detection via memory-augmented autoencoders," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 4, pp. 1274–1287, 2023.

[7] S. Arisoy, N. M. Nasrabadi, and K. Kayabol, "Unsupervised pixel-wise hyperspectral anomaly detection via autoencoding adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[8] W. Sun, C. Liu, J. Li, Y. M. Lai, and W. Li, "Low-rank and sparse matrix decomposition-based anomaly detection for hyperspectral imagery," *Journal of Applied Remote Sensing*, vol. 8, no. 1, pp. 083 641–083 641, 2014.

[9] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, "Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization," in *Proc. IEEE Int. Conf. Comput. vis. pattern recognit. (CVPR)*, 2016, pp. 5249–5257.

[10] M. Mishima, I. Kobayashi, and R. Matsuoka, "Hyperspectral anomaly detection based on local-tensor-nuclear-norm," in *Proc. IEEE Int. Conf. Geosci. Remote Sens. Symp. (IGARSS)*, 2023, pp. 2157–2160.

[11] D. Gabay and B. Mercier, "A dual algorithm for the solution of non linear variational problems via finite element approximation," *Comput. Math. App.*, vol. 2, no. 1, pp. 17–40, 1976.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[13] J. J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *C. R. Acad. Sci. Paris*, vol. 255, pp. 2897–2899, 1962.

[14] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play admm for image restoration: Fixed-point convergence and applications," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 84–98, 2016.

[15] Z. Meng, X. Yuan, and S. Jalali, "Deep unfolding for snapshot compressive imaging," *International Journal of Computer Vision*, vol. 131, no. 11, pp. 2933–2958, 2023.

Optimising Neural Networks with Fine-Grained Forward-Forward Algorithm: A Novel Backpropagation-Free Training Algorithm

James Gong, Bruce Li, and Waleed Abdulla
University of Auckland, Auckland, New Zealand
{hgon777,tli389}@aucklanduni.ac.nz
w.abdulla@auckland.ac.nz

***Abstract*—This paper introduces the Fine-Grained Forward-Forward Algorithm (FGFF), a novel backpropagation-free network training methodology that builds upon the foundational Forward-Forward algorithm. FGFF takes full advantage of being entirely backpropagation-free and enhances neural network performance through the implementation of a fine-grained significance matrix. This matrix quantitatively assesses the contribution of each neuron's activity towards the overall goodness function, thereby facilitating more explorative and precise weight adjustments. The efficacy of the algorithm is demonstrated by comparing its accuracy with the vanilla Forward-Forward algorithm and the standard backpropagation methods. Our evaluations employ three benchmark datasets: MNIST, Fashion-MNIST, and CIFAR-10. The results indicate that the FGFF achieves superior performance over the Forward-Forward algorithm, specifically showing a testing accuracy improvement of up to 25.54% with no obvious cost in the computation time, underscoring its enhanced capability in handling complex pattern recognition tasks.**

***Index Terms*—Backpropagation-free, Forward-Forward, Supervised Learning, Mnist Classification**

I. Backpropagation limitations

The backpropagation algorithm has been fundamental to the advancement of deep learning, yet it is encumbered by significant limitations that challenge both its practical application and biological plausibility. Firstly, the backpropagation algorithm is biologically implausible, as there is scant evidence to suggest that the human cortex propagates error derivatives backward or retains neural activities for subsequent processing [1]. Additionally, backpropagation necessitates a complete understanding of the forward pass to accurately calculate error derivatives, rendering it ineffective when the forward mechanism is treated as a black-box [2]. These limitations not only necessitate stringent prerequisites for deploying deep learning models but also challenge the prevailing notion that emulating the biological brain processes is a viable pathway toward achieving Artificial General Intelligence (AGI) [3].

II. The Forward-Forward Algorithm

Recent research has explored various backpropagation-free techniques for training neural networks, aiming to overcome the limitations of traditional backpropagation, such as its biological implausibility and computational inefficiency. Notable approaches include Feedback Alignment [4]–[7], and Hebbian Learning [8]–[10]. These methods offer various advantages but sometimes involve partial backpropagation and complex mechanisms. Additionally, a significant amount of work has been done to enhance the Forward-Forward algorithm, adding substantial value to its core methodology [11], [12].

The Forward-Forward algorithm [13] is an entirely backpropagation-free, greedy, and layerwise approach to training neural networks by replacing the forward and backward passes with two distinct forward passes, as indicated in Fig. 1. In the first pass, positive data are used, which are generated through the correct one-hot encoding of input data, and each instance in the dataset is accurately encoded with its true label, ensuring the network learns from correctly encoded examples. In contrast, the second pass processes negative data generated by intentionally encoding the input data with the incorrect label. Fig. 2 illustrates examples of positive data and negative data from the MNIST [14] dataset, where the indices of the first 10 pixels are employed to represent the labels for the data. Within this framework, the algorithm computes two specific measures of activity: positive activity and negative activity. Positive activity is derived by squaring the output of each neuron during the positive pass, which emphasises the activation levels when the data is correctly labelled. Conversely, negative activity is calculated identically during the negative pass, reflecting the neuron's response to incorrectly labelled data. Then, in each layer, the algorithm aims to train the mean of the positive activities far above a threshold value and the mean of the negative activities far below that.

Therefore, the loss function can be written as:

$$L_j = \frac{1}{2}\max\Big(0, \delta - (\mathrm{g}_{\mathrm{pos;i}} - \mathrm{threshold})\Big) + \max\Big(0, \delta - (\mathrm{threshold} - \mathrm{g}_{\mathrm{neg;i}})\Big) \quad (1)$$

$$g_{\mathrm{pos,\ neg;i}} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}^2 \quad (2)$$

where a hinge loss is employed to push the positive goodness and the negative goodness in the current layer to be at least $2 \times \delta$ apart centred around the threshold value. $g_{pos,neg;i}$

stands for the positive and negative goodness value calculated in the current layer i. Following the calculation of the loss, appropriate adjustments are made to the respective weights and biases using gradient descent. To ensure that large activation values from the current layer do not overly influence the activity values in the subsequent layer, L2 normalisation is applied to all output values before they enter the next layer.

During prediction, the data are initially labelled with all possible categories. After that, these data are passed into the network, and the category of data generating the highest activity sum also referred to as the goodness value, is interpreted as the outcome of the network.

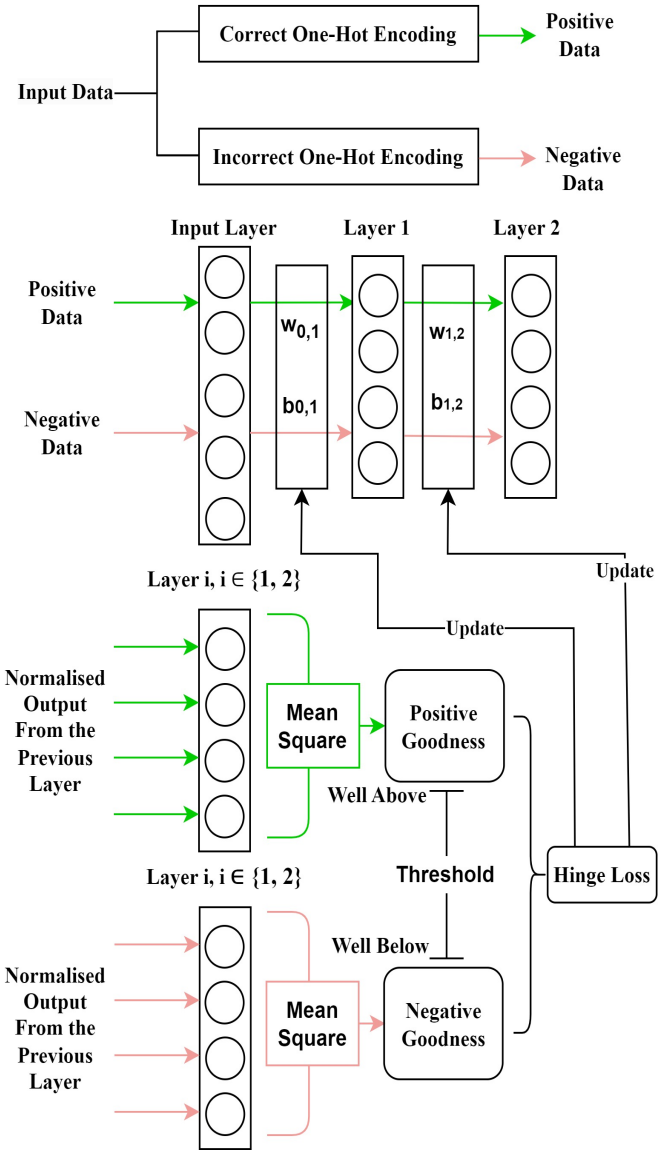
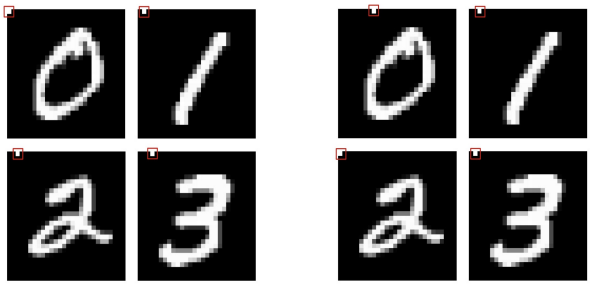


Fig. 1: Two distinct forward passes in the Forward-Forward algorithm. The green arrows demonstrate the positive forward pass with positive data input. The red arrows demonstrate the negative pass with negative data input. Layer i represents any hidden layer in the network, where i can be either 1 or 2.

From a biological perspective, the brain tends to react more vigorously to the patterns previously learned and less rigorously to the patterns they are not familiar with. Hence, the rigorousness of the neural activity is taken to compute and classify the output when a combination of patterns is presented.

However, an inherent assumption of the Forward-Forward algorithm is that each neuron and each layer contributes identically to the overall goodness function. In fact, earlier layers tend to capture more abstract features compared to the more concrete features learnt by the later layers [17]. Furthermore, within the same layer, the significance of each neuron's contribution can vary considerably. This raises critical questions about the weighting of individual neuron activities: How significant is each neuron, and to what extent should the activity value generated by each neuron influence the overall goodness function? These concerns are central to the design of FGFF, which are addressed in Section III.



(a) Positive Data Sample (b) Negative Data sample

Fig. 2: Positive and negative data visualization for the Forward-Forward algorithm. The first 10 pixels (top left corner) in the images are used for one-hot encoding. For positive data samples, the correct index corresponding to the label is encoded as white; for example, if the label is 0, the first pixel (index 0) is encoded as white. In negative data samples, the indices are randomly encoded while excluding the correct index. The encoded index has been highlighted.

III. The Fine-Grained Forward-Forward Algorithm

The Fine-Grained Forward-Forward algorithm (FGFF) builds upon the Forward-Forward algorithm by introducing a key novel element: **neuron activity scaling factors**, also referred to as the fine weights or the significance matrix in Fig. 3. These scaling factors are pivotal in modulating the activity values of neurons within each layer. By appropriately assigning these scaling factors, the concerns regarding the variable significance of individual neurons and their influence on the overall goodness function are directly addressed. The pseudo-code for FGFF is presented in Algorithm 1.

A. Introducing novel calculation of layer-wise goodness function

The neuron activity scaling factors parameter is structured as a 2D tensor within the neural network architecture. It is defined such that the number of rows corresponds to the number of

hidden layers in the network, and the number of columns in each row corresponds to the number of neurons in each respective layer. For instance, in a network with two hidden layers, each containing 500 neurons, the tensor is initialised as a shape of 2 by 500 filled with 1[1]. This 2D tensor arrangement allows for independent scaling of neuron activities across different layers during training. As the network learns from repeated exposure to stimuli, these values dynamically adjust, optimising the responsiveness and adaptability of the network over time.

Similar to the Forward-Forward algorithm, FGFF employs one loss function across each layer with the core objective remaining unchanged: to train the network such that the mean of the positive activities significantly exceeds a specified threshold, while the mean of the negative activities falls well below it. However, in FGFF, the activity values are now scaled by the corresponding neuron activity scaling factors w_{ij}, hence the new layer-wise goodness function is written as:

$$g_{\text{pos, neg};i} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}^2 w_{ij} \tag{3}$$

where w_{ij} indicates the scaling factor for the j_{th} neuron in the i_{th} layer, and x_{ij} stands for the activity output of the neuron in the layer.

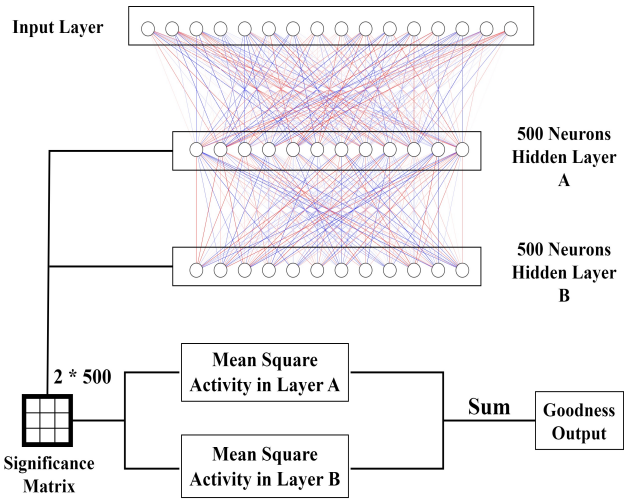


Fig. 3: A sample neural network architecture used in the Fine-Grained Forward-Forward Algorithm. The significance matrix is another terminology for the collection of the neuron activity scaling factors.

B. Learning in neuron activity scaling factors

The Fine-Grained Forward-Forward algorithm refines the traditional training process of FF by incorporating the neuron activity scaling factors. These factors are crucial trainable parameters that dynamically adjust each neuron's influence on the network's goodness function. The addition of these neuron activity scaling factors leads to an increase in the total number of trainable parameters, which is directly proportional to the product of the number of hidden layers $(m-1)$ and the number of neurons per layer (n). However, this increase is minimal compared to the core architectural parameters of the network, which amount to $(m-1) \times n^2$.

During training, each neuron outputs an activation value as inputs propagate through the network. These activations are squared and then scaled by the corresponding scaling factors. Subsequently, the scaled activations are averaged to determine the positive and negative goodness values, central to the algorithm's loss function, see equation (1). This loss function is recalculated for each epoch to ensure accurate gradient descent based on the current network state.

Updates to the neuron activity scaling factors are governed by the derivative of the loss function. From a mathematical standpoint, the adjustments made to the neuron activity scaling factors enhance the distinction between positive and negative goodness values. This increased disparity aids in sharpening the decision boundaries of the network, resulting in more precise classifications. Moreover, by incorporating the scaling factors, the algorithm fosters new connections between neighboring solutions in the solution space. This additional connectivity in the solution space reduces the number of local optima, decreasing the likelihood that the training process becomes trapped in shallow local optima. Consequently, this results in a more effective and exploratory training regime, promoting a robust learning environment that approaches better solutions.

C. Impact of neuron activity scaling factors

A critical aspect of FGFF is the role of neuron activity scaling factors during both training and prediction phases. These scaling factors enhance the learning process by allowing the network to assign varying levels of importance to different neurons, leading to a more nuanced and effective learning of patterns over the epochs.

Interestingly, even when the scaling factors are discarded during the prediction phase—utilised solely during training—the model still demonstrates performance comparable to scenarios where scaling factors are used thoroughly. This indicates that the network successfully learns the underlying patterns during training, rather than relying heavily on the scaling factors for decision-making during prediction. This robustness is particularly beneficial for a backpropagation-free neural network training algorithm like FGFF. By simplifying the prediction phase and eliminating the need for scaling factors, the FGFF achieves prediction efficiency identical to the original Forward-Forward algorithm but delivers more accurate results during deployment.

D. The fine weights optimizer

In contrast to the Forward-Forward algorithm, FGFF features two separate optimizers: one is responsible for updating the parameters of the neural network layer (weights and

[1]The scaling factors are initialized to values of one to align the fine-grained forward-forward algorithm with the standard forward-forward algorithm at the outset. This initialization ensures that, in the absence of modifications to the scaling factors, the algorithm reverts to the behavior of the original forward-forward algorithm.

biases), and another for adjusting the fine weights. This dual optimization approach introduces an additional hyperparameter—the learning rate for the fine weights optimizer.

By setting the learning rate for the fine weights optimizer to zero, FGFF disables any updates to the neuron activity scaling factors, resulting in these factors remaining at their initial value of 1. Consequently, FGFF reverts to the standard Forward-Forward algorithm. Therefore, with appropriate tuning of this hyperparameter, the worst-case performance of FGFF is equivalent to the best performance achievable by the Forward-Forward algorithm. This ensures that FGFF, at a minimum, performs no worse than the FF algorithm, while offering the potential improvements in accuracy across all datasets.

E. Mathematical parallels with feature and constraint violation detectors

In section 10 (future work) of the Forward-Forward algorithm paper [13], Hinton proposed the potential benefit of incorporating a set of feature detectors to maximise their squared activity and a set of constraint violation detectors to minimise their squared activity [18]. While the Fine-Grained Forward-Forward Algorithm (FGFF) does not explicitly employ feature detectors, its design mathematically aligns with this proposed concept. Specifically, the neuron activity scaling factors in FGFF are optimised in such a way that they maximise the squared activity of some neurons, akin to feature detectors, and minimise the squared activity of other neurons, similar to constraint violation detectors.

IV. Experiments

This section demonstrates the performance analysis of FGFF and compares it with the Forward-Forward algorithm and the traditional backpropagation algorithm on three benchmark datasets: MNIST, FashionMNIST [15], and CIFAR-10 [16].

A. Performance study of FGFF

A number of network architectures were trained on multiple datasets to evaluate the effectiveness of FGFF, FF, and BP training algorithms. This includes MNIST, FashionMNIST, and CIFAR-10.

Model	BP	FF	FGFF
2x50 Relu	97.65%	74.71%	87.13%
2x500 Relu	98.27%	85.82%	90.88%
2x1000 Relu	98.75%	85.83%	92.95%

TABLE I: Test accuracies (%) on MNIST dataset for backpropagation (BP), Forward-Forward (FF), and fine-grained Forward-Forward (FGFF) algorithms.

The results on MNIST are summarized in Table I. FGFF consistently outperformed FF across all architectures.

The results in Table II demonstrate that increasing the number of neurons in each layer tends to improve model performance on the Fashion-MNIST dataset, similar to the observations in Table I for MNIST.

Algorithm 1 The Fine-Grained Forward-Forward Algorithm

Initialize the fine-grained weight matrix w_f of dimension $m \times n$ with all ones, where m is the number of layers excluding the input layer, and n is the number of neurons in such layers.
$epochs \leftarrow 200$
$i, j \leftarrow 0$
$\mathbf{x}$, target_labels $\leftarrow$ load data
positive_data, negative_data $\leftarrow$ create data # use one-hot encoding
while $i < m$ **do**
    # Greedily train each layer, this process can be parallelized
    while $j < epochs$ **do**
        $o_p \leftarrow$ Forward(positive_data) ▷ positive output
        $o_n \leftarrow$ Forward(negative_data) ▷ negative output
        $w_{f;i} \leftarrow w_f[i, :]$
        # $Square(o_p, o_n)$ squares each element in the tensors o_p and o_n
        $g_{\text{pos}}, g_{\text{neg}} \leftarrow \text{Mean}(\text{Square}(o_p, o_n) \times \mathbf{w}_{f;i})$
        $L \leftarrow \text{Loss}(g_{\text{pos}}, g_{\text{neg}})$
        $\delta w_f \leftarrow \alpha_f \frac{dL}{dw_f},\ \delta w \leftarrow \alpha \frac{dL}{dw}, \delta b \leftarrow \frac{dL}{db}$
        # Perform weights and biases update based on L
        $w_f \leftarrow w_f - \delta w_f$ ▷ the fine-grained significance matrix
        $w \leftarrow w - \delta w$ ▷ weightings for interconnecting neurons
        $b \leftarrow b - \delta b$ ▷ biases for the neurons
        $j \leftarrow j + 1$
    end while
    positive_data $\leftarrow$ Forward(positive_data)
    negative_data $\leftarrow$ Forward(negative_data)
    $i \leftarrow i + 1$
end while

function FORWARD(x)
    $x_n \leftarrow \text{Normalize}(x)$ ▷ L2 normalization
    output $\leftarrow x_n \times W^T$
    return RELU(output)
end function

B. Experiments on CIFAR-10

Due to the intrinsic complexities of the CIFAR-10 dataset, the experiment employed a neural network with three layers, each containing 800 neurons. The learning rate was set to 0.01, and the training was conducted for 1200 epochs. The results of this experiment are presented in Table IV.

Notably, achieving comparable accuracy with FGFF necessitated an additional 1800 epochs for FF, highlighting the efficiency of FGFF in training neural networks on complex datasets such as CIFAR-10.

Model	BP	FF	FGFF
2x50 Relu	87.52%	60.76%	73.28%
2x500 Relu	89.34%	52.24%	75.63%
2x1000 Relu	89.47%	51.64%	77.18%

TABLE II: Test accuracies (%) on Fashion-MNIST dataset for back-propagation (BP), Forward-Forward (FF), fine-grained Forward-Forward (FGFF), and tuned FF models with various architectures.

Algorithm	Train Acc	Test Acc
BP	90.12%	52.08%
FF	38.33%	37.67%
FGFF	50.73%	42.67%

TABLE III: Train accuracies (%) and Test accuracies (%) on CIFAR-10 dataset for back-propagation (BP), Forward-Forward (FF), fine-grained Forward-Forward (FGFF).

C. Epoch time analysis

In this section, we evaluate the performance of FGFF relative to FF and the conventional BP by comparing the time taken per epoch for each algorithm. The findings of these comparisons are detailed in Table IV.

Algorithm	Time taken per epoch
BP	13.1754 s
FF	0.1066 s
FGFF	0.1073 s

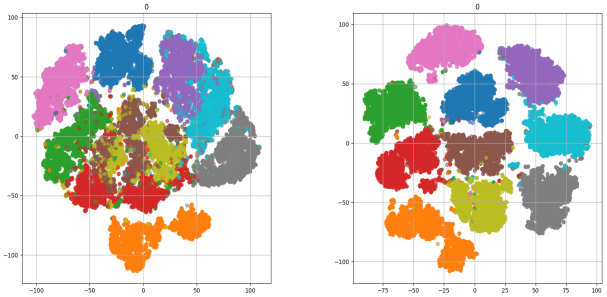
TABLE IV: Time taken per epoch (s) on Fashion-MNIST for back-propagation (BP), Forward-Forward (FF), fine-grained Forward-Forward (FGFF). The training was conducted using an NVIDIA RTX GeForce 4060 GPU on Windows 11.

BP requires a significantly longer time per epoch, as shown in the table. This extended time requirement is due to the complex nature of BP, which involves precise and symmetric weight updates that must be calculated and applied during training. These updates, while effective, are biologically implausible because such precision and symmetry are not feasible in biological neural networks. In contrast, FF and FGFF are designed to be more aligned with how learning occurs in the brain, making them biologically plausible.

From Table IV, we observe that FGFF takes approximately 0.1073 seconds per epoch, which is only around 0.65% slower than FF, which takes 0.1066 seconds per epoch. Despite the slight increase in time, FGFF's performance can be significantly better, achieving up to 25.54% improvement in accuracy as shown in Table II. This highlights FGFF's efficiency and effectiveness, making it a highly practical choice in scenarios where both computational efficiency and performance are critical.

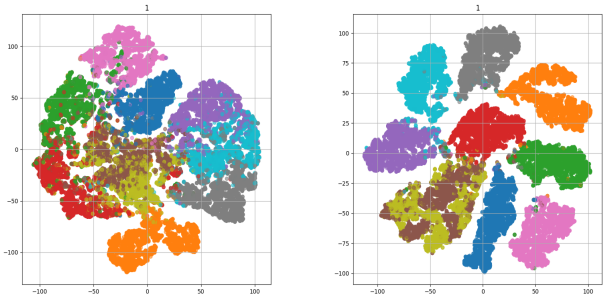
D. t-SNE visualisation

To have a deeper understanding of how our model processes and transforms the input data through each layer, and compare the feature learning capabilities of FGFF with FF, a neural network of 2x500 ReLU was trained on the MNIST dataset. The input data and layer-wise output feature map is visualised using t-SNE [19], in Fig. 4 and Fig. 5. During the training, the activations (goodness) are calculated. After training a layer, the positive goodness values are visualised.



(a) FF (b) FGFF

Fig. 4: t-SNE visualisations of first hidden layer output feature maps of positive data after 200 epochs. Different colors correspond to different classes



(a) FF (b) FGFF

Fig. 5: t-SNE visualisations of the second hidden layer output feature maps of positive data after 200 epochs. Different colors correspond to different classes

The feature maps reveal that both methods effectively learn features that distinguish between classes. However, FGFF demonstrates a more distinct data separation compared to FF.

V. Conclusion

In this paper, we introduced the Fine-Grained Forward-Forward algorithm (FGFF). A novel backpropagation-free training methodology that builds upon the Forward-Forward (FF) algorithm. FGFF incorporates neuron activity scaling factors, enabling more precise and effective weight adjustments. Our experiments on MNIST, FashionMNIST, and CIFAR-10 datasets demonstrated that FGFF consistently outperforms FF

in accuracy and efficiency, stepping closer to the traditional backpropagation algorithm (BP). Overall, FGFF represents a significant advancement in backpropagation-free neural network training, providing a viable alternative to traditional methods with potential applications in various domains.

REFERENCES

[1] Song, Yuhang, et al. "Can the Brain Do Backpropagation?—Exact Implementation of Backpropagation in Predictive Coding Networks." *Advances in Neural Information Processing Systems*, 33 (2020): 22566-22579.

[2] Rumelhart, David E., Hinton, Geoffrey E., & Williams, Ronald J. (1986). "Learning representations by back-propagating errors." *Nature*.

[3] Lake, Brenden M., Ullman, Tomer D., Tenenbaum, Joshua B., and Gershman, Samuel J. (2017). "Building Machines That Learn and Think Like People." *Behavioural and Brain Sciences*.

[4] Nøkland, Arild. (2016). "Direct feedback alignment provides learning in deep neural networks." *Advances in Neural Information Processing Systems*, 29.

[5] Crafton, Brian, Parihar, A., Gebhardt, Evan, and Raychowdhury, A. (2019). "Direct feedback alignment with sparse connections for local learning," *Frontiers in Neuroscience*, vol. 13.

[6] Lillicrap, Timothy P., Cownden, D., Tweed, D., and Akerman, C. (2016). "Random synaptic feedback weights support error backpropagation for deep learning," *Nature Communications*, vol. 7.

[7] Nøkland, Arild. (2016). "Direct feedback alignment provides learning in deep neural networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, NIPS'16, p. 1045–1053, Curran Associates Inc.

[8] Krotov, Dmitry, and Hopfield, John J. (2019). "Unsupervised learning by competing hidden units," *Proceedings of the National Academy of Sciences*, vol. 116, no. 16, pp. 7723–7731.

[9] Amato, G., Carrara, Fabio, Falchi, F., Gennaro, C., and Lagani, Gabriele. (2019). "Hebbian learning meets deep convolutional neural networks," in *ICIAP*.

[10] Gupta, M., Ambikapathi, A., & Ramasamy, S. (2021). "HebbNet: A Simplified Hebbian Learning Framework to Do Biologically Plausible Learning." In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3115-3119). IEEE. doi:10.1109/ICASSP39728.2021.9414241.

[11] Gandhi, S., Gala, R., Kornberg, J., & Sridhar, A. (2023). Extending the Forward Forward Algorithm. arXiv preprint arXiv:2307.04205.

[12] Ororbia, A., & Mali, A. (2023). The Predictive Forward-Forward Algorithm. arXiv preprint arXiv:2301.01452.

[13] Hinton, Geoffrey. (2022). "The Forward-Forward Algorithm: Some Preliminary Investigations." Google Brain.

[14] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

[15] Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. Retrieved from https://github.com/zalandoresearch/fashion-mnist

[16] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

[17] Goodfellow, Ian; Bengio, Yoshua; & Courville, Aaron. (2016). "Deep Learning." MIT Press.

[18] Welling, M., Williams, C., and Agakov, F. (2003). "Extreme components analysis." *Advances in Neural Information Processing Systems*, 16.

[19] van der Maaten, L.J.P., and Hinton, G.E. (2008). "Visualizing high-dimensional data using t-sne." *Journal of Machine Learning Research*, 9:2579–2605.

Two-Way Malaysian Sign Language Communication System for Inclusive Education

Veron Zhen Liang Hii[†*], Aaron Ken Kiat Lo[†*], Ida Pei Xin Lee[†*],
Alec Vince Gonzales Abuan[†*], Sue Han Lee[†], Patrick Hang Hui Then[†]
[†]Swinburne University of Technology Sarawak Campus, Malaysia
E-mail: 102768326@students.swinburne.edu.my, 102768452@students.swinburne.edu.my, 102766016@students.swinburne.edu.my,
105456145@swinburne.edu.my, shlee@swinburne.edu.my, pthen@swinburne.edu.my

***Abstract*—Inclusive education aims to create equal learning opportunities for all, but significant gaps still exist, particularly for the deaf community. This paper addresses these gaps by proposing a new educational platform that integrates cutting-edge technology to improve accessibility and engagement for deaf learners. Our solution introduces an AI-powered two-way sign language communication system specifically designed for integration into classrooms. With an avatar-based approach, the system focuses on simple technology transfer for Sign Language Production (SLP) and utilises advanced deep learning methods for Sign Language Recognition (SLR). This enables seamless and effective communication between deaf students and educators. To our knowledge, this is the first comprehensive approach to digital inclusion in inclusive learning that primarily addresses the specific needs of the deaf community. As part of our initiative, we have created the first Malaysian Sign Language (BIM) education dataset to serve as a benchmark in this area. A new user testing framework has also been developed to quantify the effectiveness of the system in an educational context. The results of the survey emphasise the critical importance and necessity of such an educational platform.**

I. Introduction

Inclusive education aims to remove barriers for all children and ensure access to education, active engagement and optimal academic and social outcomes [1]. Malaysia has been working towards this goal with initiatives such as the Zero Reject Policy [2]. However, the Malaysian deaf community still faces major challenges in accessing equal education. People who are deaf or hard of hearing communicate primarily through the use of the Malaysian Sign Language (BIM), which relies on body gestures such as hands, arms and facial expressions. Studies emphasise that the main barrier for deaf students in mainstream classrooms is linguistic rather than physical [3]. As a result, they often find it difficult to participate fully in a society where spoken language is predominantly used. Inclusive education policies that aim to enrol all children in mainstream schools cannot be applied directly to deaf learners. One promising approach is to integrate sign language communication into schools and to provide sign language translation tools facilitated by widely available assistive technologies.

Assistive technologies related to sign language translation aids have been investigated in recent studies [4]. These projects have proven to be successful as they have received positive feedback from the deaf community on the improvements in communication and engagement. However, they face several challenges that prevent them from being utilised in educational or communication settings. One major challenge is the high cost associated with these technologies, making them prohibitively expensive for classroom use, especially where budgets are limited. Additionally, many existing tools are still at an early stage of development and not yet publicly available. Importantly, these tools often do not support two-way communication, relying instead on translators to facilitate interaction, which limits their effectiveness in real-time communication scenarios.

To overcome these challenges, a robust two-way communication system is essential. As not all deaf people rely on spoken language, it is crucial to provide a method of converting spoken language into sign language. Conversely, it is equally important to translate sign language into spoken or written language for hearing people. As the existing systems currently do not provide support for BIM, our goal is to create an inclusive education system that utilises BIM for all deaf learners and educators in Malaysia.

To achieve this, we propose an AI-powered two-way communication system for BIM, which includes two main approaches: 1) An avatar-based approach for Sign Language Production (SLP) and 2) A deep learning approach for Sign Language Recognition (SLR).

The main contributions of our work are as follows:

1) We proposed a novel educational platform to improve inclusive education for the deaf community. This platform focuses on BIM and can be extended to cover different sign languages.
2) We introduced new digital inclusion strategies that integrate the two-way communication system in sign language. This innovative system uses an avatar-based approach for SLP and utilises deep learning methods for SLR.
3) We set a new benchmark by introducing a BIM dataset for the education sector together with a novel usability testing framework. This sets a new standard for evaluating the usability of educational tools in this field.

[*]These authors contributed equally to this work

The rest of the paper is organised as follows: II provides a literature review of related works on SLP, SLR and AI in education. In III the educational BIM dataset and a self-collected text dataset are presented. The sections IV and V deal with the methodology of SLP and SLR respectively. Section VI presents the analysis of the experiments and the results obtained. Finally, section VII concludes our study.

II. Related works

A. AI for Inclusive Education

Several studies have investigated the integration of AI in inclusive education [5]. Microsoft Translator, for example, uses a headset worn by the speaker to translate speech into subtitles for deaf learners [6]. Moreover, AI tools such as LIFEisGAME help children with Autism Spectrum Disorders (ASD) to understand facial expressions [6].

However, these tools primarily support one-way communication by translating spoken or written content into subtitles or simplified text. This limitation hinders full engagement as deaf people are unable to express their thoughts, ask questions or participate effectively in discussions without a two-way communication system [7].

B. Sign Language Production (SLP)

In the process known as SLP, descriptions in spoken language are automatically translated into sign language. This is achieved by converting sign language glosses into sign language pose sequences. This section explores the latest techniques in SLP, including NMT, conditional image and video generation and avatar-based approaches.

1) Neural machine translation (NMT) approaches: In NMT, word sequences in whole sentences are modelled and predicted with the help of Artificial Neural Networks (ANNs). Pre-processing tasks such as chunking, recognising sentence boundaries and tokenising words are a particular challenge for sign language due to differences in word order and gloss level compared to spoken language. To overcome this challenge, researchers [8] have used Convolutional Neural Networks (CNNs) together with a seq2seq model to translate spoken language sentences from sign language videos.

2) Conditional image / video generation approaches: Conditional image generation, driven by deep learning, has made progress in creating images or videos based on specific inputs. The combination of VAEs and GANs is promising for SLP as it leverages the reliability of VAEs and the discriminative capabilities of GANs for more lifelike human motion synthesis in videos. Hybrid frameworks integrating these models have been proposed for sign language video generation. However, challenges remain, particularly in the effective control of GANs [9], where issues such as gradient disappearance and mode collapse during training limits pattern diversity and hinder meaningful feature learning.

3) Avatar approaches: Recent SLP methods use animated avatars to produce sign language content online, as the JASigning [1] project shows. JASigning uses transcription languages such as HamNoSys IV-C0a and SiGMLIV-C0b as input for its sign translation model. Kim [10] uses a named entity transformer with avatar layering. However, JASigning's software is outdated and existing methods use proprietary representations that are incompatible with unknown users. To bridge this gap, we have developed a unique JSON-based representation for data and animation scripts.

C. Sign Language Recognition (SLR)

SLR, known as automatically recognises sign languages and translates them into spoken language descriptions. This section examines the latest methods of CNN, RNN and other approaches.

CNNs and RNNs are widely used in sign language recognition (SLR), with CNNs usually serving as feature extractors and RNNs performing temporal modelling. Hu, Gao, Liu and Feng [11] proposed a correlation network that uses 2D CNN for image-based feature detection, and one-dimensional CNN and BiLSTM for temporal modelling. The 2D-CNN extracts features from each image, and the 1D-CNN and BiLSTM capture temporal correlations between images, improving the network's ability to recognise characters over time.

More recent research has explored other approaches to SLR. One study [12] developed a CNN- and transformer-based multi-branch network that combines the transformer's ability to compute long-range dependencies and the CNN's ability to extract local features. This hybrid approach leverages the strengths of CNNs and transformers and improves the network's ability to recognise symbols at different temporal and spatial scales. Another study [13] introduced a novel approach to context-aware continuous sign language recognition using the GAN architecture.

III. Dataset

In this study, we collected a novel educational dataset that contains BIM and focuses on medical contexts [2]. The dataset consists of 1040 videos with 40 unique glosses and 10 sets of text annotations in Malay or Bahasa Malaysia (BM) including English (ENG) translation.

a) Dataset Modalities: The dataset consists of dual modality glosses and sentences with corpus in texts and videos.
Gloss. Gloss is a text form that can be used as an intermediate representation for the transcription of signs with spoken language words [14].
Sentence. Sentences that have been transcribed in BM/ENG are constructed in the data set using glosses so that they correspond to the approximate meanings of the sign language.
Video corpus. Video corpus is the recorded videos that contain BIM and focus on medical contexts. The video corpus is pre-processed manually using video editing software with minimal changes. For easier processing, we cut the video replays into their respective video clips. We then organise the clips and add the BM/ENG gloss annotations/rewritten sentences as class labels. This corpus contains the CSV files that contain the file

[1] https://vh.cmp.uea.ac.uk/index.php/JASigning

[2] https://arekku21.github.io/MSL-Medical/

names of the training and test data along with the class labels. We also prepare the video corpus by splitting it into training and test data in a ratio of 80:20 for model training.

Text corpus. Text corpus refers to a self-constructed corpus based on the vocabulary of the education BIM dataset. It is prepared for the text-to-gloss system that has been integrated as part of SLP. It is divided into two parts: Source and Target. The source is written in BM/ENG and serves as input text for the user, while the target consists of a combination of glosses written in BM. It consists of 1348 pairs of source and target data, which differ in length and structure. We have also done some pre-processing on the text corpus by splitting it into training and test sentences in a ratio of 80:20.

b) Signers: There are a total of seven sign language signers in this project, three of whom are experts from Sarawak Society for the Deaf (SSD) and four of whom are non-experts. The three experts were Deaf and fluent in sign language, with native-level proficiency. The other non-experts belonged to the university community and were counselled by the sign language experts and received feedback for the recording.

c) Recording Pipeline: Prior to recording, a set of 40 unique glosses and 10 sentences relating to medical contexts were agreed in advance and selected by SSD. On the day of recording, participating signers were informed about the glosses and sentences to ensure consistency. The signers were instructed to begin and end with their hands positioned at the side of their body. Once filming began, we recorded each vocabulary word and phrase three times for each signer. For two of the three recordings of the sentences, a formal approach was to be followed, in which the corpus takes into account all connectives, nouns, pronouns, etc. In the third recording, the informal, practical and more intuitive sign language was used, as recommended by the sign language experts. Although there are differences, the meaning of the sentences remains the same according to the annotations.

d) Green Screen Studio: The *Green Screen Studio* at Swinburne University of Technology Sarawak (SUTS) was the strategically chosen location to shoot the videos. We intended that the background could be edited in possible future works. Modern high-definition (HD) webcams were used for the recordings, which were aimed frontally at the participants and shot from the upper body. Both cameras used for the videos had a resolution of 1920x1080 at 30 frames per second (FPS).

e) Privacy, bias and ethical considerations: In order to protect the privacy and rights of the project participants, we have sent out a declaration of consent with the required information, which was to be read and signed.

IV. SLP METHODOLOGY

The goal of SLP is to generate sign language from the language used in the spoken language community, e.g. producing BIM from BM. In this study, the input would be the user text presented in sentences, while the output would be the sign language presented in an avatar animation. Our SLP framework consists of three stages: Text-to-Gloss, Gloss-to-Animation Mapping and Avatar Animation.

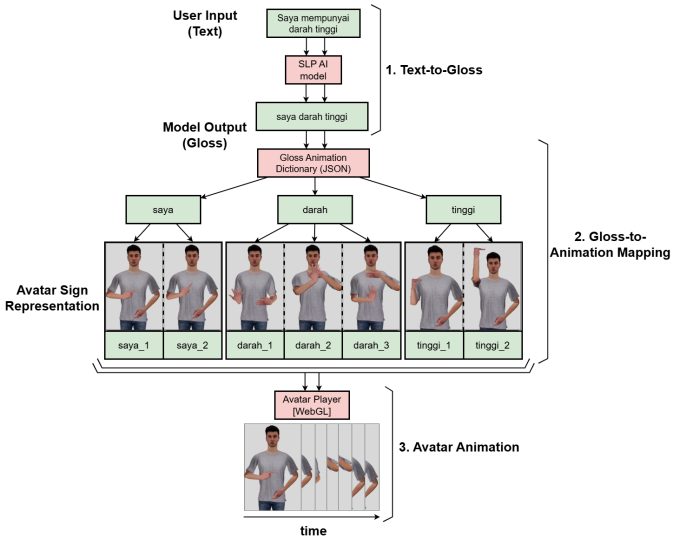


Fig. 1. Sign Language Production (SLP) Methodology

A. Text-to-Gloss

Text-to-Gloss aims to achieve the translation of user text into glosses. In our approach, we use Transformer [15], a deep learning architecture as our model architecture. Transformer consists of two parts, the encoder and the decoder.

Encoder. We chose a pre-trained BERT model trained on a public BM dataset available on GitHub [3].

Decoder. We use three different configurations for the decoder. *Config A* adopts the standard Transformer configuration with 768 hidden sizes, 8 attention heads and 6 hidden layers. *Config B* takes over the configuration of the encoder with a hidden size of 336, a number of 12 attention heads and a number of 4 hidden layers. In addition, we used the pre-trained BERT model as *Config C* of the decoder.

We use NVIDIA RTX3060 to train our models until convergence. The model is trained using the cross-entropy loss function and optimised with Adam optimiser [16]. The initial learning rates are set to 1×10^{-5} and 1×10^{-6}, with weight decay of 1×10^{-7} and a batch size of 32. Specifically, for *Config C*, the learning rate of 1×10^{-6} is used for fine-tuning purposes. During testing, the model is evaluated on the test dataset, which containing unseen sentences from the text corpus. We compare the generated sentences with the ground-truth sentences using metrics such as BLEU, ROUGE-L and accuracy.

B. Gloss-to-animation mapping

Gloss-to-animation mapping aims to map the glosses into their respective animation. We use a JSON file as dictionary to store the mapping between the glosses and their animation.

C. Avatar animation

Avatar animation indicates the movement of 3D avatar to represent the sign language. This is one of the crucial component in our methodology. The avatar animation emphasizes the

[3]https://github.com/mesolitica/malaya

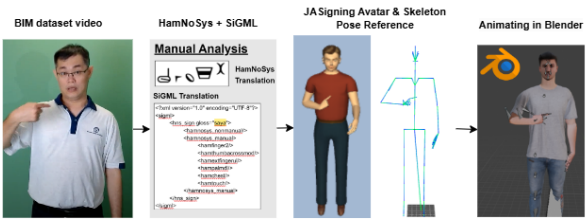


Fig. 2. Flowchart of Avatar Animation Creation: The diagram shows how video data is analyzed to create sign language animations for the avatar.

importance of accurate symbol generation, automated translation, and verification to ensure the fidelity of the resulting sign language animation. To achieve the avatar animation, there are several stages has been researched and processed, which is visualised on figure 2, as follows:

a) HamNoSys dataset analysis: Initially, each sign language video is subjected to careful manual analysis to identify the movements and positions of the hands as well as non-manual features such as facial expressions. The HamNoSys generation tool [4], is employed to transcribe these identified movements into HamNoSys symbols.

b) SiGML translation and verification: The next stage involves converting HamNoSys symbols into SiGML using an automated translation system [5] available on GitHub. This tool maps HamNoSys components to their corresponding SiGML representations, streamlining the conversion process. Subsequently, we perform skeletal model referencing and pose extraction using SiGML in a distinct avatar player to animate our 3D avatar.

c) JSON-based representation: In the final stages, our workflow involves animating the avatar using Blender. The animations are then imported into our WebGL engine within a web-based environment. To facilitate the playback of the avatar animations based on text input, we have developed a custom JSON-based representation for mapping gloss-to-animation.

V. SLR METHODOLOGY

The aim of SLR is to translate sign language into languages used by the spoken language community, e.g. translating BIM into BM. In our methodology, the input consists of sign language represented by BIM videos, while the output is a sequence of text presented in sentences. Our SLR framework consists of two stages: feature extraction, and encoding-decoding.

A. Feature extraction

For feature extraction from the video frames, we use a pre-trained CNN model. For each video, 32 frames are selected at evenly spaced intervals to ensure that the selected frames throughout the videos. Specifically, we use VGG-16 model [17] with prior knowledge from ImageNet weights as the pre-trained CNN model. In addition, we exclude the top three layers and freeze all other layers.

[4]https://www.sign-lang.uni-hamburg.de/hamnosys/input/
[5]https://github.com/carolNeves/HamNoSys2SiGML

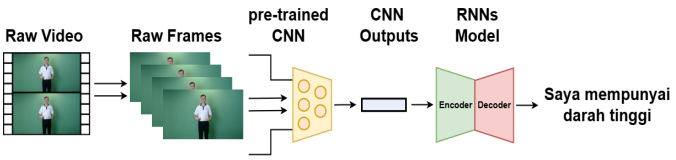


Fig. 3. Sign Language Recognition (SLR) Methodology

B. Encoding-Decoding

We input extracted features into an encoder-decoder model built with stacked RNN layers, specifically using LSTMs and GRUs with three layers each. To prepare the features for encoding, we first apply a fully connected layer to reshape them into a 512-dimensional vector for each frame. These reshaped features are then fed into the encoder, which processes them through its layers to produce a final hidden state. This hidden state is passed to the decoder, along with an initial input vector of 512 zeros for the first decoding step. As the decoder processes this input, it produces outputs that are transformed into probabilities over the vocabulary by a fully connected layer with a LogSoftmax activation function. The word with the highest probability is selected, embedded, and used as the input for subsequent decoding steps, allowing the decoder to generate the sequence of words over time.

The output of the encoder-decoder model consists of sequences of one-hot encoded vectors, each representing a word or special token (such as <BOS>, <EOS>, <PAD>, or <UNK>) from the vocabulary, with each vector having a dimension equal to the vocabulary size.

C. Training & Testing

We train the model using an NVIDIA RTX3060 until convergence, employing the cross-entropy loss function to calculate losses at the sentence level. To emphasize meaningful tokens during training, we assign a weight of 0.5 to the <PAD>token. We use the SGD optimizer with a learning rate of 0.01, a momentum of 0.9, and a batch size of 1. For testing, we evaluate the model on a test dataset consisting of unseen videos from the video corpus. The generated sentences are compared to the ground-truth sentences using metrics such as BLEU, ROUGE-L, and accuracy.

TABLE I
SLP MODEL PERFORMANCE COMPARISON.

Config	Metrics			
	BLEU-3	BLEU-4	ROUGE-L	Accuracy
A	0.8593	0.7464	0.9580	**0.9889**
B	**0.8694**	**0.7512**	**0.9659**	**0.9889**
C	0.8476	0.7316	0.9483	0.9855

VI. EXPERIMENTS

In this section, we present the experiments and results obtained. The goal is to identify the most suitable model and configuration for SLP and SLR by evaluating their performance using various metrics.

A. SLP and SLR Model Performance Analysis

For the evaluation metrics, we utilise the BLEU score [18], ROUGE score [19] and the accuracy rates. In specific, we focused on the performance of model in BLEU-3, BLEU-4, ROUGE-L and accuracy scores on test sets in both video corpus and text corpus.

SLP. We evaluate the text-to-gloss model IV-A. The table I shows that *Config B* performs exceptionally well, achieving the highest score on the metrics BLEU-3, BLEU-4, ROUGE-L and Accuracy. *Config B* performs best in all possible metrics may due to its architectural alignment with the encoder, maintaining the same hidden size of 336. In contrast, *Config C*, despite utilizing a pre-trained model, performs lower than *Config B*, which it might be due to the misalignment between the nature of the target data and the features learned in the pre-trained models.

TABLE II
SLR MODEL PERFORMANCE COMPARISON

Model	Metrics			
	BLEU-3	BLEU-4	ROUGE-L	Accuracy
LSTM	0.4645	0.3378	0.8350	0.7010
GRU	**0.5050**	**0.3643**	**0.8692**	**0.7580**

SLR. We evaluate the encoder-decoder model V-B. As shown in the table II, the *GRU* model performs exceptionally well by achieving the highest score in the metrics BLEU-3, BLEU-4, ROUGE-L and Accuracy Score. This shows that the *GRU* model performs better than the *LSTM* model in all evaluated metrics for SLR, so we used the *GRU* model as the SLR model in our BIM two-way communication system.

B. User Testing

We conducted user testing using a new framework designed to measure the effectiveness of SLP and SLR systems. Our participants, nine in total, were deaf and communicated mainly using BIM. Although their primary means of communication was sign language, most of them were proficient in written BM and some were also proficient in written English and Chinese. This diversity of language backgrounds was taken into account to provide comprehensive feedback on the systems. Participants rated their satisfaction with the systems on predetermined criteria using a scale of 1 to 5.

1) *Hand Accuracy*: How closely does the avatar's hand movements resemble those of real signers?
2) *Facial Expressions*: How realistic are the facial expressions portrayed by the avatar in videos?
3) *Speed*: How well does the speed of the translations affect user experience?
4) *Interactivity*: To what extent does the system facilitate interaction in an educational context?
5) *Complexity*: How complex is it to perform sign-to-text or text-to-sign translations without assistance?

The results shown in Table III indicate slightly higher overall satisfaction for both SLP and SLR systems. Scores range from 1 to 5, where 1 indicates poor performance and 5 indicates excellent performance. The average satisfaction scores for the five key criteria of the SLP system are as follows: *Hand Accuracy* scored 3.67, indicating users found the hand movements accurate and useful; *Facial Expressions* scored 3.11, suggesting room for improvement in conveying emotional nuances; *Speed* received a 3.22, showing moderate satisfaction with responsiveness and real-time performance; *Interactivity* scored 3.33, indicating satisfactory interaction but highlighting the need for a more seamless and intuitive experience; and *Complexity* scored 3.44, reflecting moderate satisfaction with the system's overall ease of use.

TABLE III
USABILITY SCORES FOR SLP & SLR SYSTEMS
(CRITERIA SCORES RANGING FROM 1 TO 5)

Criteria	Systems	
	SLP	SLR
Hand Accuracy	3.67	-
Facial Expressions	3.11	-
Speed	3.22	3.22
Interactivity	3.33	3.33
Complexity	3.44	3.78

The average satisfaction scores for the three key criteria of the SLR model are as follows: *Complexity* scored the highest at 3.78, reflecting that users found the system relatively straightforward to use independently, without requiring additional guidance; *Interactivity* scored 3.33, reflecting reasonably intuitive and seamless user interaction, with potential for enhanced user-friendliness; and *Speed* received a score of 3.22, suggesting that the recognition speed is somewhat lacking and could benefit from improvements in real-time responsiveness to enhance the fluency and naturalness of communication.

To evaluate the impact of our two-way sign language communication system on promoting inclusive education, we administered a survey featuring binary (yes or no) questions after participants had the opportunity to test the systems themselves. Analysis of the survey data showed that a significant majority—5 out of 9 participants (55.6%) recognized the potential of our system to promote equal educational opportunities for Deaf individuals. Moreover, all respondents unanimously agreed that the system could effectively bridge the communication gap between deaf and non-deaf individuals. Such communication is essential for enhancing accessibility within inclusive education systems.

VII. CONCLUSION

The implementation of sign language recognition and production technologies marks a substantial step towards overcoming linguistic barriers that hinder the inclusion of the deaf in mainstream education. Utilizing deep learning models, particularly the Transformer-based architecture for text-to-gloss translation, ensures high accuracy and efficiency in processing sign language data. Additionally, the avatar-based SLP module enhances the system's accessibility and usability. The successful implementation and testing of this system highlight its practicality and potential for wider application, paving the way for more inclusive educational opportunities for deaf students.

However, current limitations include a reliance on a relatively small BIM dataset, which results in low model generalization power and affects performance across diverse sign languages and contexts, indicating room for improvement towards practical application.

Future Work. Future work will focus on expanding the BIM dataset and iteratively improving the deep learning models to enhance performance for both SLP and SLR.

ACKNOWLEDGMENT

We would like to extend special gratitude to SSD for their invaluable support, contributions, and feedback as signers. Appreciation is also given to Alison Dwyne G. Abuan, Zahin Masrur Rahman, Dr. Joel Than Chia Ming and Miss Choo Ai Ling from SUTS for their critical roles in data collection and processing. Additionally, we would like to especially thank Albert Wong Tuong Chui (current Chairman of Sarawak Deaf Community Services Association and former Chairman of SSD) for his deep insights into Deaf culture and BIM.

REFERENCES

[1] E.-J. Hoogerwerf, K. Mavrou, and I. Traina, *The role of assistive technology in fostering inclusive education : strategies and tools to support change*. Routledge, Taylor & Francis Group, 2021.

[2] M. Chin, "The zero reject policy: A way forward for inclusive education in malaysia?" *International Journal of Inclusive Education*, vol. 27, pp. 1–15, Nov. 2020. DOI: 10.1080/13603116.2020.1846800.

[3] T. Tedla and D. Negassa, "The inclusive education for deaf children in primary, secondary and preparatory schools in gondar, ethiopia," *Jurnal Humaniora*, vol. 31, p. 177, Dec. 2019. DOI: 10.22146/jh.44767.

[4] M. N. Osman, K. A. Sedek, Nur, Muhamad, and M. Maghribi, "Hearing assistive technology: Sign language translation application for hearing-impaired communication," *Springer eBooks*, pp. 1–11, Jan. 2020. DOI: 10.1007/978-981-15-3434-8_1. (visited on 07/15/2024).

[5] W. Holmes and I. Tuomi, "State of the art and practice in ai in education," *European Journal of Education*, vol. 57, pp. 542–570, Oct. 2022. DOI: 10.1111/ejed.12533. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/ejed.12533.

[6] S. Garg and S. Sharma, "Impact of artificial intelligence in special need education to promote inclusive pedagogy," *International Journal of Information and Education Technology*, vol. 10, pp. 523–527, 2020. DOI: 10.18178/ijiet.2020.10.7.1418.

[7] N. D. Center, *Importance of effective communication between deaf and hearing individuals*, 2019. [Online]. Available: https://nationaldeafcenter.org/wp-content/uploads/2022/11/Importance-of-Effective-Communication-Between-Deaf-and-Hearing-Individuals.pdf.

[8] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018. DOI: 10.1109/cvpr.2018.00812.

[9] N. Vasani, P. Autee, S. Kalyani, and R. Karani, "Generation of indian sign language by sentence processing and generative adversarial networks," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2020. DOI: 10.1109/iciss49785.2020.9315979.

[10] J.-H. Kim, E. J. Hwang, S. Cho, D. H. Lee, and J. Park, *Sign language production with avatar layering: A critical use case over rare words*, N. Calzolari, F. Béchet, P. Blache, *et al.*, Eds., ACLWeb, Jun. 2022. [Online]. Available: https://aclanthology.org/2022.lrec-1.163 (visited on 07/22/2024).

[11] L. Hu, L. Gao, Z. Liu, and W. Feng, *Continuous sign language recognition with correlation network*, arXiv.org, Mar. 2023. DOI: 10.48550/arXiv.2303.03202. [Online]. Available: https://arxiv.org/abs/2303.03202.

[12] J. Shin, A. S. Musa Miah, M. A. M. Hasan, *et al.*, "Korean sign language recognition using transformer-based deep neural network," *Applied Sciences*, vol. 13, p. 3029, Jan. 2023. DOI: 10.3390/app13053029. [Online]. Available: https://www.mdpi.com/2076-3417/13/5/3029 (visited on 10/21/2023).

[13] I. Papastratis, K. Dimitropoulos, and P. Daras, "Continuous sign language recognition through a context-aware generative adversarial network," *Sensors*, vol. 21, p. 2437, Apr. 2021. DOI: 10.3390/s21072437.

[14] A. Duarte, S. Palaskar, L. Ventura, *et al.*, "How2sign: A large-scale multimodal dataset for continuous american sign language," *DIGITAL.CSIC (Spanish National Research Council (CSIC))*, Jun. 2021. DOI: 10.1109/cvpr46437.2021.00276.

[15] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, arXiv.org, Dec. 2017. DOI: 10.48550/arXiv.1706.03762. [Online]. Available: https://arxiv.org/abs/1706.03762.

[16] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv.org, Dec. 2014. [Online]. Available: https://arxiv.org/abs/1412.6980.

[17] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv.org, Apr. 2015. [Online]. Available: https://arxiv.org/abs/1409.1556.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001. DOI: 10.3115/1073083.1073135. [Online]. Available: https://dl.acm.org/citation.cfm?id=1073135.

[19] C.-Y. Lin, *Rouge: A package for automatic evaluation of summaries*, ACLWeb, Jul. 2004. [Online]. Available: https://aclanthology.org/W04-1013.

PRTGaussian: Efficient Relighting Using 3D Gaussians with Precomputed Radiance Transfer

Libo Zhang, Yuxuan Han, Wenbin Lin, Jingwang Ling, and Feng Xu
School of Software and BNRist, Tsinghua University
E-mail: zhanglb21@mails.tsinghua.edu.cn Tel: +86-15393595620

***Abstract*—We present PRTGaussian, a realtime relightable novel-view synthesis method made possible by combining 3D Gaussians and Precomputed Radiance Transfer (PRT). By fitting relightable Gaussians to multi-view OLAT data, our method enables real-time, free-viewpoint relighting. By estimating the radiance transfer based on high-order spherical harmonics, we achieve a balance between capturing detailed relighting effects and maintaining computational efficiency. We utilize a two-stage process: in the first stage, we reconstruct a coarse geometry of the object from multi-view images. In the second stage, we initialize 3D Gaussians with the obtained point cloud, then simultaneously refine the coarse geometry and learn the light transport for each Gaussian. Extensive experiments on synthetic datasets show that our approach can achieve fast and high-quality relighting for general objects. Code and data are available at https://github.com/zhanglbthu/PRTGaussian.**

I. Introduction

Relightable view synthesis has played a crucial role in computer graphics and computer vision for a long time [1]–[3]. It has many applications such as augmented reality and virtual object insertion. However, decoupling the lighting and reflectance information from the visual inputs and performing high-quality relighting is still slow, ill-posed and challenging.

Recently, some studies [2], [4] try to employ inverse rendering techniques to explicitly estimate objects' intrinsic properties (*i.e.* geometry and material) and scene lighting from images. However, these methods struggle to model complex light transport, such as subsurface scattering and indirect illumination. To tackle this problem, some methods [1], [5] have attempted to directly model the light transport of objects using multi-view one-light-at-a-time (OLAT) datasets. Nevertheless, such approaches often require cumbersome representation and intensive sampling, resulting in slow training and rendering speeds, severely limiting their application scenarios. Realtime relightable view-synthesis can improve interactive applications like VR and gaming, enabling dynamic, user-responsive lighting.

In this paper, we propose PRTGaussian, a novel framework for fast training and realtime relighting. Given multi-view OLAT images of general objects, real-time view synthesis and relighting can be realized after efficient training. Specifically, we use 3D Gaussians [6] with high-order spherical harmonics as the scene representations. Compared to implicit representations like NeRF [7], this explicit representation allows for more efficient forward rendering, resulting in faster training and inference speeds. The computation of object's appearance is based on the idea of precomputed radiance transfer. To be specific, for each Gaussian, we take its encoded position as input to regress its transport coefficients via a neural network, the final color is then obtained by combining this information with its albedo and lighting approximated by spherical harmonics. However, jointly optimizing the geometry and the appearance of objects is challenging due to the high ambiguity and complexity involved in accurately capturing both aspects simultaneously. To mitigate this problem, we propose a two-stage training strategy. In the first stage, we construct a multi-view images of fixed illuminations from the OLAT dataset and obtain the position initialization of Gaussians from it using vanilla 3D Gaussian Splatting. In the second stage, we initialize this point cloud as 3D Gaussians and proceed with further training, which refines the geometry of 3DGS and learns the light transport for each Gaussian.

In summary, the contributions of this paper can be outlined as follows:

- We propose an efficient relighting framework that utilizes 3D Gaussians for scene representation and precomputed radiance transfer for appearance modeling.
- We introduce a two-stage training process, beginning with coarse point cloud reconstruction and 3DGS initialization from multi-view images, followed by the refinement of 3D Gaussians, along with the learning of light transport.
- We conduct extensive experiments to demonstrate that our method achieves high-quality relighting effects that are comparable with SOTA while maintaining fast rendering speeds.

II. Related Work

A. 3D Scene Representation

Traditional scene representations [8], [9], including mesh-based, voxel-based, and point cloud methods, have been widely used due to their simplicity and efficiency. However, these approaches often struggle with complex geometries and detailed appearances. To address these challenges, implicit representations [7], [10] such as Neural Radiance Fields (NeRF) [7] have emerged, providing continuous and high-resolution modeling capabilities that significantly enhance the quality and realism of 3D scene representations. Specifically, NeRF represents a continuous scene as a 5D vector-valued function, with the 3D spatial position and 2D viewing direction as inputs, and the corresponding color and volume density as outputs. The image

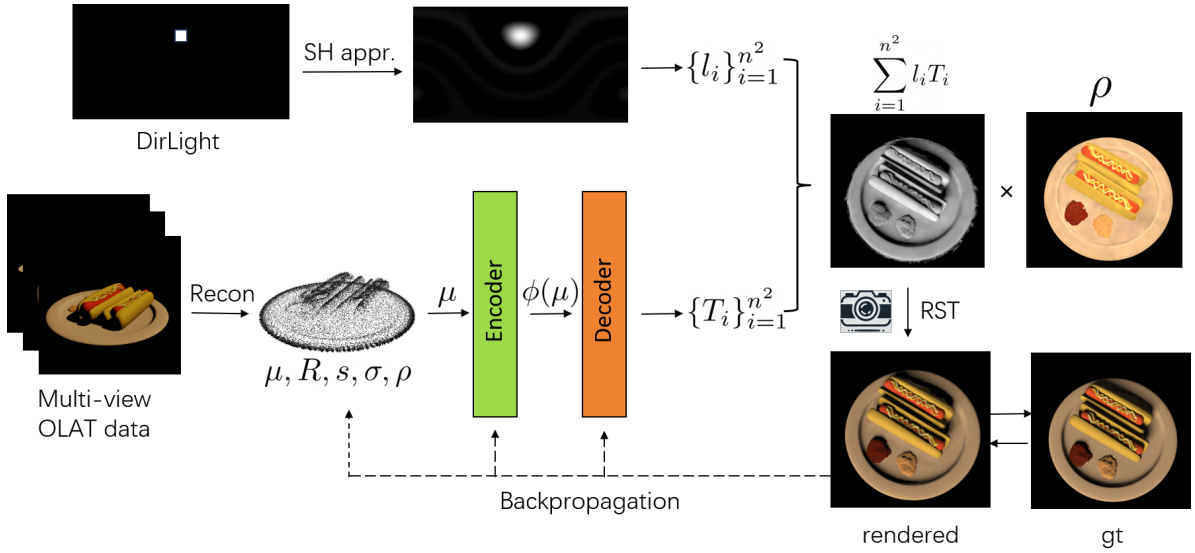


Fig. 1. **Overview of Pipeline**. Given a multi-view OLAT dataset, our method first reconstruct the initial point cloud as the initialization for the 3D Gaussians. In the next stage, the color of each Gaussian can be obtained by combining the learned radiance transfer, albedo and the light source approximated by spherical harmonics (SH). We jointly optimize the attributes of the Gaussians as well as the encoder and decoder by minimizing the loss between rendered images and ground truth images.

can then be rendered using volumetric rendering techniques. However, the sampling of rays and the use of a large Multi-Layer Perceptron significantly reduce its speed. Subsequent works [11], [12] have improved the quality and speed of NeRF, but their training and rendering time remain extremely high.

Recently, 3D Gaussians Splatting [6] has been introduced as an explicit scene representation. It uses 3D ellipsoids to represent scenes and employs a tile-based rasterization approach to render images, significantly improving training and rendering speeds while achieving high-quality novel view synthesis. The explicit nature of this representation also makes it more suitable for tasks such as scene editing and dynamic reconstruction. In this work, we use 3D Gaussians (3DGS) to represent objects.

B. Inverse Rendering

The goal of inverse rendering is to decompose the scene's geometry, material, and lighting components from images. Under fixed lighting conditions, some methods [2], [13] use differentiable rendering to estimate an object's depth, normal and material properties. Since the forward process of these methods is often based on simple physical rendering models such as the Cook-Torrance microfacet model [14], they can handle only limited geometries and materials, such as subsurface scattering, transparency, and anisotropic materials.

Other methods perform inverse rendering based on known multi-illumination conditions. NRHints [1] uses unstructured multi-view data with moving point light sources, employing two MLPs to separately represent the object's geometry and appearance. Alternative approaches [5], [15] learn appearance representations based on one-light-at-a-time (OLAT) data, but they typically focus on specific objects, such as human bodies and faces. We also utilize the OLAT dataset, but our method achieve lighting decoupling for general objects without relying on any priors.

C. Precomputed Radiance Transfer

In traditional computer graphics, rendering global illumination effects for a scene is costly due to the extensive ray tracing and numerous light bounce computations required. To achieve real-time rendering with global illumination, Sloan et al. [16] introduced the method of precomputed radiance transfer (PRT), which precomputes the coefficients of incident light projected onto spherical harmonics bases and the transfer vector at each point on the surface. The transfer vector, also represented by spherical harmonics coefficients, describes how each point on the surface interacts with incident light, including global transport effects such as reflections and shadows. At runtime, the light vector and transfer vector are integrated to obtain the radiance at each point. To address the low-frequency limitations of spherical harmonics, [17][3] introduced new representations such as nonlinear Gaussian functions and spherical Gaussians based on this foundation. However, these methods rely on known geometry and material properties.

We aim to model the radiance transfer at each point solely from images. Specifically, we use two networks to separately encode and decode the radiance transfer vector.

III. Method

Given the multi-view OLAT data, our goal is to achieve free-viewpoint relighting of the object. An overview of our method is shown in Fig.1. In the first stage, we reconstruct the initial positions for the 3D Gaussians. In the second stage, we refine the properties of 3DGS and learn the radiance transfer. In this section, we provide details on data acquisition

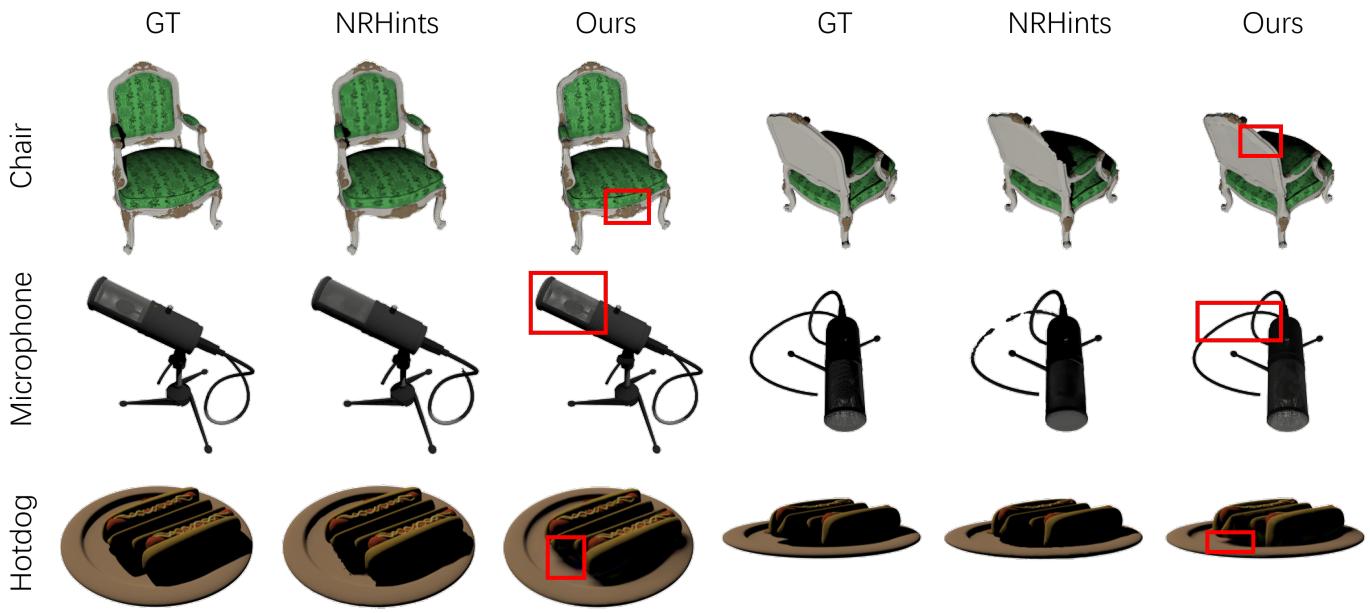


Fig. 2. **Qualitative Comparisons.** We performed qualitative comparisons with the state-of-the-art method NRHints[1]. Here we present the ground truth and the rendering results of different methods for three subjects under two novel lighting and view conditions.

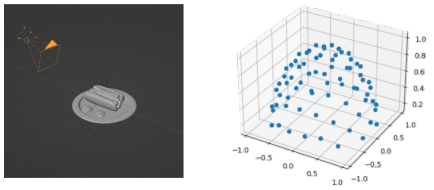


Fig. 3. Data Synthesis Setup. *Left* represents the setup in Blender and *Right* represents the sampled positions of the cameras and light sources.

(Sec.III-A), initial geometry reconstruction (Sec.III-B) and radiance transfer learning (Sec.III-C).

A. Data Acquisition

We use a one-light-at-a-time (OLAT) dataset similar to the light-stage [18] setup, where camera and light information are known. We synthesize a multi-view OLAT dataset for general objects in Blender. Specifically, we uniformly sample 25 camera positions and 200 lighting positions on the upper hemisphere of the object. The lighting is considered as directional light, with the detailed setup shown in Fig.3.

B. Initial Geometry Reconstruction

We use a set of 3D Gaussians (3DGS) to represent the object's geometry. Each Gaussian can be defined as

$$g_k = \{\boldsymbol{\mu}, \mathbf{R}, \boldsymbol{s}, \sigma, \boldsymbol{\rho}\} \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^3$ is the 3D position, $\mathbf{R} \in SO(3)$ is the rotation matrix, $\boldsymbol{s} \in \mathbb{R}^3$ is the per-axis scale factors, $\sigma \in \mathbb{R}$ is the opacity value and $\boldsymbol{\rho} \in \mathbb{R}^3$ is the albedo. The covariance matrix of each Gaussian can be represented as:

$$\Sigma = \mathbf{R}\text{diag}(\boldsymbol{s})\text{diag}(\boldsymbol{s})^T\mathbf{R}^T \tag{2}$$

In the splatting process, 3D Gaussians are first projected onto the 2D plane,

$$\Sigma' = JV\Sigma V^T J^T \tag{3}$$

where $J \in \mathbb{R}^{2\times3}$ is the Jacobian of the projective transformation, $V \in \mathbb{R}^{3\times3}$ is the viewing transformation, and $\Sigma' \in R^{2\times2}$ is the covariance matrix of the projected 2D Gaussians. The color $\mathbf{C}_p$ of each pixel can be computed using accumulated volumetric rendering as follows:

$$\mathbf{C}_p = \sum_{k\in N} \boldsymbol{c_k}\alpha_k \prod_{j=1}^{k-1}(1-\alpha_j) \tag{4}$$

where k is the index of N ordered Gaussians, $\alpha_k \in \mathbb{R}$ is estimated from the 2D covariance matrix and the opacity of each Gaussian, and $c_k \in \mathbb{R}^3$ is the color, which is determined by the albedo, radiance transfer, and incident light.

Due to the high ambiguity in jointly optimizing geometry and appearance, we first reconstruct a coarse geometry of the object from multi-view images under uniform lighting. The uniformly illuminated multi-view data is obtained by summing and averaging the illumination for each viewpoint. Then, We use the vanilla 3D Gaussian Splatting[6] method for coarse geometry reconstruction from these images. The generated point cloud is applied for Gaussian initialization in the next stage.

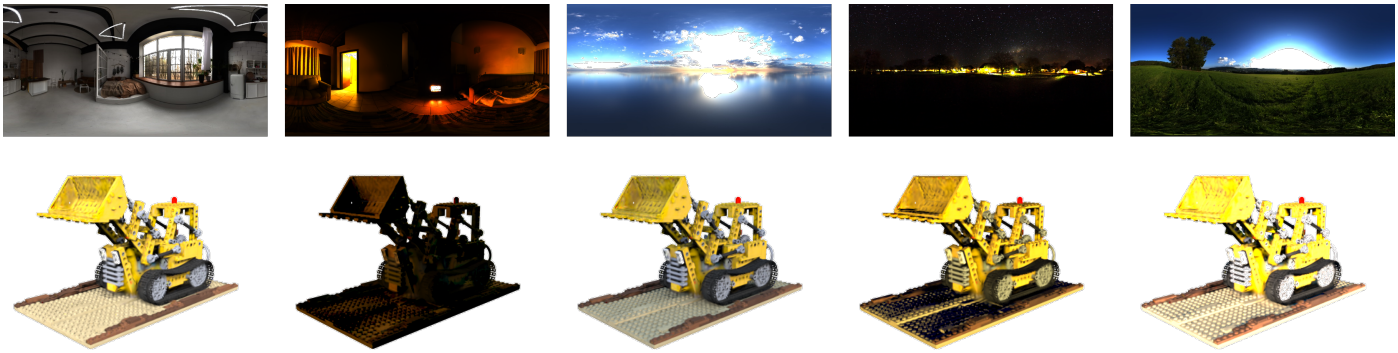


Fig. 4. **Qualitative Results.** The top shows different environment maps, while the bottom displays the corresponding results using our model trained on the OLAT dataset.

C. Radiance Transfer Learning

In the second stage, we learn the radiance transfer for each 3D Gaussian. The basic rendering equation can be written as:

$$L(o) = \int_{\Omega} L(i)V(i)\rho(i,o)\max(0, \mathrm{n}\cdot i)di \tag{5}$$

where i and o represent the incident and outgoing directions, while L, V, ρ, n represent the lighting, visibility, albedo, and normal properties, respectively.

In this work, we primarily focus on diffuse materials where each Gaussian has a specific albedo $\rho(i,o) = \rho$. Based on the concept of Precomputed Radiance Transfer (PRT), the light and radiance transfer can be precomputed as follows:

$$L(i) = \sum_{j=1}^{n^2} l_j B_j(i) \tag{6}$$

$$T_j = \int_{\Omega} B_j(i)V(i)max(0, \mathrm{n}\cdot i)di \tag{7}$$

where n is the order of the spherical harmonics, B_j is the spherical harmonics basis, $l_j \in \mathbb{R}^{n^2}$ is the coefficient of the light source projected onto SH space, and $T_j \in \mathbb{R}^{n^2}$ is the radiance transfer term. Therefore, equation 5 can be rewritten as:

$$L(o) = \rho \sum_{j=1}^{n^2} l_j T_j \tag{8}$$

The light coefficient l_j can be obtained from the input lighting by direct projection. The radiance transfer term T_i is encoded with a neural network. Specifically, we first encode the position μ of each Gaussian into a higher-dimensional vector [19] $\phi(\mu)$, and then decode it using an MLP to obtain the corresponding T_i.

During the training process, the rendered image is compared with the ground truth image to calculate the loss:

$$L = (1-\lambda)L_1 + \lambda L_{D-SSIM} \tag{9}$$

we use the gradient of this loss to update the parameters of the encoder and decoder, while simultaneously refining all the properties of each Gaussian.

IV. Experiments

A. Implementation Details

We implemented our method in Python using the PyTorch framework. Our encoder uses hash-grid encoding [19] with the number of levels, number of features per level, and hash table size set to 32, 8, and 2^{20}, respectively. Our decoder employs a four-layer fully connected MLP, with each layer consisting of 256 neurons. To capture higher frequency information, we set the order of the spherical harmonics basis to $n = 9$. During training, we used the Adam optimizer and set λ to 0.2. The entire training process was conducted on an NVIDIA RTX 4090, with a total training time of approximately 15 minutes per scene.

B. Quantitative Results

We conducted a comprehensive comparison against the state-of-the-art field-based method NRHints [1], which uses volume rendering and represents geometry in SDF. Tab. I summarizes the quantitative comparison results on synthetic data, with each metric averaged over 3 subjects. We also compared the rendering speed (per frame) and training time of these two methods. All comparison experiments are conducted under the same conditions and on the same equipment.

TABLE I
Quality Comparison.

	PSNR ↑	**SSIM ↑**	**LPIPS ↓**	**render**	**train**
NRHints [1]	30.77	**0.9729**	0.0263	17.95s	28h
Ours	**33.63**	0.9438	**0.0255**	**0.0030s**	**15m**

The 3DGS-based scene representation enables our method to facilitate efficient training and inference while simultaneously capturing fine details of apperances. The experimental results demonstrate that our approach significantly improves training and rendering speeds compared to previous method, while achieving comparable rendering quality.

C. Qualitative Results

We also conducted a qualitative comparison with the current top-performing technique (NRHints) as shown in Fig. 2.

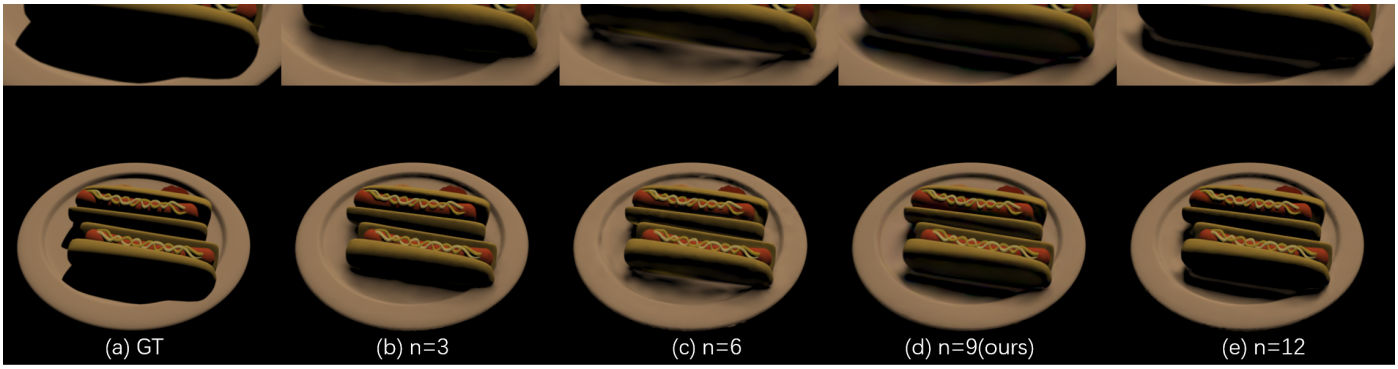


Fig. 5. **Ablation Study: The order of the SH basis.** Compared to the ground truth (a), using higher-order SH coefficients can enhance the sharpness of shadows. To balance quality and efficiency, we choose $n = 9$ in our method.

Thanks to the scene representation capability of 3DGS and the learning of radiance transfer, our method can recover fine details in both geometry and appearance, such as *Chair's patterns, Microphone's wires and Hotdog's shadows*, which other methods fails to reproduce.

Once trained on the multi-view OLAT dataset, our method can achieve relighting under arbitrary lighting conditions. In Fig. 4, we present the qualitative results of HDRI relighting. For environment maps, our approach first projects them into SH space and then performs efficient and high-quality relighting.

D. Ablation Studies

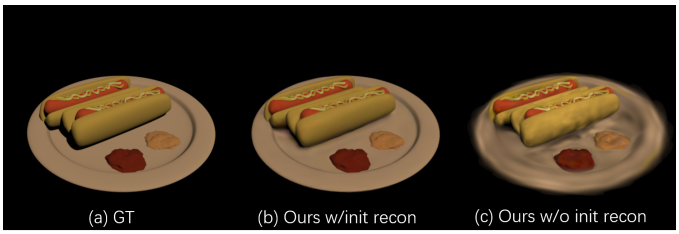


Fig. 6. **Ablation Study: Initial Geometry Reconstruction.** Compared to the ground truth (a), using initial geometry reconstruction (b) leads to more accurate rendering results than randomly initialized geometry (c).

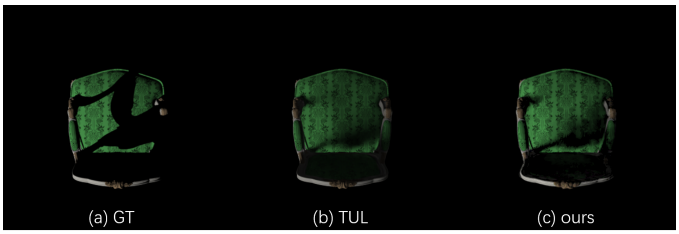


Fig. 7. **Limitations.** The left (a) is the ground truth, the middle (b) is the results rendered in Blender using directional light approximated by SH, which represents our theoretical upper limit (TUL), and the right (c) shows the results obtained by training our model with the ground truth. It can be seen that our method is inherently difficult to achieve high-frequency effects such as hard shadows limited by the low-frequency nature of SH basis.

In this section, we provide several ablation studies to validate the key design choices of our method.

Initial Geometry Reconstruction. In stage 1, we performed initial point cloud reconstruction to provide a better initialization for 3D Gaussians (3DGS). We compare our approach with one that does not use initial reconstruction but instead initializes 3DGS by sampling them uniformly in a cube to verify the necessity of this design. Fig. 6 shows our method achieves significantly better results. This improvement is due to the high ambiguity in inverse rendering and the challenge of simultaneously optimizing the geometry and appearance of 3DGS using OLAT datasets.

High Order SH Basis. In our method, both the lighting and radiance transfer are represented using spherical harmonics (SH), and the number of SH basis coefficients is positively correlated with the square of its order. To capture high-frequency effects (such as shadows) while controlling the number of SH coefficients to improve computational efficiency and fit GPU memory, our method uses spherical harmonics up to the 8th order (nine orders in total).

We compared the results with those of different orders (n=3, 6, 12) and found that our method balances the efficiency and quality. Fig.5 shows that when the order is too low, it is difficult to represent high-frequency information such as shadows, and when the order is too high, the memory and computational requirements increase significantly with minimal improvement in rendering quality.

V. Limitations

Although we have accelerated the training and inference speeds of existing relighting methods and achieved decent rendering results, our approach still has some limitations and shortcomings.

One major drawback is that we currently only consider diffuse materials and do not consider specular highlights and other view-dependent effects. Incorporating the view direction as a condition in the pipeline holds promise for addressing this issue in the future.

Another drawback is that although we use high-order SH to approximate the information of light and radiance transfer, our

method still struggles to reproduce high-frequency rendering effects such as the sharp edges in hard shadows (as shown in Fig.7). This is due to the inherently low-frequency nature of spherical harmonics. Explicitly considering the visibility of 3DGS in future work will help alleviate this situation.

VI. Conclusions

In this study, we present PRTGaussian, a novel relighting framework that combines 3D Gaussian Splatting and Precomputed Radiance Transfer to achieve efficient and high-quality relighting. Our approach demonstrates significant improvements in training and rendering speeds compared to existing techniques and efficiently utilizes multi-view OLAT data for realistic relighting. While our method excels in these areas, it is currently limited to diffuse materials and has difficulty capturing high-frequency details. Future work could address these limitations by integrating view-dependent effects and enhancing the representation of high-frequency components.

VII. Acknowledgments

This work was supported by the National Key R&D Program of China (2023YFC3305600), the Zhejiang Provincial Natural Science Foundation (No.LDT23F02024F02), the NSFC (No.61822111, 62021002), and the Key Research and Development Project of Tibet Autonomous Region (XZ202101ZY0019G). This work was also supported by THUIBCS, Tsinghua University, and BLBCI, Beijing Municipal Education Commission. Feng Xu is the corresponding author.

References

[1] C. Zeng, G. Chen, Y. Dong, P. Peers, H. Wu, and X. Tong, "Relighting neural radiance fields with shadow and highlight hints," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.

[2] J. Gao, C. Gu, Y. Lin, *et al.*, "Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing," *arXiv preprint arXiv:2311.16043*, 2023.

[3] J. Wang, P. Ren, M. Gong, J. Snyder, and B. Guo, "All-frequency rendering of dynamic, spatially-varying reflectance," in *ACM SIGGRAPH Asia 2009 papers*, 2009, pp. 1–10.

[4] Y. Shi, Y. Wu, C. Wu, *et al.*, "Gir: 3d gaussian inverse rendering for relightable scene factorization," *arXiv preprint arXiv:2312.05133*, 2023.

[5] X. Zhang, S. Fanello, Y.-T. Tsai, *et al.*, "Neural light transport for relighting and view synthesis," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 1, pp. 1–17, 2021.

[6] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[7] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[8] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *Seminal graphics: pioneering efforts that shaped the field*, 1998, pp. 347–353.

[9] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[10] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021.

[11] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5855–5864.

[12] T. Hu, S. Liu, Y. Chen, T. Shen, and J. Jia, "Efficientnerf efficient neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 902–12 911.

[13] Z. Liang, Q. Zhang, Y. Feng, Y. Shan, and K. Jia, "Gs-ir: 3d gaussian splatting for inverse rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 644–21 653.

[14] R. L. Cook and K. E. Torrance, "A reflectance model for computer graphics," *ACM Transactions on Graphics (ToG)*, vol. 1, no. 1, pp. 7–24, 1982.

[15] S. Saito, G. Schwartz, T. Simon, J. Li, and G. Nam, "Relightable gaussian codec avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 130–141.

[16] P.-P. Sloan, J. Kautz, and J. Snyder, "Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 2002, pp. 527–536.

[17] T. U. N. Gaussian, "View-dependent precomputed light transport using nonlinear gaussian function approximations," 2006.

[18] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000, pp. 145–156.

[19] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

RepViT Based Lightweight Architecture for Distracted Driving Detection

Muwei Jian [1*], Yukun Ling [1], Rui Wang [1], Yanjie Zhong [1], Huihui Huang [1], Zhuang Miao [2], Chunxiao Ren [3] and Xiaoguang Li [4]

[1] Shandong University of Finance and Economics, Jinan, China

E-mail: jianmuweihk@163.com，20181844314@mail.sdufe.edu.cn, rachelwang@mail.sdufe.edu.cn

222115038@mail.sdufe.edu.cn, huanghuihui@mail.sdufe.edu.cn

[2] Business School, The University of New South Wales, Sydney, Australia

E-mail: z5509793@zmail.unsw.edu.au

[3] Shandong International Talent Exchange Service Center, Jinan, China

E-mail: renchunxiao@gmail.com

[4] Faculty of Information Tecnology, Beijing University of Technology, Beijing, China

E-mail: lxg@bjut.edu.cn

Abstract— **Distracted driving is a significant cause of traffic accidents and has becoming an increasingly severe road safety concern, in particular with driver distractions such as using phones, talking, and makeup applications. Therefore, real-time detection of driver behavior and reactions during driving is crucial for improving traffic safety. To resolve the problems of large parameter counts and inaccurate classification in existing distracted driving detection models, this paper presents a lightweight model, named RepViT-MCFM, which features low latency and robustly detects driver distraction. Specifically, we designed a multi-scale cross-fusion denoising module (MCFM) based on Haar wavelet transform, utilized the RepViT module improved from MobileNetV3 as the backbone feature extractor, and designed the RepViT-MCFM model with a low-latency simple classifier. We have conducted quantitative and qualitative evaluations on the State-Farm dataset, extensive experiments demonstrate that the designed RepViT model, compared to existing CNN-based detection models, maintains high accuracy while achieving lower latency and fewer parameters.**

I. INTRODUCTION

According to a report by the World Health Organization (WHO), around 1.19 million people die annually in road traffic accidents. Road traffic injuries are the leading cause of death for children and young people aged 5-29. Moreover, road traffic accidents result in substantial economic losses, with the losses from these accidents amounting to 3% of the GDP in most countries. Drivers using mobile phones are about four times more likely to be involved in a collision compared to those who do not use phones. Distracted driving incidents may be underestimated, as many drivers may not admit to using their phones before an accident. Using a mobile phone while driving slows reaction times, particularly braking time, making it difficult for drivers to stay in the correct lane and maintain a proper following distance [1]. In recent years, due to the rapid advancements in computing technology and the extensive deployment of communication mechanisms, autonomous driving technology has been highly anticipated to reduce dependence on human drivers. However, the research and development of autonomous driving computing systems still face numerous challenges and unresolved issues. Therefore, real-time detection of human driver distracted behaviors remains crucial.

Although many distracted driving detection methods have shown significant effectiveness, detecting distracted drivers face numerous challenges. First, this task requires a precise and robust algorithm to detect various driver behaviors and accurately identify unsafe driving actions. Secondly, driving is an activity that requires real-time reactions, so the algorithm must exhibit low latency in detection to provide the real-time monitoring of the driver [2]. Furthermore, to ensure the model's broad applicability across diverse scenarios, it must be capable of running on resource-limited mobile devices [3]. However, many studies have not adequately considered the complexity of real-world driving scenarios and the diversity of driver distraction behaviors, leading to low classification accuracy in identifying specific categories of distracted driving behaviors. To tackle this issue, this paper proposes the RepViT-MCFM network based on a lightweight convolutional neural network and conducts extensive testing in distracted driving.

The main contributions of this paper are summarized as follows: (1) We have redesigned and trained a lightweight distracted driving detection model, RepViT-MCFM, using the lightweight convolutional neural network RepViT as the backbone network for the first time. Experiments show that compared to classic CNN networks and lightweight networks such as MobileNet, RepViT significantly reduces trainable parameters and computational cost, greatly lowering latency, (2) We designed a Haar wavelet transform-based cross-fusion

module (MCFM) to improve the accuracy of distraction detection. Experiments show that the MCFM module can significantly enhance the model's robustness and anti-interference capabilities, further increasing its applicability in complex real-world environments, (3) We extensively tested the model on the State-Farm dataset, and results indicate that, compared to existing state-of-the-art deep learning-based methods, the proposed RepViT-MCFM model has fewer parameters and higher accuracy, indicating its broad application prospects distracted driving detection.

II. Related Work

Existing distracted driving detection methods can be roughly divided into two categories: those that rely on physiological signal sensors to monitor the driver's heart rate, brain waves, and other physiological signals, and those that rely on cameras to capture the driver's facial features and body posture. In terms of detecting driver physiological signals, Wang et al. [4] used novel elastic dry electrodes to obtain ECG signals from the palm to capture driver fatigue characteristics. However, wearing sensors for extended periods can cause discomfort to drivers, affecting the driving experience, and there is an unavoidable delay in physiological signals' response to fatigue, sometimes failing to promptly reflect the driver's real-time state. In recent years, thanks to significant advancements in areas such as CNNs [5][6], attention mechanisms [7], and Transformers [8], camera-based computer vision methods have made significant strides in the field of distracted driving detection. Khan et al. [9] proposed a multi-scale convolutional neural network integrated with channel attention. Ashiqur Rahman et al. [10] introduced a distracted driving detection system that combines embedded sensors with computer vision, comprehensively monitoring the driver's eyes, mouth, and heart rate. Additionally, it can detect mask-wearing status to meet the current pandemic prevention requirements. To make the model more suitable for deployment on edge devices and expand its application range, the authors designed a distracted driving detection model using the lightweight MobileNetV2 as the backbone network [11]. Nonetheless, the MobileNet model exhibits low accuracy in recognizing certain distracted behaviors, limiting its effectiveness in real-world applications.

III. Method

A. Overall Architecture

To overcome the challenges in current research, we propose a novel lightweight distracted driving detection model, RepViT-MCFM. The model uses a Haar wavelet transform module for real-time image preprocessing, effectively reducing background noise and improving the accuracy of recognizing detailed driver actions. The designed model employs RepViT [12] as the backbone network, demonstrating significant superiority in classification tasks compared to MobileNetV3. Additionally, compared to other distracted driving detection models [13], we developed a more straightforward classification method, reducing the model's latency.

Our input images first pass through the Multi-Scale Wavelet Convolution Fusion Module (MCFM), with each convolution layer followed by a ReLU activation function, and wavelet transform is used to extract low-frequency and high-frequency features. Features from different layers are fused using learnable parameters λ. After convolutional processing, the fused features enter the backbone network RepViT, which consists of multiple RepViTBlocks and a RepViTSEBlock to extract deep features further. Finally, the features are classified by a classifier composed of a global average pooling layer and a fully connected layer, outputting the final classification results. The overall architecture of the network is shown in Fig. 1.

B. Multi-Scale Cross Fusion Module

To mitigate the limitations of single preprocessing methods in handling complex images, we designed the Multi-Scale Cross-Fusion Module (MCFM), which fully utilizes the significant advantages of wavelet transform in image denoising, image fusion, and image enhancement [14]. Through wavelet transform, background noise can be effectively reduced, significantly enhancing the accuracy of recognizing detailed features. The MCFM module combines convolutional neural networks and Haar wavelet transform to enhance feature extraction capabilities at different frequencies and scales. The model performs feature fusion through learnable parameters λ, allowing it to automatically adjust the feature fusion ratio during training, enhancing its adaptability and robustness. The model includes three convolutional layers, ReLU activation functions, and feature extraction and fusion operations based on wavelet transform. The fused features undergo further processing by passing through a convolutional layer, which refines and transforms them to produce the final output of the model.The detailed equations for each component of the model are presented below:

$$x_i = \mathrm{Re}\,LU(Conv(x_i)) \quad (i = 1, 2, 3) \tag{1}$$

$$LF_i = Wavelet(x_i) \quad (i = 1, 2, 3) \tag{2}$$

$$LF_i = Wavelet(x_i) \quad (i = 1, 2, 3) \tag{3}$$

$$Output = Convout(LF) \tag{4}$$

C. Backbone

As one of the most advanced lightweight networks, RepViT outperforms existing lightweight Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) across various vision tasks, including image classification and object detection.The RepViT model uses early convolutions [15] as the backbone, which downsamples the input by a factor of 4 through two convolutions with a stride of 2. It employs

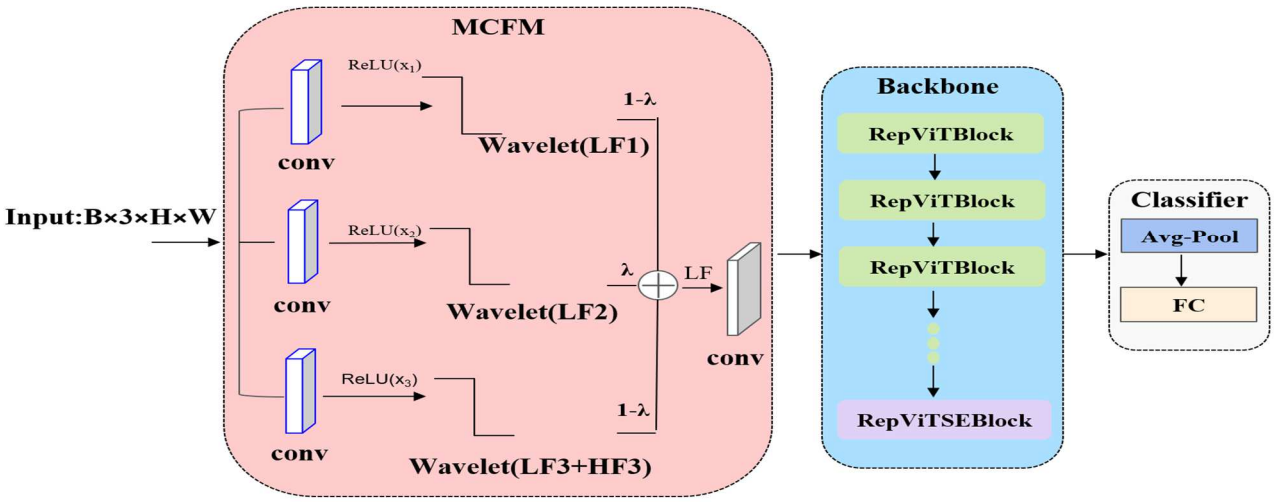


Fig. 1 Architecture of the proposed model

RepViT module, composed of structurally reparameterized depthwise convolutions [16] and a feedforward network. A depthwise downsampling module is implemented between its adjacent stages, utilizing depthwise convolutions with a stride of 2 and pointwise convolutions to perform spatial downsampling and channel dimension modulation, respectively. Moreover, squeeze-and-excitation layers [17] are applied across blocks at each stage. Compared to ViTs, the purely convolutional architecture of RepViT is more suitable for deployment on resource-constrained mobile devices, demonstrating significant advantages in latency for high-resolution vision tasks. Based on this, we selected RepViT as the backbone network. Furthermore, we balanced the relationship between the accuracy and latency of the SE layer in our design by adding the SE layer only in the last RepViT block, achieving high accuracy while maintaining low latency.

D. Classifier

We incorporate global average pooling and linear layers to accomplish the classification task. The global average pooling layer averages the spatial information of each channel in the feature map, reducing the feature dimensions and simplifying the model structure. The linear layer then maps these simplified features to the classification space. This design significantly reduces latency and computational complexity compared to more complex classifiers. Therefore, this approach is more suitable for distracted driving detection, which requires immediate response, as it ensures accuracy while enhancing the system's real-time performance and efficiency.

IV. EXPERIMENTS

A. Dataset

This study leveraged the StateFarm distracted driving detection dataset, available on Kaggle [18], one of the most commonly used datasets for detecting driver distraction. The StateFarm dataset consists of nine classes, with each image classified into one of these categories. These categories encompass the following driver behaviors: safe driving, texting with the right hand, talking on the phone with the right hand, texting with the left hand, talking on the phone with the left hand, operating the radio, reaching behind, hair and makeup, and talking to passengers, as illustrated in Fig. 2.

The State-Farm dataset primarily comprises two folders. The training folder contains 22,424 labeled images, while the test folder contains 29,700 unlabeled images. As a result, this study used the photos from the training folder (containing 22,424 images). We randomly divided the photos in the training folder into 60% for training, 20% for validation, and 10% for testing.

We enhanced the dataset according to the data augmentation strategies proposed in [19] to improve the generalization ability and robustness of the proposed detection algorithm model. We added color jitter, motion blur, Gaussian noise, horizontal flipping, and random scaling to the dataset to simulate the complex driving environment under natural conditions. Moreover, considering the post-pandemic tendency of people to wear masks while traveling, we also added masks to the faces in the dataset.

B. Training Settings

We implemented a learning rate decay strategy during the model training process to enhance training efficiency and effectiveness while preventing overfitting, with the initial learning rate set to 0.0001. In selecting the optimizer, our experiments revealed that while the model converged faster with the Adam optimizer, the training results were better with Stochastic Gradient Descent (SGD). Hence, we ultimately chose the SGD optimizer. Furthermore, we chose categorical cross-entropy as the loss function because it is well-suited for handling multi-class problems, making it ideal for training the distracted driving detection model.

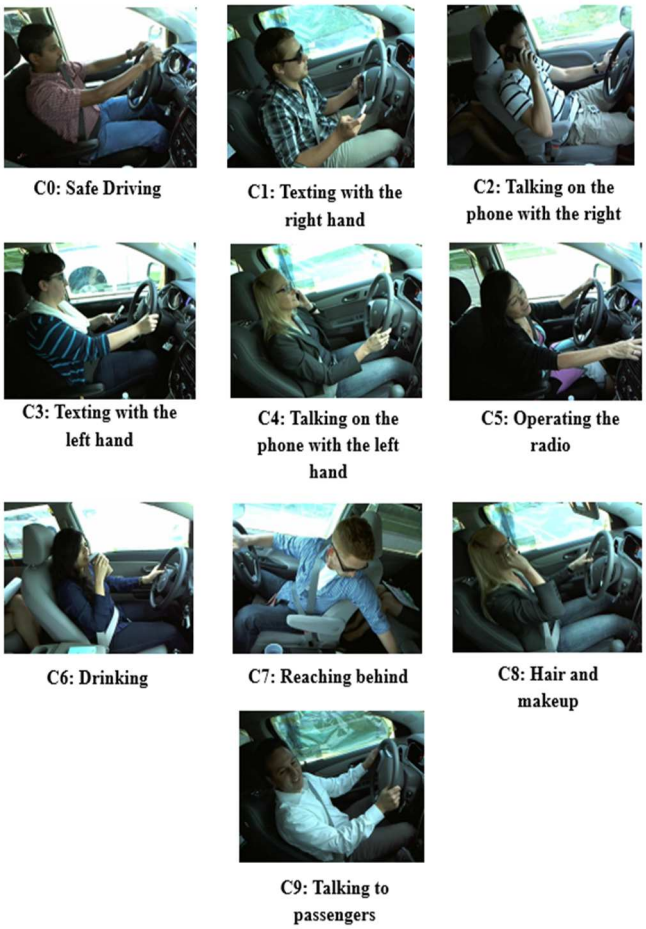


Fig. 2 Sample images representing from State Farm Dataset

C. *Results and Discussion*

We trained the model using the segmented training dataset and validated the results with the test data, performing a detailed performance analysis on the State-Farm distracted driving detection dataset. The training statistics of the model at the end of the training process are summarized in Table I. The results indicate that the model demonstrates high accuracy and low loss on the training and validation sets. This reveals that the model performs excellently on the training data and demonstrates good generalization ability on the validation data.

We also tracked the loss and accuracy of the model during the training process and visualized them upon completion. Fig. 3 and Fig. 4 show the loss curve and accuracy curve of the model during the training process. The loss curve in Fig. 3 shows that both training and validation losses rapidly decrease during the early stages of training, indicating that the model effectively learns features initially. After approximately 20 epochs, the training loss stabilizes, ultimately reaching around model shows good fitting ability on the training data while maintaining a low loss on the validation data, showcasing good generalization capability.

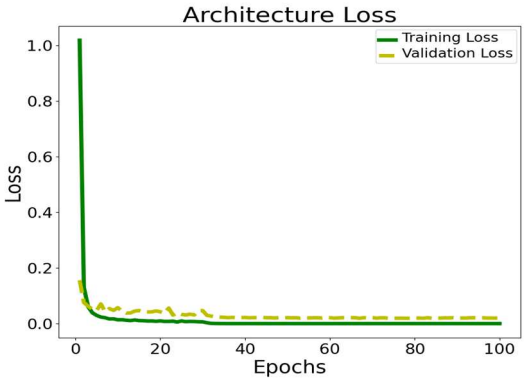


Fig. 3 Model Loss Curves

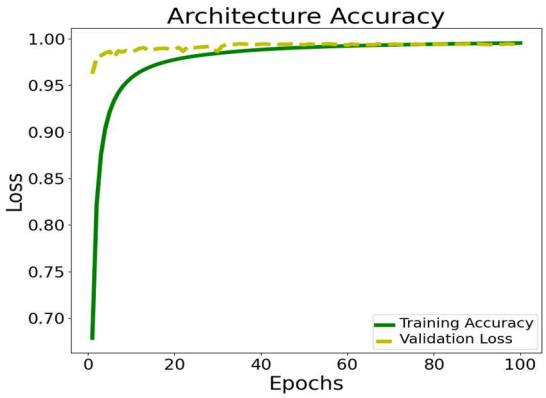


Fig. 4 Model Accuracy Curves

TABLE I. STATISTICS OF MODEL TRAINING

	Accuracy	Loss
Training	0.9953	0.0005
Validation	0.9940	0.0208

The accuracy curve in Fig. 4 reveals that training and validation accuracy rapidly increase in the early stages of training, indicating that the model effectively captures the data patterns. After about 20 epochs, the training accuracy stabilizes, ultimately reaching 0.9953, while the validation accuracy also stabilizes, with a final value of approximately 0.9940. This confirms that the model achieves extremely high accuracy on training and validation data, indicating its efficiency and robustness.

We compared our trained model with other benchmark models on the augmented segmented test set, and the experimental results are presented in Table II. The table summarizes the performance comparison of different backbone networks in the distracted driving detection task, including VGG16, MobileNet, and our RepViT-MCFM model. Specific metrics encompass the number of architecture parameters, FLOPs, test loss, and test accuracy. The experimental results

TABLE II. COMPARISON ON THE PREVIOUS WORK

Methods	Backbone	Architecture Parameters	FLOPs	Test Loss	Test Accuracy
Ref.[13]	VGG16	3.3×10^{7}	1.8585×10^{10}	0.4054	0.9203
Ref.[11]	MobileNet	4.9×10^{6}	4.4415×10^{7}	0.6189	0.8622
ours	RepViT	8.2×10^{5}	3.5025×10^{7}	0.1404	0.9550

confirm that our RepViT-MCFM model excels in all four aspects: parameter count, FLOPs, test loss, and test accuracy. The model has the fewest parameters, the lowest computational complexity, the smallest test loss, and the highest accuracy on the test data. This renders RepViT-MCFM an efficient model suitable for practical applications, showcasing great potential and wide application prospects in real driving scenarios.

The classification report of the model tested on the test set is shown in Table III. Specifically, the precision for category c0 is1.00, but the recall is relatively low at 0.59, resulting in an F1. Specifically, the precision for category c0 is 1.00, but the recall is relatively low at 0.59, resulting in an F1 score of 0.75. This indicates that although the model can correctly identify samples of category c0, a significant proportion of c0 samples are misclassified. In contrast, the precision, recall, and F1 scores for other categories (c1 to c9) are close to or reach 1.00, indicating that the model's classification performance on these categories is excellent.

TABLE III. CLASSIFICATION REPORT

	precision	recall	F1-score	support
C0	1.00	0.59	0.75	202
C1	1.00	1.00	1.00	204
C2	0.99	1.00	1.00	202
C3	0.74	1.00	0.85	216
C4	0.99	1.00	0.99	202
C5	0.96	1.00	0.98	204
C6	0.99	0.99	0.99	216
C7	0.99	1.00	1.00	234
C8	0.98	0.96	0.97	234
C9	1.00	0.99	0.99	220

TABLE IV. ANALYSES ON DWCNET PLACEMENT

MCFM	Test Accuracy
×	0.8782
√	0.9550

D. Ablation Study

To verify the effectiveness of our preprocessing module in the distracted driving detection model, we conducted ablation experiments on the segmented test set by removing the MCFM module during training, as shown in Table IV. Experiments attest that without the MCFM, the model's performance significantly decreases, underscoring the positive impact of the MCFM module on the model's performance.

V. CONCLUSIONS

In this paper, we designed a novel lightweight distracted driving detection model featuring cross-fusion preprocessing centered on Haar wavelet transform and using RepViT as the backbone. Compared to currently popular approaches, this model has fewer parameters and achieved an accuracy of 95.50% on the segmented test set, making it suitable for broad deployment on various edge devices. We believe this system can be applied to real-world driving scenarios to monitor and warn drivers in real-time to prevent distracted driving behaviors, thereby improving the safety of both passengers and drivers. The distracted driving detection model not only safeguards individual drivers' safety but also serves as an indispensable component of intelligent transportation systems, enhancing the safety, efficiency, and intelligence of transportation systems in various aspects.

VI. ACKNOWLEDGMENT

This work was supported by Taishan Young Scholars Program of Shandong Province; and Key Development Program for Basic Research of Shandong Province (ZR2020ZD44).

REFERENCES

[1] L. Liu, S. Lu, R. Zhong, et al., "Computing systems for autonomous driving: State of the art and challenges," IEEE Internet of Things Journal, vol. 8,no. 8, 2020.

[2] E. Michelaraki, C. Katrakazas, S. Kaiser, T. Brijs, and G. Yannis, "Real-time monitoring of driver distraction: state-of-the-art and future insights," Accident Analysis & Prevention, vol. 192, pp. 107241, 2023.

[3] E. Papatheocharous, C. Kaiser, J. Moser, et al., "Monitoring distracted driving behaviours with smartphones: an extended systematic literature review," Sensors, vol. 23, no.17, pp.7505, 2023.

[4] F. Wang, W. Yao, B. Lu, and R. Fu, "ECG-Based Real-Time Drivers' Fatigue Detection Using A Novel Elastic Dry Electrode," IEEE Transactions on Instrumentation and Measurement, 2023.

[5] Y. Xing, C. Lv, D. Cao and E. Velenis, "Multi-scale driver behavior modeling based on deep spatial-temporal representation for intelligent vehicles," Transp. Res. C Emerg. Technol., vol. 130, 2021.

[6] J. Methuku, "In-car driver response classification using deep learning (CNN) based computer vision," IEEE Trans. Intell. Veh. 2020.

[7] Y. Lin, D. Cao, Z. Fu, Y. Huang, and Y. Song, "Lightweight Attention-Based Network towards Distracted Driving Behavior Recognition," Applied Sciences, vol. 12, no. 9, pp. 4191, 2022.

[8] E. Akdag, Z. Zhu, E. Bondarev, "Transformer-based fusion of 2D-pose and spatio-temporal embeddings for distracted driver action recognition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[9] T. Khan, G. Choi, and S. Lee, "EFFNet-CA: an efficient driver distraction detection based on multiscale features extractions and channel attention mechanism," Sensors, vol. 23, no. 8, pp. 3835, 2023.

[10] A. Rahman, M. B. H. Hriday, and R. Khan, "Computer vision-based approach to detect fatigue driving and face mask for edge computing device," Heliyon, vol. 8, no. 10, 2022.

[11] M. A. B. Abbass, Y. Ban, "MobileNet-Based Architecture for Distracted Human Driver Detection of Autonomous Cars," Electronics, vol. 13, no. 2, pp. 365, 2024.

[12] A. Wang, H. Chen, Z. Lin, H. Pu, and G. Ding, "Repvit: Revisiting mobile cnn from vit perspective," arXiv preprint arXiv:2307.09283, 2023.

[13] Al Shalfan K A, M. Zakariah, "Detecting Driver Distraction Using Deep-Learning Approach," Computers, Materials & Continua, vol. 68, no. 1, 2021.

[14] J. Liu, H. Zhao, Z. Chen, et al., "A dynamic weights-based wavelet attention neural network for defect detection," IEEE Transactions on Neural Networks and Learning Systems, 2023.

[15] T. Xiao, M. Singh, E. Mintun, et al., "Early convolutions help transformers see better," Advances in neural information processing systems, vol. 34, pp. 30392–30400, 2021.

[16] X. Ding, X. Zhang, N. Ma, et al., "Repvgg: Making vgg-style convnets great again," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13733–13742, 2021.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141, 2018.

[18] StateFarm, State Farm Distracted Driver Detection. Kaggle, 2016.

[19] Y. Zhang, T. Li, C. Li, and X. Zhou, "A Novel Driver Distraction Behavior Detection Method Based on Self-Supervised Learning with Masked Image Modeling," IEEE Internet of Things Journal, 2023.

HSIC as Information Compression for Training Deep Neural Network

Roshan Birjais Kevin I-Kai Wang Waleed Abdulla
The University of Auckland, New Zealand
E-mail: rsof418@aucklanduni.ac.nz E-mail: kevin.wang@auckland.ac.nz E-mail: w.abdulla@auckland.ac.nz

***Abstract*—Deep Learning has made significant strides in recent years, particularly in supervised learning tasks, leading to the development of numerous architectures aimed at improving various aspects of model performance. Despite the effectiveness of backpropagation and stochastic gradient descent in training deep networks, these methods are often hindered by exploding and vanishing gradients, and significant memory overhead. Alternative training strategies that reduce complete reliance on BP are increasingly being explored to address these limitations. This paper proposes a simple architecture that uses Hilbert Schmidt Independence Criterion (HSIC) to retain relevant information from the data. The compression of information allows the neural network to focus on the essential aspects needed for accurate predictions, leading to more efficient training. Our proposed model is benchmarked against existing HSIC-only models across several datasets, including MNIST, CIFAR-10, and Fashion MNIST. Results demonstrate the superior performance of our model, in terms of computational efficiency and accuracy achieved. Additionally, we demonstrate its robustness and effectiveness on noisy data.**

I. Introduction

In recent years, Deep Learning has emerged as a prominent area of research due to its remarkable success as a machine learning technique, particularly in supervised learning tasks. Numerous upgraded versions of architectures have been developed, each striving to improve upon its predecessors in various aspects. Although training a model using backpropagation (BP) and stochastic gradient descent, or its variants, is effective, it proves to be highly computationally expensive, particularly when applied to deep networks. Additionally, challenges such as exploding gradients, vanishing gradients, and the need for update locking further compound the issue. Fig. 1 illustrates how training commences with BP which has two passes. In forward propagation, input data is processed sequentially through successive network layers to generate output. This output is subsequently employed to evaluate the loss function, which quantifies the divergence between the predicted outputs and the actual target values. In the subsequent backward propagation phase, the loss function's gradients with respect to each parameter are calculated using the chain rule. BP is the most conventional method used in supervised learning. However, the limitations of BP such as vanishing gradients, getting stuck in local minima especially when the network is deep, and also the memory overhead are the key problems.

A need to find a technique that eliminates the limitation of BP while maintaining performance is required. This involves focusing on information approximation, where only relevant information passes through the layers. A significant advancement in this direction was demonstrated by Naftali and Noga, highlighting the information-theoretic tradeoff between compression and relevance and demonstrating that the quantification of Deep Neural Networks (DNNs) can be achieved through the mutual information between the layers and the inputs and outputs.[1] Mutual information[2] is used to find the dependence between two random variables x and y, and can be expressed using Shannon's entropy as $I(x, y) = H(y) - H(y|x)$ where $H(y|x)$ is the conditional entropy. Considering x as the representation of the input and y being the label in a network, the mutual information offers a training objective that engages with x directly without involving BP through cross-entropy. Calculating mutual information is difficult [3], Ma et al [4] substituted mutual information with the Hilbert Schmidt Information Criterion (HSIC) for training the model without using backpropagation and cross-entropy loss. HSIC helps to mitigate the limitations of BP and maintain the performance of the network. It addresses the issues of exploding and vanishing gradients, thereby enabling the training of very deep networks without the need for skip connections. In the proposed model, HSIC is used to compression information during the training process. Only relevant features are passed as the training progresses through the layers. We introduce a model consisting of blocks. Each block consists of compression layer followed by a fully connected layer. Compression layers are trained using HSIC and linear layers using BP. Training done in this way allows a deep network to train on relevant features required to do the predictions.

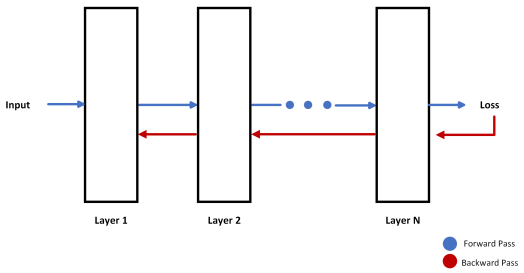


Fig. 1. Illustration of the forward pass and backward pass in a neural network.

Deep neural networks trained on large, well-annotated datasets have achieved remarkable results in image classification and other tasks. However, the collection of such

high-quality datasets can be both time-consuming and costly, leading to a growing interest in leveraging larger but noisier datasets that are more readily available. In this paper, we have evaluated our model for robustness and noise adaptation, showing how well the model maintains its performance under noisy conditions, demonstrating its resilience and reliability. In Section 2, we provide background information, detailing the techniques employed to design the architecture. Section 3 delves into the architecture's structure and operation, offering a comprehensive overview of its components and functionalities. Finally, in Section 4, we present the results obtained, providing a detailed analysis of its performance and effectiveness. The contributions of this research are as follows:

- This research introduces a novel approach that uses Hilbert-Schmidt Independence Criterion to compress the information which allows the neural network to focus on the essential aspects needed for accurate predictions, leading to more efficient training. The proposed architecture demonstrates better performance to existing models while reducing complexity overhead.
- The network trained using this approach exhibits improved resilience to noise. The model has been tested for robustness and noise adaptation, highlighting its effectiveness in maintaining performance under noisy conditions.

II. Background

This section elucidates the theoretical foundations underpinning the techniques employed in our architectural design. We begin by examining the concepts of relative entropy and mutual information, proceed to discuss the Hilbert-Schmidt Independence Criterion (HSIC).

A. Relative Entropy and Mutual Information

The entropy of a random variable is the uncertainty associated with it, representing the average amount of information needed to describe it, and the measure of distance between two distributions $p(x)$ and $p(y)$ is the relative entropy. Mutual information represents a specific instance within it. Mutual information serves as a comprehensive metric for gauging the relationship between two random variables. In a discrete domain, it quantifies the dependence between two random variables X and Y.

B. Hilbert Schmidt Independence Criterion

HSIC is a way to calculate mutual information between two distributions. Consider F, a Reproducing Kernel Hilbert Space (RKHS) of functions mapping from X to $\mathbb{R}$, where X is a separable metric space. For any two points x and x' in X, we can associate elements $\phi(x)$ and $\phi(x')$ in F (referred to as feature mapping) such that the associated reproducing F kernel $k(x, x') = \langle \phi(x), \phi(x') \rangle$.

In Reproducing Kernel Hilbert Space (RKHS), HSIC can be understood as a measure of the distance between the embeddings of the joint distribution P_{XY} and the product of the marginal distributions $P_X P_Y$, where X and Y are the random variables. When characteristic RKHSs are employed, the independence of the random variables can be inferred by checking if the value of HSIC is zero. Compared to other traditional independence measures, HSIC offers multiple advantages, including its ease of computation, quick convergence, and minimal estimation bias when working with limited data samples. HSIC-based learning methods revolve around the concept of leveraging HSIC to gauge the correlation or linkage between the various elements being analyzed, like the inputs and outputs. The objective is to optimize the solutions by either increasing or decreasing these associations. Let X and Y be the random variables, $k(\cdot, \cdot)$ and $l(\cdot, \cdot)$ the kernel functions, then

$$\begin{aligned} \mathrm{HSIC}(X, Y) = \; & \mathbb{E}_{(x,x',y,y')}[k(x, x')l(y, y')] \\ & + \mathbb{E}_{(x,x')}[k(x, x')] \cdot \mathbb{E}_{(y,y')}[l(y, y')] \\ & - 2\mathbb{E}_{(x,y)}\left[\mathbb{E}_{(x')}[k(x, x')] \cdot \mathbb{E}_y[l(y, y')]\right] \end{aligned} \tag{1}$$

Where $\mathbb{E}_{(x,x',y,y')}$ denote expectation over independent pairs (x, y) and x', y'. A kernel function transforms the input data into a feature space with a high dimension, where the statistical correlation between two random variables can be more conveniently assessed.

III. Methodology

In this section, we introduce the proposed architecture. As shown in Fig. 2, the proposed architecture consists of one or many blocks. Generally, each block consists of one or more compression layers followed by one or more linear layers. The number of blocks used in a network depends on the problem, just like in any other neural network; the complexity of the task and data, modelling goals like generalisation and overfitting decide on what and how many layers are needed.

A. Datasets

This section contains information about the datasets used for the experimentation. 1) The MNIST dataset is a widely used benchmark in the field of Deep learning, consisting of 70,000 grayscale images of handwritten digits from 0 to 9, each sized at 28x28 pixels. It is split into 60,000 training images and 10,000 testing images. 2) The MNIST Fashion dataset serves as a more challenging alternative to MNIST, containing 70,000 grayscale images of 10 different clothing items, also divided into a training set of 60,000 images and a test set of 10,000 images. 3) CIFAR-10 is another popular dataset comprising 60,000 32x32 color images in 10 classes, with 50,000 images for training and 10,000 for testing.

B. Proposed Architecture

In this research, we introduce a streamlined deep learning architecture that ensures layers are trained on the relevant information as it progresses through the network. As illustrated in Fig. 3, the model is composed of several blocks. In this experiment we created a model of two blocks. Each block contains one compression layer followed by one linear layer. The compression layers (also called as HSIC layers) are trained locally using HSIC without BP. The linear layers use ReLU

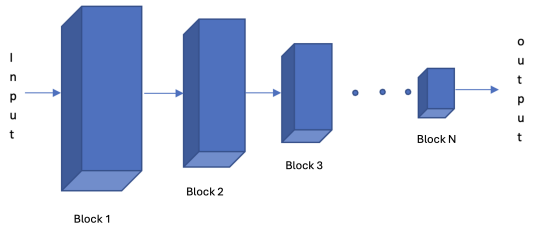


Fig. 2. Block diagram of the network which consists of one or more blocks.

as the activation function, with softmax applied to the output layer, and are trained using BP. The HSIC layers play a crucial role in compressing the data, filtering out only the most relevant information required for doing the prediction. This refined information is then passed to the next layer, allowing the linear layer to be trained more effectively with more important information.

The number of blocks used depends on the specific problem, similar to any neural network. Factors such as task complexity, data characteristics, and modeling goals like generalisation and avoiding overfitting determine the type and number of layers required in a block. It is better understood from Algorithm 1. We have tested the model on various datasets and have observed that the model performs better than the architecture trained using HSIC only found in the literature. Having a linear layer after every HSIC layer allows for capturing complex relationships in the data and potentially enhancing the model's ability to learn and discriminate between different patterns in the data.

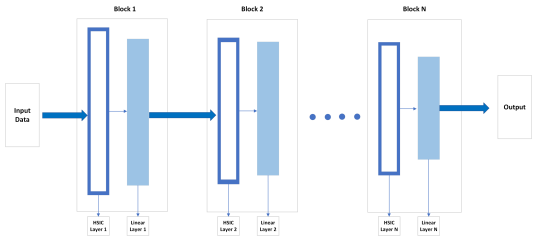


Fig. 3. Illustration of proposed network architecture.

Consider a neural network comprising L hidden layers, for the layers, denoted as $T_i(\cdot) : \mathbb{R}^{d_{i-1}} \rightarrow \mathbb{R}^{d_i}$, resulting in hidden representations $Z_i \in \mathbb{R}^{m \times d_i}$, where $i \in \{1, \ldots, L\}$, and m represents the batch size, then

$$Z_i = \arg\min_{Z_i} \left(nHSIC(Z_i, X) - \beta nHSIC(Z_i, Y)\right) \quad (2)$$

The input data $X \in \mathbb{R}^{m \times dx}$ and labels $Y \in \mathbb{R}^{m \times dy}$, where m denotes the batch size, i ranges from 0 to L, and L represents the total number of hidden layers. The dimensions dx and dy correspond to the input and output variables, respectively. The parameter β modulates the trade-off between the IB objectives that aims to compress input data while retaining as much relevant information as possible for predicting the output. nHSIC is given as

$$nHSIC(Z_i, X) = \text{tr}(K_e Z_i K_e X) \quad (3)$$

$$nHSIC(Z_i, Y) = \text{tr}(K_e Z_i K_e Y) \quad (4)$$

The above equations suggest that the optimal hidden representation Z_i achieves a delicate equilibrium between independence from irrelevant input details and association with the output. K_e is the centered kernel matrix derived from the original kernel matrix by removing the mean effects. The convergence of equation (3) ideally preserves the essential information required for label prediction while eliminating extraneous details that might lead to overfitting. The features derived from the HSIC layers serve as input to the BP-led layers. As the data traverses through the network comprising L layers, the final output is produced.

Algorithm 1 Training an MLP with HSIC and Backpropagation

Input: X (input data), Y (labels), n (number of samples), m (batch size), L (number of layers), α (learning rate), β (HSIC regularization parameter), T_{i-1} and F_i are parametrized by $\{\theta_i \mid W_i, b_i\}$ and $\{\phi_i \mid W_i, b_i\}$ respectively.
Initialize: θ_i (HSIC layer parameters), ϕ_i (BP layer parameters) for $i \in \{1, \ldots, L\}$
for $j \in \{1, \ldots, n/m\}$ **do**
  $Z_0 \leftarrow X_j$
  for $i \in \{1, \ldots, L\}$ **do**
    if $i \in$ HSIC Layers **then**
      $Z_i \leftarrow T_{i-1}(Z_{i-1})$
      $g_i \leftarrow \nabla_{\theta_i}(nHSIC(Z_i, X_j) - \beta \times nHSIC(Z_i, Y_j))$
      $\theta_i \leftarrow \theta_i - \alpha g_i$
    else
      $V_i \leftarrow F_i(Z_{i-1})$
      $h_i \leftarrow \nabla_{\phi_i}(\text{Loss}(V_i, Y_j))$
      $\phi_i \leftarrow \phi_i - \alpha h_i$
      $Z_i \leftarrow V_i$
    end if
  end for
end for

IV. Results

The performance of proposed model was assessed across several datasets, including MNIST, Fashion MNIST, and CIFAR-10. We evaluated and compared the results with those reported in [4] as shown in Fig. 4. The observed accuracies from the proposed model highlight its effectiveness when compared to models where layers are trained exclusively using HSIC. The testing results demonstrate that the proposed architecture outperforms the HSIC-only model, suggesting that incorporating linear layers between HSIC layers enhances the model's ability to learn and capture essential features for

accurate prediction. This indicates that the integration of linear layers contributes to a more robust feature extraction process, ultimately leading to improved performance.

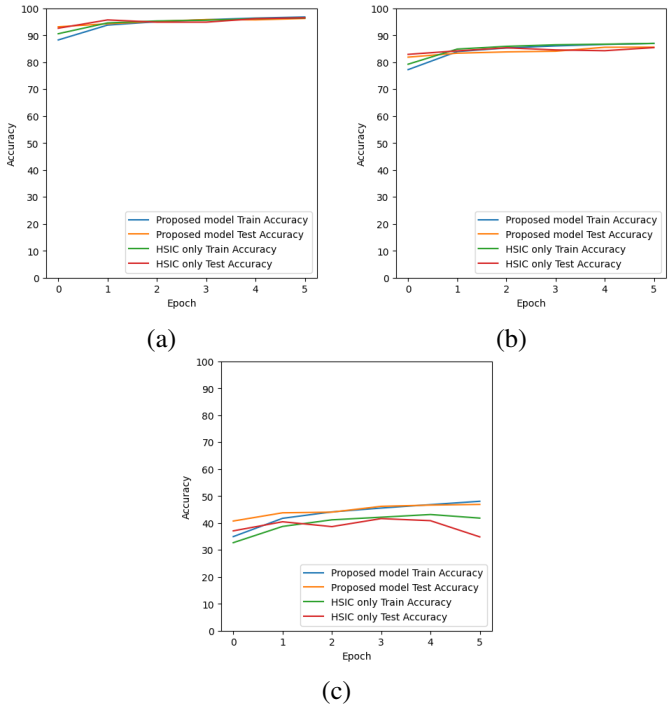


Fig. 4. Performance comparison between the HSIC-trained architecture and our proposed method: (a) Results on MNIST, (b) Results on FashionMNIST, (c) Results on CIFAR-10.

In addition to achieving high accuracy, computational efficiency is paramount for any model. We conducted a comparative analysis of architectures based on several key metrics number of layers, and nodes per layer, and evaluated the efficiency based on execution time and memory usage as detailed in Table I. The proposed model is structured as 784-256-256-128-10. This configuration is more compact compared to the 784-256-256-256-256-256-10 architecture described in [4]. Despite using only 3 hidden layers with fewer nodes per layer, our model achieves better accuracy, indicating its better efficiency. We also computed the difference in FLOPs (Floating Point Operations) between the two models. The proposed model attained 601216 estimated flops and HSIC-only model attained 930816, indicating that the proposed model is computationally more efficient than the HSIC-only model.

TABLE I
COMPARATIVE ANALYSIS OF TWO ARCHITECTURES BASED ON ACCURACY (%), EXECUTION TIME (SEC) AND MEMORY USED (MB).

Method	Hidden Layers	Memory	Time	Accuracy
HSIC Only	5	1.79	130.16	93.95
Proposed model	3	1.16	104.11	94.64

A. *Experimental results with Noise*

The presence of noise is a well-known issue that significantly impacts data quality, particularly in classification tasks [5], [6]. In various application domains such as medical image processing [7], speech recognition [8], and stock market analysis [9], classification accuracy is a crucial metric. Deep neural networks are usually trained through supervised learning with large, meticulously curated datasets that are free from noise.

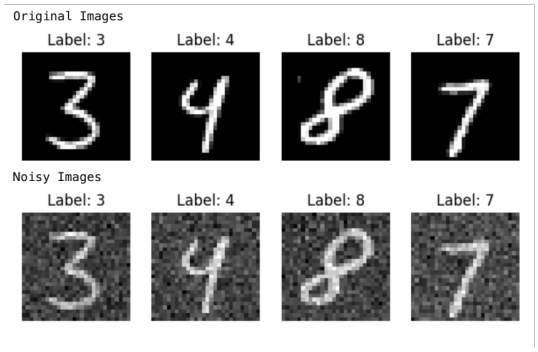


Fig. 5. Images before and after the Gaussian noise is added to it (standard deviation = 0.05).

However, the requirement for such datasets limits the range of problems that can be tackled because in real life, getting data that is clean is not most of the time true. This has resulted in a surge of deep learning studies focused on the same tasks using these familiar datasets. We utilized the MNIST dataset and introduced Gaussian noise to the data. Fig. 5 shows the image before and after the noise is added to the images.

This approach allowed us to assess the performance of our model under conditions where data is not perfectly clean, providing insights into its robustness and generalisation. By simulating real-world scenarios where noise may be present, we aimed to better understand the model's effectiveness in practical applications.

The results indicate that our model exhibits greater robustness. This enhanced resilience suggests that the model is capable of maintaining high performance even when confronted with noisy or imperfect data, which is critical for its deployment in real-world environments where data quality can vary. Table II indicates that there is minimal difference in the model's accuracy when trained and tested on noisy data. This demonstrates the robustness of our model, suggesting that it maintains performance even under challenging conditions.

TABLE II
PERFORMANCE COMPARISON OF THE PROPOSED METHOD ON DIFFERENT NOISY DATA FOR BOTH TRAINING AND TESTING.

	Std = 0.05	Std = 0.5	Std = 0.8
Train Accuracy (%)	96.30	93.98	91.31
Test Accuracy (%)	95.97	93.22	91.28

We also evaluated our model by testing it with noisy test data while using clean data for training. The results revealed

minimal differences in accuracy, indicating that the model remains resilient and performs well even when faced with noisy data during testing as shown in Table III. Hence, checking the model for both: 1) Evaluating noise adaptation: Train on noisy data and test on noisy data. 2) Robustness testing: Train on clean data and test on noisy data.

TABLE III
PERFORMANCE EVALUATION OF PROPOSED METHOD WHEN ONLY TEST DATA IS NOISY DATA.

	Std = 0.5
Train Accuracy	96.21
Test Accuracy	92.61

V. CONCLUSION

In this research, we propose a simple deep learning architecture that ensures layers are trained on the relevant information as it progresses through the network. Hilbert Schmidt Information Criterion is used to compress the information between the layers which helps linear layers trained on only relevant information. Our approach addresses several limitations of traditional BP training, such as the challenges posed by vanishing and exploding gradients, and the substantial memory overhead associated with deep networks.

The experimental results across various datasets, including MNIST, CIFAR-10, and Fashion MNIST, demonstrate that our model outperforms the HSIC-only approach. This indicates that the addition of linear layers enhances the network's ability to learn and retain crucial features, contributing to more accurate predictions. Furthermore, our architecture achieves comparable accuracy to existing models but with a more compact and memory-efficient design, highlighting its advantages in resource-constrained environments. Future research could delve deeper into the architecture and explore the impact of utilising multiple HSIC or fully connected layers within a block, as it may offer additional improvements to the model's performance.

REFERENCES

[1] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015.

[2] P. Czyż, F. Grabowski, J. Vogt, N. Beerenwinkel, and A. Marx, "Beyond normal: On the evaluation of mutual information estimators," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *CoRR*, vol. abs/1612.00410, 2016.

[4] W.-D. K. Ma, J. Lewis, and W. B. Kleijn, "The hsic bottleneck: Deep learning without back-propagation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 5085–5092, 2020.

[5] H. Khan, H. Liu, and C. Liu, "Missing label imputation through inception-based semi-supervised ensemble learning," *Advances in Computational Intelligence*, vol. 2, no. 1, p. 10, 2022.

[6] Z. Zhu, T. Liu, and Y. Liu, "A second-order approach to learning with instance-dependent label noise," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10113–10123, 2021.

[7] O. Albahri, A. Zaidan, A. Albahri, B. Zaidan, K. H. Abdulkareem, Z. Al-Qaysi, A. Alamoodi, A. Aleesa, M. Chyad, R. Alesa, *et al.*, "Systematic review of artificial intelligence techniques in the detection and classification of covid-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects," *Journal of infection and public health*, vol. 13, no. 10, pp. 1381–1396, 2020.

[8] T. Nakatani, "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," in *proc. INTERSPEECH*, vol. 2019, pp. 1408–1412, 2019.

[9] M. Nabipour, P. Nayyeri, H. Jabani, S. Shahab, and A. Mosavi, "Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis," *Ieee Access*, vol. 8, pp. 150199–150212, 2020.

Zero-Shot Learning for Haze Removal Using Fusion of Near-Infrared and Color Images

Onhi KATO* and Akira KUBOTA†
* Chuo University, Tokyo, Japan
E-mail: a17.ffca@g.chuo-u.ac.jp
† Chuo University, Tokyo, Japan
E-mail: kubota@elect.chuo-u.ac.jp

***Abstract*—Outdoor images are often degraded by haze. In this study, we propose a new method to remove haze by fusing near-infrared (NIR) images and color images using zero-shot learning. Unlike most learning-based methods that require a large amount of training data, our method employs zero-shot learning, which learns using the input image only. The proposed method removes haze from both NIR and color images. In edge fusion, a weight map is generated and added to the NIR image, which is then multiplied by the lightness image of the haze-removed color image. This process reduces the differences in characteristics between the NIR and lightness images, achieving a natural fusion of edges. Our qualitative and quantitative evaluations using 59 pairs of NIR and color images demonstrated that our method can restore color and edges more naturally than conventional methods.**

I. Introduction

Images captured outdoors are often degraded by atmospheric particles such as fog and haze. These particles cause loss of colors and edges, especially in distant regions. This image quality degradation is a problem in various computer vision tasks, such as depth estimation and object detection.

Haze removal methods are classified into those using a single captured color image [1]–[8] and those fusing a near-infrared (NIR) image and a color image [9]–[16]. Most single-image methods are based on atmospheric scattering models [1], [2] and effectively restore color. In contrast, fusion-based methods restore edges better because NIR images are less affected by scattering and contain more apparent edges than color images. However, simply fusing NIR and color images may result in unnatural color shifts in images due to their different characteristics.

The fusion-based methods solve this by manipulating lightness images [9], using optimization methods [10], and using sparse approximation methods [11]. Learning-based methods have also been proposed recently [12]–[16]. However, these methods focus on fusing infrared images [12] or denoising [14]–[16] and do not specialize in haze removal. Therefore, they often produce images with overly bright colors, especially in plant areas. This is because NIR images have different characteristics from color images. In addition, many learning-based methods require and depend on a large amount of training data.

Unlike these learning-based methods, zero-shot learning can learn using the input image only and generate an output image without a large amount of training data. Therefore, haze removal methods using zero-shot learning have been proposed [3]–[8]. These are single-image methods. A fusion-based method has also been proposed [17]. It fuses NIR and color images by colorizing the NIR image. However, to our knowledge, the fusion-based methods specific to haze removal have not been sufficiently studied.

This paper proposes a new method for haze removal by fusing color images with NIR images using zero-shot learning. The proposed method has three main features. First, it uses zero-shot learning, which does not require a large amount of training data. Second, it removes haze on both NIR and color images, enhancing edge restoration. Third, it introduces a novel approach for fusing NIR and lightness images. In this fusion process, a weight map is appropriately estimated using zero-shot learning, added to the NIR image, and then multiplied by the lightness image. This enables more natural color and edge restoration.

II. Related work

A. Observed Haze Model

A commonly used observation model of haze images is based on the atmospheric scattering model [1], [2]. In this model, the observed haze image I_c is modeled by

$$I_c(\mathbf{x}) = t(\mathbf{x})J_c(\mathbf{x}) + (1 - t(\mathbf{x}))A_c, \quad c \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}, \quad (1)$$

where J_c denotes the haze-removed color image, t the transmission map, A_c the atmospheric light value, c the channel index, and $\mathbf{x}$ the image coordinate. By estimating the transmission map and atmospheric light from the haze image (denoted $\hat{t}(\mathbf{x})$ and $\hat{A}_c$, respectively) and inversely solving this model, we can obtain the haze-removed image $\hat{J}_c$:

$$\hat{J}_c(\mathbf{x}) = \frac{I_c(\mathbf{x}) - \hat{A}_c}{\hat{t}(\mathbf{x})} + \hat{A}_c. \quad (2)$$

B. Zero-Restore

Zero-shot learning generates output images directly from input images without a large amount of training data. This method is effective when training data is difficult to collect. In haze removal, zero-shot learning is useful because capturing hazy and non-hazy images is difficult.

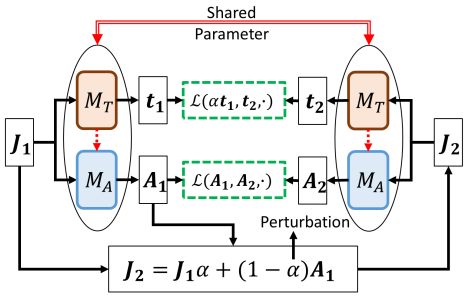


Fig. 1: Overview of Zero-Restore [6]. Note that J_1 represents the input haze image and J_2 represents the haze-enhanced image.

Zero-Restore [6] uses zero-shot learning to remove haze from captured color images. Fig. 1 shows an overview of this method. First, the transmission map t_1 and atmospheric light A_1 are obtained from the input haze image J_1 using modules M_T and M_A. Then, the haze-enhanced image J_2 is generated using the α-multiplied transmission map t_1 and atmospheric light A_1. The transmission map t_2 and atmospheric light A_2 are estimated again from the haze-enhanced image using the same modules. The module learns to minimize the difference between the α-multiplied transmission map t_1, αt_1, and the re-estimated transmission map t_2, and the difference between A_1 and A_2 of the atmospheric light.

In the final inference, the transmission map and atmospheric light are obtained from the input haze image and substituted into (2) to obtain the haze-removed image.

III. Presented Haze Removal Method

This paper presents an image fusion method for haze removal using zero-shot learning. The proposed method consists of two main processing steps: haze removal step and edge fusion step, as shown in Fig. 2. Here, I_N and I_{RGB} represent the input NIR and color images, respectively. Examples are shown in Fig. 3(a), (b). $\hat{J}_N$ and $\hat{J}_{RGB}$ represent the haze-removed NIR and color images. Examples are shown in Fig. 3(c), (d). $\hat{F}_{RGB}$ is the edge-fused color image. An example is shown in Fig. 3(e).

A. Haze removal step

The haze removal step consists of two key steps: obtaining the transmission map and atmospheric light from the input NIR and color images and generating the haze-removed image from

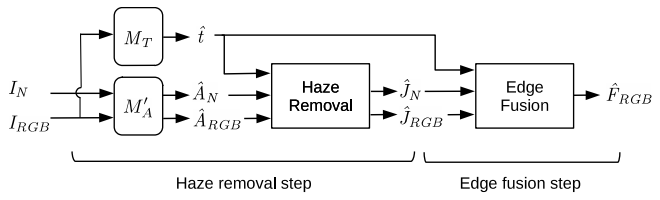
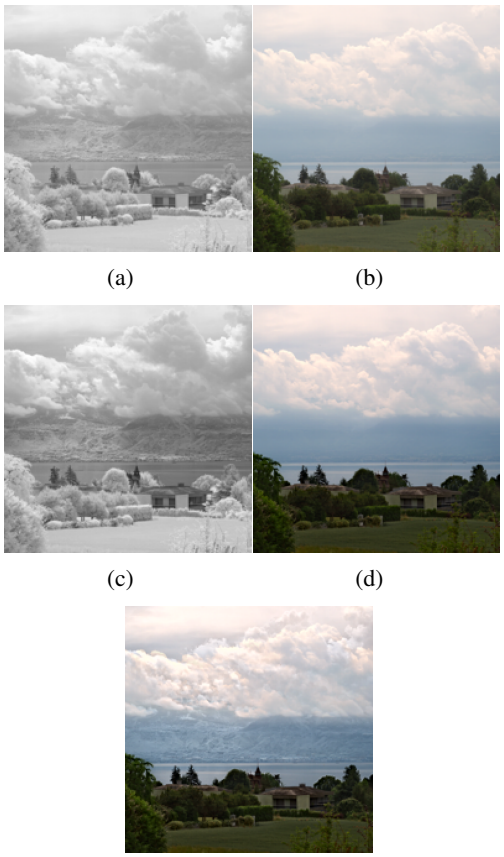


Fig. 2: Block diagram of the proposed haze removal method



(a) (b) (c) (d) (e)

Fig. 3: Examples of input image pair, haze-removed image pair, and edge-fused image. (a) Input NIR image I_N, (b) Input color image I_{RGB}, (c) haze-removed NIR image $\hat{J}_N$, (d) haze-removed color image $\hat{J}_{RGB}$, (e) Edge-fused color image $\hat{F}_{RGB}$.

these. In this process, zero-shot learning is used to effectively remove haze in both NIR and color images, restoring colors and enhancing edges.

First, the transmission map $\hat{t}$, the atmospheric light $\hat{A}_N$ (for the NIR image) and $\hat{A}_{RGB}$ (for the color image) are obtained from the input NIR image I_N and color image I_{RGB} using modules M_T and M'_A. These modules are designed and trained based on the conventional method of Zero-Restore [6]. The module M_T is the same as in Zero-Restore, while M'_A was modified from M_A by changing the output of the final linear layer from 3 to 4 channels to estimate both NIR and RGB atmospheric light.

Second, the obtained transmission map $\hat{t}$, the atmospheric light $\hat{A}_N$ and $\hat{A}_{RGB}$ are substituted into (2). This equation was originally designed for color images, but we also apply it to NIR images in the same way. This allows us to obtain the haze-removed NIR image $\hat{J}_N$ (Fig. 3(c)).

Unlike conventional methods, our method removes haze from both NIR and color images. Removing the haze in the NIR image preserves and enhances more edges, allowing for the fusion of more detailed edges in the edge fusion step.

B. Edge fusion step

The block diagram of edge fusion is shown in Fig. 4. First, the haze-removed color image $\hat{J}_{RGB}$ is converted into two chrominance components a^* and b^*, and a lightness image L^* (Fig. 5(a)). Next, the lightness image L^* and the haze-removed NIR image $\hat{J}_N$ are fused using the module M_L. This produces the edge-fused lightness image $\hat{L}^*$ (Fig.5(b)). Finally, the image is converted into the RGB color image. The edge-fused lightness image $\hat{L}^*$ and the original color components a^* and b^* are used for this conversion. This results in the final edge-fused color image $\hat{F}_{RGB}$ with color restored.

We explain the module M_L used for edge fusion. The flow of generating the edge-fused lightness image $\hat{L}^*$ in module M_L is shown in Fig. 6. In this module M_L, a weight map $\hat{w}$ is generated using zero-shot learning, with input from the haze-removed NIR image $\hat{J}_N$, the transmission map $\hat{t}$, and the lightness image L^*. The generated weight map is added to the haze-removed NIR image $\hat{J}_N$ and multiplied by the lightness image L^* to obtain the edge-fusion lightness image $\hat{L}^*$.

Image generation using zero-shot learning has its limitations in generating details but is good at generating the overall structure. Therefore, even if a rough weight map generated by zero-shot learning is added to the NIR image, the details of the NIR image will be preserved. NIR images have different reflection characteristics from color images. To reduce the effect of this difference during fusion, a weighting map is added to the NIR image. This is then multiplied by the lightness image. This allows for the suppression of changes in a lightness image and the fusion of natural edges.

C. Loss functions

This section describes the loss function used to train the module M_L. The loss function is designed to penalize changes in lightness by the amount of haze, thus suppressing changes in lightness in the edge-fused lightness image $\hat{L}^*$.

Specifically, the larger the transmission map t value, the less the haze effect. Therefore, the edge-fused lightness image $\hat{L}^*$ and the lightness image L^* are penalized for large changes in lightness. This makes the edge-fused lightness image $\hat{L}^*$ closer to the lightness of the lightness image L^*. The penalty is set in the regions where the influence of haze is small, as follows:

$$\mathcal{L}_{PL} = \mathrm{MSE}(t(\mathbf{x})\hat{L}^*(\mathbf{x}), t(\mathbf{x})L^*(\mathbf{x})), \tag{3}$$

where MSE is the mean squared error.

Conversely, the smaller the transmission map t value, the greater the haze effect. Therefore, changes in lightness between

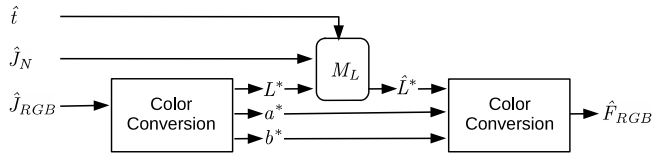


Fig. 4: Block diagram of edge fusion.



(a) (b)

Fig. 5: Examples of the edge fusion process in module M_L. (a) Lightness image L^*, (b) Edge-fused lightness image $\hat{L}^*$.

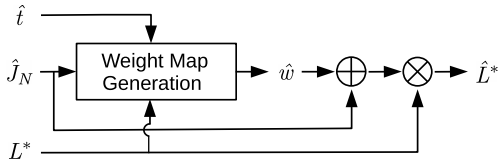


Fig. 6: Block diagram of module M_L.

the edge-fused lightness image $\hat{L}^*$ and the haze-removed NIR image $\hat{J}_N$ are more heavily penalized. This makes the edge-fused lightness image $\hat{L}^*$ closer to the lightness of the haze-removed NIR image $\hat{J}_N$. The penalty is set in the regions where the influence of haze is large, as follows:

$$\mathcal{L}_{PN} = \mathrm{MSE}((1 - t(\mathbf{x}))\hat{L}^*(\mathbf{x}), (1 - t(\mathbf{x}))\hat{J}_N(\mathbf{x})). \tag{4}$$

The final loss function $\mathcal{L}$ is defined as the sum of these penalties:

$$\mathcal{L} = \mathcal{L}_{PL} + \mathcal{L}_{PN}. \tag{5}$$

IV. Experimental results and discussions

To evaluate the effectiveness of the proposed method, we conducted qualitative and quantitative comparison experiments with three conventional methods. The first method is Zero-Restore which removes haze using zero-shot learning[6]. Although it does not fuse NIR images, it was used for color restoration and included in the comparison. The second method is ConvSSA which fuses NIR images using sparse approximation [11]. It successfully fuses NIR images without deep learning, so it was included. The third method is SwinFusion which uses deep learning with a Transformer-based architecture to fuse NIR images [13]. It was included for its effectiveness in fusing NIR images using deep learning. In addition, we also conducted experiments to evaluate the effectiveness of the edge fusion in the proposed method.

The images used in the experiments were 59 pairs of NIR and color images from the RGB-NIR Scene Dataset [18] with relatively weak haze. We defined here the relatively weak haze images as the images in which the edges of distant objects are barely visible in the color images but clearer in the NIR images. We cropped the 512 $\times$ 512 pixels region in the center of the original images.

We used the Adam optimizer with a learning rate of 0.001. We trained our model in 500 epochs at maximum with early stopping applied if the loss function did not decrease for 40 epochs. In the haze removal step, we used the loss function used in Zero-Restore [6]. In the edge fusion step, we used the loss function in (5).

A. Qualitative evaluation

Fig. 7 compares the haze-removed images generated by the three conventional methods and our method. Zero-Restore [6] removes haze from color images only and thus can restore color, but it can hardly restore edges lost in the original color image. ConvSSA, SwinFusion, and our method could restore edges because they fuse the NIR images. However, ConvSSA and SwinFusion could not restore color sufficiently, and the SwinFusion caused unnatural color shifts in the generated images. In contrast, our method succeeded in restoring both color and edges naturally.

B. Quantitative evaluation using image quality assessment metrics

For quantitative evaluation, we used three image quality metrics: two reference image quality metrics and one no-reference image quality metric. The rate of edges newly visible (denoted as e) and the contrast restoration (denoted as r) were calculated using Hautiere et al.'s blind measures [19], which evaluate the dehazed image by referencing the input hazy image. The no-reference image quality scores (denoted as NIMA) were calculated using Neural Image Assessment [20], which uses convolutional neural networks to assess the image quality, including its naturalness, and predicts a subjective human score. The larger e, r and NIMA are, the better the quality of the image.

Table I shows the averaged values of the three metrics over 59 generated images. We can see that our method achieves the highest value of e. In ConvSSA [11] and SwinFusion [13], the e value is low, although the edges of the NIR images are fused. This is because of differences in the characteristics of NIR and color images. These differences make the fused images overly bright and reduce the edges. This result demonstrates the effectiveness of our method in fusing NIR images.

The value of r is the largest for ConvSSA [11] and the next largest for our method. Our method is slightly inferior to ConvSSA, but improves the contrast high enough.

The results also show that our method achieved the highest value of NIMA, suggesting that our method can generate more visually natural and higher quality images than the conventional methods.

C. Evaluation of the proposed method

We conducted a comparative experiment of four cases: before edge fusion, edge fusion without adding a weight map, edge fusion without removing the haze from NIR, and our method. Examples are shown in Fig. 8. Quantitative evaluation results using 59 images are shown in Table II.

First, we evaluate the effectiveness of adding a weight map. Compared to the image before edge fusion (Fig. 8(a)), the image without a weight map (Fig. 8(b)) is darker overall. In contrast, our method (Fig. 8(d)) has a natural lightness. The quantitative results in Table II show that without a weight map, the value of e is the highest, but r and NIMA values are lower than before edge fusion. This indicates more edges but degraded overall image quality. However, our method improves all e, r, and NIMA values, enhancing edges, contrast, and overall quality. These results demonstrate the effectiveness of adding a weight map.

Next, we evaluate the effectiveness of NIR haze removal. Compared to the image without NIR haze removal (Fig. 8(c)), our method (Fig. 8(d)) can see more edges. The quantitative results in Table II show that without NIR haze removal, the NIMA value is the highest, but e and r values are lower than those of our method. This indicates that our method is slightly inferior to NIR haze removal in naturalness, but that it shows more edges. These results show the effectiveness of haze removal for NIR images.

V. Conclusions

This paper proposes a new method to remove haze by fusing NIR and color images using zero-shot learning. The proposed method learns using the input image only and removes the haze from both NIR and color images. In addition, zero-shot learning is good at capturing the overall structure. Using this feature, we generated a weight map, added it to the NIR image, and multiplied it by the lightness image of the haze-removed color image to fuse the edges. This process of adding the weight map reduced the differences in characteristics between the NIR image and the lightness image and achieved a natural fusion of edges. Qualitative and quantitative evaluation results show that the proposed method restores color and edges more naturally than conventional methods.

References

[1] S. G. Narasimhan and S. K. Nayar, "Removing weather effects from monochrome images," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, IEEE, vol. 2, 2001, pp. II–II.

[2] R. T. Tan, "Visibility in bad weather from a single image," in *2008 IEEE conference on computer vision and pattern recognition*, IEEE, 2008, pp. 1–8.

[3] B. Li, Y. Gou, J. Z. Liu, H. Zhu, J. T. Zhou, and X. Peng, "Zero-shot image dehazing," *IEEE Transactions on Image Processing*, vol. 29, pp. 8457–8466, 2020.

[4] J. Wei, Y. Cao, K. Yang, L. Chen, and Y. Wu, "Self-supervised remote sensing image dehazing network based on zero-shot learning," *Remote Sensing*, vol. 15, no. 11, p. 2732, 2023.

[5] J. Wei, Y. Wu, L. Chen, K. Yang, and R. Lian, "Zero-shot remote sensing image dehazing based on a re-degradation haze imaging model," *Remote Sensing*, vol. 14, no. 22, p. 5737, 2022.



(a) NIR (b) Color (c) Zero-Restore [6] (d) ConvSSA [11] (e) SwinFusion [13] (f) Ours

Fig. 7: Comparison of the haze removed images between the conventional methods and ours. The images in the even-numbered rows are magnified for better comparison.

TABLE I: Quantitative evaluation using the average of three image quality metrics for 59 generated images.

	Input	Zero-restore [6]	convSSA [11]	Swinfusion [13]	Ours
$e \uparrow$		0.2106	0.1676	-0.0026	0.3697
$r \uparrow$		1.2641	1.9913	1.3350	1.5318
NIMA $\uparrow$	7.1103	7.4356	7.4304	6.9591	7.4713

Red text indicates the best, blue text indicates the second best.



(a) Before edge fusion (b) Without adding weight maps (c) Without NIR haze removal (d) Ours

Fig. 8: Comparison of Four Cases: Before Edge Fusion, Edge Fusion Without Adding Weight Map, Edge Fusion Without Removing Haze from NIR images, and Our Method.

TABLE II: Quantitative evaluation of the effectiveness of our method using average scores from three image quality metrics on 59 images.

	Before edge restoration	Without adding weight maps	Without NIR haze removal	Ours
$e \uparrow$	0.2373	0.5318	0.3172	0.3697
$r \uparrow$	1.2666	0.9915	1.4660	1.5318
NIMA $\uparrow$	7.4157	7.1955	7.4743	7.4713

Red text indicates the best, blue text indicates the second best.

[6] A. Kar, S. K. Dhara, D. Sen, and P. K. Biswas, "Zero-shot single image restoration through controlled perturbation of koschmieder's model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 205–16 215.

[7] Y. Gandelsman, A. Shocher, and M. Irani, ""double-dip": Unsupervised image decomposition via coupled deep-image-priors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 026–11 035.

[8] B. Li, Y. Gou, S. Gu, J. Z. Liu, J. T. Zhou, and X. Peng, "You only look yourself: Unsupervised and untrained single image dehazing neural network," *International Journal of Computer Vision*, vol. 129, pp. 1754–1767, 2021.

[9] L. Schaul, C. Fredembach, and S. Süsstrunk, "Color image dehazing using the near-infrared," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2009, pp. 1629–1632.

[10] C. Feng, S. Zhuo, X. Zhang, L. Shen, and S. Süsstrunk, "Near-infrared guided color image dehazing," in *2013 IEEE international conference on image processing*, IEEE, 2013, pp. 2363–2367.

[11] F. G. Veshki and S. A. Vorobyov, "Convolutional simultaneous sparse approximation with applications to rgb-nir image fusion," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, IEEE, 2022, pp. 872–876.

[12] X. Zhang and Y. Demiris, "Visible and infrared image fusion using deep learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[13] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[14] Q. Han and C. Jung, "Deep selective fusion of visible and near-infrared images using unsupervised u-net," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[15] L. Mei and C. Jung, "Deep fusion of rgb and nir paired images using convolutional neural networks," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 6802–6803.

[16] C. Jung, K. Zhou, and J. Feng, "Fusionnet: Multispectral fusion of rgb and nir images using two stage convolutional neural networks," *IEEE Access*, vol. 8, pp. 23 912–23 919, 2020.

[17] C. Wei, H. Chen, L. Bai, J. Han, and X. Chen, "Infrared colorization with cross-modality zero-shot learning," *Neurocomputing*, p. 127 449, 2024.

[18] M. Brown and S. Süsstrunk, "Multi-spectral sift for scene category recognition," in *CVPR 2011*, IEEE, 2011, pp. 177–184.

[19] N. Hautiere, J.-P. Tarel, D. Aubert, and E. Dumont, "Blind contrast enhancement assessment by gradient ratioing at visible edges," *Image Analysis and Stereology*, vol. 27, no. 2, pp. 87–95, 2008.

[20] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

Color Enhancement for the Colorblind Using Color Correction Intensity Map and Pix2pix Image Conversion

Shu KOMATSU* and Akira KUBOTA†
* Chuo University, Tokyo, Japan
E-mail: a19.tncg@g.chuo-u.ac.jp
† Chuo University, Tokyo, Japan
E-mail: kubota@elect.chuo-u.ac.jp

***Abstract*—In this paper, we propose a method for creating easier-to-see images for colorblind individuals by estimating regions where color discrimination is difficult for people with color blindness and performing color correction. There are various prior studies on color correction of images for the colorblind. However, in many cases, the goal is only to identify text or objects, and thus the hue of the image is often unnatural compared to the original image. To solve this problem, we estimated the color discrimination gap between colorblind and normal subjects based on confusion loci theory, and performed color correction using the pix2pix model. The aim is to prevent unnecessary hue changes by estimating the correction intensity, and to achieve highly accurate color correction by deep learning. From the validation using a colorblindness simulator, we obtained the result that color discrimination becomes easier for colorblind people while maintaining the hue close to that of the original image. This method allows easy color correction of existing images for the colorblind without any loss of hue.**

I. Introduction

Photoreceptor cells in the human eye are classified into three types according to their sensitivity to light wavelengths[1], [2]. S-cone cells respond to blue, M-cone cells to red, and L-cone cells to green. Color perception is based on the relative excitation levels of these three types of cone cells, which allows us to distinguish a wide range of colors. When one of the photoreceptor cells is damaged or loses its function, either by birth or later in life, we receive only a part of the spectral information of light, limiting the colors we can perceive. 8% of white men, 4% of black men, and 5% of yellow men have color vision characteristics that make it difficult to perceive a specific range of colors. As an example, a person who lacks L-cone function is called Protanope and cannot distinguish between red and green. Those lacking M-cones are called Duteranope, and those lacking S-cones are called Tritanope. These color-vision characteristics may cause inconveniences such as difficulty in reading letters and signs in certain color combinations. Therefore, colorblind people need assistance to distinguish colors more effectively.

A solution to this problem is to adjust the color contrast using aids such as color vision aids such as spectacles, but there are problems such as the time of professional testing and physical constraints Another method is color correction of images by image processing. This method requires no inspection and has a wide range of applications. Huang[3] proposed a color correction method based on mapping by the relationship between color blindness images and normal images. Dody[4] proposed a color correction algorithm for color deficient images using RGB color clusters and graph coloring.

Most color correction algorithms are only intended to help colorblind people distinguish colors. Therefore, the corrected images are unnatural for normal people, for example, a red flower becomes blue. In addition, there is a problem that color correction is applied to the background color, which changes the saliency of the image. To solve this problem, we propose an estimation of correction intensity based on confusion loci theory and color correction by pix2pix. The present study will focus Duteranope, which has the highest percentage of color blindness. In our method, the color components of an image that are difficult for colorblind people to distinguish are estimated for each pixel, and the intensity of color correction is estimated. Next, a color-corrected image is generated using a pix2pix model that has been trained to learn color corrections that are easy for colorblind people to distinguish. Then, the original image and the corrected image by pi2xpix are blended using the color correction intensities. By using the color correction intensity, saliency is preserved and unnecessary coloring is prevented. Pix2pix model allows natural color correction by referring to the color scheme of the original image rather than a random color scheme.

II. related works

A. Simulation of color blindness(Duteranope)

The spectral sensitivity values of human L, S, M-cones can be obtained from the CIE-1931XYZ, as follows:

$$\begin{pmatrix} L \\ S \\ M \end{pmatrix} = \begin{pmatrix} 0.40024 & 0.70760 & -0.08081 \\ -0.22630 & 1.16532 & 0.04570 \\ 0 & 0 & 0.91822 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \quad (1)$$

From Brettel's formula[5], assuming LSM cone response values as L_d, M_d, S_d, the color vision of Duteranope is expressed

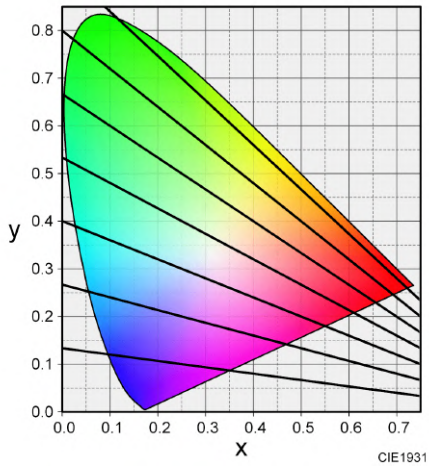


Fig. 1: Confusion loci in xy chromaticity diagram (Duteranope vision)

as follows

$$\begin{cases} \begin{pmatrix} L_d \\ S_d \\ M_d \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0.82781 & 0 & 0.17216 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} L \\ S \\ M \end{pmatrix} & \text{if } S \leq L \\ \begin{pmatrix} L_d \\ S_d \\ M_d \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0.81951 & 0 & 0.18046 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} L \\ S \\ M \end{pmatrix} & \text{if } S \geq L \end{cases} \tag{2}$$

By using the inverse matrix of equation (3) for L_d, S_d, M_d, we can obtain the pixel values X_d, Y_d, Z_d, Duteranope vision in XYZ.

$$\begin{pmatrix} X_d \\ Y_d \\ Z_d \end{pmatrix} = \begin{pmatrix} 1.85995 & -1.12939 & 0.21990 \\ 0.36119 & 0.63881 & 0 \\ 0 & 0 & 1.08906 \end{pmatrix} \begin{pmatrix} L_d \\ S_d \\ M_d \end{pmatrix} \tag{3}$$

By converting X_d, Y_d, Z_d to RGB, it is possible to create a simulated image of color blindness.

B. Confusion Loci Theory

Colorblind people have coordinates called confusion centers at different locations on the xy chromaticity diagram, depending on the type of color vision characteristic. The coordinates of the confusion center for Duteranope obtained by Judd[6] are $x_d = 1.000, y_d = 0.000$. From the confusion center, multiple lines (confusion loci) can be drawn on the chromaticity diagram. Colors that fall along the same confusion locus are indistinguishable for individuals with that specific type of color blindness. It is known that colors located on the same confusion loci on the xy chromaticity diagram are difficult to distinguish for Duteranope. Therefore, it is possible to estimate color combinations that are difficult for colorblind people to discriminate by examining colors that belong to the same confusion loci.

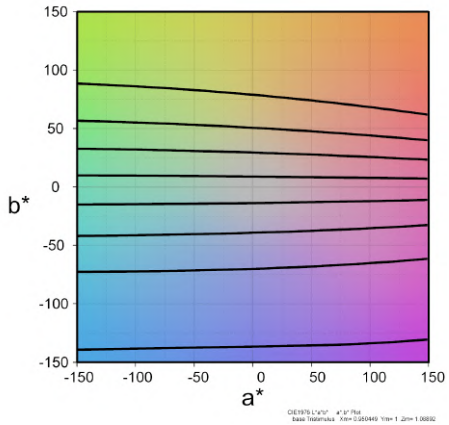


Fig. 2: Confusion loci in a^*b^* chromaticity diagram (Duteranope vision)

III. Mathod

A. Color Correction Intensity Map(CCImap)

We estimate the regions in the input images where the colorblind person has difficulty in discriminating colors compared to the normal person. The estimation is done using a* and b* components in $L^*a^*b^*$ color space. The geometric distance in the a^*b^* chromaticity diagram is proportional to the difference in color as perceived by humans. Fig. 2 shows a^*b^* chromaticity diagram with the confusion loci of Duteranope. Colorblind people cannot perceive hue changes parallel to the confusion loci, but they can perceive changes perpendicular to the confusion loci as well as normal people. From Fig. 2, we assume that Duteranope cannot perceive the a^* change, but can perceive the b^* change with sensitivity close to that of normal subjects. Therefore, we estimate the saliency of the a^*, b^* components of the input image.

We reconstruct the Gaussian pyramid of a* and b* component images with eight levels, $(I_{a^*1}, ...I_{a^*8})$ and $(I_{b^*1}, ...I_{b^*8})$. We then resize these images to the same size as the original. This produces a set of images in which the a^*, b^*-component images are gradually blurred. For each group of images in I_{a^*}, I_{b^*}, we compute the sum of difference of Gaussian:

$$M_{a^*} = \sum_{k=1}^{7} |I_{a^*k} - I_{a^*k+1}| \tag{4}$$

$$M_{b^*} = \sum_{k=1}^{7} |I_{b^*k} - I_{b^*k+1}| \tag{5}$$

Equations (4) and (5) extract the difference between one image and the next from the first level of the pyramid, and repeats the operation of adding them together. This allows pixels that are different from their surroundings to stand out more prominently. This method is based on the method for producing saliency map by Itti et al[7]. Thus, M_{a^*}, M_{b^*} obtained for each of the I_{a^*}, I_{b^*} image groups is a grayscale image representing saliency in each hue component of the original image. From

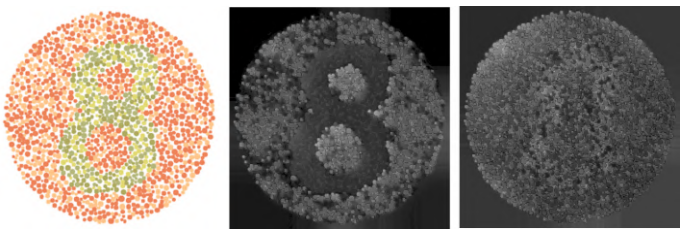


Fig. 3: Example of a^*, b^* component saliency
From left to right: original image, M_{a^*} image, M_{b^*} image

the assumption, the regions where a^* components are more prominent and b^* components are less prominent are the regions where the gap in color discrimination between normal and colorblind people is large. Therefore, the size of the gap is used as an index of correction intency. The formula for determining the correction intency is as follows

$$I_{\text{CCImap}} = M_{a^*} + |M_{a^*} - M_{b^*}| \tag{6}$$

The M_{a^*} is used as the basis, and the regions where the change of M_{a^*} is strong with respect to M_{b^*} are emphasized. The two terms $M_{a^*}, |M_{a^*} - M_{b^*}|$ are pre-normalized so that the range of pixel values is 0-255 and then added together. The resulting image is used as Color Correction Intensity Map(CCImap).

B. Color Correction by Pix2pix

For color correction of images, we use pix2pix[8]. The training dataset is a pair of the original image and a degraded image. Degradation is done in such a way that it becomes difficult for Duteranope to distinguish colors. By having the model learn the inverse transformation from the degraded image to the original image, the model learns a color correction that makes color discrimination easier for Duteranope. From Fig. 2, the color perception of Duteranope strongly depends on the vertical direction, i.e., b^* component, of confusion loci. Therefore, the smaller the range of change in b^*, the more difficult it is for Duteranope to distinguish colors. Therefore, an image with a narrower range of b^* values in the $L^*a^*b^*$ color space is created for the input image, and is used as the degraded image. If all b^* values in the input image were fixed to 0, the hue of the image would be almost imperceptible to Duteranope, and the image would be similar to a grayscale image.

When performing model-based color correction, we adjust the range and intensity of the correction with reference to CCImap. This avoids unnecessary color changes, preserves saliency, and keeps the image natural for normal people. Normalize CCImap pixel values to continuous values between 0 and 1. Using this value, CCImap performs alpha blending between the original image and the pix2pix color-corrected image. Pixels with a CCImap pixel value of 0 are corrected to be the same as the original image, and pixels with a CCImap pixel value of 1 are corrected perfectly to the model. Thus, the output image is as follows

$$I_{\text{out}} = (1 - I_{\text{CCSmap}})I_{\text{orignal}} + I_{\text{CCSmap}}I_{\text{pix2pix}} \tag{7}$$

IV. Results and Discussion

500 pairs of color and degraded images were used to train the model. The degraded image is created in $L^*a^*b^*$ color space by reducing the b^* range of the original image from $-127 \leq b^* \leq 127$ to $-50 \leq b^* \leq 50$. Learning was performed with batch size=64, epoch=400, learning rate = 0.0002. The results are evaluated by creating color blindness simulated image from the original image and the output image and comparing them.

Some example is shown in Fig. 4. In the first image, the “1” in the color blindness simulation image is difficult to distinguish. In the second image, the bear is assimilated into the background color. The CCImaps have larger values in the regions of colors that are indistinguishable for the colorblind. The corrected images retain their natural hue compared to the original images, and normal people do not feel much discrepancy in the corrected images. From the corrected color blindness simulation images, we can see that the “1” and the bears can be distinguished from the background colors more easily than in the original images. The results show that CCImap and pix2pix can be used to perform color correction for colorblind people in a natural way when viewed by normal people.

In addition, we also experimented with gamma correction of CCImap and compared the output images by changing the gamma value. The results are shown in Fig. 5. The CCImap is becoming more bright and dark as the gamma value is increased. Comparison of the color blindness simulation images shows that as the gamma value is increased, the central bear and the background are more easily distinguished. Comparing the original images, the higher the gamma value, the more bluish the bear in the center becomes, resulting in unnatural tints. From these results, it is considered that the gamma correction to CCImap and increasing the gamma value will improve the ease of color discrimination for Duteranope. However, a large gamma value may degrade the naturalness of the image, so it is necessary to estimate the appropriate gamma value for each input image for the future work.

By applying this method to existing posters and websites, we can expect applications such as changing them to take color blindness into consideration without compromising the design. One of the problems is that there is a limitation in improving the ease of color discrimination for the colorblind compared to the method of re-coloring without regard to the hue of the original image. It is also necessary to develop a quantitative evaluation method of the results and to conduct a questionnaire survey of colorblind people.

V. Conclusions

In this paper, we examined a method using image processing to solve the inconvenience in daily life caused by color combinations that are difficult to distinguish for people with color blindness. We have therefore estimated the regions in which colorblind people have more difficulty in color discrimination than normal people in images based on confusion loci theory.



Fig. 4: Color Enhancement Result

From left to right: original image, our method image, original image (simulation of Duteranope vision), our method image (simulation of Duteranope vision), CCImap

We also proposed a method for creating images that are easy to see for people with color blindness while maintaining the hue and saliency of the original image by combining it with color correction based on learning pix2pix models. The result is a color correction for the colorblind while maintaining a more natural appearance.

REFERENCES

[1] R. W. G. Hunt, *Colour Standards and Calculations*. Hoboken, NJ, USA: Wiley, 2004, pp. 92–125.

[2] J. Nathans, D. Thomas, and D. Hogness, "Molecular genetics of human color vision: the genes encoding blue, green, and red pigments," *Science*, vol. 232, no. 4747, pp. 193–202, 1986.

[3] J. B. Huang, S. Y. Wu, and C. S. Chen, "Enhancing color representation for the color vision impaired," *in Proc. Workshop Comput. Vis. Appl. for Visually Impaired*, pp. 1–13, 2008.

[4] Q. U. Dody, L. R. M. Tati, M. Richard, P. G. Andika, and N. A. Tauhid, "Rgb color cluster and graph coloring algorithm for partial color blind correction," *MATEC Web Conf*, vol. 255, no. 01002, Oct, 2019.

[5] H. Brettel, F. Viénot, and J. D. Mollon, "Computerized simulation of color appearance for dichromats," *Journal of the Optical Society of America A*, vol. 14, no. 10, pp. 2647–2655, Oct. 1997.

[6] D. B. Judd, "Standard response functions for protanopic and deuteranopic vision," *Journal of the Optical Society of America*, vol. 35, pp. 199–221, 1944.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, Nov. 1998.

[8] P. Isola and J. Y. Zhu, "Pix2pix: Image-to-image translation with conditional adversarial networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

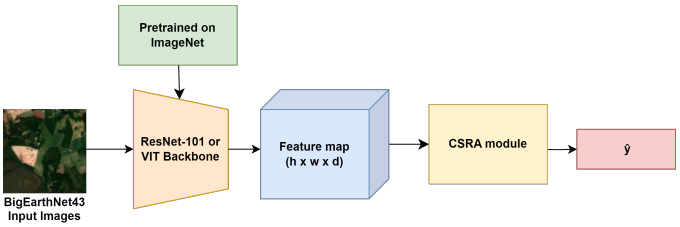


Fig. 5: Example of applying gamma correction to CCImap
From the top row, CCImap, our method image, our method image (simulation of Duteranope vision)

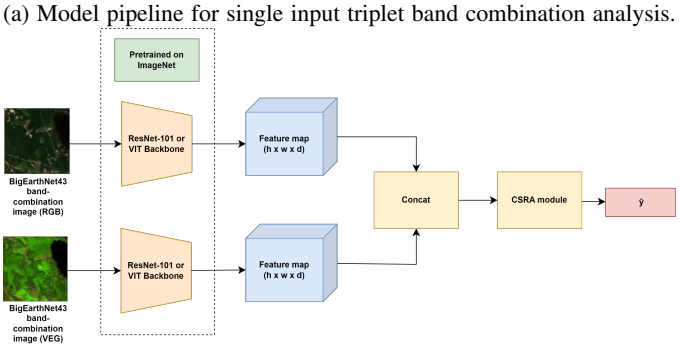
Multi-band Satellite Image Analysis for Multi-label Classification

Sarah Shahmina Abdul Rauf, Mas Ira Syafila Mohd Hilmi Tan, and Yuen Peng Loh
Faculty of Computing and Informatics, Multimedia University, Malaysia
E-mail: {1201102479@student, masirasyafila@, yploh@}.mmu.edu.my

***Abstract*—Multispectral satellite imagery captures rich information of the Earth surface from an extensive range of wavelengths that enhances the understanding and discrimination of Earth objects. When integrated with multi-label classification, it has great potential to provide perceptual comprehension of land cover that exceeds the details observable from the visible light spectrum (RGB). However, the variety of spectral bands each capture a different aspect of the earth, thus can be challenging to identify the optimal spectral band, and their combinations for a classification task. To overcome these issues, this study leverages deep learning by exploring various band combinations and neural network architectures for multi-label classification. Specifically, the performance of triplet band combinations was explored and compared against standard RGB imagery, using ResNet101 as the backbone and the incorporation of the Class Specific Residual Attention (CSRA) mechanism. We then propose the multi-stream triplet band fusion model with Vision Transformer (ViT) backbone and CSRA for multi-label satellite image classification. We found that in the single triple band input approach, the ShortWave InfraRed (SWIR1 or SWIR2) combinations is able to improve F1 scores by 1.25% to 2.02% compared to models using only RGB bands. More notably, our proposed multi-stream approach that fuses the triplets of RGB and Vegetation bands outperformed all other models with an F1 score of 0.7484, consistently surpassing the ResNet101 backbone baseline with similar configurations. These findings reveal the potential of band combination approaches to enhance the Earth object discrimination and land cover analysis.**



(a) Model pipeline for single input triplet band combination analysis.



(b) Proposed multistream model with CSRA for two triplet inputs for improved multi-label classification.

Fig. 1: Multi-band multi-label satellite image classification frameworks.

I. Introduction

Multispectral satellite imagery, which captures a wide range of spectral bands beyond the visible spectrum, has become an invaluable tool for detailed Earth observation. Unlike traditional RGB images, multispectral images offer enhanced discrimination of various land cover types as they provide additional spectral information that can differentiate objects appearing homogeneous in RGB imagery [1], [2]. This added spectral resolution allows for more precise analysis and classification of land cover, which is important for applications like environmental monitoring, agriculture, and urban planning [2].

The varied reflectance behaviors of different Earth objects across multiple wavelengths enable the experimentation with different band combinations to improve object differentiation. However, utilising multispectral imagery for classification poses significant challenges. For instance, using all available bands for the classification of remote sensing images can lead to the curse of dimensionality [3], where the increase in noise worsens classification performance [4]. Reducing the number of bands is a way to avoid this problem, but the selection of optimal bands can be difficult due to the sheer number of spectral bands available [5]. Additionally, in multi-label datasets, the multiple labels can co-occur in complex patterns, where a single image may contain multiple classes such as water, vegetation, and urban areas, which can appear simultaneously but with varying spatial distributions. This makes it difficult for a model to correctly identify and classify all relevant labels in a single representation.

In this study, we aim to explore the effect of different combinations of multispectral bands for multi-label satellite imagery classification. We utilized the BigEarthNet43 [6] multispectral dataset for experimentation with various band combinations and trained various multi-label classifiers by adopting the ResNet101 [7] as the baseline backbone and the Class Specific Attention (CSRA) module [8] for multi-label classification. Lastly, we designed a multi-stream model with ViT backbone, that takes in two band triplets as "images" and concatenates their feature maps for the final multi-label classification by the CSRA module as shown in Fig. 1.

The contributions of the work can be summarised as follows:

- A comparative study on multispectral band combinations for enhancing multi-label satellite image classification. To the best of our knowledge, there has yet to be any work that explores such band combinations for multi-label classification using multispectral data.
- An empirical analysis on band combinations and their contributions to specific classes in multi-label classification. We found that combinations containing ShortWave InfraRed (SWIR1 or SWIR2) improved F1-scores by 1.25-2.02% over RGB-only models, consistently.
- A novel multi-stream approach using the Vision Transformer (ViT) backbone, fusing RGB and vegetation (VEG) band triplets with CSRA module for multi-label satellite image classification. The model achieved an F1 score of 0.7484, outperforming all other tested models.

II. Related Work

With the recent rise in the number of large scale multi-label satellite image datasets, this opens up more avenues for exploring multi-label satellite image classification. Despite this, the research in multi-label classification specifically for multispectral satellite datasets with a large number of samples and class labels is still lacking. Moreover, in the image classification domain, RGB images are most widely used.

For instance, [9] introduced a CNN-based model for hierarchical multi-label annotation, incorporating attention modules and skip-layer connections to enhance discriminative capabilities. Using Inception-ResNet-v2 as a feature extractor pretrained with ImageNet weights, the model integrates attention modules and skip-layer connections in the classification branch, and employs embedding learning to preserve scene-level semantic similarity. It achieved significant improvements in the harmonic mean of label-based and example-based F1 scores (H-F1) when evaluated on datasets like RGB UC-Merced, Ankara, and RGB Multi-label AID.

Alternatively, optimal band combinations had been explored for Landsat-8 in land use and land cover (LULC) classification using Support Vector Machine (SVM) [3]. Addressing the 'curse of dimensionality', the study compared various three- and four-channel band combinations to using all bands, selecting combinations through correlation analysis. The SVM model was specifically trained and evaluated on Landsat-8 with 4 LULC classes in single label classification. Similarly, [5] used a Mask-RCNN to explore optimal band combinations for mapping permafrost tundra landforms, targeting tussock and non-tussock sedge vegetation. The WorldView-02 satellite data was evaluated with five three-band combinations. The Mask-RCNN model, with a ResNet101 backbone pre-trained on the COCO dataset and retrained on 40,000 annotated ice-wedge polygons, processed 200x200 pixel image tiles. They found that triplets of band combinations were most effective.

As a whole, while there have been advancements in multi-label satellite image classification and optimal band combinations, there is considerable potential for further exploration in multispectral datasets with extensive samples and diverse class labels. Thus, our study aims to address this gap by investigating novel band combinations and advanced models to enhance classification accuracy and leverage the full potential of multispectral data.

TABLE I: Sentinel-2 Spectral Bands [11].

Band	Description	Resolution (m)	Center (nm)	Band (nm)
B01	Coastal aerosol	60	443	20
B02	Blue	10	490	65
B03	Green	10	560	35
B04	Red	10	665	30
B05	Vegetation red edge	20	705	15
B06	Vegetation red edge	20	740	15
B07	Vegetation red edge	20	783	20
B08	NIR (Near Infrared)	10	842	115
B09	Water vapor	60	940	20
B11	SWIR1	20	1610	90
B12	SWIR2	20	2190	180
B8A	Narrow NIR (Near Infrared)	20	865	20

III. Methodology

A. Band Combinations

In multispectral satellite imagery, combining and rearranging spectral bands into RGB channels is a commonly used technique [1]. This technique produces false colour composites images, which make discerning certain classes easier as it is able to highlight various different Earth objects, but its use in deep learning-based methods is still limited and mainly focused on classifications of specific classes.

Table I shows the types of spectral bands captured by the Sentinel-2 satellite that is used to collect the BigEarthNet43 dataset. Spatial resolution for the Sentinel-2 ranges from 10m to 60m, where the 10m bands are standard satellite image bands used for land cover classification, and 20m bands are used for vegetation monitoring, whereas bands relating to atmospheric correction have a resolution of 60m [10].

The main hypothesis of this exploration is the positive impact of band combinations on classification performance, where specific combinations can result in significant improvements in accuracy [12]. As such, various triplet sets of band combinations were formed based on their use in satellite imagery analysis for both Sentinel-2 and Landsat satellites, where each band combination serves to emphasise distinct features, as described in [13]. Table II shows the specific combinations for our investigations, and Fig. 2 illustrates the visualization of color composites created from these band combinations showing different resolutions and compositions that may not be as intuitive for human observation.

To elaborate, the RGB composite closely mirrors human vision, making it effective for general analysis, where healthy vegetation appears green, while unhealthy vegetation looks brown or yellow. It is particularly useful for analysing water bodies but often appears low in contrast due to atmospheric scattering [13]. False Color (Urban) composite, is used for distinguishing urban areas from vegetation, displaying water bodies as blue-black and urban areas as grey-purple [13]. The Color Infrared composite, is used in differentiating vegetation

TABLE II: Band combinations

Colour-composite	Abbreviation	Sentinel-2 Band combination (RGB channel)
RGB Colour	RGB	4, 3, 2
False Colour (Urban)	FC	12, 11, 4
Colour Infrared	IR	8, 4, 3
Agriculture	AGRI	11, 8, 3
Land/Water	LW	8, 11, 4
Natural (Atmospheric Removal)	NWAR	12, 8, 3
Shortwave Infrared	SWIR	12, 8, 4
Vegetation Analysis	VEG	11, 8, 4

types and health, showing healthy vegetation in red hues [12], [13]. The Agriculture composite is used for crop health assessment, with vibrant greens indicating healthy crops [13]. The Land/Water composite is often used for distinguishing land from water, useful for highlighting flooded areas [13]. The Natural (Atmospheric Removal) composite provides a clearer view by reducing atmospheric interference, helping in agricultural and post-fire analysis [13]. The Shortwave Infrared composite is useful in vegetation studies, highlighting healthy vegetation in deep red [13]. Lastly, the Vegetation (VEG) Analysis composite is effective for soil moisture and vegetation health analysis, making it suitable for detecting plant vigor and water bodies [13].

B. Multi-label Classification

To address the challenge of multilable classification of images consisting varying object locations, and sizes, we explore both CNN and ViT backbones for feature extraction and adopted the Class-Specific Residual Attention (CSRA) module [8] that enhances spatial attention for each object class, as shown in Fig. 1. Spatial attention directs focus to relevant areas, like artificial structures or forests, while disregarding less irrelevant regions, to improve analysis of complex landscapes.

Particularly, the pipeline begins with an image input into the backbone, which extracts image features and outputs a feature tensor of dimensions $d \times h \times w$. This tensor is then fed into a classifier composed of a fully connected layer with a kernel size of 1×1, producing score tensors of shape $C \times h \times w$, where C is the number of classes. These tensors provide class-specific attention scores for each location within the feature tensor, representing the probability of each class being present at specific locations. Residual attention is applied to these score tensors using either a single-head or multi-head attention mechanism. Single-head attention computes one set of class-specific attention scores for the entire tensor, whereas multi-head attention generates multiple sets of scores, resulting in various logits combined to form the final prediction.

IV. Experiments

In our experiments, we conducted comprehensive evaluations using the BigEarthNet43 dataset [6], consisting 519,284 multispectral images across 43 class labels with diverse environmental and land-use categories. As the dataset features severe imbalanced class distributions, we randomly sub-sampled 118,065 images for our study. Our experimental setup included a consistent 70/15/15 train-validate-test split, resulting in 82,645 training images, 17,710 validation images, and 17,710 testing images that were normalized to the range [0, 1].

A. Implementation

1) Single Input Triplet Model: The baseline for our experimentation involves using the single-headed CSRA module as a classifier, combined with either the CNN (ResNet101) or ViT-B16 backbones, as shown in 1a. Two baseline models were created and trained with BigEarthNet43 RGB images: one using a ResNet101 backbone and the other using the ViT backbone. Both models were trained for 100 epochs, with a learning rate of 0.01 and a batch size of 16. Furthermore, the ResNet101-based model used an input size of 120×120, whereas ViT-based model used 224×224 to support a reasonable resolution for the patch-splitting mechanism.

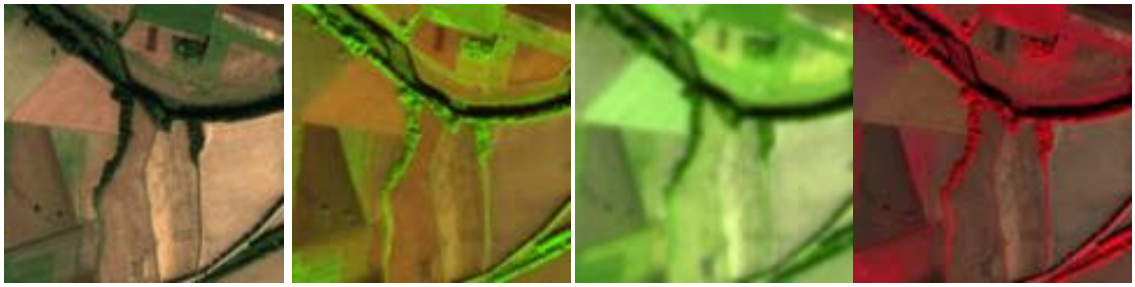
2) Proposed Model: In our investigation, we aim to explore two aspects: (1) identifying the optimal band combination that yields the best performance for the BigEarthNet43 image dataset, and (2) examining the impact on performance when two different color-composite images are input into the model.

For our first approach, we utilised the same pipeline as in Fig. 1a, with the ResNet101 based backbone. For the input, we used the band combinations listed in Table II to create eight false colour-composite image. The hyperparameters of this model follows that of the ResNet101 baseline, using 120×120 image inputs. For the second approach as shown in Fig. 1b, we implemented a two-input multi-stream model with feature map concatenation, using the single-headed CSRA classifier with both ResNet101 and ViT backbones. This multi-stream approach captures more features, enhancing the model's ability to differentiate similar classes compared to a traditional CNN with a single input. The images for this model use band combinations (4, 3, 2) and (11, 8, 4), corresponding to the RGB and Vegetation Analysis colour composites respectively. The different colour composites are passed through the backbone separately, and the feature maps generated are concatenated together by the channel dimension. Just before being passed to the CSRA classifier, a dropout layer is used for regularisation. This model were trained for 100 epochs, with learning rate of 0.001, a batch size of 16 and a dropout value of 0.5. Images with dimensions 120×120 and 224×224 are used, respectively for the models using the ResNet101 and ViT backbones.

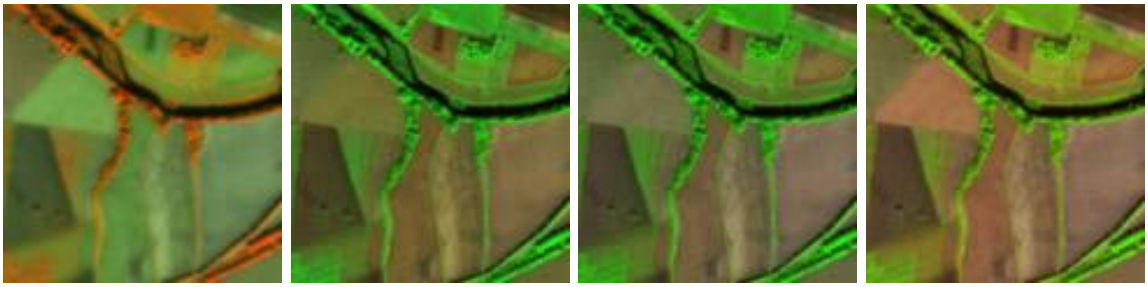
B. Evaluation

This section presents an evaluation of both baseline and proposed approaches. The performance metrics used for the evaluation are commonly used in multi-label classification tasks: mean average precision (mAP), overall precision, overall recall and overall F1 score. For clarity, we indicate each different models trained with the respective band combinations based on the convention of $\{$backbone$\}$-$\{$band combination$\}$ in the subsequent results and discussions.

As seen in Table III, the ViT-RGB was able to outperform ResNet101-RGB across all metrics. ViT-RGB achieved



(a) RGB (b) AGRI (c) FC (d) IR



(e) LW (f) NWAR (g) SWIR (h) VEG

Fig. 2: Visualisation of colour composites created from band combinations in Table II.

a higher mAP of 0.5465 compared to 0.4791 for ResNet101-RGB, suggesting better overall performance in terms of ranking relevant instances higher. Additionally, ViT-RGB showed higher precision (0.8090), and higher recall (0.6846) which indicates that more accurate and relevant predictions are being returned. The F1 Score, was also higher for ViT-RGB (0.7416 vs. 0.7039). These results suggest that the transformer-based ViT architecture is more effective than the ResNet101 architecture when combined with the CSRA module for the classification of BigEarthNet43 satellite images. However, there is a trade-off in computational efficiency. ResNet101-RGB completed each epoch in 1760.04 seconds, while ViT-RGB took 2540.97 seconds. This increased training time with ViT-RGB raises considerations regarding scalability.

Table III provides detailed evaluation metrics for the proposed solutions. In the single image models, the most notable observation is that, band combinations containing bands SWIR1 or SWIR2 consistently outperform those without those bands (RGB and IR), with the F1-score being up to 1.250-2.024% higher when compared to RGB and 1.479-2.255% higher when compared to IR. The ResNet101-VEG model achieves the highest performance amongst the single input models, with an mAP of 0.5162 and the top F1 Score of 0.7242. Furthermore, the ViT-RGB+VEG model demonstrates superior performance with an F1 Score of 0.7484, however the ResNet101-RGB+VEG did not demonstrate this same pattern, with lower scores in all metrics, suggesting that the ViT architecture may be better at integrating the diverse information provided by multiple spectral bands compared to the ResNet101 architecture. When comparing the ViT-RGB+VEG model with ViT-RGB, the former achieved higher scores in all but the recall metric (0.6790 vs 0.6846). The higher mAP and precision indicate that the ViT-RGB+VEG model is better at distinguishing between classes and making accurate classifications but some true positives might be missed.

TABLE III: Evaluation results of multi-label classification with baseline models using RGB bands only, and the proposed multistream models with RGB+VEG fusion.

Models	**Metrics**			
	mAP	**Precision**	**Recall**	F_1 **Score**
ResNet101-RGB	0.4791	0.7857	0.6375	0.7039
ViT-RGB	0.5465	0.8090	0.6846	0.7416
ResNet101-FC	0.5101	0.7816	0.6740	0.7238
ResNet101-SWIR	0.5096	0.7829	0.6686	0.7212
ResNet101-NWAR	0.5124	0.7780	0.6674	0.7185
ResNet101-IR	0.5023	0.7661	0.6472	0.7016
ResNet101-LW	0.5074	0.7806	0.6700	0.7210
ResNet101-AGRI	0.5019	0.7751	0.6660	0.7164
ResNet101-VEG	0.5162	0.7849	0.6721	0.7242
ResNet101-RGB+VEG	0.4667	0.7618	0.5885	0.6640
ViT-RGB+VEG	**0.5942**	**0.8335**	**0.6790**	**0.7484**

The evaluation of F1-scores across various land use classes, as shown in Table IV reveals that the ViT model with RGB+VEG features outperforms ResNet101 models for many classes, particularly in classes such as coastal lagoons and continuous urban fabric. High F1 scores are observed in classes with high intra-class similarity or high occurrence like ‘sea and ocean’ (16309 sample images) and ‘coniferous forest’ (42337 sample images), with the ViT model achieving an F1 score of 0.974 for sea and ocean and 0.869 for coniferous forest. Conversely, certain classes have lower samples or are relating to artificial object tend to have lower intra-class similarity,

including airports (183 sample images) and port areas (108 sample images) exhibit notably low or zero scores.

Table V shows sample images and comparison of their respective ground truth labels with predictions from different models for three images. For the first image, ViT-RGB+VEG predicted all labels correctly, while ViT-RGB and ResNet101-RGB+VEG missing some labels. In the second image, ViT-RGB+VEG likely confused 'industrial or commercial unit' or 'airport' with 'sport and leisure facility'. In the third image, ViT-RGB+VEG missed the 'airport' label but correctly identified the other classes. The two images with 'airport' show very little similarity to each other and are not easily visible, particularly in the third image indicating a challenge for models to perform the classification, thus highlights that classes concerning artificial objects is still a challenge.

In overall, different band combinations are able to boost the performance of multi-label classification, specifically in improving the mAP, Recall, and F1-scores irregardless of the backbone feature extractor. Furthermore, the fusion of these band combinations also shows good potential at improving the performance across all metrics, hence, signifying that such multispectral band combinations and fusion have room for further investigation to solve a complex task as multi-label classification of satellite imagery.

V. Conclusion

This study has demonstrated the efficacy of utilising multispectral satellite imagery and advanced deep learning techniques for multi-label classification tasks. Our experiments highlighted the importance of band combinations in creating effective false-color composites for satellite image analysis. We found that specific band combinations, such as those incorporating bands SWIR1 or SWIR2, consistently outperformed others in terms of F1 score and mAP. We further showed that the fusion composite input could improve the performance on different classes in multi-label classification.

We believe that future studies could explore advanced methodologies incorporating such band combination and fusion approach. Class imbalance mitigation strategies can be explored, such as leveraging techniques like class weighting by assign higher weights to minority classes in the loss function. Additionally, integrating higher-resolution spectral bands could provide richer data for more precise classification and analysis.

References

[1] F. A. Al-Wassai and N. V. Kalyankar, "Major limitations of satellite images," *CoRR*, vol. abs/1307.2434, 2013. arXiv: 1307.2434. [Online]. Available: http://arxiv.org/abs/1307.2434.

[2] S. Borra, D. Thanki, and N. Dey, *Satellite Image Analysis: Clustering and Classification*. Jan. 2019, ISBN: 9811364230. DOI: 10.1007/978-981-13-6424-2.

[3] Z. Yu, L. Di, R. Yang, *et al.*, "Selection of landsat 8 oli band combinations for land use and land cover classification," Jul. 2019, pp. 1–5. DOI: 10.1109/Agro-Geoinformatics.2019.8820595.

[4] L. P. Osco, J. Marcato Junior, A. P. Marques Ramos, *et al.*, "A review on deep learning in uav remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102 456, 2021, ISSN: 1569-8432. DOI: https://doi.org/10.1016/j.jag.2021.102456. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030324342100163X.

[5] M. A. E. Bhuiyan, C. Witharana, A. K. Liljedahl, *et al.*, "Understanding the effects of optimal combination of spectral bands on deep learning model predictions: A case study based on permafrost tundra landform mapping using high resolution multispectral satellite imagery," *Journal of Imaging*, vol. 6, no. 9, 2020, ISSN: 2313-433X. DOI: 10.3390/jimaging6090097. [Online]. Available: https://www.mdpi.com/2313-433X/6/9/97.

[6] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5901–5904. DOI: 10.1109/IGARSS.2019.8900532.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385.

[8] K. Zhu and J. Wu, "Residual attention: A simple but effective method for multi-label recognition," *CoRR*, vol. abs/2108.02456, 2021. arXiv: 2108.02456. [Online]. Available: https://arxiv.org/abs/2108.02456.

[9] P. Zhu, Y. Tan, L. Zhang, *et al.*, "Deep learning for multilabel remote sensing image annotation with dual-level semantic concepts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 4047–4060, 2020. DOI: 10.1109/TGRS.2019.2960466.

[10] F. Spoto, O. Sy, P. Laberinti, *et al.*, "Overview of sentinel-2," in *2012 IEEE International Geoscience and Remote Sensing Symposium*, 2012, pp. 1707–1710. DOI: 10.1109/IGARSS.2012.6351195.

[11] M. Drusch, U. Del Bello, S. Carlier, *et al.*, "Sentinel-2: Esa's optical high-resolution mission for gmes operational services," *Remote Sensing of Environment*, vol. 120, pp. 25–36, 2012, The Sentinel Missions - New Opportunities for Science, ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2011.11.026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425712000636.

[12] J. Tamouk, N. Lotfi, and M. Farmanbar, *Satellite image classification methods and landsat 5tm bands*, 2013. arXiv: 1308.1801 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1308.1801.

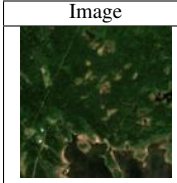
[13] E. D. Analytics, *Satellite band combinations: Analytical methods for imagery*, eos.com, 2021. [Online]. Available: https://eos.com/make-an-analysis/.

TABLE IV: F1-Scores for land use classes based with respect to the band combinations. [Bold indicates the best band for the class.]

Class	ResNet101									ViT	
	AGRI	FC	LW	NWAR	SWIR	VEG	IR	RGB	RGB+VEG	RGB	RGB+VEG
Mixed forest	0.806	0.803	0.796	0.804	0.801	0.803	0.790	0.766	0.781	0.810	**0.821**
Non-irrigated arable land	0.793	0.804	0.816	0.796	0.811	0.817	0.784	0.777	0.729	**0.827**	0.826
Broad-leaved forest	0.715	0.720	0.718	0.711	0.715	0.725	0.678	0.679	0.667	0.724	**0.740**
Complex cultivation patterns	0.587	0.622	0.616	0.607	0.622	0.611	0.558	0.615	0.519	0.662	**0.655**
Water bodies	0.818	0.829	0.820	0.807	0.825	0.826	0.803	0.788	0.789	0.834	**0.846**
Discontinuous urban fabric	0.665	0.704	0.690	0.702	0.702	0.700	0.702	0.705	0.688	0.748	**0.755**
Peatbogs	0.551	0.572	**0.608**	0.538	0.567	0.572	0.578	0.548	0.497	0.594	0.587
Industrial or commercial units	0.438	0.482	0.439	0.480	0.462	0.466	0.505	0.504	0.443	0.506	**0.523**
Olive groves	**0.393**	0.265	0.183	0.252	0.294	0.282	0.244	0.319	0.026	0.296	0.352
Continuous urban fabric	0.327	0.462	0.493	0.359	0.395	0.404	0.350	0.494	0.116	0.617	**0.669**
Vineyards	0.144	0.172	0.267	0.242	0.193	0.185	0.247	**0.339**	0.073	0.253	0.283
Inland marshes	0.170	0.152	**0.185**	0.158	0.161	0.179	0.142	0.063	0.087	0.056	0.066
Sport and leisure facilities	0.122	0.098	0.202	0.074	0.086	**0.167**	0.133	0.025	0.158	0.087	0.134
Mineral extraction sites	0.305	0.323	**0.357**	0.295	0.333	0.299	0.145	0.232	0.272	0.240	0.268
Road and rail networks	0.151	**0.256**	0.233	0.234	0.215	0.234	0.105	0.073	0.172	0.108	0.172
Green urban areas	0.031	0.116	0.031	**0.169**	0.087	0.032	0.085	0.062	0.000	0.000	0.149
Sparsely vegetated areas	0.078	0.172	0.167	0.083	0.151	0.157	0.080	0.172	0.043	0.218	**0.255**
Coastal lagoons	0.549	0.343	0.371	0.568	0.435	0.519	0.480	0.611	0.600	**0.750**	0.711
Estuaries	0.280	0.213	0.226	0.292	0.182	0.200	0.379	**0.438**	0.133	0.353	0.429
Airports	0.000	0.000	**0.059**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Port areas	0.000	0.000	0.000	0.000	**0.118**	0.000	0.000	0.000	0.000	0.000	0.000
Burnt areas	0.000	**0.182**	0.000	**0.182**	**0.182**	0.000	0.000	0.000	0.000	0.000	0.000
Coniferous forest	0.852	0.856	0.852	0.860	0.855	0.859	0.839	0.842	0.806	0.857	**0.869**
Transitional woodland	0.616	0.620	0.605	0.598	0.604	0.610	0.623	0.605	0.568	**0.625**	0.614
LPA-ANV[1]	0.565	0.573	0.569	0.570	0.584	0.558	0.484	0.544	0.350	**0.625**	0.613
Pastures	0.713	0.722	**0.729**	0.709	0.714	0.718	0.704	0.708	0.670	0.721	0.723
Sea and ocean	0.938	0.930	0.928	0.944	0.924	0.925	0.937	0.946	0.928	0.970	**0.974**
Agro-forestry areas	0.739	0.750	0.749	0.738	0.740	0.747	0.751	0.732	0.736	0.771	**0.791**
Permanently irrigated land	0.504	0.480	0.547	0.504	0.474	0.541	0.535	0.478	0.309	0.545	**0.565**
Natural grassland	0.424	0.456	0.431	0.372	0.419	0.438	0.390	**0.467**	0.327	0.432	0.447
Sclerophyllous vegetation	0.441	0.426	0.380	0.390	0.458	0.458	0.341	0.380	0.070	0.540	**0.576**
Water courses	0.698	0.714	0.680	**0.721**	0.696	0.693	0.662	0.584	0.668	0.602	**0.721**
Annual permanent crops	0.405	0.368	0.375	0.386	0.401	0.380	0.348	**0.428**	0.000	0.420	0.399
Moors and heathlands	**0.341**	0.273	0.278	0.310	0.304	0.276	0.271	0.316	0.141	0.306	0.287
Fruit trees and berry plantations	0.126	0.048	0.025	0.134	0.036	0.061	0.085	**0.188**	0.000	0.161	0.153
Rice fields	0.262	0.389	0.365	0.372	0.414	0.382	**0.422**	0.413	0.264	0.327	0.369
Bare rock	0.417	**0.578**	0.421	0.423	0.522	0.489	0.444	0.535	0.385	0.388	0.521
Beaches, dunes, sands	0.611	**0.642**	0.582	0.617	0.585	0.634	0.629	0.641	0.620	0.629	0.640
Salt marshes	0.274	0.194	0.242	0.310	0.187	0.232	0.257	0.282	0.000	0.190	**0.358**
Construction sites	0.000	0.000	0.000	0.000	0.000	0.000	**0.069**	0.000	0.000	0.000	0.000
Intertidal flats	0.154	0.217	0.318	0.195	0.208	0.238	0.150	0.267	0.273	0.341	**0.381**
Dump sites	0.061	**0.063**	0.000	0.057	0.059	0.000	0.000	0.000	0.000	0.000	**0.063**
Salines	0.435	0.560	0.667	0.500	0.615	0.615	0.667	0.417	0.105	0.286	**0.714**

[1] Land principally occupied by agriculture, with significant areas of natural vegetation.

TABLE V: Sample predictions. [Blue: correct predictions; Red: Incorrect predictions.]

Image	Groundtruth	ViT-RGB	ViT-RGB+VEG	ResNet101-RGB+VEG
	Coniferous forest, Mixed forest, Sea and ocean	Mixed forest	Coniferous forest, Mixed forest, Sea and ocean	Coniferous forest, Mixed forest
	Industrial or commercial units, Airports, Non-irrigated arable land, Pastures	Pastures	Non-irrigated arable land, Sport and leisure facilities, Pastures	Non-irrigated arable land, Pastures
	Airports, Coniferous forest, Mixed forest, Transitional woodland/shrub	Coniferous forest, Land principally occupied by agriculture, with significant areas of natural vegetation	Coniferous forest, Mixed forest, Transitional woodland/shrub	Mixed forest, broad-leaved forest, Transitional woodland/shrub

LoFLAT: Local Feature Matching using Focused Linear Attention Transformer

Naijian Cao, Renjie He*, Yuchao Dai, and Mingyi He
Northwestern Polytechnical University, Xian, China
E-mail: caonaijian@mail.nwpu.edu.cn, {davidhrj; daiyuchao; myhe}@nwpu.edu.cn

***Abstract*—Local feature matching is an essential technique in image matching and plays a critical role in a wide range of vision-based applications. However, existing Transformer-based detector-free local feature matching methods encounter challenges due to the quadratic computational complexity of attention mechanisms, especially at high resolutions. However, while existing Transformer-based detector-free local feature matching methods have reduced computational costs using linear attention mechanisms, they still struggle to capture detailed local interactions, which affects the accuracy and robustness of precise local correspondences. In order to enhance representations of attention mechanisms while preserving low computational complexity, we propose the LoFLAT, a novel Local Feature matching using Focused Linear Attention Transformer in this paper. Our LoFLAT consists of three main modules: the Feature Extraction Module, the Feature Transformer Module, and the Matching Module. Specifically, the Feature Extraction Module firstly uses ResNet and a Feature Pyramid Network to extract hierarchical features. The Feature Transformer Module further employs the Focused Linear Attention to refine attention distribution with a focused mapping function and to enhance feature diversity with a depth-wise convolution. Finally, the Matching Module predicts accurate and robust matches through a coarse-to-fine strategy. Extensive experimental evaluations demonstrate that the proposed LoFLAT outperforms the LoFTR method in terms of both efficiency and accuracy.**

I. Introduction

Image matching is a fundamental task in computer vision, and is crucial for various applications including object recognition [1], 3D reconstruction [2], [3], and autonomous navigation [4]. This process entails the alignment and comparison of images to accurately identify corresponding features or regions, which is essential for achieving precise visual correspondence across varying viewpoints, scales, and lighting conditions. Among the various techniques employed for image matching, local feature matching stands out as a key approach. This technique focuses on the identification and comparison of distinctive local features within images, effectively managing variations in viewpoint, scale, and illumination by leveraging local descriptors that encapsulate unique aspects of the image content. Generally, existing local feature matching methods can be broadly classified into two categories: detector-based and detector-free approaches [5].

This work is partially supported by National Key Research and Development Project of China (2023YFF0906203), Shaanxi Key R&D Program (2023-ZDLGY-46)
* denotes the corresponding author

Detector-based methods for local feature matching are grounded in the use of specific algorithms to identify salient points in images, such as corners or edges, which are robust to various transformations including scale, rotation, and changes in illumination. After these key points are detected, feature descriptors are computed to encapsulate the local image information around each point, facilitating the matching process by comparing these descriptors between different images. Methods such as SIFT [6], SURF [7], and ORB [8] exemplify this category, offering high robustness and accuracy in identifying correspondences in structured and well-textured environments. However, these methods have notable limitations, particularly in scenarios with repetitive patterns, texture-less regions, or significant viewpoint changes, where the reliance on distinct key points may lead to inadequate matches and decreased performance.

On the other hand, detector-free methods, also known as dense or end-to-end learning-based methods, avoid the explicit detection of key points [9]. Instead, they focus on matching features across all pixels or dense grids within images, thereby leveraging the full image content. These methods utilize deep learning architectures, such as Convolutional Neural Networks (CNNs) and Transformers [10], to jointly learn feature representations and matching functions. By training on large datasets, these approaches can effectively handle a wide range of challenging conditions, including texture-less regions, complex lighting, and severe geometric transformations. Techniques like FlowNet [11] and SuperGlue [12] exemplify this paradigm, offering superior flexibility and adaptability in diverse visual tasks. More recently, methods based on the Transformer backbone have emerged to better model long-range dependencies [13]–[16]. Through self-attention mechanisms, Transformers can simultaneously consider all pixels or feature points in an image, enabling feature matching within a global context. This capability is particularly beneficial for addressing significant viewpoint changes or complex scenes. As one representative work, LoFTR [15] employs self-attention and cross-attention blocks to update cross-view features. Notably, LoFTR integrates a linear Transformer to replace the global full attention mechanism, thereby achieving a manageable computational cost. It also employs self-attention and cross-attention blocks to update cross-view features. However, studies have shown that the cross-attention maps generated by the linear Transformer in LoFTR tend to spread over larger areas rather than concentrate on the actual corresponding areas.

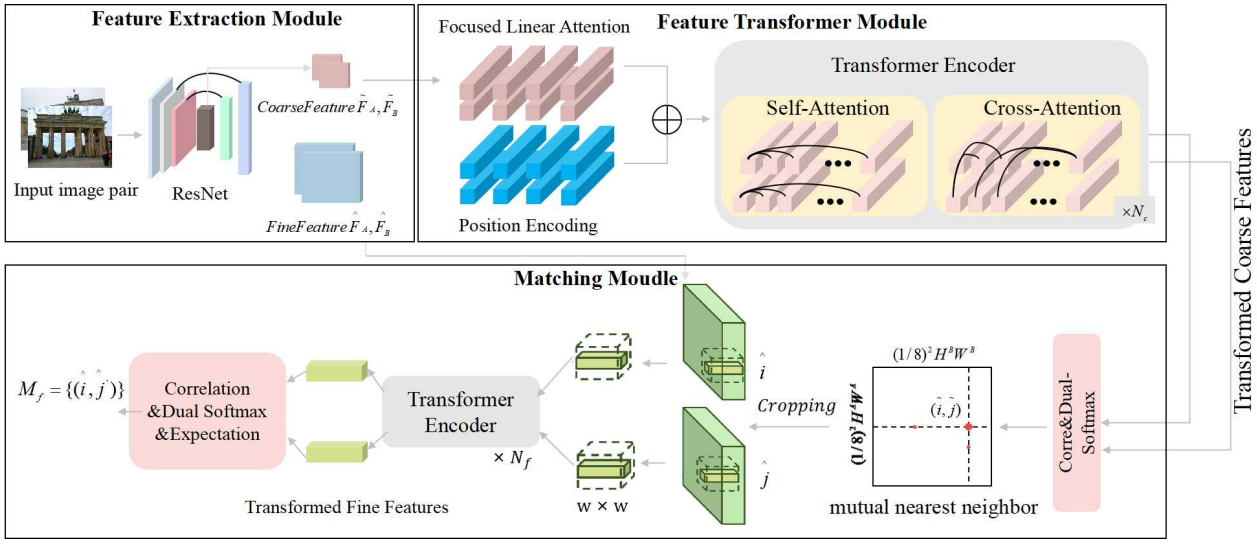


Fig. 1. Overview of the proposed LoFLAT.

Consequently, LoFTR encounters a significant limitation in extracting highly accurate and stable local correspondences due to the deficiency in detailed local interactions between pixel labels.

To address this issue, this paper introduces a novel local feature matching method based on Focused Linear Attention Transformer. The proposed LoFLAT consists of three key modules: the feature extraction module (FEM), the feature Transformer module (FTM), and the matching module (MM). Specifically, the FEM firstly employs ResNet and a Feature Pyramid Network (FPN) to extract multi-level features from the two images to be matched. After that, the FTM is applied to further capture contextually relevant and position-dependent local features. Specifically, a focused mapping function is utilized to ensure a sharp attention distribution by adjusting the directions of each query and key feature while simultaneously repelling dissimilar pairs. Moreover, depth-wise convolution is incorporated to enhance feature diversity and capture the deep contextual structure of the input images. Finally, the MM predicts accurate and robust feature matches using a coarse-to-fine framework. Experimental results on the MegaDepth dataset demonstrate that the proposed LoFLAT outperforms LoFTR in terms of efficiency and feature representation quality.

II. Proposed Method

As shown in Fig. 1, the feature extraction module first employs ResNet [17] as the backbone network for feature extraction. In order to extract features at different resolutions, a feature pyramid network (FPN) [18] is also incorporated to construct a multi-scale feature pyramid. After that, features are transformed by the feature transformer module, which leverages self-attention and cross-attention layers to capture local features that are both contextually relevant and position-dependent. Finally, the matching module outputs precise and robust feature matches through a coarse-to-fine process.

A. Feature Extraction Module

As illustrated in Fig. 1, the FEM utilizes the ResNet with a feature pyramid network to extract multi-scale feature maps. In our work, coarse-level features $\tilde{F}_A, \tilde{F}_B$ are extracted with $\frac{1}{8}$ of the original image size, and fine-level features $\hat{F}_B, \hat{F}_B$ are obtained with $\frac{1}{2}$ of the original image size, respectively.

B. Feature Transformer Module

After the extraction of local features, the Feature Transformer module further is subsequently applied to transform the initial features $\tilde{F}_A$ and $\tilde{F}_B$ into $\tilde{F}_A^t$ and $\tilde{F}_B^t$ for improved representation. The objective of this module is to integrate positional and contextual information, thereby augmenting the robustness and discriminative power of the features during the matching process. The transformed feature representations $\tilde{F}_A^t$ and $\tilde{F}_B^t$ facilitate easier matching, significantly improving both the accuracy and robustness of the matching process. It considers not only the intrinsic properties of local features but also the inter-feature relationships and global contextual information, thereby providing a more holistic capture of the structural details within the images.

The core component of Vision Transformer (ViT) is the Transformer encoder, which comprises multiple self-attention layers and feed-forward neural network layers designed to capture global contextual information. By employing the self-attention mechanism, ViT is able to aggregate and interact with information across the entire image, thus extending beyond the constraints of a local receptive field. Typically, these attention layers employ either standard Softmax attention or Linear attention.

1) Softmax Attention: Softmax Attention is the most basic form of attention in transformer, which assigns weights to value elements by computing the similarity between queries and keys, generating output vectors that allow the model to process information in parallel and capture global context. The

Softmax Attention function can be defined as follows:

$$\mathrm{Att}(Q,K,V)=\mathrm{softmax}\left(QK^T\right)V, \tag{1}$$

where Q, K, and V refer to the query, key and value vectors, respectively. However, in high-resolution image feature matching tasks, the number of feature points significantly increases, leading to a substantial growth in token size. This makes the employment of Softmax Attention impractical due to the computational demands.

2) Linear Attention: In practice, directly employing Softmax attention becomes challenging and often results in prohibitively high computational costs due to the quadratic complexity of Softmax attention with a global receptive field. As a result, most approaches primarily employ Linear attention as an efficient alternative due to its superior computational efficiency, lower memory requirements, and effectiveness in real-time processing. In contrast to the $\mathcal{O}\left(N^2\right)$ complexity of Softmax attention, Linear attention reduces the complexity to $\mathcal{O}\left(N\right)$, significantly enhancing the efficiency of processing large-scale image data. This makes Linear attention particularly suitable for resource-constrained environments and applications with stringent real-time requirements. The calculation of Linear attention is expressed as follows:

$$\mathrm{Att}_{\mathrm{L}}(Q,K,V)=\phi(Q)\phi(K)^TV, \tag{2}$$

where $\phi(x)$ denotes the kernel function that maps data into a new space through nonlinear transformations, while also structurally altering the computation order. Linear attention constrains the model to consider only a small subset of key points most relevant to each query point, which reduces the number of necessary comparisons and mitigates the computational bottleneck faced by traditional attention mechanisms when dealing with large-scale datasets.

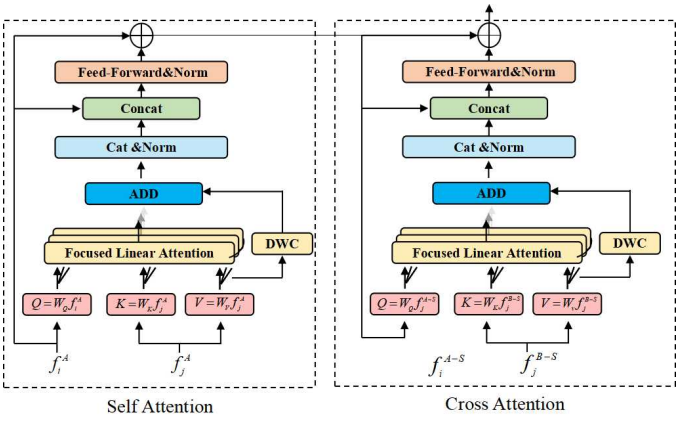


Fig. 2. Architecture of the Feature Transformer Module.

However, existing Linear Attention methods still require a balance between model complexity and representational capacity. While using simple approximations, such as ReLU and ELU activations, may lead to a decrease in performance, more complex kernel function designs or matrix decomposition techniques can result in increased computational costs. Furthermore, Linear Attention may be insufficient in capturing long-range dependencies, which can negatively impact the accuracy of feature matching.

3) Focused Linear Attention based Transformer: Previous studies have demonstrated that Linear attention exhibits a relatively smoother distribution compared with Softmax attention. This suggests that the attention weights in Linear attention are more uniformly distributed across multiple input features. However, for coarse matching, the sharper distribution characteristic of Softmax attention enables the model to focus on a smaller set of key features, thereby enhancing its discriminative capability. To address both model complexity and representational effectiveness, we propose applying Focused Linear attention [19] in coarse matching. In the design of the Feature Transformer model, we initially employ a focused mapping function to distill key features from the queries Q and keys K. Following this, we adhere to the principles of Linear Attention to reduce computational cost while maintaining the accuracy of the attention layer output. Concurrently, depth-wise convolution is applied to the values V to capture detailed local features, enhancing the model's sensitivity to fine-grained information. Ultimately, we integrate the outputs from these two processes to produce a feature representation that balances global context with rich local details. The architecture of the Focused Linear Attention based Transformer is illustrated in Fig. 2. This approach not only significantly enhances the performance of the attention mechanism but also ensures that the model remains computationally efficient while effectively processing information.

Specifically, the Focused Linear attention can be modeled as follows:

$$\mathrm{Att}_{\mathrm{FL}}(Q,K,V)=\phi_p(Q)\phi_p(K)^TV+\mathrm{DWConv}(V) \tag{3}$$

In (3), the first term incorporates a focused function ϕ_p to replace the original kernel function, which is defined as follows:

$$\begin{aligned}\phi_p(x)&=f_p(\mathrm{ReLU}(x)),\\ f_p(x)&=\frac{\|x\|}{\|x^{**p}\|}x^{**p},\end{aligned} \tag{4}$$

where x^{**p} represents element-wise power p of x.

The inclusion of the ReLU function ensures that all input values are non-negative and valid. This modification is proved to affect only the direction of the feature vectors. By adjusting the directions of query and key features, it effectively aligns similar query-key pairs while distancing dissimilar ones. This refined adjustment amplifies the focus on pertinent query-key pairs and diminishes attention to irrelevant pairs, thereby resulting in a more pronounced and sharper attention distribution within the Linear Attention framework.

The second term in (3) is the depth-wise convolution (DWConv), which is employed to further improve the model's capability in capturing local features and increasing feature diversity. The DWConv component acts as a localized attention mechanism, enabling each query to focus solely on a limited

set of features within its spatial neighborhood rather than considering all features globally. The benefit of this localized attention is that, even if two queries receive identical attention weights in a traditional Linear Attention mechanism, DWConv ensures that they extract information from distinct local feature subsets.

C. Matching Module

With both position and context dependent local features obtained, we proceed to establish accurate and robust matching pairs through a coarse-to-fine process. Specifically, the dual-Softmax operator is firstly applied to determine a coarse level matches by estimating a probability $\mathcal{P}_{coarse}$, which is calculated as follows:

$$\mathcal{P}_{coarse}(i,j) = \mathrm{softmax}(\mathcal{S}(i,\cdot))_j \cdot \mathrm{softmax}(\mathcal{S}(\cdot,j))_i, \quad (5)$$

where $\mathcal{S}$ represents the similarity score matrix, which is calculated by taking the dot product of two feature vectors $\tilde{F}_{\mathrm{t}}^{A}(i)$ and $\tilde{F}_{\mathrm{t}}^{B}(j)$ as follows:

$$\mathcal{S}(i,j) = \frac{1}{\tau} \cdot \left\langle \tilde{F}_t^A(i), \tilde{F}_t^B(j) \right\rangle, \quad (6)$$

where τ is a parameter to adjust the smoothness of Softmax operation.

Coarse matches are initially obtained based on the probability matrix $\mathcal{P}_{coarse}$. The mutual nearest neighbor criterion is then employed to filter out potential outliers by ensuring bidirectional consistency in the matching pairs. For each coarse match, the positions are first localized on fine-grained feature maps. Subsequently, local windows are cropped around these localized feature points. A smaller Transformer module is applied to these local feature windows to capture contextual information within the local area. Within the local window, the correlation between a feature point and all other points in the same window is calculated, resulting in a heatmap where each pixel value represents the probability of matching with the reference point. By computing the expected value over the probability distribution on the heatmap, the final matching position is obtained with sub-pixel accuracy. Ultimately, the refined positions of each coarse match pair are output with sub-pixel precision, providing a more accurate correspondence between the two images.

III. Experiments and Analysis

A. Datasets

We evaluate the proposed method on the MegaDepth dataset [20], which consists of more than one million landmark photographs from diverse outdoor scenes captured from the This dataset includes a wide range of lighting conditions and scale variations, featuring both day and night images. The MegaDepth dataset is used to construct three-dimensional reconstructions of approximately 200 world-renowned landmarks. These reconstructions are created using the open-source software COLMAP [2], which processes the images through Structure-from-Motion to derive the camera's internal and external parameters, scene structural information. The depth maps are then generated through the Multi-View Stereo process. For detailed analysis and experimentation, we select tens of thousands high-quality images from the dataset. The primary challenge presented by the MegaDepth dataset is achieving accurate matching despite extreme viewpoint changes and repetitive patterns. This challenge is addressed by leveraging stereo image pairs derived from the three-dimensional scene information and camera parameters to ensure precise correspondences at the pixel level.

B. Evaluation Metrics

Following previous work [12], [15], [21], [22], we report the AUC of the pose error at thresholds of $5^\circ, 10^\circ$, and 20° in our experiments. . The pose error is measured as the maximum of the angular errors in both rotation and translation. To determine the camera pose, we use RANSAC to estimate the essential matrix from the predicted matches. Moreover, we do not use matching precision for comparison, as there is currently no well-established metric for evaluating detector-free image matching methods.

C. Implementation details

During the training phase, we utilized 4 NVIDIA GeForce RTX 3090 GPUs, each with a batch size of 1, resulting in a total effective batch size of 4. The entire training process spanned approximately 69 hours and consisted of 30 epochs. We employed the AdamW optimizer with a weight decay parameter of 0.1. The initial learning rate was set to 0.006, with dynamic adjustment of the actual learning rate during runtime. To mitigate the risk of gradient explosion, we applied gradient clipping with a threshold of 0.5. Additionally, a random seed of 66 was established to ensure reproducibility in data sampling. For pose estimation, we also adopted the RANSAC method [23].

The relative pose between image pairs is estimated based on the computed matching positions, and the accuracy of the matching is assessed using AUC. Pose error is defined as the maximum value among the angular errors in rotation and translation.

D. Experimental Results

For fair comparisons, we retrained LoFTR as the baseline and resized the original images to 500 $\times$ 500 pixels. The comparison of matching between LoFTR and the proposed LoFLAT on MegaDepth dataset is illustrated in Fig. 3. It is observed in Fig. 3 (a) and (b) that both methods are capable of achieving dense matching results in scenarios with small viewpoint changes. However, our method produces a greater number of matches with higher accuracy. In Fig. 3 (c) and (d), it is demonstrated that in scenarios with significant viewpoint and scale changes, the accuracy of LoFTR noticeably decreases, leading to an increased number of erroneous matches. Our method, on the other hand, significantly improves matching accuracy in these challenging cases. Overall, our approach outperforms LoFTR in both match density and accuracy, demonstrating superior robustness and stability.

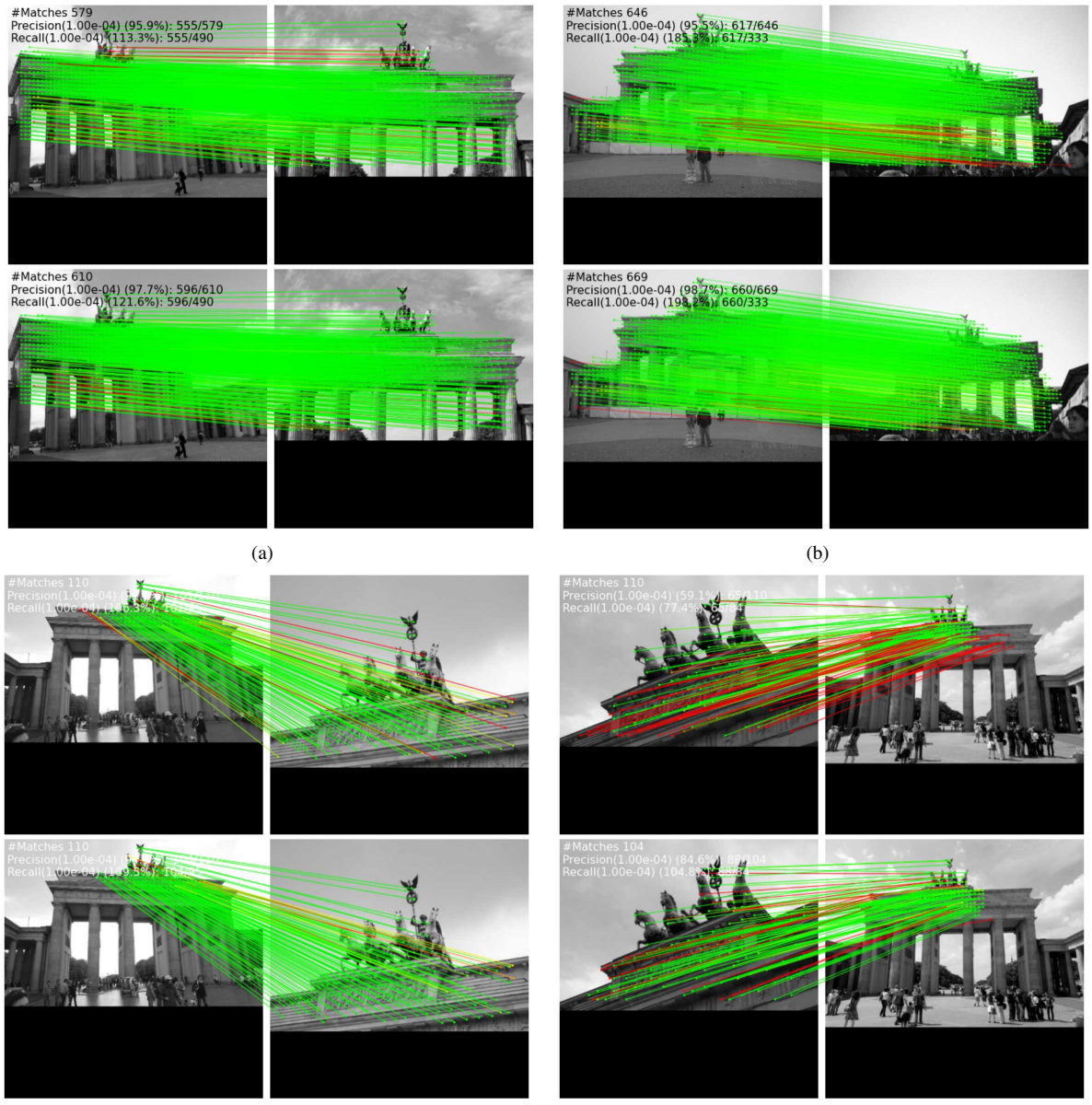


Fig. 3. Qualitative comparison with LoFTR. top: results of LoFTR, bottom: results of LoFLAT

TABLE I
EVALUATION OF RELATIVE POSE ESTIMATION ON MEGADEPTH DATASET

	AUC@5°	AUC@10°	AUC@20°
LoFTR	42.9	60.6	74.8
LoFTR+Focus	44.9	61.7	75.1
LoFLAT	**45.6**	**62.5**	**75.9**

We further compare the performance of LoFTR and our proposed LoFLAT in terms of AUC metric at different angular error thresholds ($5°, 10°$, and $20°$). It is observed in Table I that our model outperforms the baseline method, LoFTR, across all three thresholds, with improvements of 2.7%, 1.9%, and 0.9%, respectively. Notably, the performance gains are more pronounced under stricter threshold conditions (i.e., smaller angular errors). This suggests that our method excels particularly well in tasks requiring precise matching.

IV. CONCLUSIONS

In this paper, we introduced a novel local feature matching method based on the Focused Linear Attention Transformer, to address the limitations of LoFTR. Our method comprises three main modules: the Feature Extraction Module, the Feature Transformer Module, and the Matching Module. The Feature Extraction Module leverages ResNet and a Feature Pyramid Network to extract hierarchical features, while the Feature Transformer Module refines attention distribution using a focused mapping function and depth-wise convolution to enhance feature diversity. The Matching Module employs a coarse-to-fine strategy to predict accurate and robust matches.

Extensive experimental evaluations demonstrate that the proposed LoFLAT significantly outperforms the LoFTR method in terms of both efficiency and accuracy.

REFERENCES

[1] B. Hou, Y. Liu, N. Ling, Y. Ren, and L. Liu, "A survey of efficient deep learning models for moving object segmentation," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 1, 2023.

[2] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[3] Z. Rao, M. He, Y. Dai, Z. Zhu, B. Li, and R. He, "Nlca-net: A non-local context attention network for stereo matching," *APSIPA Transactions on Signal and Information Processing*, vol. 9, e18, 2020.

[4] Z. Zhang, J. Zhao, C. Huang, and L. Li, "Learning visual semantic map-matching for loosely multi-sensor fusion localization of autonomous vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 358–367, 2022.

[5] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 23–79, 2021.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[7] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, Springer, 2006, pp. 404–417.

[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*, Ieee, 2011, pp. 2564–2571.

[9] X. Li, K. Han, S. Li, and V. Prisacariu, "Dual-resolution correspondence networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 346–17 357, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] A. Dosovitskiy, P. Fischer, E. Ilg, *et al.*, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[12] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[13] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. M. Yi, "Cotr: Correspondence transformer for matching across images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6207–6217.

[14] H. Chen, Z. Luo, L. Zhou, *et al.*, "Aspanformer: Detector-free image matching with adaptive span transformer," in *European Conference on Computer Vision*, Springer, 2022, pp. 20–36.

[15] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[16] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou, "Efficient loftr: Semi-dense local feature matching with sparse-like speed," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 666–21 675.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[19] D. Han, X. Pan, Y. Han, S. Song, and G. Huang, "Flatten transformer: Vision transformer using focused linear attention," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 5961–5971.

[20] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2041–2050.

[21] E. Brachmann and C. Rother, "Neural-guided ransac: Learning where to sample model hypotheses," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4322–4331.

[22] J. Zhang, D. Sun, Z. Luo, *et al.*, "Learning two-view correspondences and geometry using order-aware network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5845–5854.

[23] R. Raguram, J.-M. Frahm, and M. Pollefeys, "A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus," in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10*, Springer, 2008, pp. 500–513.

New approach on Smiling faces with Domain Transfer in Latent Space

Felix Ming-Fei Duan *, Wan-Chi Siu *# *Life-FIEEE*, and Chun-Chuen Hui *#
* Saint Francis University, Hong Kong
# The Hong Kong Polytechnic University



Fig. 1 Faces manipulated with smiling by our domain transfer approach.
For each face, we get its manipulation result from 10 adjacent domains of our model, indicating different degree of smiling.

***Abstract*— Sometimes we do not satisfy with our own old photo without smiling. Do you want to make it smile? One possible approach traditionally is to make use of the manipulation technique in latent space to achieve this. However, often we find that this may produce some errors. In this paper, we propose a new smile manipulation approach using Domain Transfer in Latent Space (DTLS) approach. We define various degrees of smile as a series of domains transfer in the DTLS. Our experimental results show that with this of our approach, manipulated photo shows less errors than traditional method, visual comparison can be found in this paper.**

I. INTRODUCTION

Face manipulation technology is a cutting-edge advancement in the fields of computer vision and image processing. It stands out for its ability to manipulate facial features in images, including expressions, age, and mouth openness [1-3]. This innovative technology has captured the attention of academia and industry alike due to its extensive range of potential applications, sparking interest in its capabilities and possibilities.

The influence of face manipulation technology extends across various sectors, notably leaving its mark on digital art, entertainment, and human-computer interaction. In these domains, it plays a crucial role in enhancing visual elements such as movie special effects, video games, and virtual reality experiences, while also contributing to advancements in character design and performance, enriching the overall user experience.

Looking beyond its current applications, the realm of face manipulation technology opens up a myriad of creative opportunities. For example, personalized image editing empowers individuals to tailor features like smiles and age in their photos to achieve desired effects. Fig. 1 shows our smiling manipulation using domain transfer technique which has great potential in editing human expression. This not only caters to the needs of social media and photography enthusiasts but also paves the way for unique avenues in personal branding and creative expression, fostering individuality and artistic exploration.

In the past, the conventional methodology for face manipulation involved the utilization of high-performance generative models such as DDPM and StyleGAN [4-6], which operate within a distinct latent space to pinpoint semantic vectors associated with specific features, thus enabling the manipulation of facial attributes [1, 2]. Specifically, the latent space refers to input space of generative models where inputs sampled from this space represent different faces. A powerful latent space can decouple different kinds of attributes, making

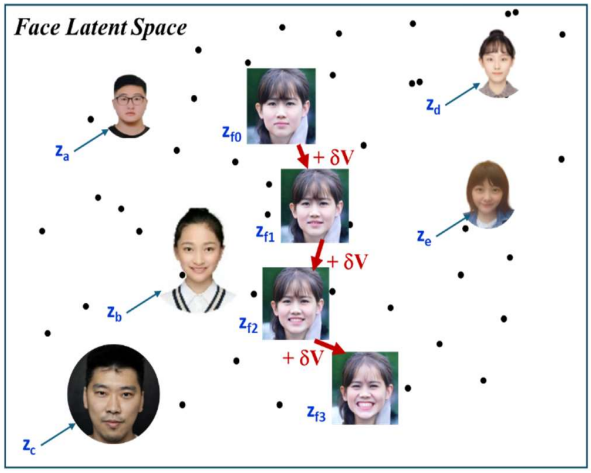


Fig. 2 Manipulate in latent space with smiling vector

manipulation possible. Semantic vectors, on the other hand, are specific vectors within this space that encode meaningful information about certain features, like a smile or other facial characteristics. Manipulating images by adjusting these semantic vectors involves changing these attributes in the latent space to reflect desired modifications in the output image. The procedure of manipulating with semantic vector is shown at Fig. 2, while to make sure the semantic vector to change smiling exists, a functional and powerful latent space is required. Moreover, the translation of real images into this space often introduces errors, hindering the efficacy of this technique for manipulating real facial images with precision and accuracy [7].

To overcome these challenges and reduce dependence on latent space navigation, and inspired by the domain transfer approach for super-resolution[8], in this paper we suggest a novel approach with domain transfer for smiling operation. As shown in Fig. 1, by creating distinct domains that represent varying degrees of specific features, and making transition training, this method facilitates precise control over feature transition, ultimately resulting in better performance and effective face manipulation tasks. We believe that this domain transfer approach marks a significant step forward advancing the capabilities and potential of face manipulation technology in diverse industrial applications academic studies.

Compared to the previous methods based on the latent space navigation, our model can directly modify input real images without the need to map them explicitly into the latent space. This reduces computational costs and ensures that results of facial manipulations closely resemble the original image, avoiding errors that may arise from mapping. Furthermore, we suggest to use a U-net with domain identifier t to separate the degree of smiling, and it is obviously similar to other U-Net with encoder-decoder architecture and convolutional characteristics [9].

In terms of training data, we do not rely on excessive amounts of data. The model can be trained with small-scale of data, thereby saving computational costs. In contrast to traditional convolutional networks, our model supports migration between different levels of modification and can extract results from each domain, offering greater flexibility.

II. METHODOLOGY

In this work, we make use of Domain Transfer in Latent

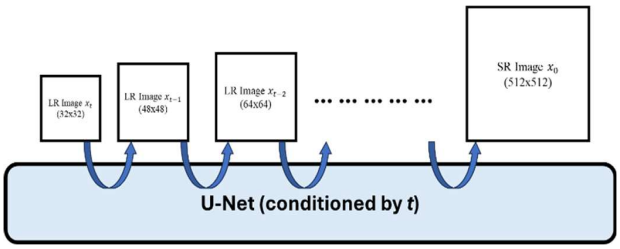


Fig. 3 Structure of DTLS for image super-resolution

Space (DTLS) technique to transform unsmiling human faces into smiling ones. DTLS was originally used as iterative domain transfer approach for image super-resolution, as shown in Fig. 3. As shown in the figure the DTLS can transfer a low-resolution image incrementally to a higher-resolution image, such as from 32x32 to 48x48, or from 64x64 to 80x80, and … to 512x512, etc.).

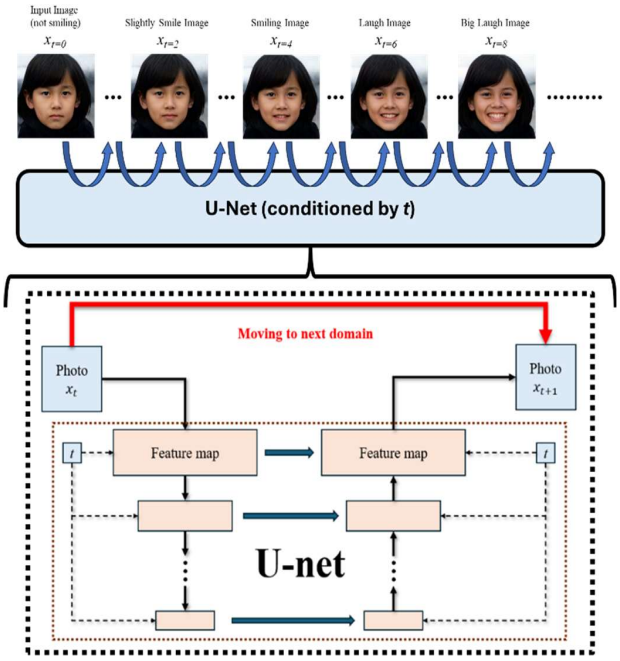


Fig. 4 Detail structure of our smile domain transfer model where U-net is applied to transfer photo among domains. U-Net is conditioned by input t which indicates the current domain.

Fig. 4 shows a diagram to illustrate some details of the structure for our approach for smiling manipulation. To turn a non-smiling image, $x_{t=0}$, into a big laughing image, $x_{t=8}$, we utilize a U-Net which is conditioned by t to transfer the image from one domain to another domain progressively. Therefore, each forward domain in our approach should represent a larger degree of smile compared to its previous domain.

As for domain transformation, the U-Net is a very important part of the design and the lower part of Fig.4 shows the details of its structure. The U-Net is conditioned by a variable t which

indicates the current smiling degree, and is used set the degree of smiling image transition, hence each smiling objective is conditioned by *t*. This enables the U-Net to understand the degree of smiling across all deep feature levels of input images and to target output images.

The U-Net architecture consists of four downsampling and four upsampling blocks to process input image features. Each block comprises two convolutional layers where feature manipulation takes place. As shown in Fig. 5, an input domain representation, *t*, undergoes a sinusoidal transformation followed by processing through three fully connected layers. This transformed *t* is then channel-wise multiplied with and added to the feature maps within each convolutional block. This mechanism enables the U-Net to manipulate image features across various domains separately and effectively, and achieve

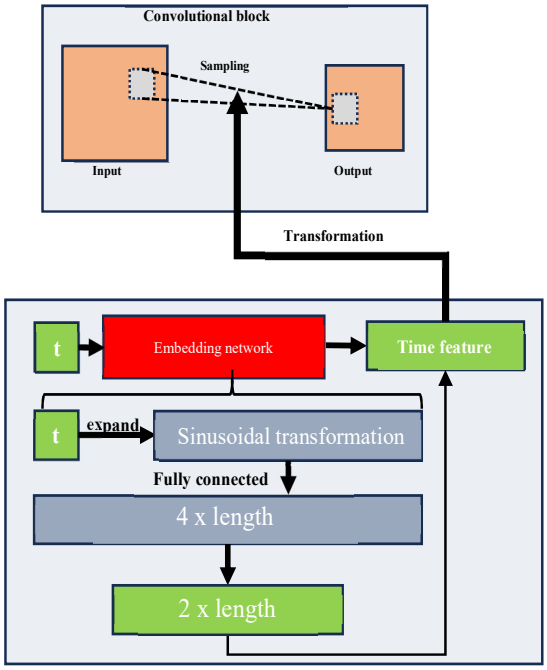


Fig. 5 The procedure of injecting time information into U-Net

domain transfer.

A key aspect of domain transfer is its ability to learn the statistical properties of images across a wide range of domains. This is achieved through training iterations with multiple pairs of input and target domain images, such that the DTLS can learn the statistical representation of each assigned domain. This comprehensive training allows the model to understand and replicate the intricate details and patterns present at various scales, enabling it to produce high-quality results across a broad spectrum of image-to-image transfer.

A. Model Structure

Inference Flow

Input: $x_{t=0}$, T
for $t = 0,\ 1, \ldots, T$ **do**
$x_{t+1} \leftarrow U(x_t, t)$
return x_T

The inference flow of our model is shown above. We denote x_t as smiling photo in domain t. $U(x, t)$ is our U-net with input photo x and current domain t, which functions as moving photo x from domain t to domain $t+1$. Therefore, to move a face photo x from domain 0 (not smiling) to the domain T (big laugh smile), we have repeatedly to use the trained U-Net to shift from current domain to next domain until $t = T$.

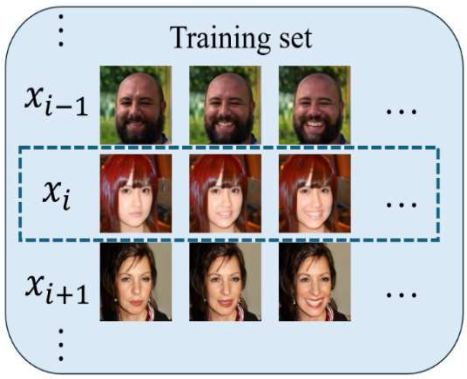


Fig. 6 For each face, the model needs different degree of smiling for training.

B. Training and Loss function

Training Flow

repeat:
$t \sim Randomly\ sample\ from\ \{0 \ldots T-1\}$
$x_0 \sim Randomly\ sample\ from\ \boldsymbol{Dataset}\{x\}$
$y \leftarrow U(x_{0,t}, t)$
$Loss \leftarrow |y - x_{0,t+1}|$
Backpropagate
Until break

For the training procedure, in each iteration, we randomly fetch a face x_0 from the dataset. Then the target domain t is also chosen randomly. The domain transfer process from domain t and domain $t+1$ is to be trained. The face photo x_0 at domain t is fed into the network, and photo at domain $t+1$ is the ground truth to be compared with. Therefore, for each face, a large scale of different degrees of smiling images is required for training. Fig. 6 shows the training dataset of our model. The horizontal level indicates different degrees of smiling and the vertical level indicates different faces.

This appears a one to many problem, since we have various degrees of smiling. MSE loss is used in this work, but it is only a small part of the overall loss function. To enhance the authenticity of the generated images, an adversarial network has to be used. This is to integrate into the overall loss aiming to enrich the realism and lifelikeness of the resulting images.

For image-to-image tasks, Mean Squared Error (MSE) is commonly used for target matching. However, since the U-Net needs to learn multiple representations of the training image in various domains, it processes multiple similar training images during training. This may lead to overfitting of the dataset when

using MSE solely. Therefore, to increase the robustness and photo-realism of output images, using multiple loss functions is proposed.

The training of our module comprises three loss functions. Initially, for the U-Net, domain transfer represents an image-to-image problem with accessible ground truth data. The Mean Squared Error (MSE) loss is driven by the need to accurately align the transferred images with the desired ground truth, ensuring a more precise image-to-image transformation process.

For the second component: the Generative Adversarial Network (GAN) loss, we have to use a Binary-Cross-Entropy (BCE) formulation, taking the realism score given from the discriminator which is another network only applied in training. The BCE equation calculates the losses for both the

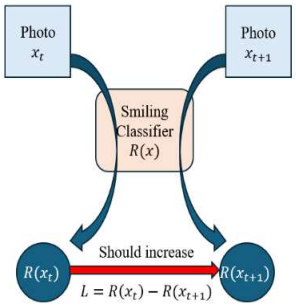


Fig. 7 Smiling Loss to ensuring gradual change on

discriminator and U-Net training [10].

To guarantee the directional movement of images towards smiling expressions during domain transfer, we a smile classifier is suggested as the third component of the overall loss function. This classifier ensures that the degree of smiles in the images increases appropriately throughout the domain transfer process, aligning with the desired facial attribute modifications. Specifically, we a ResNet-50 trained on the CelebA dataset can be used as the binary classifier to discern smiles, assigning a score between 0 and 1 to indicate the extent of the smile in an image [11, 12]. The objective of the training is to allow the smile intensity to increase after undergoing domain transfer via U-Net. Assuming that $R(x)$ denotes the smile score obtained for image x by this ResNet network, we aim to maximize $R(x_t) - R(x_{t+1})$ to ensure that the smile intensity rises. As shown in Fig. 7, therefore, we define $L_{smiling} = R(x_t) - R(x_{t+1})$ as our smiling loss. We have utilized these three loss functions to ensure the correctness of the model's domain transfer process and to maintain the authenticity of the generated images as much as possible, as shown below.

III. EXPERIMENTAL WORK

A. *Training set*

Our model requires large scale human faces with various degrees of smiles. We utilized the StyleGAN2 model to generate our dataset[7]. Specifically, we employ the StyleGAN2 model trained on the FFHQ dataset to create our dataset [1, 7, 13]. To achieve varying degrees of smiles for the same individual, we used an SVM-based approach to identify the semantic vector related to smiling within the latent space of StyleGAN2 [1]. Subsequently, by manipulating smiles in the



Fig. 8 Samples of training set, generated by StyleGAN2

latent space, we generate data across different domains, as illustrated in Fig. 8, where $x_{i,0}$ is the set of initial faces and $x_{i,1}, x_{i,2}, \ldots, x_{i,n}$ are sets of smiling faces with increasing degree of smiles.

We generated our domain transfer smiling dataset which contains 50,000 faces with resolution of 128x128, and each of the faces has respective smiles in 11 domains of smiling. Thus, the total amount of images in generated dataset is 550,000.

B. *Implementation details*

We utilized a U-Net with 128x128 three-channel input images, incorporating the current

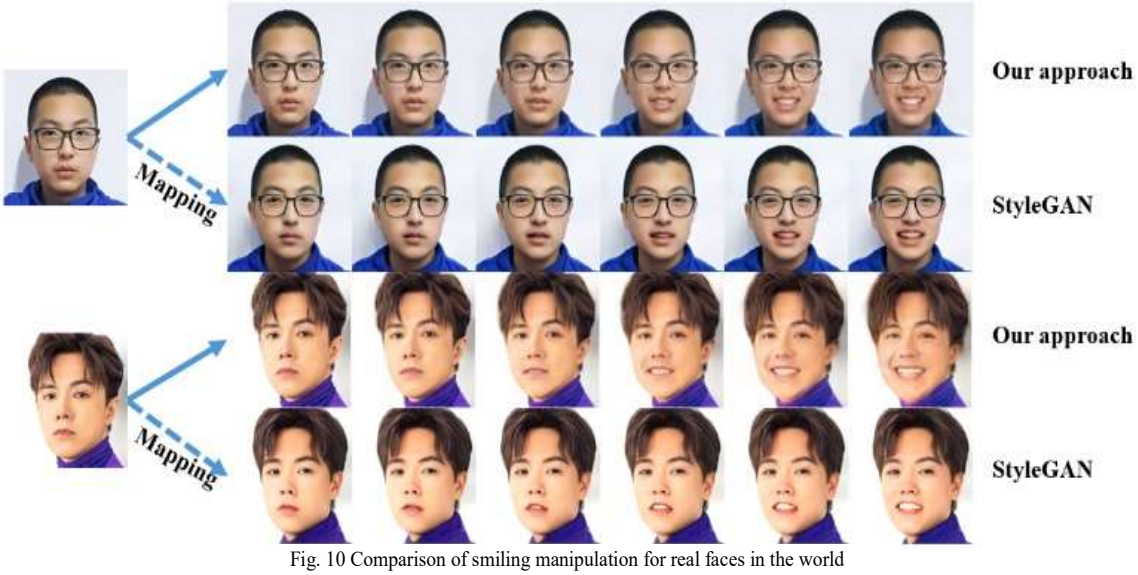


Fig. 10 Comparison of smiling manipulation for real faces in the world

domain signature t as an embedding to provide additional domain information. The output consists of smiling images of the same dimensions. The U-Net architecture employed channel widths of {64, 128, 256, 512}. We utilized the Adam optimizer with betas {0.9, 0.999} and an epsilon value of $1e^{-8}$, while setting the learning rate to $2e^{-5}$. The training involved 100,000 iterations with a batch size of 16.

C. *Evaluation and Comparison*

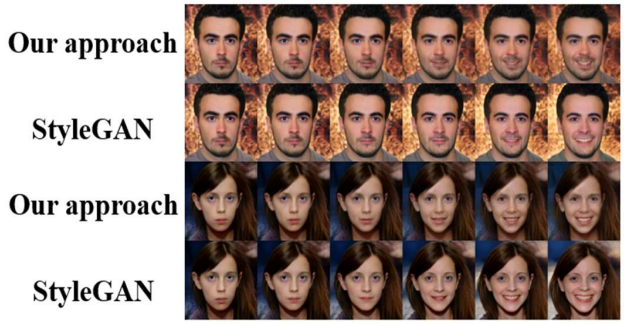


Fig. 9 Comparison of smiling manipulation for fake faces generated by StyleGAN2

Our model performs well in both making fake faces and real faces smile. Unlike StyleGAN2, which is based on latent space manipulation, it only performed well using fake faces generated by StyleGAN2 itself according to our tests. Our model shows superior quality in making both face and real photos smile. In addition, for each inference process, our model only requires 7ms with a single NVIDIA GeForce RTX 4090 GPU, whilst the StyleGAN2 model would take 20ms on average for modifying the latent vector. Hence, our model is more efficient on actual application.

Fake face manipulation is shown in Fig. 9. It shows that using our approach, more facial identities can be retained after face manipulation. You can identify that the manipulated face is still the same person. However, in StyleGAN2, after using semantic vectors in the latent space, the manipulated face demonstrated facial features of another person.

This phenomenon is more obvious in real face manipulation, although mapping these faces to the latent space of GAN network could be a simple procedure [10]. The possible significant errors may come from the over-simplified procedure to just use a single (or a few) δ-vector for the mapping process. Furthermore, the positions in the latent space to which real images are mapped could be relatively sparse, and during training, there might not be sufficient data to effectively disentangle smile features. Consequently, as shown in Fig. 10, our model exhibits noticeably better performance in modifying real photos.

IV. Conclusion

This paper presented a novel approach for manipulating facial expressions, mainly modifying the degree of smiles, using the Domain Transfer in Latent Space (DTLS) model. Unlike traditional methods that operate within the latent space of generative models like StyleGAN2, our technique directly transfers faces across domains representing varying degrees of smiling. Compared to semantic vector manipulation in the StyleGAN2 latent space, our model exhibits superior performance in editing authentic photographs, highlighting its practical utility.

The approach can be generalized to other facial manipulations. Hence this proposed DTLS-based facial editing technique opens up exciting possibilities for various applications, including digital media creation, entertainment, social media, and human-computer interaction. Its ability to generate realistic, controllable facial expressions paves the way for enhancing user experiences, creative self-expression, and novel ways of engaging with digital content.

V. Acknowledgment

This work has been supported by Saint Francis University (under Grant No. ISG200206), The Hong Kong Polytechnic University, and Research Grants Council of the Hong Kong Special Administrative Region, China (under Grant No. UGC/FDS11/E05/22)

VI. Reference

[1] Y. Shen, C. Yang, X. Tang and B. Zhou, "InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 44, no. 04, 2022, pp. 2004-2018.

[2] G. Yang, N. Fei, M. Ding, G. Liu, Z. Lu, and T. Xiang, "L2m-gan: Learning to manipulate latent space semantics for facial attribute editing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2951-2960, 2021.

[3] C. H. Lin, Y. L. Pan and J. L. Wu, "Attribute-Based Facial Image Manipulation on Latent Space," 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Washington, DC, USA, 2021, pp. 1-8.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20), Article No. 574, Dec. 2020, pp. 6840-6851.

[5] T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 43, no. 12, 2021, pp. 4217-422.

[6] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8107-8116.

[7] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "Restyle: A residual-based StyleGAN encoder via iterative refinement," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6711-6720.

[8] C. C. Hui, W.C. Siu, and N. F. Law, "Domain Transfer in Latent Space (DTLS) Wins on Image Super-Resolution -- a Non-Denoising Model," arXiv:2311.02358 [eess.IV], 2023.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234-241.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," in Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), 2014, pp. 2672-2680

[11] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3730-3738.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[13] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 4, pp. 985-994, April 2019.

[14] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in International Conference on Learning Representations (ICLR), 2015.

Data-Driven Tuning for Weighted Least Square of BLE-AoA-based Indoor Localization

Ginji Ohashi*, Shinsuke Ibi*, Takumi Takahashi†, and Hisato Iwai*
* Faculty of Science and Engineering, Doshisha University, 1-3 Tataramiyakodani, Kyotanabe, 610-0394, Japan
E-mail: ctwj0344@mail4.doshisha.ac.jp, {sibi, iwai}@mail.doshisha.ac.jp
† Graduate School of Engineering, Osaka University, Suita, Japan
E-mail: takahashi@comm.eng.osaka-u.ac.jp

***Abstract*—This paper proposes a novel positioning algorithm based on angle of arrival (AoA) in Bluetooth low energy (BLE), utilizing machine learning (ML) to achieve high-precision indoor positioning. Although various positioning methods use BLE, such as radio fingerprinting and proximity methods, the AoA method is known for its superior positioning accuracy and has received significant attention. However, similar to other methods, it is susceptible to the degradation of the positioning accuracy caused by the reflection, obstruction, and diffraction of radio waves by walls and furniture. To address this issue, we exploit ML techniques in the weighted least square (WLS) method in BLE-AoA systems. Using ML models trained on AoA datasets with supervised learning, the weights in the WLS method can be adjusted in a data-driven manner, thereby enhancing positioning accuracy. As the ML models, this paper utilizes weight learning model (WLM), random forest (RF), and multi-layer perceptron (MLP) and validates their effectiveness through computer simulations using publicly available datasets.**

I. Introduction

Indoor positioning, a technology for tracking and identifying the location of people or devices within indoor spaces, has seen an increasing demand over the past decade [1]. Its usage ranges widely from indoor navigation (in hospitals, office buildings, and airports) to object localization (in warehouses for goods and household items) [2]. Global navigation satellite system (GNSS) is a highly accurate and effective technology for use in outdoor environments. However, positioning accuracy is significantly degraded in indoor environments, where exterior walls create non-line-of-sight (NLOS) conditions [3]. Consequently, various indoor positioning technologies have been developed to replace GNSS, many of which utilize wireless technologies, such as ultra-wideband (UWB), Wi-Fi, and Bluetooth [4]. Among these technologies, Bluetooth low energy (BLE) is anticipated to play a crucial role in building wireless communication infrastructure for Internet of things (IoT) technologies owing to its low power consumption [5].

BLE offers multiple positioning metrics, each of which has its own advantages and disadvantages. Received signal strength indicator (RSSI) methods include proximity and fingerprinting; however, both achieve only meter-level accuracy, which is insufficient [6]. Additionally, methods such as time difference of arrival (TDoA) in GNSS or time of flight (TOF) in light detection and ranging (LiDAR) and ultrasonic sensors can lead to significant positioning errors owing to minor clock discrepancies between the transmitter and receiver, making them impractical for BLE, which assumes inexpensive devices [7]. However, BLE in Bluetooth 5.1 and beyond is equipped with direction-finding capabilities, enabling the acquisition of angle of arrival (AoA) using BLE devices [8]. Positioning methods using AoA can achieve centimeter-level accuracy; therefore, they are considered the most precise among BLE-based positioning methods and are attracting significant attention [9]. Therefore, this paper focuses on the AoA method for BLE. In the AoA method, two or more receivers with known positions receive direction-finding signals transmitted by a transmitter, and measure the AoA. The measured AoA is aggregated through the network into an upstream centralized system, where the estimated position coordinates of the transmitter are calculated.

Algorithms such as multiple signal classification (MUSIC) method calculate the AoA, and algorithms based on triangulation or the least square (LS) method calculate the estimated position coordinates [10]. Although the AoA method boasts high positioning accuracy, it is not immune to the effects of reflection, diffraction, and obstruction of radio waves by walls and furniture, which degrade positioning accuracy [11]. These radio propagation characteristics vary depending on the dimensions and layout of the indoor environment, which makes it challenging to address them using statistical methods. Therefore, many studies have explored the application of machine learning (ML) techniques to solve this issue. For example, [12] effectively addressed the AoA measurement error by integrating ML into the MUSIC algorithm.

In this study, we aim to enhance positioning accuracy by introducing ML into existing positioning algorithms. Specifically, we focus on the latest positioning algorithms based on the weighted LS (WLS) method [13], and propose a method to adjust the weights using ML models in a data-driven manner. As the ML models, this paper utilizes three ML models: weight learning model (WLM), random forest (RF), and multi-layer perceptron (MLP), to adjust the weights and evaluate each. The main contribution of this paper is to demonstrate the effectiveness of applying ML to the positioning algorithm of the BLE-AoA method through computer simulations using publicly available datasets created with commercially available BLE devices [14].

The remainder of this paper is organized as follows. Sect. II presents the system configuration and positioning algorithm

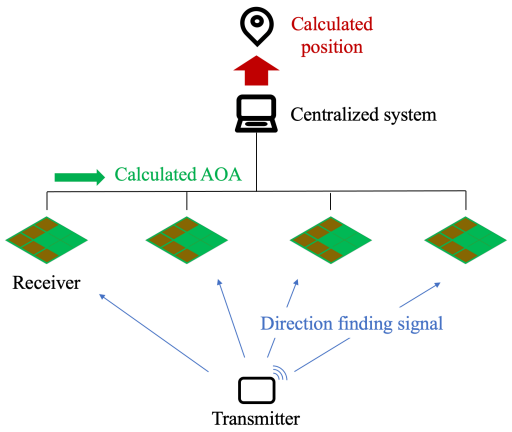


Fig. 1: Overview of AoA-Based positioning system.

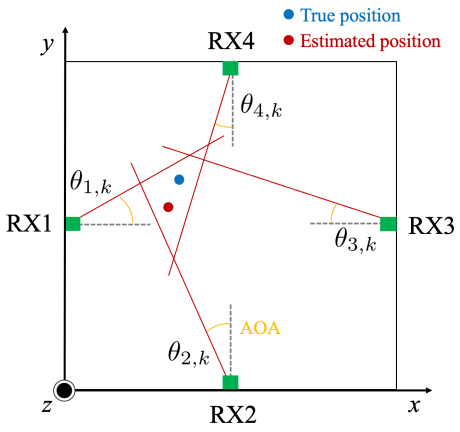


Fig. 2: AoA-Based estimation of transmitter position.

of the AoA-based positioning method. Sect. III explains the proposed machine-learning-based positioning algorithm. Sect. IV first describes the dataset and ML settings used for the simulations and then discusses the effectiveness of the proposed method based on the simulation results. Finally, Sect. V provides a summary and conclusion.

II. Indoor Positioning Using AoA

A. Positioning system configuration

Fig. 1 shows the configuration of the positioning system assumed in this paper. The transmitter, installed at an arbitrary location within the positioning environment, transmits direction-finding signals to the receivers fixed on walls and other surfaces. Each receiver calculates the AoA based on the received signal, and the calculated AoA is aggregated into a centralized system through the network. In the centralized system, the estimated position coordinates of the transmitter are calculated using positioning algorithms based on the triangulation method or LS method. In this study, we conducted offline simulations using a publicly available dataset that recorded the AoAs aggregated in a centralized system. Additionally, BLE allocates channels 37, 38, and 39 to transmit signals in the advertising channels. However, this paper only uses data from channel 37 to ignore the characteristic changes due to the different frequencies used.

B. Positioning algorithm

Fig. 2 is an overview of the AoA-based indoor positioning environment. The green squares represent the fixed receivers (RX1 to RX4), the blue plot represents the transmitter (TX). Let the total number of observed AoA data points be K, and the k-th AoA data point observed at the j-th RX be denoted by $\theta_{j,k}$ [degree]. In vector form, $\boldsymbol{\theta}_k = [\theta_{1,k}, \theta_{2,k}, \theta_{3,k}, \theta_{4,k}]$. The RX and TX position vectors are denoted as $\boldsymbol{a}_j = \left[x^{\mathrm{RX}j}, y^{\mathrm{RX}j}\right]^{\mathsf{T}} (j \in \{1,2,3,4\})$, $\boldsymbol{t}_k = \left[x_k^{\mathrm{TX}}, y_k^{\mathrm{TX}}\right]^{\mathsf{T}}$, respectively. The role of positioning is to estimate the position of TX $\hat{\boldsymbol{p}}_k = [\hat{x}_k, \hat{y}_k]^{\mathsf{T}}$ from observed AoA data $\boldsymbol{\theta}_k$, which is represented by red plot.

In the AoA method, the intersection of the AoAs is considered the estimated position coordinate. However, due to measurement errors in the AoAs, a single intersection point is rarely obtained, and only a candidate region for the estimated position coordinates is identified. Therefore, algorithms such as the LS method are used to determine the estimated position coordinates from the candidate region. In this algorithm, the estimated position vector is determined as the positions that minimize the sum of the squared perpendicular distances D_k from the red lines corresponding to $\boldsymbol{n}_{j,k}$, which is a unit direction vector given by

$$\boldsymbol{n}_{j,k} = \left[\cos\left(\frac{\theta_{j,k}}{180}\pi\right), \sin\left(\frac{\theta_{j,k}}{180}\pi\right)\right]^{\mathsf{T}}. \tag{1}$$

The distance D_k is expressed as

$$D_k = \sum_{j=1}^{4} (\boldsymbol{a}_j - \boldsymbol{p})^{\mathsf{T}} \left(\boldsymbol{I}_2 - \boldsymbol{n}_{j,k}\boldsymbol{n}_{j,k}^{\mathsf{T}}\right) (\boldsymbol{a}_j - \boldsymbol{p}), \tag{2}$$

where $\boldsymbol{p}$ is a position vector and $\boldsymbol{I}_2$ is the identity matrix with size of two. According to the LS approach, the position vector $\boldsymbol{p}$ that minimizes D_k, denoted as $\hat{\boldsymbol{p}}_k$, is defined as

$$\hat{\boldsymbol{p}}_k = \arg\min_{\boldsymbol{p}} D_k. \tag{3}$$

To find the optimal vector $\hat{\boldsymbol{p}}_k$, D_k is partially differentiated concerning $\boldsymbol{p}$, and the solution with the minimum value D_k is formulated as

$$\sum_{j=1}^{4} \left(\boldsymbol{I}_2 - \boldsymbol{n}_{j,k}\boldsymbol{n}_{j,k}^{\mathsf{T}}\right) (\boldsymbol{a}_j - \hat{\boldsymbol{p}}_k) = \boldsymbol{0}. \tag{4}$$

Denoting matrix and vector as

$$\boldsymbol{R}_k = \sum_{j=1}^{4} \left(\boldsymbol{I}_2 - \boldsymbol{n}_{j,k}\boldsymbol{n}_{j,k}^{\mathsf{T}}\right), \tag{5}$$

$$\boldsymbol{q}_k = \sum_{j=1}^{4} \left(\boldsymbol{I}_2 - \boldsymbol{n}_{j,k}\boldsymbol{n}_{j,k}^{\mathsf{T}}\right) \boldsymbol{a}_j, \tag{6}$$

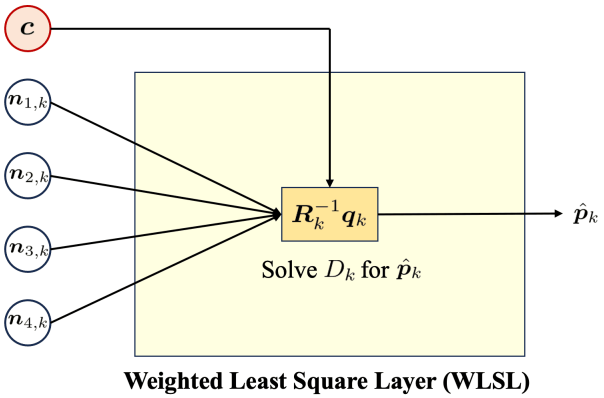


Fig. 3: Schematic of WLM-aided positioning algorithm.

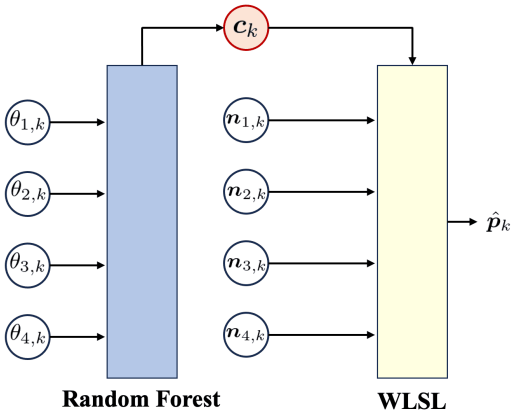


Fig. 4: Schematic of RF-aided positioning algorithm.

(4) is simply expressed as

$$\hat{\boldsymbol{p}}_k = \boldsymbol{R}_k^{-1}\boldsymbol{q}_k. \tag{7}$$

However, LS is vulnerable to uncertainties in the observed values and cannot minimize the mean square error (MSE), *i.e.*, the variance of the estimate [15].

WLS is an effective method to address this problem, and it utilizes

$$\boldsymbol{R}_k = \sum_{j=1}^{4} c_j \left(\boldsymbol{I}_2 - \boldsymbol{n}_{j,k}\boldsymbol{n}_{j,k}^{\mathsf{T}}\right), \tag{8}$$

$$\boldsymbol{q}_k = \sum_{j=1}^{4} c_j \left(\boldsymbol{I} - \boldsymbol{n}_{j,k}\boldsymbol{n}_{j,k}^{\mathsf{T}}\right)\boldsymbol{a}_j, \tag{9}$$

instead of (6). Here, $c_j \in [0,1]$ is a weight coefficient for each AoA, and $c_j = 0$ means that RXj is not used in the calculation. Appropriately adjusting c_j can improve positioning accuracy; however, the adjustment method has not been sufficiently studied yet. If the statistical model of the observations is well-formulated, minimum MSE (MMSE) filtering should be applied. Otherwise, it is difficult to determine the optimal weight coefficients analytically. Therefore, this paper proposes a novel positioning algorithm that adjusts the weight c_j in a data-driven manner using ML models.

III. Machine Learning-Aided Positioning

A. Weight learning model

Fig. 3 is a schematic diagram of the positioning algorithm using WLM. In this algorithm, the weight vector $\boldsymbol{c} = [c_1, c_2, c_3, c_4]$ are trainable parameters, and the positioning accuracy is improved by using the optimal vectors obtained through the ML-based search. ML involves a training phase, during which the model is trained using training and validation data, and a test phase, during which the model is evaluated using test data. Yielding the unit direction vectors $\boldsymbol{n}_{j,k}$ into WLS layer (WLSL), the estimated position vector $\hat{\boldsymbol{p}}_k$ is calculated by (7) with the assistance of weight vector $\boldsymbol{c}$.

To adjust the weight $\boldsymbol{c}$, supervised learning is conducted in the training phase to minimize MSE loss function, which is expressed as

$$E = \frac{1}{K}\sum_{k=1}^{K} \|\hat{\boldsymbol{p}}_k - \boldsymbol{t}_k\|^2. \tag{10}$$

The gradient of $\boldsymbol{c}$ based on the error function is then calculated using the backpropagation method. The weights $\boldsymbol{c}$ are updated by an optimization algorithm using the obtained gradients. In the test phase, the weights are fixed to the optimized vectors obtained in the training phase.

B. Random forest

Fig. 4 shows a schematic diagram of the positioning algorithm using RF. The major difference of RF from the weight vector $\boldsymbol{c}$ of WLM is that the weight vector $\boldsymbol{c}_k = [c_{1,k}, c_{2,k}, c_{3,k}, c_{4,k}]$ can be adjusted for each data point k. RF is a ML model consisting of many decision trees. In this study, 100 decision trees are prepared. When $\boldsymbol{\theta}_k$ is yielded into the RF, each decision tree locally predicts weight vector, and the resultant output weight vector $\boldsymbol{c}_k$ of RF is selected by taking the majority vote of these local predictions.

Here, $c_j \in \{0,1\}$ allowing the RF to classify whether each receiver is used ($c_j = 1$) or not ($c_j = 0$) for each $\boldsymbol{\theta}_k$. As c_j is a binary value, there are 16 $\left(2^4\right)$ candidates of weight vector $\boldsymbol{c}$. However, considering a constraint that at least two receivers are required for calculating the two-dimensional estimated position vector $\hat{\boldsymbol{p}}_k$, five cases where three or more $c_{j,k}$ are 0 are excluded, resulting in 11 candidate vectors: $\boldsymbol{c} \in \{[0,0,1,1],[0,1,0,1],[0,1,1,0],[0,1,1,1],[1,0,0,1],[1,0,1,0],[1,0,1,1],[1,1,0,0],[1,1,0,1],[1,1,1,0],[1,1,1,1]\}$.

For preparing labeled training data $\tilde{\boldsymbol{c}}_k$ of RF,

$$\tilde{\boldsymbol{c}}_k = \arg\min_{\boldsymbol{c}} \|\hat{\boldsymbol{p}}_k - \boldsymbol{t}_k\| \tag{11}$$

is found among 11 candidates, then dataset $(\boldsymbol{\theta}_k, \tilde{\boldsymbol{c}}_k)$ is provided into RF for supervised training.

In the training phase, the model parameters within RF are updated to ensure that the RF output vector $\boldsymbol{c}_k$ matches the labeled training data $\tilde{\boldsymbol{c}}_k$. In the test phase, the RF output vector

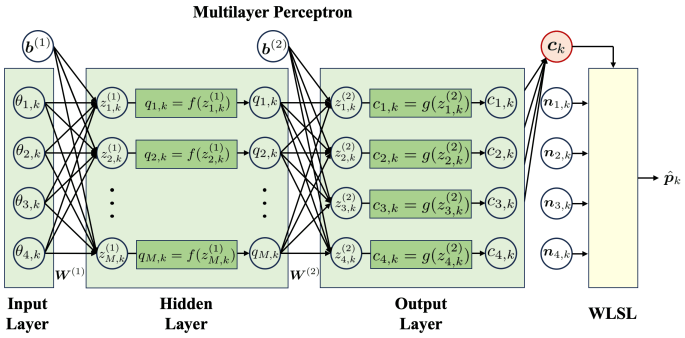


Fig. 5: Schematic of MLP-aided positioning algorithm.

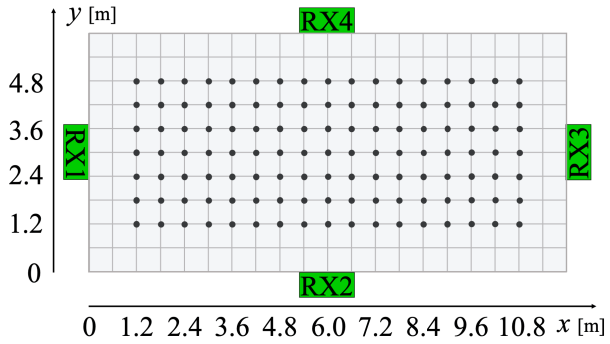


Fig. 6: Positioning environments.

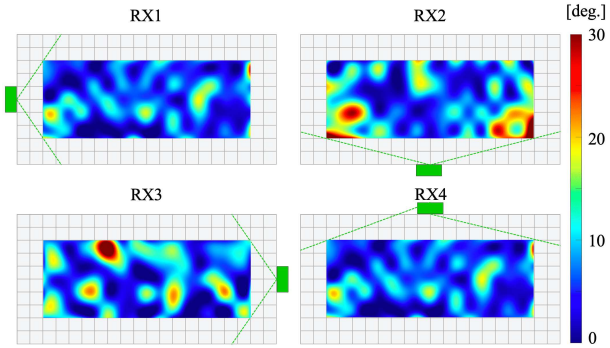


Fig. 7: Heatmaps of AoA measurement errors.

$\boldsymbol{c}_k$ obtained by yielding $\boldsymbol{\theta}_k$ into the trained RF and the unit direction vectors of the AoA $\boldsymbol{n}_{j,k}$ are provided into the WLSL for position estimation.

C. Multi-layer perceptron

Fig. 5 is a schematic diagram of the positioning algorithm using MLP. The MLP considered consists of an input layer, a hidden layer, and an output layer. Here, we define a weight matrix $\boldsymbol{W}^{(1)}$ of size $M \times 4$. The (j,m)-th element $w_{j,m}^{(1)}$ of $\boldsymbol{W}^{(1)}$ is the weight between the m-th node of the hidden layer and the j-th node of the input layer. Additionally, $\boldsymbol{b}^{(1)} = [b_1^{(1)}, \ldots, b_m^{(1)}, \ldots, b_M^{(1)}]^\mathsf{T}$ is a bias vector of size $M \times 1$. When $\boldsymbol{\theta}_k$ is yielded into the input layer, the input $z_{m,k}^{(1)}$ to the hidden layer is calculated as

$$z_{m,k}^{(1)} = \sum_{j=1}^{4} w_{m,j}^{(1)} \theta_{j,k} + b_m^{(1)}. \tag{12}$$

Thus, the input $\boldsymbol{z}_k^{(1)} = [z_{1,k}^{(1)}, \ldots, z_{m,k}^{(1)}, \ldots, z_{M,k}^{(1)}]^\mathsf{T}$ to the hidden layer is expressed as

$$\boldsymbol{z}_k^{(1)} = \boldsymbol{W}^{(1)} \boldsymbol{\theta}_k + \boldsymbol{b}^{(1)}. \tag{13}$$

$\boldsymbol{z}_k^{(1)}$ is then input to the activation function as

$$\boldsymbol{q}_k = f\left(\boldsymbol{z}_k^{(1)}\right) = \max\left(0, \boldsymbol{z}_k^{(1)}\right), \tag{14}$$

where the activation function $f(\cdot)$ of the hidden layer is the ReLU function.

In MLP, linear processing with weights and biases and nonlinear processing with activation functions are performed alternately, as described above. The input to the output layer is $\boldsymbol{z}_k^{(2)} = \boldsymbol{W}^{(2)} \boldsymbol{q}_k^{(1)} + \boldsymbol{b}^{(2)}$, where $\boldsymbol{W}^{(2)}$ and $\boldsymbol{b}^{(2)}$ are the weight matrix and bias vector between the hidden and the output layers, respectively. By using the sigmoid function as the activation function of the output layer $g(\cdot)$, the weights $\boldsymbol{c}_k$ are output as a vector with each component taking a value between 0 and 1 as

$$\boldsymbol{c}_k = g\left(\boldsymbol{z}_k^{(2)}\right) = \frac{1}{1 + \exp\left(-\boldsymbol{z}_k^{(2)}\right)}, \tag{15}$$

where the exponential function $\exp(\cdot)$ is extended to a vector-valued function. The obtained weights $\boldsymbol{c}_k$ and unit direction vectors of the AoA are input into the WLSL, and the estimated position coordinates $\hat{\boldsymbol{p}}_k$ are obtained.

In the training phase, similar to the explanation in Subsect. 3.A, the weights $\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}$ and biases $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}$ between each layer of the MLP are updated iteratively to minimize the error function. In the test phase, positioning is performed by yielding $\boldsymbol{\theta}_k$ into the MLP, which has fixed weights and biases optimized in the training phase, and the obtained weights and unit direction vectors of the AoA into WLSL.

IV. Simulation Result

A. Simulation setup

This paper evaluates the proposed method using a publicly available dataset created in [14]. This dataset contains AoA data collected using four receivers (RX1 to RX4) and one transmitter (TX) in the room space of (12, 6, 3.1) meters, as shown in Fig. 6. To ignore the effects of human bodies, the transmitter and receivers were mounted on tripods and fixed at heights of 2.3 meters and 1.1 meters from the floor, respectively. In addition, the orientation of TX was fixed during the AoA measurements to eliminate the influence of its directionality and installation direction. The grid spacing on the floor was 60 cm, and the transmitter remained stationary for one minute at each of the 119 points indicated by black plots. The transmit power was set to 0 dBm, and the direction-finding signal transmission period was 50 Hz. However, due to packet loss, the AoA measured by the receivers is approximately 20

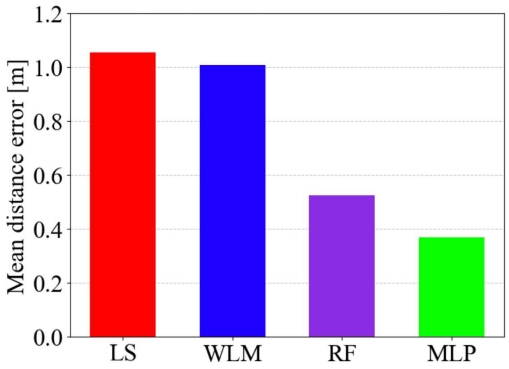


Fig. 8: MDE of estimated position coordinates.

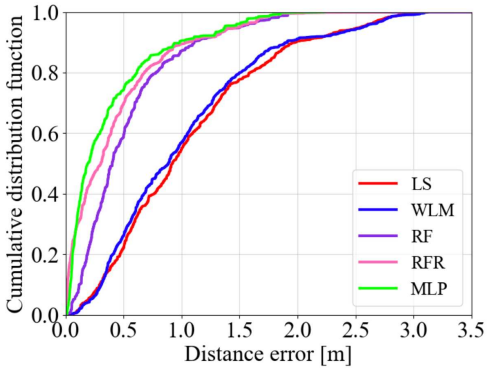


Fig. 9: CDF of DE in estimated position coordinates.

to 30 samples per second. As mentioned in Sect. II, this paper uses only data from channel 37 advertising channels; therefore, channel 37 data were extracted from the entire dataset. Furthermore, because the receivers were not time-synchronized during AoA measurements, temporal synchronization was performed on the post-measurement data by taking the median every second. Consequently, with $\boldsymbol{\theta}_k$ considered one sample, 60 samples were obtained per point, resulting in 7,140 samples across all data points (*i.e.*, $K = 7,140$). The entire dataset was randomly divided into training (50 %), validation (25 %), and test (25 %) data. The ML models were trained using the training and validation data, and then each method was evaluated using the test data.

Fig. 7 shows heatmaps of each receiver's mean absolute error (MAE) of the AoA. The MAE ranges from a minimum of approximately $0°$ to a maximum of approximately $60°$, and positioning accuracy is expected to decrease in areas with high MAE. RX2 and RX4, installed along the long sides, show a common trend of increasing MAE as the distance from the center increases, whereas RX1 and RX3, installed along the short sides, do not show a common trend. This analysis suggests that the characteristics of the AoA cannot be predicted based on the installation positions of the receivers. The proposed method aims to improve positioning accuracy by adjusting the weights of the WLS method using a ML model trained on complex AoA characteristics. The following subsection discusses the evaluation results of the conventional and proposed methods using test data.

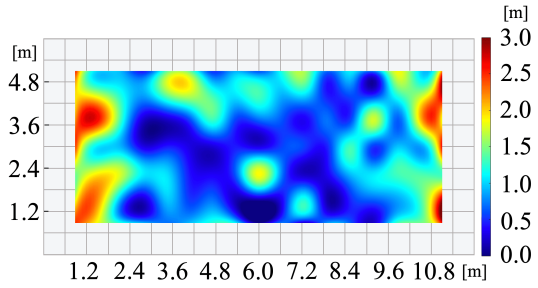
B. Evaluation using test data

The positioning accuracy of the conventional LS algorithm and the proposed methods using WLM, RF, and MLP is evaluated from the perspective of the distance error (DE) $|\hat{\boldsymbol{p}}_k - \boldsymbol{t}_k|$.

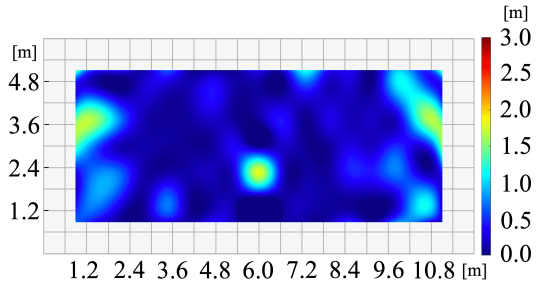
Fig. 8 shows the mean DE (MDE) for all 1,785 test data samples. The MDE of LS is 1.05 m, WLM is 1.01 m, RF is 0.52 m, and MLP is 0.37 m, indicating improved positioning accuracy for all proposed methods compared with conventional LS. MLP achieved about an approximately 65 % improvement in accuracy compared with the LS method. Moreover, the MDE follows the order of MLP <RF <WLM, indicating that RF is more accurate than WLM, and that MLP is more precise than RF. In the test phase, WLM has a lower degree of freedom in weight adjustment than the other two methods, as the weights are fixed over K samples. RF has a lower degree of freedom in weight adjustment than MLP because the possible values of each weight component are only 0 or 1. From this perspective, it is inferred that higher degrees of freedom in weight adjustment lead to improved positioning accuracy.

Fig. 9 shows cumulative distribution function (CDF) characteristics of DE for each method. While the BLE-AoA method is capable of centimeter-level positioning, only 55 % of the conventional LS methods resulted in less than 1 m DE. In comparison, 91 % of MLP and 87 % of RF have DE below 1 meter, demonstrating stable centimeter-level positioning. However, the maximum DE is about 2 to 3 m: 3.08 m for LS, 2.80 m for RF, and 2.23 m for MLP. Addressing these outliers remains a challenge for future works. On the other hand, in the figure, RFR represents a method using a Random Forest Regressor (RFR), where $c_{j,k}$ can take values of $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ as a simple extension of RF, increasing the number of weight candidates in RF to 1,225. As the number of weight candidates increases, improvements in CDF characteristics are observed, asymptotically approaching the performance of MLP. This result confirmed that, as mentioned earlier, the degree of freedom in weight adjustment is crucial for improving positioning accuracy. Furthermore, increasing the number of weight candidates in RF is expected to approach positioning accuracy comparable to MLP. However, as the number of weight candidates increases, the computational complexity required to label the training data increases exponentially, making MLP preferable considering the time cost.

Fig. 10 shows the DE heatmaps within the positioning area for LS and MLP. This figure demonstrates how the positioning accuracy improves at each point. While MLP shows an overall improvement in positioning accuracy, points with high DE in LS also have high DE in MLP. Similar trends were observed for WLM and RF. As ML learns the characteristics of the entire dataset, it may generalize poorly to data that deviate from the overall trend. Fig. 7 suggests that AoA measured in regions with significant positioning errors in LS likely has different



(a) Least square



(b) MLP-aided positioning algorithm

Fig. 10: Heatmaps of DE in estimated position coordinates.

characteristics from the other areas. The ML model might have generalized poorly to such data. Future studies should focus on generalizing the ML model to the entire area using various methods, such as increasing the number of datasets used for training and integrating the positioning results of multiple ML models.

V. Conclusions

In this paper, we proposed a positioning algorithm that uses ML to adjust the weights of the WLS method in a data-driven manner to improve the positioning accuracy of the BLE-AoA positioning method. Our research has significantly improved positioning accuracy across all three proposed methods. This confirms the effectiveness of integrating ML into the positioning algorithm. Additionally, it was found that the degree of freedom in weight adjustment is crucial for improving positioning accuracy. Acquiring generalization performance for the entire positioning area remains a challenge for the future.

Acknowledgment

A part of this work was financially supported by JSPS KAKENHI Grant Number JP23K20935, Japan.

References

[1] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019.

[2] A. Koutris, T. Siozos, Y. Kopsinis, *et al.*, "Deep learning-based indoor localization using multi-view BLE signal," *Sensors*, vol. 22, no. 7, p. 2759, 2022.

[3] D. Dardari, P. Closas, and P. M. Djurić, "Indoor tracking: Theory, methods, and technologies," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 4, pp. 1263–1278, 2015.

[4] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, 2007.

[5] C. Huang, Y. Zhuang, H. Liu, J. Li, and W. Wang, "A performance evaluation framework for direction finding using BLE AoA/AoD receivers," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3331–3345, 2021.

[6] L. Klinkusoom, T. Sanpechuda, K. Maneerat, K. Chinda, and L.-O. Kovavisaruch, "Accuracy comparison of Bluetooth Low Energy indoor positioning system based on measurement techniques," in *2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 2022, pp. 1–5.

[7] K. Zhao, T. Zhao, Z. Zheng, *et al.*, "Optimization of time synchronization and algorithms with TDOA based indoor positioning technique for Internet of Things," *Sensors*, vol. 20, no. 22, p. 6513, 2020.

[8] P. Mohaghegh, A. Boegli, and Y. Perriard, "Bluetooth Low Energy direction finding principle," in *2021 24th International Conference on Electrical Machines and Systems (ICEMS)*, 2021, pp. 830–834.

[9] Y. Zhuang, C. Zhang, J. Huai, Y. Li, L. Chen, and R. Chen, "Bluetooth localization technology: Principles, applications, and future trends," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23 506–23 524, 2022.

[10] G. Pau, F. Arena, Y. E. Gebremariam, and I. You, "Bluetooth 5.1: An analysis of direction finding capability for high-precision location services," *Sensors*, vol. 21, no. 11, p. 3589, 2021.

[11] A. Yassin, Y. Nasser, M. Awad, *et al.*, "Recent advances in indoor localization: A survey on theoretical approaches and applications," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 1327–1346, 2017.

[12] A. Khan, S. Wang, and Z. Zhu, "Angle-of-Arrival estimation using an adaptive machine learning framework," *IEEE Communications Letters*, vol. 23, no. 2, pp. 294–297, 2019.

[13] N. Paulino and L. M. Pessoa, "Self-localization via circular Bluetooth 5.1 antenna array receiver," *IEEE Access*, vol. 11, pp. 365–395, 2023.

[14] M. Girolami, F. Furfari, P. Barsocchi, and F. Mavilia, "A Bluetooth 5.1 dataset based on Angle of Arrival and RSS for indoor localization," *IEEE Access*, vol. 11, pp. 81 763–81 776, 2023.

[15] S. O. Haykin, *Adaptive Filter Theory*, 5th. Hoboken: Pearson, 2013.

Observation of the Terrestrial Radio Environment Using the Low Earth Orbit Satellite Constellation

Takatoshi OBATA* and Osamu TAKYU* and Kei INAGE† and Takeo FUJII‡ and
Kohei YOSHIDA§ and Masayuki ARIYOSHI§
* Shinshu University, Japan
E-mail: takyu@shinshu-u.ac.jp Tel/Fax: 026-269-5255
† Tokyo Metropolitan College of Industrial Technology, Japan
‡ The University of Electro-Communications, Japan
§ NEC Corporation, Japan

***Abstract*—In recent years, spectrum sharing is one of the technologies to solve the depletion of spectrum resources due to the increasing demands for wireless communications. In spectrum sharing using a database, there is a method to use radio wave environmental data in order to improve the accuracy of interference calculation. When making the radio wave environmental data into statistics, it is necessary to measure the received power intensity at each position. However, this is difficult for a large area. Therefore, in this paper, the concept of a method to measure the received electric field intensity on the ground using a satellite constellation is proposed and examined.**

I. Introduction

In recent years, due to the increasing demand for wireless communications, the depletion of spectrum resources has become a problem. As a solution to this problem, spectrum sharing technology is attracting attention. In particular, dynamic spectrum access is a technology in which multiple systems use one frequency band by dividing time and space. This makes it possible to effectively use free spectrum resources. At present, it is being examined, tested, implemented, and put into practical use in countries around the world [1],[2],[3]. In dynamic spectrum access, there are existing systems with the highest priority and secondary systems that can use frequencies only in the temporal and spatial areas not used by existing systems. The secondary systems must not cause radio interference to existing systems and require advanced interference control. There are two main types of control methods to prevent interference with existing systems: the sensing method and the database method [4].

In the database method assumed in this paper, the primary system registers the utilization information in the database, performs interference calculation based on the information, and the secondary system uses the vacant resources in time and space that are available for secondary use. When the secondary system is a mobile station, accumulated interference occurs due to simultaneous radiation of multiple radio stations, and the interference power increases. Therefore, there is a concern that unexpected interference may occur and the utilization efficiency of spectrum resources may decrease due to excessive radio wave stop. Therefore, attention is paid to a method called radio map, which measures the received power value for each place and averages it for each two dimensional mesh to make statistics [5]. It is possible to improve the accuracy of interference calculation by utilizing this radio map and acquiring the radio wave environment unique to each place. However, at present, the measurement of a wide range is difficult because the measurement of the radio map is done by hand. Furthermore, it is also possible to contribute to the accuracy improvement of interference calculation by measuring the position of the radio wave source and the transmission power in real time.

The authors have proposed a system model for acquiring terrestrial radio wave environments from space far above the ground by utilizing satellite constellation in order to measure a wide range in a short time, and clarified the spatial resolution [6]. By using the proposed method, it was shown that the terrestrial radio signal strength could be observed stably and with high accuracy in comparison with the case in which the terrestrial radio signal strength was observed only by a single satellite. The spatial resolution when the beam was irradiated toward the ground from the satellite was clarified by simulation. In this paper, the proposed method is analyzed in detail, and the effectiveness of the proposed method is shown more clearly.

II. Proposed System Model

This section describes a system model and a proposed method for remote sensing of terrestrial radio signal strength from a satellite.

The system model is shown in Fig. 1. This system model consists of the combination of two proposed methods, i.e., the measurement method and the estimation method. As a premise, it is assumed that the satellites to be used are multiple satellites orbiting in low Earth orbit at an altitude of 300 [km] and constitute a low Earth orbit satellite constellation. Each satellite is equipped with an antenna that generates an ideal single beam with high gain, and the beam direction and reception time can be adjusted arbitrarily.

In a low-Earth orbit satellite such as the one assumed in this paper, the beam irradiated from the satellite does not spread on the ground surface and is directed to a specific narrow range. Even if the beam width is widened by reducing the size of the antenna, the resolution can be increased by narrowing the coverage range on the ground surface. Therefore, it is possible to reduce the size and cost of the antenna mounted on the satellite, which leads to the launching of many low-cost satellites. In addition, in a constellation, simultaneous observation by multiple satellites becomes possible, and the degree of freedom in measurement increases.

A. Proposed measurement method

As for the beam irradiation method, multiple satellites within the range of visibility from a certain ground point are used from among the satellites constituting the constellation. The beams are simultaneously directed to the same point from these satellites and the received power values are measured by each satellite. The received power values measured by each satellite is converted from logarithmic values to true values and then averaged. The averaged values are used as the received power value measured by the satellites and used to estimate the terrestrial radio signal strength. As a result, even if there is interference from outside the beam coverage, the effect can be suppressed. In addition, when only one satellite is used, the measured value varies due to fading, etc. Therefore, by using this proposed measurement method, beams can be irradiated from multiple directions, and improvement of measurement accuracy and stabilization of accuracy can be realized by space diversity.

B. Proposed estimation method

The received power values measured by the satellite does not become the terrestrial radio signal strength as it is. For the estimation, two methods were examined: the method using the free-space path loss and the method using the anchor.

First, a method using free-space path loss is described. The attenuation between the satellite and the ground is estimated by free-space path loss by distance. Since the distance from the altitude of the orbit around the satellite to the ground is known, the free-space path loss is calculated based on the distance. The antenna gain by the beam generated by the satellite is subtracted from the propagation attenuation value. This is used as a correction value, and the averaged received power value by the measurement method proposed in the preceding paragraph is added to it to estimate the terrestrial radio signal strength.

Second, an anchor, which is a transmission point whose transmission power and position are known, is separately arranged. A beam is generated from the satellite to the anchor in advance, and the propagation attenuation of the propagation path is measured. The transmission power of the anchor is subtracted from the measured value when the beam is directed at the anchor. This is used as a correction value, and the averaged received power value by the measurement method proposed in the preceding paragraph, as in the case of using free-space path loss, is added to it to estimate the terrestrial radio signal strength.

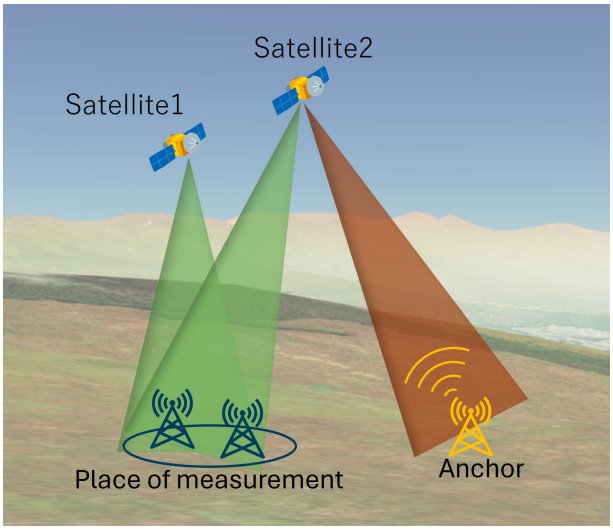


Fig. 1. Proposed system model

III. Simulation Overview

Understanding of the beam coverage and the estimation accuracy of the proposed estimation method were evaluated by the ray tracing simulation. The outline of the simulation is shown in Fig. 2, and the specifications are shown in Table I. The radio wave propagation condition was analyzed by setting the simulation range around Maeyama, Saku City, Nagano by WirelessInSite 3.4.4.12. In this paper, in order to reduce the calculation amount of the simulation software, the transmission and reception are reversed by utilizing the reversibility of the radio wave propagation. That is, originally, the terrestrial terminals become the radio wave source, and the satellite receives the radio wave. However, in this simulation, the satellite becomes the transmitter and transmits the radio wave toward the ground, and then the terrestrial terminals receive the radio wave from the satellite. As for the measured value, the received power value obtained by each terrestrial terminal is used as the received power value of the satellite when each terrestrial terminal transmits. The following estimated value is the value calculated by estimating the transmission power of the terrestrial terminals using the proposed system in the state before the transmission and reception are reversed.

The assumed satellite constellation orbits 300 [km] above the ground, and the antenna generates an ideal single beam with a half-power beamwidth of 5 degrees and a maximum gain of 30 [dBi]. Two satellites are placed, and all satellites generate beams toward one point. Satellite1 is placed directly above the beam irradiation center, and this satellite generates beams vertically from itself. Satellite2 is placed 5 [km] north of satellite1, and the beam is tilted 0.955 degrees south. This causes the beam to face the beam irradiation center. Since transmission and reception are reversed, the radio wave

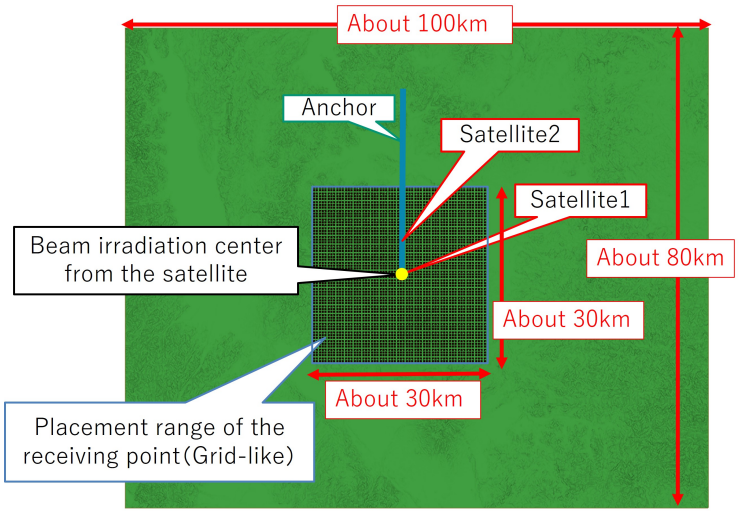


Fig. 2. Simulation Overview

TABLE I
SIMULATION DATA

Item	Value
Receiver(Terrestrial terminals)	
Center frequency	3650 [MHz] (3.7 [GHz] band)
Bbandwidth	100 [MHz]
Directivity	Omni-directional
Arrangement	500 [m] grid
Transmitter(Satellite)	
Center frequency	3650 [MHz] (3.7GHz band)
Bbandwidth	100 [MHz]
Transmit power	0 [dBm]
Directivity	Ideal single beam
Antenna maximum gain	30 [dBi]
Half-Power beamwidth	5 degrees
Altitude	300 [km]
Polarization	Left-hand circular
Channel Model	
Propagation model	X3D
Number of reflections	3
Number of transmissions	0
Number of diffractions	1

from the satellite uses a Left-hand circular polarized wave. As for the terrestrial receivers, the receivers are uniformly placed at 500 [m] intervals within a range of about 30 [km] $\times$ about 30 [km] square around the irradiation point of the beam. In this paper, in order to grasp the beam coverage and clarify the estimation accuracy for each position, there are no multiple transmitters and there is no interference. For each receiver, measurement using averaging and estimation by the two methods shown in the proposed method are performed. In order to perform estimation using anchors, anchors are placed linearly at 500 [m] intervals north of the beam irradiation center as shown in Fig. 2.

IV. Simulation Result

This section presents the simulation results separately for the proposed measurement method and the proposed estimation method.

A. Proposed measurement method

Fig. 3 and Fig. 4 show the absolute value of the difference between the estimated value by the measurement method and the correct value when the estimation method using free-space propagation attenuation is applied. The correct value is the value actually transmitted by the transmitter. Fig. 3 shows the result of applying the proposed estimation method of taking the average value of two satellites for each receiver. Fig. 4 shows the result of using the value measured by one satellite for each receiver. Each point of the heat map is the position of the receiver placed on the ground, and the error between the actual transmission power and the estimated value is shown in color. The value is a positive number because the absolute value of the error is taken in this heat map. The darker the blue, the smaller the error, and the darker the red, the larger the error. The range is the smallest at the deepest blue, with an error of 0 [dB], and the largest at the deepest red, with an error of 20 [dB]. Comparing these figures, when the average value of two satellites is taken, the points where the error is large are reduced, and the color changes relatively gently from the center to the periphery. This simulation is conducted in a place where the undulation of the landform is severe, but the change according to the landform is not observed.

Fig. 5 shows the relationship between the distance from the beam irradiation center and the estimation error. The error is plotted on the vertical axis and the distance from the beam irradiation center at each terrestrial receiver is plotted on the horizontal axis. This is not the absolute value of the difference from the correct value, but the value obtained by subtracting the correct value from the estimated value. Fig. 5 shows that the error increases as the distance from the beam irradiation center increases for both the case where the average of two satellites is taken and the case where only a single satellite is taken. However, the magnitude of the error is still less than 14 [dB] even at a distance of about 20000 [m]. It is also shown that the error fluctuation is suppressed by taking the average of two satellites.

From these, the spread of the error in generating the beam with the directivity from the satellite which is 300 [km] up to the ground was able to be confirmed. Since the change of the error by the undulation of the landform is not observed, it is considered that the effect of the landform on the error is slight. From the heat map of Fig. 3 and Fig. 4 and the graph of Fig. 5, the fluctuation of the error is suppressed in the proposed measuring method. Therefore, it is possible to improve the stability and accuracy of the measured value by measuring by multiple satellites simultaneously.

B. Proposed estimation method

Fig. 6 shows the heat map of the estimation method using the anchor. The heat map of Fig. 6 was prepared under the

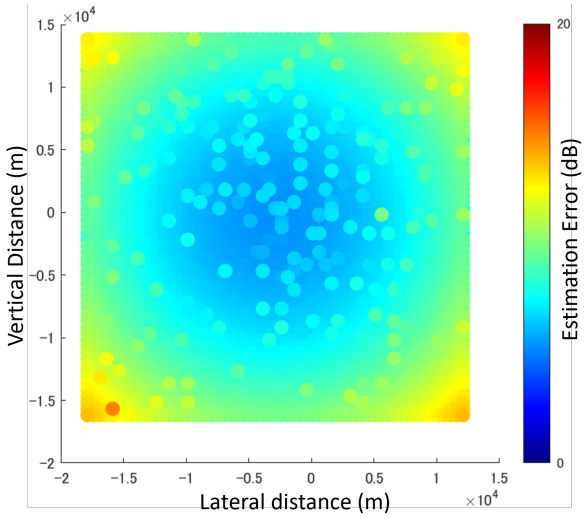


Fig. 3. Heat map of estimation error when using the average value of two satellites(Free-Space Path Loss)

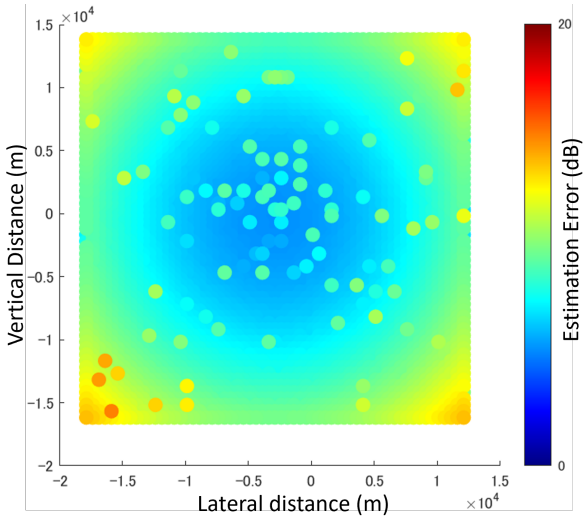


Fig. 4. Heat Map of Estimation Error Using Received Power Value of Only One Satellite(Free-Space Path Loss)

same conditions as Fig. 3 except for the estimation method. In this fig. 6, the anchor was placed at the beam irradiation center of the satellite. In other words, since the anchor is placed at the position where the measurement is desired, the received power on the ground surface is estimated with accurate propagation attenuation information obtained. Comparing Fig. 6 with Fig. 3, it can be seen that the error of the estimation method using anchors is overwhelmingly small.

In Fig. 7, the distribution of the error due to the distance from the center of the beam irradiation is shown for each estimation method. As for the estimation using the anchor, as in Fig. 6, the anchor position is at the center of the beam irradiation. This shows that the estimation using the anchor has about 5 [dB] less error than the estimation using the free-space

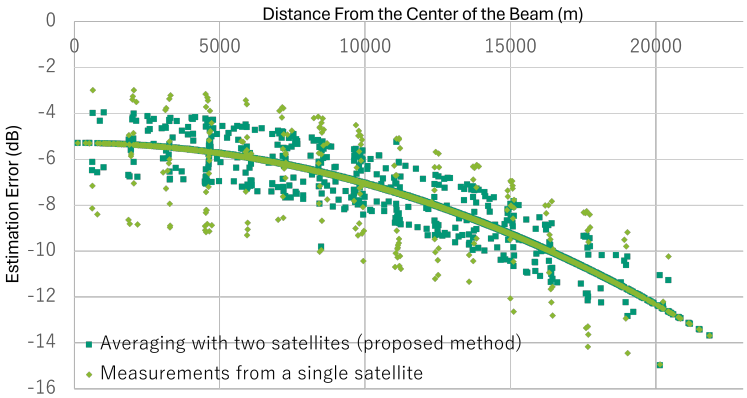


Fig. 5. Relationship between the distance from the center of the beam and the estimation error according to the number of satellites used

path loss.

Fig. 8 shows the change in the heat map when the position of the anchor is moved away from the beam irradiation center. As shown in Fig. 2, multiple anchors are arranged in a straight line toward the north from the beam irradiation center. Therefore, this time, the heat map was created by selecting only one ground reception point to be used as an anchor at every 1500m toward the north from the beam center. From these figures, it can be seen that the position where the error is small spreads outward as the distance from the beam irradiation center increases, and it changes outward in a doughnut shape when the distance exceeds 6000 [m].

As a result, in the estimation method using the anchor, the accuracy was greatly improved in the estimation in the most ideal condition in which the anchor was placed at the position to be measured, compared with the estimation method using the free-space path loss, and it was possible to confirm the distribution change of the error by the position of the anchor from Fig. 8.

V. Summary

A system model and estimation method for grasping the terrestrial radio wave environment from the satellite are proposed. Visualization of the beam coverage for each estimation method and accuracy analysis of the measurement method are carried out by simulation. It is shown that the proposed measurement method can suppress the fluctuation of the error in comparison with the case in which one satellite is used, and that stable and high-precise measurement is possible. Comparison between the proposed 2 kinds of estimation methods and the change of the error by the position of the anchor are shown. In the future, analysis in the case in which multiple terrestrial terminals are simultaneously operated and narrowing of the coverage and high-precision are aimed by combining multiple beams.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP24K00881.

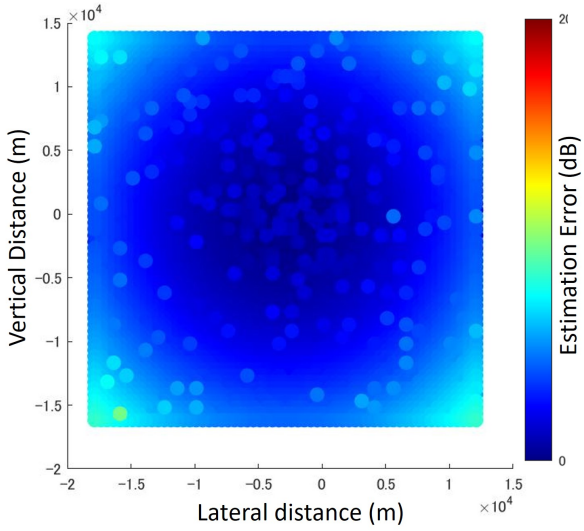


Fig. 6. Heat map of estimation error when using the average value of two satellites(Using an anchor)

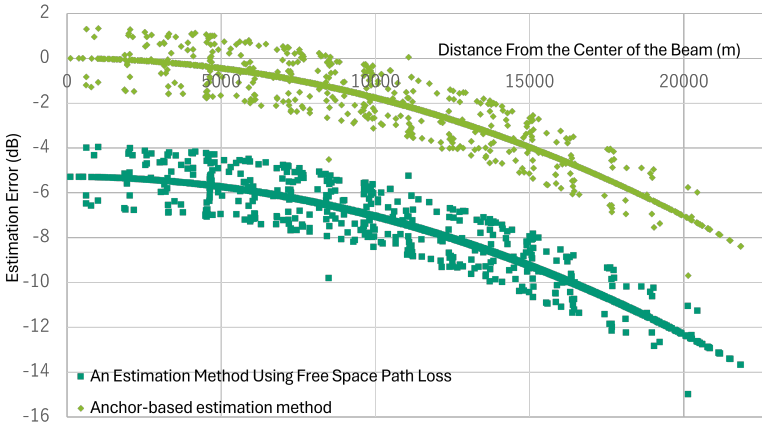


Fig. 7. Comparison of estimation methods for distance from center of the beam and estimation error

REFERENCES

[1] CSMAC (2013). Final report: Working group 1 – 1695–1710 MHz Meteorological-Satellite. Tech. rep., Commerce Spectrum Management Advisory Committee. (”WG-1 Report”)

[2] Ryo SAWAI, ”Current status and its future advanced solutions of database control & management in dynamic spectrum sharing/access techn”, IEICE Technical Report SR2020-46(2020-11) , 2020 , p152-158

[3] P. Amirshahi, et al.,”Comparison of Centralized and Distributed Spectrum Monitoring Methods for Enabling Spectrum Sharing Between Weather Satellites and Terrestrial Networks”, DySPAN, Dec.2019

[4] Takatoshi Obata, Osamu Takyu, ”Radio Sensor Detection of Interference to Satellite Earth Station in Frequency Spectrum Sharing”, International Conference on Ubiquitous and Future Networks 2023 (ICUFN 2023)

[5] Sunao Miyamoto, Takeo Fujii, ”Interference Power Estimation Method Using the Radio Map and the 3D Map for Spectrum Sharing”, IEICE Technical Report, vol. 121, no. 392, SR2021-93, pp. 36-41, March. 2022.

[6] Takatoshi Obata, Osamu Takyu, Kei Inage, Takeo Fujii, ”[Short Paper]Method for Sensing the Radio Wave Environments on the Ground Using satellite constellation”, IEICE Technical Report, vol. 123, no. 435, SR2023-108, pp. 66-68, March. 2024.

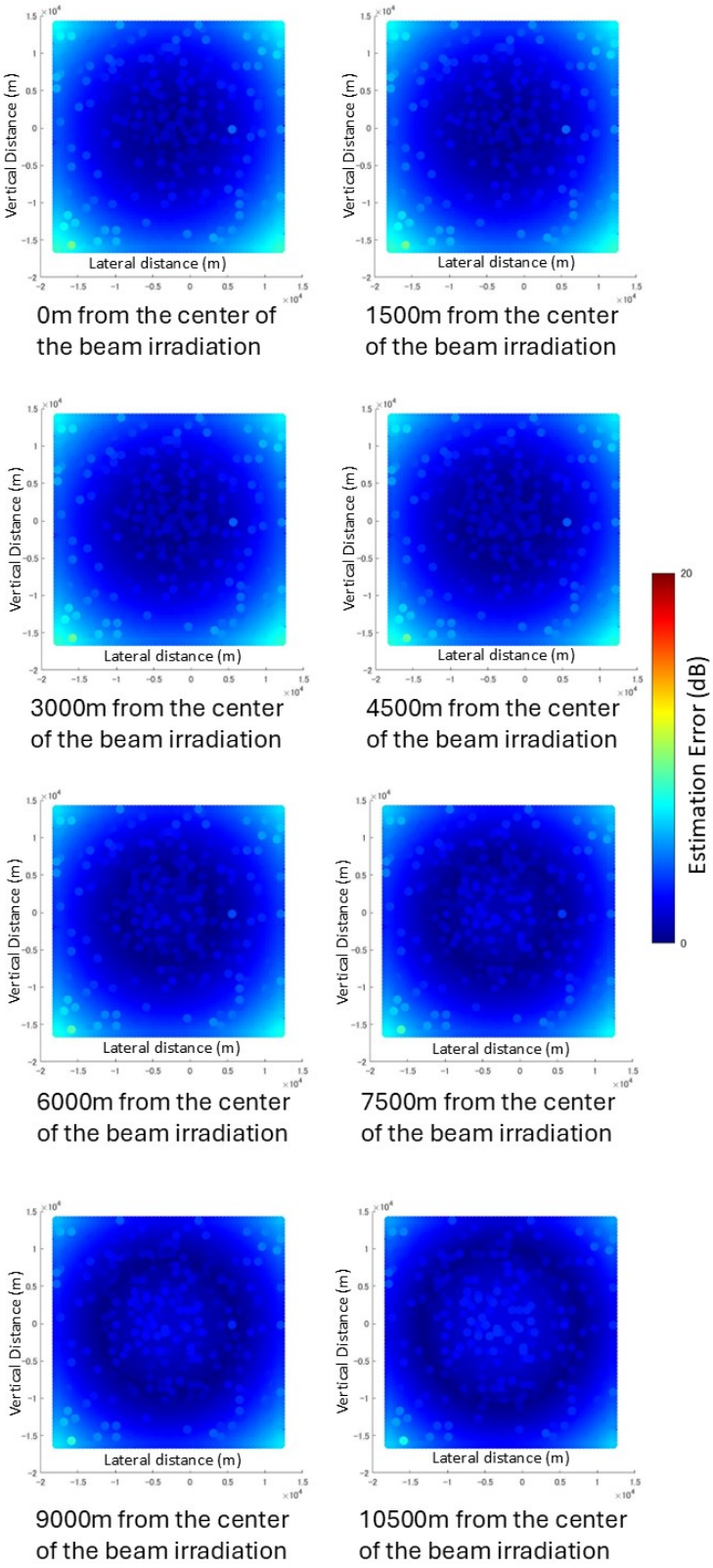


Fig. 8. Heat map changes when anchor position is moved

Deep Unfolding Aided Parameter Optimization for Multi-task Diffusion LMS Algorithm

Xiaoqing Tong* and Kazunori Hayashi†
* Kyoto University, Kyoto
E-mail: tong.xiaoqing.75d@st.kyoto-u.ac.jp
† Kyoto University, Kyoto
E-mail: hayashi.kazunori.4w@kyoto-u.ac.jp

***Abstract*—The paper proposes a novel algorithm termed TMDLMS (Trainable Multi-task Diffusion Least Mean Squares) to collaboratively estimate multiple vectors, which are assumed to have certain correlations, from observations obtained at multiple nodes in the network. Adopting the concept of deep unfolding, the proposed algorithm is obtained by unfolding iterative process of the conventional MDLMS (Multi-task Diffusion Least Mean Squares) algorithm, resulting in a multi-layered signal flow graph reminiscent of neural networks. In this structure, the parameters in the algorithm such as the step size of each layer are considered as trainable parameters, allowing for optimization via machine learning techniques, such as stochastic gradient descent (SGD). Numerical experimental results indicate that, compared to the conventional MDLMS using fixed parameters, the proposed algorithm can boast the convergence rates and achieve lower steady-state errors at the same time. The results demonstrate the validity of the proposed approach under various conditions.**

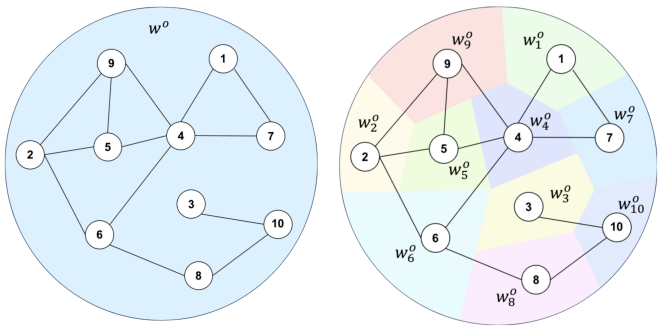
I. Introduction

In the fields of adaptive signal processing, Least Mean Squares (LMS) algorithm [1] would be the most prominent adaptive algorithm because of its simplicity and robustness. While LMS has been widely used in various applications, it utilizes its own measurement only in the update equation to obtain the estimate of the unknown vector. This leads to the development of diffusion LMS (DLMS) [2] for collaborative estimation of a common unknown vector in distributed environment. Moreover, DLMS has been extended to multi-task diffusion LMS (MDLMS) [3], which can estimate different unknown vectors but with certain correlations in collaborative manner to cope with more intricate networks. However, LMS based algorithms are known to suffer from slow rate of convergence in general.

Model-driven machine learning (ML) [4] has become a popular tool in communication systems in recent years. The concept of model-driven ML is to merge principled algorithms that have performance guarantees with tools derived from ML techniques, aiming to combine the strengths of both approaches. Deep unfolding [5] is one of powerful model-driven ML methods that fuses iterative optimization algorithms with tools from neural networks to efficiently solve various tasks in adaptive signal processing. As stated in [6], deep unfolding takes an existing iterative algorithm with a predefined number of iterations, unfolds it in the direction of time axis with the multi-layer structure, and introduces a set of trainable parameters. These parameters are then learned using deep learning techniques, utilizing suitable loss functions, stochastic gradient descent (SGD), and backpropagation. With learned parameters, the unfolded algorithm can achieve better convergence performance than that of original iterative algorithm in range of tasks. Currently, deep unfolding has been applied to various applications. For instance, hyperparameters of the iterative shrinkage thresholding algorithm (ISTA) [7] has been determined by using deep unfolding leading to the trainable iterative shrinkage thresholding algorithm (TISTA) [8]. It has been demonstrated that TISTA achieves a significantly faster convergence rate compared to the conventional ISTA.

Previous research has integrated deep unfolding with LMS based algorithms as well. Sasaki et al. combined deep unfolding with both LMS and normalized LMS (NLMS) algorithms, named TLMS and TNLMS, respectively [9]. Grounded in the concept of deep unfolding, they treated the step size as a learnable parameter within the iterative process. Numerical results have indicated that this approach could enhance convergence rates and reduce steady-state error with the TNLMS algorithm showing particularly notable performance. Similarly, NISHIHATA et al. [10] applied deep unfolding to the DLMS algorithm, which has demonstrated an improved convergence rate compared to the original DLMS by optimizing the step size and edge weights.

Inspired by the prior works, this paper applies the deep unfolding technique to the MDLMS algorithm, and proposes a new algorithm termed "Trainable Multi-task Diffusion Least Mean Squares" (TMDLMS). The algorithm is to collaboratively estimate multiple unknown vectors, which are presumed to have certain correlations, from observation vectors collected by multiple nodes in a network. We have unfolded the iterative process of the MDLMS algorithm, resulting in a multi-layered signal flow graph reminiscent of neural networks. Within this structure, parameters in the algorithm, such as the step size, are considered trainable. For optimizing these parameters, we have employed optimization algorithms such as Adam [11]. To assess the practical performance of the TMDLMS algorithm, numerical experiments have been conducted under various correlation conditions. The results have indicated that, compared to MDLMS, the proposed algorithm boasts convergence rates and achieves lower steady-state errors at the same time.



(a) (b)

Fig. 1. Examples of two types of learning problems. (a) Single-task learning problem. (b) Multi-task learning problem. [3]

II. Review of LMS, DLMS and MDLMS algorithms

Here, we briefly review LMS, DLMS and MDLMS algorithms implemented for the estimation of an unknown vector (or unknown vectors) in the network as the basis of the discussion in this paper.

A. Problem Setting

We consider a sensor network represented by an undirected graph with N nodes. The presence of an edge between any two nodes signifies the existence of a communication link between them.

Assuming that the unknown vector to be estimated by node i is $\boldsymbol{w}_i^{\mathrm{o}} \in R^{L\times 1}$, the linear measurement of the unknown vector $\boldsymbol{w}_i^{\mathrm{o}}$ observed at node i at time $k \geq 0$ is given by

$$d_i^{(k)} = \boldsymbol{u}_i^{(k)\mathsf{T}} \boldsymbol{w}_i^{\mathrm{o}} + v_i^{(k)}, \tag{1}$$

where $\boldsymbol{u}_i^{(k)} \in R^{L\times 1}$ is a vector used for the linear measurement by node i, and $v_i^{(k)}$ is an additive white Gaussian noise (AWGN) with mean of 0 and variance of σ_v^2.

Depending on the relation among $\boldsymbol{w}_i^{\mathrm{o}}$, the estimation problem of $\boldsymbol{w}_i^{\mathrm{o}}$ can be classified into a single-task learning problem and a multi-task learning problem as shown in Fig. 1:

- **Single-task learning problem:** All nodes have a common estimation target vector $\boldsymbol{w}^{\mathrm{o}}$. That is, in this case we have

$$\boldsymbol{w}_i^{\mathrm{o}} = \boldsymbol{w}^{\mathrm{o}}, \quad \forall i \in \{1, \ldots, N\}$$

- **Multi-task learning problem:** Each node i tries to estimate its own target vector $\boldsymbol{w}_i^{\mathrm{o}}$, which is different from the target vectors of other nodes. However, a certain correlation among $\boldsymbol{w}_i^{\mathrm{o}}$ is often assumed, which enables us to achieve collaborative estimation.

B. LMS Algorithm

Although the LMS algorithm is not designed for the estimation in the network environment, it can be applied for the problem by implementing the LMS algorithm at each node and simply ignoring the existence of communication links among the nodes.

In the LMS algorithm, the estimated vector $\boldsymbol{\phi}_i^{(k)}$ at time k and node i is updated according to the following equation:

$$\boldsymbol{\phi}_i^{(k+1)} = \boldsymbol{\phi}_i^{(k)} + \mu \boldsymbol{u}_i^{(k)} \left(d_i^{(k)} - \boldsymbol{u}_i^{(k)\mathsf{T}} \boldsymbol{\phi}_i^{(k)} \right), \tag{2}$$

where the initial estimate for the vector $\boldsymbol{\phi}_i^{(0)}$ is often set to $\mathbf{0}$.

C. DLMS Algorithm

The DLMS algorithm [2] can achieve a collaborative estimation for the single-task learning problem. The DLMS algorithm consists of the **LMS step** responsible for signal estimation and the **averaging step** that facilitates information exchange between nodes.

In the LMS step, the estimated vector $\boldsymbol{\psi}_i^{(k)}$ at time k and node i is updated by using its own measurement as

$$\boldsymbol{\psi}_i^{(k+1)} = \boldsymbol{\phi}_i^{(k)} + \mu \boldsymbol{u}_i^{(k)} \left(d_i^{(k)} - \boldsymbol{u}_i^{(k)\mathsf{T}} \boldsymbol{\phi}_i^{(k)} \right), \tag{3}$$

which is basically the same as the update equation of the LMS algorithm.

Each node exchanges $\boldsymbol{\psi}_i^{(k)}$ obtained in the LMS step with its neighbor nodes $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is a set of neighbors of node i including i itself. Then, the updated estimate of the vector in the averaging step is obtained at each node i by the weighted average of $\boldsymbol{\psi}_i^{(k)}$ based on the idea of the averaging consensus technique. Specifically, the update equation in the averaging step is given by

$$\boldsymbol{\phi}_i^{(k)} = \sum_{j\in\mathcal{N}_i} c_{ij} \boldsymbol{\psi}_j^{(k)}, \tag{4}$$

where c_{ij} is the weight given for $\boldsymbol{\psi}_j^{(k)}$ obtained from node j in the averaging update at node i. The choice of the weight c_{ij} has an impact on the convergence performance of DLMS. In this paper, we employ an established approach named Metropolis rule [12] given by

$$c_{ij} = \begin{cases} \frac{1}{\max(|\mathcal{N}_i|, |\mathcal{N}_j|)} & \text{if } j \in \mathcal{N}_i \text{ and } i \neq j \\ 1 - \sum_{j\in\mathcal{N}_i^-} c_{ij} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where $|\mathcal{N}_i|$ denotes the cardinality of the set $\mathcal{N}_i$, and $\mathcal{N}_i^-$ denotes the set $\mathcal{N}_i$ but without node i.

D. MDLMS Algorithm

The MDLMS algorithm has been proposed by Chen et al. in [3] for the multi-task learning problem. To be more precise, they have considered more general problem of the clustered multi-task learning problem than the multi-task learning problem, where nodes in the same cluster share a common target vector. Thus, the multi-task learning problem considered in this paper can be regarded as a special case of the clustered multi-task learning problem, where the number of nodes in each cluster is fixed to one.

The update equation of the MDLMS algorithm for the multi-task learning problem is given by

$$\boldsymbol{\phi}_i^{(k+1)} = \boldsymbol{\phi}_i^{(k)} + \mu \boldsymbol{u}_i^{(k)} \left[d_i^{(k)} - \boldsymbol{u}_i^{(k)\mathsf{T}} \boldsymbol{\phi}_j^{(k)} \right] + \eta\mu \sum_{j \in \mathcal{N}_i^-} \rho_{ij} \left(\boldsymbol{\phi}_j^{(k)} - \boldsymbol{\phi}_i^{(k)} \right), \tag{6}$$

where a regularization term to delineate the interrelation of estimated vectors across neighbor nodes is incorporated.The non-negative coefficient ρ_{ij} modulate the regularization intensity, which is achieved by applying the Metropolis rule [12], as specified in Equation (5). Recognizing inter-task relationships is crucial, offering potential enhancements in estimation precision.

III. Proposed TMDLMS Algorithms

Here, we propose TMDLMS algorithm by treating parameters of the conventional MDLMS algorithm in each iteration to be independent learnable parameters, and optimize them by supervised learning using the pair of the measurement vector and the corresponding desired signal as training data with some SGD based algorithm.

We introduce the proposed TMDLMS algorithm by treating the parameters μ and η of the MDLMS algorithm to be the independent learnable parameters. Namely, in the proposed algorithm, μ and η in different iteration become distinct learnable parameters denoted as $\mu^{(k)}$ and $\eta^{(k)}$. Thus, the update equation of the TMDLMS algorithm is given by

$$\boldsymbol{\phi}_i^{(k+1)} = \boldsymbol{\phi}_i^{(k)} + \mu^{(k)} \boldsymbol{u}_i^{(k)} \left[d_i^{(k)} - \boldsymbol{u}_i^{(k)\mathsf{T}} \boldsymbol{\phi}_j^{(k)} \right] + \eta^{(k)}\mu^{(k)} \sum_{j \in \mathcal{N}_i^-} \rho_{ij} \left(\boldsymbol{\phi}_j^{(k)} - \boldsymbol{\phi}_i^{(k)} \right). \tag{7}$$

The time variant parameters $\mu^{(k)}$ and $\eta^{(k)}$ are optimized by using SGD, where the loss function is defined as

$$J_i^{(k)} = \frac{1}{N} \sum_{i=0}^{N-1} \left\| d_i^{(k)} - y_i^{(k)} \right\|_2^2, \tag{8}$$

where

$$y_i^{(k)} = \boldsymbol{u}_i^{(k)\mathsf{T}} \boldsymbol{\phi}_i^{(k-1)}. \tag{9}$$

Then, we describe the gradient descent-based approach in the TMDLMS algorithm, assuming the size of the batch to be $B = 1$ for simplicity. The partial differentiation of the loss function $J_i^{(t)}$ at the t-th round of the incremental training with respect to the step size $\mu_i^{(k-\tau)}$ and the regularization coefficient $\eta_i^{(k-\tau)}$ are given by

$$\begin{aligned} \frac{\partial J_i^{(t)}}{\partial \mu_i^{(t-\tau)}} &= \frac{\partial J_i^{(t)}}{\partial y_i^{(t)}} \frac{\partial y_i^{(t)}}{\partial y_i^{(t-1)}} \frac{\partial y_i^{(t-1)}}{\partial y_i^{(t-2)}} \cdots \\ &\quad \cdots \frac{\partial y_i^{(t-\tau+2)}}{\partial y_i^{(t-\tau+1)}} \frac{\partial y_i^{(t-\tau+1)}}{\partial \mu_i^{(t-\tau)}} \\ &= (-2)^\tau \cdot e_i^{(t)} {e_i^{(t-\tau)}}^2 \boldsymbol{u}_i^{(t-\tau+1)\mathsf{T}} \boldsymbol{u}_i^{(t-\tau)} \\ &\quad \cdot \prod_{k=t-\tau+1}^{t-1} \left[\mu_i^{(k)} \boldsymbol{u}_i^{(k+1)\mathsf{T}} \boldsymbol{u}_i^{(k)\mathsf{T}} \right], \end{aligned} \tag{10}$$

$$\begin{aligned} \frac{\partial J_i^{(t)}}{\partial \eta_i^{(t-\tau)}} &= \frac{\partial J_i^{(t)}}{\partial y_i^{(t)}} \frac{\partial y_i^{(t)}}{\partial y_i^{(t-1)}} \frac{\partial y_i^{(t-1)}}{\partial y_i^{(t-2)}} \cdots \\ &\quad \cdots \frac{\partial y_i^{(t-\tau+2)}}{\partial y_i^{(t-\tau+1)}} \frac{\partial y_i^{(t-\tau+1)}}{\partial \eta_i^{(t-\tau)}} \\ &= (-2)^\tau \cdot e_i^{(t)} \left(y_j^{(t-1)} - y_i^{(t-1)} \right) \\ &\quad \cdot \prod_{k=t-\tau+1}^{t-1} \left[\mu_i^{(k)} \boldsymbol{u}_i^{(k+1)\mathsf{T}} \boldsymbol{u}_i^{(k)\mathsf{T}} \right]. \end{aligned} \tag{11}$$

IV. Numerical Results

We have conducted numerical experiments of the system identification problem as an example of the multi-task learning problem in order to evaluate the performance of the proposed TMDLMS algorithm in comparison with that of conventional algorithms such as LMS, DLMS and MDLMS.

A. Experimental Setup

In the numerical evaluations, we consider to identify the system characterized by the target vectors $\boldsymbol{w}_i^{\mathrm{o}} \in R^{L \times 1}$ $(i = 1, \ldots, N)$, which corresponds to node i in the network of $N = 10$ nodes, from the linear observations of

$$d_i^{(k)} = \boldsymbol{u}_i^{(k)\mathsf{T}} \boldsymbol{w}_i^{\mathrm{o}} + v_i^{(k)}, \tag{12}$$

where the measurement noise $v_i^{(k)}$ is assumed to follow $\mathcal{N}(0, 0.1)$, and the length of the target vector is set to $L = 10$. The vector $\boldsymbol{u}_i^{(k)}$ is used to obtain the linear measurement $d_i^{(k)}$ at node i and time k, whose elements are generated from independent Gaussian distribution with mean of 0 and variance of 1.

For the network structure, we have randomly generated the 10-node network, where all nodes are randomly positioned within a square region of side length $1/\sqrt{3}$. Edges are established between nodes if their distance was less than or equal to 0.25. We have further ensured that each node had at least one neighboring node.

The target vectors $\boldsymbol{w}_i^{\mathrm{o}}$ are systematically generated based on the network structure as follows: Initially, we generate the target vector $\boldsymbol{w}_1^{\mathrm{o}}$ of node 1 as a sample from the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Then, for the other nodes, if

node j is adjacent to node i whose target vector $\boldsymbol{w}_i^{\mathrm{o}}$ has been already generated, we generate the target vector of node j as

$$\boldsymbol{w}_j^{\mathrm{o}} = \frac{1}{\sqrt{1+\sigma^2}}(\boldsymbol{w}_i^{\mathrm{o}} + \delta\boldsymbol{w}_j^*), \tag{13}$$

where $\delta\boldsymbol{w}_j^*$ is a sample from the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Note that the value of σ^2 can control the similarity of the target vectors of the neighboring nodes. Namely, if σ^2 is smaller, the target vectors of different nodes are similar, and if σ^2 is larger, the target vectors are more different.

In the proposed TMDLMS, we have employed Adam [11] as the optimization algorithm and utilized Optuna [13] to determine the hyper-parameters, such ad the learning rate and the initial value. Optuna [13] is employed to optimize the loss function $J_i^{(k)}$, which quantifies the error between the model output and the target values. The optimization process is conducted through multiple trials, where Optuna selects a set of initial values for the hyperparameters, such as $\mu^{(k)}$ and $\eta^{(k)}$, from predefined ranges in each trial. These hyperparameters are then used to update the model, and the loss value for the current trial is computed. By iteratively exploring different hyperparameter combinations and using the loss value feedback from each trial, Optuna gradually narrows the search space.As the trials progress, Optuna dynamically adjusts its hyperparameter search strategy based on historical data, steadily converging toward the optimal solution.

B. Performance Evaluation

Figures 2-4 depict the learning curves of the proposed TMDLMS algorithms, accompanied by the profiles of the step size $\mu^{(k)}$ and $\eta^{(k)}$ obtained by the proposed deep unfolding approach under the condition of $\sigma^2 = 0.01$. In the figures, "TMDLMS ($\mu^{(k)}$)" and "TMDLMS ($\mu^{(k)}$ and $\eta^{(k)}$)" stand for the performance of the proposed TMDLMS with the learnable parameter(s) of $\mu^{(k)}$ only (η is fixed to 0.1) and of $\mu^{(k)}$ and $\eta^{(k)}$, respectively. For a comprehensive evaluation, we have also presented the performance of the LMS, the DLMS, and the MDLMS algorithm with fixed parameters in the figures. As observed in Figure 2, both proposed algorithms achieve better convergence performance compared to the conventional schemes, with the trainable η achieving the lowest steady-state error. Figure 3 reveals that the learned step size $\mu^{(k)}$ tends to decrease as iterations progress, aligning qualitatively with conventional heuristic variable step size approaches.

Figures 5-7 illustrate the learning curves and the profiles of the parameters of the proposed TMDLMS algorithm under the condition of $\sigma^2 = 0.001$, which is smaller than the previous case. As seen in Figures 5, the DLMS achieves better performance than the standalone LMS algorithm. This is because the DLMS algorithm can share information with neighboring nodes, which improves the accuracy of weight estimates. The proposed TMDLMS algorithms can outperform other existing schemes in terms of both the convergence rate and the steady-state error in this case as well.

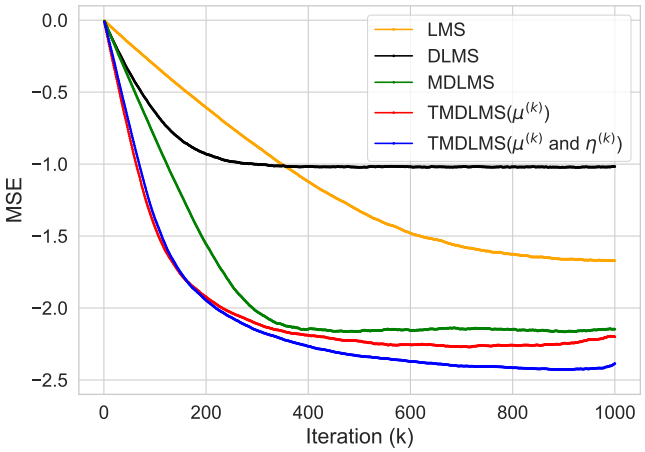


Fig. 2. Learning curve of proposed method ($\sigma^2 = 0.01$)

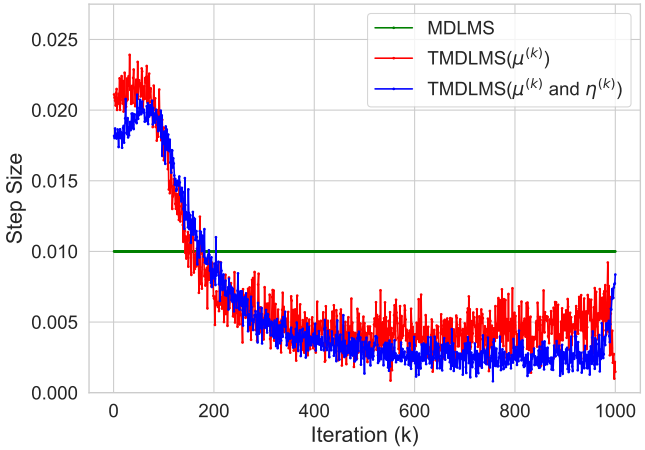


Fig. 3. Profile of step size $\mu^{(k)}$ of proposed TMDLMS ($\sigma^2 = 0.01$)

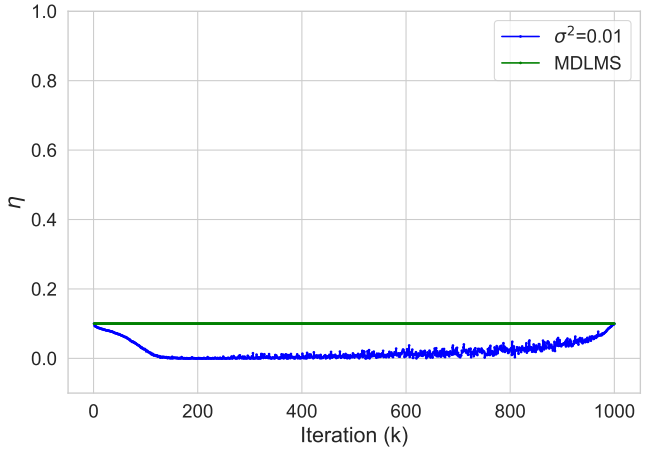


Fig. 4. Profile of η of proposed TMDLMS ($\sigma^2 = 0.01$)

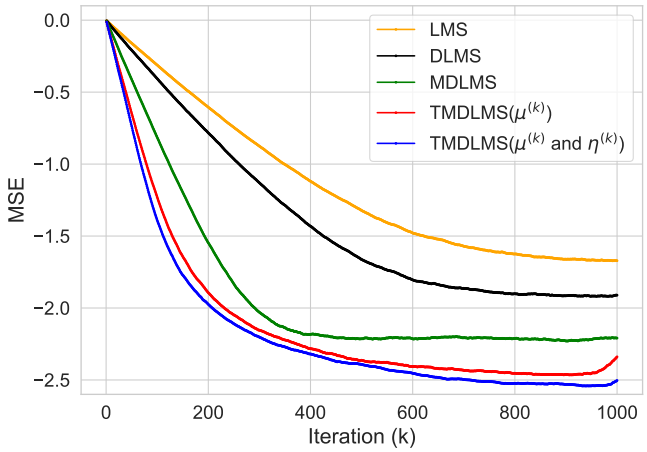


Fig. 5. Learning curve of proposed method ($\sigma^2 = 0.001$)

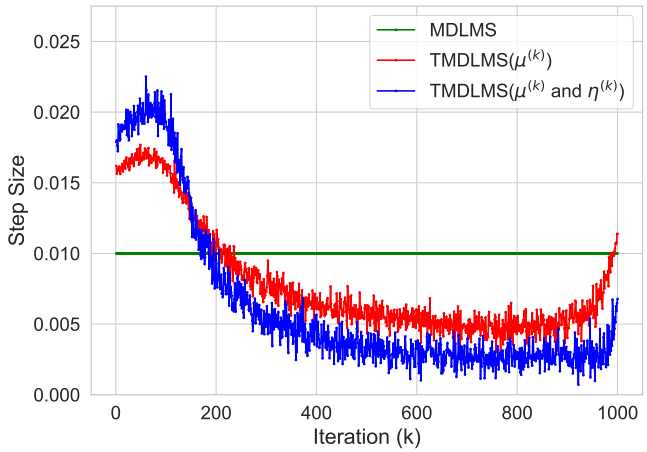


Fig. 6. Profile of step size $\mu^{(k)}$ of proposed TMDLMS ($\sigma^2 = 0.001$)

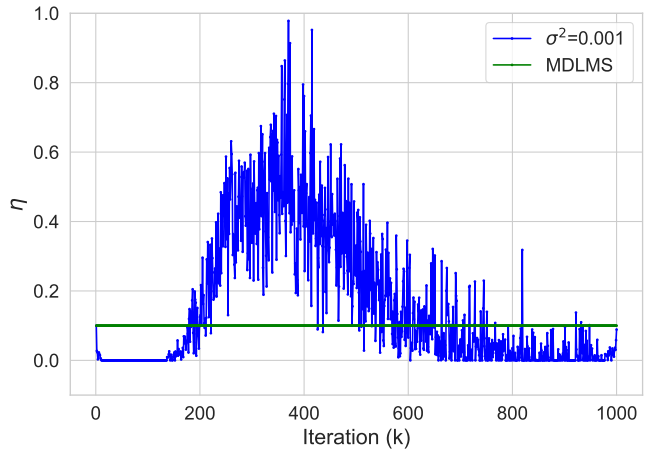


Fig. 7. Profile of η of proposed TMDLMS ($\sigma^2 = 0.001$)

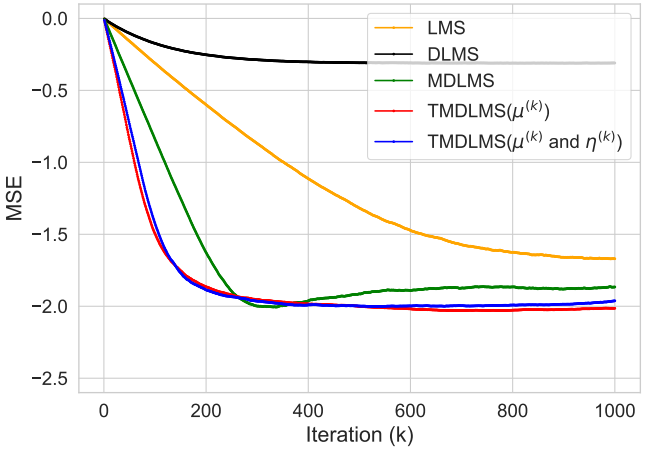


Fig. 8. Learning curve of proposed method ($\sigma^2 = 0.1$)

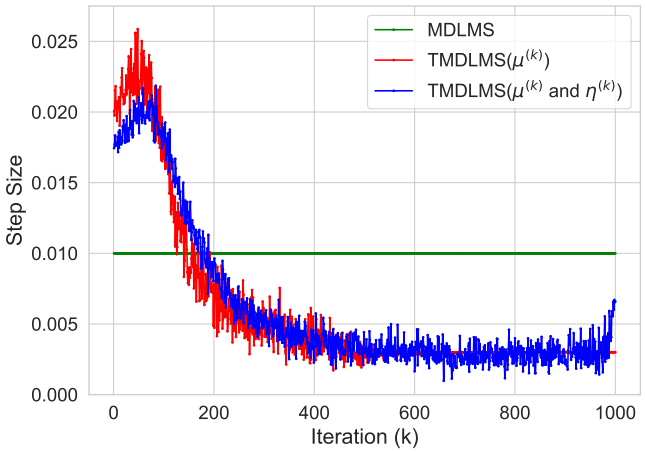


Fig. 9. Profile of step size $\mu^{(k)}$ of proposed TMDLMS ($\sigma^2 = 0.1$)

Figures 8-10 present the learning curves and the profiles of the parameters of the proposed TMDLMS algorithm under the condition of $\sigma^2 = 0.1$.

As illustrated in Figure 8, when σ^2 is high, the performance of the DLMS is degraded significantly, because it assumes that the target vectors of all nodes are in common.

The MDLMS algorithm, on the other hand, can achieve better performance than the standalone LMS algorithm even in such a case. While they get information from other nodes, they can cope with the difference of target vector of each node. The proposed TMDLMS algorithms demonstrate a faster convergence and a lower steady-state error again compared to the existing algorithm. However, in this case, it's not apparent which of the two scenarios of the proposed TMDLMS, trainable η or non-trainable η, performs better. It might be said that, under the condition of $\sigma^2 = 0.1$, the trainable η has small impact on the convergence performance.

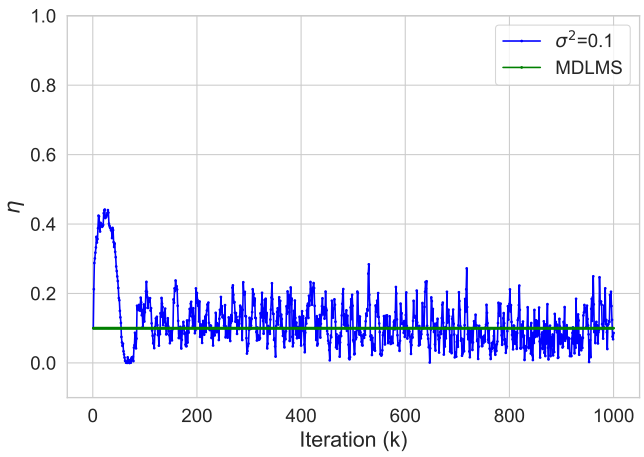


Fig. 10. Profile of η of proposed TMDLMS ($\sigma^2 = 0.1$)

V. Conclusion

We have proposed the TMDLMS algorithms by introducing learnable variable parameters into the MDLMS algorithm and optimizing those parameters with the machine learning approach using the idea of deep unfolding. Through numerical experiments of the system identification problem, we have verified the validity of the proposed approach, demonstrating that they can achieve a faster convergence rate and a reduced steady-state error concurrently. Additionally, the trainable parameter η plays a role in influencing the convergence speed and steady-state error. However, the exact nature of this influence warrants further in-depth experimentation.

Acknowledgment

This work has been supported in part by JSPS KAKENHI Grant Number JP22H00514.

References

[1] B. Widrow, M. E. Hoff, *et al.*, "Adaptive switching circuits," in *IRE WESCON convention record*, New York, vol. 4, 1960, pp. 96–104.

[2] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3122–3136, 2008.

[3] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4129–4144, 2014.

[4] H. He, S. Jin, C.-K. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 77–83, 2019.

[5] J. R. Hershey, J. L. Roux, and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," *arXiv preprint arXiv:1409.2574*, 2014.

[6] A. Balatsoukas-Stimming and C. Studer, "Deep unfolding for communications systems: A survey and some new directions," in *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*, IEEE, 2019, pp. 266–271.

[7] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[8] D. Ito, S. Takabe, and T. Wadayama, "Trainable ista for sparse signal recovery," *IEEE Transactions on Signal Processing*, vol. 67, no. 12, pp. 3113–3125, 2019.

[9] K. Hayashi, K. Shiohara, and T. Sasaki, "Differentiable programming based step size optimization for lms and nlms algorithms," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1721–1727. DOI: 10.1109/APSIPAASC47483.2019.9023175.

[10] Y. Nishihata and K. Ishii, "Deep-unfolding aided optimization of edge weights and step sizes for diffusion lms algorithm," *IEICE Technical Report; IEICE Tech. Rep.*, vol. 121, no. 101, pp. 1–6, 2021 (in Japanese).

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[12] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

Reduced-dimensional MUSIC Algorithm for Frequency Diverse Array in MIMO Radar System

Beizuo Zhu*, Kazunori Hayashi† and Hiroki Mori‡
* Graduate School of Informatics, Kyoto University, Japan
E-mail:zhu.beizuo.54t@st.kyoto-u.ac.jp
† Institute for Liberal Arts and Sciences, Kyoto University, Japan
E-mail: hayashi.kazunori.4w@kyoto-u.ac.jp
‡ Toshiba Corporation, Japan
E-mail: hiroki2.mori@toshiba.co.jp

***Abstract*—The paper proposes a reduced-dimensional multiple signal classification (RD-MUSIC) algorithm for frequency diverse array (FDA) multi-input multi-output (MIMO) radar. The proposed method firstly constructs a constrained minimization problem based on the two-dimensional (2D) MUSIC spectrum function. Then, we reduce the dimension of the minimization problem by calculating the partial derivative with respect to the range parameter, which results in the minimization problem only about the direction of arrival (DOA) parameter. Finally, we substitute the DOA estimates into the steering vector and directly obtain the closed-form solutions of range estimates based on the least squares (LS) method. The proposed algorithm requires only one 1D peak search for DOA parameters, which greatly reduces the computational complexity. In addition, 1D peak searches can be used for the range estimation as well, achieving better estimation accuracy at the cost of the increased complexity compared with the LS solutions. Moreover, the proposed algorithm can be applied to not only uniform linear array (ULA) but also sparse array, demonstrating high versatility. Numerical results show that the proposed RD-MUSIC algorithm is effectiveness regardless of the array configuration and that it can outperform the conventional algorithms including atomic norm minimization (ANM) based method and MUSIC based method.**

I. Introduction

In array signal processing, direction of arrival (DOA) estimation based on the multi-input multi-output (MIMO) radar [1] is widely used in various fields due to the higher spatial degrees of freedom compared to passive antenna arrays. Recently, frequency diverse array (FDA) MIMO radar [2][3] has received attention because of its DOA-range independent beampattern, which enables us to simultaneously obtain DOA and range estimates of multiple targets. Many traditional algorithms have been applied for the joint DOA and range estimation of the FDA MIMO radar so far, such as multiple signal classification (MUSIC) [4], estimation of signal parameters via rotational invariance techniques (ESPRIT) [5], and parallel factor analysis (PARAFAC) [6] based algorithms.

In order to improve the estimation performance, compressed sensing (CS) [7] technique has been applied for the FDA MIMO radar. A representative CS approach called atomic norm minimization (ANM), which is a grid-free algorithm, has been applied to the FDA MIMO radar in [8]. The CS based methods generally can cope with a small number of snapshots, but the algorithm requires prohibitive computational complexity, since the estimation problem is formulated as a semi-definite programming (SDP) problem [9]. To reduce the complexity while keeping the estimation performance, the idea of sparse array has been utilized in [10], but it still requires high computational complexity. Moreover, the algorithm requires the transmitting array to be a uniform linear aray (ULA) in order to ensure that the Vandermonde decomposition of the correlation matrix is possible.

Some improved algorithms based on the traditional approaches also has been proposed for the FDA MIMO radar. An improved ESPRIT based algorithm is proposed in [11], where DOA estimates are used to recover the range steering vector from the estimated receive-transmit array manifold. The method requires slightly higher complexity than the classical ESPRIT algorithm, but achieves better range estimation performance when nonlinear frequency increments are applied for the FDA. One of restrictions of the algorithm is that it requires the array structure to be the ULA. On the other hand, a MUSIC based algorithm for the FDA MIMO radar is proposed in [12], called reduced-dimensional (RD) MUSIC algorithm, where a two-dimensional (2D) peak search the MUSIC spectrum function for the DOA and range estimation is converted into several one-dimensional (1D) peak searches by the proposed constrained optimization problem, which can reduce the computational complexity of the original 2D MUSIC algorithm. The algorithm does not require the ULA, and can be applied for sparse arrays to improve the performance. However, the required number of the 1D peak searches for the range estimates is the same as the number of targets, whose computational complexity might not be acceptable in some applications.

In this paper, to further reduce the computational complexity of the RD MUSIC in [12], an improved MUSIC based algorithm for the FDA MIMO radar is proposed. Different from [12], the proposed method uses the ℓ_2-norm of the transmitting

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2123.

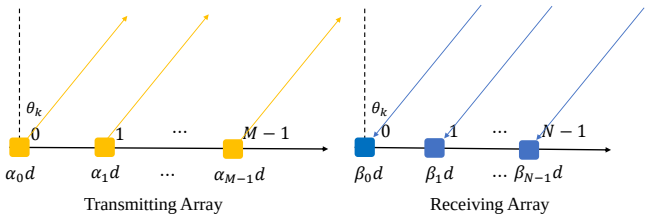


Fig. 1. An example of FDA for co-located MIMO radar

steering vector as a constraint of the optimization problem, which enables us to fully utilize all information in the received signal in the DOA peack search process. Moreover, for the computational complexity reduction, the range estimates are given by closed-form solutions based on the LS method. Additionally, 1D peak searches can be also applied to obtain the range estimates with higher accuracy at the cost of increased complexity compared with the closed-form LS solutions. The performance of the proposed algorithm for both ULA and sparse arrays is verified by computer simulations comparing with that of conventional methods.

In the rest of paper, $\mathbb{N}$ and $\mathbb{C}$ denote sets of all natural numbers and complex numbers, respectively. We use lower-case bold characters to denote column vectors and upper-case bold characters for matrices. $(\cdot)^T$ and $(\cdot)^H$ are the transpose and conjugate transpose of a matrix or vector, respectively. $(\cdot)^{-1}$ denotes the inverse of a matrix. $\oplus$, $\otimes$, and $\odot$ represent Hadamard product, Kronecker product, and Khatri-Rao product, respectively. $\|\cdot\|_2$ denotes the ℓ_2-norm and $|\cdot|$ is the absolute value. $\ln(\cdot)$ denotes the natural logarithm.

II. System Model of FDA-MIMO Radar

Fig. 1 illustrates the co-located FDA-MIMO radar considered in this paper. The transmitting array is a linear array consisting of M antenna elements. Suppose that the 0-th transmitting antenna is located at the origin, and the location set of the transmitting antenna elements is defined as

$$\Phi_t = \{\alpha_0 d, \alpha_1 d, \ldots, \alpha_{M-1} d\}, \tag{1}$$

where $\alpha_0 = 0$, $\alpha_m \in \mathbb{N}$ $(m = 1, \ldots, M-1)$, and d denotes unit spacing between antenna elements.

A linear frequency shift corresponding to the antenna location is employed for the FDA, and the transmitting frequency of the m-th antenna is given by

$$f_m = f_0 + \alpha_m \Delta f, \ \ (m = 0, 1, \ldots, M-1), \tag{2}$$

where f_0 denotes the carrier frequency at the 0-th transmitting antenna and $\Delta f (\ll f_0)$ is the frequency shift corresponding to the unit antenna spacing d. To avoid phase ambiguity in DOA estimation, d is set to be the half of the smallest wavelength among the transmitted signals as

$$d = \frac{c}{2(f_0 + \alpha_{M-1}\Delta f)}, \tag{3}$$

where c is the speed of light. Note that, for the case with a sparse array, at least one antenna spacing must be d as a limitation.

Based on the assumptions above, the transmitted signal from the m-th transmitting antenna can be written as

$$x_m(t) = \phi_m(t) e^{j2\pi(f_0 + \alpha_m \Delta f)t}, \ \ (0 \le t \le T), \tag{4}$$

where T is the duration and $\phi_m(t)$ is the orthogonal baseband waveform from the m-th transmitting antenna, which has properties of

$$\int_0^T \phi_{m_1}(t)\phi_{m_2}^*(t-\tau)dt = \begin{cases} 0 & (m_1 \neq m_2) \\ 0 & (m_1 = m_2, \tau \neq 0) \\ 1 & (m_1 = m_2, \tau = 0) \end{cases}. \tag{5}$$

The receiving array depicted in Fig. 1 having N antenna elements is also assumed to be a linear array, and the location set of the receiving antenna elements is expressed as

$$\Phi_r = \{\beta_0 d, \beta_1 d, \ldots, \beta_{N-1} d\}, \tag{6}$$

where β_0 is assumed as 0 and $\beta_n \in \mathbb{N}$ $(n = 1, \ldots, N-1)$.

Suppose that there exist K far field targets in the space with DOA and range parameters of (θ_k, r_k), $(k = 0, 1, \cdots, K-1)$ with respect to the antenna origin. All the targets are illuminated by the signals transmitted from the FDA, and reflect them back to the receiving array. During the process, taking the 0-th transmitting and receiving elements as the references, for a specific k-th target with DOA and range (θ_k, r_k), the time delay from the m-th transmitting antenna to the n-th receiving antenna is given by

$$\tau_{m,n,k} = \tau_{0,0,k} - \left(\frac{\alpha_m d \sin\theta_k}{c} + \frac{\beta_n d \sin\theta_k}{c} \right), \tag{7}$$

where $m = 0, 1, \ldots, M-1$ and $n = 0, 1, \ldots, N-1$. The first term is the reference time delay, namely $\tau_{0,0,k} = 2r_k/c$. The second and third term are resulted from the location of transmitting and receiving antennas, respectively.

Combining (4) and (7) while ignoring the additive noise, the received signal at the n-th receiving antenna corresponding to the k-th target from all the transmitting antennas is given by

$$\bar{y}_{n,k}(t) = \sum_{m=0}^{M-1} \phi_m(t - \tau_{m,n,k}) e^{j2\pi(f_0 + \alpha_m \Delta f)(t - \tau_{m,n,k})}. \tag{8}$$

Assuming transmitted signals to be narrowband, namely $\phi_m(t - \tau_{m,n,k}) = \phi_m(t - \tau_{0,0,k})$, (8) can be rewritten as

$$\begin{aligned} &\bar{y}_{n,k}(t) \\ &= \sum_{m=0}^{M-1} \phi_m(t - \tau_{0,0,k}) e^{j2\pi f_0 t} e^{-j2\pi f_0 \frac{2r_k}{c}} e^{j2\pi f_0 \frac{\alpha_m d \sin\theta_k}{c}} \\ &\quad \cdot e^{j2\pi f_0 \frac{\beta_n d \sin\theta_k}{c}} e^{j2\pi \alpha_m \Delta f t} e^{-j2\pi \alpha_m \Delta f \frac{2r_k}{c}} \\ &\quad \cdot e^{j2\pi \alpha_m \Delta f \frac{\alpha_m d \sin\theta_k}{c}} e^{j2\pi \alpha_m \Delta f \frac{\beta_n d \sin\theta_k}{c}}. \end{aligned} \tag{9}$$

The received signal can be separated by matched filters using the orthogonal properties in (5). Specifically, the received

signal at the n-th antenna from the m-th transmitting antenna reflected at the k-th target is obtained by

$$\begin{aligned} & y_{m,n,k} \\ &= \int_0^T \phi_m^* \left(t-\tau_{0,0,k}\right) e^{-j2\pi(f_0+\alpha_m\Delta f)t} \bar{y}_{n,k}(t)dt \\ &= e^{-j2\pi f_0 \frac{2r_k}{c}} e^{j2\pi(f_0+\alpha_m\Delta f)\frac{\alpha_m d\sin\theta_k}{c}} e^{j2\pi(f_0+\alpha_m\Delta f)\frac{\beta_n d\sin\theta_k}{c}} \\ &\quad \cdot e^{-j2\pi\alpha_m\Delta f\frac{2r_k}{c}}. \end{aligned} \tag{10}$$

Since we assume $\Delta f \ll f_0$, we have approximate formulas of

$$\begin{aligned} e^{j2\pi(f_0+\alpha_m\Delta f)\frac{\alpha_m d\sin\theta_k}{c}} &\approx e^{j2\pi f_0\frac{\alpha_m d\sin\theta_k}{c}}, \\ e^{j2\pi(f_0+\alpha_m\Delta f)\frac{\beta_n d\sin\theta_k}{c}} &\approx e^{j2\pi f_0\frac{\beta_n d\sin\theta_k}{c}}. \end{aligned}$$

Besides, if we define $\xi_k = e^{-j4\pi f_0 r_k/c}$, the received signal can be expressed as

$$y_{m,n,k} = \xi_k e^{-j4\pi\frac{\alpha_m\Delta f r_k}{c}} e^{j2\pi f_0\frac{\alpha_m d\sin\theta_k}{c}} e^{j2\pi f_0\frac{\beta_n d\sin\theta_k}{c}}. \tag{11}$$

Since we have N receiving antennas and M matched filter outputs for each receiving antenna, the received signal corresponding to the k-th target can be written in a matrix form as

$$\mathbf{Y}_k = \begin{bmatrix} y_{0,0,k} & y_{0,1,k} & \cdots & y_{0,N-1,k} \\ y_{1,0,k} & y_{1,1,k} & \cdots & y_{1,N-1,k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{M-1,0,k} & y_{M-1,1,k} & \cdots & y_{M-1,N-1,k} \end{bmatrix}. \tag{12}$$

According to the exponential terms in (11), we can rewrite (12) by a matrix-vector form as

$$\mathbf{Y}_k = \xi_k \left[\boldsymbol{b}_{\mathrm{t}}\left(\theta_k\right) \oplus \boldsymbol{c}_{\mathrm{t}}\left(r_k\right)\right] \boldsymbol{a}_{\mathrm{r}}^T\left(\theta_k\right), \tag{13}$$

where

$$\begin{aligned} \boldsymbol{b}_{\mathrm{t}}(\theta_k) &= \left[e^{j2\pi\alpha_0 d\sin(\theta_k)/\lambda}, \cdots, e^{j2\pi\alpha_{M-1}d\sin(\theta_k)/\lambda}\right]^T, \\ \boldsymbol{c}_{\mathrm{t}}(r_k) &= \left[e^{-j4\pi\alpha_0\Delta f r_k/c}, \cdots, e^{-j4\pi\alpha_{M-1}\Delta f r_k/c}\right]^T, \\ \boldsymbol{a}_{\mathrm{r}}(\theta_k) &= \left[e^{j2\pi\beta_0 d\sin(\theta_k)/\lambda}, \cdots, e^{j2\pi\beta_{N-1}d\sin(\theta_k)/\lambda}\right]^T, \end{aligned}$$

and $\lambda = c/f_0$.

Taking the impact of the additive Gaussian noise on the received signal into consideration and summing up all the reflected signals from the K targets, the received signal model of the l-th $(l = 0, 1, \cdots, L-1)$ snapshot is denoted as

$$\mathbf{Y}(l) = \sum_{k=0}^{K-1} \xi_k(l) \left[\boldsymbol{b}_{\mathrm{t}}\left(\theta_k\right) \oplus \boldsymbol{c}_{\mathrm{t}}\left(r_k\right)\right] \boldsymbol{a}_{\mathrm{r}}^T\left(\theta_k\right) + \mathbf{N}(l), \tag{14}$$

where $\mathbf{N}(l) \in \mathbb{C}^{M\times N}$ is a zero-mean white circular complex Gaussian noise matrix in the l-th snapshot. We give the index l to ξ_k because it takes different value for each snapshot.

By using the property of $\mathrm{vec}(\boldsymbol{a}\boldsymbol{b}^T) = \boldsymbol{b} \otimes \boldsymbol{a}$ where $\mathrm{vec}(\cdot)$ denotes the vectorization of a matrix, (14) can be rewritten by vector form with size $MN \times 1$ as

$$\begin{aligned} \boldsymbol{y}(l) &= \mathrm{vec}(\mathbf{Y}(l)) \\ &= \sum_{k=0}^{K-1} \xi_k(l) \boldsymbol{a}_{\mathrm{r}}\left(\theta_k\right) \otimes \left(\boldsymbol{b}_{\mathrm{t}}\left(\theta_k\right) \oplus \boldsymbol{c}_{\mathrm{t}}\left(r_k\right)\right) + \boldsymbol{n}(l) \\ &= \sum_{k=0}^{K-1} \xi_k(l) \boldsymbol{a}_{\mathrm{r}}\left(\theta_k\right) \otimes \boldsymbol{a}_{\mathrm{t}}\left(\theta_k, r_k\right) + \boldsymbol{n}(l) \\ &= \mathbf{A}_{\mathrm{r}}\left(\boldsymbol{\theta}\right) \odot \mathbf{A}_{\mathrm{t}}\left(\boldsymbol{\theta}, \boldsymbol{r}\right) \boldsymbol{\xi}(l) + \boldsymbol{n}(l), \end{aligned} \tag{15}$$

where

$$\boldsymbol{n}(l) = \mathrm{vec}(\mathbf{N}(l)), \tag{16}$$

$$\boldsymbol{a}_{\mathrm{t}}\left(\theta_k, r_k\right) = \boldsymbol{b}_{\mathrm{t}}\left(\theta_k\right) \oplus \boldsymbol{c}_{\mathrm{t}}\left(r_k\right), \tag{17}$$

$$\mathbf{A}_{\mathrm{t}}\left(\boldsymbol{\theta}, \boldsymbol{r}\right) = \left[\boldsymbol{a}_{\mathrm{t}}\left(\theta_0, r_0\right), \cdots, \boldsymbol{a}_{\mathrm{t}}\left(\theta_{K-1}, r_{K-1}\right)\right], \tag{18}$$

$$\mathbf{A}_{\mathrm{r}}\left(\boldsymbol{\theta}\right) = \left[\boldsymbol{a}_{\mathrm{r}}\left(\theta_0\right), \cdots, \boldsymbol{a}_{\mathrm{r}}\left(\theta_{K-1}\right)\right], \tag{19}$$

$$\boldsymbol{\xi}(l) = \left[\xi_0(l), \cdots, \xi_{K-1}(l)\right]^T, \tag{20}$$

and $\boldsymbol{\theta} = [\theta_0, \cdots, \theta_{K-1}]^T$, $\boldsymbol{r} = [r_0, \cdots, r_{K-1}]^T$. By collecting L snapshots, the final received signal is written as

$$\bar{\mathbf{Y}} = \mathbf{A}_{\mathrm{r}}\left(\boldsymbol{\theta}\right) \odot \mathbf{A}_{\mathrm{t}}\left(\boldsymbol{\theta}, \boldsymbol{r}\right) \boldsymbol{\xi} + \mathbf{N} \in \mathbb{C}^{MN\times L}, \tag{21}$$

where

$$\begin{aligned} \bar{\mathbf{Y}} &= \left[\boldsymbol{y}(0), \cdots, \boldsymbol{y}(L-1)\right], \\ \boldsymbol{\xi} &= \left[\boldsymbol{\xi}(0), \cdots, \boldsymbol{\xi}(L-1)\right], \\ \mathbf{N} &= \left[\boldsymbol{n}(0), \cdots, \boldsymbol{n}(L-1)\right]. \end{aligned}$$

III. Reduced-Dimensional MUSIC Algorithm for DOA and Range Estimation

In this section, an improved RD MUSIC algorithm is proposed for FDA in MIMO radar system introduced in Sect. II. The proposed algorithm can be applied not only to ULA but also sparse arrays. First, the MUSIC spectrum function is reshaped to separate DOA and range parameter. The reformed function is transformed into a constrained minimization problem and the range parameter in the search function is eliminated by using the property of the vanishing derivative. After 1D search for DOA estimates, they are substituted into the minimization problem and the range estimates are obtained by the least squares method. In addition, 1D search can also be supported for range estimates with higher performance.

A. DOA estimation by 1D search

The covariance matrix of the final received signal in (21) can be approximately given by the sample covariance matrix using L snapshots as

$$\mathbf{R} \approx \bar{\mathbf{Y}}\bar{\mathbf{Y}}^H/L. \tag{22}$$

By eigendecomposition, the covariance matrix can be written as

$$\mathbf{R} = \mathbf{U}_{\mathrm{s}}\boldsymbol{\Lambda}_{\mathrm{s}}\mathbf{U}_{\mathrm{s}}^H + \mathbf{U}_{\mathrm{n}}\boldsymbol{\Lambda}_{\mathrm{n}}\mathbf{U}_{\mathrm{n}}^H, \tag{23}$$

where $\mathbf{U}_{\mathrm{s}} \in \mathbb{C}^{MN\times K}$ and $\mathbf{U}_{\mathrm{n}} \in \mathbb{C}^{MN\times(MN-K)}$ denote the matrices whose column vectors span signal subspace and noise subspace, respectively, while $\mathbf{\Lambda}_{\mathrm{s}}$ and $\mathbf{\Lambda}_{\mathrm{n}}$ are the diagonal matrices with eigenvalues corresponding to the signal subspace and the noise subspace, respectively.

Applying the received signal model in the previous section, the traditional MUSIC spectrum function for peak search is expressed as

$$F(\theta,r)=\frac{1}{\left[\boldsymbol{a}_{\mathrm{r}}(\theta)\otimes\boldsymbol{a}_{\mathrm{t}}(\theta,r)\right]^{H}\mathbf{U}_{\mathrm{n}}\mathbf{U}_{\mathrm{n}}^{H}\left[\boldsymbol{a}_{\mathrm{r}}(\theta)\otimes\boldsymbol{a}_{\mathrm{t}}(\theta,r)\right]}. \tag{24}$$

Since we have $\boldsymbol{a}_{\mathrm{r}}(\theta)\otimes\boldsymbol{a}_{\mathrm{t}}(\theta,r)=\left[\boldsymbol{a}_{\mathrm{r}}\otimes\mathbf{I}_M\right]\boldsymbol{a}_{\mathrm{t}}(\theta,r)$, where $\mathbf{I}_M$ is an identity matrix of size $M\times M$, the MUSIC spectrum function (24) can be rewritten as

$$F(\theta,r)=\frac{1}{\boldsymbol{a}_{\mathrm{t}}(\theta,r)^{H}\mathbf{Q}(\theta)\boldsymbol{a}_{\mathrm{t}}(\theta,r)}, \tag{25}$$

where

$$\mathbf{Q}(\theta)=\left[\boldsymbol{a}_{\mathrm{r}}(\theta)\otimes\mathbf{I}_M\right]^{H}\mathbf{U}_n\mathbf{U}_n^{H}\left[\boldsymbol{a}_{\mathrm{r}}(\theta)\otimes\mathbf{I}_M\right] \tag{26}$$

is an $M\times M$ Hermitian matrix depending only on the DOA parameter.

Since the direct 2D search for the MUSIC spectrum function requires prohibitive computational complexity, we transform the MUSIC search process into a minimization problem with constraint as

$$\begin{cases}\min_{\theta,r}\boldsymbol{a}_{\mathrm{t}}(\theta,r)^{H}\mathbf{Q}(\theta)\boldsymbol{a}_{\mathrm{t}}(\theta,r)\\ \text{s.t. }||\boldsymbol{a}_{\mathrm{t}}(\theta,r)||_2=M\end{cases}, \tag{27}$$

where the constraint is added to avoid the trivial solution of $\boldsymbol{a}_{\mathrm{t}}(\theta,r)=0$. The minimization problem with the constraint can be reformulated as the unconstrained minimization problem whose cost function is given by using a Lagrange multiplier μ as

$$L(\theta,r)=\boldsymbol{a}_{\mathrm{t}}(\theta,r)^{H}\mathbf{Q}(\theta)\boldsymbol{a}_{\mathrm{t}}(\theta,r)-\mu\left(||\boldsymbol{a}_{\mathrm{t}}(\theta,r)||_2-M\right). \tag{28}$$

The partial derivative of $L(\theta,r)$ with respect to r is given by

$$\begin{aligned}\frac{\partial L(\theta,r)}{\partial r}=&\left[\boldsymbol{a}_{\mathrm{t}}(\theta,r)^{H}\mathbf{Q}(\theta)-\mu\boldsymbol{a}_{\mathrm{t}}(\theta,r)^{H}\right]\mathrm{diag}\left(\boldsymbol{b}_{\mathrm{t}}(\theta)\right)\\ &\cdot(-j4\pi\Delta f/c)\begin{bmatrix}\alpha_0e^{-j4\pi\alpha_0\Delta fr/c}\\ \vdots\\ \alpha_{M-1}e^{-j4\pi\alpha_{M-1}\Delta fr/c}\end{bmatrix}.\end{aligned} \tag{29}$$

The condition of the vanishing derivative gives

$$\mathbf{Q}(\theta)\boldsymbol{a}_{\mathrm{t}}(\theta,r)=\mu\boldsymbol{a}_{\mathrm{t}}(\theta,r), \tag{30}$$

which means that $\boldsymbol{a}_{\mathrm{t}}(\theta,r)$ is an eigenvector of $\mathbf{Q}(\theta)$ corresponding to the eigenvalue of μ. If we substitute (30) into (25), we have

$$F(\theta)=\frac{1}{\mu M}. \tag{31}$$

Thus, we need to choose the smallest eigenvalue $\mu_{\min}(\theta)$ of $\mathbf{Q}(\theta)$ out of M eigenvalues as μ in (31) to maximize the MUSIC spectrum function.

Through the above steps, we have transformed the peak search of the 2D MUSIC spectrum function (24) into the 1D search of (31) with the smallest eigenvalue $\mu_{\min}(\theta)$, which gives K DOA estimates $\hat{\boldsymbol{\theta}}=[\hat{\theta}_0,\cdots,\hat{\theta}_{K-1}]^T$ and their corresponding eigenvectors $[\overline{\boldsymbol{a}}_{\mathrm{t}}(\theta_0,r_0),\cdots,\overline{\boldsymbol{a}}_{\mathrm{t}}(\theta_{K-1},r_{K-1})]$.

B. Range estimation by least squares method

The remaining problem is how to obtain the range estimates from $\overline{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)$ and $\hat{\theta}_k$, $k=0,1,\cdots,K-1$. Basically, we can estimate the ranges by using (17), but the eigenvectors $\overline{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)$ and the actual steering vectors $\boldsymbol{a}_{\mathrm{t}}(\theta_k,r_k)$ can be different by a complex constant multiplication factor. In addition, due to the additive noise and/or the estimation error of the covariance matrix, the absolute value of each element in the estimated $\overline{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)$ is not necessarily 1. Thus, in our approach, we propose to use the modified eigenvectors $\hat{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)$, whose m-th element is defined as

$$\hat{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)_m=\frac{\overline{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)_m}{|\overline{a}_{\mathrm{t}}(\theta_k,r_k)_m|\frac{\overline{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)_0}{|\overline{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)_0|}}, \tag{32}$$

where $\overline{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)_m$ denotes the m-th element of $\overline{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)$.

According to (17), $\boldsymbol{c}_{\mathrm{t}}(r_k)$, which contains the information of the range r_k in the power of each element, can be given by

$$\boldsymbol{c}_{\mathrm{t}}(r_k)=\boldsymbol{a}_{\mathrm{t}}(\theta_k,r_k)./\boldsymbol{b}_{\mathrm{t}}(\theta_k), \tag{33}$$

where $./$ denotes the element-wise division. Based on the error between the theoretical value and the estimated value of $\boldsymbol{c}_{\mathrm{t}}(r_k)$ under natural logarithm, least squares (LS) method is used for range estimation as

$$\begin{aligned}\hat{r}_k&=\arg\min_{r_k}||\ln\hat{\boldsymbol{c}}_{\mathrm{t}}(r_k)-\ln(\hat{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)./\hat{\boldsymbol{b}}_{\mathrm{t}}(\theta_k))||_2\\ &=\arg\min_{r_k}||\boldsymbol{p}r_k-\boldsymbol{h}(\theta_k,r_k)||_2,\end{aligned} \tag{34}$$

where

$$\begin{aligned}\hat{\boldsymbol{b}}_{\mathrm{t}}(\theta_k)&=[e^{j2\pi\alpha_0d\sin(\hat{\theta}_k)/\lambda},\cdots,e^{j2\pi\alpha_{M-1}d\sin(\hat{\theta}_k)/\lambda}]^T,\\ \boldsymbol{p}&=[-j4\pi\alpha_0\Delta f/c,\cdots,-j4\pi\alpha_{0M-1}\Delta f/c]^T,\\ \boldsymbol{h}(\theta_k,r_k)&=\ln(\hat{\boldsymbol{a}}_{\mathrm{t}}(\theta_k,r_k)./\hat{\boldsymbol{b}}_{\mathrm{t}}(\theta_k)).\end{aligned}$$

The closed-form solution can be obtained as

$$\hat{r}_k=\frac{\boldsymbol{p}^H\boldsymbol{h}(\theta_k,r_k)+\boldsymbol{h}(\theta_k,r_k)^H\boldsymbol{p}}{2\boldsymbol{p}^H\boldsymbol{p}},\ (k=0,1,\cdots,K-1). \tag{35}$$

C. Range estimation by search method

The proposed LS range estimation method requires low computational complexity, but better range estimation performance can by achieved by the 1D searches for the range estimation at the cost of increased computational complexity.

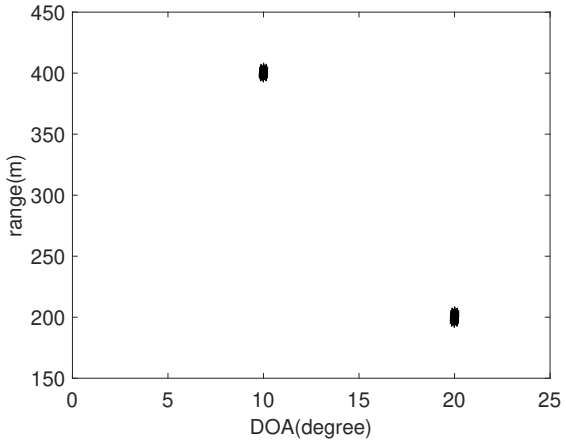


Fig. 2. Scatter plot of the DOA and range estimates

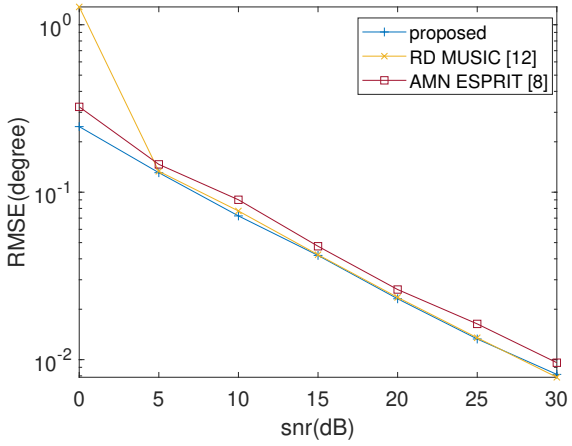


Fig. 3. RMSE of DOA estimates (ULA)

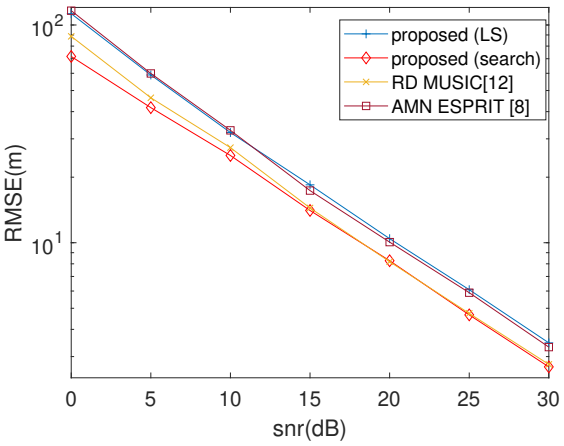


Fig. 4. RMSE of range estimates (ULA)

For each DOA estimate $\hat{\theta}_k, k = 1, 2, \cdots, K$, search function for the corresponding range parameter can be established based on (25) as

$$F(r) = \frac{1}{\boldsymbol{a}_{\mathrm{t}}(\hat{\theta}_k, r)^H \mathbf{Q}(\hat{\theta}_k) \boldsymbol{a}_{\mathrm{t}}(\hat{\theta}_k, r)}, \tag{36}$$

and the range estimate $\hat{r}_k, k = 1, 2, \cdots, K$ can be obtained from the maximum peak point of (36). Note that this method requires K times 1D searches, which require higher computational cost, for better estimation performance.

IV. Simulation Results

In this section, the effectiveness of the proposed RD MUSIC algorithm is verified through computer simulations with various conditions comparing with other algorithms. The performance of the algorithms is evaluated by root-mean-square error (RMSE) of the DOA estimation and the range estimation. For simplicity, we assume a co-located FDA MIMO radar, where the transmitting and receiving arrays share the same physical antennas. f_0 and Δf are set to be 10GHz and 5kHz, respectively. Two far-field targets with DOA and range parameters of $(\theta_0, r_0) = (10^\circ, 400\mathrm{m})$ and $(\theta_1, r_1) = (20^\circ, 200\mathrm{m})$ are assumed.

Fig. 2 shows the scatter plot of the DOA and range estimates pairs obtained by the proposed algorithm, where array structure is 1-level nest array (NA) [13] with the number of antenna elements $M = N = 10$. Signal-to-noise ratio (SNR) and the number of snapshots are set to be 20dB and $L = 100$, respectively. Search step for DOA estimates is $\Delta_\theta = 0.01^\circ$ and range estimates are obtained by the closed-form LS solution. 100 independent simulation trials are performed, and each estimation result is shown as a point in Fig. 2. From the figure, it can be seen that all estimation results are close to the true points, which demonstrates the effectiveness of the proposed algorithm.

Figs. 3 and 4 demonstrate the RMSE performance of DOA and range estimation by the proposed and conventional methods using ULA with $M = N = 10$, where $\Phi_t = \Phi_r = \{0d, 1d, \ldots, 9d\}$. The number of snapshot of the received signal is $L = 10$. Search steps for DOA and range estimates are $\Delta_\theta = 0.01^\circ$ and $\Delta_\mathrm{r} = 0.1$ m, respectively, for the algorithms which require peak search. Yellow line marked with cross is the performance of the conventional RD MUSIC algorithm [12], which require separate DOA and range search for each target. Red line marked with square is the performance of ANM algorithm [8], obtaining the covariance matrix by solving a SDP problem and apply ESPRIT for estimates. From Fig. 3, we can see that the DOA estimation performance of the proposed algorithm is better than that of the conventional RD MUSIC algorithm under low SNR region, and is comparable to the ANM ESPRIT algorithm under high SNR region. Moreover, Fig. 4 shows that the proposed low-complexity LS range estimation method can achieve comparable performance to the ANM ESPRIT algorithm, and that the proposed 1D search method for the range estimation can outperform the conventional RD MUSIC algorithm.

Figs. 5 and 6 illustrate the RMSE performance of DOA and range estimation by the proposed and conventional methods

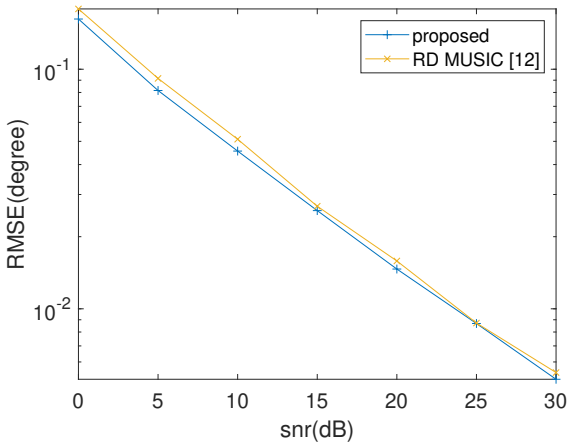


Fig. 5. RMSE of DOA estimates (1-level NA)

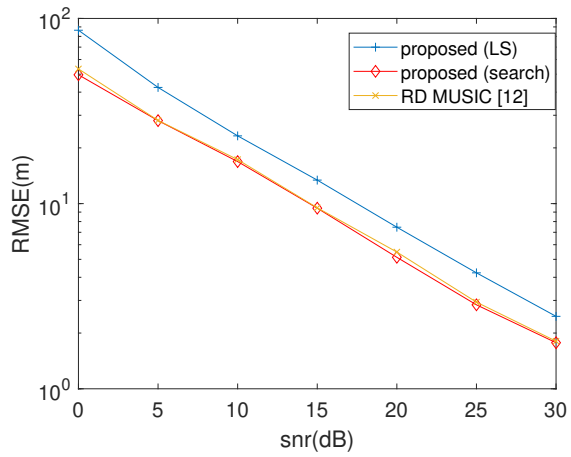


Fig. 6. RMSE of range estimates (1-level NA)

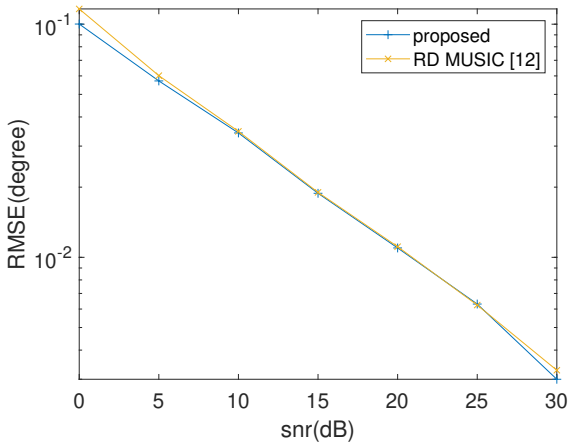


Fig. 7. RMSE of DOA estimates (2-level NA)

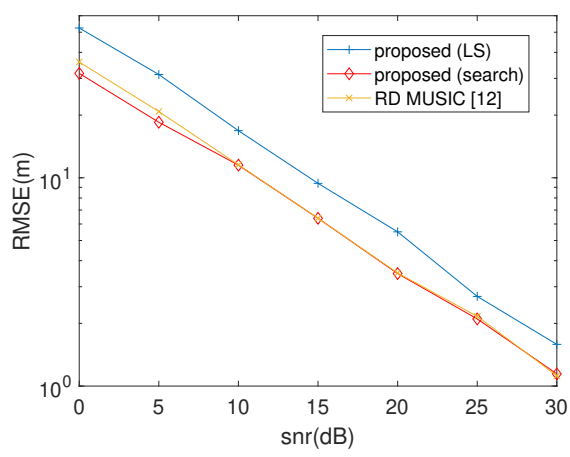


Fig. 8. RMSE of range estimates (2-level NA)

using 1-level NA with $M = N = 10$, where $\Phi_t = \Phi_r = \{0d, 1d, \ldots, 8d, 17d\}$ and the number of snapshot of the received signal is $L = 10$. Search steps are the same as the previous case. Note that the performance of the ANM approach [8] is not shown here, because it requires the transmitting array to be ULA. From the results, we can see that the DOA estimation performance of the proposed algorithm is better than that of the conventional RD MUSIC algorithm in [12]. As for the range estimation performance, the proposed range estimation method with 1D search achieves the best performance for all SNR.

Figs. 7 and 8 show the RMSE performance for the case with 2-level NA with $M = N = 10$, where $\Phi_t = \Phi_r = \{0d, 1d, \ldots, 7d, 15d, 23d\}$. Search steps are the same as the previous cases. Compared with the case of ULA or 1-level NA, it can be found that the RMSE performance of both DOA and range estimation is greatly improved. This is because, as the virtual array aperture increases, the estimation accuracy of the DOA improves, which indirectly leads to the improvement of the range estimation performance.

Finally, Figs. 9 and 10 show the RMSE performance of the DOA and range estimation for the case with co-prime array (CA) [14], which is composed by two sub-arrays given by $\Phi_1 = \{0d, 6d, 12d, 18d, 24d\}$ and $\Phi_2 = \{0d, 5d, 10d, 15d, 20d, 25d\}$. All the physical antennas are used for transmitting array and receiving array, written as $\Phi_t = \Phi_r = \Phi_1 \cup \Phi_2$, thus $M = N = 10$. Search steps are the same as the previous cases. From the results, the estimation performance for the case with CA is close to that with 2-level NA, verifying the effectiveness of the proposed algorithm on sparse arrays. Besides, the RMSE of the proposed algorithm can also outperform that of the conventional RD MUSIC algorithm regardless of the array configuration.

V. Conclusions

In this paper, an improved RD MUSIC algorithm is proposed for FDA MIMO radar, which can be applied for arbitrary linear array. The proposed approach reshapes the MUSIC spectrum function and transforms the peak search operation on it into the minimization problem with the ℓ_2-norm constraint. After some

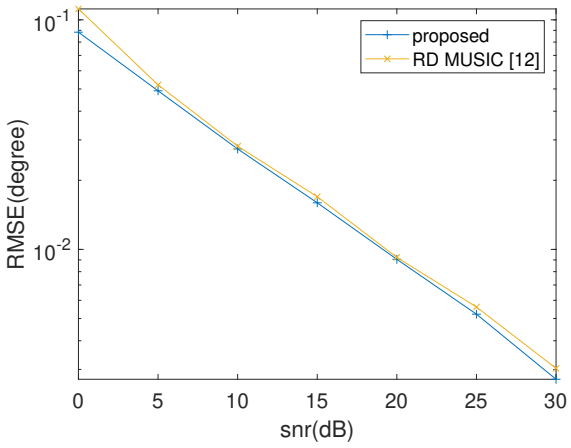


Fig. 9. RMSE of range estimates (CA)

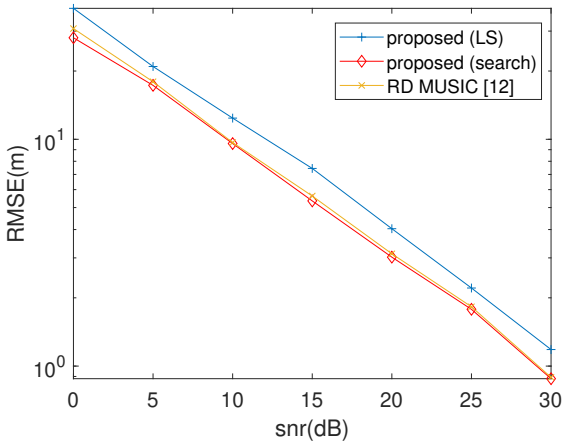


Fig. 10. RMSE of DOA estimates (CA)

manipulations, we get the search function only about DOA, and perform a 1D search to obtain DOA estimates. Finally, we use the LS method to get the closed-form solution for the range estimates. In addition, 1D search about range can perform better than the LS method at the cost of increased computational complexity. Computer simulation results prove that the algorithm is effective regardless of the array configuration.

REFERENCES

[1] E. Fishler, A. Haimovich, R. Blum, D. Chizhik, L. Cimini, and R. Valenzuela, "MIMO radar: An idea whose time has come," in *Proceedings of the 2004 IEEE Radar Conference*, 2004, pp. 71–78.

[2] P. Antonik, M. C. Wicks, H. D. Griffiths, and C. J. Baker, "Frequency diverse array radars," in *2006 IEEE Conference on Radar*, IEEE, 2006, pp. 215–217.

[3] M. Secmen, S. Demir, A. Hizal, and T. Eker, "Frequency diverse array antenna with periodic time modulated pattern in range and angle," in *2007 IEEE Radar Conference*, IEEE, 2007, pp. 427–430.

[4] J. Xiong, W. Wang, and K. Gao, "FDA-MIMO radar range–angle estimation: CRLB, MSE, and resolution analysis," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 1, pp. 284–294, Feb. 2018.

[5] Y. Yan, J. Cai, and W. Wang, "Two-stage ESPRIT for unambiguous angle and range estimation in FDA-MIMO radar," *Digital Signal Processing*, vol. 92, pp. 151–165, Sep. 2019.

[6] T. Zhang, F. Xu, and S. A. Vorobyov, "Transmit energy focusing for parameter estimation in transmit beamspace slow-time MIMO radar," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2023, pp. 1–5.

[7] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[8] W. Tang, H. Jiang, and Q. Zhang, "Range-angle decoupling and estimation for FDA-MIMO radar via atomic norm minimization and accelerated proximal gradient," *IEEE Signal Processing Letters*, vol. 27, pp. 366–370, Mar. 2020.

[9] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM review*, vol. 38, no. 1, pp. 49–95, Mar. 1996.

[10] X. Wu, Y. Liu, and X. Jia, "Gridless target localization for FDA-MIMO radar with sparse arrays," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2023, pp. 1–5.

[11] M. Feng, Y. Yang, Q. Shu, and R. Yang, "An improved ESPRIT-based algorithm for monostatic FDA-MIMO radar with linear or nonlinear frequency increments," *IEEE Communications Letters*, vol. 25, no. 7, pp. 2375–2379, Apr. 2021.

[12] H. Chen, X. Tian, Z. Li, and X. Jiang, "Reduced-dimension target parameter estimation for conformal FDA-MIMO radar," *Journal of Radars*, vol. 10, no. 6, pp. 811–821, Dec. 2021.

[13] P. Pal and P. P. Vaidyanathan, "Nested arrays: A novel approach to array processing with enhanced degrees of freedom," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4167–4181, Aug. 2010.

[14] P. P. Vaidyanathan and P. Pal, "Sparse sensing with coprime arrays," in *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers*, 2010, pp. 1405–1409.

Collection of Correlated Information from Superimposed Multiple Chirp Signals

Koki Aoyama* and Koichi Adachi*
* Advanced Wireless and Communication Research Center, The University of Electro-Communications, Tokyo, Japan
E-mail: {aoyama, adachi}@awcc.uec.ac.jp

***Abstract*—Wireless Sensor Networks (WSNs) based on Long-Range Wide Area Networks (LoRaWAN) are being used in various applications. In WSNs, in addition to periodic packet transmission, sensor nodes transmit packets when they detect an event. This action enables real-time information aggregation. When multiple sensor nodes detect an event at the same time, they send signals simultaneously, which may result in signal collisions at the receiver. This paper proposes a symbol detection method that allows information aggregation even when the GW receives superimposed chirp signals. Furthermore, this paper introduces an event source estimation method as a use case for collecting aggregate correlated information. Computer simulation results show that the proposed method increases the probability of achieving the absolute source estimation error of 10 degrees by 11 points compared to conventional methods.**

I. Introduction

Advances in wireless communication technology have led to the use of the Internet-of-Things (IoT) in a variety of fields. In particular, the IoT-based Wireless Sensor Networks (WSNs) technology has attracted attention, and various research works have been conducted in both industrial and academic fields [1]. WSN is a network consisting of many sensor nodes with wireless communication capability and a GateWay (GW) that collects information from the sensor nodes. Sensor nodes consisting of WSN are preferred to be inexpensive because they are installed in many locations. Due to environmental constraints, they are usually powered by batteries. Therefore, maintenance costs and battery life are important in WSN, and communication systems with low power consumption are preferred [2]. One of the communication systems used in WSN is the Long-Range Wide Area Network (LoRaWAN). LoRaWAN enables long-range communication with low power consumption by adopting Chirp Spread Spectrum (CSS) modulation as its physical layer technology. Furthermore, since LoRaWAN operates in the unlicensed band, it is used in various WSN scenarios [3], [4].

Local regulators introduce several restrictions, including maximum Time-on-Air (ToA) per transmission and Duty Cycle (DC) to prevent uncontrolled use of the frequency bands. DC limits the total transmission time during a specific period. In addition, the battery-powered nature of the nodes makes it difficult for them to exchange information with the GW frequently. Thus, information collection in WSNs is often initiated by sensor nodes. The sensor nodes generally periodically transmit the generated packet containing the measured data to the GW. However, periodic transmission cannot handle real-time information collection, such as sudden environmental changes. Therefore, an *event-driven transmission* is introduced in addition to the periodic transmission. In event-driven transmission, sensor nodes transmit the measured data immediately once it satisfies the predetermined criteria. This event-driven transmission allows GWs to detect events and collect real-time information. However, when an event are observed by multiple sensor nodes, they send information to the GW at the same time, resulting in signal collisions. Since nodes in WSN generally do not retransmit signals to save battery, information about events cannot be collected if such signal collision happens.

One approach to this problem is to adjust the transmission timing of each sensor node to avoid signal collision through some techniques, including Carrier Sense (CS). However, CS and corresponding backoff may incur delays for sending nodes to transmit information to the GW, which takes long time to collect information from all nodes. Therefore, this approach is not suitable for LoRaWAN, which requires a long transmission time for each packet transmission. There is an early work that considers simultaneous transmission of multiple nodes such as over-the-air computing-based wireless information aggregation [5] and sequential interference cancellation [6]. These methods enable information collection from multiple nodes through simultaneous transmission. However, the former requires strict time synchronization, while the latter does not consider reception in situations where many nodes transmit the same data simultaneously. Since event-driven packets contain information about the same event detected by sensor nodes, the transmitted data is expected to be the same or at least highly correlated. Therefore, this paper proposes a symbol detection method for superimposed chirp signal waveforms that takes advantage of the characteristics of event-driven data. Through the computer simulation, the proposed method is shown to enable statistical acquisition of event information in a single transmission time. This paper also introduces event source estimation as one of the use cases for exploiting information aggregation through simultaneous transmission.

This paper is organized as follows. Section 2 presents the system model considered in this study. Section 3 describes the considered modulation method, LoRa and chirp signal superposition. Section 4 describes the proposed method. Section 5 shows the results of computer simulations, and Section 6 concludes the paper.

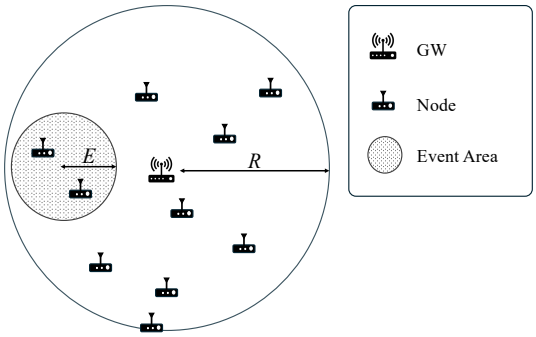


Fig. 1. System model

II. LoRa Modulation & Demodulation

LoRa modulation is based on CSS modulation, which enables long-distance communication with high noise immunity. LoRa signals are characterized by bandwidth W [Hz] and Spreading Factor (SF) S that indicates the number of bits transmitted by one CSS symbol. A high SF improves noise immunity and hence achieves longer communication distance at the cost of longer symbol length. In the following sections, we focus on the physical layer of LoRa and discuss the transmission and reception of chirp signals.

The base up-chirp signal, $\psi_0(t)$, at time t is given by

$$\psi_0(t) = \begin{cases} \frac{1}{\sqrt{2^S}} \left(j2\pi t^2 \frac{W^2}{2^{S+1}} \right) & 0 \le t < T_\mathrm{s} \\ 0 & \text{otherwise} \end{cases}, \qquad (1)$$

where $T_\mathrm{s} = 2^S/W$ [sec] is the symbol length. Different symbols are represented by cyclically shifting this base up-chirp signal in the time domain. Therefore, the transmitted signal $\psi_m(t)$ of symbol m at time t is expressed as

$$\psi_m(t) = \frac{1}{\sqrt{2^S}} \left(j2\pi \left((t + m \cdot T_\mathrm{c}) \bmod T_\mathrm{s} \right)^2 \frac{W^2}{2^{S+1}} \right), \qquad (2)$$

where $T_\mathrm{c} = 1/W$ [sec] is the chip length and mod denotes the modulo operation. In the following, we consider the case $0 \le t < T_\mathrm{s}$ only for notation simplicity.

Once a receiver receives a chirp signal, it performs *dechirping*, *Discrete Fourier Transform (DFT) operation*, and *symbol detection*, which are to be explained in Section III.B in detail.

III. System Model

Figure 1 shows the system model considered in this paper, where N nodes are randomly and uniformly placed within the area of interest with a radius of R [m] centered on the GW. An event occurs at a random location in the area. The nodes, which are located within a radius of E [m] centered on the event location, simultaneously detect the event. Upon detecting the event, a node generates a packet, performs CS, and transmits the generated packet to the GW if it does not detect the ongoing transmission. For simplicity, this paper assumes that nodes transmit event-driven packets only and DC constraints due to periodic transmission are not considered.

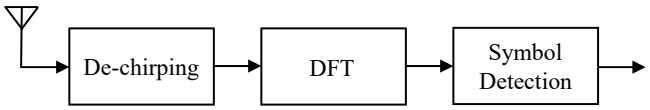


Fig. 2. Receiver model

A. Transmitter Model

Upon detecting the event, a sensor node immediately tries to send the measured data to the GW. The measured value is first quantized into a dicrete value using a predetermined quantization technique. The obtained quantized value is converted to a binary number, which is then modulated to a CSS symbol. This paper does not consider any whitening, channel coding, and interleaving [7]. Therefore, if there is a slight difference or error in the measured value, the neighboring CSS symbols are selected at each transmitter, resulting in multiple neighboring peaks of the DFT output at the receiver.

B. Receiver Model

The simplified receiver model considered in this paper is shown in Fig. 2, which consists of three steps: *de-chirping*, *DFT*, and *symbol detection*.

The chirp signal waveforms transmitted from multiple sensor nodes are superimposed upon the reception at the receiver. Assume that the symbol m is transmitted by multiple nodes (set $\mathcal{A}$) that detected the event. The superimposed signal $r[\ell]$ at the ℓth sampling point can be expressed as

$$r_m[\ell] = \sum_{i \in \mathcal{A}} \sqrt{\frac{P_{r,i}(d_{\mathrm{GW},i})}{2}} \psi_m[\ell - \tau_i] + w[\ell], \qquad (3)$$

where $P_{\mathrm{r},i}(d_{\mathrm{GW},i})$ is the received power of node $i \in \mathcal{A}$ located at $d_{\mathrm{GW},i}$ [m] away from the GW, $\psi_m(\cdot)$ is the transmitted chirp signal, and $w[\ell]$ is the additive white Gaussian noise (AWGN) that follows $\mathcal{CN}(0, N_0)$ (N_0 is the one-sided power spectral density of the noise).

The τ_i in Eq. (3) denotes the propagation delay for node i. Assume that the internal processing time of all transmitters is constant (or zero). Given event propagation speed v_event, distance between event source and node $d_{\mathrm{event},i}$, and distance between node and GW $d_{\mathrm{GW},i}$, τ_i is determined as follows

$$\tau_i = \frac{d_{\mathrm{GW},i}}{c} + \frac{d_{\mathrm{event},i}}{v_\mathrm{event}}, \qquad (4)$$

where c [m/sec] is the speed of light.

The received superimposed chirp signal is subjected to dechirping, DFT processing, and symbol detection.

a) De-chirping: The de-chirping is carried out as

$$r_m[\ell]\psi_0[\ell]^* = \left\{ \sum_{i \in \mathcal{A}} \sqrt{\frac{P_{\mathrm{r},i}(d_{\mathrm{GW},i})}{2}} \psi_m[\ell - \tau_i] + w[\ell] \right\} \psi_0[\ell]^*, \qquad (5)$$

where $\{.\}^*$ denotes the complex conjugate operation.

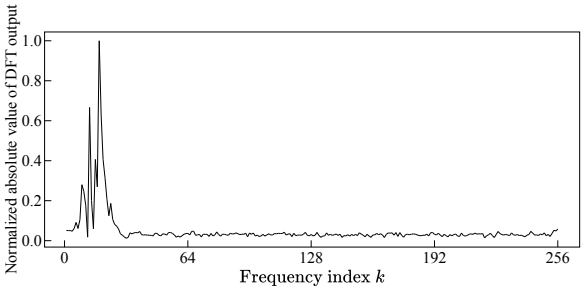


Fig. 3. DFT output spectrum of superimposed received chirp signals.

b) DFT: The kth frequency spectrum obtained by DFT is given by

$$F[k] = \sum_{\ell=0}^{2^S-1} r_m[\ell]\psi_0[\ell]^* \left(-j2\pi \frac{k}{2^S}\ell\right). \tag{6}$$

c) Symbol Detection: The symbol detection at an ordinary LoRa receiver is performed as

$$m^\star = \underset{k\in[0,2,\cdots,2^S-1]}{\operatorname{argmax}} |F[k]|. \tag{7}$$

Conventional symbol detection is optimized for the case where a signal transmitted by a single node is received. When multiple nodes transmit signals simultaneously, the difference in symbols transmitted by each node, propagation delays, and bias in node placement may result in incorrect symbol decision.

IV. Proposed Method

If multiple nodes transmit at the same time, it is likely that they detected the same event. Therefore, the information transmitted by each node is expected to be either the same or at least highly correlated. As described in the previous section, even if there is a measurement error or delay, multiple symbols arrive at the GW within a relatively short time interval. Therefore, if signals from multiple nodes are received simultaneously, the central limit theorem can be used to statistically collect information on events in a single communication time.

This paper proposes a symbol detection method suitable for superimposed chirp signals and an efficient information collection method. Specifically, it aims to improve the symbol detection process after the DFT process to achieve statistically correct symbol detection even when multiple chirp signals are superimposed. We also present an event source estimation as a use case of the proposed symbol detection method.

A. Symbol Estimation

It is highly likely that the DFT output, which usually exhibits a single strong peak, may have multiple peaks within some frequency range due to delay or measurement error. Therefore, the spectrum with the width of the DFT output is used to estimate the transmitted symbols of the superimposed chirp signal. Symbol estimation is difficult because the DFT output is affected by noise in addition to the effect of the superimposed chirp signal. Thus, the proposed symbol estimation method consists of three steps: *noise suppression*, *smoothing*, and *peak width-based symbol detection*.

a) Noise Suppression Step: The noise suppression process is expressed as

$$\alpha[k] = \begin{cases} |F[k]| & \text{if } F[k] > \text{mean}\,(F) \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

Since the noise is spread over the entire frequency bandwidth after the dechirping process, the spectrum with a small amplitude value represents highly likely the noise. Thus, this paper uses the average value of DFT output as the threshold to cut out the noise.

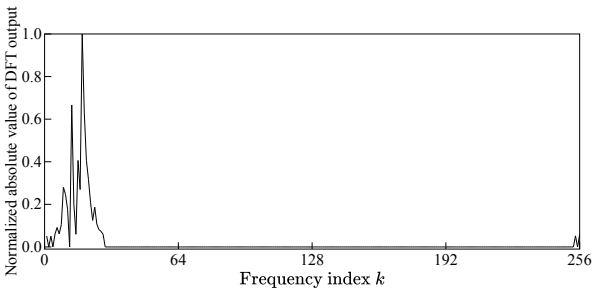
b) Smoothing Step: The smoothing is expressed by

$$\mu_n[k] = \frac{1}{n}\{\alpha[k] + \sum_{j=1}^{(n-1)/2} (\alpha[(k+j)_{\bmod 2^S}] + \alpha[(k-j)_{\bmod 2^S}])\}. \tag{9}$$

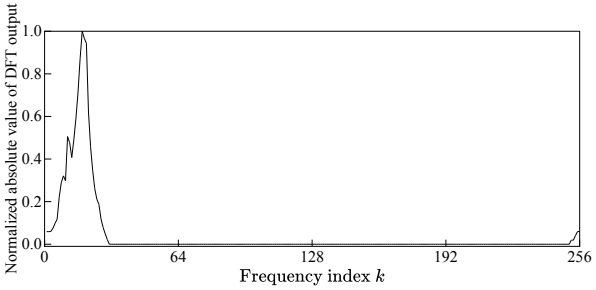
For a superimposed chirp signal, the DFT output exhibits several consecutive peaks. However, if a small number of sensor nodes transmit simultaneously, the DFT output may have nonconsecutive peaks. Therefore, a smoothing process with a tap number of T is applied to supplement the peak continuity. As the number of smoothing taps, T, increases, the spectrum becomes smoother. However, when the received Signal-to-Noise Ratio (SNR) is low, spectra other than the received signal may appear in the DFT output because the noise suppression process cannot completely suppress the noise. In such cases, if the number of smoothing taps, T, is large, the peaks of the desired signal and those due to noise are indistinguishable, and the estimation accuracy degrades. Thus, the number of taps, T, in the smoothing process may have a significant impact on symbol detection performance.

c) Symbol Detection Step: The post-smoothed spectrum of the superimposed chirp signal is shown in Fig. 4. The figure shows that, in addition to the desired symbol spectrum, some other spectra have relatively large amplitude values due to noise. Therefore, it is necessary to distinguish between the desired signal's spectrum and the noise spectrum. To overcome this difficulty, this paper proposes *peak width-based detection*. The peak width-based detection algorithm is shown in Algorithm 1. Since a LoRa signal is cyclic because it is represented by cyclic shifting the base up-chirp signal, the first half of the DFT spectrum is copied and added to the second half, and then the symbol is determined.

When multiple chirp signals are simultaneously transmitted from multiple sensor nodes, there are time shifts between the received signals due to propagation delay. Thus, the DFT output peak has a width rather than a sharp peak at the desired symbol index, which indicates that the spectra with the largest peak width contain the desired signal's spectrum. Thus, in the proposed peak width-based detection, the DFT spectrum with the largest width is detected as the target spectrum. When only



(a) before smooth process



(b) after smooth process

Fig. 4. Smooth process

a single signal is received or when the propagation delay is extremely small, the DFT output peak appears as a single peak. To ensure correct operation even in such cases, when the peak widths of the spectra are the same, the strongest peak is considered as the target spectrum. After deciding the target peak width, the symbol at the center of the peaks is output as the estimated symbol based on the assumption that the superimposed chirp signal obeys the central limit theorem.

B. Event Source Estimation

The proposed method enables symbol detection even for superimposed signals if the correlated information is transmitted. Thus, the receiver cannot detect the data if there is no correlation among the transmitted data such as device IDs. Therefore, this paper considers a way to let the transmitted data have some correlation. In this paper, one method of correlating data is to use the fact that the nodes that detect an event source are located in a circle around the source of the event. Specifically, the device ID is assigned based on the azimuth angle viewed from the GW as

$$\mathrm{dev.ID_H} = \left\lceil \frac{\theta}{2\pi} 2^S \right\rceil . \tag{10}$$

Since there are 3 bytes address prefixes for the device by default as shown in Section V-B, the first 1 byte of the address prefix is assigned to each azimuth angle and the last 2 bytes are used to represent a unique address.

Algorithm 1 Symbol detection :Peak width detection

function PEAKWIDTHDETEC($\mu = [\mu_0, \mu_1, \cdots, \mu_{2^S-1}]$)
  $\mathcal{S} \leftarrow \mathrm{vcat}(\mu, \mu[1 : 2^{S-1}/2])$
  $flag \leftarrow 0$
  for all $index \leftarrow \mathcal{S}$ **do**
    if $\mathcal{S}[index]! = 0$ **then**
      if $flag == 0$ **then**
        $flag \leftarrow 1$
        $tempStart \leftarrow index$
      end if
    else
      if $flag == 1$ **then**
        $flag \leftarrow 0$
        $tempWidth \leftarrow index - tempStart$
        if $tempWidth > indexWidth$ **then**
          $startIndex \leftarrow tempStart$
          $indexWidth \leftarrow tempWidth$
        else if $tempWidth == indexWidth$ **then**
          if $\max(\mathcal{S}[tempStart : index]) > \max(\mathcal{S}[startIndex : startIndex + indexWidth])$ **then**
            $startIndex \leftarrow tempStart$
            $indexWidth \leftarrow tempWidth$
          end if
        end if
      end if
    end if
  end for
  $Symbol \leftarrow startIndex + \mathrm{ceil}(indexWidth/2)$
  return $Symbol \mod 2^S$
end function

V. Simulation and Results

A. Propagation Model

This paper adopts a static channel model that considers path loss and shadowing. The received power, $P_{\mathrm{r},i}(d_{\mathrm{GW},i})$, of node i at GW is calculated as

$$P_{\mathrm{r},i}(d_{\mathrm{GW},i}) = P_{\mathrm{t}} - L_{\mathrm{path}}(d_{\mathrm{GW},i}) - L_{\mathrm{s}}, \tag{11}$$

where P_{t} [dBm] is the common transmit power for all nodes, $L_{\mathrm{path}}(d_{\mathrm{r},i})$ [dB] is space attenuation due to path loss, L_s [dB] is variation due to shadowing. Shadowing is a model that follows a log-normal distribution with spatial correlations [8].

Path loss $L_{\mathrm{path}}(d_{\mathrm{GW},i})$ is calculated using the ITU-R model given by [9]

$$L_{\mathrm{path}}\left(d_{\mathrm{GW},i}\right) = 10\alpha \log_{10}\left(d_{\mathrm{GW},i}\right) + \beta + 10\gamma \log_{10}\left(f_{\mathrm{c}}\right), \tag{12}$$

where α, β, γ are parameters related to path loss and f_{c} is the center frequency of the CSS modulation signals.

B. Frame Format

The LoRaWAN frame format is shown in Fig. 5. The communication standard for LoRaWAN is defined by the LoRa

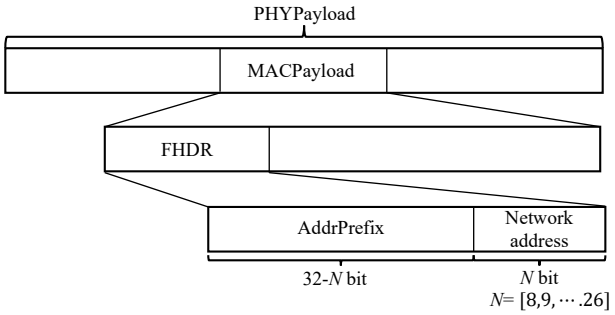


Fig. 5. Frame Format

Alliance, and the frame format shall also conform to the standard [10]. The device address (4 bytes) is stored in the MAC payload in the frame header.

As shown in Fig. 5, the device address consists of a node unique AddrPrefix and a network address totaling 32 bits and is unique within the network. The length of the AddrPrefix is determined by the length of the network address, which is set to 8 bits (1 byte) [10].

C. Simulation Parameters and Performance Metrics

This section provides computer simulation results to show the effectiveness of the proposed method. The simulation parameters are shown in Table 1. The path loss parameters are set to $\alpha = 4.0$, $\beta = 9.5$, and $\gamma = 4.5$, respectively, assuming communication in a Non-Line-of-Sight (NLoS) environment [9]. Symbol synchronization during reception is assumed to be ideal for the first received signal. In this paper, we adopt the Cumulative Distribution Function (CDF) to characterize the change in information aggregation accuracy with the number of concurrent aggregation nodes and the absolute estimation error of the event source estimation. For the change in information aggregation accuracy with the number of concurrently collected nodes, we assume the mean $\mu = 100$ and variance $\sigma^2 = 5$ due to measurement error, and verify the information collection accuracy when the number of concurrently collected nodes is set to $N = \{1, 10, 100, 1000\}$. For the event source estimation, we evaluate the estimation results and the CDF of the estimation error after 10^6 trials. As a benchmark method, the DFT peak detection given by Eq. (7) is used as the symbol detection method.

D. Simulation Results

Figure 6 shows the impact of the number of transmitting nodes on the absolute error of the event source estimation. The figure shows that information collection accuracy increases with the number of superimposed LoRa signals.

Figures 7(a) and 7(b) show the event source estimation results of the benchmark method and those of the proposed method. Figure 7(b) shows that the proposed method estimate the center of the event source better than the conventional method. Because the proposed method is not affected by the DFT output peak bias due to the superimposed chirp signal.

TABLE I
SIMULATION PARAMETERS

Parameters	Value
Center Frequency f_c	920 [MHz]
Band Width W	125 [kHz]
Spreading Factor S	8
Transmit Power P_t	13 [dBm]
Noise power spectrum density N_0	-174 [dBm/Hz]
Shadowing standard deviation σ^2	7.6
Transmit symbol	10 [Symbol]
Number of GW	1
Number of ENs	1,000
Simulation area radius R	800 [m]
Event area radius E	200 [m]
Number of trials	10^6

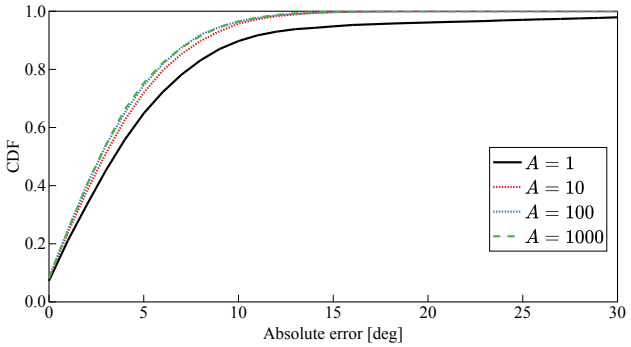
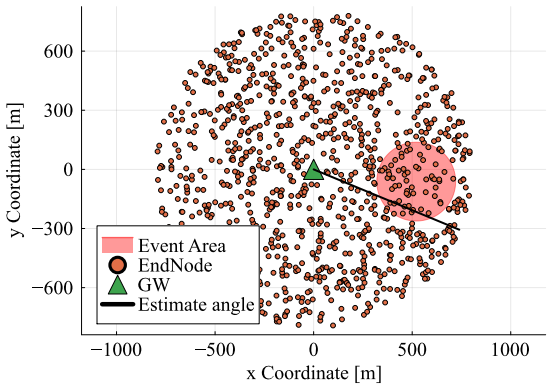


Fig. 6. Impact of the number of superimposed chirp signals on estimation accuracy ($A = |\mathcal{A}|$).

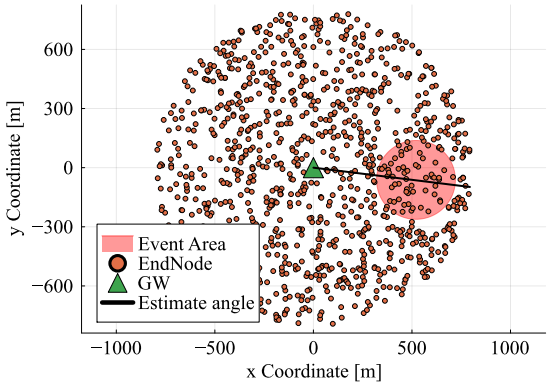
The CDF of the absolute estimation error of the proposed method with varying the number of smoothing taps, T, is shown in Fig. 8. Figure 8 shows that the proposed method outperforms the benchmark method in estimation accuracy; specifically, the proposed method with $T = 5$ taps improves the CDF value by 11 points at ± 10 degrees and by 5 points at ± 15 degrees compared to the benchmark method. The reason why the estimation accuracy differs depending on the number of taps T lies in the fact that the peak width is detected to be wide due to noise or that the smoothing of peaks is insufficient, and the number of taps T in the smoothing process needs to be adjusted according to the average received SNR. In addition, both the benchmark and the proposed methods show an absolute error of ± 30 degrees or more, and there are cases where the event source estimation does not work correctly. This may be due to the fact that the sensor nodes detecting the event are located in the vicinity of the GW and the superimposed signals are composed of uncorrelated data. In the case where the event happened on edge to distributions of end node, estimation accuracy degrades due to SNR degradation, but estimation itself can be performed.

VI. CONCLUSION

This paper examined statistical information collection from superimposed chirp signals transmitted by the nodes that have detected the event. Since the delay of the superimposed signal



(a) Benchmark method



(b) Proposed method

Fig. 7. An example of event source estimation result.

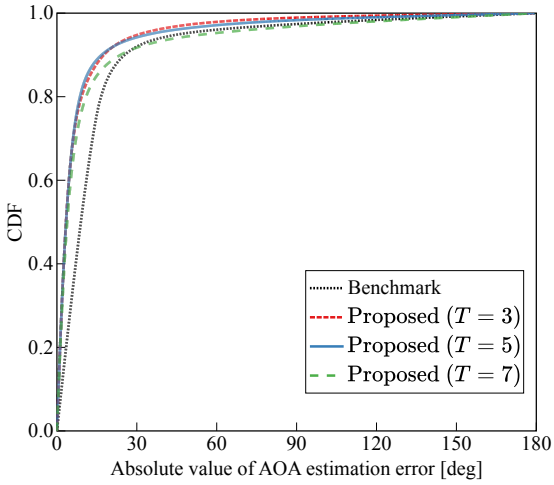


Fig. 8. Impact of smoothing tap T on the absolute estimate error.

degrades the accuracy of information collection, this paper proposed a symbol detection method using the peak width of the DFT spectrum, and event source estimation was identified as a use case. Computer simulation results have shown that the proposed method improves the absolute error in event source estimation by up to 11 points for a CDF value of 10 degrees compared to the conventional symbol detection method, thus improving the accuracy of information collection.

REFERENCES

[1] K. Thirunavukkarasu, L. Raju, and S. Sathishbabu, "A Survey on LoRaWAN for Smart Medical and Industries," in *2023 9th Int. Conf. on Electrical Energy Systems (ICEES)*, 2023, pp. 34–40. DOI: 10.1109/ICEES57979.2023.10110099.

[2] C. Milarokostas, D. Tsolkas, N. Passas, and L. Merakos, "A Comprehensive Study on LPWANs With a Focus on the Potential of LoRa/LoRaWAN Systems," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 1, pp. 825–867, 2023. DOI: 10.1109/COMST.2022.3229846.

[3] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low Power Wide Area Networks: An Overview," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 2, pp. 855–873, 2017. DOI: 10.1109/COMST.2017.2652320.

[4] W. Ayoub, A. E. Samhat, F. Nouvel, M. Mroue, and J.-C. Prévotet, "Internet of Mobile Things: Overview of LoRaWAN, DASH7, and NB-IoT in LPWANs Standards and Supported Mobility," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 2, pp. 1561–1581, 2019. DOI: 10.1109/COMST.2018.2877382.

[5] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-Bit Over-the-Air Aggregation for Communication-Efficient Federated Edge Learning: Design and Convergence Analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, 2021. DOI: 10.1109/TWC.2020.3039309.

[6] J. Bukhari and Z. Zhang, "Simple Peak Interference Cancellation (SPIC): Interference Cancellation Prior to Packet Decoding in LoRa Networks," in *2023 IEEE Latin-American Conf. on Commun. (LATINCOM)*, 2023, pp. 1–6. DOI: 10.1109/LATINCOM59467.2023.10361900.

[7] M. Jouhari, N. Saeed, M.-S. Alouini, and E. M. Amhoud, "A Survey on Scalable LoRaWAN for Massive IoT: Recent Advances, Potentials, and Challenges," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 3, pp. 1841–1876, 2023. DOI: 10.1109/COMST.2023.3274934.

[8] Z. Wang, E. K. Tameh, and A. R. Nix, "Joint Shadowing Process in Urban Peer-to-Peer Radio Channels," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 52–64, 2008. DOI: 10.1109/TVT.2007.904513.

[9] ITU-R Recommendation P.1411-12, "Propagation data and prediction methods for the planning of short-range outdoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 100 GHz," Sep. 2021.

[10] LoRa Alliance, "LoRaWAN Certification Protocol Specification TS009-1.0.0," 2020.

A Study on Packet-Level Index Modulation Using Frequency Offsets within a LoRaWAN Channel

Mai Ohta* Hiroki Matsuura†‡ and Takeo Fujii§
* Fukuoka University, Fukuoka, Japan
E-mail: maiohta@fukuoka-u.ac.jp
† NATANE eICT Lab., Tokyo, Japan
‡ The University of Electro-Communications, Tokyo, Japan
§ Advanced Wireless & Communication Research Center, The University of Electro-Communications, Tokyo, Japan

***Abstract*—Low Power Wide Area (LPWA), which is increasingly utilized in Internet of Things (IoT) systems, is attracting significant attention. In LPWA systems, Long Range Wide Area Network (LoRaWAN) is one of the most widely adopted communication methods for IoT systems. Since LoRaWAN relies on long range modulation, increasing the data rate poses significant challenges. Therefore, this study addresses the limitations in the number of channels owing to the restricted number of transmission bits by proposing a Packet-level index modulation (PLIM) scheme. The frequency bands allocated to the LPWA networks vary by region and country, and the number of channels is limited. To increase the number of transmission bits as effectively as possible, this study proposes the PLIM scheme, leveraging frequency offsets within the LoRaWAN channel.**

I. Introduction

In this study, we focus on a low-power wide area (LPWA) system, which is employed in Internet-of-Things (IoT) systems, and has become increasingly popular. In LPWA systems, various communication schemes have been proposed to achieve low power consumption and long-distance communication. LoRaWAN, which is used as an LPWA communication scheme, utilizes a chirp-spread spectrum modulation (CSS) called LoRa modulation. A higher spreading factor in LoRa modulation leads to longer transmitted signals, better tolerance to interference, and implements quasi-orthogonality between different spreading factors. LoRaWAN is a standard specified by the LoRa Alliance [**1**], [**2**].

Packet-level index modulation (PLIM) has been proposed to add transmission bits to the time-slot and channel indexes of packets with LoRaWAN [**3**]. In a PLIM system, time synchronization between the transmitter and receiver must be performed in advance when information is added to the time-slot index. If the start time of a frame in the PLIM system is not synchronized between the transmitter and receiver, the time-slot index cannot be correctly detected, resulting in PLIM symbol errors. Despite time synchronization, if the oscillator in the microcontroller regulating the timing of packet transmission at the node is inaccurate, the effect of clock drift accumulates over time, requiring its estimation and periodic correction of the clock drift. Clock drift estimation is performed on the base station side, where computational resources are less limited. The clock drift values generated at the node are modeled as a normal distribution with mean μ and variance σ^2, and the estimated values are calculated using the received packets [**4**].

However, it is widely known that oscillators generate clock drifts that vary with temperature. It has been suggested that the mean value of the normal distribution model of the clock drift varies significantly with changes in temperature, causing errors when demodulating the PLIM symbols [**5**]. When the interval between packet transmissions is long, estimation and correction become challenging, owing to the corresponding increase in packet reception intervals. However, the PLIM system requires an increase in the number of indexes and a longer frame length to increase the number of bits in the PLIM symbol. The number of frequency indexes cannot be easily increased, owing to the limitation in the number of channels to the originally allocated bandwidth. Although the number of time-slot indexes can theoretically be increased indefinitely, in practice, the data transmission interval of packets tends to be system-dependent. Therefore, the aforementioned clock drift estimation is most effective when performed at transmission intervals in which the temperature is kept constant.

As mentioned above, the number of time-slot indexes can be increased by increasing the frame time length, thereby increasing the number of bits that can be transmitted. However, processing tasks, such as time synchronization, are necessary, and studies have been conducted on estimation and correction techniques for addressing clock drift caused by inaccurate oscillators. Therefore, this study focuses on channel indexes that have been strictly limited by the number of channels, and we propose a new method to increase the number of indexes within each channel bandwidth. This study assumes that the transmitter and receiver use LoRaWAN to transmit packets. The proposed method realizes PLIM using frequency offsets within a LoRaWAN channel.

II. Previous Work

In this paper, a packet-level index modulation (PLIM) method is focused because the number of conveyed bits can increase. In PLIM, both the time-slot and channel indexes are assigned to different PLIM symbols. However, the detection of the time-slot index in PLIM is affected by inaccurate oscillators.

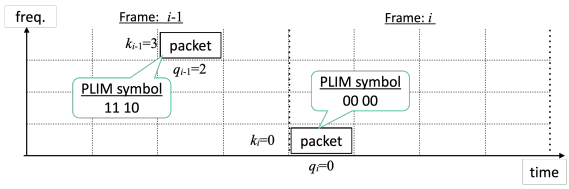


Fig. 1. Image of PLIM at a transmitter.

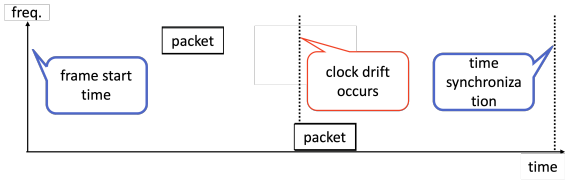


Fig. 2. Image of PLIM at a receiver.

A. Packet-Level Index Modulation (PLIM)

It is difficult to increase the data rate in a LoRaWAN using LoRa modulation by devising a modulation scheme owing to the low channel frequency, narrow bandwidth, and the fact that DC regulations must be strictly adhered to. In LPWA systems, when using modulation schemes such as LoRaWAN modulation, where data rate improvement is challenging, PLIM has been proposed as a method for increasing the number of transmission bits [**3**]. PLIM, wherein symbols are assigned to the time-slot and channel indexes as packets, enables nodes to increase the number of transmitted bits in addition to the amount of information in the packet. The time-slot and channel indexes to which a PLIM symbol is allocated can be designed as a frame such that the packet length to be transmitted relative to the frame length satisfies the duty cycle (DC) regulations.

As described above, the PLIM can transmit more information than the packet transmission volume without changing the modulation scheme of the packet itself, while satisfying DC regulation. Therefore, the practical application of PLIM is expected to be relatively simple. However, the nodes used in IoT and other applications are expected to be numerous and forming large networks, while also being cost-effective. Thus, it is expected that the circuit configuration or individual elements constituting a node will be affordable and have a minimum performance guarantee. In the PLIM system, a particular concern is regarding the accuracy of the oscillator in the microcontroller that controls the timing of transmission, which is a factor in PLIM symbol errors.

In PLIM, symbols are assigned to time slot and channel indexes. A PLIM frame is defined as a period in which a packet transmits in time domain and channel domain. Nodes transmit a packet at the corresponding time slot and channel to convey PLIM symbols.

B. Symbol demodulation in PLIM

Time synchronizations between a transmitter and receiver are performed in advance. Therefore, the receiver can demodulate PLIM symbols by detecting the time slot and channel indexes. The channel index is detected from the channel frequency of the received packet. In addition, the time slot index is detected from the frame start time and the packet reception time. In the time slot index detection, PLIM symbol in time domain is given by

$$\hat{q}_i = \frac{t_{\mathrm{r}} - t_{\mathrm{FST}}}{T_{slot}}, \tag{1}$$

where i is the frame number, t_{r} is the reception time of the packet, t_{FST} is the frame start time, and T_{slot} is the slot length.

In the time domain, a clock drift occurs because nodes are operated by a microcontroller without GPS system. The clock drift at the transmitter causes PLIM demodulation errors at the receiver. In general, the impact on the PLIM errors is greater for clock drift by the microcontroller than by the radio. Therefore, the receiver in PLIM periodically performs time synchronization.

III. PLIM Using Frequency Offsets with LoRaWAN

The proposed method is based on PLIM using LoRaWAN. LoRaWAN, which is an LPWA network, is used for IoT. In general, IoT nodes consist of inaccurate elements and devices owing to their large number of nodes. When using the PLIM, a clock drift is caused by the inaccurate oscillator of the transmit node because time synchronization is required to derive the PLIM time-slot index from the packet's received time.

However, IoT nodes are stricter on the frequency offset than on the clock drift because there are strict rules regarding microwave emissions. The signals were emitted by a wireless device equipped with a higher-precision oscillator than that of the microcontroller equipped for controlling the transmission timing. Therefore, we propose a PLIM method that uses frequency offsets without time synchronization, because it is assumed that the packet is modulated by LoRaWAN.

In the basic PLIM method, PLIM symbols is assigned to the time-slot and channel indexes. By contrast, in our proposed method, PLIM is performed by shifting the center frequency of a LoRaWAN signal within the channel bandwidth. In this proposed method, PLIM symbols assign to the amount of frequency offsets. This section presents the spreading and despreading methods, and the proposed method of frequency offset mapping using packet-level index modulation.

First, the basic LoRa modulation method is explained because it is an important technology for the method proposed in this study.

A. LPWA and LoRa Modulation

It is necessary to achieve low-power communication because IoT systems are assumed to be battery powered. Consequently, the LPWA is highly compatible with IoT systems, where it is difficult to ensure a sufficient power supply for each node because the LPWA has a small transmission power of approximately 20 mW and assumes low-power communication. Moreover, to achieve long-distance communication, LPWA systems are allocated a relatively low-frequency band of 800 MHz to 900 MHz, and the allocated frequency band differs

Preamble	Sync word	Payload (PHDR+CRC+PHY Payload)	CRC (only up link)

Fig. 3. Packet structure in LoRa PHY layer.

depending on the region or country. In the LoRaWAN, signal bandwidths of 125, 250, and 500 kHz were used, and when the bandwidth was 125 kHz, the channel center frequencies were specified at 200 kHz intervals.

LoRaWAN uses CSS, which continuously changes the frequency by performing a cycle within the channel bandwidth. The CSS in a LoRaWAN system is called LoRa modulation. The LoRaWAN system selects and uses a channel with a bandwidth of 200 kHz and can determine the LoRaWAN signal bandwidth at 125, 250, and 500 kHz. If the bandwidth of the LoRaWAN signal was 125 kHz, a guard band of 37.5 kHz was provided at both ends of each channel.

In LoRaWAN, packets are modulated using LoRa modulation, and the preamble of the signal is added to the head of the packets. The packet structure of the LoRa physical layer is illustrated in Fig. 3. LoRaWAN symbols are spread and modulated using CSS modulation. If the basic chirp is an up-chirp, the sync word consists only of down-chirps and the other symbols are up-chirps.

The received LoRaWAN signals were demodulated by de-spreading with an inverse-basic-chirp, and the LoRa symbols were then obtained by FFT or DFT processing. In this subsection, the theories of spread processing are explained.

1) Spreading on LoRaWAN: The transmission LoRaWAN signal is obtained by multiplying the waveform $w_{\mathrm{S}}(t)$ of the symbols before spreading by the basic waveform $w_b(t)$ while spreading. The baseband waveform is as follows:

$$w(t) \quad = \quad w_{\mathrm{S}}(t) \cdot w_b(t), \tag{2}$$

where, $w_{\mathrm{S}}(t)$ is the LoRaWAN signal and is also the target signal to be spread. By contrast, $w_b(t)$ is the waveform used to spread the target signal, that is, the basic chirp. $w_{\mathrm{S}}(t)$ is given by

$$w_{\mathrm{S}}(t) \quad = \quad \exp(j2\pi \cdot D\frac{t}{2^{SF}T}) \tag{3}$$

$$= \quad \exp(j2\pi \cdot D\frac{B}{2^{SF}}t), \tag{4}$$

where j is the imaginary unit, T is the sampling interval, and D is the index of the data that can be transmitted with the SF [bit]. D is expressed as follows:

$$D = 0, 1, .., 2^{SF} - 1. \tag{5}$$

Moreover, T is rewritten as $T = 1/B$ using the sampling frequency B, which is equal to the LoRaWAN signal bandwidth (i.e., the chirping bandwidth). In the LoRaWAN, the sampling frequency and the LoRa signal bandwidth were set equal to facilitate FFT processing and LoRaWAN symbol detection. Consequently, the length of one LoRaWAN symbol is equal to the length of 2^{SF} samples (one symbol length $= 2^{SF} \cdot T$), because the sampling frequency B determines the sample length T. When performing FFT processing for 2^{SF} points to the waveform of the baseband signal $w(t)$, the bin with the largest value is the Dth FFT bin, indicating that demodulation is possible by performing FFT processing of the signal before the spectrum spreads.

The LoRaWAN symbols to be transmitted are spread using an upchirp of 2^{SF} sample length. The up-chirp changes linearly in frequency from $-B/2$ [Hz] to $B/2$ [Hz] during the symbol length $2^{SF} \cdot T$. Therefore, the angular frequency at time t is expressed as follows:

$$\omega(t) \quad = \quad 2\pi \left(\frac{B}{2^{SF} \cdot T}t - \frac{B}{2} \right) \tag{6}$$

$$= \quad 2\pi \left(\frac{B^2}{2^{SF}}t - \frac{B}{2} \right) \text{ [rad/s]}. \tag{7}$$

Therefore, the phase of the baseband signal with an angular frequency $\omega(t)$ described in Eq. (7) is as follows:

$$\phi(t) \quad = \quad \int_0^t \omega(t)\mathrm{d}t \tag{8}$$

$$= \quad 2\pi \left(\frac{B^2}{2 \cdot 2^{SF}}t^2 - \frac{B}{2}t \right) \text{ [rad]}. \tag{9}$$

In addition, the baseband signal of the waveform underlying the spread spectrum can be written as

$$w_b(t) \quad = \quad \exp(j\phi(t)) \tag{10}$$

$$= \quad \exp\left\{ j2\pi \left(\frac{B^2}{2 \cdot 2^{SF}}t^2 - \frac{B}{2}t \right) \right\}. \tag{11}$$

By Eq. (2), the transmitted LoRa signal is rewritten as

$$w(t) \quad = \quad \exp\left\{ j2\pi \left(D\frac{B}{2^{SF}}t + \frac{B^2}{2 \cdot 2^{SF}}t^2 - \frac{B}{2}t \right) \right\} \tag{12}$$

Note that when the frequency of the transmitted signal that is the up-chirp exceeds by $B/2$ [Hz], it starts from $-B/2$ [Hz] again, because the frequency bandwidth of the transmission signals is limited from $-B/2$ [Hz] to $B/2$ [Hz]. However, if the sampling frequency is B [Hz], the sampling theorem allows the equation above to be used without modification.

2) De-spreading on LoRaWAN: The signal received at the receiver is the transmitted waveform plus the change in phase θ owing to the transmission path, as shown in the following equation:

$$\hat{w}(t) \quad = \quad w(t) \cdot \exp(j\theta). \tag{13}$$

Here, changes in amplitude were not considered.

In de-spreading, by multiplying the received LoRa signal, $\hat{w}(t)$, and the inverse of the baseband signal, $1/w_b(t)$, the receiver can obtain the desired waveforms, as shown in

$$\hat{w_{\mathrm{S}}}(t) \quad = \quad w(t) \cdot \exp(j\theta) \cdot \frac{1}{w_b(t)} \tag{14}$$

$$= \quad w_{\mathrm{S}}(t) \cdot w_b(t)\frac{1}{w_b(t)} \cdot \exp(j\theta) \tag{15}$$

$$= \quad \exp(j2\pi \cdot D\frac{B}{2^{SF}}t) \cdot \exp(j\theta). \tag{16}$$

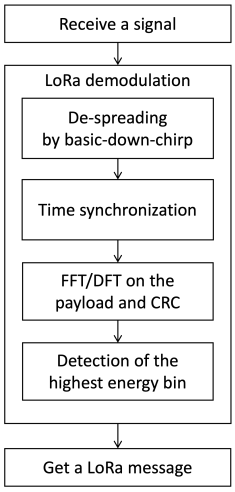


Fig. 4. Flow of the original LoRa demodulation at the receiver using LoRaWAN.

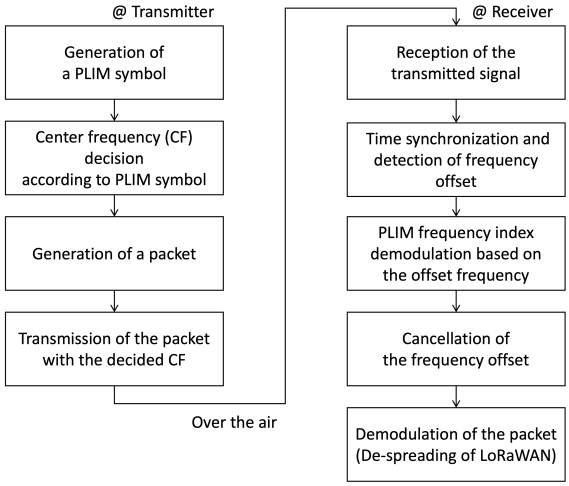


Fig. 5. Flow from the transmitter to the receiver, in the proposed method.

Note that $1/w_b(t)$ is a basic down-chirp because the positive and negative phases of $w_b(t)$ are reversed. Using the de-spread waveform expressed by Eq. (16), when performing the FFT for 2^{SF} points, the receiver obtains the result with the maximum absolute value in the Dth FFT bin, and the transmitted data can then be demodulated.

Fig. 4 illustrates the LoRa demodulation flow at the receiver. In LoRa demodulation, sync words consisting of basic-down chirps are used for synchronization in the time domain because the receiver can detect the peaks of the sync words by de-chirping the received signal.

B. Procedure of LoRaWAN with Frequency Offset Using PLIM

Fig. 5 illustrates the flow of the proposed method from the transmitter to the receiver. Fig. 6 illustrates the procedure for the proposed method at the receiver.

In the proposed scheme, the transmitter selects a channel and determines the offset according to the PLIM symbol. The signal was transmitted at the center frequency by considering

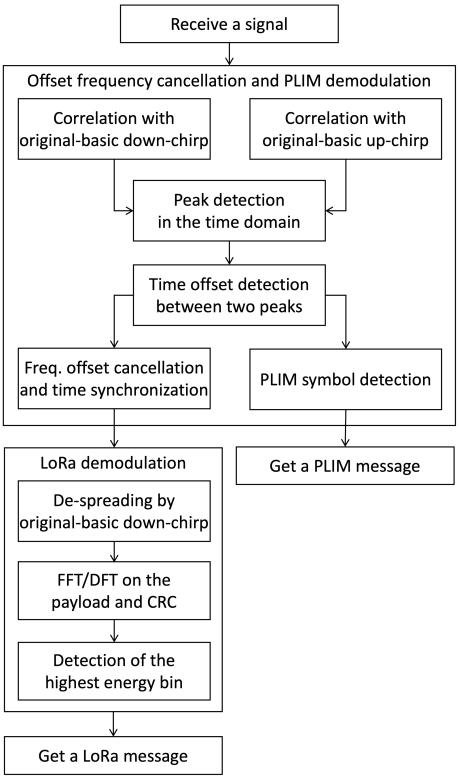


Fig. 6. Procedure of the proposed method.

the offsets. At the receiver, the PLIM symbol is demodulated by detecting the offset of the received signal. Moreover, the data transmitted by the packet can be obtained by processing LoRa demodulation after the cancellation of its detected offset.

C. Frequency Offset Acquisition by Correlation Peaks

The LoRa modulation scheme uses a basic-up-chirp in the preamble and basic-down-chirp in the sync word. In this subsection, we demonstrate that the offset of the transmission frequency can be obtained from the peak of the correlation function of the received up- and down-chirps.

The basic-up-chirp in the preamble is described by

$$w_u(t) = w_b(t) \tag{17}$$

$$= \exp\left\{j2\pi\left(\frac{B^2}{2\cdot 2^{SF}}t^2 - \frac{B}{2}t\right)\right\}. \tag{18}$$

The basic-down-chirp in the sync word is given by

$$w_d(t) = {w_b}^*(t) \tag{19}$$

$$= \exp\left\{j2\pi\left(-\frac{B^2}{2\cdot 2^{SF}}t^2 + \frac{B}{2}t\right)\right\}, \tag{20}$$

where * denotes the complex conjugation.

Fig. 7 shows an image of the original basic chirp and the shifted-chirps in the proposed method. The transmitter shifts the frequencies of the up-chirp and down-chirp by $f_{\rm os}\,[\rm Hz]$. Because $f_{\rm os}\,[\rm Hz]$ is the frequency offset for transmitting the

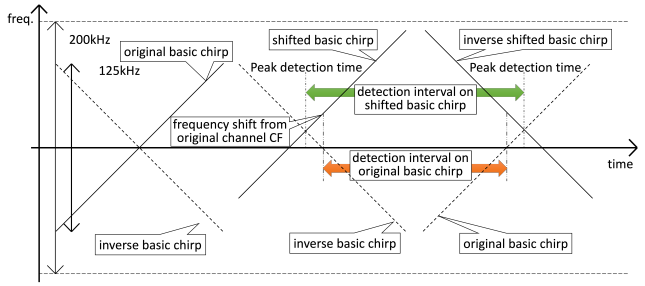


Fig. 7. Image of original-basic-chirp and shifted-chirp of the proposed method.

PLIM data, the transmission waveforms of the up- and down-chirps with offsets are as follows:

$$w_{u_{\rm os}}(t) = w_u(t)\cdot\exp(j2\pi f_{\rm os}t) \quad (21)$$

$$= \exp\left\{j2\pi\left(\frac{B^2}{2\cdot 2^{SF}}t^2-\frac{B}{2}t+f_{\rm os}t\right)\right\}, \quad (22)$$

and

$$w_{d_{\rm os}}(t) = w_d(t)\cdot\exp(j2\pi f_{\rm os}t) \quad (23)$$

$$= \exp\left\{j2\pi\left(\frac{B^2}{2\cdot 2^{SF}}t^2-\frac{B}{2}t+f_{\rm os}t\right)\right\}. \quad (24)$$

Moreover, the received signal of the up-chirp, whose frequency is offset by $f_{\rm os}$ [Hz] is described by

$$\hat{w_{u_{\rm os}}}(t) = w_{u_{\rm os}}(t)\cdot\exp(j2\pi f_{\rm os}t)\cdot\exp(j\theta) \quad (25)$$

$$= \exp\left\{j2\pi\left(\frac{B^2}{2\cdot 2^{SF}}t^2-\frac{B}{2}t+f_{\rm os}t\right)+j\theta\right\}. \quad (26)$$

The received signal of the downchirp, whose frequency is offset by $f_{\rm os}$ [Hz] is described as

$$\hat{w_{d_{\rm os}}}(t) = w_{d_{\rm os}}(t)\cdot\exp(j2\pi f_{\rm os}t)\cdot\exp(j\theta) \quad (27)$$

$$= \exp\left\{j2\pi\left(-\frac{B^2}{2\cdot 2^{SF}}t^2+\frac{B}{2}t+f_{\rm os}t\right)+j\theta\right\}. \quad (28)$$

Note that these waveforms exist in the range $0 \le t \le 2^{SF}/B$; however, for simplicity, the defined range is not considered.

The correlation function for the up-chirp is given by

$$R_u(t) = \int_{-\infty}^{\infty} \hat{w_{u_{\rm os}}}(\tau)\cdot {w_u}^*(\tau-t)\mathrm{d}\tau. \quad (29)$$

The correlation function $R_u(t)$ has a constant absolute value for $\hat{w_u}(\tau)$ and constant value for ${w_u}^*(\tau-t)$. Therefore, the absolute value of $R_u(t)$ is maximized at t where Eq. (31) holds, that is, when the following equation does not rotate on the complex plane according to τ.

$$\hat{w_{u_{\rm os}}}(\tau)\cdot {w_u}^*(\tau-t) \quad (30)$$

$$\frac{\partial}{\partial\tau}\arg(\hat{w_{u_{\rm os}}}(\tau)\cdot {w_u}^*(\tau-t)) = 0 \quad (31)$$

Solving for Eq. (31) as shown in the followings,

$$0 = 2\pi\left(\frac{B^2}{2^{SF}}\tau-\frac{B}{2}+f_{\rm os}\right)-2\pi\left(\frac{B^2}{2^{SF}}(\tau-t)-\frac{B}{2}\right) \quad (32)$$

$$= 2\pi\left(\frac{B^2}{2^{SF}}t+f_{\rm os}\right), \quad (33)$$

Time t is derived as follows:

$$t = -f_{\rm os}\frac{2^{SF}}{B^2}. \quad (34)$$

As a result, it ensures that when there is an offset of $f_{\rm os}$ [Hz] in the up-chirp, the correlation peaks appear at a time $f_{\rm os}\cdot 2^{SF}/B^2$ [s] earlier. In addition, the correlation function for the down chirp is given by

$$R_d(t) = \int_{-\infty}^{\infty} \hat{w_{d_{\rm os}}}(\tau)\cdot {w_d}^*(\tau-t)\mathrm{d}\tau. \quad (35)$$

Time t for which $R_d(t)$ is maximum, can be solved in the same manner, as shown in

$$t = f_{\rm os}\frac{2^{SF}}{B^2}. \quad (36)$$

From this result, when there is an offset of $f_{\rm os}$ [Hz] in the down-chirp, it is evident that the correlation peaks appear at a time $f_{\rm os}\cdot 2^{SF}/B^2$ [s] later.

Thus, the receiver can detect the offset frequency of the transmitted LoRaWAN signal in packets in wherein the up- and down-chirps are transmitted consecutively. That is, when the time difference between the respective correlation peaks of the consecutive up and down chirps is Δt, the offset frequency can be expressed by the following equation:

$$f = \frac{1}{2}\left(\Delta t-\frac{2^{SF}}{B}\right)\frac{B^2}{2^{SF}}. \quad (37)$$

The baseband signal of the transmitted signal when there is an offset of $f_{\rm os}$ [Hz] can be written as follows:

$$w_{\rm os}(t) = w(t)\cdot\exp(j2\pi f_{\rm os}t). \quad (38)$$

Therefore, the received signal with no offset can be obtained by multiplying $\exp(-j2\pi f_{\rm os}t)$ on the receiving side as follows:

$$\hat{w}(t) = w_{\rm os}(t)\cdot\exp(j\theta)\cdot\exp(-j2\pi f_{\rm os}t) \quad (39)$$

$$= w(t)\cdot\exp(j\theta). \quad (40)$$

This is equal to Eq. (13). In other words, to cancel the offset detected by the receiver, the received signal with the offset should be multiplied by $\exp(-j2\pi f_{\rm os}t)$.

IV. Simulation Result

This study confirmed that the offset frequency could be detected by the time difference between the peaks of the correlated up-chirp and down-chirp waveforms of the LoRaWAN preamble using simulation results. Table I lists the parameters used in the simulations.

In this simulation, the baseband signal that is the preamble of the LoRaWAN signal is constructed by consecutive chirp

TABLE I
SIMULATION PARAMETERS

Generation frequency of transmit waveforms	500 [ksamples/s]
Operation frequency of receive filter	500 [ksamples/s]
Down sampling frequency	125 [ksamples/s] ($= B$)
Receive filter	Raised-Cosine (FIR Filter: 27 taps)
Spreading factor (SF)	7
Number of FFT bins	128 [points]
Frequency offsets	0 [Hz] or 29297 [Hz]

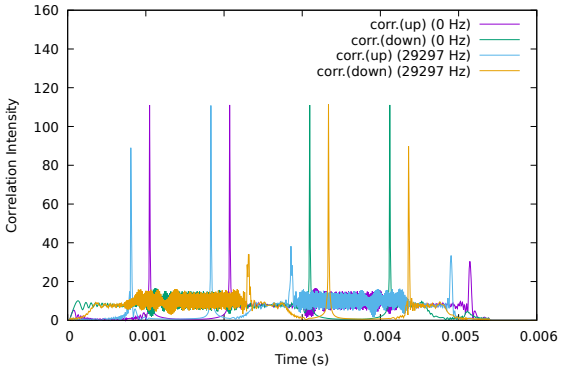


Fig. 8. Correlation peaks of basic-chirp and shifted-chirp in the time domain of the proposed method.

signals that consist of two up-chirps, two down-chirps. An up-chirp of 0.25 was used, as referenced in [**6**].

The offset frequencies f_{os} were set to $0\,[\mathrm{Hz}]$ and $29297\,[\mathrm{Hz}]$. $f_{\mathrm{os}} = 0\,[\mathrm{Hz}]$ indicates that the signal is transmitted with the original LoRa modulation, and $f_{\mathrm{os}} = 29297\,[\mathrm{Hz}]$ indicates that the signal is shifted by the shifted basic chirp to the PLIM and then transmitted. $f_{\mathrm{os}} = 29297\,[\mathrm{Hz}]$ is derived from

$$f_{\mathrm{os}} = (D - 8 + 0.5) \cdot \left(4\frac{125\mathrm{k}}{128}\right) = 29297\,[\mathrm{Hz}]. \qquad (41)$$

Here, D was set to 15. $D = 15$ implies that when the number of bits conveyed with the proposed PLIM is $4\,[\mathrm{bits}]$, the frequency offset has the maximum value. In this case, the width of FFT bins is $4 \cdot 125\mathrm{k}/128\,[\mathrm{Hz}]$, which is around $3.9\,[\mathrm{kHz}]$. When the channel frequency is $920\,[\mathrm{MHz}]$ and the oscillator accuracy of the transceivers is $\pm 1\,[\mathrm{ppm}]$, the range of the clock drift of the oscillator is twice the oscillator accuracy; therefore, we must consider that $\pm 2\,[\mathrm{ppm}]$ occurrs between the transmitter and receiver. Then, the width of the FFT bins was set to approximately $3.9\,[\mathrm{kHz}]$, which is approximately equal to $4\,[\mathrm{ppm}]$ at $920\,[\mathrm{MHz}]$.

Fig. 8 shows the simulation results used to confirm the correlation peaks of each chirp. Based on Fig. 8, the simulation results indicate that the interval between the peaks when the offset frequency is $0\,[\mathrm{Hz}]$ is $1024.0\,[\mathrm{ms}]$ and the interval between the peaks when the offset frequency is $29297\,[\mathrm{Hz}]$ is $1504.0\,[\mathrm{ms}]$. These results indicate that frequency offsets can be detected when they are $0.0000\,[\mathrm{Hz}]$ and $29297\,[\mathrm{Hz}]$.

V. CONCLUSION

This study focused on LPWA, which is used in IoT systems. The LoRaWAN is a communication method for IoT systems. Because LoRaWAN uses LoRa modulation, it is difficult to increase the data rate. Therefore, we addressed the limitations in the number of channels owing to the restricted number of transmission bits by using the proposed PLIM scheme. To increase the number of transmission bits as effectively as possible, we propose a PLIM that leverages frequency offsets within the LoRaWAN channel. This paper presented simulation results to confirm that the channel offsets assigned to PLIM symbols can be detected using our detection algorithm.

ACKNOWLEDGMENT

This research and development work was supported by the MIC/SCOPE #JP235004002.

REFERENCES

[1] LoRa Alliance Technical Committee, "LoRaWAN L2 1.0.4 Specification (TS001–1.0.4), 2020. [Online]. Available: https://resources.lora-alliance.org/technical-specifications/ts001-1-0-4-lorawan-l2-1-0-4-specification.

[2] LoRa Alliance Technical Committee, "RP002-1.0.1 LoRaWAN Regional Parameters, 2020. [Online]. Available: https://resources.lora-alliance.org/technical-specifications/rp2-1-0-1-lorawan-regional-parameters.

[3] Koichi Adachi, Kohei Tsurumi, Aoto Kaburaki, Osamu Takyu, Mai Ohta, and Takeo Fujii, "Packet-Level Index Modulation for LoRaWAN, IEEE Access, vol.9, no.1, pp.12601–12610, Jan. 2021, DOI: 10.1109/ACCESS.2021.3051933.

[4] Kohei Tsurumi, Aoto Kaburaki, Koichi Adachi, Osamu Takyu, Mai Ohta, and Takeo Fujii, "Simple Clock Drift Estimation & Compensation for Packet-Level Index Modulation and its Implementation in LoRaWAN, IEEE IoTJ, vol. 9, no. 16, pp. 15089–15099, Aug. 2022.

[5] H. Yamasaki, H. Matsuura, M. Ohta and M. Taromaru, "Impact of Inexpensive Oscillator in Wireless Sensor Node on Packet-Level Index Modulation Method under Real Environment, The 15th International Conference on Ubiquitous and Future Networks (ICUFN 2024), July 2024.

[6] X. Tang, Y. Zhang, Y. Wang, D. Zhao, H. Li and X. Zhao, "Performance Analysis of Preamble Detection of LoRa System, 2019 International Conference on Electronic Engineering and Informatics (EEI), pp. 175-180, 2019, doi: 10.1109/EEI48997.2019.00046.

ET-SSM: Linear-Time Encrypted Traffic Classification Method Based On Structured State Space Model

Yanjun Li*, Xiangyu Zhao*, Zhengpeng Zha*†, Zhenhua Ling‡
* Institute of Advanced Technology, University of Science and Technology of China
Email: {ballinli, xyzhao23}@mail.ustc.edu.cn,
† Corresponding author. Email: zhazp@ustc.edu.cn,
‡ School of Information Science and Technology, University of Science and Technology of China
E-mail: zhling@ustc.edu.cn

***Abstract*—Encrypted traffic classification involves categorizing data traffic that has been encrypted for privacy and security. In this domain, Transformers have been essential for successful pretraining. However, their quadratic time complexity makes them costly and slow for inference. While many works tried to reduce memory overhead and speed up inference, they often fail to match the classification performance of existing methods. To tackle this challenge, we present an Encrypted Traffic State Space Model (ET-SSM), utilizing recent advances in sequence routing based on state space models. The ET-SSM initially integrates SSM layers with a multiplicative gated architecture, which has proven effective in simplified sequence modeling frameworks. This structure learns static linear layers that do not account for pairwise traffic interactions. To overcome this limitation, we employ convolution layers to capture adjacent traffic context interactions while maintaining overall linear time complexity. Extensive experiments demonstrate that ET-SSM not only significantly reduces memory overhead and accelerates inference time but also achieves state-of-the-art classification results. This makes it a promising candidate for the pretraining model in encrypted traffic classification for realistic production environments. The source code will be released upon paper acceptance.**

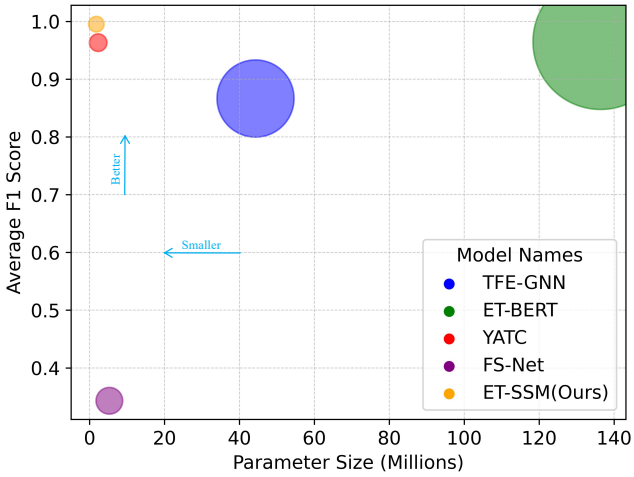


Fig. 1. **The average F1 Score *vs*. parameter size (millions) comparison.** Our ET-SSM achieves the state-of-the-art average F1-score on six encrypted traffic classification datasets while having the smallest parameter size (i.e. The size of the bubbles is depicted in the figure.), compared with cutting-edge encrypted traffic classification model, *e.g.* TFE-GNN [3], ET-BERT [4], YAYC [5], FS-Net [6].

I. Introduction

Recently, the widespread use of traffic encryption has become instrumental in protecting the privacy and anonymity of Internet users [1], [2]. While this advancement is vital for security and confidentiality, it concurrently presents significant challenges to traffic classification. The increasing utilization of privacy-enhancing encryption techniques, such as Tor and VPNs, by both legitimate users and malicious actors, complicates the task of distinguishing benign from harmful traffic. Encrypted traffic classification thus emerges as a crucial tool in this landscape. It enables the identification and mitigation of malware and cybercriminal activities that exploit encryption to bypass surveillance systems, without compromising the privacy and integrity of legitimate communications. This delicate balance between user privacy and cybersecurity underscores the indispensable role of sophisticated encrypted traffic classification methodologies in maintaining a secure digital environment.

In this area, Transformers [7] have become the *de facto* architecture for pretraining models in encrypted traffic classification. Since the introduction of ET-BERT [4], Transformers have proven central to encrypted traffic classification due to their ability to learn effectively from large unlabeled datasets. Specifically, using attention mechanisms [8] as a core routing component has been critical to achieving empirical success in downstream encrypted traffic classification tasks. Moreover, YaTC introduces a masked autoencoder-based traffic transformer [9] with multi-level flow representation to learn more efficient encrypted traffic representations. However, despite the significant success of these attention-based models in encrypted traffic classification, their high memory cost and slow inference time make them challenging to deploy in realistic production environments [1]. In the field of cybersecurity, where encrypted traffic classification is a crucial downstream task, there is a greater emphasis on achieving high throughput rates at a low cost [10]–[12]. Some efforts address the high memory costs and slow inference throughput of pretrain-

ing models for encrypted traffic classification by employing knowledge distillation methods. For instance, XENTC [13] improved the DistilBERT [14] leverages knowledge distillation to enable the use of a compressed model while maintaining good performance. Subsequently, a streamlined packet structure, consisting of a header and a partial payload, is employed to achieve effective feature reduction. Despite these efforts to create lightweight Transformers for encrypted traffic classification, they cannot achieve comparable classification performance. Given the quadratic computational complexity of Transformer models, which escalates with the sequence length, the Transformers cannot balance the effectiveness and efficiency. **Are Transformers the best solution to effective and efficient encrypted traffic classification?**

State Space Models (SSMs) for deep learning offer a promising alternative for sequence modeling. Recent research indicates that SSMs are a competitive architecture for handling sequences, as demonstrated in [15]. Various adaptations of SSMs have demonstrated success across diverse domains. They effectively handle continuous signals, such as audio [16] and visual data [17], as well as discrete signals, such as textual content [18]. SASHIMI [16] is a new architecture for modeling waveforms that yield state-of-the-art performance on conditional audio generation benchmarks. S4ND [17] shows SSMs modeling large-scale visual data across 1D, 2D, and 3D formats as continuous multidimensional signals. Mamba [18] is a promising language model that not only outperforms Transformers of equivalent size but also matches the performance of Transformers twice its size. In addition to delivering strong performance, SSM-based routing avoids the quadratic complexity associated with increasing sequence lengths. Specifically, the model facilitates the capture of recurrent neural network (RNN) long-range dependencies while maintaining convolutional neural networks (CNN) training speeds. Compared to the sparse information density typical of visual data and the denser information density found in speech, text-based SSMs [18] more closely resemble encrypted traffic. However, encrypted traffic classification requires lower computational costs and higher inference speeds compared to the text domain to ensure privacy and anonymity.

The encrypted traffic classification field urgently requires a model that combines low cost, rapid processing speeds, and high classification performance. In this work, we propose an architecture, ET-SSM for applying SSMs for encrypted traffic classification. Specifically, to enhance the model's representative capacity, we have developed a multiplicative gating architecture [19], [20]. This approach simplifies the routing mechanism while maintaining its effectiveness in modeling necessary interactions. Moreover, to better model the adjacent encrypted traffic context without increasing computational costs, we have incorporated CNN blocks into the ET-SSM framework. As observed in Fig. 1, ET-SSM achieves the state-of-the-art average F1-score on six encrypted traffic classification datasets while having the smallest parameter size. Furthermore, compared to existing methods, ET-SSM improves inference speed while maintaining low GPU memory usage. Our contributions are outlined as follows:

- We propose ET-SSM, the first SSM-based method for encrypted traffic classification. Compared to existing classification methods, our approach achieves the lowest GPU memory cost and the fastest inference speed.
- The ET-SSM achieves average state-of-the-art F1 scores across six encrypted traffic classification datasets, consistently outperforming other models on all these datasets.
- Extensive experiments were conducted across a range of encrypted traffic classification tasks to evaluate our methods, including detailed evaluations such as efficiency analyses, and ablation studies.

II. Methods

A. ET-SSM Model Pre-training

Recent work has demonstrated that gating can enhance the performance of models with simplified routing. For instance, linear time attention models have been shown to benefit from improved gating mechanisms [21]. Inspired by the work of [21], we incorporate gated units as the core structural components of our model. The complete ET-SSM model is shown in Figure 2. In the initial phase of our study, we utilize the BURST structure, identified as a sequence of temporally contiguous network packets emanating from either a request or a response in a single session flow. This structure is employed to precisely depict encrypted traffic, thereby forming the input for our pre-trained ET-BERT model, mirroring the strategy delineated in [4]. Subsequently, we combine token embedding, position embedding, and segment embedding to obtain the final encrypted traffic token representation, which serves as the input to the ET-SSM encoder. The final MLP classification head processes the output from the ET-SSM encoder to determine the encrypted traffic categories.

The architecture of ET-SSM is a bidirectional adaptation of the gated unit of [21]. Specifically, let $\mathbf{X}_{i-1} \in \mathbb{R}^{L\times d}$ be activations at the $i-1$-th layer where the length is L, and the model size is d. We use the activation GELU [22] for σ. The ET-SSM layer begins as follows,

$$\begin{aligned}
\mathbf{X}'_{i-1} &= \text{LayerNorm}(\mathbf{X}_{i-1}) && \in \mathbb{R}^{L\times d},\\
\mathbf{Z} &= \sigma(CONV(\mathbf{X_{i-1}})) && \in \mathbb{R}^{L\times 3d},\\
\mathbf{F} &= \sigma(CONV(\mathbf{X}'_{i-1})) && \in \mathbb{R}^{L\times d},\\
\mathbf{B} &= \sigma(CONV(\text{Flip}(\mathbf{X}'_{i-1})) && \in \mathbb{R}^{L\times d}.
\end{aligned} \tag{1}$$

To emulate the bidirectional information extraction capability of BERT [8], we employ two sequential blocks, namely, a forward and a backward SSM layer, as presented,

$$\begin{aligned}
\mathbf{U}_1 &= CONV(\text{SSM}(\mathbf{F})) && \in \mathbb{R}^{L\times d},\\
\mathbf{U}_2 &= CONV(\text{SSM}(\mathbf{B})) && \in \mathbb{R}^{L\times d},\\
\mathbf{U} &= \sigma(CONV((\mathbf{U}_1 \otimes \text{Flip}(\mathbf{U}_2)))) && \in \mathbb{R}^{L\times 3d}.
\end{aligned} \tag{2}$$

Finally, we utilize a residual connection [23] to avoid the gradient vanishing problem. Specifically, we sum the output

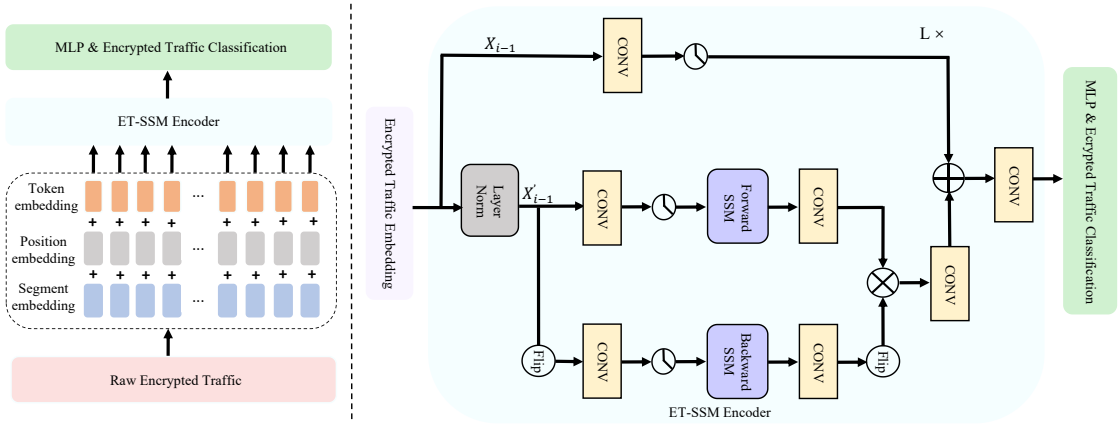


Fig. 2. **The overview of the proposed ET-SSM model.** Initially, we segment the raw encrypted traffic into tokens and incorporate both position and segment embeddings to generate the encrypted traffic embedding, following the approach described by Lin et al. [4]. This embedding is then inputted into our proposed ET-SSM encoder, which processes the sequence bidirectionally. To enhance processing speed and more effectively capture adjacent information, we employ CONV blocks within the encoder. For inference, the MLP classification heads operate following the methodology outlined by Devlin et al. [8], predicting encrypted traffic classifications.

$\mathbf{U}$ with the original input $\mathbf{V}$, as the input $\mathbf{X}_{i+1}$ of the next layer $i+1$.

$$\mathbf{O} = CONV(\mathbf{U} + \mathbf{Z}) \quad \in \mathbb{R}^{L \times d}. \tag{3}$$

However, the multiplicative gated architecture has a significant limitation. It employs static linear layers that are constrained in their ability to account for pairwise traffic interactions. To address this limitation, we incorporate CONV layers to capture adjacent traffic context interactions. By doing so, we enhance the model's ability to understand the relationships between neighboring traffic data points. Importantly, this approach maintains overall linear time complexity, ensuring that the model remains efficient and scalable despite the added complexity of capturing these interactions.

For the feature extraction modules, we use convolutional (CONV) layers instead of fully connected (FC) layers as used in [21] for two reasons. Firstly, the CONV layer is better optimized by compilers than the FC layer [24]. Additionally, With the same number of parameters, a model with CONV 1×1 runs approximately 27% faster than one with FC layers. Finally, suppose FC layers are used as the feature extraction modules. In such cases, additional Permute operations are required to adjust the output feature dimensions to be compatible with the CONV layers, since FC and CONV layers handle different tensor dimensions. However, these Reshape operations may not be hardware-friendly on certain mobile devices [25], [26].

B. ET-SSM Model Fine-tuning

All encoder parameters, including embedding modules and SSM blocks, are loaded from pretraining for downstream tasks. To classify labeled traffic data, the decoder is replaced with an MLP head. Given that all tokens are visible, the fine-tuning of ET-SSM is performed in a supervised manner, as detailed below,

$$\begin{aligned} \mathbf{X} &= \text{Encoder}(\mathbf{X}_0) && \in \mathbb{R}^{L \times D_{\text{enc}}}, \\ \mathbf{f} &= \mathbf{X}[L, :] && \in \mathbb{R}^{D_{\text{enc}}}, \\ \hat{\mathbf{y}} &= \text{MLP}(\text{Norm}(\mathbf{f})). \end{aligned} \tag{4}$$

Here, $\mathbf{f}$ denotes the trailing class token, and $\hat{\mathbf{y}} \in \mathbb{R}^C$ represents the prediction distribution, where C is the number of traffic categories. The classification process is then optimized by minimizing the cross-entropy loss between the prediction distribution $\hat{\mathbf{y}}$ and the ground-truth label $\mathbf{y}$ as below,

$$L_{\text{cls}} = \text{CrossEntropy}(\hat{\mathbf{y}}, \mathbf{y}). \tag{5}$$

III. Experiments

A. Dataset And Metrics

To validate the efficacy and broad applicability of ET-SSM, we conducted a series of experiments on encrypted traffic classification tasks, utilizing six publicly accessible datasets. Due to page limitations, the details of the datasets are provided in Appendix C of the supplementary materials. We evaluate and compare the performance of our model using four typical metrics: Accuracy (AC), Precision (PR), Recall (RC), and F1 Score (F1), as described in [27], [28].

B. Implementation Details

In our approach, we implemented two distinct learning strategies for the ET-SSM model to adapt to different levels of traffic data granularity, the ET-SSM (flow) and the ET-SSM (packet). Due to page limitations, the details of the implementation details are provided in Appendix C of the supplementary materials.

C. Comparison With Existing Methods

Tables I and II showcase the performance comparison of our framework with existing frameworks. The best results are highlighted with an orange line, and the model with the smallest parameters is shown in bold. Our framework sets a

TABLE I

PERFORMANCE COMPARISON OF DIFFERENT METHODS ON CROSS-PLATFORM(IOS), CROSS-PLATFORM(ANDROID), ICSX-VPN-SERVICE DATASETS.THE BEST RESULTS ARE HIGHLIGHTED WITH AN ORANGE LINE, AND THE MODEL WITH THE SMALLEST PARAMETERS IS SHOWN IN **BOLD**.

Dataset	**Params(M)**	**Cross-Platform(iOS)**				**Cross-Platform(Android)**				**ISCX-VPN-Service**			
Method		AC	PR	RC	F1	AC	PR	RC	F1	AC	PR	RC	F1
AppScanner [29]	-	0.3205	0.2103	0.2173	0.2030	0.3868	0.2523	0.2594	0.2440	0.7182	0.7339	0.7225	0.7197
CUMUL [30]	-	0.2910	0.1917	0.2081	0.1875	0.3525	0.2221	0.2409	0.2189	0.5610	0.5883	0.5676	0.5668
BIND [31]	-	0.3770	0.2566	0.2715	0.2484	0.4728	0.3126	0.3253	0.3026	0.7534	0.7583	0.7488	0.7420
K-fp [32]	-	0.2155	0.2037	0.2069	0.2003	0.2248	0.2113	0.2104	0.2052	0.6430	0.6492	0.6417	0.6395
FlowPrint [27]	-	0.9254	0.9438	0.9254	0.9260	0.8698	0.9007	0.8698	0.8702	0.7962	0.8042	0.7812	0.7820
DF [33]	-	0.3106	0.2232	0.2179	0.2140	0.3862	0.2595	0.2620	0.2527	0.7154	0.7192	0.7104	0.7102
FS-Net [34]	5.3	0.3712	0.2845	0.2754	0.2655	0.4846	0.3544	0.3365	0.3343	0.7205	0.7502	0.7238	0.7131
GraphDApp [35]	-	0.3245	0.2450	0.2392	0.2297	0.4031	0.2842	0.2786	0.2703	0.5977	0.6045	0.6220	0.6036
TSCRNN [36]	-	-	-	-	-	-	-	-	-	-	-	-	-
Deeppacket [37]	-	0.9204	0.8963	0.8872	0.9034	0.8805	0.8004	0.7567	0.8138	0.9329	0.9377	0.9306	0.9321
PERT [38]	-	0.9789	0.9621	0.9611	0.9584	0.9772	0.8628	0.8591	0.8550	0.9352	0.9400	0.9349	0.9368
ET-BERT(flow) [4]	187.4	0.9844	0.9701	0.9632	0.9643	0.9865	0.9324	0.9266	0.9246	0.9729	0.9756	0.9731	0.9733
ET-BERT(packet) [4]	187.4	0.9810	0.9757	0.9772	0.9754	0.9728	0.9439	0.9119	0.9206	0.9890	0.9891	0.9890	0.9890
YaTC[39]	2.3	0.9310	0.9307	0.9310	0.9295	0.9042	0.9081	0.9042	0.9042	-	-	-	-
TFE-GNN	-	0.8241	0.8326	0.8241	0.8130	0.8141	0.8308	0.8141	0.8067	-	-	-	-
Ours(flow)	**1.7**	0.9887	0.9932	0.9866	0.9868	0.9894	0.9571	0.9587	0.9579	0.9814	0.9874	0.9877	0.9875
Ours(packet)	**1.7**	0.9954	0.9997	0.9951	0.9974	0.9977	0.9829	0.9891	0.9860	0.9987	0.9934	0.9947	0.9940

TABLE II

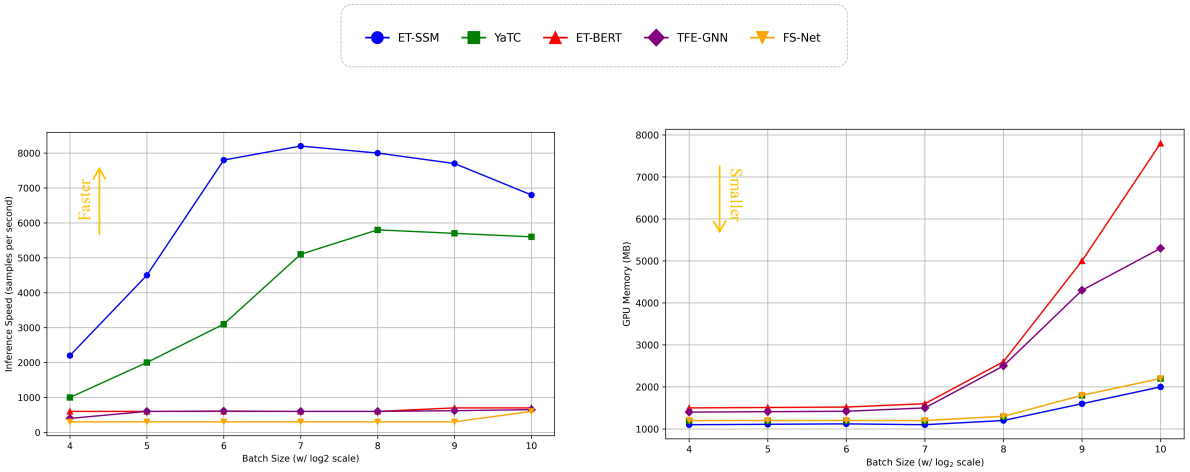
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON ISCX-VPN-APP, ISCX-TOR, USTC-TFC DATASETS.THE BEST RESULTS ARE HIGHLIGHTED WITH AN ORANGE LINE, AND THE MODEL WITH THE SMALLEST PARAMETERS IS SHOWN IN **BOLD**.

Dataset	**Params(M)**	**ISCX-VPN-App**				**ISCX-Tor**				**USTC-TFC**			
Method		AC	PR	RC	F1	AC	PR	RC	F1	AC	PR	RC	F1
AppScanner[29]	-	0.6266	0.4864	0.5198	0.4935	0.6722	0.3756	0.4422	0.3913	0.8954	0.8984	0.8968	0.8892
CUMUL [30]	-	0.5365	0.4129	0.4535	0.4236	0.6606	0.3850	0.4416	0.3918	0.5675	0.6171	0.5738	0.5513
BIND [31]	-	0.6767	0.5152	0.5153	0.4965	0.7185	0.4598	0.4515	0.4511	0.8457	0.8681	0.8382	0.8396
K-fp [32]	-	0.6070	0.5478	0.5430	0.5303	0.6472	0.5576	0.5849	0.5522	-	-	-	-
FlowPrint [27]	-	0.8767	0.6697	0.6651	0.6531	0.9092	0.3820	0.3661	0.3654	0.8146	0.6434	0.7002	0.6573
DF [33]	-	0.6116	0.5706	0.4752	0.4799	0.7533	0.6228	0.6010	0.5850	0.7787	0.7883	0.7819	0.7593
FS-Net [34]	5.3	0.6647	0.4819	0.4848	0.4737	0.6071	0.5080	0.5350	0.4590	0.8846	0.8846	0.8920	0.8840
GraphDApp [35]	-	0.6328	0.5900	0.5472	0.5558	0.6836	0.4864	0.4823	0.4488	0.8789	0.8226	0.8260	0.8234
TSCRNN [36]	-	-	0.9490	0.9480	0.9480	-	0.9870	0.9860	0.9870	-	-	-	-
Deeppacket [37]	-	0.9758	0.9785	0.9745	0.9765	0.7449	0.7549	0.7399	0.7473	0.9640	0.9650	0.9631	0.9641
PERT [38]	-	0.8229	0.7092	0.7173	0.6992	0.7682	0.4424	0.4446	0.4345	0.9909	0.9911	0.9910	0.9911
ET-BERT(flow) [4]	187.4	0.8519	0.7508	0.7294	0.7306	0.8311	0.5564	0.6448	0.5886	0.9929	0.9930	0.9930	0.9930
ET-BERT(packet) [4]	187.4	0.9962	0.9936	0.9938	0.9937	0.9921	0.9923	0.9921	0.9921	0.9915	0.9915	0.9916	0.9916
TFE-GNN	-	-	-	-	-	0.7692	0.8030	0.7692	0.7618	0.9747	0.9747	0.9747	0.9734
YaTC [39]	2.3	0.9819	0.9820	0.9819	0.9819	0.9959	0.9959	0.9959	0.9959	0.9947	0.9749	0.9747	0.9734
Ours(flow)	**1.7**	0.8635	0.7874	0.7597	0.7733	0.8616	0.5876	0.6787	0.6299	0.9934	0.9947	0.9945	0.9946
Ours(packet)	**1.7**	0.9944	0.9978	0.9979	0.9978	0.9969	0.9974	0.9975	0.9976	0.9958	0.9975	0.9976	0.9975

new benchmark in state-of-the-art performance, outperforming the preceding leading method in the F1-score across all six datasets. The margin of enhancement ranges from **+0.16%** to **+6.54%**, as substantiated by the results enumerated in both tables. Moreover, it is imperative to highlight that the results of our experiments surpass all previous state-of-the-art methods in terms of F1-score in the flow-level and packet-level fine-tuning. With state-of-the-art performance in encrypted traffic classification, our ET-SSM model also achieves the smallest parameter count. These outstanding classification results demonstrate that our ET-SSM structure excels in incorporating inductive bias and computational efficiency. For ease of comparison, subsequent experiments and analyses designate ET-SSM(packet) and ET-BERT(packet) as ET-SSM and ET-BERT, respectively.

D. Efficiency Evaluation

Figure 3 presents the experimental analysis of inference speed and GPU memory utilization for the ET-SSM on the largest encrypted traffic classification dataset, Cross-Platform(iOS). As depicted in Figure 3(a), ET-SSM achieved the fastest inference speed among all evaluated methods across various input batch sizes. Specifically, it enhanced inference speed by a factor of 2.2 to 2.5 compared with the previous fastest method YaTC. This significant speed advantage was especially notable given the substantial model parameters and less efficient architectural designs observed in models like ET-BERT, TFE-GNN, and FS-Net. This superior performance highlighted ET-SSM's lower computational complexity than other encrypted traffic classification models. In Figure 3(b), ET-SSM exhibited the lowest GPU memory consumption compared to previous encrypted traffic classification models, even using large batch sizes. Compared to the relatively low-



(a) Inference Speed Comparison

(b) GPU Memory Comparison

Fig. 3. The Inference Speed and GPU Memory Comparison of ET-SSM, YaTC, ET-BERT, TFE-GNN and FS-Net on Cross-Platform(iOS)

cost GPU memory methods FS-Net and YaTC, FS-Net was hindered by significantly slower inference speeds and subpar classification performance. And YaTC was still based on attention structure, hindering its inference speed and GPU memory. Our ET-SSM distinguished itself by significantly enhancing memory efficiency compared to other encrypted traffic classification models.

E. Ablation Studies

Due to page limitations, the details of the ablation studies are provided in Appendix C of the supplementary materials.

IV. Conclusion

In this work, we introduce ET-SSM, a novel framework for encrypted traffic classification designed to elevate both efficiency and effectiveness. This framework integrates SSM layers with a multiplicative gated architecture, which has demonstrated efficacy in simplified sequence modeling contexts. The core architecture of ET-SSM comprises static linear layers that initially do not accommodate pairwise traffic interactions. To address this limitation, we incorporated CNN to capture adjacent traffic context interactions while preserving the overall linear time complexity. Extensive experimental results validate the efficiency and effectiveness of ET-SSM in handling encrypted traffic classification.

Supplementary Materials

V. Appendix A: Related Work

A. State Space Models

State space models (SSMs) [40], [41] have emerged as a promising architecture for sequence modeling. These models synthesize the characteristics of RNNs and CNNs [42], drawing inspiration from classical state space models. SSMs can be implemented efficiently either as a recurrence or convolution, offering linear or near-linear scaling to sequence length. Various adaptations of SSMs have demonstrated success in diverse domains, handling both continuous signals, such as audio [16] and visual [17] data, and discrete signals, such as textual content [18]. SASHIMI [16] is based upon SSMs, resulting in a state-of-the-art performance for both autoregressive and non-autoregressive audio generation. S4ND [17] is an innovative multidimensional SSM layer that extends SSM capabilities to multidimensional data, including images and videos. S4ND excels at modeling large-scale visual data across 1D, 2D, and 3D dimensions as continuous signals. Mamba [18] integrates selective structured state space models into a streamlined, attention-free end-to-end neural network architecture, achieving significant success in the text domain. This model offers rapid inference and linear scaling with sequence lengths, showing enhanced performance on real datasets involving sequences of up to a million elements. These innovative SSMs research efforts inspire us to develop a model based on SSMs that features reduced hyperparameter complexity and accelerated inference speeds for encrypted traffic classification.

B. Encrypted Traffic Classification

Encrypted traffic classification aims to discern the services operating behind obfuscated network traffic, enhancing network service quality and security assurance. Given the remarkable success of the BERT [8] model within the natural language processing community, researchers [4] are exploring the application of its structural principles in the realm of encrypted traffic classification through pre-training approaches. An enhanced traffic classification framework [5] leverages a formatted traffic representation matrix, an efficient Traffic Transformer with hierarchical attention mechanisms, and an MAE-based [9] pre-training paradigm to achieve superior performance with limited labeled data. Although this attention-based [7] pre-training model achieves great classification performance, researchers found that the pre-training model suffers from long processing time and large memory consumption. FastTraffic [43] introduces a novel N-gram feature embedding approach to encapsulate network packet structure and sequential characteristics. Additionally, they design a three-layer MLP that achieves rapid and efficient classification. Furthermore, XENTC [13] enhances the DistilBERT [14] architecture through knowledge distillation, enabling the use of a compressed model while maintaining good performance. Subsequently, they employed a streamlined packet structure, consisting of a header and a partial payload, to achieve effective feature reduction. Despite these two methodologies being less computationally demanding than the attention-based counterparts for encrypted traffic classification, they do not attain comparable levels of performance. What's more, the inherent non-linearity of both the MLP and BERT models, restricts their efficacy in lightweight settings.

Unlike the sparse information density characteristic of visual data and the denser information typical of speech, text-based SSMs [18] more closely resemble encrypted traffic. Nevertheless, encrypted traffic classification demands even lower computational costs and higher inference speeds than those required in the text domain, crucial for ensuring privacy and anonymity. In response, we introduce ET-SSM, a novel framework founded on the principles of SSMs. This approach not only matches state-of-the-art classification performance when compared with existing methods but also reduces GPU memory usage and enhances inference speed. The complete framework of ET-SSM is depicted in Fig. 2.

VI. Appendix B: Methods

A. Preliminaries

A state space model (SSM) is a general-purpose tool for describing the relationship between a continuous-time scalar input $u(t)$ and a scalar output $y(t)$ using the following differential equations,

$$\begin{aligned} x'(t) &= \boldsymbol{A}x(t) + \boldsymbol{B}u(t), \\ y(t) &= \boldsymbol{C}x(t) + \boldsymbol{D}u(t), \end{aligned} \tag{6}$$

where $x(t) \in \mathbb{R}^N$ is a continuous-time state vector, $x'(t)$ is its derivative, and the equation is parameterized by $\boldsymbol{A} \in \mathbb{R}^{N\times N}, \boldsymbol{B} \in \mathbb{R}^{N\times 1}, \boldsymbol{C} \in \mathbb{R}^{1\times N}, \boldsymbol{D} \in \mathbb{R}^{1\times 1}$.

When applied to a discrete-time scalar input sequence $u_1, \dots u_L$, the SSM equations, and parameters can be discretized, leading to the following recursion,

$$x_k = \overline{\boldsymbol{A}}x_{k-1} + \overline{\boldsymbol{B}}u_k, \quad y_k = \overline{\boldsymbol{C}}x_k + \overline{\boldsymbol{D}}u_k, \tag{7}$$

where $\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \overline{\boldsymbol{C}}, \overline{\boldsymbol{D}}$ are functions of the original parameters and a discretization rate.

This equation can be computed similarly to an RNN, where $x_k \in \mathbb{R}^N$ is a hidden state at time step k. However, unlike an RNN, the linearity of the recursion allows $y_1 \dots y_L$ to be computed directly using a convolution with precomputed kernel $\overline{\boldsymbol{K}} \in \mathbb{R}^L$,

$$\begin{aligned} \overline{\boldsymbol{K}} &= (\overline{\boldsymbol{C}\boldsymbol{B}}, \overline{\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}}, \dots, \overline{\boldsymbol{C}\boldsymbol{A}}^{L-1}\overline{\boldsymbol{B}}), \\ y &= \overline{\boldsymbol{K}} * u. \end{aligned} \tag{8}$$

An effective approach for integrating SSMs into neural networks has been demonstrated by the works of [15], [41]. Their core insight is the parameterization of the transition matrix $\boldsymbol{A}$, known as HiPPO,

$$\boldsymbol{A}_{nk} = -\begin{cases} (2n+1)^{1/2}(2k+1)^{1/2} & \text{if } n > k, \\ n+1 & \text{if } n = k, \\ 0 & \text{if } n < k. \end{cases} \tag{9}$$

This matrix yields a stable and efficient training regime. The full model, S4, retains the SSM's ability to model long-term sequences while being more training-efficient than RNNs.

Recently, researchers [44], [45] have proposed simplified diagonalized versions of S4, which achieve comparable results with a more straightforward approximation of the original parameterization. In our preliminary experiments, we tested several different S4 parameterizations and found no significant difference in accuracy. Therefore, we use S4D as the parameterization throughout this work.

B. The Forward Pass Algorithm Description

The operational process of the ET-SSM block forward pass is detailed in Algorithm 1. For a given input token sequence $\mathbf{X}_{i-1}$ with a batch size B and sequence length L from the $(i-1)$-th ET-SSM block, we begin by normalizing it and then projecting it linearly into $\mathbf{x}$ and $\mathbf{z}$, both with a dimension size of E. We subsequently apply causal 1-D convolution to $\mathbf{x}$, resulting in $\mathbf{x}'$. Based on $\mathbf{x}'$, we compute the input-dependent step size Δ, as well as the projection parameters $\mathbf{B}$ and $\mathbf{C}$ having a dimension size of N. We then discretize $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ using Δ. Following this, we calculate $\mathbf{y}$ employing a hardware-aware State Space Model. Finally, $\mathbf{y}$ is gated by $\mathbf{z}$ and added residually to $\mathbf{X}_{i-1}$, resulting in the output token sequence $\mathbf{X}_i$ for the i-th ET-SSM block.

Algorithm 1 ET-SSM Block Process

Require: token sequence $\mathbf{X}_{i-1} : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
Ensure: token sequence $\mathbf{X}_i : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
/* normalize the input sequence $\mathbf{X}'_{i-1}$ */
$\mathbf{X}'_{i-1} : (\mathtt{B}, \mathtt{L}, \mathtt{D}) \leftarrow \mathbf{Norm}(\mathbf{X}_{i-1})$
$\mathbf{x} : (\mathtt{B}, \mathtt{L}, \mathtt{E}) \leftarrow \mathbf{Conv1d}^{\mathbf{x}}(\mathbf{X}'_{i-1})$
$\mathbf{z} : (\mathtt{B}, \mathtt{L}, \mathtt{E}) \leftarrow \mathbf{SiLU}(\mathbf{Conv1d}^{\mathbf{z}}(\mathbf{X}'_{i-1}))$
/* process with different direction */
for o in {forward, backward} **do**
  $\mathbf{x}'_o : (\mathtt{B}, \mathtt{L}, \mathtt{E}) \leftarrow \mathbf{SiLU}(\mathbf{Conv1d}_o(\mathbf{x}))$
  $\mathbf{B}_o : (\mathtt{B}, \mathtt{L}, \mathtt{N}) \leftarrow \mathbf{Linear}_o^{\mathbf{B}}(\mathbf{x}'_o)$
  $\mathbf{C}_o : (\mathtt{B}, \mathtt{L}, \mathtt{N}) \leftarrow \mathbf{Linear}_o^{\mathbf{C}}(\mathbf{x}'_o)$
  /* softplus ensures positive $\boldsymbol{\Delta}_o$ */
  $\boldsymbol{\Delta}_o : (\mathtt{B}, \mathtt{L}, \mathtt{E}) \leftarrow \log(1 + \exp(\mathbf{Linear}_o^{\boldsymbol{\Delta}}(\mathbf{x}'_o) + \mathbf{Parameter}_o^{\boldsymbol{\Delta}}))$
  /* shape of $\mathbf{Parameter}_o^{\mathbf{A}}$ is $(\mathtt{E}, \mathtt{N})$ */
  $\overline{\mathbf{A}}_o : (\mathtt{B}, \mathtt{L}, \mathtt{E}, \mathtt{N}) \leftarrow \boldsymbol{\Delta}_o \bigotimes \mathbf{Parameter}_o^{\mathbf{A}}$
  $\overline{\mathbf{B}}_o : (\mathtt{B}, \mathtt{L}, \mathtt{E}, \mathtt{N}) \leftarrow \boldsymbol{\Delta}_o \bigotimes \mathbf{B}_o$
  $\mathbf{y}_o : (\mathtt{B}, \mathtt{L}, \mathtt{E}) \leftarrow \mathbf{SSM}(\overline{\mathbf{A}}_o, \overline{\mathbf{B}}_o, \mathbf{C}_o)(\mathbf{x}'_o)$
end for
/* get gated $\mathbf{y}_o$ */
$\mathbf{y}'_{forward} : (\mathtt{B}, \mathtt{L}, \mathtt{E}) \leftarrow \mathbf{Conv1d}^{\mathbf{forward}}(\mathbf{y}_{forward})$
$\mathbf{y}'_{backward} : (\mathtt{B}, \mathtt{L}, \mathtt{E}) \leftarrow \mathbf{Conv1d}^{\mathbf{backward}}(\mathbf{y}_{backward})$
$\mathbf{y}_{hadamard} : (\mathtt{B}, \mathtt{L}, \mathtt{E}) \leftarrow \mathbf{y}'_{forward} \odot \mathbf{y}'_{backward}$
$\mathbf{y} : (\mathtt{B}, \mathtt{L}, \mathtt{E}) \leftarrow \leftarrow \mathbf{Conv1d}^{\mathbf{y_{hadamard}}}(\mathbf{y}_{hadamard})$
$\mathbf{T}_i : (\mathtt{B}, \mathtt{L}, \mathtt{D}) \leftarrow (\mathbf{z} + \mathbf{y})$
/* residual connection */
Return: $\mathbf{T}_i$

VII. Appendix C: Experiments

A. Dataset And Metrics

To validate the efficacy and broad applicability of ET-SSM, we conducted a series of experiments on encrypted traffic classification tasks, utilizing six publicly accessible datasets. Table III delineates the specifics of these datasets. The General Encrypted Application Classification [27] task categorizes application traffic under standard encryption protocols. Our evaluations were conducted on the Cross-Platform datasets for both iOS and Android, encompassing 196 and 215 applications respectively. The Encrypted Malware Classification [46] task involves the analysis of encrypted traffic comprising both malware and benign applications. In this context, the USTC-TFC dataset is particularly noteworthy, as it features 10 categories each of benign and malicious traffic, providing a comprehensive framework for assessing encryption-based malware detection capabilities. The Encrypted Traffic Classification on VPN [47] task focuses on classifying encrypted traffic that utilizes Virtual Private Networks (VPNs) for network communication. We employ the widely-used ISCX-VPN dataset, which comprises data from six communication applications, captured through the Canadian Institute for Cybersecurity in both VPN and non-VPN settings. The Encrypted Application Classification on Tor [48] task is centered around classifying encrypted traffic using the Onion Router to enhance communication privacy. The relevant dataset, ISCX-Tor, comprises 16 distinct applications, offering a unique landscape for assessing privacy-preserved encrypted traffic analysis.

We evaluate and compare the performance of our model using four typical metrics: Accuracy (AC), Precision (PR), Recall (RC), and F1 Score (F1), as described in [27], [28].

To mitigate the effects of data imbalance across multiple categories, we employ Macro Averaging [49] which calculates the mean value of AC, PR, RC, and F1 for each category.

B. Implementation Details

During the pre-training, approximately 30GB of traffic data was utilized for this procedure. The dataset was divided into two segments: (1) roughly 15GB of traffic data sourced from public datasets [47], [50]; (2) an equivalent volume of traffic data, approximately 15GB, obtained through passive collection within the China Science and Technology Network (CSTNET). Furthermore, we adopted the training data and strategies outlined by Izsak et al. [51]. In line with the approach used by RoBERTa [52], we exclusively employed masked language modeling, preceding next-sentence prediction. The batch size was set at 128, and the total number of steps was 500,000. The learning rate was established at 1×10^{-5}, with a warmup ratio of 0.1. For fine-tuning, we utilized the AdamW optimizer across 10 epochs, applying a learning rate of 6×10^{-5} for flow-level and 2×10^{-5} for packet-level tasks. The batch size remained at 32, and the dropout rate was set to 0.5.

TABLE III
SUMMARY OF DATASETS USED IN ENCRYPTED TRAFFIC CLASSIFICATION EXPERIMENTS

Task	Dataset	Flow	Packet	Label
General Encrypted Application Classification	Cross-Platform (iOS) [27]	20,858	707,717	196
	Cross-Platform (Android) [27]	27,846	656,044	215
Encrypted Malware Classification	USTC-TFC [46]	9,853	97,115	20
Encrypted Traffic Classification on VPN	ISCX-VPN-Service [47]	3,694	60,000	12
	ISCX-VPN-App [47]	2,329	77,163	17
Encrypted Application Classification on Tor	ISCX-Tor [48]	3,021	80,000	16

TABLE IV
COMPARISON OF DIFFERENT METHODS, THEIR CORE STRUCTURES, AND THE COMPLEXITY PER LAYER.

Symbolic method	Core structure	Sequential operations	Complexity per layer
DeepPacket [37]	CNN	$\mathcal{O}(1)$	$\mathcal{O}(k \cdot \mathcal{L} \cdot \mathcal{D}^2)$
TSCRNN [36]	RNN	$\mathcal{O}(\mathcal{L})$	$\mathcal{O}(\mathcal{L} \cdot \mathcal{D}^2)$
FastTraffic [43]	MLP	$\mathcal{O}(1)$	$\mathcal{O}(\mathcal{L} \cdot \mathcal{H})$
ET-BERT [4]	MHA	$\mathcal{O}(1)$	$\mathcal{O}(h \cdot \mathcal{L}^2 \cdot \mathcal{D})$
Ours(ET-SSM)	SSMs	$\mathcal{O}(1)$	$\mathcal{O}(h \cdot \mathcal{L} \cdot \mathcal{D})$

All experiments were conducted using Pytorch 1.8.0 on eight NVIDIA V100 GPUs. In our approach, we implemented two distinct learning strategies for the ET-SSM model to adapt to different levels of traffic data granularity, the ET-SSM (flow) and the ET-SSM (packet).

For testing, we maintained consistency in the dataset across both strategies, ensuring a fair and objective comparison with other methodologies. The pivotal difference between these strategies lay in the granularity of the fine-tuning input traffic information. Our method employed a dataset comprising a concatenated sequence of M consecutive packets within a flow, where M is predefined as 5 in our experimental setup.

C. Efficiency Evaluation

We conducted a theoretical analysis of the time complexity. The computational complexity for each layer was detailed in Table IV. In this table, we assigned a symbolic method to each core architecture type and subsequently analyzed their per-layer time complexities. Our ET-SSM's core architecture was based on State Space Models (SSMs), contrasting with other encrypted traffic classification methods that utilized CNN, RNN, MLP, and Multi-Head Attention (MHA). The parameters defined in the table included $\mathcal{L}$, which represents the input length; k, the kernel size of the convolution; $\mathcal{D}$, the representation dimension, and h, the number of attention heads, where H is typically less than $\mathcal{D}^2$. Consequently, ET-SSM's per-layer computational complexity was notably lower compared to CNN, RNN, MLP, and MHA. Furthermore, MLP, CNN, and MHA support parallel operations through vector multiplication, reducing their sequential operation complexity to merely $\mathcal{O}(1)$. In contrast, RNN requires $\mathcal{O}(\mathcal{L})$, due to its dependence on sequential execution at each time step. This sequential execution analysis shows that SSM-based routing does not exhibit quadratic complexity as the sequence length increases. Specifically, the ET-SSM facilitates achieving RNN-like long-range dependencies with a training speed comparable to CNNs.

D. Ablation Studies

In this study, we evaluated the impact of individual components within our framework at the packet level across six encrypted traffic classification datasets, as detailed in Table V. We established the original ET-SSM as our baseline for comparison. Initially, replacing the bidirectional SSM with a unidirectional SSM resulted in a minimum decline of **-3.93%** in the F1-score for the ISCX-VPN-Service dataset. This decline likely stemmed from the inability of the unidirectional SSM to capture bidirectional information as effectively as architectures like BERT [8] do. Additionally, substituting the convolutional module with a linear projection led to a decrease in the F1 score across all datasets, with a maximum drop of **-1.40%** observed in the Cross-Platform(iOS) dataset. This reduction may be attributed to the linear projection's inferior capability to capture adjacent information, which was crucial for enhancing the model's performance. Lastly, reverting to the original masking and next sentence prediction strategies employed in BERT [8] resulted in a maximum reduction of **-1.28%** in the F1-score for the Cross-Platform(iOS). This outcome underscored the effectiveness of masking strategies, similar to those proposed in [52], in encrypted traffic classification tasks.

E. Different Label Data Size Analysis

We conducted an extensive analysis to assess the impact of varying percentages of fine-tuning dataset volumes, with the variance in F1-score illustrated by the size of the error bands under packet-level fine-tuning, as depicted in Figures 4 and 5. These figures highlight the comparative performance of ET-SSM, YaTC, and ET-BERT across six datasets. It is observed that as dataset volumes decrease, ET-SSM consistently outperforms ET-BERT, achieving a minimum increase of **+0.41%** in classification performance. This improvement becomes even more pronounced at reduced dataset volumes, particularly below 40%, as shown in Figures 4 and 5. Furthermore, the

TABLE V
ABLATION STUDY ON SIX ENCRYPTED TRAFFIC CLASSIFICATION DATASETS

Method	Cross-Platform (iOS)		Cross-Platform (Android)		ISCX-VPN-Service		ISCX-VPN-App		ISCX-Tor		USTC-TFC	
	AC	F1	AC	F1	AC	F1	AC	F1	AC	F1	AC	F1
ET-SSM	0.9954	0.9974	0.9977	0.9860	0.9987	0.9940	0.9944	0.9978	0.9969	0.9976	0.9958	0.9975
w/o Bidirection	0.9571	0.9581	0.9417	0.9401	0.9412	0.9387	0.9391	0.9402	0.9423	0.9487	0.9413	0.9434
w/o Conv	0.9814	0.9834	0.9825	0.9764	0.9841	0.9864	0.9832	0.9864	0.9879	0.9893	0.9874	0.9865
w/o RoBERTa masking	0.9871	0.9857	0.9848	0.9804	0.9854	0.9823	0.9821	0.9879	0.9871	0.9903	0.9846	0.9847

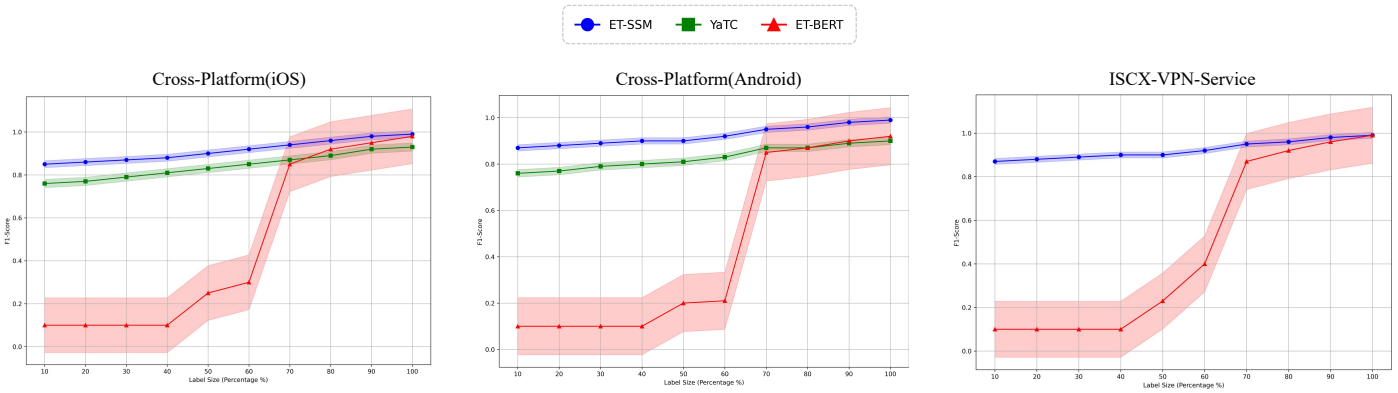


Fig. 4. F1-scores of ET-SSM, YaTC, and ET-BERT using varying percentages of fine-tuning samples on Cross-Platform(iOS), Cross-Platform(Android), and ISCX-VPN-Service datasets

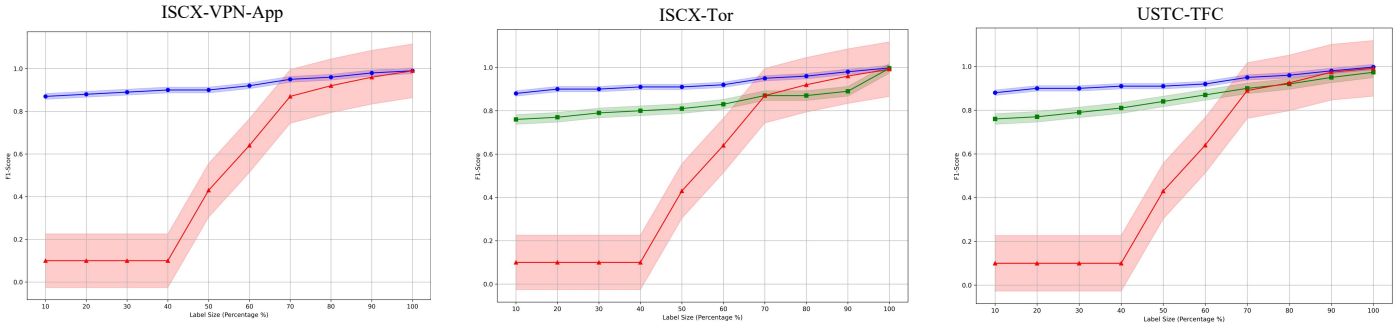


Fig. 5. F1-scores of ET-SSM, YaTC, and ET-BERT using varying percentages of fine-tuning samples on ISCX-VPN-App, ICSX-Tor, and USTC-TFC datasets

error band associated with ET-BERT is broader than that of ET-SSM, indicating greater variance in ET-BERT's F1 scores compared to those of ET-SSM. This observation underscores the superior robustness and enhanced classification efficacy of our proposed framework relative to ET-BERT when applied to packet-level fine-tuning. Conversely, while YaTC demonstrates comparably low variance, akin to ET-SSM, and performs well under limited label data scenarios, it does not match the robust encrypted traffic classification capabilities of ET-SSM across all dataset volumes. These findings further illustrate the superior robustness and improved classification efficacy of our framework compared to YaTC under similar conditions. YaTC did not test performance on the ISCX-VPN-Service and ISVC-VPN-App datasets under varying percentages of fine-tuning samples. Instead, we only compared our method with ET-BERT on these datasets.

REFERENCES

[1] S. Rezaei and X. Liu, "Deep learning for encrypted traffic classification: An overview," *IEEE communications magazine*, vol. 57, no. 5, pp. 76–81, 2019.

[2] P. Velan, M. Čermák, P. Čeleda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," *International Journal of Network Management (IJNM)*, vol. 25, no. 5, pp. 355–374, 2015.

[3] H. Zhang, L. Yu, X. Xiao, *et al.*, "Tfe-gnn: A temporal fusion encoder using graph neural networks for fine-grained encrypted traffic classification," in *Proceedings of the 32nd World Wide Web Conference (WWW)*, ACM, 2023, pp. 2066–2075.

[4] X. Lin, G. Xiong, G. Gou, Z. Li, J. Shi, and J. Yu, "Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classifica-

tion," in *Proceedings of the ACM Web Conference 2022 (WWW)*, 2022, pp. 633–642.

[5] R. Zhao, M. Zhan, X. Deng, *et al.*, "Yet another traffic classifier: A masked autoencoder based traffic transformer with multi-level flow representation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 37, 2023, pp. 5420–5427.

[6] C. Liu, L. He, G. Xiong, Z. Cao, and Z. Li, "Fs-net: A flow sequence network for encrypted traffic classification," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, IEEE, 2019, pp. 1171–1179.

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 000–16 009.

[10] U. Tariq, I. Ahmed, A. K. Bashir, and K. Shaukat, "A critical cybersecurity analysis and future research directions for the internet of things: A comprehensive review," *Sensors*, vol. 23, no. 8, p. 4117, 2023.

[11] A. Handa, A. Sharma, and S. K. Shukla, "Machine learning in cybersecurity: A review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1306, 2019.

[12] O. Abdulkader, A. M. Bamhdi, V. Thayananthan, F. Elbouraey, and B. Al-Ghamdi, "A lightweight blockchain based cybersecurity for iot environments," in *2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, IEEE, 2019, pp. 139–144.

[13] C.-Y. Shin, J.-T. Park, U.-J. Baek, and M.-S. Kim, "A feasible and explainable network traffic classifier utilizing distilbert," *IEEE Access*, 2023.

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[15] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, "Hippo: Recurrent memory with optimal polynomial projections," *Advances in Neural Information Orocessing systems (NIPS)*, vol. 33, pp. 1474–1487, 2020.

[16] K. Goel, A. Gu, C. Donahue, and C. Ré, "It's raw! audio generation with state-space models," in *International Conference on Machine Learning (ICML)*, PMLR, 2022, pp. 7616–7633.

[17] E. Nguyen, K. Goel, A. Gu, *et al.*, "S4nd: Modeling images and videos as multidimensional signals using state spaces," *arXiv preprint arXiv:2210.06583*, 2022.

[18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[19] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning (ICML)*, PMLR, 2017, pp. 933–941.

[20] H. Mehta, A. Gupta, A. Cutkosky, and B. Neyshabur, "Long range language modeling via gated state spaces," *arXiv preprint arXiv:2206.13947*, 2022.

[21] W. Hua, Z. Dai, H. Liu, and Q. Le, "Transformer quality in linear time," in *International Conference on Machine Learning (ICML)*, PMLR, 2022, pp. 9099–9117.

[22] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[24] X. Liu, J. Pool, S. Han, and W. J. Dally, "Efficient sparse-winograd convolutional neural networks," *arXiv preprint arXiv:1802.06367*, 2018.

[25] Y. Li, J. Hu, Y. Wen, *et al.*, "Rethinking vision transformers for mobilenet size and speed," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023, pp. 16 889–16 900.

[26] Y. Li, G. Yuan, Y. Wen, *et al.*, "Efficientformer: Vision transformers at mobilenet speed," *Advances in Neural Information Processing Systems (NIPS)*, vol. 35, pp. 12 934–12 949, 2022.

[27] T. Van Ede, R. Bortolameotti, A. Continella, *et al.*, "Flowprint: Semi-supervised mobile-app fingerprinting on encrypted network traffic," in *Network and distributed system security symposium (NDSS)*, vol. 27, 2020.

[28] W. Zheng, C. Gou, L. Yan, and S. Mo, "Learning to classify: A flow-based relation network for encrypted traffic classification," in *Proceedings of The Web Conference 2020*, 2020, pp. 13–22.

[29] V. F. Taylor, R. Spolaor, M. Conti, and I. Martinovic, "Robust smartphone app identification via encrypted network traffic analysis," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 13, no. 1, pp. 63–78, 2017.

[30] A. Panchenko, F. Lanze, J. Pennekamp, *et al.*, "Website fingerprinting at internet scale.," in *NDSS*, 2016.

[31] K. Al-Naami, S. Chandra, A. Mustafa, *et al.*, "Adaptive encrypted traffic fingerprinting with bi-directional dependence," in *Proceedings of the 32nd Annual Conference on Computer Security Applications*, 2016, pp. 177–188.

[32] J. Hayes and G. Danezis, "K-fingerprinting: A robust scalable website fingerprinting technique," in *25th USENIX Security Symposium (USENIX)*, 2016, pp. 1187–1203.

[33] P. Sirinam, M. Imani, M. Juarez, and M. Wright, "Deep fingerprinting: Undermining website fingerprinting defenses with deep learning," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 1928–1943.

[34] C. Liu, L. He, G. Xiong, Z. Cao, and Z. Li, "Fs-net: A flow sequence network for encrypted traffic classification," in *IEEE INFOCOM 2019-IEEE Conference On Computer Communications*, IEEE, 2019, pp. 1171–1179.

[35] M. Shen, J. Zhang, L. Zhu, K. Xu, and X. Du, "Accurate decentralized application identification via encrypted traffic analysis using graph neural networks," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 16, pp. 2367–2380, 2021.

[36] K. Lin, X. Xu, and H. Gao, "Tscrnn: A novel classification scheme of encrypted traffic based on flow spatiotemporal features for efficient management of iiot," *Computer Networks*, vol. 190, p. 107 974, 2021.

[37] M. Lotfollahi, M. Jafari Siavoshani, R. Shirali Hossein Zade, and M. Saberian, "Deep packet: A novel approach for encrypted traffic classification using deep learning," *Soft Computing*, vol. 24, no. 3, pp. 1999–2012, 2020.

[38] H. Y. He, Z. G. Yang, and X. N. Chen, "Pert: Payload encoding representation from transformer for encrypted traffic classification," in *2020 ITU Kaleidoscope: Industry-Driven Digital Transformation (ITU K)*, IEEE, 2020, pp. 1–8.

[39] R. Zhao, M. Zhan, X. Deng, *et al.*, "Yet another traffic classifier: A masked autoencoder based traffic transformer with multi-level flow representation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, AAAI Press, 2023, pp. 5420–5427.

[40] A. Gu, I. Johnson, K. Goel, *et al.*, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in Neural Information Processing Systems (NIPS)*, vol. 34, pp. 572–585, 2021.

[41] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[42] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[43] Y. Xu, J. Cao, K. Song, Q. Xiang, and G. Cheng, "Fasttraffic: A lightweight method for encrypted traffic fast classification," *Computer Networks*, vol. 235, p. 109 965, 2023.

[44] A. Gu, A. Gupta, K. Goel, and C. Ré, "On the parameterization and initialization of diagonal state space models," *arXiv preprint arXiv:2206.11893*, 2022.

[45] A. Gupta, "Diagonal state spaces are as effective as structured state spaces," *arXiv preprint arXiv:2203.14343*, 2022.

[46] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning," in *2017 International conference on information networking (ICOIN)*, IEEE, 2017, pp. 712–717.

[47] G. D. Gil, A. H. Lashkari, M. Mamun, and A. A. Ghorbani, "Characterization of encrypted and vpn traffic using time-related features," in *Proceedings of the 2nd international conference on information systems security and privacy (ICISSP)*, SciTePress Setúbal, Portugal, 2016, pp. 407–414.

[48] A. H. Lashkari, G. D. Gil, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of tor traffic using time based features," in *International Conference on Information Systems Security and Privacy (ICISSP)*, SciTePress, vol. 2, 2017, pp. 253–262.

[49] C. Liu, W. Wang, M. Wang, F. Lv, and M. Konan, "An efficient instance selection algorithm to reconstruct training set for support vector machine," *Knowledge-Based Systems*, vol. 116, pp. 58–73, 2017.

[50] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, *et al.*, "Toward generating a new intrusion detection dataset and intrusion traffic characterization.," *ICISSP*, vol. 1, pp. 108–116, 2018.

[51] P. Izsak, M. Berchansky, and O. Levy, "How to train bert with an academic budget," *arXiv preprint arXiv:2104.07705*, 2021.

[52] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

Toward Universal Detector for Synthesized Images by Estimating Generative AI Models

Ryota Seo*, Minoru Kuribayashi†, Akinobu Ura*, Antoine Mallet‡, Rémi Cogranne‡,
Wojciech Mazurczyk§ David Megías¶
* Okayama University, Japan, E-mail: {pr0765pj, pl843doc}@s.okayama-u.ac.jp
† Tohoku University, Japan, E-mail: kminoru@tohoku.ac.jp
‡ Troyes University of Technology, France, E-mail: {antoine.mallet, remi.cogranne}@utt.fr
§ Warsaw University of Technology, Poland, E-mail: wojciech.mazurczyk@pw.edu.pl
¶ Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Spain, E-mail: dmegias@uoc.edu

***Abstract*—One of the vulnerabilities in discriminators for AI-generated images is that the classification accuracy degrades when dealing with images generated using methods other than those they were trained on. As a countermeasure, in this study, we propose an image generation method estimator. The process of discrimination involves the input of an image to the estimator, which estimates the method used for its generation. Subsequently, a specialized fake image discriminator tailored to the estimated image generation method is used to identify the authenticity of the image. The activation functions are also considered according to the estimation results and analyzed for those discriminators. Discrimination scores are weighted and aggregated according to the estimation results, and the final decision is output. Our experimental results showed that the estimator achieved a classification accuracy of approximately 90% for 18 types of AI-generated images. Furthermore, by selecting the top two estimations in order of confidence, the accuracy increased to around 98%.**

I. Introduction

Since the introduction of Generative Adversarial Networks (GANs) in 2014 [1], numerous image generators have been developed [2]–[6], generating images so sophisticated that they are indistinguishable from those created by humans. Unlike the GAN-based models, diffusion models such as Stable Diffusion and Midjourney use forward and reverse processes to generate a composite image [7].

As a forensic technique, it has been studied to assess whether a given content is naturally captured by a camera or artificially produced. At the early stage of image generators, some artifacts are observed by human eyes if we carefully observe the synthesized images. With the development of machine learning techniques, such obvious artifacts no longer appear. CNN-based detectors have been studied as alternatives to the human eye, and quantitative and qualitative evaluations have been conducted on various types of datasets.

However, it raises a new challenge for the practical application of forensic techniques for analyzing artifacts, such that discrepancies between training and test data lead to poor detector performance. In the case of a deepfake detector, several detection methods achieve high test accuracy on one dataset, but show low detection accuracy on datasets that differ from the training phase. If the detector recognizes artifacts outside the training data set, the performance in terms of classification accuracy significantly degrades. It is thus necessary to evaluate the generalization of the classifiers.

A weakness of the currently available fake image detectors is that their classification accuracy drops significantly for the images generated using methods different from those encountered during their training. Mismatches in the training and testing phases are considerably difficult problems when discriminating against fake images. On the other hand, if a CNN-based detector is trained with various images generated by several generation methods, the classification accuracy becomes lower.

This study proposes a new universal fake image detection framework that achieves high classification accuracy, even with input generated by a method different from the training datasets. One of the advantages is the low computational cost of retraining new image generation methods that may emerge in the future. We introduce the idea of an "Estimator" that classifies an input image into groups according to how they were generated. The proposed framework consists of two main steps. First, an estimate is made of the input image on how the image was generated. Then, the image is classified as fake or real using each discriminator specialized for the corresponding generation method.

The discriminator dedicated to identifying specific still image generators is not limited to only the method with the highest confidence output by the Estimator but utilizes multiple discriminators based on activation output. This approach compensates for any misestimations made by the Estimator at the discrimination step. For such activation, three different algorithms are considered in this study. The first method is "Static," which focuses on each value of the confidence vector output by the Estimator. The second method, "Top-β," focuses on the number of special discriminators used for each input image. The third method is "Dynamic," which focuses on the sum of the values in the confidence vector output by the Estimator for the image generation method used for activation.

In the experiments, we design the estimator using a fine-tuned model of XceptionNet [8] pre-trained on ImageNet [9]. We validate the proposed estimator's ability to infer the generation method from a given image by employing 18 image generation methods, including GAN [1] models such as

StyleGAN [6] and GigaGAN [10], as well as diffusion models [11] such as Stable Diffusion [12] and DALL-E [13].

II. Related Works

In this section, we provide an overview of image generation techniques relevant to this study and review the difficulty of the forensics study.

A. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [1] involve two networks: the Generator, which generates images from random noise, and the Discriminator, which distinguishes between generated images and original images. The Generator aims to generate images that can deceive the Discriminator into classifying them as genuine images rather than fake ones. On the other hand, the Discriminator strives to minimize a loss function that enables it to accurately classify the images generated by the Generator as fake images.

1) StyleGAN: In StyleGAN [6], the structure of the Generator is modified. In traditional GANs, the Generator produces images from random noise. In StyleGAN, the input layer is omitted, and generation starts from pre-trained constants. Noise images are supplied to each synthesis network's layer. After each convolution in the Generator, normalization is performed using Adaptive Instance Normalization (AdaIN) [14]. Similar to PGGAN, StyleGAN also utilizes Progressive Growing for training.

StyleGAN has undergone various improvements over time. To eliminate the droplet-like noise observed in generated images, StyleGAN2 [15] was introduced, which utilizes weight demodulation with standard deviation-based normalization instead of AdaIN [14]. Lastly, there is StyleGAN3 [16], which ensures that all layers of the Generator are equivariant with respect to continuous signals, intending to generate images even more natural than those produced by StyleGAN2.

B. Latent Diffusion Models

A diffusion model [11] defines a diffusion process by simply adding easily manageable noise to input data and learning the data distribution by reversing the diffusion process over a finite time. Diffusion models using denoising autoencoders can generate highly sophisticated images. However, optimizing such a powerful diffusion model requires an enormous amount of time and resources. Therefore, a method called Latent Diffusion Models (LDM) was proposed in [17] to achieve learning of diffusion models under limited computational resources. Numerous improved models of LDMs [18]–[21] have been published, including Stable Diffusion [12], which reduces computational time and resources by separating the compression learning and generative learning phases. Other models include DALL-E 2 [13], which combines the multimodal model CLIP [22] with the diffusion model GLIDE [23].

C. Mismatch at Training and Testing

The study in [24] examined the effectiveness of various feature extraction algorithms in four automated facial manipulation techniques (Deepfakes, Face2Face, FaceSwap, NaturalTextures) [25]. High recognition accuracy is demonstrated when the training and testing methods are identical. However, when different methods are used for training and testing, the recognition accuracy decreases significantly.

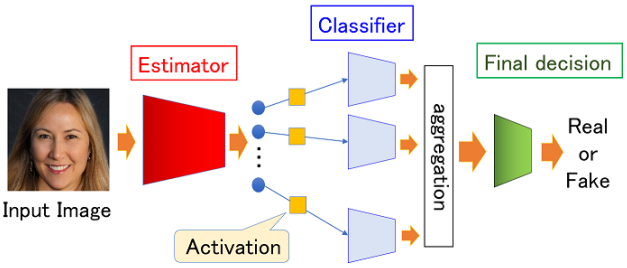


Fig. 1. Overview of proposed detector composed of the estimator of generation method and classifier.

Furthermore, the approach in [26] involves using a classifier trained using an enhanced spectrum and an RGB image-trained detector for assessment. When trained using images generated by the Progressive Growing of GANs (PGGAN), this method exhibits high recognition accuracy across multiple models with network structures similar to PGGAN. Data augmentation is introduced in [27], [28] to overcome dataset bias. However, the empirically designed augmentation strategies have certain limitations due to the poor generalizations.

III. Proposed Method

In this section, we propose an architecture for assessing images generated by image-generating AI systems. It comprises an *Estimator* and a *Classifier*. In the proposed method, we employ an estimation model $\mathrm{Est}(\cdot)$ and a set of discriminator models $(D_1(\cdot), D_2(\cdot) \ldots, D_n(\cdot))$, and the overall process is illustrated in Fig.1.

The Estimator assesses the method used to generate the input image. The estimated generation method corresponds to a fake image discriminator that evaluates whether the image is real or fake. In the classifier process, the estimated generation methods are not limited to the one with the highest output probability; instead, multiple top-ranked methods are selected and provided to the discriminator for assessment.

A. Estimator of Generation Method

An input image I is assessed using an estimation model $\mathrm{Est}(\cdot)$ to determine which generation method was applied to produce it. The model is trained using images generated by n different generation methods and outputs a confidence vector $\boldsymbol{p}$ of length n, where the i-th element is the probability that the input image was generated by the i-th generation method:

$$\boldsymbol{p} = (p_1, \ldots, p_n) = \mathrm{Est}(I), \tag{1}$$

where $\sum p_i = 1$.

In this study, we employ fine-tuning on XceptionNet, which has been pre-trained on ImageNet, to generate the estimation model $\mathrm{Est}(\cdot)$. In fine-tuning, the fully connected layer is removed and a newly designed fully connected layer is added along with a dropout layer and dense connections to output the confidence vector $\boldsymbol{p}$. As the default input image size for

TABLE I
HYPER-PARAMETERS IN THE PROPOSED ACTIVATION FUNCTIONS.

Method	Hyper-parameter
Static	T
Top-β	$\beta \in \mathbb{Z}_n$
Dynamic	$0 < \gamma \leq 1$

XceptionNet is 299×299, respectively, an input image I is resized to adjust these dimensions before being processed. It is worth mentioning that the fine-tuning models employed in this study can be replaced with a sophisticated model to improve performance.

When the input image is not produced by a generation method but is actually a captured photograph, the expected outcome for the confidence outputs is random. However, it does not affect the final decision due to the use of the classifier mentioned in the following section.

B. Classifier with Some Discriminators

First, we apply an activation function $\mathrm{Act}(\cdot)$ to the confidence vector $\boldsymbol{p}$. It selects the discriminator $D_i(\cdot)$ corresponding to each image generation method based on $\boldsymbol{p}$. According to a threshold T, $(0 < T \leq 1)$ determined in $\mathrm{Act}(\cdot)$, we calculate the binary vector $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) = \mathrm{Act}(\boldsymbol{p}) \tag{2}$$

where

$$\alpha_i = \begin{cases} 1 & \text{if } p_i \geq T \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Hence, the selection is not limited to one specific discriminator, but a group of discriminators. The details of the activation function are explained in the following section.

Next, for $1 \leq i \leq n$, the binary classification result $b_i \in \{0, 1\}$ is calculated by using the i-th discriminator $D_i(\cdot)$:

$$b_i = \alpha_i D_i(I) \tag{4}$$

Note that $b_i = 1$ implies that I is judged as fake by the discriminator $D_i(\cdot)$. If the input image is actually a real photograph, the result is expected to be $b_i = 0$ for all discriminators, hence, the "error" at the Estimator does not matter during the final decision.

At the final decision, the final probability P is calculated by aggregating the binary results $\boldsymbol{b}$ with the confidence vector $\boldsymbol{p}$ as follows:

$$P = \frac{\sum_{i=1}^{n} b_i p_i}{\sum_{i=1}^{n} \alpha_i p_i}. \tag{5}$$

If $P > 0.5$, the input image I is judged fake; otherwise real.

It is worth mentioning that each discriminator $D_i(\cdot)$ can be selected from conventional works, considering both the classification accuracy and computational complexity.

C. Activation

In this study, the following three methods for setting the threshold T are devised and compared. Table I enumerates the hyper-parameters utilized in these methods.

TABLE II
ACCURACY OF ESTIMATOR'S TOP-RANKED CONFIDENCE LEVELS.

	Estimator
Top1 Accuracy	0.900
Top2 Accuracy	0.982
Top3 Accuracy	0.994

1) Static Method: It is a simple method that employs a pre-determined threshold T without considering the confidence vector $\boldsymbol{p}$.

2) Top-β Method: In this method, the elements with the top β confidence values p_i are activated. Here β is a pre-determined parameter. At first, the confidence vector $\boldsymbol{p}$ is sorted in descending order:

$$\bar{\boldsymbol{p}} = (\bar{p}_1, \ldots, \bar{p}_n) = \mathrm{sort}(\boldsymbol{p}), \tag{6}$$

where $\mathrm{sort}(\cdot)$ denotes sorting the elements in descending order. Then, the threshold is selected as the β-th elements from $\bar{\boldsymbol{p}}$.

$$T = \bar{p}_\beta \tag{7}$$

3) Dynamic Method: Different from the Top-β technique, the dynamic method determines β according to the sum of their confidence values.

First, we calculate the sorted confidence vector $\bar{\boldsymbol{p}}$. Then, the sum of top confidence values is calculated, and β is determined for a pre-determined parameter γ:

$$\beta = \arg\min_i \left(\sum_i \bar{p}_i \geq \gamma \right) \tag{8}$$

Finally, the threshold T is selected using Eq.(7).

IV. EXPERIMENTS

In our experimental evaluation, we measure the classification accuracy of the estimator designed by fine-tuning XceptionNet. We used datasets consisting of images generated by 18 different image generation methods and real-world images.

We conducted two experiments. First, we investigated the accuracy of the Estimator. Second, we examined the overall accuracy of the proposed method in the presence of unknown image generators.

A. Estimation of Image Generation Method

In this experiment, we used 24-bit RGB color images generated by $n = 18$ different image generation methods. The resolution of generated images can be 512×512, 768×768, or 1024×1024 pixels, which are larger than the input size of XceptionNet (299×299 pixels). Therefore, each image was cut into 9 blocks of 299×299 pixels chosen from the top left, top center, top right, middle left, middle center, middle right, bottom left, bottom center, and bottom right of the image. We use 63,000 cropped blocks from 7,000 images for training, while 18,000 blocks from 2,000 images are used for testing. The estimated generation method was determined by averaging the confidence scores from the 9 cropped blocks.

Figure 2 describes a confusion matrix for identifying image generation methods. It arranges actual generation method on

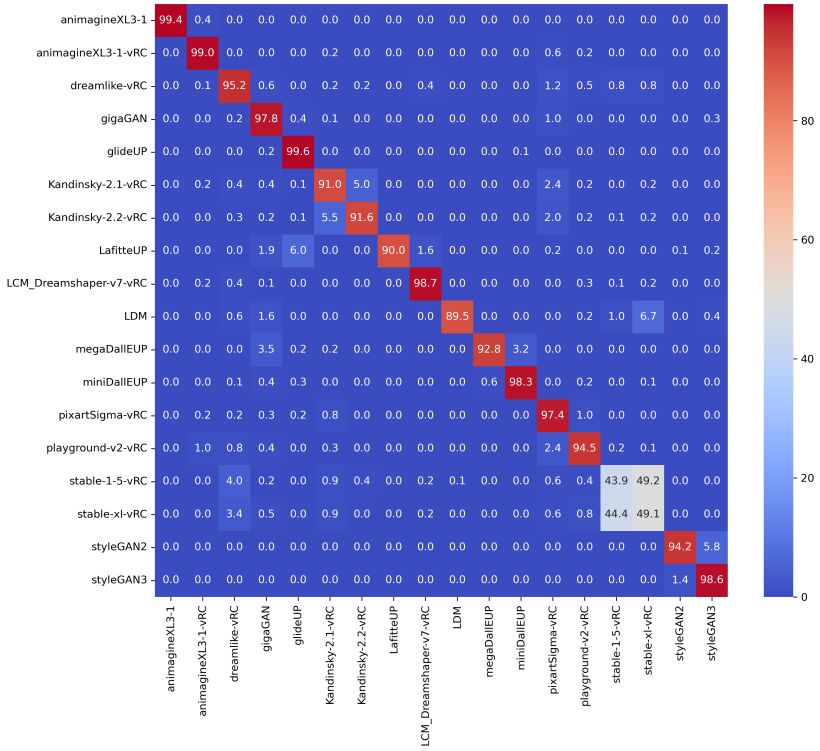


Fig. 2. Confusion matrix for estimator

the vertical axis and estimated generators on the horizontal axis, showing the accuracy with which the Estimator correctly classified image generation method output by each label ($2,000$ images per label). The classification accuracy was approximately 90%. In the case of two versions of stable diffusion method, it occurs misclassification due to the similar characteristics appeared in the generated images. In the second step of the proposed method, such misclassification is not a problem because the corresponding discriminator $D_i(\cdot)$ can output binary classification results with high accuracy. Additionally, as shown in Table II, the "Top2 Accuracy" was around 98%, indicating that the estimator accurately classified the image generation method within the Top2 class. Based on the above results, the Estimator demonstrates high classification accuracy, allowing the proposed method to achieve high accuracy comparable to using all classifiers specialized for each image generation method, while incurring lower computational costs.

B. Comparison of Activation Methods

If an image generated by an unknown image generation method, the confidence vector $\boldsymbol{p}$ calculated by Estimator is expected to vary across n classes and each discriminator $D_i(\cdot)$ may not accurately classify whether it is generated. To simulate such a situation, we train the Estimator excluding one specified image generation method and measure the performance of the proposed method for the images generated by the excluded image generation method.

In this experiment, we compared three activation methods

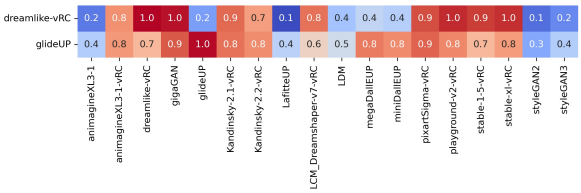


Fig. 3. Comparison of confidence value as outputs of two discriminators "Dreamlike" and "Glide".

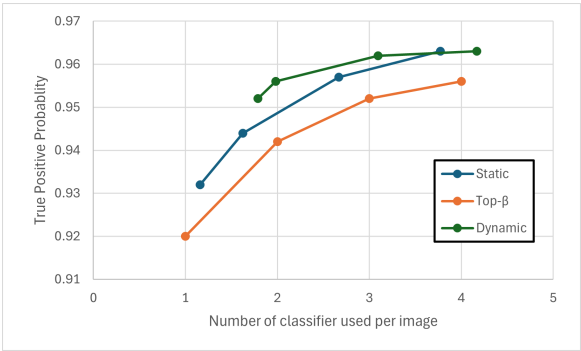


Fig. 4. Comparison of three activation methods in terms of true positive probability by using images generated by "Glide".

for images generated by 18 different image generation methods and real photographs. We train each discriminator $D_i(\cdot)$, which is a fine-tuned version of the XceptionNet, using the images generated by each image generation method. These

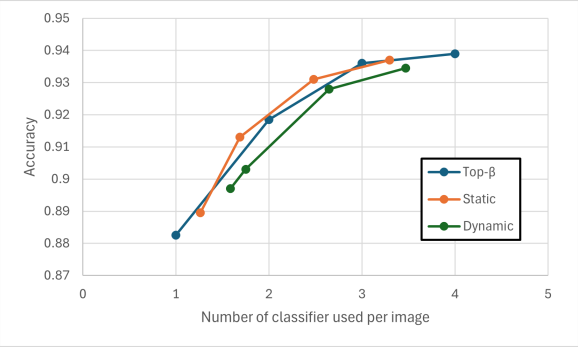


Fig. 5. Comparison of three activation methods in terms of classification accuracy in the case of "Glide".

discriminators were designed to perform binary classification between images generated by each image generation method and real photographs from the FFHQ dataset [29]. For training, we used 63,000 cropped images of size 299×299 pixels for each label, using the same method as the Estimator. Note that, when activation was not used, the accuracy was approximately 99%.

First, as a preliminary experiment, we investigated how confident each discriminator, specialized in a particular image generation method, was when classifying images generated by other methods. The investigation was conducted using 2000 images for each method. Due to the page limitation, we only show the results of two image generation methods "Dreamlike[1]" and "Glide[2]" in Fig. 3. It is observed that the classification accuracy is not necessarily low for images generated by unknown image generation methods. A discriminator specialized for one image generation method can be said to be compatible to a certain extent with other image generation methods in terms of classification accuracy.

In both cases, the Estimator is trained by excluding a specific image generation method and outputs $\boldsymbol{p}$ with $n = 17$. In the case of "Dreamlike," some mismatched discriminators show similar classification accuracy to the matched discriminators. Therefore, the overall accuracy is close to the matched (Dreamlike) case, and there was no difference in accuracy due to the activation methods.

On the other hand, in the case of "Glide," the classification accuracy of some top discriminators tends to be lower than that of the matched discriminator. Therefore, the selection of a good $D_i(\cdot)$ plays an important role, which is controlled by $\mathrm{Act}(\boldsymbol{p})$. The true positive probability using 1000 images generated by "Glide" is shown in Fig. 4, and the classification accuracy using 1000 real images and 1000 images generated by "Glide" is shown in Fig. 5. For images generated by "Glide," the Top-β method specifies a fixed number of $D_i(\cdot)$, thus ignoring the information extracted by the Estimator. While the Static method does not consider the choice of threshold T, the Dynamic method provides flexibility in selecting the threshold. This improves the stability of predictions and the appropriateness of the number of activated discriminators in the proposed method. On the other hand, for real photographs, it is found that the Dynamic method needs to activate relatively more classifiers compared to the other two methods.

Experimental results show that when a particular discriminator, such as "Dreamlike", exhibits high classification accuracy, there is little need for Activation. This is because the misclassification correction capability provided by Activation is well complemented by other discriminators and only results in additional computational costs. On the other hand, when the classification accuracy of some of the top discriminators is lower than that of the matched discriminators, as seen in "Glide", Activation has been confirmed to be effective in improving classification accuracy and can be used in such a situation.

Based on the results, the Static method is considered the most suitable among the three proposed methods for both real photographs and generated images. However, as the Dynamic method shows better accuracy for generated images, developing more flexible algorithms is necessary. In particular, it is important to optimize the balance between computational efficiency and accuracy by more effectively utilizing the confidence information obtained from the Estimator to improve adaptability to unknown image generation methods. Additionally, the real photograph data used may be biased, so further experiments with more diverse, raw data are required to improve the reproducibility of the results.

V. Conclusions

In this paper, we proposed a two-stage classification architecture for detecting synthesized images created by a generative AI model. An estimator was introduced to improve the classification accuracy of a detector for synthesized images. We envision a workflow that estimates several possible image generation methods from a suspicious image and aggregates the binary classification results obtained from each discriminator to determine whether the image is a fake image or not.

In the experiment, we designed an estimator using a model that is a fine-tuned version of the XceptionNet. The use of other models and tuning the parameters to further improve accuracy is a future task. In this experiment, the estimator was also classified by image generation method. However, in the future, by integrating image generation methods with similar features, the estimator aims to reduce the number of classes to classify and at the same time improve the accuracy.

Acknowledgment

The authors acknowledge the funding obtained from the EIG CONCERT-Japan call to the project Detection of fake newS on SocIal MedIa pLAtfoRms "DISSIMILAR" through grants PCI2020-120689-2 (Agencia Estatal de Investigación, Spain), EIG CONCERT-JAPAN/05/2021 (National Centre for Research and Development, Poland), and JPMJSC20C3 (Japan Science and Technology Agency, Japan).

The work of D. Megías was supported, in part, by the "SECURING" project (PID2021-125962OB-C31) funded by

[1]https://huggingface.co/dreamlike-art/dreamlike-diffusion-1.0
[2]https://github.com/openai/glide-text2im

the Ministry of Science and Innovation, the Agencia Estatal de Investigación, and the European Regional Development Fund (ERDF), and by the ARTEMISA International Chair of Cybersecurity (C057/23) and the DANGER Strategic Project of Cybersecurity (C062/23), both funded by the Spanish National Institute of Cybersecurity through the European Union – NextGenerationEU and the Recovery, Transformation, and Resilience Plan. The work of M. Kuribayashi was supported by the JSPS KAKENHI (22K19777).

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *CoRR*, 2017. arXiv: 1710.10196.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV'17*, 2017, pp. 2242–2251.

[4] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. CVPR'18*, 2018, pp. 8789–8797.

[5] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. ICLR'19*, 2019.

[6] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2021.

[7] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion models in generative ai: A survey," *CoRR*, 2023. arXiv: 2303.07909.

[8] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR'17*, 2017, pp. 1800–1807.

[9] J. Deng, W. Dong, R. Socher, K. L. L. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR'09*, 2009, pp. 248–255.

[10] M. Kang, J.-Y. Zhu, R. Zhang, *et al.*, "Scaling up gans for text-to-image synthesis," in *Proc. CVPR'23*, 2023, pp. 10 124–10 134.

[11] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. ICML'15*, vol. 37, 2015, pp. 2256–2265.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, 2021. arXiv: 2112.10752.

[13] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *CoRR*, 2022. arXiv: 2204.06125.

[14] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. ICCV'17*, 2017, pp. 1510–1519.

[15] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR'20*, 2020, pp. 8107–8116.

[16] T. Karras, M. Aittala, S. Laine, *et al.*, "Alias-free generative adversarial networks," in *Proc. NeurIPS'21*, 2021.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR'22*, 2022, pp. 10 674–10 685.

[18] Y. Zhou, R. Zhang, C. Chen, *et al.*, "Towards language-free training for text-to-image generation," in *Proc. CVPR'22*, 2022, pp. 17 907–17 917.

[19] J. Chen, J. Yu, C. Ge, *et al.*, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," *CoRR*, 2023. arXiv: 2310.00426.

[20] J. Chen, C. Ge, E. Xie, *et al.*, "Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation," *CoRR*, 2024. arXiv: 2403.04692.

[21] A. Razzhigaev, A. Shakhmatov, A. Maltseva, *et al.*, "Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion," *CoRR*, 2023. arXiv: 2310.03502.

[22] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML'21*, 2021, pp. 8748–8763.

[23] A. Nichol, P. Dhariwal, A. Ramesh, *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *CoRR*, 2021. arXiv: 2112.10741.

[24] Y. Xu and S. Y. Yayilgan, "When handcrafted features and deep features meet mismatched training and test sets for deepfake detection," *CoRR*, 2022. arXiv: 2209.13289.

[25] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics++: Learning to detect manipulated facial images," in *Proc. ICCV'19*, 2019, pp. 1–11.

[26] M. Tanaka, S. Shiota, and H. Kiya, "A universal detector of CNN-generated images using properties of checkerboard artifacts in the frequency domain," *CoRR*, 2021. arXiv: 2108.01892.

[27] L. Li, J. Bao, T. Zhang, *et al.*, "Face X-ray for more general face forgery detection," in *Proc. CVPR'20*, 2020, pp. 5000–5009.

[28] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proc. ICCV'21*, 2021, pp. 15 003–15 013.

[29] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR'19*, 2019, pp. 4401–4410.

Innovative Information Hiding in H.266/VVC Using Sub-Block Transform Technique

Hau Joan[#], Yiqi Tew[*] and Li Peng Tan[+]
Faculty of Computing and Information Technology
Tunku Abdul Rahman University of Management and Technology
Kuala Lumpur, Malaysia
E-mail: {[#]joanh-wm18@student., [*]yiqi@, [+]tanlp@} tarc.edu.my

***Abstract*— This paper presents a novel method for information hiding in H.266/Versatile Video Coding (VVC) using the Sub-Block Transform (SBT) technique. The proposed method embeds data by selectively manipulating SBT modes to achieve minimal impact on video quality and compression efficiency. By manipulating SBT selections, we aim to achieve minimal degradation in video quality and compression efficiency while embedding information. By limiting the embedding to SBT modes 1 (SBT_VER_HALF) and 2 (SBT_HOR_HALF), the method ensures higher hiding capacity while preserving video integrity. Performance evaluation shows that the proposed method achieves a favorable trade-off between hiding capacity, video quality, and compression efficiency, outperforming benchmark methods like Selective Quantization Technique (SQT) and Multiple Transform Selection (MTS). Experimental results confirm that our approach offers a robust solution for information hiding with minimal visual degradation and higher information hiding capacity. To the best of our knowledge, this research presents the first information hiding technique specifically designed for H.266/VVC that leverages SBT.**

***Keywords*— Information Hiding, Versatile Video Coding (VVC), Sub-Block Transform (SBT), Selective Quantization Technique (SQT)**

I. INTRODUCTION

Video has become an indispensable component of contemporary society, serving as a primary medium for communication and information dissemination across diverse platforms, including social media, education, business, and entertainment. The medium's capacity to engage audiences through visually compelling and informative content has contributed to its widespread adoption. The ubiquity of high-speed internet and mobile devices has further accelerated video consumption, making it a powerful tool for knowledge sharing and public awareness. Research by Syamsulaini and Mashitoh [1] underscores the pedagogical efficacy of video in educational settings. Nevertheless, the ease with which video content can be created and distributed raises concerns about its reliability as a medium for personal and professional communication in the digital age [2]. This study aims to address this challenge by exploring the potential of the Versatile Video Coding (VVC) standard and advanced information hiding techniques to fortify the security of digital ecosystems. By investigating these areas, this research contributes to the development of robust solutions that safeguard privacy and confidentiality in digital transactions and content delivery.

Video technology has emerged as a pivotal force in contemporary society, permeating diverse sectors and reshaping human interaction. Its applications span education, economics, science, healthcare, and entertainment, among others [3 - 5]. As a cornerstone of the information age, video has transformed how individuals consume, create, and share content. Its impact on education, in particular, is undeniable, with video-based instruction enhancing engagement and learning outcomes. Moreover, the entertainment industry has been revolutionized by immersive experiences offered by virtual reality [6]. The proliferation of social media platforms has further solidified video's role in daily life, enabling global connectivity and content creation [7 - 9]. Recent advancements in steganography and watermarking techniques have explored distortion-less and energy efficient approaches for embedding data in multimedia content, particularly in images and videos, as demonstrated in [10], [11].

The rapid evolution of video coding standards and the growing importance of digital content security have created a compelling need for robust video protection mechanisms. From early standards like MPEG-2 to the advanced High Efficiency Video Coding (HEVC) and the latest VVC [12], the industry has consistently sought to improve compression efficiency while maintaining visual quality. However, the increasing accessibility and value of video content have made it a prime target for cyberattacks, including unauthorized access and tampering [13], [14], [15].

Information hiding offers a promising countermeasure by embedding data within video content without compromising its perceived quality or functionality [16], [17]. While the concept has a rich history, its integration into the complex structure of VVC remains a relatively unexplored area. The imperative for this research is driven by the escalating demand for both privacy and robustness in digital communications, offering

solutions for enhancing security, protecting copyright, and enabling covert communication. As data breaches pose increasingly severe risks to individuals and organizations, the development of sophisticated information hiding techniques within the VVC framework becomes paramount. This study delves into the theoretical underpinnings and practical implementation of embedding and extracting covert information within VVC video streams, assessing their efficacy across diverse scenarios.

II. Information Hiding in Video Coding

Existing research has extensively explored information hiding techniques within the HEVC compression standard. These methods can be classified according to their embedding strategies: partition modes, prediction units, transform coding, and syntax elements. Partition mode-based techniques modify the Prediction Unit (PU) partition modes within Coding Unit (CU) [18] . However, these methods are limited to 16x16 and 8x8 CUs. Another partition mode-based approach [19] extends this concept to 8x8 CU by modifying PU partition modes during the inter-prediction process. Additionally, a PU-based technique [20] modifies PUs to embed secret messages, employing a non-standard in-loop filter within the HM encoder. Transform coding-based methods [21] modify transform block partitions for information hiding but are restricted to 8x8 Transform Blocks (TBs) and a limited number of frames. Syntax element-based techniques, such as encrypting quantized transform coefficients (TC) and motion vector differences (MVDs) during Context-based Adaptive Binary Arithmetic Coding (CABAC) and inter-prediction processes [16], have also been explored. Nevertheless, these methods primarily focus on HEVC videos in low-delay mode, with limited evaluation in other video modes.

Building upon the successes of information hiding techniques within previous video compression standards, researchers have continued to explore this domain within the VVC framework. Partition mode-based methods have been extended by extracting CU partitions during VVC encoding to create binary masks [22]. Additionally, Li et al. [23] proposed a technique modifying chroma CU partition modes for information embedding, albeit with limitations in terms of frame rate and quantization parameter (QP) values. Both approaches exhibit constraints in terms of embedding capacity.

Transform coding-based techniques have been further explored by modifying transform modes within Multiple Transform Selection (MTS) and Cross-Component Linear Model for I-frames [24]. However, this approach is limited to I-frames, restricting the amount of embeddable information. Syntax element-based information hiding techniques have been integrated into VVC by encrypting critical components within the CABAC process. These components include TC, motion vectors (MVs), adaptive loop filters, inter-prediction modes, sample adaptive offset filters, and intra-prediction modes. The application of specific QP values (e.g., 17, 22, 27, 32, 37) within this approach can impact overall video quality [25]. Alternatively, a selective encryption method targeting luma intra prediction modes, MVDs, and residual signs has been proposed, potentially resulting in significant visual distortions [26]. Building on these approaches, a data embedding technique that modifies the quantization parameter has also been introduced in VVC, further expanding the possibilities for information hiding [27], [28].

While a plethora of information hiding techniques have been introduced within the VVC standard, a notable gap exists in the literature regarding the exploitation of sub-block transforms (SBT) for covert data embedding. This research aims to address this limitation by proposing a novel information hiding approach that leverages selective SBT within the VVC framework. It achieves a better trade-off among hidden capacity, video quality and compression efficiency.

III. The Proposed Method

This paper presents a novel information hiding technique tailored to the VVC standard by exploiting SBT. Selective SBT techniques employ a strategic modification of SBT values to 1 or 2 for embedding binary information within the video stream. While VVC offers four distinct SBT modes (SBT $\in \{1, 2, 3, 4\}$), which include SBT_VER_HALF, SBT_HOR_HALF, SBT_VER_QUAD and SBT_HOR_QUAD, experimental analysis revealed that restricting SBT values to 1 and 2 optimizes embedding capacity without compromising video quality. These modes are advantageous for embedding binary data because their larger sub-block size introduces minimal distortion when modified. In contrast, the SBT_VER_QUAD and SBT_HOR_QUAD modes, which split the blocks into smaller segments, are more prone to affecting video quality when altered. By restricting data embedding to SBT modes 1 and 2, our approach maintains a higher hiding capacity with minimal perceptual degradation. This selective approach allows for efficient data concealment without significantly affecting visual quality or compression efficiency.

A. Selective Sub-Block Transform Encoding Technique

This methodology ensures that alterations to the stream are conducted only when there are enough bits available for embedding within the video stream. The upper limit of bits that can be embedded is represented by η, which is determined based on the size of the file being embedded whenever necessary. In situations where data concealment is not required, η is adjusted to zero to ensure minimal degradation, modifying the SBT only when there are bits to be concealed. The SBT modification strategy involves switching the existing SBT to '1' for encoding a '0' bit, and to '2' for a '1' bit. Alterations to the

SBT are omitted if data concealment is unnecessary or if the limit of embeddable bits has been met. A comprehensive account of the process for encoding binary data within the video stream is provided subsequently.

Let's define:

$\mathscr{b}_i$: The i-th bit of the binary representation of the message intended for concealment, where $\mathscr{b}_i \in \{0,1\}$

$\mathcal{S}_{orig,i}$: The original SBT of the i-th slice.

$\mathfrak{s}_i$: The modified SBT of the i-th slice.

η : The maximum number of bits designated for embedding.

Modification of SBT based on detected bits:

$$\mathfrak{s}_i = \begin{cases} 1, & if\ \mathscr{b}_i == 0\ \wedge i < \eta \\ 2, & if\ \mathscr{b}_i == 1\ \wedge\ i < \eta \\ \mathcal{S}_{orig,i}, & if\ \ i \geq \eta \vee \eta\ ==\ 0 \end{cases} \quad (1)$$

The modification of SBT might not be fully exploited for the purpose of information hiding during the transformation stage. Therefore, during the CABAC stage, this technique evaluates the potential to embed additional bits into the current transform block. The method is described in the pseudocode presented below:

Algorithm 1: Algorithm to Find the Opportunity to Hide Bits in CABAC

Input

α: Current coding unit data.

β: Array of encoded text data, with each element containing a bit and an encoding status ('N' for not encoded, 'Y' for encoded).

γ: Index to the first non-encoded text element (set to -1 for no operation).

$\mathfrak{s}$: Current modified SBT.

$\mathfrak{e}$: Encoding status for each coding unit.

$\mathfrak{b}$: Binary bits, where $\mathfrak{b} \in \{0,1\}$.

Output

Encoded bits status updated, returns true if bits were hidden successfully.

Initialization

If $\gamma \neq -1$, proceed to check the encoding status.

SBT Mode Verification

If $\alpha_{\mathfrak{s}} = SBT_VER_HALF$:

  Iterate over $\beta_i\ from\ 0\ to\ \beta_{i_{\max-1}}$

    If $\beta_{i_{\mathfrak{b}}} = '0'$ and $\beta_{i_{\mathfrak{e}}} = 'N'$

      Set $\beta_{i_{\mathfrak{e}}} \leftarrow$ 'Y'

      Write the position i to the text file.

      Return true and terminate the algorithm.

Alternative SBT Mode:

Else, if $\alpha_{\mathfrak{s}} = SBT_HOR_HALF$:

  Iterate over $\beta_i\ from\ 0\ to\ \beta_{i_{\max-1}}$

    If $\beta_{i_{\mathfrak{b}}} = '1'$ and $\beta_{i_{\mathfrak{e}}} = 'N'$

      Set $\beta_{i_{\mathfrak{e}}} \leftarrow$ 'Y'

      Write the position i to the text file.

      Return true and terminate the algorithm.

Completion:

Write a termination indicator to the file with value -1.

Return true.

Algorithm 1 is intended to handle the current CU and modify an array of encoded text data based on certain conditions. This algorithm is vital in the field of information hiding within VVC, especially in deciding how and when to embed specific bits into the text data array. The algorithm starts by checking the validity of the index parameter (γ). If γ is deemed invalid $(\gamma = -1)$, the function exits without making any alterations. Otherwise, it continues to inspect the SBT mode of the current CU $(\alpha_{\mathfrak{s}})$ to decide the next steps. In the case of a vertical half SBT $(\alpha_{\mathfrak{s}} = SBT_VER_HALF)$, the algorithm looks for the first '0' bit that has not been encoded in the array, which includes both binary data and the status of each bit's encoding (β). Once it finds a valid bit, it updates the encoding status $(\mathfrak{e})$ and records the index (i) in a text file through the writeToFile function. For a horizontal half sub-block $(\alpha_{\mathfrak{s}} = SBT_HOR_HALF)$, the algorithm searches for the first unencoded '1' bit in the β array and updates accordingly. If no suitable bits are found for encoding in either case, a default value of -1 is written to the file using writeToFile(-1). The function returns true if the operation is complete, regardless of whether a bit was encoded

B. Selective Sub-Block Transform Decoding Technique

The decoding algorithm begins by loading all index values from the sequence text file into an array. If the SBT is '1' and the corresponding index in the sequence is not '-1' (indicating the presence of a hidden bit), a '0' is appended to the decoded text array. Conversely, if the SBT is '2', a '1' is appended. A counter is used to track the number of processed locations. This decoding mechanism depends on the presence of hidden bits at specific indices. If no bit is concealed at a given index (as indicated by a sequence value of '-1'), no decoding occurs, and the counter is incremented and move to the next location. The following algorithm outlines the extraction of binary information using the selective SBT technique:

Let:

$\mathbb{d} = \{d_{i\,i=1}^{M}\}$: The array of decoded text, where M is the length of the text.

$\mathbb{m} = \{n_{i\,i=1}^{N}\}$: The array of loaded indices from the sequences file (text file), where N is the total number of indices.

$\mathfrak{s}$: The current SBT mode

Extraction of binary bits based on detected SBT:

$$d_{n_i} = \begin{cases} 0, & if\ \mathfrak{s} = SBT_VER_HALF\ and\ n_i \neq -1 \\ 1, & if\ \mathfrak{s} = SBT_HOR_HALF\ and\ n_i \neq -1 \\ No\ Operation, if\ n_i = -1. \end{cases} \quad (2)$$

IV. Experiment Results and Discussions

In this study, the encoder with a random access configuration profile is utilized to evaluate the proposed information hiding technique at bitrates ranging from 100 kbps to 5000 kbps. 60 frames are encoded from selected test video sequences to

ensure a comprehensive analysis of diverse content types. The random access configuration enabled efficient frame navigation and retrieval, crucial for applications requiring frequent access to various video segments, allowing for a balanced assessment of visual quality and information hiding capacity while supporting inter-frame dependencies.

All proposed methods, along with benchmark techniques, were implemented within the same coding environment—VVC reference software VTM 23.1. This consistency eliminates potential bias from using different coding standards, ensuring fair comparisons. The research objective is to demonstrate the effectiveness of the selective SBT technique in achieving a higher information hiding capacity within the H.266/VVC compared to other strategies. Results in Tables 1 and 2 provide a comprehensive comparison of our method against the MTS-based [24] and the SQT [27] approaches.

To evaluate the effectiveness of the proposed method, three key metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and information hiding capacity are employed. PSNR measures visual quality retention, with higher values indicating better preservation. SSIM assesses the structural and perceptual quality maintained from the viewer's perspective. Lastly, information hiding capacity quantifies the amount of embedded data without significant degradation, with higher capacity facilitating greater data concealment while preserving video quality.

Table 1 presents quality metrics, specifically SSIM and PSNR, for three methods applied to two test video sequences ("Blowing Bubbles" and "People On Street"), under conditions of embedding both '0' and '1' bits. The selective SBT technique consistently achieves high SSIM values across all test cases, indicating minimal perceptual degradation. For example, in the "Blowing Bubbles" sequence, the SSIM values for the proposed method remain above 0.9157, peaking at 0.9984. This performance is comparable to the SQT method and generally superior to the MTS-based approach, which shows slightly lower SSIM values. In the "People On Street" sequence, characterized by more complex motion and texture, the proposed method still performs strongly, with SSIM values ranging from 0.8453 to 0.9678, thereby demonstrating robust visual quality preservation even in challenging scenarios. PSNR values, reflecting objective quality, indicate that the proposed method exhibits competitive performance, slightly trailing the SQT approach but outperforming the MTS-based method in most instances. For instance, in the "Blowing Bubbles" sequence, PSNR values for the proposed method range from 27.8335 dB to 43.5494 dB, closely aligning with SQT values with a maximum difference of 0.1 dB and significantly better than MTS-based results. In the "People On Street" sequence, the SBT method achieves SSIM and PSNR values of 0.8663 and 24.3804, respectively, effectively maintaining high video quality and structural integrity even with substantial data embedding.

Table 2 highlights the information hiding capacity of the three methods, comparing the number of bits that can be embedded at different targeted bitrates. The MTS-based approach exhibits considerable variation in capacity based on video content and bitrate. In "Blowing Bubbles", the capacity ranges from 122 to 1028 bits for '0' bits and from 244 to 794 bits for '1' bits. Conversely, in "People On Street", the capacity increases significantly, indicating a sensitivity to video complexity and motion. The SQT method demonstrates a fixed and notably low embedding capacity of 60 bits per frame for both '0' and '1' bits across all conditions, which limits its applicability in scenarios requiring higher data embedding. In contrast, the proposed technique outperforms both alternatives, particularly at higher bitrates. For "Blowing Bubbles", the method achieves capacities ranging from 61 to 1929 bits for '0' bits and from 50 to 1831 bits for '1' bits. Notably, in "People On Street", the proposed method significantly surpasses the other techniques, reaching capacities of up to 9188 bits for '0' bits and 6186 bits for '1' bits. This substantial increase in capacity demonstrates the effectiveness of the selective SBT technique in leveraging the structural aspects of VVC to embed a larger amount of information without significantly compromising video quality.

From the graph in Fig. 1, it is evident that the SQT generally achieves the highest PSNR across the range of bitrates, indicating superior preservation of video quality. This suggests that SQT involves minimal modification of the video signal, resulting in only slight degradation when embedding information. The Selective SBT method, represented by a distinct line, shows slightly lower PSNR values compared to SQT. However, the PSNR difference is notably minimal, with a maximum divergence of around 0.1 dB, particularly evident in the zoomed-in section of the graph for bitrates between 3.0 and 3.1 kbps.

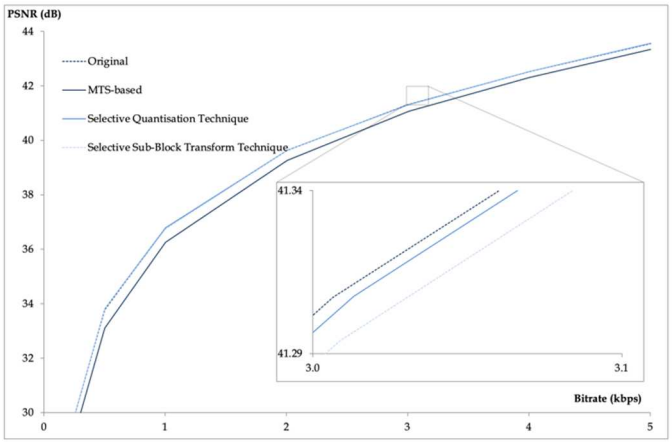


Fig. 1: RD Curve of BlowingBubbles with All Bit ‘0’ Hidden: A Focus on Original Vs Proposed Algorithm

Table 1: Comparison of Embedding Strategies for MTS-based, SQT, and Proposed Method with All Bit '0' and '1' Hidden

Video Sequences	MTS-based [24]				SQT [27]				Proposed Method (Selective Sub-Block Transform)			
	Bit '0'		Bit '1'		Bit '0'		Bit '1'		Bit '0'		Bit '1'	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
Blowing Bubbles	0.9026	27.3536	0.9024	27.3352	0.9152	27.8385	0.9175	27.8594	0.9157	27.8424	0.9157	27.8335
	0.9785	33.1087	0.9782	33.0761	0.9831	33.8170	0.9827	33.7846	0.9827	33.7971	0.9825	33.7668
	0.9912	36.2484	0.9912	36.2541	0.9924	36.7852	0.9923	36.7652	0.9923	36.7445	0.9924	36.7616
	0.9959	39.2690	0.9959	39.2749	0.9962	39.6325	0.9962	39.6350	0.9962	39.6183	0.9962	39.6158
	0.9973	41.0683	0.9973	41.0693	0.9974	41.3076	0.9974	41.3097	0.9974	41.2939	0.9974	41.2950
	0.9979	42.3036	0.9979	42.3075	0.9980	42.5312	0.9980	42.5277	0.9980	42.5215	0.9980	42.5204
	0.9983	43.3325	0.9983	43.3549	0.9984	43.5696	0.9984	43.5532	0.9984	43.5491	0.9984	43.5494
People On Street	0.8222	23.2314	0.8226	23.2353	0.8457	23.7342	0.8446	23.7016	0.8456	23.7220	0.8453	23.7250
	0.8444	23.8586	0.8440	23.8483	0.8632	24.2880	0.8622	24.2645	0.8663	24.3804	0.8653	24.3613
	0.8801	25.2264	0.8802	25.2368	0.8946	25.6738	0.8942	25.6617	0.8944	25.6578	0.8940	25.6483
	0.9170	27.4074	0.9166	27.4043	0.9276	27.9910	0.9269	27.9321	0.9272	27.9294	0.9270	27.9275
	0.9361	28.9803	0.9363	28.9834	0.9452	29.6455	0.9462	29.7340	0.9450	29.5754	0.9449	29.5754
	0.9472	30.0635	0.9471	30.0547	0.9585	31.0900	0.9567	30.8925	0.9565	30.8637	0.9566	30.8636
	0.9575	31.3081	0.9576	31.3103	0.9678	32.4626	0.9655	32.1310	0.9655	32.1171	0.9653	32.0964

This slight difference underscores that while SQT might slightly outperform SBT in terms of video quality retention, the degradation introduced by SBT is marginal and likely imperceptible to most viewers. This suggests that the SBT method, despite embedding more information, maintains a high level of video quality.

The experimental results validate the effectiveness of the selective sub-block transform technique for information hiding in H.266/VVC. The proposed method not only maintains high visual quality, as evidenced by competitive SSIM and PSNR values, but also significantly enhances the information hiding capacity, particularly in more complex video sequences. The findings underscore the potential of this technique to provide a robust and efficient solution for applications requiring high-capacity data embedding with minimal visual distortion.

V. Conclusion

In this research, a novel information hiding technique for H.266/VVC using the SBT method is introduced. This approach leverages the advanced encoding capabilities of VVC to conceal data efficiently without significantly impacting the visual quality of the video. By selectively modifying the sub-block transform modes, a minimal perceptual degradation is ensured while maintaining a high capacity for hidden information. The experimental results demonstrate that the proposed method effectively embeds and extracts hidden data, providing a secure and robust means of communication.

Table 1: Comparison of Information Hiding Capacity for MTS-based, SQT and Proposed Method with All Bit '0' and '1' Hidden

Video Sequences	Targeted Bitrate (kbps)	Information Hiding Capacity					
		MTS-based [24]		SQT [27]		Proposed Method (Selective Sub-Block Transform)	
		Bit '0'	Bit '1'	Bit '0'	Bit '1'	Bit '0'	Bit '1'
Blowing Bubbles	100	122	244	60	60	61	50
	500	578	754	60	60	527	444
	1000	726	760	60	60	1085	905
	2000	906	790	60	60	1569	1447
	3000	940	676	60	60	1815	1706
	4000	1028	770	60	60	**1929**	1831
	5000	974	794	60	60	1892	1827
People On Street	100	2	72	60	60	328	282
	500	244	1128	60	60	419	290
	1000	694	2660	60	60	860	627
	2000	1776	4286	60	60	2215	1618
	3000	2612	5106	60	60	4058	2827
	4000	3868	6718	60	60	6093	4065
	5000	5076	7654	60	60	**9188**	6186

Additionally, the algorithm's adaptability allows for seamless integration with existing VVC encoding pipelines, making it a practical solution for real-world applications. For future research, investigation on the feasibility of real-time information hiding and extraction will be conducted. This includes the implementation of proposed methods in live video streams presents unique challenges, such as minimizing latency and optimizing resource usage.

VI. Acknowledgment

We would like to extend our sincere gratitude to everyone who has supported us throughout this research. We were deeply appreciative of the assistance provided by Tunku Abdul Rahman University of Management and Technology (TAR UMT). The university's valuable resources, including its library, online databases, and laboratory facilities, were instrumental in the completion of this study.

References

[1] R. Syaripuddin, A. Razaq Ahmad, and M. Mahzan Awang, "The Use of Video in Teaching and Learning 21st Century History Education in Malaysia," Redwhite Press, Oct. 2019, pp. 182–186. doi: 10.32698/gcs.0194.

[2] F. & D. Y. & Y. A. Topsümer, "Media and Communication in the Digital Age: Changes and Dynamics," 2023. [Online]. Available: https://www.researchgate.net/publication/375116310.

[3] W. Bumia, "Emerging Technologies For 21st Century Education —Strategies For The Use Of Video Mediated Instruction," 2018.

[4] B. Ilyassova, L. S. Zarkenova, and G. K. Alshinbekova, "Information technology and its importance in modern teaching methods," 2018.

[5] I. Mambetova, "The Use of Multimedia Technologies in the Teaching Process," *Bulletin of Science and Practice*, vol. 6, no. 11, pp. 427–432, Nov. 2020, doi: 10.33619/2414-2948/60/53.

[6] O. R. Deshmukh, "Virtual Reality Next-gen Gaming Experience," *Int J Res Appl Sci Eng Technol*, vol. 10, no. 6, pp. 3442–3443, Jun. 2022, doi: 10.22214/IJRASET.2022.44662.

[7] V. B. Dzobelova, L. A. Aguzarova, A. V. Olisaeva, and E. E. Kornilova, "Digital Technologies in Education and Their

Influence on Modern Society," *Proceedings of the "New Silk Road: Business Cooperation and Prospective of Economic Development" (NSRBCPED 2019)*, Apr. 2020, doi: 10.2991/AEBMR.K.200324.030.

[8] M. A. Quainoo and T. Pasawano, "The Use of Digital Media Platforms Education and Supporting Self During The COVID," 2020.

[9] G. Wang, G. Zhou, and Z. Li, "Research on the Application of Chinese Teaching Based on Social Media Video Platforms," *International Journal of Distance Education Technologies*, vol. 20, no. 1, pp. 1–16, Mar. 2022, doi: 10.4018/IJDET.296699.

[10] R. Noor, A. Khan, and A. Sarfaraz, "High Performance and Energy Efficient Image Watermarking for Video Using a Mobile Device," *Wirel Pers Commun*, vol. 104, no. 4, pp. 1535–1551, Feb. 2019, doi: 10.1007/S11277-018-6097-3/FIGURES/8.

[11] A. Khan and A. Sarfaraz, "FFT-ETM based distortion less and high payload image steganography," *Multimed Tools Appl*, vol. 78, no. 18, pp. 25999–26022, Sep. 2019, doi: 10.1007/S11042-019-7664-7.

[12] D. R. Bull and F. Zhang, "Video coding standards and formats," *Intelligent Image and Video Compression*, pp. 435–484, Jan. 2021, doi: 10.1016/B978-0-12-820353-8.00021-9.

[13] M. Conti, E. Losiouk, and A. Visintin, "What You See is Not What You Get: a Man-in-the-Middle Attack Applied to Video Channels," *Proceedings of the ACM Symposium on Applied Computing*, pp. 1723–1726, Apr. 2022, doi: 10.1145/3477314.3507233.

[14] Y. Huang, X. Li, W. Wang, T. Jiang, and Q. Zhang, "Forgery Attack Detection in Surveillance Video Streams Using Wi-Fi Channel State Information," *IEEE Trans Wirel Commun*, vol. 21, no. 6, pp. 4340–4349, Jun. 2022, doi: 10.1109/TWC.2021.3129188.

[15] Y. Huang, S. Yang, T. Ma, Z. Liu, Z. Ding, and Q. An, "A Low-cost Active Defense Mechanism against DDoS Attacks in Streaming Media Scenarios," *Proceedings - 2022 International Conference on Networking and Network Applications, NaNA 2022*, pp. 130–135, 2022, doi: 10.1109/NANA56854.2022.00030.

[16] A. Elhadad, S. Hamad, A. Khalifa, and A. Ghareeb, "High capacity information hiding for privacy protection in digital video files," *Neural Comput Appl*, vol. 28, no. 1, pp. 91–95, Dec. 2017, doi: 10.1007/S00521-016-2323-7/FIGURES/4.

[17] D. Xu, "Data hiding in partially encrypted HEVC video," *ETRI Journal*, vol. 42, no. 3, pp. 446–458, Jun. 2020, doi: 10.4218/ETRIJ.2018-0484.

[18] Z. Li, L. Meng, X. Jiang, and Z. Li, "High Capacity HEVC Video Hiding Algorithm Based on EMD Coded PU Partition Modes," *Symmetry 2019, Vol. 11, Page 1015*, vol. 11, no. 8, p. 1015, Aug. 2019, doi: 10.3390/SYM11081015.

[19] Y. Yang, Z. Li, W. Xie, and Z. Zhang, "High capacity and multilevel information hiding algorithm based on pu partition modes for HEVC videos," *Multimed Tools Appl*, vol. 78, no. 7, pp. 8423–8446, Apr. 2019, doi: 10.1007/s11042-018-6859-7.

[20] Z. Li, X. Jiang, Y. Dong, L. Meng, and T. Sun, "An Anti-Steganalysis HEVC Video Steganography With High Performance Based on CNN and PU Partition Modes," *IEEE Trans Dependable Secure Comput*, vol. 20, no. 1, pp. 606–619, Jan. 2023, doi: 10.1109/TDSC.2022.3140899.

[21] H. Zhao, Y. Liu, Y. Wang, S. Liu, and C. Feng, "A Video Steganography Method Based on Transform Block Decision for H.265/HEVC," *IEEE Access*, vol. 9, pp. 55506–55521, 2021, doi: 10.1109/ACCESS.2021.3059654.

[22] H. Chai, Z. Li, F. Li, and Z. Zhang, "An End-to-End Video Steganography Network Based on a Coding Unit Mask," *Electronics 2022, Vol. 11, Page 1142*, vol. 11, no. 7, p. 1142, Apr. 2022, doi: 10.3390/ELECTRONICS11071142.

[23] M. Li, Z. Li, and Z. Zhang, "A VVC Video Steganography Based on Coding Units in Chroma Components with a Deep Learning Network," *Symmetry 2023, Vol. 15, Page 116*, vol. 15, no. 1, p. 116, Dec. 2022, doi: 10.3390/SYM15010116.

[24] X. Liu, K. Shi, A. Li, H. Zhang, J. Ming, and H. Fang, "A Novel Information Hiding Method for H.266/VVC Based on Selections of Luminance Transform and Chrominance Prediction Modes," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 3158–3163. doi: 10.1109/SMC52423.2021.9659244.

[25] M. Farajallah, G. Gautier, W. Hamidouche, O. Deforges, and S. El Assad, "Selective Encryption of the Versatile Video Coding Standard," *IEEE Access*, vol. 10, pp. 21821–21835, 2022, doi: 10.1109/ACCESS.2022.3149599.

[26] A. Fotovvat and K. A. Wahid, "Selective Encryption of VVC Encoded Video Streams for the Internet of Video Things," Mar. 2021, Accessed: Aug. 05, 2023. [Online]. Available: https://arxiv.org/abs/2103.14844v1.

[27] H. Joan, L. P. Tan, and Y. Tew, "Data Embedding with Dynamic Slice Quantisation in Versatile Video Coding," in *2023 IEEE International Conference on Computing, ICOCO 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 41–46. doi: 10.1109/ICOCO59262.2023.10397877.

[28] H. Joan, L. P. Tan, and Y. Tew, "Quick Response (QR) codes embedding in VVC using Quantisation Parameter Manipulation," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1697–1704. doi: 10.1109/APSIPAASC58517.2023.10317417.

GGMDDC: A GAN-Guided Multilingual Audio Deepfake Detection Dataset

Ravindrakumar M. Purohit, Arth J. Shah, and Hemant A. Patil
Speech Research Lab, DA-IICT, Gandhinagar, India
E-mail: {202321002, 202101154, hemant_patil}@daiict.ac.in

***Abstract*—Audio Deepfake Detection (ADD) is of critical concern due to its social relevance. The development of ADD solution faces the limitation of the low diversity of language resources and is restricted to a limited number of speakers. However, there is a need for a more reliable ADD system due to advancements in deepfake audio generation by the attackers. Most of the existing datasets for ADD task are limited to a language count, so their usage remains limited to restricted number of speakers. Also, in most of the existing datasets for ADD, the number of issue of some speakers' real and deepfake exists, and others are unavailable. In this study, we implemented a configuring version of Generative Adversarial Networks (GANs), namely, High Fidelity-GAN, to create dataset, GAN-Guideded Multilingual Deepfake Detection Corpus(GGMDDC). We achieved Mean Openion Score (MOS) of 4.52 for the proposed system, which proved to be better than MOS of the existing systems.**

***Index Terms* - HiFi-GAN, Mel Spectrogram, Multilingual Deepfakes, Audio Deepfake Detection.**

I. Introduction

Deepfake is the synthetic data generated, converted, or edited using the Deep Learning (DL) algorithm with the goal of making fake audio, video, or image that sounds real. In recent years, deepfake content in image, video, and audio formats has become a boom with advancement in technology. In the earlier days, it took a lot of work to generate deepfake because of the complex nature of data patterns or the limited computational resources.

Deepfake signals are generated by preserving the voice and sentence structure, in order to make them close to real speech signals. Innovation of Generative Adversarial Networks (GANs) in 2014, resulted in a solution to many problems e.g., data augmentation using generative models [1], [2]. The potential of GAN for ADD task has also been explored previously [3]. It is motivated by recent advancements in GAN, particularly, High-Fidelity Generative Adversarial Networks (HiFi-GAN), to generate deepfake audio from real utterances. HiFi-GAN (High fidelity means, closely matches the natural human voice in terms of clarity, accuracy, and realism) works in two phases. In the first phase, it predicts the low resolution intermediate representation of linguistic features in terms of the Mel spectrogram [4], and in the last phase, it creates the synthetic raw waveform from the intermediate representation of the Mel spectrogram [5]. During this process, we took sampling frequency as 22050 samples per second, all representing the 16-bits fidelity. As output, raw audio will closely resemble the qualities of human speech. We created a HiFi-GAN model to generate the dataset. At first, we trained a model using the LJ Speech dataset, in which we employed a generator (G), and discriminator (D). We trained the model with the LJ Speech audio dataset containing 13,100 samples. After 600k steps, we got an optimized model, which was further used to generate deepfake files. For inference purposes, the samples pass through the upsampling (22050 Hz from 16000 Hz) phase, where a pre-trained feature extraction module extracts Mel spectrograms, which are further passed through a generator to extract and learn the speaker's properties. Further, it will pass through a discriminator to test the quality of the generated signal.

A. Related Work

Many studies have been lately focusing on Audio Deepfake Detection (ADD) task [7]. However, attackers succeed to attack many times due to the perceptually similar creation of deepfake audio, which creates threat to audio insecurity. Many researchers have started working on ADD tasks, and many datasets have been developed to safeguard society against the adverse effects of deepfakes. However, most of the existing datasets are restricted to one language (English) speakers only, resulting in difficulty to detect deepfakes in another language. One of the most popular existing dataset for ADD task, namely, FoR (Fake or Real) is also restricted to English dialect only [8]. Other popular datasets, namely, in-the-wild (ITW) [9], Singfake [10], and ASVSpoof [11], [12] datasets are also released with English and Mandarin language speakers recordings only.

Recently released dataset, namely, WaveFake [13], contains audio samples in two languages (English and Japanese). Many a times, attackers plan to generate deepfake audio of celebrities, politicians, or businessman in native languages, resulting in a failure of good performance models trained on one language only. In highly populated countries, such as China, and India, native language-based deepfakes seems as real to locals, creating a threat to subjects' personality, security, and business. In this study, we propose a dataset (containing samples collected from 10 different languages), namely, GAN-guideded Multilingual Corpus for Deepfake Detection Corpus (GGMDDC). This dataset is aimed to boost ADD research activity, due to its property of being robust to speakers' dialect. It contains some of the most popular languages of world, such as Russian, French, Hindi, etc. resulting in its broad application perspective for ADD task.

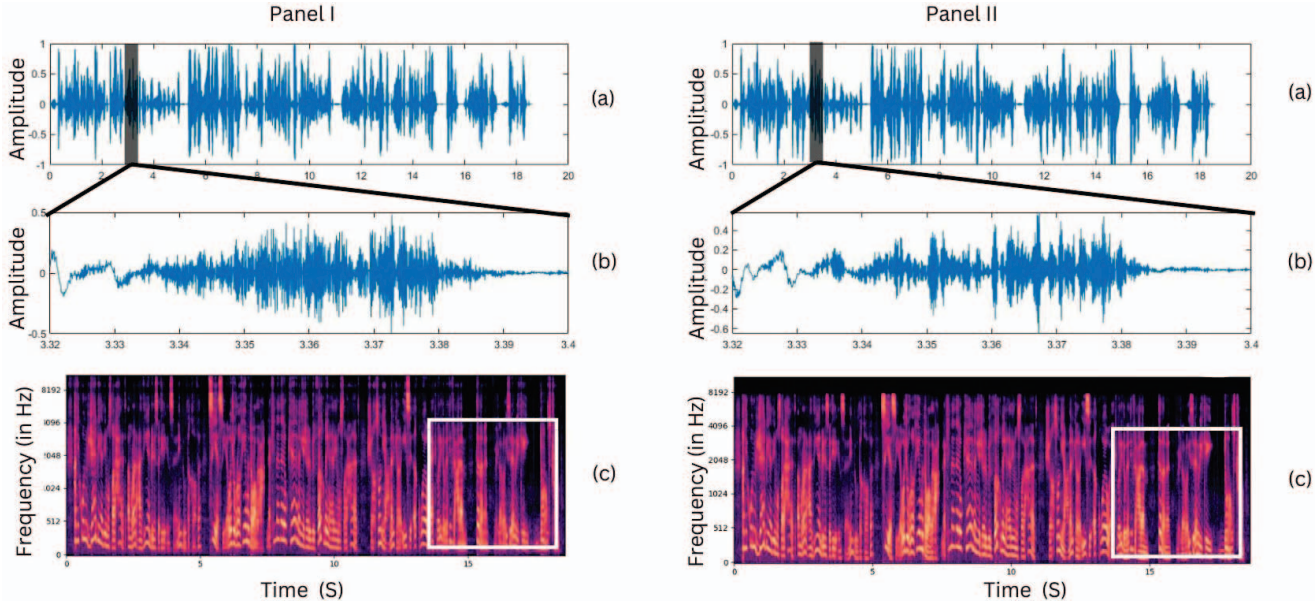


Fig. 1. Panel I (deepfake signal), and Panel II (real speech): (a), (b), and (c) shows the waveform, short speech segment, and Mel spectrogram, respectively (White boxes in Panel I (c) and Panel II (c) highlight the similarities between the deepfake and real speech samples).

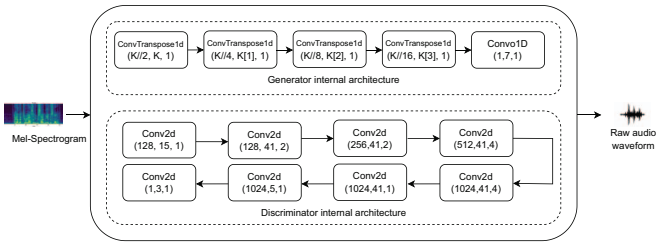


Fig. 2. Architecture of HiFi-GAN. Adapted from [6].

Fig. 1 (a) shows that mostly both signals (deepfake as well as real) are similar to each other. However, a significant difference can be noted in the amplitude variations in both waveforms. It can be observed that the deepfake signal has more high amplitude samples than a low number of high amplitude samples. Also, in Fig. 1, it can be observed that Panel I and Panel II show the variational differences in amplitude over time. In the generated speech spectrogram, there are three bands with a broader frequency range compared to Panel II, or real audio. The real audio displays the thin and well-defined bands, indirectly showing the narrower frequency range as more focused and natural speech. The distribution of the frequency and amplitude over time helps identify the deepfake audio from the spectrogram characteristics.

II. Proposed Methodology

A. Why HiFi-GANs?

In the case of deepfake detection, the dataset should contain high quality and controllable speech samples. To create GGMDDC, we employ HiFi-GAN speech synthesis architecture, to create deepfake audio samples from the real datasets. To choose the appropriate architecture, we surveyed several architectures, such as Tacotron [14], Tacotron2 [15], Deep Voice 3 [16], clariNet [17], Parallel WaveGAN [18], and many more (as mentioned in Table I). Tacotron is an end-to-end TTS model, which follows the *seq2seq* paradism to synthesis the speech [14]. Tacotron predicts the Mel spectrogram from the text and later on uses the Griffin-Lim reconstruction algorithm [19] to construct the raw waveforms. However, model struggle with unusual or unexpected text inputs, resulting in incorrect or nonsensical outputs, and achived the MOS of 3.82 out of the *5*-scale. The Tacotron2 [15] uses the recurrent *seq2seq* network with an attention mechanism to predict the Mel spectrograms, and later on vocoder transforms the Mel spectrogram to the raw waveform. Deep Voice 3 [16], follows the fully-convolutional architecture with monotonic attention mechanism, to minimise the limitation of the *seq2seq* models for waveform synthesis, again, the control over the prosody is the less as compared to the Tacotron2 and achieves the MOS of the 3.62, which was less than the first version of the Tacotron. ClariNet [17] and Parallel WaveGAN [18] are limited to single-stage speech synthesis models. Where ClariNet [17] is built on an autoregressive (AR) [18] Text-to-Wave (T2W) neural network, and Parallel WaveGAN employs a non-autoregressive (NAR) WaveNet architecture. ClariNet provides faster inference speed because of its AR structure. However, because of its fully parallel design, Parallel WaveGAN [18] trains far more quickly than ClariNet [17]. In particular, it takes just 2.8 days compared to the 12.7 days for ClariNet [17]. Despite this discrepancy, both models, i.e., ClariNet [17] [17] and parallel WaveGAN [18] achieve most of 4.15 and 4.16, respectively, and achieve comparable quality while synthesizing raw waveforms. One of the major issues in almost every existing work is Real-Time Capability (RTC), which is provided by a few models. The RTC models have low MOS compared to HiFiGAN, forcing us to employ HiFi-GAN for deepfake generation. As

shown in Table I, Transformer TTS [20], HiFiGAN [6], FreGAN [21], Fastspeech2 [22], Glow-TTS [23], and BigVGAN [24]. All models perform a significant amount of the speech synthesis quality but in terms of the other parameters, e.g., Quality-of-Speech (QoS), inference speed, real-time capability, computational cost, and multiple modalities. Recent models have significantly improved upon previous works, bringing the quality of synthesized audio closer to natural human speech. However, many models, such as Tacotron [14], lack the RTC or require significant computations for fast inference. In contrast, HiFi-GAN performs high inference speed without generating poor-quality samples [6]. In such cases, handling multi-speakers and diverse speech modalities is essential to generate robust deepfake audio, which tends to struggle for AR and NAR when a dataset containing multiple speakers or different speech styles. To balance the quality, speed, and robustness against various inputs, make the HiFiGAN a superior choice over other architectures like Tacotron [14], clariNET [17], and parallel WaveGAN [5].

Additionally, they have greatly reduced training and inference times without degrading speech synthesis performance. We initially started with the LJ Speech dataset, which includes the speech of a single female speaker for model training. However, despite its high quality recordings, the dataset lacks speaker diversity, and the result could have been more optimal. Consequently, we fine-tuned the model on the VCTK dataset (which contains recordings from 109 speakers with various accents and speaking styles) using transfer learning. This approach significantly improved the ability of the model to generate deepfake audio that closely resembles the original speech samples in the real dataset. Fig. 1 illustrates the high degree of similarity achieved between the original and synthetic samples.

The generator architecture [6] is initially built using the convolutional neural networks as shown in Fig. 2. It starts with the convolution layers (ConvTranspose1D) and weight normalization (each layer followed by a nonlinear activation function $f(\cdot)$), which helps to extract the different aspects of the Mel spectrogram from the input features by decoupling the magnitude of the weight vector from the Mel spectrogram, also it progressively increases the temporal resolutions of the signal, and do feature mapping-stretches over time according to the specific upsample rates for the high frequency details. The feature maps are refined to the final audio waveform with the single output channel at the final layer by normalizing the weights. However, the MRF (Multiple Receptive Field-fusion) [26] block is also used in generator architecture because it identifies the various lengths of hidden patterns from the Mel spectrograms. A large receptive field helps to identify the phonemes more effectively [4], [26]. Resblock1 and Resblock2 are multi-scale architectures that apply multiple convolutional operations to the different delations rated to capture *local* and *global* temporal dependencies from nonlinear transformations and residual connections. The feature maps are refined to the final audio waveform with the single output channel at the final layer by normalizing the weights.

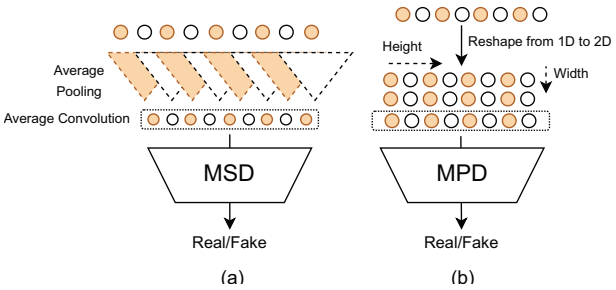


Fig. 3. Architecture of discriminator. After [6].

1) Discriminator: The discriminator D identifies the short and long-term dependencies on the raw waveform made by the generator. To achieve natural-sounding speech synthesis, it is crucial to understand the long-term dependencies in speech signal. This ensures that the generated speech closely mimics natural human speech. Studies reported in [4], [14] treat audio as a sinusoidal signal with varying periods to capture the nonlinear correlations between different phonemes.

To capture various repeating patterns, the discriminator (refer Fig. 3) is divided into two sub-discriminators: **(1) Multi-Period Discriminator (MPD), and (2) Multi-Scale Discriminator (MSD)**. D_{MPD} is a mixture of multiple sub-discriminators. It tries to capture the implicit features of the raw waveform with a different period sizes to avoid overlapping from a length, $\boldsymbol{T}$. Mathematical representation as follows in Eq. (1), where $\boldsymbol{N}$ represents the number of sub-discriminators, $\boldsymbol{D}$ denotes the contribution of each sub-discriminator to the overall discriminator's decision, i.e.,

$$D_{MPD} = \sum_{i=1}^{N} D_i. \tag{1}$$

D_{MSD} architecture was proposed first by Kumar *et al.* [26]. MSD combines the three sub-discriminators, e.g., raw audio, *2 average-pooled audio, and *4 average-pooled audio. Each of the discriminator convolutions is covered using the leaky ReLU. Discriminator uses the *weight* and *spectral* normalization [27] to do stable training, and avoid exploding gradient problems in discriminator networks.

2) Loss Functions: As the original papers of the GANs [28], [29] utilize minimax as shown in Eq. (2) and Wasserstein loss functions, operating on distance or probability distributions between generated and real data. Minimax loss establishes a competitive game for realistic sample generation. Wasserstein loss in WGANs focuses on minimizing Wasserstein distance for smoother gradients. In particular,

$$\max_{D} V(D) = E_x[\log(D(x))] + E_z[\log(1 - D(G(z)))]. \tag{2}$$

Loss function makes a balance between the generator and discriminator. WGAN refines Wasserstein loss with a gradient penalty, enhancing stability in GAN training. The loss function of both networks is next.

1) **Generator Loss Function**: Let G denote the generator, D denote the discriminator, and α be a reward/penalty

TABLE I
COMPARISON OF VARIOUS ARCHITECTURES BASED ON VARIOUS EVALUATION FACTORS

Architecture	Dataset	5-Scale MOS	QoS (MOS)	Inference Speed	RTC[$]	Multiple Modalities	Computational Cost	Limitations
Tacotron [14]	NAE	3.82±0.08*	Fair	Moderate	No	Moderate	Moderate	Alignment issues and artifacts issues generates mechanical-sounding voice.
Tacotron 2 [15]	LJ Speech	4.52±0.06*	Good	Moderate	No	Good	High	Slower inference speed and higher computational cost.
Deep Voice 3 [16]	VCTK	3.01±0.29*	Fair	Moderate	No	Good	High	Higher computational requirements.
ClariNet [17]	IES	4.15±0.25*	Good	Moderate	No	Excellent	Moderate	Training can be complex and resource-intensive.
Parallel WaveGAN [18]	Custom Dataset	4.16±0.09*	Good	Very Fast	Yes	Good	Low	May not capture fine speech details as well as others.
Transformer TTS [20]	US English female	4.44±0.05*	Good	Moderate	No	Good	Moderate	Requires significant training data.
HiFi-GAN [25]	**LJSpeech**	**4.36±0.07***	**Good**	**1,186×**	**Yes**	**Excellent**	**Moderate**	**Can be resource-intensive in some cases.**
Fre-GAN [21]	LJ Speech	4.25±0.04	Good	Moderate	No	Good	Moderate	Requires careful tuning for best results.
FastSpeech 2 [22]	LJ Speech	3.83±0.08	Fair	Very Fast	Yes	Good	Low	Lower expressiveness compared to some models.
Glow-TTS [23]	LJ Speech	4.01±0.08	Good	Moderate	No	Good	Moderate	More complex training process.
BigVGAN [24]	LibriTTS	4.11±0.09	Good	Fast	Yes	Good	Low	May struggle with very diverse speakers.

* shows confidence interval is 95%. [$] RTC - Real-Time Capability, NAE - North Americal English, QoS = Quality of Speech.

factor. The generator loss (L_G) can be formulated, in simple words as:

$$G_L = [\alpha \cdot \text{reward} + (1-\alpha) \cdot \text{penalty}]. \tag{3}$$

In the other words,

$$G_L = \theta_g \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(z^{(i)}))). \tag{4}$$

2) **Discriminator Loss Function**: Let G denote the generator, D denote the discriminator, and α be a reward/penalty factor. The discriminator loss (L_D) can be formulated as follows, in simple words:

$$D_L = \frac{1}{m} \sum_{i=1}^{m} \left[\log D(x^{(i)}) + \log(1 - D(G(z^{(i)}))) \right]. \tag{5}$$

It penalizes itself for misclassifying a real instance as fake or a fake instance (created by the generator) as real by maximizing the Eq. (5).

3) Training Paradigm: The training paradigm of the HiFi-GAN follows the Goodfellow *et al.* [28]. The primary goal of GAN training is to find an equilibrium, where the generator produces realistic data that fool the discriminator. As per Eq. (6):

$$\begin{aligned} \min_G \max_D V(D,G) = E_{x \sim p_{data}(x)}[\log D(x)] \\ + E_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \end{aligned} \tag{6}$$

The training of the HiFi-GAN is being used for speech-related tasks, e.g., ASR, Text-to-Speech (TTS) synthesis [26]. Training is set up using the multiple loss functions described earlier.

III. DATASET FORMATION

This Section proposes the presents dataset structure and its design details. Due to the limited language resources, the authors were unable to collect data manually as real samples were collected for our dataset. Alternatively, in this study, we propose a dataset in which we generated 40,000 deepfake samples of real utterances that were collected from the VoxLingua107 dataset [30], which is one of the most popular largest open source multilingual dataset for Spoken Language Identification (SLID) task. VoxLingua107 was formed by recording utterances from 107 different languages and data from 6628 hours. Limited to storage resources, authors could not create a dataset for more than 10 languages, namely, Russian, French, Arabic, Spanish, Vietnamese, Mandarin Chinese, English, Hindi, Portuguese, and Sanskrit. Around 11.35 hours of data was selected from each language having based on time duration statistics. Comprising 10 languages in the GGMDDC, it is also robust to speaker's dialects. To our best knowledge and belief, this is the first study of its kind that employs and proposes the corpus, which contains both the real and the corresponding fake utterances of each speaker w.r.t. text material used for the recordings. The total number of utterances in the proposed dataset is 80,000 (40,000 real and 40,000 fake), making it one of the largest datasets among currently available open source datasets in the literature. Dataset statistics and demo is publicly available at [1].

A. Real Data

First, we collected all the samples available from the VoxLingua107 dataset (which is an open source freely available dataset) [30]. We labeled them into 5 classes based on the audio duration of particular samples, namely, A (0-5 seconds),

[1]For more details, https://iamshreeji-copy1.github.io/submission_website/

B (5-10 seconds), C (10-15 seconds), D (15-20 seconds), and E ($>$ 20 seconds). After that, we selected 1,000 random samples (except in one (Sanskrit) class) from each class collectively to form a dataset of total of 111.16 hours (40,000 samples) of real data. We eliminated the issue due to audio sample size dependencies by selecting the variable length audio. In order to generalize sampling rate to *16 kHz*, all audios of VoxLingua107 were resampled to 16 kHz before generating deepfake files from it. The resampling process was carried out in order to generalize the dataset.

B. Fake Data

We use the model based on HiFi-GANs to generate deepfakes from the real signal (described in subsection II-A). We employed to process the real audio and generate the deepfake audio of the same speaker with the same utterance spoken in the original sample. HiFi-GAN generated deepfake illustrate almost similar properties as the real signals. Audio-based comparison of generated deepfakes and real, as discussed in sub-Section I-A, illustrates the difficulties in distinguishing between fake *vs.* real. Due to their perfect creation (i.e., high *perceptual* similarity), these generated deepfakes are extremely difficult to distinguish between human ears. As the dataset is generated by sophisticated machine learning/DL methods, it also aims to fool the ADD system and humans.

C. Results

Table II compares the performance of various datasets using several metrics, namely *Perceptual Evaluation of Speech Quality* (PESQ), *Short-Time Objective Intelligibility* (STOI), *Fréchet Distance, Mean Opinion Score* (MOS), *Subjective Mean Opinion Score* (SMOS), and *Mel Cepstral Distortion* (MCD). The proposed dataset demonstrates superior performance across these metrics. Specifically, it achieves a PESQ score of 1.0213, which is competitive with the other datasets. The STOI score of 0.1224 is notably higher, indicating better intelligibility. Additionally, the proposed dataset records the lowest Frechet distance of 220.22, suggesting it generates deepfakes that are perceptually closest to the real audio. The Mean Opinion Score (MOS) of 4.52, and Subjective Mean Opinion Score (SMOS) of 4.39 are the highest among all datasets, reflecting higher quality and naturalness of the generated audio. Lastly, the Mel Cepstral Distortion (MCD) score of 23.638 is the lowest, indicating better spectral quality. These results collectively demonstrate that the proposed dataset, when used with HiFi-GAN, is capable of generating more convincing and high quality deepfakes compared to the existing datasets.

TABLE II
COMPARISON W.R.T. VARIOUS EXISTING DATASETS

Dataset	PESQ	STOI	Frechet Distance	MCD	MOS	SMOS
FoR [8]	1.02	-0.01	228.74	4.12	4.02	24.95
ITW [9]	1.06	0.22	287.66	34.20	3.64	3.46
Singfake [10]	1.32	0.07	324.31	34.22	4.36	4.38
Wavfake [13]	**2.91**	0.95	276.3	34.82	4.4	4.11
Proposed	1.02	**0.12**	**220.22**	**23.63**	**4.52**	**4.39**

IV. SUMMARY AND CONCLUSIONS

In this study, we proposed GGMDDC dataset, which is generated using recently proposed, HiFi-GAN. Real data in GGMDDC dataset was acquired from VoxLingua107 dataset, whose deepfakes were generated GGMDDC. Total of 40,000 files collectively form GGMDDC dataset. A proposed dataset is a multilingual dataset, which consists of 10 different languages, resulting into a wide application of dataset. We also proposed a few experiments on GGMDDC dataset, resulting into a comparable accuracy. Future works involve more detailed experimentation and analysis on proposed dataset using advanced DL methods. One key limitation of this study would be the GAN architecture used in this work use pretrained models in English to generate multilingual voices in 10 different languages. Limitations of this study include training GANs on the LJ Speech and VCTK dataset, i.e., a single language dataset. Capturing the critical phoneme patterns of different languages is challenging when using a dataset from just one language. In the future, We aim to address this limitation by training the model on diverse languages and large-scale datasets to overcome this issue. However, the model's effectiveness will increase when trained on the multilingual dataset and provides the generalized diverse and dialects linguistic characteristics to identify the deepfake. The rapid involvement of the deepfake generation is already a challenge for deepfake detection. Implementing the robust multilingual dataset is computationally intensive due to the detection in diverse scenarios in real-time, which makes ongoing research on ADD more complicated and challenging, which remains an open research challenge associated with our detection process.

REFERENCES

[1] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017, {Last Accessed Date : 1st August, 2024}.

[2] D. Baby and S. Verhulst, "SERGAN: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, 106–110, Brighton, United Kingdom.

[3] T. P. Doan, K. Hong, and S. Jung, "GAN discriminator based audio deepfake detection," in *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, 2023, Melbourne, Australia, pp. 29–32.

[4] A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016, {Last Accessed Date : 1st August, 2024}.

[5] A. Oord, Y. Li, I. Babuschkin, *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International Conference on Machine Learning (ICML)*, PMLR, 2018, Stockholm, Sweden, pp. 3918–3926.

[6] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 17 022–17 033, 2020, {Last Accessed Date : 1st August, 2024}.

[7] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection," in *Proc. INTERSPEECH*, vol. 2023, 2023, 2808–2812, Dublin, Ireland.

[8] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2019, Timisoara, Romania, pp. 1–10.

[9] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *arXiv preprint arXiv:2203.16263*, 2022, {Last Accessed Date : 9th March, 2024}.

[10] Y. Zang, Y. Zhang, M. Heydari, and Z. Duan, "Singfake: Singing voice deepfake detection," *arXiv preprint arXiv:2309.07525*, 2023 {Last Accessed Date : 9th March, 2024}.

[11] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101–114, 2020.

[12] A. Chintha, B. Thai, S. J. Sohrawardi, *et al.*, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1024–1037, 2020.

[13] J. Frank and L. Schönherr, "Wavefake: A dataset to facilitate audio deepfake detection," *arXiv preprint arXiv:2111.02813*, 2021 {Last Accessed Date : 9th March, 2024}.

[14] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017 {Last Accessed Date : 1st August, 2024}.

[15] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, Calgary, Canada, pp. 4779–4783.

[16] W. Ping, K. Peng, A. Gibiansky, *et al.*, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017 {Last Accessed Date : 1st August, 2024}.

[17] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018 {Last Accessed Date : 1st August, 2024}.

[18] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, Virtual Barcelona, pp. 6199–6203.

[19] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[20] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, {Last Accessed Date : 1st August, 2024}, pp. 6706–6713.

[21] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, "Fre-GAN: Adversarial frequency-consistent audio synthesis," *arXiv preprint arXiv:2106.02297*, 2021, {Last Accessed Date : 1st August, 2024}.

[22] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020 {Last Accessed Date : 1st August, 2024}.

[23] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 8067–8077, 2020, Vancouver, Canada.

[24] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022 {Last Accessed Date : 1st August, 2024}.

[25] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *arXiv preprint arXiv:2006.05694*, 2020, {Last Accessed Date : 1st August, 2024}.

[26] K. Kumar, R. Kumar, T. De Boissiere, *et al.*, "MelGAN: Generative adversarial networks for conditional waveform synthesis," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 14910-14921, 2019, Vancouver, Canada.

[27] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018 {Last Accessed Date : 1st August, 2024}.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Neural Information Processing Systems (NIPS)*, vol. 27, pp. 2672-2680, 2014, Montreal, Canada.

[29] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016 {Last Accessed Date : 1st August, 2024}.

[30] J. Valk and T. Alumäe, "Voxlingua107: A dataset for spoken language recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, Shenzhen, China, pp. 652–658.

MGVul: a Multi-Granularity Detection Framework for Software Vulnerability

Xiangyu Zhao*, Yanjun Li*, Zhengpeng Zha*†, Zhenhua Ling‡
* Institute of Advanced Technology, University of Science and Technology of China
Email: {xyzhao23, ballinli}@mail.ustc.edu.cn,
† Corresponding author. Email: zhazp@ustc.edu.cn,
‡ School of Information Science and Technology, University of Science and Technology of China
E-mail: zhling@ustc.edu.cn

***Abstract*—Source code vulnerability detection is a critical issue in software security. Existing detection methods primarily focus on the function-granularity, neglecting inter-function call information and lacking the ability to detect inter-function vulnerability. The isolation of functions is one of the reasons that limits the performance of deep learning models. In this work, we propose MGVul: a Multi-granularity detection framework for software vulnerability. This framework treats a project as a whole graph, files as subgraphs, functions as nodes, and function call relationships as edges. By using a multi-expert mixture graph neural network to learn inter-function call information, it not only improves the model's detection performance at the function-granularity but also is capable of handling detection tasks at the file-granularity and project-granularity. Preliminary evaluation shows that introducing inter-function call relationships bring an average F1-score improvement of 7.99% for function-level detection. Meanwhile, our framework also shows promising results at both the file-granularity and the project-granularity. The source code will be released upon paper acceptance.**

I. Introduction

The development of the Internet industry has led to the widespread influence of open-source software. According to information from CVE (Common Vulnerabilities and Exposures) [1], the number of vulnerabilities has rapidly increased in recent years [2]. Research on source code vulnerability detection is also growing steadily. With the development of artificial intelligence technology, detection methods have shifted from expert-dominated static and dynamic analysis to machine learning and deep learning models [3].

In recent years, most research based on deep learning, using either pre-trained language models or graph neural networks, has focused on individual functions[4]–[7]. Specifically, the models take a single function as input and output whether the function contains a vulnerability. Pre-trained language models treat functions as text and perform classification through steps such as tokenization and embedding [6], [8]. Graph neural networks detect vulnerabilities by learning from the Abstract Syntax Tree (AST) and other representations parsed from a single function [9], [10]. Despite significant progress in function-granularity detection in recent years, there remains an insurmountable barrier [7]. We argue that the problem lies in the absence of some critical information: the function call relationships are not provided to the model for learning and the model cannot observe the internal conditions of related functions. Since vulnerabilities often occur during function calls [11]–[13], it is reasonable to view a project as a whole rather than as multiple individual entities.

In this paper, we propose MGVul, which enhances the model's detection capability at the function-granularity by using function call graphs. Additionally, by learning the graph structure from different levels (whole graph, subgraphs, and nodes), we can naturally accomplish vulnerability detection tasks at multiple granularities (project, file, and function) by aggregating information at the corresponding level. Firstly, we use code parsing tools to analyze software projects and capture the complex call relationships between multiple functions in multiple files. Then, we encode functions as fixed-length vectors using pre-trained language models, treating them as function features. Moreover, we utilize a graph neural network (GNN) with a multi-expert mixture model to learn the function call graph from different perspectives. Furthermore, we introduce a pairwise training method to deepen the model's understanding of vulnerabilities in the training parse. Finally, we treat the project as a whole graph, files as subgraphs, and functions as nodes, conducting classification tasks at all these granularities for vulnerability detection.

Experimental results show that MGVul can effectively improve function-granularity detection performance compared to baseline models. Additionally, our framework is also competent for file-granularity and project-granularity detection.

In summary, the contributions of this work are as follows:

- We introduce function call information into function-level vulnerability detection tasks by using GNN to learn the function call graph.
- By aggregating information at the corresponding granularity, our framework is capable of handling detection tasks at the file-granularity and project-granularity.
- Through preliminary evaluation, incorporating function call information improves the model's function-granularity detection performance, with an average F1-score improvement of 7.99%. For the other two granularity levels of detection, MGVul also demonstrates encouraging results.

II. PRELIMINARY

This section introduces intra-function vulnerability and inter-function vulnerability with their examples. Additionally, we provide the definition of the multi-granularity source code vulnerability detection.

A. Intra-function Vulnerability and Inter-function Vulnerability

Intra-function vulnerability refers to a vulnerability that occurs within a single function. Figure 1(a) illustrates an example (CVE-2014-1266) [14]. Due to the extra line *"goto fail;"*, the signature verification always fails. Another scenario is when vulnerabilities arise during a function call, even though both the caller and callee function are flawless. Figure 1(b) provides another example (CVE-2014-3513) [15]. Both the *ssl3_read_n()* and *ssl3_get_record()* functions are benign on their own. However, when *ssl3_get_record()* calls *ssl3_read_n()*, if the length field of the read data packet is incorrect, *ssl3_read_n()* will enter an infinite loop, resulting in a Denial of Service (DoS) attack.

```
static OSStatus SSLVerifySignedServerKeyExchange(
    SSLContext *ctx, bool isRsa, SSLBuffer signedParams,
    uint8_t *signature, UInt16 signatureLen)
{
    OSStatus err;
    ...
    if ((err =
    SSLHashSHA1.update(&hashCtx, &serverRandom)) != 0)
        goto fail;
    if ((err =
    SSLHashSHA1.update(&hashCtx, &signedParams)) != 0)
        goto fail;
        /* MISSING BRACE - leads to always failing */
        goto fail;
    ...
}
```

(a) An example of the intra-function vulnerability.

```
int ssl3_read_n(SSL *s, int n, int max, int extend) {
    ...
    if (s->packet_length < (unsigned int)n) {
        ret = ssl3_read_n(
        s, n-s->packet_length, max-s->packet_length, 1);
    }
    ...
    return ret;
}

int ssl3_get_record(SSL *s) {
    int n;
    ...
    n = ssl3_read_n(
    s, SSL3_RT_HEADER_LENGTH, SSL3_RT_HEADER_LENGTH, 0);
    ...
}
```

(b) An example of the inter-function vulnerability.

Fig. 1: Examples of different categories of function vulnerabilities, the blue-highlighted lines cause the vulnerability.

B. Multi-granularity Source Code Vulnerability Detection

Given a project $P = \{F_1, F_2, \ldots, F_T\}$, where F_t is a file belonging to P and T is the number of files. Each file F_t contains a set of functions $\{f_t^1, f_t^2, \ldots, f_t^{N_t}\}$, where f_t^n is a function belonging to F_t and N_t is the number of functions of F_t. We aim to learn a mapping to predict whether the project P, each file F_t, and each function f_t^n contain vulnerabilities. Each of these predictions is a binary classification problem.

III. METHODOLOGY

A. Overview

We use a code parsing tool to transform the project P into a function call graph $G = \{SG_1, SG_2, \ldots, SG_T\}$, where each subgraph SG_t corresponds to a file F_t in P. Each subgraph $SG_t = \{ND_t^1, ND_t^2, \ldots, ND_t^{N_t}\}$ consists of nodes ND_t^n that correspond to a function f_t^n. The edges in G represent the function call relationships. We employ a GNN to learn from G and perform classification tasks at three granularities: the whole graph level, the subgraph level and the node level. The framework of our approach is illustrated in Figure 2. The framework consists of three main components: function call graph generation, input representation for code, MGVul training and prediction.

B. Function Call Graph Generation

A clear function call relationships is crucial for vulnerability detection. We chose Cflow [16], a widely used and free parsing software. It can resolve not only intra-file call relationships, but also inter-file function calls. By parsing the files to be detected in a project, we obtained the function call graph, where each function represents a node, and the nodes belonging to the same file form a subgraph.

C. Input Representation

The goal of this phase is to convert functions into compact, uniform-length feature vectors that preserve semantic and syntactic information. This uniform length is crucial because the feature vectors are used as node features in the graph.

To achieve this, we utilize pre-trained language models for code representation. Trained on large code corpora, these models can effectively capture the intricate semantic and syntactic details of the source code. The code is input into the model, which outputs a sequence of contextual embeddings. Such embeddings are then converted to code features.

D. MGVul Training and Prediction

The goal of this phase is to leverage the aggregation capabilities of GNN to learn from functions with call relationships. Since function calls can have multiple layers and long-range dependencies, we use Gated Graph Neural Network (GGNN) [17] to effectively learn and integrate the embeddings due to their excellent depth perception.

Our multi-granularity detection requires learning features from multiple levels, a single network may not effectively meet these requirements. We need to aggregate information from different perspectives (granularities) to suit different focuses. Therefore, we introduce the multi-expert mixture model[18]. Specifically, each f_t^n is transformed into $\mathbf{e}_t^n$ after encoding.

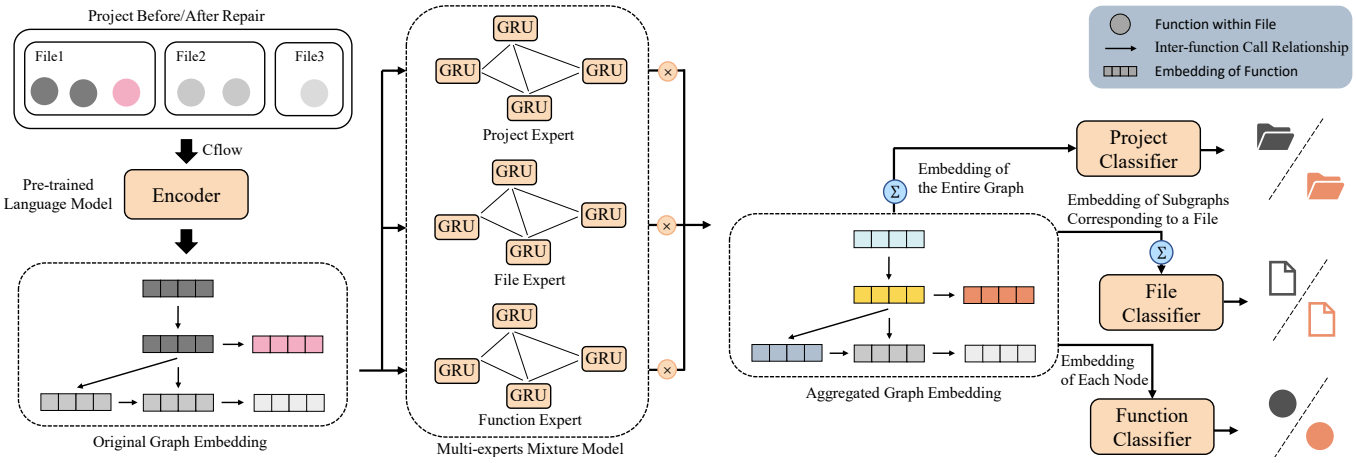


Fig. 2: Overview of the MGVul detection framework.

These embeddings $\mathbf{e}_t^n$ are then processed through three independent expert networks (GGNN) for node communication and feature integration, generating $(\mathbf{e}_t^n)_1, (\mathbf{e}_t^n)_2, (\mathbf{e}3_t^n)_3$, the final representation of the node $(\mathbf{e}_t^n)'$ can be formulated as:

$$(\mathbf{e}_t^n)' = \sum_{i=1}^{3} g(x)_i (\mathbf{e}_t^n)_i. \tag{1}$$

Here, $(\mathbf{e}_t^n)_i, i = 1, 2, 3$ are 3 expert GGNNs and g represents a gating network that ensembles the results from all experts. The term $\sum_{i=1}^{3} g(x)_i = 1$, and $g(x)_i$, the i-th logit of the output of $g(x)$, indicates the weight for representation $(\mathbf{e}_t^n)_i$.

After obtaining the final representation of the nodes, we generate the features of the subgraph $(\mathbf{e}_t)'$ based on the nodes' affiliation. Specifically, we compute $(\mathbf{e}_t)'$ as the average of $(\mathbf{e}_t^n)'$:

$$(\mathbf{e}_t)' = \frac{1}{n} \sum_{n=1}^{N} (\mathbf{e}_t^n)'. \tag{2}$$

The feature of the entire graph $(\mathbf{e}_P)'$ is generated by averaging all the node features. Specifically, we compute $(e_P)'$ as follows:

$$(\mathbf{e}_P)' = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{N_t} \sum_{n=1}^{N_t} (\mathbf{e}_t^n)' \right). \tag{3}$$

After obtaining the representations at each granularity, we complete our task by adding classification layers separately for each granularity. During training, we compute the weighted sum of the loss functions for the three granularities.

Since many of the vulnerability detection datasets are built by crawling historical patches [12], [19], the advantage of these approaches is that they provide code changes before and after vulnerabilities are fixed, and we hope that the model can observe such subtle changes in order to enhance the model's understanding of the vulnerabilities. To this end, we propose a pairwise training strategy. Specifically, we train the model by packing pairs of data in a batch so that the model learns not only to detect anomalies in the vulnerability code but also to understand the logic of fixing the code. $Loss_{all}$ is the sum of $Loss_{before}$ and $Loss_{after}$. The specific computation can be formulated as:

$$Loss_{all} = Loss_{before} + Loss_{after} \tag{4}$$

$$Loss_{stage} = \alpha * Loss_{graph} + \beta * Loss_{subgraph} + \gamma * Loss_{node}, \tag{5}$$

where α, β, and γ are hyperparameters indicating the contribution of different granularities. The *stage* can be either *before* or *after*. During the inference phase, the input data consists of individual graphs rather than pairs, so the pairwise training strateg is only used in the training phase.

IV. EVALUATION

- **RQ1:** Can learning inter-function call relationships help improve the model's function-granularity vulnerability detection capability?
- **RQ2:** How effective are the multi-expert mixture model and the pairwise training strategy in our proposed method on multi-granularity detection?

A. Dataset Details

Most current datasets are still at the function-granularity [8], [20], where individual functions and their labels are used for training. To the best of our knowledge, ReposVul [12] is the first repository-level vulnerability dataset. It contains a large number of real security incidents, meeting our requirements for call parsing. We believe that this form of data organisation is more reasonable for the vulnerability detection task. We selected the C/C++ portion of ReposVul for our experiments.

We treated each record in ReposVul, consisting of files related to C/C++ before and after fixing, as a project pair. After parsing, each file in these projects is regarded as a subgraph, and each function within these files is regarded as a node. By contacting the authors, we determined the organisation of the dataset with the meaning of the relevant targets. We excluded records with files marked as -1, indicating discrepancies between large language models and static code analysis tools. Similarly, we excluded records where Cflow

parsing was unsuccessful. The target of a project depends on the files it includes: if it contains any file with a target of 1, the project's target is 1 (indicating the presence of one or more vulnerabilities), otherwise, it is 0 (indicating no vulnerabilities). After processing, we obtained 2,220 graphs (project), containing 3,933 subgraphs (file), and a total of 132,088 nodes (function). We split the dataset into train, validation, and test sets in an 8:1:1 ratio.

B. Experimental Settings and Metrics

For the pre-trained language model, we used the embedding of the special token ([CLS]) as the input representation of the function, which served as the node feature for the GGNN, with a dimension of 768. The hyperparameters α, β, and γ were set to 50, 10 and 1. We used a two-layer MLP for the classification layers. The loss function was the Cross-entropy loss, and the optimizer was Adam with a learning rate of 2×10^{-5}. The batch size for training was set to 32. All experiments were conducted on an NVIDIA RTX 3090 GPU.

For evaluation metrics, we used Precision, Recall, and F1-score. Precision is the proportion of correctly predicted positive samples among all predicted positive samples. Recall is the proportion of correctly identified positive samples among all actual positive samples. The F1-score is the harmonic mean of Precision and Recall.

C. Effectiveness of Introducing Inter-function Call Information at Function Granularity (RQ1)

To the best of our knowledge, vulnerability detection at the function granularity is still the mainstream nowadays. To address RQ1, we validate the effectiveness by comparing to baseline models. For fairness, our encoder is set to be the same as the baseline. Specifically, for the baseline models, we use the encoder to convert functions into vectors and then directly apply a classifier. For the MGVul framework, we use the same encoder to convert functions into vectors, then use GGNN for feature communication and aggregation, and finally apply the classification layer on the node features. We selected four pre-trained code models: CodeBERT [21], GraphCodeBERT [22], PDBERT [5] and UniXcoder [23]. These models are widely used as encoders in vulnerability detection tasks.

Table 1 presents our experiment results. We observed that all four baseline models showed significant improvements in function-granularity detection performance after incorporating inter-function call information. The UniXcoder model exhibited the most substantial improvement, with an F1-score increase of 12.80%. Compared to recall, the precision rate was generally lower, indicating a higher likelihood of false positives in the current dataset. Surprisingly, CodeBERT achieved the highest recall. One possible reason is that the structural information introduced during the pre-training phase may have misled the model.

TABLE I: Evaluation results for the effectiveness of MGVul.

Method	Precision	Recall	F1-score
CodeBERT	3.06	57.50	5.81
GraphCodeBERT	5.17	30.23	8.83
PDBERT	4.39	32.50	7.74
UniXcoder	10.62	36.25	16.43
MGVul			
+CodeBERT	4.79	**61.25**	9.74
+GraphCodeBERT	6.35	52.33	11.32
+PDBERT	12.63	54.00	20.47
+UniXcoder	**20.10**	53.55	**29.23**

TABLE II: Evaluation results of ablation studies on the components of MGVul.

Method	Node_F1	Subgraph_F1	Graph_F1
MGVul	**29.23**	**31.25**	**33.33**
w/o PTS	28.79	29.73	30.30
w/o MeMM	28.77	31.15	29.09
w/o PTS & MeMM	26.64	29.73	29.19

Observation 1. Learning inter-function call relationships using GGNN improves the model's function-granularity vulnerability detection capability.

D. Effectiveness of MGVul and the ablation studies of its components (RQ2)

MGVul can perform detection tasks not only at the function-granularity but also at the file and project granularity. To answer RQ2, we use UniXcoder as the encoder to validate the detection capability of the MGVul framework and verify the effectiveness of the multi-expert mixture model (MeMM) and the pairwise training strategy (PTS). For the case without MEMM, we use only a single GGNN. For the case without PTS, we use disrupted data.

Table 2 presents our experiment results. We observed that both MeMM and PTS can improve the model's detection capability at three function granularities. Specifically, MeMM focuses on enhancing detection at the project-granularity, while PTS focuses on improving file-granularity detection. When both methods are used together, the model achieves the best performance. The F1-score increased by 2.59%, 1.52%, and 4.14%, respectively.

Observation 2. The MGVul framework shows promising results across all three detection granularities. And MeMM and PTS are useful for our framework.

V. RELATED WORK

In recent work, learning-based detection methods have gradually become mainstream [24]–[26]. Among these methods, most operate at the function level. VulCNN [27] learns vulnerability features by converting functions into images. MVulD

[28] proposes a multimodal function-level vulnerability detection method. Meanwhile, some studies operate at other granularities. Vuldeepecker [29] uses code gadgets to represent programs for vulnerability detection. Ghaffarian et al. [30]provide a detection method at the file level using parsed graphs and graph neural networks. The work at different granularity levels lacks a unified framework, which our work attempts to address. Our work designs the vulnerability detection task as a unified framework by learning function call relationships through graph neural networks and code semantics through pre-trained models.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose the MGVul framework, which enhances detection performance at the function-granularity and shows promising results at the file and project granularities. Our framework combines pre-trained language models and GGNN, using a multi-expert mixture model and pairwise training strategy. Evaluations show our method outperforms baseline models. However, precision is notably lower than recall, possibly due to the imbalance in the number of vulnerable and non-vulnerable classes. In the future, we plan to adopt data augmentation methods to reduce class imbalance. At the same time, the introduction of graph neural networks may lead to efficiency problems, which is the focus of the future research.

REFERENCES

[1] MITRE Corporation, *Common Vulnerabilities and Exposures*, https://cve.mitre.org/, Accessed: 2024-06-09, 2024.

[2] H. Hanif, M. H. N. M. Nasir, M. F. Ab Razak, A. Firdaus, and N. B. Anuar, "The rise of software vulnerability: Taxonomy of software vulnerabilities detection and machine learning approaches," *Journal of Network and Computer Applications*, vol. 179, p. 103 009, 2021.

[3] B. Wu and F. Zou, "Code vulnerability detection based on deep sequence and graph models: A survey," *Security and Communication Networks*, vol. 2022, no. 1, p. 1 176 898, 2022.

[4] Y. Zhou, S. Liu, J. Siow, X. Du, and Y. Liu, "Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[5] Z. Liu, Z. Tang, J. Zhang, X. Xia, and X. Yang, "Pre-training by predicting program dependencies for vulnerability analysis tasks," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.

[6] H. Hanif and S. Maffeis, "Vulberta: Simplified source code pre-training for vulnerability detection," in *2022 International joint conference on neural networks (IJCNN)*, IEEE, 2022, pp. 1–8.

[7] S. Chakraborty, R. Krishna, Y. Ding, and B. Ray, "Deep learning based vulnerability detection: Are we there yet?" *IEEE Transactions on Software Engineering*, vol. 48, no. 9, pp. 3280–3296, 2021.

[8] Y. Chen, Z. Ding, L. Alowain, X. Chen, and D. Wagner, "Diversevul: A new vulnerable source code dataset for deep learning based vulnerability detection," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, 2023, pp. 654–668.

[9] X. Wang, Y. Wang, Y. Wan, *et al.*, *Code-mvp: Learning to represent source code from multiple views with contrastive pre-training. corr abs/2205.02029 (2022)*, 2022.

[10] X. Duan, J. Wu, S. Ji, *et al.*, "Vulsniper: Focus your attention to shoot fine-grained vulnerabilities.," in *IJCAI*, 2019, pp. 4665–4671.

[11] Z. Li, N. Wang, D. Zou, *et al.*, "On the effectiveness of function-level vulnerability detectors for inter-procedural vulnerabilities," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–12.

[12] X. Wang, R. Hu, C. Gao, X.-C. Wen, Y. Chen, and Q. Liao, "Reposvul: A repository-level high-quality vulnerability dataset," in *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings*, 2024, pp. 472–483.

[13] J. Sun, J. Chen, Z. Xing, Q. Lu, X. Xu, and L. Zhu, "Where is it? tracing the vulnerability-relevant files from vulnerability reports," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.

[14] Apple, *Improper Certificate Validation*, https://nvd.nist.gov/vuln/detail/CVE-2014-1266, 2024.

[15] OpenSSL, *Improper Input Validation*, https://nvd.nist.gov/vuln/detail/CVE-2014-3513, 2023.

[16] S. Poznyakoff, *Gnu cflow*, https://www.gnu.org/software/cflow/, 2005.

[17] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," *arXiv preprint arXiv:1511.05493*, 2015.

[18] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1930–1939.

[19] G. Bhandari, A. Naseer, and L. Moonen, "Cvefixes: Automated collection of vulnerabilities and their fixes from open-source software," in *Proceedings of the 17th International Conference on Predictive Models and Data Analytics in Software Engineering*, 2021, pp. 30–39.

[20] J. Fan, Y. Li, S. Wang, and T. N. Nguyen, "Ac/c++ code vulnerability dataset with code changes and cve summaries," in *Proceedings of the 17th International Conference on Mining Software Repositories*, 2020, pp. 508–512.

[21] Z. Feng, D. Guo, D. Tang, *et al.*, "Codebert: A pre-trained model for programming and natural languages," *arXiv preprint arXiv:2002.08155*, 2020.

[22] D. Guo, S. Ren, S. Lu, *et al.*, "Graphcodebert: Pre-training code representations with data flow," *arXiv preprint arXiv:2009.08366*, 2020.

[23] D. Guo, S. Lu, N. Duan, Y. Wang, M. Zhou, and J. Yin, "Unixcoder: Unified cross-modal pre-training for code representation," *arXiv preprint arXiv:2203.03850*, 2022.

[24] X.-C. Wen, X. Wang, C. Gao, S. Wang, Y. Liu, and Z. Gu, "When less is enough: Positive and unlabeled learning model for vulnerability detection," in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, IEEE, 2023, pp. 345–357.

[25] S. Wang, T. Liu, and L. Tan, "Automatically learning semantic features for defect prediction," in *Proceedings of the 38th International Conference on Software Engineering*, 2016, pp. 297–308.

[26] Z. Li, D. Zou, S. Xu, H. Jin, Y. Zhu, and Z. Chen, "Sysevr: A framework for using deep learning to detect software vulnerabilities," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 4, pp. 2244–2258, 2021.

[27] Y. Wu, D. Zou, S. Dou, W. Yang, D. Xu, and H. Jin, "Vulcnn: An image-inspired scalable vulnerability detection system," in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 2365–2376.

[28] C. Ni, X. Guo, Y. Zhu, X. Xu, and X. Yang, "Function-level vulnerability detection through fusing multi-modal knowledge," in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, IEEE, 2023, pp. 1911–1918.

[29] Z. Li, D. Zou, S. Xu, *et al.*, "Vuldeepecker: A deep learning-based system for vulnerability detection," *arXiv preprint arXiv:1801.01681*, 2018.

[30] S. M. Ghaffarian and H. R. Shahriari, "Neural software vulnerability analysis using rich intermediate graph representations of programs," *Information Sciences*, vol. 553, pp. 189–207, 2021.

Enhancing Security Using Random Binary Weights in Privacy-Preserving Federated Learning

Hiroto Sawada*, Shoko Imaizumi†, and Hitoshi Kiya‡
* Graduate School of Science and Engineering, Chiba University, Chiba, Japan E-mail: 24wm3202@student.gs.chiba-u.jp
† Graduate School of Engineering, Chiba University, Chiba, Japan E-mail: imaizumi@chiba-u.jp
‡ Faculty of System Design, Tokyo Metropolitan University, Tokyo, Japan E-mail: kiya@tmu.ac.jp

***Abstract*—In this paper, we propose a novel method for enhancing security in privacy-preserving federated learning using the Vision Transformer. In federated learning, learning is performed by collecting updated information without collecting raw data from each client. However, the problem is that this raw data may be inferred from the updated information. Conventional data-guessing countermeasures (security enhancement methods) for addressing this issue have a trade-off relationship between privacy protection strength and learning efficiency, and they generally degrade model performance. In this paper, we propose a novel method of federated learning that does not degrade model performance and that is robust against data-guessing attacks on updated information. In the proposed method, each client independently prepares a sequence of binary (0 or 1) random numbers, multiplies it by the updated information, and sends it to a server for model learning. In experiments, the effectiveness of the proposed method is confirmed in terms of model performance and resistance to the APRIL (Attention PRIvacy Leakage) restoration attack.**

I. Introduction

The rapid development of AI technology has accelerated the growth of businesses and services that use deep learning. However, training a model with many parameters requires a large amount of training data, and preparing this training data takes a lot of time and effort. Additionally, we have to take into account the privacy information contained in the training data. In addressing this issue, federated learning, which is a distributed learning method, has attracted much attention [1]–[3].

In federated learning, multiple parties cooperate for deep learning, as shown in Fig. 1. Here, we define a server as a party that provides a model and clients as parties that have training data. Each client independently trains the model shared by the server using its own training data. The clients send their updated information obtained through the training to the server, and the server integrates all the information to update the model. As described above, federated learning can efficiently train high-performance models with a large number of parameters. However, attacks [4]–[7] have been proposed to infer training data on the basis of such updated information.

For this reason, security enhancement methods have been actively researched to protect privacy against such attacks. One method [8] can compute in the encryption domain by using homomorphic cryptography [9] and retain accuracy; however, it is computationally expensive. In another method [10], updated information is concealed by combining model subsampling, model shuffling, and blanket noise. The method is, however, vulnerable to collusion between a server and a party shuffling the model. Consequently, this approach causes a reduction in accuracy. There is a different type of method based on differential privacy [11]–[15]. In this type, each client adds a random number sequence following a specific distribution to the updated information. To enhance security, however, the variance in the distribution has to be enlarged. Thus, the model performance tends to decrease.

There also exists research on the use of the Vision Transformer (ViT) [16] for federated learning. ViT is an image classification model known for its high performance. However, a gradient leakage attack called Attention Privacy Leakage (APRIL) [7], which attempts to recover training images on the basis of updated information for ViT, has become a serious problem. Against APRIL, Aso et al. proposed an effective model encryption method [17]. This method maintains accuracy but does not possess resistance against attacks from internal parties such as clients. Lu et al. locked the positional embedding layer of ViT to prevent it from being updated [7]. Although this method is robust against internal attacks, the performance is degraded when using a model without pre-training.

To tackle the above issue, we propose a novel security enhancement method for ViT-based federated learning without performance deterioration. In the proposed method, each gradient of ViT is multiplied by random binary weights, and all parameters including those in the positional embedding layer can be updated. Simulation results show the effectiveness of our method through an evaluation on the classification accuracy and resistance against APRIL.

II. Preparation

We enhance security for federated learning with ViT, assuming APRIL as a gradient leakage attack. In this section, we first give brief explanations of ViT and APRIL, and we then summarize several previous methods that have resistance against gradient leakage attacks.

A. Vision Transformer

ViT with self-attention mechanisms has attracted much attention for its high performance in the field of image recognition and classification [16]. On the other hand, in the federated learning field, there have been studies on attacks

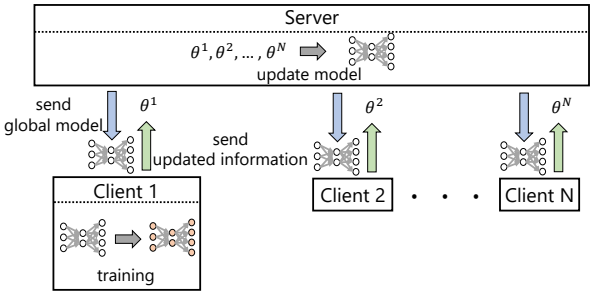
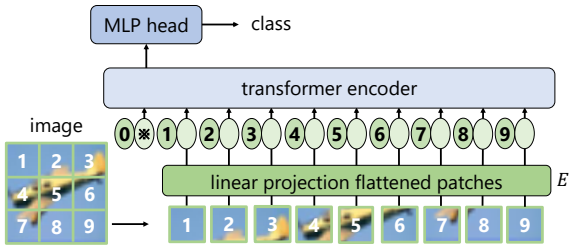
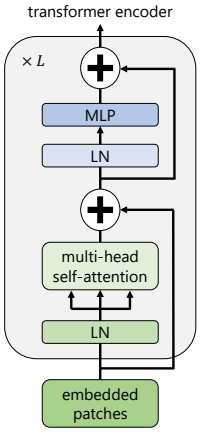


Fig. 1. Overview of federated learning



(a) Overview



(b) transformer encoder layer

Fig. 2. ViT structure

targeting ViT, and a powerful restoration attack called APRIL has been proposed [7].

Fig. 2(a) illustrates the procedure for ViT. ViT first divides an input image $x \in \mathbb{R}^{H\times W\times C}$ into S patches with a size of $P \times P$, where H, W, and C are the height, width, and number of channels. Here, S is determined by HW/P^2. E is an embedding that then dimensionally transforms each patch to be adequate for the transformer encoder layer. In addition, a class token $x_{class} \in \mathbb{R}^D$ is set at the beginning of the patches to represent the features of the entire image. Next, the positional embedding layer $E_{pos} \in \mathbb{R}^{(S+1)\times D}$ embeds information on the location relationship among class tokens and patches. The calculation result $z_0 \in \mathbb{R}^{(S+1)\times D}$ is given by

$$z_0 = [x_{class}; x_P^1 E; x_P^2 E; \cdots; x_P^S E] + E_{pos}, \tag{1}$$

and it is input to the transformer encoder layer.

As shown in Fig. 2(b), the transformer encoder layer is further decomposed into three layers: layer normalization (LN), multi-head self-attention (MSA), and multi-layer perceptron (MLP) [16]. Let z_{l-1} be the input and z_l be the output in the transformer encoder layer in the $l \in \{1, 2, \cdots, L\}$-th layer. We explain the calculation process for obtaining z_l from z_{l-1} in the l-th transformer encoder layer. First, z_{l-1} is normalized by LN. In LN, the output $LN(z_{l-1})$ is obtained by

$$LN(z_{l-1}) = \gamma \frac{z_{l-1} - E[z_{l-1}]}{\sqrt{Var[z_{l-1}] + \epsilon}} + \beta, \tag{2}$$

where γ and β are learnable parameters, and ϵ is a small constant for ensuring that the denominator never takes zero. Next, MSA carries out computation to obtain the output $SA(LN(z_{l-1}))$ by using multiple self-attentions (SAs):

$$SA(LN(z_{l-1})) = softmax\left(\frac{q_l k_l^T}{\sqrt{D_h}}\right) v_l. \tag{3}$$

In the computation of SA, (q_l, k_l, v_l) for the input $LN(z_{l-1})$ are determined by

$$\begin{cases} q_l = LN(z_{l-1})U_{ql}, \\ k_l = LN(z_{l-1})U_{kl}, \\ v_l = LN(z_{l-1})U_{vl}. \end{cases} \tag{4}$$

Here, $(U_{ql}, U_{kl}, U_{vl}) \in \mathbb{R}^{D\times D_h}$ are learnable matrices in the l-th SA layer, and D_h is the vector length D divided by the number of heads. Additionally, (q_l, k_l, v_l) are the input queries, keys, and values linearly transformed by (U_{ql}, U_{kl}, U_{vl}). From (3), in MSA, the output $MSA(LN(z_{l-1}))$ is given by

$$MSA(LN(z_{l-1})) = \left[SA_1; SA_2; \cdots; SA_T,\right] U_{msa}, \tag{5}$$

where SA_t is the output of the $t \in \{1, 2, \cdots, T\}$-th SA, and $U_{msa} \in \mathbb{R}^{k\cdot D_h \times D}$ is a matrix for returning the shape of the MSA output to that of the input. According to (5), in the transformer encoder layer, the output z_l is obtained by

$$\begin{aligned} z_l' &= MSA(LN(z_{l-1})) + z_{l-1}, \\ z_l &= MLP(LN(z_l')) + z_l', \end{aligned} \tag{6}$$

where $MLP(LN(z_l'))$ is the output when $LN(z_l')$ is input into MLP. Finally, we obtain the output y of ViT, which can be calculated from LN by inputting only the class token z_l^0 from the outputs of the L-th transformer encoder layer.

$$y = LN(z_l^0). \tag{7}$$

ViT predicts the label of the input image on the basis of the output y.

B. Attention Privacy Leakage

In federated learning, methods for updating a model can be broadly classified into two categories: Federated Stochastic Gradient Descent (FedSGD) and Federated Averaging (FedAvg) [1]. As mentioned above, we focus on APRIL [7], which is a kind of Gradient Inversion Attack (GIA), because a representative GIA for ViT is APRIL. APRIL targets FedSGD-based learning. We summarize the attack below.

First, attackers calculate the input to the transformer encoder layer z_0:

$$\begin{aligned} \frac{\partial l}{\partial z_0} z_0^T &= U_{q1}^T \cdot \frac{\partial l}{\partial U_{q1}} + U_{v1}^T \cdot \frac{\partial l}{\partial U_{v1}} + U_{k1}^T \cdot \frac{\partial l}{\partial U_{k1}}, \\ \frac{\partial l}{\partial z_0} &= \frac{\partial l}{\partial E_{pos}}. \end{aligned} \tag{8}$$

Here, $\frac{\partial l}{\partial *}$ is the gradient of the parameter $*$ on the loss function l. Next, the inferred image x' is obtained by inversion of (1):

$$x' = E \times (z_0 - E_{pos})^T. \tag{9}$$

In (9), E and E_{pos} refer to the parameters of the global model before being updated. APRIL attempts to reconstruct the input images for learning by using the above procedure.

C. Previous Method

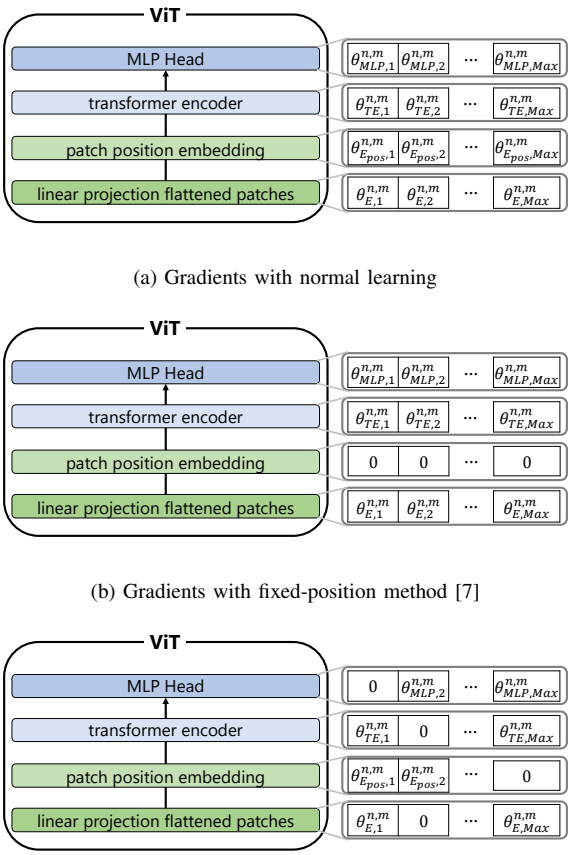
Several security enhancement methods for federated learning have been proposed. A representative method is differential privacy [11]–[15]. In differential privacy, we add noise to the updated information so as to enhance security at a low computational cost. However, differential privacy degrades model performance. Another method lets a third party scramble the updated information for high anonymity for each client [10]. This approach requires the assumption that the third party is absolutely credible.

In contrast, there is another security enhancement method for federated learning with ViT, which is called the fixed-position method [7]. This method is implemented assuming FedSGD. In FedSGD, the update of parameters is expressed by

$$w_{A,i}^{m+1} = w_{A,i}^m - \eta \frac{\sum_{n=1}^{N} \theta_{A,i}^{n,m}}{N}, \tag{10}$$

where $w_{A,i}^{m+1}$ and $w_{A,i}^m$ are the i-th global model parameters of the A-th layer after the $m+1$-th or m-th updates, η is the learning rate, N is the number of clients, and $\theta_{A,i}^{n,m}$ is the gradient for updating $w_{A,i}^m$ calculated by the n-th client. The parameter update is carried out for each batch training. Fig. 3(a) depicts the gradients obtained from training using a plain FedSGD. In this figure, Max is the number of gradients in each layer.

The fixed-position method enhances security but does not learn the positional embeddin layer. Fig. 3(b) shows gradients transmitted in federated learning with the fixed-position method. All the gradients in the positional embedding layer are changed to 0. In this case, the learning process is described as follows.



(a) Gradients with normal learning

(b) Gradients with fixed-position method [7]

(c) Gradients with random binary weight (proposed)

Fig. 3. Update of gradients

Step 1: A server sends global model information to each client.
Step 2: Each client trains each local model using his/her own dataset.
Step 3: Each client converts gradients in the positional embedding layer to 0.
Step 4: Each client sends the updated gradients to the server.
Step 5: The server integrates the information sent from all the clients and updates the global model.

We repeat the steps a specified number of times. The fixed-position method is highly robust against APRIL. The attacker can be a server, one of the clients, or an external third party; the security enhancement method is effective against all of them. The method, however, changes all gradients in the positional embedding layer to zero, so we cannot update the parameters. This leads to severe degradation in model performance when learning a model without pre-training.

III. Proposed Method

We propose a novel method that enhances security for federated learning using ViT. The proposed method is robust

against APRIL without affecting model performance.

A. Main Procedure

The following series of steps is the main procedure for the federated learning to which the proposed method is applied.

- **Step 1:** A server sends global model information to each client.
- **Step 2:** Each client trains each local model using his/her own dataset.
- **Step 3:** Each client multiplies gradients by a random binary sequence.
- **Step 4:** Each client sends the updated gradients to the server.
- **Step 5:** The server integrates the information sent from all the clients and updates the global model.

We repeat this procedure a predefined number of times. The main difference from the previous method [7] is Step 3. Next, we explain Steps 3 and 5 in more detail.

B. Multiplication of Random Binary Weights

Fig. 3(c) illustrates a simplified gradient processing for each layer of ViT in the proposed method. In Step 3, each client multiplies gradients by random binary weights $B_{A,i}^{n,m} \in \{0, 1\}$. Here, a weighted gradient $\theta_{A,i}^{n,m'}$ is given by

$$\theta_{A,i}^{n,m'} = B_{A,i}^{n,m} \times \theta_{A,i}^{n,m}, \tag{11}$$

where $B_{A,i}^{n,m}$ is a binary weight that is multiplied by $\theta_{A,i}^{n,m}$. In the case that the occurrence probability of zeros $R \in [0, 1]$ is set higher, more robust privacy protection would be provided; however, the model may not be updated correctly. We tackle this issue by changing the random binary weights $B_{A,i}^{n,m}$ at each epoch.

It is possible to define a different probability for each layer. The positional embedding layer E_{pos} and linear layer E are layers that include the major information of an image, so APRIL uses these layers for image restoration. Thus, for instance, it would be useful to assign high probabilities for E_{pos} and E while assigning low probabilities for other gradients. In this paper, we apply the same probabilities R to all layers to simplify the argument.

C. Gradient Integration

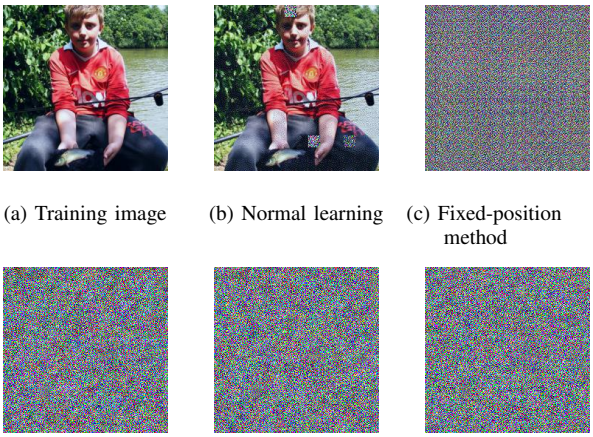
Next, we elaborate on the gradient integration in Step 5. The proposed method integrates gradients on the basis of FedSGD. However, it is not appropriate to directly apply (10) to the proposed method because each client randomly converts the gradients to 0. The gradient integration is implemented as:

$$w_{A,i}^{m+1} = \begin{cases} w_{A,i}^{m} - \eta \frac{\sum_{n=1}^{N} \theta_{A,i}^{n,m'}}{\sum_{n=1}^{N} B_{A,i}^{n,m}} & \text{if } 0 < \sum_{n=1}^{N} B_{A,i}^{n,m} \leq N \\ w_{A,i}^{m} & \text{if } \sum_{n=1}^{N} B_{A,i}^{n,m} = 0. \end{cases} \tag{12}$$

Note that zero gradients are excluded when calculating the mean of the gradients. As an exception, parameters cannot be updated in the case of $\sum_{n=1}^{N} B_{A,i}^{n,m} = 0$.

TABLE I: Fundamental conditions

# of clients	5
Pre-trained model	vit_small_patch16_224
Patch size	16
Learning rate	0.0001
Dataset	CIFAR10
Occurrence probability of zeros	0.2, 0.5, and 0.8



(a) Training image (b) Normal learning (c) Fixed-position method

(d) Proposed method (R=0.2) (e) Proposed method (R=0.5) (f) Proposed method (R=0.8)

Fig. 4. Images inferred by APRIL

IV. Experimental Results and Discussion

In this section, we clarify the effectiveness of the proposed method through experiments from the aspects of attack resistance and model performance.

A. Setup

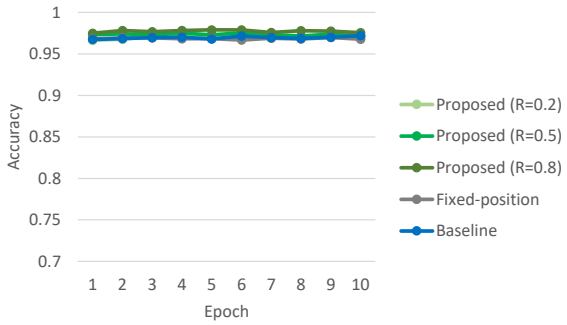
Table I lists the fundamental conditions of our experiments. We virtually configured a server and five clients on a single machine. For a pre-trained model, we used vit_small_patch16_224, where the patch size P and learning rate η were 16 and 0.0001, respectively. We further set the occurrence probability of zeros R to 0.2, 0.5, and 0.8. The experiments were conducted using the CIFAR10 dataset, which consists of 50,000 training images and 10,000 test images. The images were resized from $32 \times 32 \times 3$ to $224 \times 224 \times 3$ by bilinear transformation to correspond to the input size of ViT.

We used 16 images from the CIFAR10 dataset to evaluate robustness against APRIL. The code of APRIL was referenced from [18]. The images were input to the model fine-tuned with CIFAR10, and the inferred images were generated by APRIL from the updated information.

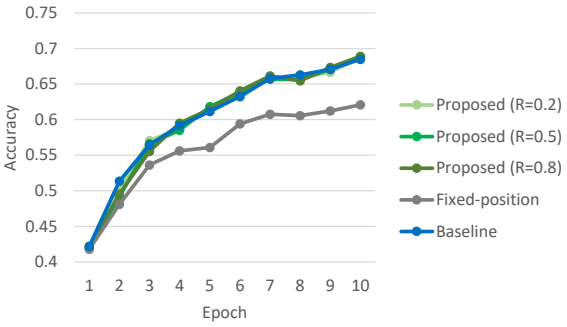
Additionally, we evaluated model performance. Using a model pre-trained using ImageNet and another model without pre-training, we verified the classification accuracy using the 10,000 test images at the end of each epoch.

B. Robustness against APRIL

Fig. 4 depicts the restoration results by APRIL using the updated information. Fig. 4(a) is a training image, while 4(b)



(a) W/ pre-training



(b) W/o pre-training

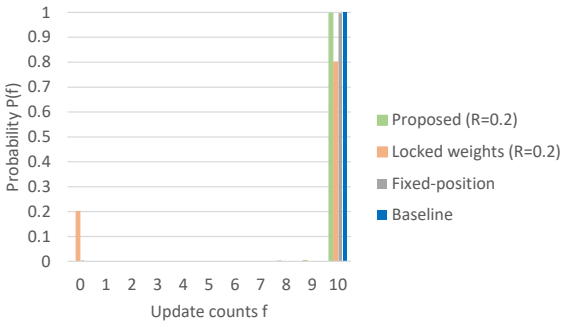
Fig. 5. Comparison of classification accuracy

is an inferred image when using a plain model. In this case, the visual information of the training image was fully disclosed by APRIL. In contrast, as shown in Fig. 4(c), the fixed-position method prevented APRIL from successful restoration. Similarly, from Figs. 4(d), (e), and (f), which are the images inferred by the proposed method, it was difficult for APRIL to successfully restore the training image even at $R = 0.2$. Thus, we can claim that the proposed method is as robust to APRIL as the fixed-position method.

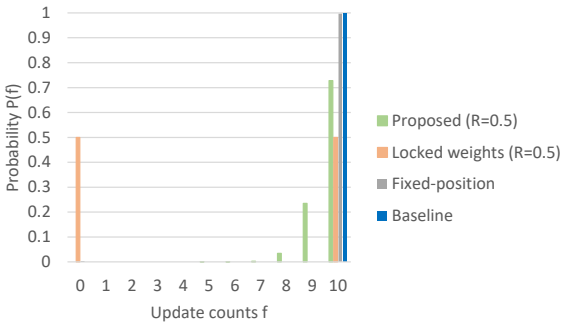
C. Image Classification Performance

We assessed the influence of security enhancement with the proposed and fixed-position methods on classification accuracy. Fig. 5(b) shows the transition in classification accuracy for each epoch in the case of using the model pre-trained with ImageNet. In this figure, the baseline is federated learning with the plain model. It is clear that both the proposed and fixed-position methods retain similar accuracy to the baseline.

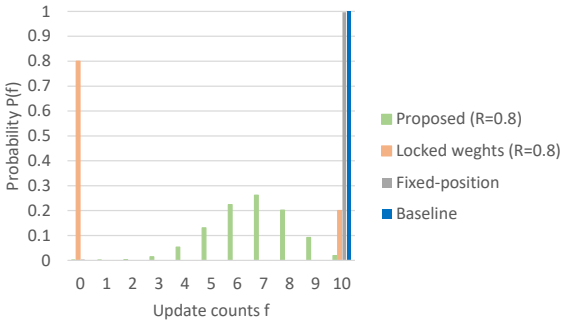
Fig. 5(c) is the transition in classification accuracy for each epoch using the model without pre-training. Although the proposed method could maintain accuracy, the fixed-position method significantly reduced it. In the case of using a model without pre-training, we should update all layers of ViT because the initial parameters are generally not suitable for classifying desired images. The fixed-position method does not update the positional embedding layer; this is the reason for the reduction in accuracy. In contrast, the proposed method updates all the layers, so accuracy can be improved.



(a) $R = 0.2$



(b) $R = 0.5$



(c) $R = 0.8$

Fig. 6. Probability distribution of update counts

D. Discussion

We discuss why the classification accuracy was not degraded with the proposed method. If the random binary weight of each client for a given parameter is zero, the parameter is not updated. Thus, in the case that each client uses a common sequence of random binary weights throughout all epochs, some of the parameters may never be updated. We call this a locked binary weights method. In contrast, in the proposed method, the clients use a different sequence of random binary weights varying from epoch to epoch. This significantly reduces the probability that parameters will never be updated. In the proposed method, the probability $P(f)$ that a parameter is updated f times is obtained by

$$P(f) = {}_mC_f \times (1 - R^n)^f \times R^{n \times (1-f)}, \qquad (13)$$

where m, R, and n denote the number of epochs, the occurrence probability of zeros, and the number of clients, respectively. Additionally, ${}_nC_r$ is the number of combinatons

that is the number of ways of choosing r elements from a group of n elements. From this equation, the update frequency of each parameter will get larger as the number of clients and epochs increases. Consequently, the proposed method is expected to train the model more effectively compared with the locked binary weights method.

Fig. 6 shows the probability distribution of update counts with $m = 10$ and $n = 5$. Here, in the proposed method and the locked binary weights method, R was defined as 0.2, 0.5, and 0.8. From the results, it is clear that most of the parameters were constantly updated through 10 epochs with the fixed-position method. This method, however, never updates the positional embedding layer. Updating the positional embedding layer is important for improving the classification accuracy, so the method degrades the accuracy. Note that it is difficult to recognize this in Fig.6 since the number of parameters in the positional embedding layer is significantly small. The proposed method has a large number of updated parameters, although not as large as the fixed-position method. Since non-updated parameters are not concentrated in a particular layer, our method can also obtain the updates of the positional embedding layer, which is a serious issue with the fixed-position method. Furthermore, the use of an independent sequence of binary weights for each epoch enabled the proposed method to update more parameters than the locked binary weights method at the values of R defined in this section, that is, $R = 0.2, 0.5, and, 0.8$. For these reasons, the proposed method provides not only enhanced security against attacks but also updates of a large number of parameters.

V. Conclusion

We proposed a new security enhancement method for ViT-based federated learning. With most previous methods, the classification accuracy degrades. The proposed method not only solves this issue but is also made robust to the APRIL attack by multiplying updated information by random binary weights. Through our simulations, we confirmed resistance to APRIL and retention of model performance. Our future work involves the application of this method to other models such as CNN models.

VI. Acknowledgemet

This work was partially supported by JSPS KAKENHI Grant Number JP21H01327.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B.A.Y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Proc. Int. Conf. Artif. Intell. Stat.,* pp.1273-1282, 2017.

[2] L. Li, Y. Fan, M. Tse, and K. Y. Lin, "A review of applications in federated learning," *Computers & Industrial Engineering,* vol.149, 106854, Elsevier, 2020.

[3] R. Aso, S. Sayaka, and H. Kiya, "Disposable-key-based image encryption for collaborative learning of Vision Transformer, " *arXiv preprint arXiv:2408.05737.,* 2024,

[4] Y. Wen, J. Geiping, L. Fowl, M. Goldblum, and T. Goldstein, "Fishing for User Data in Large-Batch Federated Leaarning via Gradient Magnification, " *Proc. Int. Conf. Machine Learning,* pp.23668-23684, 2022.

[5] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients, " *Proc. Int. Conf. NeurIPS.,* pp.14774-14784, 2019.

[6] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting Unintended Feature Leakage in Collaborative Learning, " *Proc. Int. Conf. IEEE Secur. Priv.,* pp.691-706, 2019.

[7] J. Lu, X. S. Zhang, T. Zhao, X. He, and J. Chang, "April: Finding the Achilles' Heel on Privacy for Vision Transformers, " *Proc. CVPR.,* pp.10051-10060, 2022.

[8] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-Preserving Deep Learning via Additively Homomorphic Encryption, " *IEEE Trans. Inf. Forensics Security,* vol.13, pp.1333-1345, 2017.

[9] A. C. Yao, "Protocols for Secure Computations, " *Proc. Int. Conf. IEEE FOCS.,* pp.160-164, 1982.

[10] R. Liu, Y. Cao, H. Chen, R. Guo, and M. Yoshizawa, "Flame: Differentially Private Federated Learning in the Shuffle Model, " *Proc. Int. AAAI Conf. Artif. Intell.,* vol.35, no.10, pp.8688-8696, 2021.

[11] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially Private Empirical Risk Minimization, " *Journal of Machine Learning Research,* vol.12, pp.1069-1109, 2011.

[12] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, et al., "A Hybrid Approach to Privacy-Preserving Federated Learning, " *Proc. Int. Conf. AlSec.,* pp.1-11, 2019.

[13] A. Triastcyn, and B. Faltings, "Federated Learning with Bayesian Differential Privacy, " *Proc. Int. Conf. IEEE Big Data,* pp.2587-2596, 2019.

[14] A. E. Ouadrhiri, and A. Abdelhadi, "Differential Privacy for Deep and Federated Learning: A Survey, " *IEEE Access,* vol.10, pp.22359-22380, 2022.

[15] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, et al., "Federated Learning with Differential Privacy: Algorithms and Performance Analysis, " *IEEE Trans. Inf. Forensics Security,* vol.15, pp.3454-3469, 2020.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, " *Proc. ICLR., Virtual Event, Austria,* pp.1-21, 2021.

[17] R. Aso, S. Shiota, and H. Kiya, "Enhanced Security with Encrypted Vision Transformer in Federatd Learning. " *Proc. Int. Conf. IEEE Global Conference on Consumer Electronics* pp.819-822, 2023.

[18] [Online]. Available: https://github.com/JonasGeiping/breaching

Estimation of rotation angle and anisotropic scaling rate using pilot signals for watermarking

Rinka Kawano and Masaki Kawamura
* Graduate School of Sciences and Technology for Innovation, Yamaguchi University
E-mail: {rinka1005,m.kawamura}@m.ieice.org Tel: +81-83-933-5701

***Abstract*—We propose a method for simultaneously estimating the rotation angle and anisotropic scaling rates. Conventional watermarking methods are robust against non-geometric attacks. However, these methods need to synchronize to resist geometric attacks. Invariant features, such as SIFT, have been used for the synchronization. In this paper, we focus on pilot signals for estimating geometric attacks. By embedding a grid-shaped pilot signal and detecting its change, individual attacks on the scaling rate and rotation angle can be estimated. However, it has been difficult to estimate combined attacks. Our proposed method for estimating rotation angle and anisotropic scaling rates uses Radon coefficients. A Radon transform is applied to the signal obtained by QIM from a stego image. Periodic strong peaks appear in the coefficients at rotation angles. The period of the peaks is also calculated. From the rotation angle and the period at which the peaks appear, the rotation angle and the anisotropic scaling rates can be estimated. Computational results show that the rotation angle can be successfully estimated with an absolute error below 0.2 degrees. In more than half of the results, the absolute error of the anisotropic scaling rates was almost zero.**

I. Introduction

Unauthorized use of digital content posted on social media has become a problem. Digital watermarking [1] has been focused on as one of the technologies to protect digital content. Digital watermarking is a technique for secretly embedding other information in digital content. The embedded information is called a watermark. An image in which the watermark is embedded is called a stego image. It is possible to manage content and detect unauthorized use by embedding user IDs and other information in distributed images. However, illegal users often alter distributed images by scaling, rotating, cropping, etc. They also compress the images when they are saved. This leads to image degradation. These alterations (attacks) to stego images can cause the watermark to disappear or make it difficult to detect the watermark. Thus, the watermark needs to be robust against various attacks.

There are two main types of attacks on images: non-geometric attacks and geometric attacks. Non-geometric attacks, such as JPEG compression and noise addition, alter the pixel values. The watermark is also degraded when a non-geometric attack is applied to an image. As a result, the watermark cannot be extracted. Methods robust against non-geometric attacks include those using error-correcting code [2] and spread spectrum [3], [4]. By applying redundancy to the watermark, these methods can correct errors in degraded watermarks.

In contrast, geometric attacks such as scaling, rotating, cropping, etc. change the position of pixels. It is necessary to detect the embedding position because a geometric attack on an image changes the coordinates at which the watermark is embedded. As a method robust against geometric attacks, some watermarking methods using the SIFT features [5], [6], [7], [8], have been proposed. By embedding watermarking around multiple SIFT features, it is possible to improve the resistance to geometric attacks. An embedding method that is robust to linear transformations by normalizing images to a universal shape was proposed by P. Dong et al. [9]. However, both the image and the watermark itself can be degraded by the normalization.

Many conventional methods require knowledge on the parameters of inverse transformation or involve computationally expensive processes such as brute force against geometric attacks. As a result, these methods lack sufficient tolerance to unknown geometric attacks. If we could estimate the type and strength of geometric attacks, we could detect the exact location of the watermark, reducing the number of watermarking errors. For this reason, we have focused on the framework of communication channel estimation in order to estimate the type and strength of the attacks [10]. There is a method that uses a signal different from the message, called a pilot signal, for communication channel estimation. When a pilot signal is transmitted, it is degraded by the communication channel, so the degree of degradation can be used to estimate the parameters of the channel. By introducing the pilot signal into the watermarking scheme, pilot signal degradation can be used to estimate the attack parameters.

A watermarking method using a pilot signal was proposed by Su *et al.* [11]. In their method, a pilot signal was embedded to detect the locations of embedding regions around SIFT feature points. The direction of the embedding regions was determined using the gradient of the SIFT features. The watermarks were embedded into the rotated embedding regions. This method is robust with respect to the change in coordinates. However, it is necessary to rotate the regions in the gradient direction during watermark embedding and extraction. Since these processes are also geometric attacks, they degrade the watermark and image quality.

A new approach is needed to estimate the type and intensity of geometric attacks. To estimate attacks from pilot signals, it is necessary to determine the appropriate embedding method and the shape of the signal for each attack. Since there are different types of attacks, it is necessary to generate signals and develop extraction methods for each type. In our previous

work, we developed a method for estimating attacks with the use of a grid-shaped pilot signal. We showed that individual attacks, i.e., scaling and rotation attacks, can be estimated using the pilot signal [12]. The watermark could also be estimated from the estimated scaling rate [13]. However, it is difficult to estimate these combined attacks.

As an example of estimating a simple composite attack, we proposed a method for estimating the rotation angle and scaling rate simultaneously using the Radon transform [10]. The Radon transform is an integral transform that line integrates an image from different angles. When a grid-shaped pilot signal is Radon transformed, strong intensities appear in the Radon coefficients due to line integration at the same angle as the grid. Therefore, by determining the angle at which strong intensities appear in the Radon coefficient, the rotation angle can be estimated. Also, these strong intensities appear with the same period as the grid interval. Thus, the scaling rate can be estimated by detecting this interval and comparing it with the original grid interval. However, this method is only effective for attack detection when the horizontal and vertical scaling rates are equal. When an anisotropic scaling is performed on an image, the orthogonality of the grid lines is lost, so the anisotropic scaling rate and the rotation angle cannot be estimated simultaneously. In this paper, we propose a method for simultaneously estimating the rotation angle and the anisotropic scaling rate. We theoretically derive equations for estimating these attack parameters from Radon coefficients. We also validate these equations with computer simulations.

This paper is organized as follows. Section II outlines the method for estimating the rotation angle and anisotropic scaling rate using the pilot signal. Theoretical equations for the rotation angle and the grid interval estimated from the Radon coefficient are derived. Section III describes the results of computer simulations to verify the theoretical equations, and Sec. IV concludes our study.

II. Geometric transformations and Radon transform

In the proposed method, a grid-shaped pilot signal is embedded in the original image. By detecting changes in the signal, geometric attacks such as rotation and anisotropic scaling are estimated. From the Radon coefficients obtained by Radon transform of the pilot signal obtained from the attacked image, the rotation angle and the scaling rate are estimated. The estimation procedure is carried out as follows. (i) The pilot signal is extracted from the attacked image. (ii) The pilot signal is transformed by Radon transform to obtain the detection angle and the detection interval. (iii) The rotation angle and scaling rate are estimated using both the detection angle and the detection interval.

The assumed attack combinations are rotation and anisotropic scaling. Let T_θ be the rotation matrix when the rotation angle is θ. Let $T_S = \mathrm{diag}(S_x, S_y)$ be the scaling matrix when the scaling rate on the x-axis is S_x and the scaling rate on the y-axis is S_y, where $\mathrm{diag}(\cdot)$ represents a diagonal matrix. The transformation matrix when the image is scaled and then rotated is given by

$$T_\theta T_S = \begin{pmatrix} S_x \cos\theta & -S_y \sin\theta \\ S_x \sin\theta & S_y \cos\theta \end{pmatrix}. \quad (1)$$

The transformation matrix when the image is rotated and then scaled is given by

$$T_S T_\theta = \begin{pmatrix} S_x \cos\theta & -S_x \sin\theta \\ S_y \sin\theta & S_y \cos\theta \end{pmatrix}. \quad (2)$$

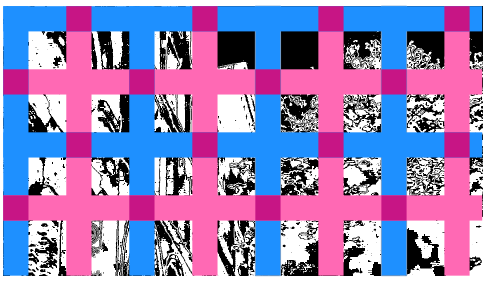


Fig. 1. Pilot signal instead of watermark. Note that pilot signal $p \in 0, 1$ is embedded in pixel value $U(i, j)$.

A. Pilot signal

The image is decomposed into a YUV color space. A pilot signal is embedded into the U-component of the image using QIM [14]. The shape of the pilot signal is a grid as shown in Figure 1 and consists of two directional lines, one at an angle of $\psi_h = 0$ to the x-axis and the other at an angle of $\psi_v = \pi/2$. The interval between the lines of the same color is set to $\gamma = 100$ pixels. The width of the lines is set to 5 pixels. The values on the lines take two forms. The values 0 and 1 are embedded in the pilot signals of the blue and red lines, respectively. Alternating values of 1 and 0 are embedded in the purple area where the two types of lines intersect. No changes are applied to the regions other than the pilot signal. This means that during detection, image-specific components are detected in the regions where no pilot signal is embedded.

B. Extraction of pilot signals and Radon transformation

First, the stego image is decomposed into the YUV color space. From the U-component image $U'(x, y)$, a binary signal is obtained with QIM [14]. This signal is called the extracted signal $\hat{p}(x, y)$. Applying the Radon transform to the extracted signal $\hat{p}(x, y)$, the slope of lines in the signal can be detected as a feature. Let us consider the case where only rotation is applied to the image. Let ϕ_1 and ϕ_2 be the detection angles obtained from the Radon coefficients when the image is rotated by the angle $0 \leq \theta < \pi/2$. Figure 2 shows a schematic diagram of the Radon coefficients of the pilot signal extracted from the rotated image. The vertical axis is the projection position, and the horizontal axis is the projection angle. When a grid-shaped pilot signal, consisting of straight lines in two orthogonal directions, is Radon-transformed, strong intensities appear at two angles, ϕ_1 and ϕ_2 ($\phi_1 < \phi_2$), as shown in the figure.

The method used to estimate the detection angle and detection interval from the Radon coefficients is an adapted version of the

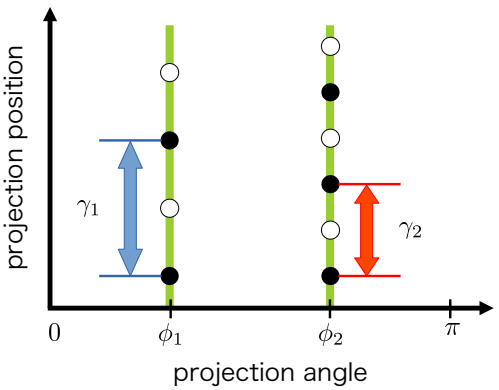


Fig. 2. Schematic diagram of Radon coefficients

method in our previous study [10]. Here, we briefly describe how to detect strong intensities from Radon coefficients and calculate detection intervals. Strong intensities appear at equal intervals at detection angles ϕ_1 and ϕ_2. From the variance of the Radon coefficients for each projection angle, we can see that the variance is large at the projection angles ϕ_1 and ϕ_2. Therefore, the detection angles ϕ_1 and ϕ_2 can be obtained by detecting the peaks from the variance of the Radon coefficient at each projection angle. At angles ϕ_1 and ϕ_2, there are strong periodic intensities in the Radon coefficients. Each interval is called a detection interval, denoted by γ_1 and γ_2. The Radon coefficients contain not only the pilot signal but also components of the original image. Therefore, it is difficult to determine the detection interval automatically. After calculating the autocorrelation of the Radon coefficients, the detection interval is obtained by using DFT on the autocorrelation coefficients. The grid interval of the pilot signal coincides with the detection interval only in the case of rotation. The relationship between the detection angles ϕ_1, ϕ_2 and the rotation angle θ is derived theoretically in this section.

In the Radon transform, the projection plane is rotated clockwise from the y-axis in order to detect components perpendicular to the projection plane. This means that the projection plane of the Radon transform is shifted by $\pi/2$. In contrast, the geometric transform rotates counterclockwise. Note that due to the difference in the direction of rotation between the geometric and Radon transform, the angle of rotation on the Radon coefficients is represented as $-\theta$. Let ϕ_h and ϕ_v be the angles after θ rotation of the horizontal component $\psi_h = 0$ and the vertical component $\psi_v = \pi/2$ of the pilot signal. They are given by

$$\phi_h = \frac{\pi}{2} - \theta + \psi_h = \frac{\pi}{2} - \theta \bmod \pi, \tag{3}$$

$$\phi_v = \pi - \theta + \psi_v = \frac{\pi}{2} - \theta \bmod \pi. \tag{4}$$

Thus, the estimated rotation angle $\hat{\theta}$ is obtained by

$$\hat{\theta} = \frac{1}{2}\left(\phi_h + \phi_v - \frac{3\pi}{2}\right). \tag{5}$$

Since the pilot signal is rotationally symmetric, rotations by whole integer multiples of $\pi/2$ are indistinguishable.

C. Combined attack of rotation and anisotropic scaling

As expressed in (1) and (2), when an image is attacked by rotation and anisotropic scaling, the results depend on the order of the attacks. When the image is rotated after anisotropic scaling, the rotation angle $\hat{\theta}$ is given by (5) since equations (3) and (4) hold. The scaling rates can be estimated by using the detection intervals γ_1 and γ_2 on the detection angles ϕ_1 and ϕ_2. When the horizontal component ϕ_h of the pilot signal is detected as the detection angle ϕ_1, the detection interval γ_1 on ϕ_1 corresponds to the grid interval of the vertical component ψ_v. Therefore, when the original grid interval is γ, the estimated scaling rates $\hat{S}_x$ and $\hat{S}_y$ are calculated by

$$\hat{S}_x = \frac{\gamma_2}{\gamma}, \ \hat{S}_y = \frac{\gamma_1}{\gamma}. \tag{6}$$

On the other hand, when anisotropic scaling is performed after image rotation, the orthogonality of the grid lines is broken, and Eqs. (3) and (4) do not hold. In this paper, we derive an estimation method for this case. Let us consider the case where the rotation angle is between 0 and $\pi/2$. In this case, the detection angles for the horizontal component ψ_h and the vertical component ψ_v are $\phi_h < \phi_v$, so $\phi_h = \phi_1$ and $\phi_v = \phi_2$. The derivation is described in Appendix A. From (18), the tangent of the estimated rotation angle $\hat{\theta}$ is given by

$$\tan\hat{\theta} = \mathrm{sgn}\left(\frac{1}{\tan\phi_1}\right)\sqrt{-\frac{\tan\phi_2}{\tan\phi_1}}, \tag{7}$$

where $\mathrm{sgn}(x)$ is the sign function. The actual rotation angle θ is between 0 and 2π. Therefore, there are two possible solutions to (7). From eqs. (30) and (31) in Appendix B, the estimated scaling rates $\hat{S}_x$ and $\hat{S}_y$ are given by

$$\hat{S}_x = \sqrt{\frac{\gamma_1\gamma_2 \tan\phi_1 \tan\hat{\theta}}{\gamma^2 \sin|\phi_2 - \phi_1|}}, \tag{8}$$

$$\hat{S}_y = \sqrt{\frac{\gamma_1\gamma_2}{\gamma^2 \tan\phi_1 \tan\hat{\theta} \sin|\phi_2 - \phi_1|}}. \tag{9}$$

D. In case of right angles

If the detection angle ϕ_1 is 0 or π, then $\tan\phi_1 = 0$. Therefore, (7) is not computable. Similarly, $\tan\hat{\theta}$ in (9) includes $\tan\phi_2$. Therefore, if the detection angle ϕ_2 is 0 or π, then $\tan\phi_2 = 0$. Therefore, (9) is not computable. In this paper, rotation and anisotropic scaling are applied to the image once each. As a result, the detection angle ϕ_1 or ϕ_2 becomes zero only if the rotation angle satisfies $\theta = n\pi/2, n = 0, 1, 2, 3$. Therefore, when the detection angle ϕ_1 or ϕ_2 is 0 or π, the estimated rotation angle and the estimated scaling rates are obtained by (5) as in the case of rotation after anisotropic scaling of the image. The estimated scaling rates are given by (6).

III. COMPUTER SIMULATION

We derived equations to estimate the attack parameters by detecting the pilot signal from the Radon coefficients. This section demonstrates that our theoretical equations are correct.

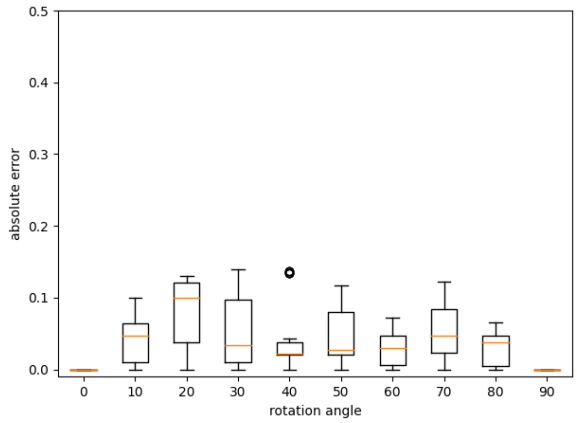


Fig. 3. Absolute error $|\hat{\theta} - \theta|$ for each angle: Orange line represents medians of results.

Specifically, we evaluated our method for estimating the rotation angle and the anisotropic scaling rates with absolute error when the stego image was rotated and then anisotropically scaled.

A. Experimental conditions

The pilot signal was embedded in the U-component of six IHC standard images of 4608×3456 pixels [15]. We considered attacks in which the stego images were rotated, anisotropically scaled, and then cropped. The angles of rotation were set to $0, 10, 20, \cdots, 90$ degrees. In addition, the scaling rate for the x-axis S_x was set to 1, and the scaling rates for the y-axis were assumed to be $S_y = 0.5, 0.6, \cdots, 1.5$. Since the Radon transform is performed on a square surface, the stego images were cropped to a size of 1080×1080 pixels from the center of the image after the geometric transformation.

B. Experimental results

First, we evaluated the estimated rotation angle $\hat{\theta}$ in terms of the absolute error $|\hat{\theta} - \theta|$. Since our pilot signal is grid-shaped, four candidates are detected as estimated rotation angles. These are indistinguishable. Therefore, the angle closest to the true rotation angle was treated as the estimated rotation angle. There were $11 \times 6 = 66$ results since there were 11 different scaling rates in the y-axis direction. Figure 3 shows their box-and-whisker plots. The horizontal axis is the rotation angle θ. The vertical axis is the absolute error $|\hat{\theta} - \theta|$. The orange line represents the medians of the results. All rotation angles were successfully estimated with an absolute error of less than 0.2 degrees.

Next, the estimated scaling rates were evaluated. The absolute errors of the estimated scaling rates $\hat{S}_x$ and $\hat{S}_y$ are shown in Figure 4 as box-and-whisker plots. The horizontal axis is the scaling rate S_y. Now remember that we set $S_x = 1.0$. The vertical axis represents the absolute errors of (a) the estimated scaling rates $\hat{S}_x$ and (b) $\hat{S}_y$. The orange line represents their medians. For all scaling rates, the medians were close to zero. However, we found that many results had a large absolute error for scaling rates greater than 1.0.

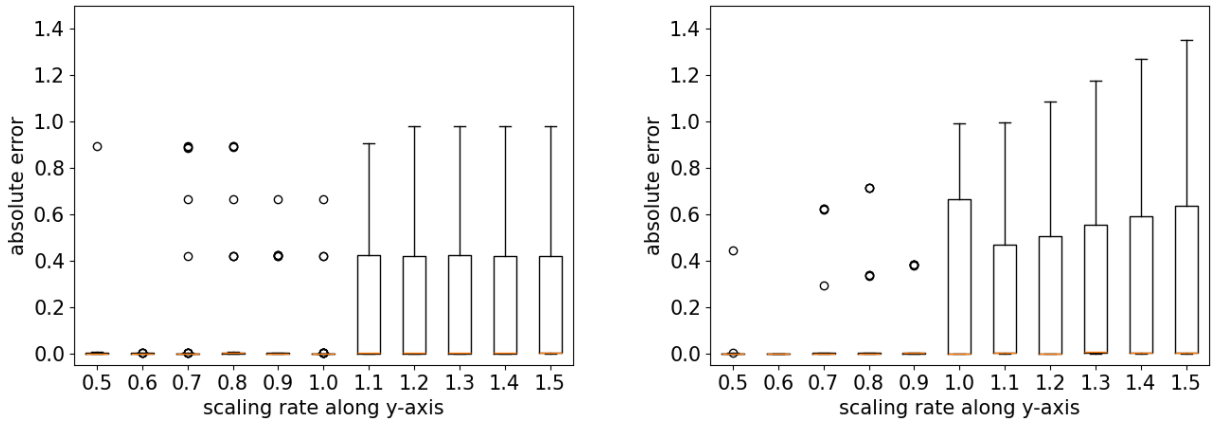
We considered the reason why the estimation errors are more frequent when the scaling rate is large. From our previous study [10], we know that the larger the scaling rate, the more likely it is that the grid interval γ_1 and γ_2 will fail detection. This is due to the fact that when a scaled image is cropped to a given size, the larger the scaling rate, the fewer the number of grid lines within the cropped image. In addition, if the components of the original image are not smooth, the components may cause the peak to not be detected, resulting in incorrect period estimation. In summary, the reason why the absolute error of the scaling rate was large, while the estimated rotation angle was calculated accurately, is due to the inaccurate detection of the grid intervals γ_1 and γ_2.

IV. Conclusion

Blind watermarking methods need to be robust against geometric attacks. Methods with redundant watermark embedding are ineffective against geometric attacks. Therefore, methods that assume geometric attacks by using SIFT feature points and the polar coordinate system have been proposed. We proposed a grid-shaped pilot signal for estimating geometric attacks. So far, we have been successful in estimating a single attack, either a rotational attack or an isotropic scaling attack [10]. The angle and interval of the grid can be obtained by applying the Radon transform to the pilot signal extracted from the image.

In this paper, we proposed a method for estimating both the rotation angle and the anisotropic scaling rates from images attacked by a combined rotation attack and anisotropic scaling attack. We theoretically derived the estimated rotation angle and the estimated scaling rates from the Radon coefficients and the geometric transformation matrix. Using both the detection angles and the detection intervals obtained from the Radon coefficients, the estimated rotation angle and the estimated scaling rates can be calculated. To demonstrate the correctness of the estimators obtained from our theory, we validated them in computer simulations. In particular, we verified the accuracy of the estimators for the case of anisotropic scaling after rotating stego images. As a result, it was possible to estimate the rotation angle with an absolute error of less than 0.2 degrees. The median of the estimates was zero for the estimation of anisotropic scaling. That is, the absolute error was zero for more than half of the results. In particular, the estimation was almost accurate when the scaling rate was less than 1. In contrast, when the scaling rate was greater than 1, the number of cases with large absolute errors increased. This is due to inaccuracies in the estimates of the detection intervals γ_1 and γ_2.

The proposed method could estimate the combined rotation and scaling attacks. However, the scaling rate needs to be estimated more accurately. It is necessary to improve the detection of peaks from the Radon coefficients to accurately detect the detection intervals γ_1 and γ_2. Improving the detection algorithm and estimating general geometric transformations including shear is our future work.



(a) Scaling rate in x-axis direction, $\hat{S}_x$ (b) Scaling rate in y-axis direction, $\hat{S}_y$

Fig. 4. Absolute error of estimated scaling rates

V. Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP20K11973 and JP24K15106, the Telecommunications Advancement Foundation, the Yamaguchi University Fund, and the Cooperative Research Project Program, R04/B09 Research on Multifunctional Multimedia Production, of the RIEC, Tohoku University. A part of the computations were performed using the supercomputer facilities at the Research Institute for Information Technology, Kyushu University.

References

[1] F. A. Petitcolas, R. J. Anderson, M. G. Kuhn, "Information hiding-a survey," Proceedings of the IEEE, vol. 87, no. 7, pp. 1062–1078, 1999
[2] H. Kang, K. Iwamura, "Watermarking based on the difference of discrete cosine transform coefficients and an error-correcting code," Proc. of IWIHC, pp. 9–17, 2014
[3] I. J. Cox, J. Kilian, F. T. Leighton and T. Shamoon, "Secure spread spectrum watermarking for multimedia," in IEEE Transactions on Image Processing, vol. 6, no. 12, pp. 1673–1687, 1997
[4] J. J.K.O. Ruanaidh, T. Pun, "Rotation, scale and translation invariant spread spectrum digital image watermarking," Signal Process., vol. 66, Issue 3, pp. 303–317, 1998
[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International J. Compute Vision, vol. 60, no. 2, pp. 91–110, 2004
[6] M. Hayashi, M. Kawamura, "Improved SIFT feature-based watermarking method for IHC ver. 5," Proc. of APSIPA ASC 2018, pp. 1536–1543, 2018
[7] V. Solachidis, I. Pitas, "Circularly symmetric watermark embedding in 2-D DFT domain," IEEE Trans Image Process, vol. 10, no. 11, pp. 1741–1752, 2001
[8] J. S. Seo, C. D. Yoo, "Localized image watermarking based on feature points of scale-space representation," Pattern Recognition, vol. 37, Issue 7, pp. 1365–1375, 2004
[9] P. Dong, J. G. Brankov, N. P. Galatsanos, Y. Yang and F. Davoine, "Digital watermarking robust to geometric distortions," in IEEE Trans. on Image Processing, vol. 14, no. 12, pp. 2140–2150, 2005
[10] R. Kawano, M. Kawamura, "Estimation of scaling rate and rotation angle using pilot signal for image watermarking," IEICE Technical Report, vol. 123, no. 407, pp. 28–33, 2024
[11] P. C. Su, Y. C. Chang and C. Y. Wu, "Geometrically resilient digital image watermarking by using interest point extraction and extended pilot signals," IEEE Trans. Inf. Forensics Secur., vol. 8, no. 12, pp. 1897–1908, 2013
[12] R. Kawano, M. Kawamura, "Detection of periodic pilot signal in image watermarking," APSIPA ASC 2021, SS-IVM-1.6, pp. 1647–1652, 2021
[13] R. Kawano, M. Kawamura, "Watermarking method with scaling rate estimation using pilot signal," IEICE Trans. Inf. & Syst., vol. E107-D, no. 9, 2024
[14] B. Chen, G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," IEEE Trans. Inf. Theory, vol. 47, no. 4, pp. 1423–1443, 2001
[15] Information Hiding and its Criteria for evaluation, https://www.ieice.org/iss/emm/ihc/en/image/image.php (20th February, 2023 access)

Appendix

A. Estimation of rotation angle

We will consider the case where the rotation angle θ is in $0 \leq \theta < \pi/2$. The detection angles obtained from the Radon coefficients are ϕ_1 and ϕ_2 ($\phi_1 < \phi_2$). The projective plane of the Radon transformation is shifted by $\pi/2$. The direction of rotation is opposite that of the rotation transformation. Let $A'(x_a, y_a)$, $B'(x_b, y_b)$, and $C'(x_c, y_c)$ be points $A(\gamma, 0)$ on the horizontal component ϕ_h of the pilot signal, $B(0, \gamma)$ on the vertical component ϕ_v and $C(\gamma, \gamma)$ after the transformation. Let α be the angle between the x-axis and line OA'. Let β be the angle between the x-axis and line OB'.

Figure 5 shows these angles α, β, ϕ_1, and ϕ_2. The angles α and β are represented by

$$\alpha = \frac{\pi}{2} - \phi_1,\ \beta = \frac{3\pi}{2} - \phi_2. \tag{10}$$

The tangent of these angles is therefore given by

$$\tan\alpha = \frac{1}{\tan\phi_1},\ \ \tan\beta = -\frac{1}{\tan\phi_2}. \tag{11}$$

Therefore, the angle $\angle A'OB' = |\beta - \alpha|$ between line OA' and line OB' is given by

$$|\beta - \alpha| = |\pi + \phi_1 - \phi_2|. \tag{12}$$

Point A' projected from point A is given by

$$\begin{pmatrix} x_a \\ y_a \end{pmatrix} = T_\theta T_S \begin{pmatrix} \gamma \\ 0 \end{pmatrix} = \begin{pmatrix} \gamma S_x \cos\theta \\ \gamma S_y \sin\theta \end{pmatrix}. \tag{13}$$

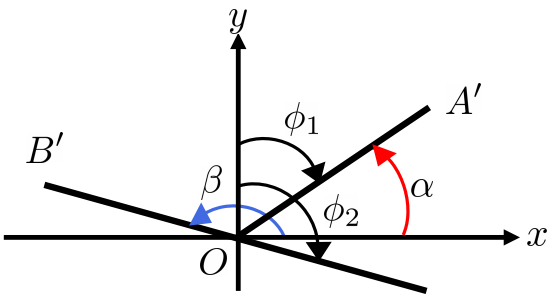


Fig. 5. Relationship between angles α, β, ϕ_1, and ϕ_2.

Therefore, the tangent of the slope of line OA', $\tan\alpha$, is given by

$$\tan\alpha = \frac{S_y}{S_x}\tan\theta. \tag{14}$$

Similarly, point B' projected from point B is given by

$$\begin{pmatrix} x_b \\ y_b \end{pmatrix} = T_\theta T_S \begin{pmatrix} 0 \\ \gamma \end{pmatrix} = \begin{pmatrix} -\gamma S_x \gamma \sin\theta \\ \gamma S_y \gamma \cos\theta \end{pmatrix}. \tag{15}$$

Therefore, the tangent of the slope of line OB', $\tan\beta$, is given by

$$\tan\beta = \frac{S_y}{S_x}\left(-\frac{1}{\tan\theta}\right). \tag{16}$$

From (14) and (16), we obtain

$$-\frac{\tan\alpha}{\tan\beta} = -\frac{\tan\phi_2}{\tan\phi_1} = \tan^2\theta. \tag{17}$$

The signs of the tangents of the two detection angles ϕ_1 and ϕ_2 are always different, even if the orthogonal grids are rotated and scaled anisotropically. Therefore, the sign of tangent $\tan\hat{\theta}$ of the estimated rotation angle and that of tangent $\tan\alpha$ of the detection angle are always the same. Therefore, from (11), the tangent of the estimated rotation angle $\hat{\theta}$ is given by

$$\tan\hat{\theta} = \mathrm{sgn}\left(\frac{1}{\tan\phi_1}\right)\sqrt{-\frac{\tan\phi_2}{\tan\phi_1}}, \tag{18}$$

where $\mathrm{sgn}(x)$ is the sign function.

B. Estimation of scaling rates

Let us first consider the area Q of the quadrilateral $OA'C'B'$ to find the scaling rates. Since the quadrangle $OA'C'B'$ is a parallelogram, the area of the parallelogram is expressed as

$$Q = |OA'||OB'|\sin|\beta-\alpha|. \tag{19}$$

The parallelogram $OA'C'B'$ and the grid intervals γ_1, γ_2 are shown in Figure 6. The distance between line OA' and line $B'C'$ is represented by the grid interval γ_1, and the distance between line OB' and line $A'C'$ is represented by the grid interval γ_2. The lengths of edges $|OA'|$ and $|OB'|$ are therefore given by

$$|OA'| = \frac{\gamma_2}{\sin|\beta-\alpha|},\ |OB'| = \frac{\gamma_2}{\sin|\beta-\alpha|}. \tag{20}$$

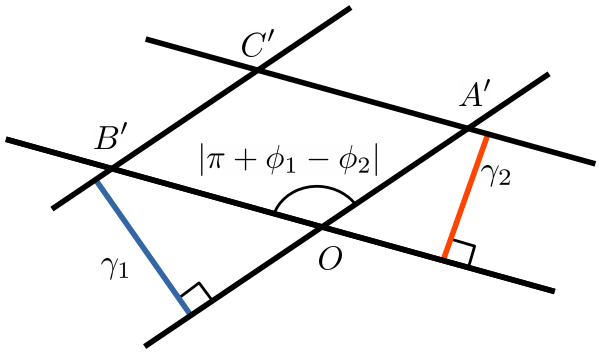


Fig. 6. Relationship between quadrilateral $OA'C'B'$ and grid intervals γ_1, γ_2.

From (19), the area Q is expressed as

$$Q = \frac{\gamma_1\gamma_2}{\sin|\beta-\alpha|}. \tag{21}$$

Next, we calculate the area Q of the parallelogram in a different way, focusing on (13) and (15). From (12), $\sin|\beta-\alpha|$ can be expanded as

$$\begin{aligned}
\sin|\beta-\alpha| &= \sin|\pi+\phi_1-\phi_2| && (22)\\
&= \sin|\phi_2-\phi_1| && (23)\\
&= \sin\phi_2\cos\phi_1 - \cos\phi_2\sin\phi_1 && (24)\\
&= \frac{y_b}{|OB'|}\cdot\frac{x_a}{|OA'|} - \frac{x_b}{|OB'|}\cdot\frac{y_a}{|OA'|} && (25)\\
&= \frac{\gamma^2 S_x S_y}{|OA'||OB'|}. && (26)
\end{aligned}$$

Therefore, from (19), the area Q is given as

$$Q = |OA'||OB'|\cdot\frac{\gamma^2 S_x S_y}{|OA'||OB'|} = S_x S_y \gamma^2. \tag{27}$$

This equation states that the area γ^2 before the transformation is equal to the area multiplied by $S_x S_y$ after the transformation at the anisotropic scaling rates.

Finally, from (21) and (27) we obtain

$$S_x S_y = \frac{\gamma_1\gamma_2}{\gamma^2\sin|\phi_2-\phi_1|}. \tag{28}$$

Using (11) and (14), we obtain

$$S_y = \frac{S_x}{\tan\phi_1\tan\theta}. \tag{29}$$

Substituting this into (28), the estimated scaling rate $\hat{S}_x$ is given by

$$\hat{S}_x = \sqrt{\frac{\gamma_1\gamma_2\tan\phi_1\tan\hat{\theta}}{\gamma^2\sin|\phi_2-\phi_1|}}. \tag{30}$$

Similarly, the estimated scaling rate $\hat{S}_y$ is given by

$$\hat{S}_y = \sqrt{\frac{\gamma_1\gamma_2}{\gamma^2\tan\phi_1\tan\hat{\theta}\sin|\phi_2-\phi_1|}}. \tag{31}$$

On the Security of Bitstream-level JPEG Encryption with Restart Markers

Mare Hirose*, Shoko Imaizumi* and Hitoshi Kiya†
* Chiba University, Chiba, Japan
E-mail: marehirose@chiba-u.jp, imaizumi@chiba-u.jp
† Tokyo Metropolitan University, Tokyo, Japan
E-mail: kiya@tmu.ac.jp

***Abstract*—This paper aims to evaluate the security of a bitstream-level JPEG encryption method using restart (RST) markers, where encrypted image can keep the JPEG file format with the same file size as non-encrypted image. Data encrypted using this method can be decoded without altering header information by employing a standard JPEG decoder. Moreover, the use of RST markers enables the definition of extended blocks divided by the markers, so spatially partial encryption and block-permutation-based encryption can be carried out. However, the security of the method was evaluated only with respect to the key space analysis for brute-force attacks and other limited attacks. Accordingly, in this paper, we evaluated the security of the method with respect to robustness against ciphertext-only attacks including state-of-the-art attacks. In experiments, the method is compared with conventional encryption methods, and it is confirmed to be robust against ciphertext-only attacks if parameters used for image encryption are carefully chosen.**

I. Introduction

The use of JPEG images has grown due to the expansion of social networking services. These platforms generally restrict the use of various file formats. In many cases, images have to adhere to the JPEG standard. Furthermore, these platforms are not reliable owing to incidents, including information leakage. Accordingly, many encryption methods that keep the JPEG format have been proposed to date. Recently, a bitstream-level encryption method for JPEG compression was proposed, and it was confirmed to outperform conventional encryption methods in some important respects [1]. However, it has not been verified enough yet with respect to the robustness of the encryption against various attacks. Accordingly, we aim to evaluate the security of the method against ciphertext-only attacks (COAs) in this paper.

Bitstream-level encryption is one of the types of encryption combined with image compression. Compared with other types, this type of encryption has some advantages. For example, it permits us not just to keep the JPEG file format but also to use standard JPEG encoders. In addition, some methods of this type ensure that the file size remains the same before and after encryption, so encryption can be carried out by hooking images within a transmission channel. One state-of-the-art method [1] can also generate partially encrypted images, but all other bitstream-level methods cannot.

In [1], the restart (RST) marker, placed at fixed intervals between minimum coded units (MCUs), is used for encryption. This marker contributes to the mixing of encrypted and unencrypted regions. Furthermore, MCUs divided by RST markers can be defined as extended blocks, and the visibility of the encrypted image can be reduced by permuting the position of these blocks. This block permutation is also expected to enhance security strength.

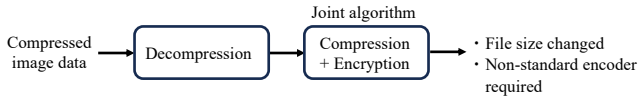
Accordingly, we focus on the above state-of-the-art method to evaluate the attack resistance of the method against ciphertext-only attacks (COAs) in this paper. Key space analysis, key sensitivity analysis, the non-zero counting attack (NZCA), and histogram analysis are performed, and the relationship between the use of RST markers and the attack resistance is discussed. In experiments, it is demonstrated that the RST marker, used by the method used in image encryption for the first time, can enhance the robustness against a variety of attacks in general, and the attack resistance of encrypted images depends on the selection of restart interval values.

II. Related Work

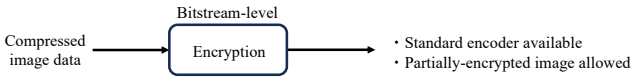
The encryption method discussed in this paper is a method coupled with image compression that is able to produce JPEG files containing obscured visual information [1]. Accordingly, conventional encryption methods coupled with image compression are reviewed here. Furthermore, the structure of JPEG bitstreams is briefly explained.

A. Combined Use of Encryption and Compression

Encryption methods coupled with image compression are categorized into four types as follows. Type 1 is compression-then-encryption (CtE) using standard cryptography, in which images are compressed and then encrypted using a standard cryptography including AES (advanced encryption standard). Type 2 is encryption-then-compression (EtC), in which the visual information of images is secured through a perceptual encryption method, referred to as compressible encryption, and afterward the encrypted images undergo compression [2], [3]. Type 3 combines compression and encryption. For type 3, encryption and compression are processed at the same time (see Fig. 1(a)) [4]–[8]. Therefore, when JPEG images are provided before encryption, the images need to be decompressed prior to performing encryption. Furthermore, standard encoders cannot be used in this framework. Type 4 is CtE using a bitstream-level encryption method (see Fig. 1(b)), in which images



(a) Type 3



(b) Type 4

Fig. 1. Types of encryption.



Fig. 2. Structure of JPEG bitstream.

undergo compression and then encryption using a bitstream-level encryption method [1], [9]–[12]. Type 4 permits us not just to keep the JPEG format but also to use standard JPEG decoders.

We focus on a state-of-the-art method [1], which is a type 4 method for JPEG images. The method enables us to produce encrypted images with different levels of visibility, such as partially encrypted images.

B. JPEG Bitstream and Marker Codes

Here, we describe the structure of the JPEG bitstream. Fig. 2 illustrates an example of the structure of JPEG bitstreams [14]. Codes called SOI and EOI, which respectively signify the beginning and end of a bitstream, are allocated. These are two-byte codes called marker codes. Segments include information necessary for decoding, such as Huffman tables and quantization tables. Marker codes are also allocated at the beginning of each segment. The first byte of the marker code is fixed at $\mathrm{FF}_{(16)}$, and the second byte indicates the type of each marker. The second byte of the marker code distinguishes each segment. Note that $\mathrm{FF00}_{(16)}$ is not defined as a marker code.

Fig. 3 depicts the structure of the image data of a JPEG bitstream for the case of 4:2:0 color subsampling. Image data is divided and processed in units called minimum coded units (MCUs). Each MCU is composed of four luminance components (Y) and two chroma components (Cb, Cr). Each component contains a DC coefficient and AC coefficients, and the DC coefficient is recorded as the difference value from the previous DC coefficient. Finally, these bit sequences are divided and stored byte by byte. Due to this process, $\mathrm{FF}_{(16)}$ may occur in the image data. Therefore, the JPEG encoder inserts $00_{(16)}$ immediately after $\mathrm{FF}_{(16)}$ occurs in the data to tell it apart from the first byte of a marker code. Additionally, when the JPEG decoder detects $\mathrm{FF00}_{(16)}$, it reads only $\mathrm{FF}_{(16)}$ and skips $00_{(16)}$. This process is called "byte stuffing."

One of the JPEG marker codes is called the RST marker, which can be placed at fixed intervals among MCUs. The restart interval (RI), which is the interval at which the RST

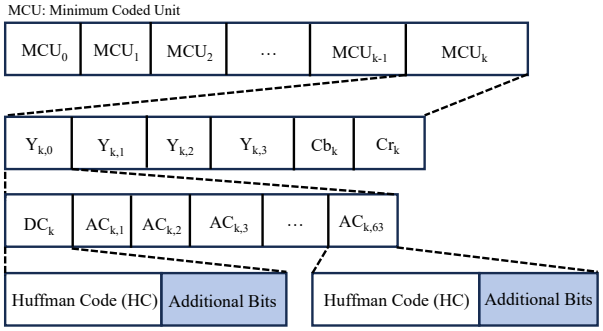


Fig. 3. Structure of image data.

marker is placed, is set during JPEG encoding. When a RST marker is not used, an error that arises in the DC coefficient in a bitstream propagates to the end of the bitstream. In contrast, the value of a DC coefficient following a RST marker is stored as its original value. Therefore, using RST markers prevents errors from propagating among DC coefficients in bitstreams.

III. Security Evaluation

A. Bitstream-level JPEG Encryption with RST Markers

The bitstream-level JPEG encryption method with RST markers [1] is summarized here. Fig. 4 shows the encryption process of the method [1]. The method utilizes RST markers for encryption, where a sequence of MCUs divided by RST markers is referred to as an extended block (see Fig. 5). The encryption procedure for this method is shown below.

- Step 1: Select an RI and prepare a bitstream including RST markers to be encrypted.
- Step 2: Select encryption regions.
- Step 3: Extract bytes that satisfy encryption conditions from the extended blocks to maintain the same file size as that before encryption.
- Step 4: Prepare a binary sequence of pseudorandom numbers (PRNs) by using secret K1. Perform an exclusive-or (XOR) operation between additional bits of the extracted bytes and the sequence of PRNs.
- Step 5: Replace the additional bits with the XOR results.
- Step 6: Randomly permute the positions of the extended blocks defined in Step 1 with secret K2.

The encryption conditions in Step 3 were discussed in [12], [13], in which the bytes of a bitstream are categorized into five cases, as shown in Fig. 6, where the white, black, and red areas denote Huffman codes (HCs), additional bits, and bytes placed immediately after $\mathrm{FF}_{(16)}$, respectively. Each pattern is defined below.

- Pattern 1: Consists solely of HCs.
- Pattern 2: Consists solely of additional bits.
- Pattern 3: Consists of HCs and additional bits, and every bit in the HC is 1.
- Pattern 4: Consists of HCs and additional bits, and the HC contains 0.

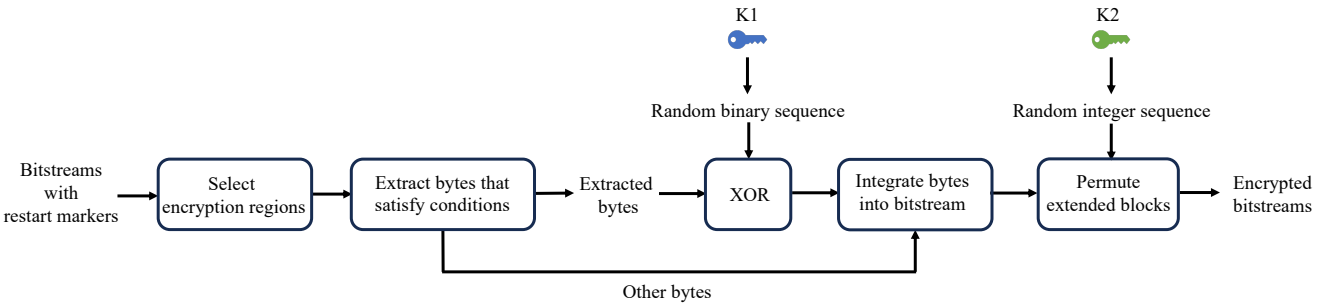


Fig. 4. Block diagram of encryption process.

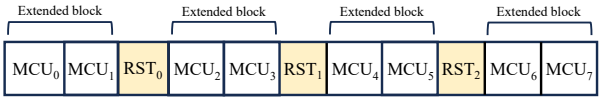


Fig. 5. Bitstreams after insertion of RST markers (RI = 2).

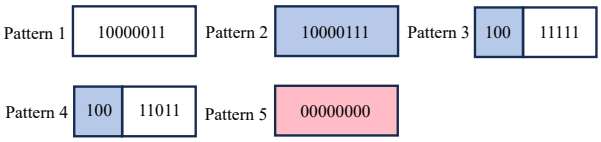


Fig. 6. Five types of bytes.

Pattern 5: Consists solely of 0, and the byte is placed immediately after $FF_{(16)}$.

Among these patterns, only additional bits in Pattern 4 can be encrypted.

In [1], each extended block is encrypted separately. Thus, it is possible to generate a partially encrypted image that contains both unencrypted and encrypted regions. Fig. 7 illustrates an example of images encrypted with this method.

B. Threat Model

A threat model includes a set of assumptions, such as an adversary's goals, knowledge, and capabilities. The aim of an attacker is to restore visual information from encrypted data. We assume that the attacker is able to use encrypted data and the encryption algorithm but does not have the secret key. Accordingly, the attacker can only perform ciphertext-only attacks (COAs) using encrypted images.

C. Security Analysis

Several COAs have been studied to restore visual information from encrypted images [4], [10], [11], [13]. In this paper, we use key space analysis, key sensitivity analysis, the non-zero counting attack (NZCA), and histogram analysis for security evaluation. The relationship between the use of RST markers and the attack resistance is discussed.

1) Key space analysis: Key space is the total number of patterns that can be produced by a encryption algorithm. Here, we analyze the encryption method in the case where all DCT coefficients in the whole image are encrypted. The algorithm encrypts additional bits within bytes that satisfy specific conditions and permutes the positions of the extended blocks, as shown in III-A. We assume that JPEG encoding is applied to an image with a size of $M \times N$ under RI = r and 4:2:0 color subsampling. First, considering that the number of bytes to be encrypted is T, the minimum and maximum sizes of the key space derived from the encryption of the additional bits S_{enc_min} and S_{enc_max} are given by

$$S_{enc_min} = 2^{T}, \tag{1}$$

$$S_{enc_max} = 2^{7T}. \tag{2}$$

Note that the minimum number of bits to be encrypted is one, and the maximum is seven within a byte that satisfies the encryption conditions. The key space obtained from the permutation of the extended blocks, S_{bp}, is expressed as

$$S_{bp} = \left\lfloor \left(\left\lceil \frac{M}{16} \right\rceil \times \left\lceil \frac{N}{16} \right\rceil \times \frac{1}{r} \right) \right\rfloor !. \tag{3}$$

Accordingly, the overall minimum key space S_{min} and maximum key space S_{max} are given by

$$\begin{aligned} S_{min} &= S_{enc_min} \times S_{bp} \\ &= 2^{T} \times \left\lfloor \left(\left\lceil \frac{M}{16} \right\rceil \times \left\lceil \frac{N}{16} \right\rceil \times \frac{1}{r} \right) \right\rfloor !, \end{aligned} \tag{4}$$

$$\begin{aligned} S_{max} &= S_{enc_max} \times S_{bp} \\ &= 2^{7T} \times \left\lfloor \left(\left\lceil \frac{M}{16} \right\rceil \times \left\lceil \frac{N}{16} \right\rceil \times \frac{1}{r} \right) \right\rfloor !. \end{aligned} \tag{5}$$

Equation (3) shows that the number of extended blocks increases as RI becomes shorter. Therefore, an encrypted image with a shorter RI has a larger key space. The algorithm for the method with RST markers can use a shorter RI and thus can provide a larger key space. For example, if the image in Fig 7(a) is encoded with r = 4 as an RI, T = 37,031 is given. Thus, the minimum key space S_{min} can be expressed by

$$S_{min} = 2^{37,031} \times \left\lfloor \left(\left\lceil \frac{384}{16} \right\rceil \times \left\lceil \frac{512}{16} \right\rceil \times \frac{1}{4} \right) \right\rfloor ! > 2^{256}. \tag{6}$$

This is sufficiently larger than 2^{256}, which is the space of a key with 256 bits. If we use r = 2, S_{min} is larger than in



(a) Original (b) Encrypted ($RI = 2$) (c) Partially-encrypted ($RI = 2$) (d) Partially-encrypted ($RI = 4$)

Fig. 7. Examples of encrypted images (ucid00459).

the case of r = 4. Accordingly, the bitstream-level encryption method with RST markers has enough key space.

2) Key sensitivity analysis: Attack-resistant encryption methods should be sensitive even to a slight change in the key. We analyze key sensitivity by considering two cases:

Case 1: Encryption using an encryption key that differs from the original key by only one bit.

Case 2: Decryption using an incorrect key that differs from the correct key by only one bit.

In Case 1, the encrypted image must be wholly different from the image encrypted with the original key. In Case 2, the image decrypted with an incorrect key must be wholly different from the original image.

3) NZCA: The sketch attack tries to obtain contour information of the original one from the encrypted image. We use the non-zero-counting attack (NZCA), which is a sketch attack for encrypted JPEG images [4]. This attack attempts to restore the original image's contours on the basis of the number of non-zero coefficients in each MCU of the encrypted image. The use of RST markers will be evaluated to generate encrypted images robust against NZCA.

4) Histogram analysis: The histogram is very useful for an adversary to analyze a data distribution graphically in the encrypted domain. An adversary tries to analyze the distribution of pixel frequencies in encrypted images to apply attacks. Accordingly, attack-resistant encryption methods should provide encrypted images with histograms similar to those of other encrypted images.

IV. Experimental Results

A. Setup

In experiments, we used 1,338 test images with 384 × 512 pixels or 512 × 384 pixels from Uncompressed Colour Image Database [15]. JPEG compression was performed using libjpeg [16], with the color subsampling set to 4:2:0 and the quality factor Q set to 80. Binary sequences of PRNs were prepared using HMAC_DRBG [17] with a 384-bit key. For all images, we encrypted the whole image and permuted the positions of all extended blocks.

B. Protection of Visual Information

The encryption method with RST markers can permute the positions of blocks at random, thereby strongly protecting the visual information of images, compared with previous methods without block permutation [5], [12], [13]. For objectively evaluating the protection strength, peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) values of encrypted images are shown as boxplots in Fig. 8. From the figure, it is confirmed that the method generates encrypted images without identifiable visual information.

C. Key Sensitivity

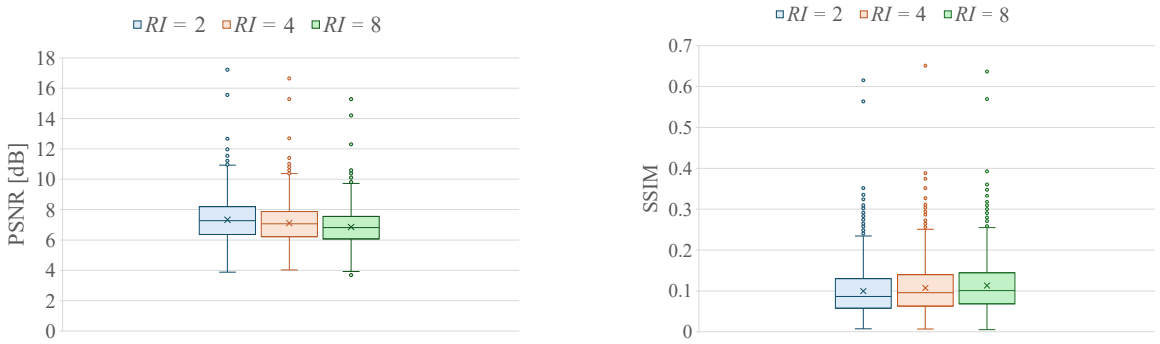
Fig. 9 illustrates the results of the key sensitivity analysis, where the SSIM values indicate the difference between two images. In Fig. 9(a), two images were encrypted with keys that differed by a single bit. In contrast, in Fig. 9(b), the two images were decoded from an encrypted one by using the correct key and an incorrect key that differed from the correct key by a single bit. From the figure, the method was verified to have sufficiently high key sensitivity.

D. NZCA

NZCA was carried out on the luminance component of an encrypted image. As illustrated in Fig. 10, in the previous method without RST markers [12], [13], the original image was revealed in its contour. In contrast, the method with RST markers concealed the original image. This was attained by replacing extended blocks. However, when a large RI was adopted, the outline of the original image was revealed in some areas. Accordingly, the use of RST markers is important for enhancing robustness against attacks, but the value of RI should be carefully selected.

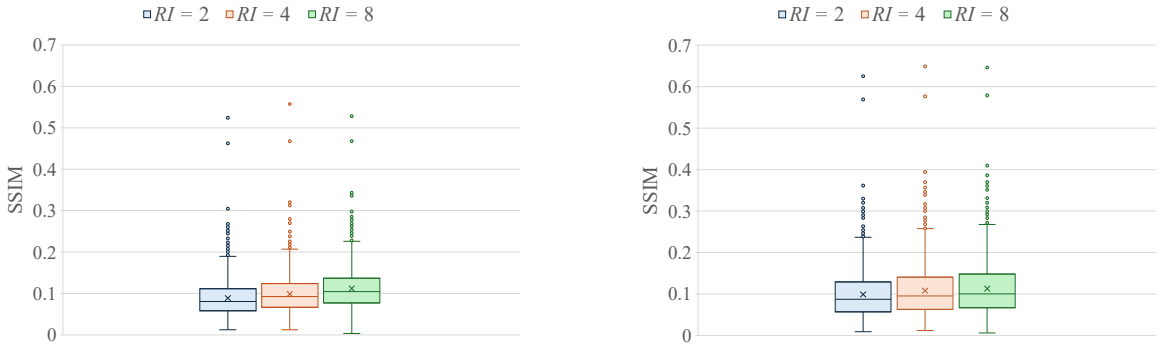
E. Histogram Analysis

Fig. 11 illustrates the R, G, and B histograms of the original ucid00459 image, which is shown in Fig. 7(a), and its encrypted images. In the figure, there are histograms for four encrypted images. Three of them were derived by using our method with RST markers, and the other was derived by using the previous method. In our method with RST markers, RI was defined as 2, 4, and 8. The images encrypted using this method were compared with the original image and those encrypted with the previous method. Attack-resistant encryption methods should provide an encrypted image with a totally different histogram from that of the original image and analogous histograms among the three color channels. From the figure,



(a) PSNR (b) SSIM

Fig. 8. Evaluation of visual protection. Top of box, bottom of box, and line in middle represent 75th, 25th, and 50th percentile, respectively. Whiskers denote highest and lowest values that are not outliers. Circles are outliers, and crosses are means.



(a) Encryption with different keys (Case 1) (b) Decryption with incorrect keys (Case 2)

Fig. 9. Key sensitivity analysis results. Top of box, bottom of box, and line in middle represent 75th, 25th, and 50th percentile, respectively. Whiskers denote highest and lowest values that are not outliers. Circles are outliers, and crosses are means.

we can see that the R, G, and B channels in the encrypted image have analogous histograms. In addition, the use of a small RI can strongly reduce the identifying characteristics of the image. Other images were verified to have similar trends.

V. Conclusion

In this paper, we conducted security analyses on a state-of-the art encryption method with RST markers against ciphertext-only attacks. It was confirmed that the RST markers used in the method contribute to enhancing security in terms of key space, key sensitivity, NZCA, and histogram analysis. As a result, the method was demonstrated not only to outperform conventional JPEG encryption methods in some important respects but also to maintain a high security level. Furthermore, it was shown that the attack resistance varies with the length of the restart interval.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number JP21H01327.

References

[1] M. Hirose, S. Imaizumi, and H. Kiya, "Encryption Method for JPEG Bitstreams for Partially Disclosing Visual Information," *Electronics,* vol. 13, no. 11, 2024.

[2] W. Sirichotedumrong, T. Chuman, S. Imaizumi, and H. Kiya, "Grayscale-Based Block Scrambling Image Encryption for Social Networking Services," in *Proc. IEEE Int. Conf. Multimed. Expo,* pp. 1–6, 2018.

[3] H. Kiya, A. P. M. Maung, Y. Kinoshita, S. Imaizumi, and S. Shiota, "An overview of compressible and learnable image transformation with secret key and its applications," *APSIPA Trans. Signal Inf. Process.,* vol. 11, no. 1, 2022.

[4] Y. Peng, C. Fu, G. Cao, W. Song, J. Chen, and C. W. Sham, "JPEG-compatible Joint Image Compression and Encryption Algorithm with File Size Preservation," *ACM Trans. Multimed. Comput. Commun. Appl.,* vol. 20, no. 4, pp. 1–20, 2024.

[5] K. Shimizu and T. Suzuki, "Finely Tunable Bitcuboid-based Encryption with Exception-free Signed Binarization for Jpeg Standard," *IEEE Trans. Inf. Forensics Secur.,* vol. 16, pp. 4895–4908, 2021.

[6] W. Cao, X. Leng, T. Yu, X. Gu, and Q. Liu, "A Joint Encryption and Compression Algorithm for Multiband Remote Sensing Image Transmission," *Sensors,* vol. 23, no. 17, 2023.

[7] O. Watanabe, A. Nakazaki, and H. Kiya, "A fast image-scramble method using public-key encryption allowing backward compatibility with JPEG2000" in *Proc. Int. Conf. Image Process.,* pp. 3435–3438, 2004.

[8] P. Li, Z. Sun, Z. Situ, M. He, and T. Song, "Joint JPEG compression and encryption scheme based on order-8-16 block transform," *IEEE Trans. Intell. Transp. Syst.,* vol. 24, no. 7, pp. 7687-7696, 2023.

[9] Y. Yuan, H. He, Y. Yang, N. Mao, F. Chen, and M. Ali, "JPEG Image

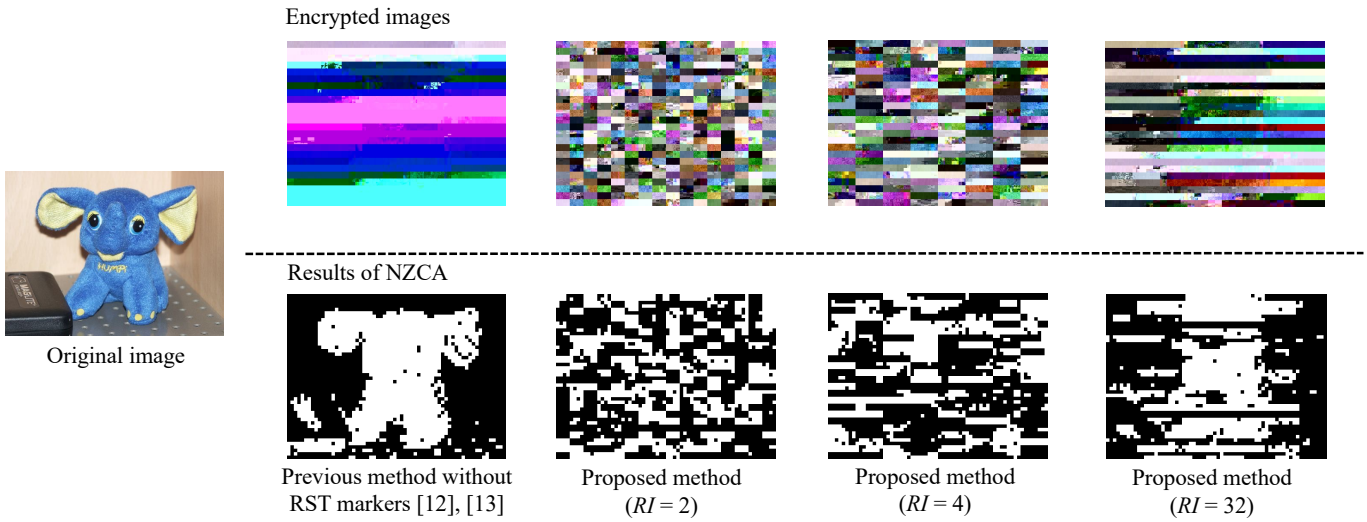


Fig. 10. Results of non-zero-counting attack (ucid00294).

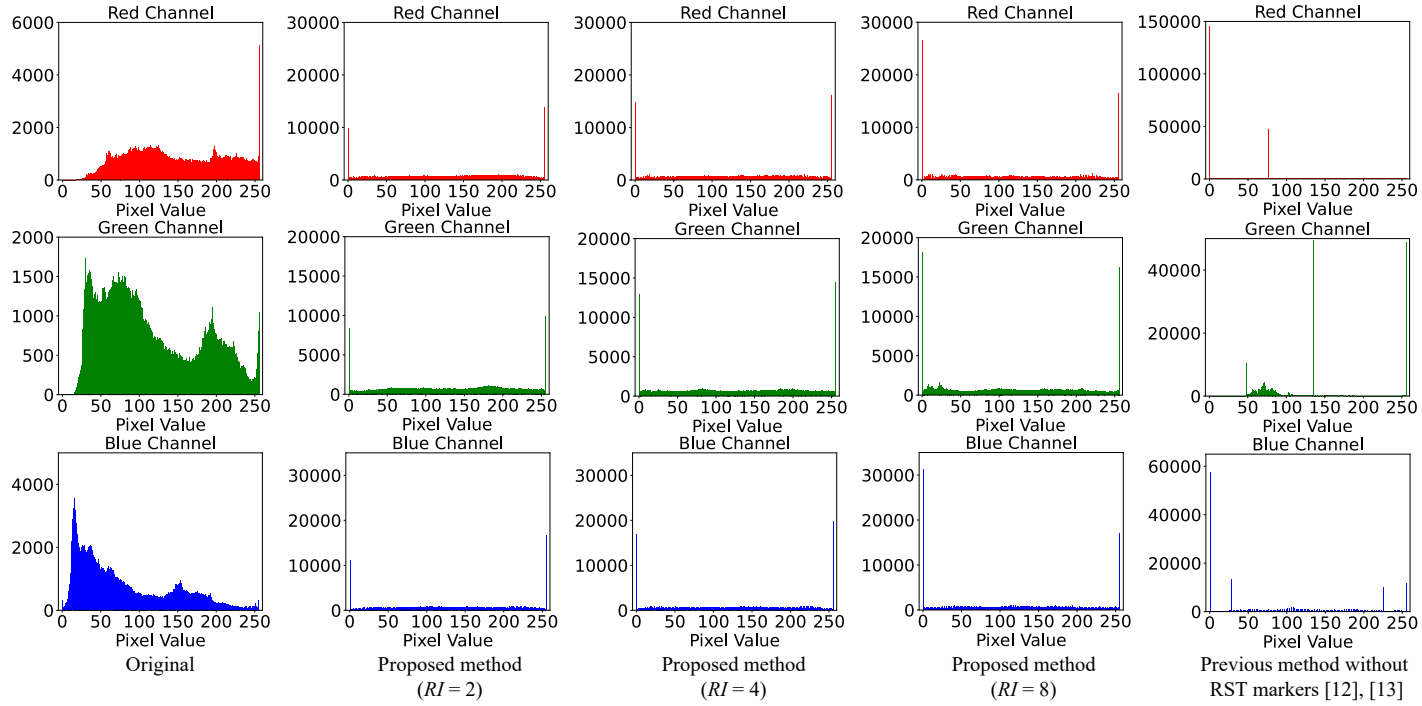


Fig. 11. Histogram analysis (ucid00459).

Encryption with Grouping Coefficients Based on Entropy Coding," *J. Vis. Commun. Image Represent.,* vol. 97, 2023.

[10] J. He, S. Huang, S. Tang, and J. Huang, "JPEG Image Encryption With Improved Format Compatibility and File Size Preservation," *IEEE Trans. Multimedia,* vol. 20, no. 10, pp. 2645–2658, 2018.

[11] C. Qin, J. Hu, F. Li, Z. Qian, and X. Zhang, "JPEG Image Encryption with Adaptive DC Coefficient Prediction and RS Pair Permutation," *IEEE Trans. Multimedia,* vol. 25, pp. 2528–2542, 2022.

[12] H. Kobayashi and H. Kiya, "Bitstream-based Jpeg Image encryption with File-size Preserving," in *Proc. IEEE Glob. Conf. Consum. Electron.,* pp. 384–387, 2018.

[13] H. Kobayashi and H. Kiya, "File-Size Preserving Encryption of JPEG Images in the Bitstream Domain," *IEICE Trans. Inf.& Syst. Jpn. Ed.,* vol. J102-D, no. 12, pp. 787–795, 2019.

[14] *Information technology—Digital Compression and Coding of Continuous-Tone still Images:Requirements and Guidelines.* International Organization for Standardization ISO/IEC IS-10918-1, 1994.

[15] G. Schaefer and M. Stich, "UCID: An Uncompressed Color Image Database," in *Proc. SPIE,* vol. 5307, pp. 472–480, 2003.

[16] [Online] Available:URL:https://www.ijg.org

[17] Barker, E. B. and Kelsey, J. M, *SP 800-90A. Recommendation for Random Number Generation Using Deterministic Random Bit Generators.* NIST: Gaithersburg, MD, USA, 2012.

Improved Ultimate Link without Markers for Projective Transformation

Keiji Yamadera* and Michiharu Niimi†
* Graduate School of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: yamadera.keiji334@mail.kyutech.jp
† Faculty of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: niimi@ai.kyutech.ac.jp

***Abstract*—In our previous work, we have proposed a method that links character string having a certain semantic meaning on papers in real world to the digital content in cyberspace, and call it "Ultimate Link." Ultimate Link adds additional information to characters by drawing colored circles on top of the characters without changing the shape of the characters. A URL can be generated by connecting additional information of semantically organized character strings from an image taken with a digital device. The conventional Ultimate Link drew markers to normalize the drawn circles (markers for projective transformation), similar to the finder pattern in a QR code. In order to achieve a more realistic application, it is desirable it. We report on an improved Ultimate Link that can handle complex multiple character strings. Specifically, improvements were made to enable extraction of information by identifying character strings with embedded additional information from sheets of paper printed with multiple character strings. As a result of experiments, we confirmed that information could be successfully embedded and extracted for both character strings in the target image.**

I. Introduction

Today, there is a growing movement toward "Digital Transformation (DX)," which utilizes information technology to enrich people's lives and transform existing values and frameworks in all areas of society. Then, DX is also being promoted in the field of education. For example, mobile information terminals such as tablets and notebooks are being introduced into educational settings to enable students to learn more effectively. However, several studies have shown the superiority of the learning effects of paper-based learning over information terminal-based learning[1]. We believe that if URLs can be attached to paper documents, it is possible to benefit from DX while efficiently deepening understanding of things.

In order to realize this, it is necessary to superimpose link information on a single semantic coherent string of text on the paper. For example, we assume that each rendered character of "Mt.Fuji" has additional information and is printed on papers, then the printed characters are captured by digital camera. From the captured image, a URL reconstructed by the additional information extracted from each rendered characters may be assumed to a web site of explaining Mt.Fuji.

As the method to do this, we have proposed Ultimate Link[2] that links semantically coherent character strings printed on paper with digital contents in cyberspace, allowing anyone to easily superimpose a large amount of information on the printed text without deforming the shape of the characters and extract information in real time by using standard cameras of common electronic devices. In [2], an image rendering only a single character string, which means there is no objects in the image without characters consisting a certain meaning, to be embedded with information was prepared, and then information was superimposed on the character and markers for the projective transformation necessary for image normalization during extraction were drawn around the character string to generate an information embedded image. Ultimate Link expresses bit information by arranging multiple colored circles on a font. In other words, the bit information corresponds to the color information. In this sense, we call a circle that retains bit information a bit-colored circle. In [2], the Ultimate Link system was evaluated through experiments of information superimposition and extraction using various types of character strings, such as multiple katakana-only character strings and multiple character strings containing both kanji and hiragana, in Japanese. However, a printed document containing only a single string of text is not realistic, and markers for projective transformation drawn around the string can disrupt the layout of the printed document. In this paper, we report on an improved Ultimate Link system , which aims to improve the Ultimate Link system for more realistic applications. Specifically, in the information embedding process, the circles conventionally used as bit-colored circles are given a function equivalent to projection markers. These circles are called position detection markers. In an image in which multiple character strings are rendered, information is embedded by drawing bit-color circles and position detection markers on the font only for the character strings in which the information is to be embedded. In the information extraction process, the printed information-embedded image is captured by a smartphone camera so that the entire image is captured, position detection markers are detected from the captured image, and information is extracted by identifying the character string in which the information is embedded.

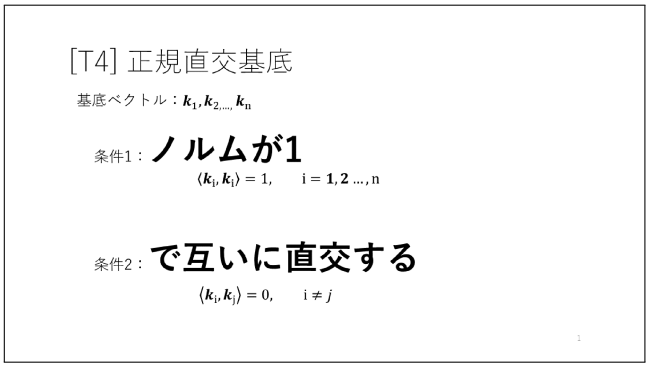


Fig. 1: Example of image for embedding information

In the following sections of this paper, an overview of Ultimate Link and a method to realize an improved Ultimate Link are described in Section II, experiments are conducted in Section III, discussions are given in Section III-D. Finally, in Section IV, we summarize the results and discuss future issues.

II. Ultimate Link

Ultimate Link is a method that links a certain semantic cohesion in the real world to accompanying/supplementary information in cyberspace. This method is consisting of two parts : superimposing (embedding) and extracting binary data on rendered characters. In embedding process, information is embedded by drawing bit-colored circles on the text that correspond to color information and bit information in order to superimpose bit information on the text. On the other hand, in extracting process, information is extracted by detecting bit-colored circles on the text and arranging them in an appropriate order. Each process is described below.

A. Embedding Process

The embedding process follows the flow below.

(1) Character recognition
(2) Character segmentation
(3) Character image thinning
(4) Pseudo-stroke order calculation
(5) Grid points assignment
(6) Grid points for position detection assignment
(7) Bit-colored circle drawing
(8) Final processing

We assume that the improved Ultimate Link deals with an image data that multiple Japanese characters are rendered, which is called a cover image. There are several character strings having certain meaning in the characters, and Fig.1 is an example of this. We explain each procedure listed above in detail using the example cover image. In addition, we use a character string : "ノルムが 1" (in Japanese) as a target string, which is a character string for information to be embedded, and denote it as TS. Note that TS is assumed to be a single line and written horizontally, so it is not possible to select TS that extends over multiple lines or a character string that is written vertically. In addition, we need to define the information to be embedded, which is a shortened URL.

(1) Character recognition
Character recognition by Vision API[3], which is an image recognition service provided by Google, is performed to the entire cover image. Firstly, the entire characters in cover image are recognized. Then, only the part of TS is extracted from the character recognition result.

(2) Character segmentation
TS consisting of multiple characters is divided into individual characters. In [2], this was performed by image processing for entire cover image, but it sometime failed when the cover image contains characters other than TS. Therefore, we use the response of the character recognition function of Vision API which includes the rectangular area information of each recognized character.

(3) Character image thinning Before performing the thinning of each segmented character image, the erosion is performed for the characters. After erosion of each character, we perform thinning to it. A binary image with a line width of 1 is obtained from this process. This information is used for the position of drawing circles.

(4) Pseudo-stroke order calculation
Ultimate Link expresses bit information by drawing bit-colored circles on characters. Here we introduce the concept of "pseudo-stroke order" of characters to define the order of bit information, i.e., we define the order in which bit-colored circles are drawn on the characters.
The pseudo-stroke order is calculated by tracing the order of foreground pixels from an image of a thinned character image. This order based on the contour tracking method and is calculated by the following procedure.

Step 1 : The first pixel which is found in the foreground by raster scanning is set to the target pixel.
Step 2 : The next target pixel is search by a counter-clockwise direction within the adjacent pixel of the target pixel.
Step 3 : Until there are no more adjacent pixels, Step 2 is repeated, and then return to Step 1.
Step 4 : Until all foreground pixels is searched, steps of 1 through 3 is repeated.

The pseudo-stroke order is calculated by the above procedure. The obtained pseudo-stroke order is saved for each character.

(5) Grid points assignment
By using pseudo-stroke order, we can scan the foreground pixels of thinning image with unique order. Following this order, we draw a circle of radius r with d as the distance between circles. In some cases, the circles may overlap depending on the character shape and scanning order. In such cases, we skip the foreground pixel and then we search the next one that the drawn circle is not overlapped with pseudo-stroke order. After this procedure, we can determine the center of circle on the foreground pixels of thinning image, which is called "grid point." Fig.2 shows

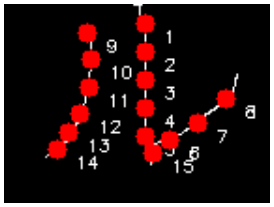
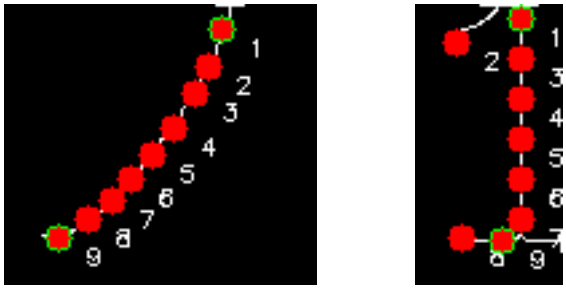


Fig. 2: Example of grid points



(a) The first character of TS **(b)** The last character of TS

Fig. 3: Grid points for position detection (2 for each character)

an example of drawing grid points.

(6) Grid points for position detection assignment
In order to identify the region of TS with the information is embedded in the extraction process, we define position detection markers on grid points. The markers are consisted of four grid points obtained by the first and last grid points of the first character and that of the last character. In the TS example, the first and last grid points for "ノ" and the first and last grid points for "1." Fig.3 shows the results of the grid points for position detection. In the figure, the grid points for position detection are circled by green circles.

(7) Bit-colored circle drawing
For the obtained grid points, colored circles which color is selected by the embedded information are drawn. We call it a bit-colored circle. The position at which bit-colored circle is drawn is randomly selected from among the grid points with a fixed random seed according to the length of the embedding information. One circle represents two bits (four patterns), and five colors should be in total, since colors are also needed for markers for position detection. In the HSV color space, especially in the OpenCV COLOR BGR2HSV FULL color space, each HSV value is described as 0 to 255, and within this space, five colors are set by changing the H value as shown in Table I, where S = 255 and V = 255. The reason why we chose these colors is that it makes the proposed method easy to implement. The colors in Table I are, from top to bottom, markers for position detection and embedded bit information: $(00)_2$, $(01)_2$, $(10)_2$ and $(11)_2$.

(8) Final processing
As bit-colored circles are superimposed on rendered character images, we replace original characters with it. In addition, we draw position detection markers on the first and last characters at the grid points for position detection

TABLE I: Colors for grid points/detection markers

H value	Actual color	Role
16		detection marker
46		00
106		01
166		10
226		11



Fig. 4: Example of information embedding

obtained in (7). Fig.4 shows an example of a string with embedded information. Multiple colored circles are drawn on the characters according to the embedded information.

B. Extraction Process

An image with embedded information created in Section II-A is printed on low-quality paper, and the embedded information is extracted from an image of the printed paper captured by a camera of a digital device such as a smartphone (hereinafter referred to as "captured image").

The extraction process follows the flow below.

(1) Search the position of a string with embedded information in the captured image
(2) Image binarization
(3) Character recognition
(4) Character segmentation
(5) Bit-colored circle detection
(6) Generate temporal character fonts
(7) Drawing the reference grid points on the temporal character font
(8) Drawing white circles at the center coordinates of the detected bit-colored circles
(9) Alignment based on detected bit-colored circles and reference grid points.
(10) Nearest circle detection
(11) Bit information extraction

The following explanation of each procedure uses, as an example, an image printed with information embedded in "ノルムが 1" created in Section II-A and captured with a smartphone camera.

(1) Search the position of a string with embedded information in the captured image
We search the position of the string with embedded information in the captured image. The position of a character string in which information is embedded is identified by detecting the position detection markers drawn on the first and last characters of the character string. The position detection markers are detected by generating a mask image in the color of the position detection marker set in advance in the HSV color space, expanding the mask image, and

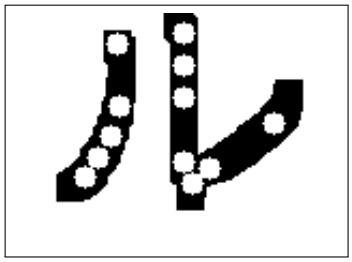


Fig. 5: Drawing reference grid points



Fig. 6: Drawing detected circles

detecting circles by Hough transform on the expanded mask image. This identifies the coordinates of the four position detection marker points and fills in the area outside of the four points with white color. This process identifies the character string in which the information is embedded because only the target character string remains in the image, and all other character strings are filled with white.

(2) Image binarization
After processing (1), binarization is performed using the predefined colors in HSV color space to improve the accuracy of character recognition.

(3) Character recognition
Character recognition is performed on the binarized image created in (2). Since only the character string to be embedded is available in the image, the result of character recognition for the entire image can be used as that of the target string recognition.

(4) Character segmentation
The string is divided into individual characters in the same way as during the embedding process.

(5) Bit-colored circle detection
The method for detecting bit-colored circles is almost the same as the method for position detection markers in (1), but an opening process is applied to the expanded mask image before the Hough transform to remove noise.

(6) Generate temporal character fonts
The pseudo-stroke order information saved during the embedding process generates a binary image with a line width of 1. Base on this information, we make a temporal font image by expanding it, which means like a reverse step of thinning image. The image generated by this process is called a temporal character font.

(7) Drawing the reference grid points on the temporal character font
The grid points are obtained from the pseudo-stroke order information saved at the time of embedding, and the grid points used to embed the information are obtained using the same random seed as that used for embedding. These grid points are reference grid points. These reference grid points are drawn as white circles on the temporal character font. Fig. 5 shows an example of an image with reference grid points drawn.

(8) Drawing detected bit-colored circles
The center coordinates of the bit-colored circle detected in (5) are drawn as a white circle as shown in Fig. 6.

(9) Alignment based on detected bit-colored circles and reference grid points
Based on Fig. 5 and Fig. 6, the detected bit-colored circle position is aligned to the reference grid point position. This procedure is one of the most important steps in the extraction process. In (5), the bit information embedded in the string itself was extracted by detecting bit color circles, but the order of the bit information was not recovered. The order of the bit information is recovered by finding the reference grid point closest to the center coordinates of the detected bit-colored circle in (10), described below, and arranging the bit information corresponding to the color of the bit-colored circle in the order of the reference grid points in (11). In this case, the coordinates of the bit-colored circle obtained in (5) are obtained from the coordinate system of the image from which the string was segmented from the captured image, and thus no normalization to the coordinate system of the reference grid point is performed. In this paper, normalization is performed by aligning one character at a time using the coordinates of the bit-colored circle detected for each character and the coordinates of the reference grid point. Therefore, if the alignment is not performed accurately, the mapping between the detected bit-color circle and the reference grid point in (10) and (11) fails, and the embedded information cannot be recovered. Thus, this alignment procedure is an important step in the extraction process.
The alignment method is based on Procrustes analysis[4]. Procrustes analysis is a method of superimposing two point groups that correspond to each other in such a way that the squared error between the point groups is minimized. Thus, we apply this to the coordinates of the detected bit-colored circles and reference grid points. Therefore the transformation matrix that minimizes the squared error between the coordinates of the detected bit-colored circles and the coordinates of the reference grid points can be calculated, then the matrix is applied to the coordinates of the detected bit-colored circles and superimposed on the coordinates of the reference grid points.

(10) Nearest circle detection
The nearest neighbor circle is obtained by finding the

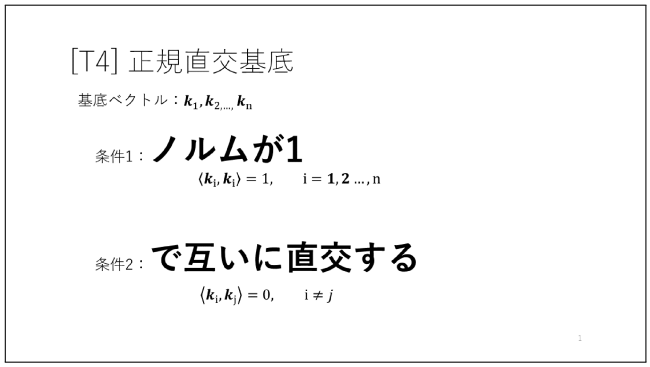
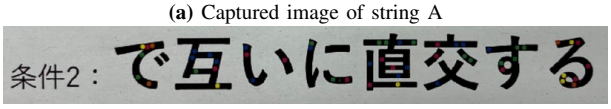


Fig. 7: Image for embedding information



(a) Captured image of string A



(b) Captured image of string B

Fig. 8: Captured images

center coordinates of the detected bit-colored circle that has been aligned and the center coordinates of the nearest reference grid point.

(11) Bit information extraction
The bit sequences corresponding to the colors of the nearest neighbor circles are ordered. In this way, the embedded information is extracted.

III. Experiments

Two experiments were conducted to evaluate the improved Ultimate Link proposed in Section II. In the first experiment referred as Experiment I , we conducted information embedding and extraction processing experiments on two Japanese character strings A: “ノルムが 1” and B: “で互いに直交する” in an image containing multiple character strings as shown in Fig.7. The second experiment referred as Experiment Ⅱ was conducted to investigate the performance of the Ultimate Link system. Specifically, experiments were conducted by changing the distance between the camera and an image printed on a paper surface containing embedded information denoted as “target paper” , to investigate whether the extraction process is possible or not by changing the conditions under which the target paper is captured.

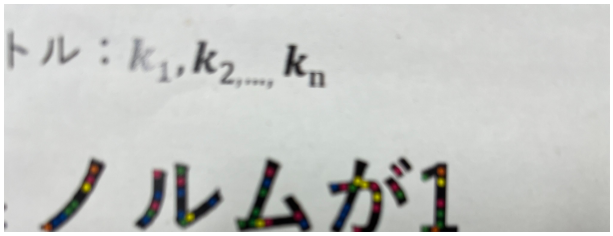
A. Common settings for each experiment

The target image is a 2339 × 1315 binary image, and two of the target strings are rendered in bold gothic font with a font size of 54pt. In the goal of our research, we aim to be able to use a variety of fonts and character sizes. However, in this experiments, we chose the font type and size so that it makes the proposed method easy to implement without considering additional technical improvements.

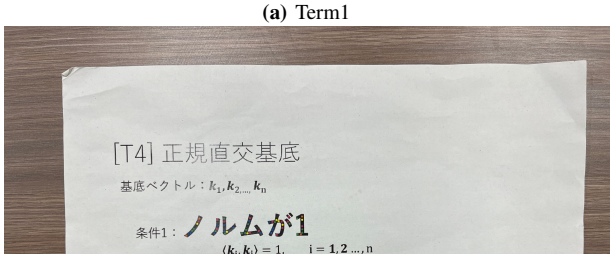
We prepared shortened URL E_1 and E_2 as embedding information, and embedded E_1 : “bit.ly/4aO6jiU” in string A and E_2: “bit.ly/3U5rYgp” in string B.

Embedding was performed on two target character strings, and images with embedded information were printed on A4 paper and captured with a smartphone camera. The smartphone used was an iPhone 13. Fig. 8(a) and Fig. 8(b) are partially enlarged images, and the full image size is 4032 × 3024.

The system was implemented using Windows 10 (CPU: Intel(R) Core(TM) i7-8565U, memory: 8GB), Python 3.11 as the programming language, and OpenCV as the image data processing environment.



(a) Term1



(b) Term4

Fig. 9: Captured images for Term1 and Term4

B. Experiment I

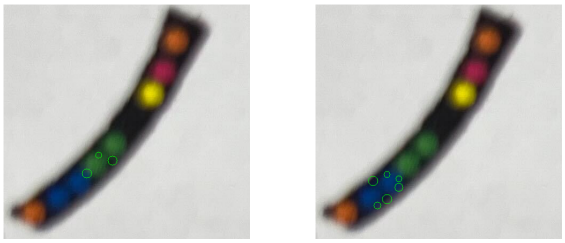
We confirmed that both strings A and B were successfully embedded and extracted, and that the URL could be recovered. The extraction process took 3.11 seconds for the string A and 3.26 seconds for the string B.

C. Experiment Ⅱ

This experiment was conducted on the extraction process by capturing the target paper from the front and changing the distance between the target paper and the camera to confirm whether the information was extracted correctly. The string A was chosen as the target string.

The distance between the target paper and the smartphone camera was set under the following four conditions : Term1, Term2, Term3 and Term4 that the distance between the target paper and the camera is about 10 cm, 15 cm, 20 cm and 30 cm.

Fig. 9 shows two examples of captured image for Term1 and Term4. The experimental results confirmed that the information was extracted correctly under the terms 2 and 3, where the camera focused on the target character string and no other



(a) Example A **(b)** Example B

Fig. 10: Circle detection failure

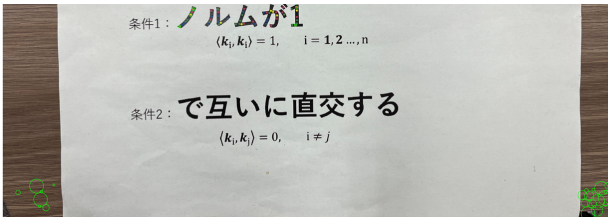


Fig. 11: Failure to identify the embedding position

images other than the target paper were captured within the camera's angle of view.

In Term 1, as shown in Fig.9(a), the camera was not focused on the target character string and it was blurred, so the detection of the colored circles failed as shown in Fig.10, and the information was not extracted correctly. In Term 4, as shown in Fig.9(b), there are objects other than the target paper in the camera's angle of view, so as shown in Fig.11, many false positives occurred when detecting the position detection marker, and parts other than the target character string remained, so the information was not extracted correctly.

D. Discussion

In Experiment I, we confirmed that both strings A and B successfully embed information. This is because the high character recognition accuracy of the Vision API's character recognition function enables it to recognize characters across the entire image even when there are multiple character strings, extract the recognition results for the target character string from the overall results, determine the rectangular area for each character, divide the character string, and perform information embedding processing for each character. In the extraction process, both strings A and B successfully identified the location of the string in which the information was embedded. This is because the position detection markers detected during the extraction process to identify the position of the character string in which the information was embedded from the captured image. Furthermore, we confirmed that both strings A and B can be correctly extracted without markers for projective transformation. This means that the alignment by Procrustes analysis based on the reference grid points and the detected bit-colored circles eliminates the need to normalize the image by a projective transformation using markers for the projective transformation.

In Experiment Ⅱ, the results show that information was successfully extracted at a distance of approximately 15 to 20 cm, which is the distance at which the camera was able to focus on the target character string and at which no objects other than the target paper appeared within the angle of view of the camera. The extraction process failed at distances of about 10 cm, where the camera could not focus on the target string and the image was blurred, and at about 30 cm, where the camera's angle of view included objects other than the target paper. Based on the results of this experiment, it is considered that a distance of approximately 15 to 20 cm between the target paper and the camera is appropriate for the current extraction processing system.

IV. Conclusions

In this paper, we propose an improved Ultimate Link method for realistic applications that can embed and extract information in target character strings in character images containing multiple character strings without using markers for projective transformation. Experimental results show that the proposed method succeeds in embedding and extracting information in two target character strings. We also investigated the characteristics of the Ultimate Link extraction process by changing the conditions under which the target paper was captured. In this experiment, we conducted the experiment on letters printed on paper with a white background, so it is necessary to conduct the experiment on posters and other postings with a colored background. In addition, since the experiment was conducted under a single lighting condition, i.e., under fluorescent light, it is necessary to conduct the experiment under multiple lighting conditions, such as under sunlight or in a dark room.

In the method used in this experiment, the images captured by the smartphone camera were once imported to a PC for extraction processing. For future practicality, we think it is necessary to develop an application that can complete the information extraction process and display a Web page within a smartphone and evaluate its practicality.

Acknowledgment

This research was supported by JSPS Grant-in-Aid for Scientific Research JP 19K12879.

References

[1] J. Noyes and K. Garland, "VDT versus paper-based text: Reply to mayes, sims and koonce," *International Journal of Industrial Ergonomics*, vol. 31, no. 6, pp. 411–423, 2003.

[2] K. Yamadera and M. Niimi, "Ultimatelink Between Characters Having a Certain Meaning in Physical Space to URL in Cyberspace with Robust Print and Scan," in *Proceedings of the 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Oct. 2023, pp. 514–518.

[3] *Vision API*, https://cloud.google.com/vision?hl=ja.

[4] S. P. H., "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1996.

Detection of Diffusion-Generated Images Using Sparse Coding

Daishi TANAKA* and Michiharu NIIMI†
* Graduate School of Computer Science and Systems Engineering,
Kyushu Institute of Technology, 680-4 Kawazu, Iizuka-shi, Fukuoka, 820-8502 Japan
tanaka.daishi864@mail.kyutech.jp
† Faculty of Computer Science and Systems Engineering,
Kyushu Institute of Technology, 680-4, Kawazu, Iizuka-shi, Fukuoka, 820-8502 Japan
niimi@ai.kyutech.ac.jp

***Abstract*—This paper proposes a method for detecting images generated by diffusion models using sparse coding. In the diffusion model, an image can be generated by removing noise from a noisy image. This different generation process from real images leads us to believe that there may be a statistical difference in pixels between the real and the generated images. Specifically, the image is divided into small patch regions, and all patch images are reconstructed using the basis image. In this process, sparse coefficients that contain many zeros are obtained using sparse coding, and features are calculated from the obtained coefficients. Then, a simple discriminator using the features as input is trained with a small number of data to discriminate the diffusion-generated images. In our experiments, we evaluated the proposed method on six datasets created using three diffusion models and two real image datasets. Experiments were also conducted to evaluate the robustness of the proposed method against JPEG compression. Experimental results show that our proposed method is sufficiently robust against JPEG compression with as few as 1800 training data.**

I. Introduction

In recent years, it has become difficult even for the human eye to distinguish between generated and real images. In addition, as AI that generates images from text, such as Stable Diffusion, has become available to everyone, the number of cases of generated images containing false information being intentionally spread has been increasing. In fact, in Japan in 2022, a false image of typhoon flooding in Shizuoka Prefecture was generated using Stable Diffusion, and in the U.S. in 2023, images of former President Trump being arrested using Midjourney was spread on X (formerly Twitter), respectively.

This ability to generate high-quality images from text has been made possible by the emergence of powerful image generation techniques such as diffusion models. This rapid progress in image generation technology has led to growing concerns about malicious use of the generated images. For this reason, there has been a lot of recent research on diffusion-generated image detection. Bammey et al.[1] proposed a detection method focusing on artifacts in the frequency domain of diffusion-generated images. Wang et al.[2] proposed a detection method using the reconstruction error between the reconstructed image and the original image obtained from a learned diffusion model. As an improvement, Luo et al.[3] proposed a very fast reconstruction-based detection method in the latent space . On the other hand, several detection methods have been proposed that use a large-scale vision-language model CLIP (Contrastive Language-Image Pre-training) as a feature extraction model[4][5]. As evaluation methods for diffusion-generated image detection, Mingjian et al.[6] presented two evaluation methods for detection methods similar to real-world scenarios using a newly constructed large dataset GenImage. As one of the methods, they proposed a method to evaluate detectors on images degraded by JPEG compression and so on. For some of the detection methods[1][4][5] shown earlier, the impact of JPEG compression on image degradation classification has been evaluated, and it can be confirmed that the identification accuracy decreases due to JPEG compression. Similarly, DIRE[2] has also been tested for robustness against JPEG compression, and while it performed well, it requires a large amount of data for detector training, which is an issue. One remaining detection method[3] has not been evaluated for classification of image degradation.

Considering these previous works, we aim to propose a new method for detecting diffusion-generated images that is robust in real-world scenarios with less data and high robustness against JPEG compression. As an approach to our detection method, we focused on the process of generating diffusion-generated images. The diffusion process gradually adds Gaussian noise to the image x_0, and finally obtains pure Gaussian noise x_T. On the other hand, the inverse diffusion process is a process that, given a pure Gaussian noise x_T, gradually removes the noise from it to finally obtain a clear image x_0. From this fact, we considered that the method of steganalysis, which discriminates whether an image has been subjected to steganography that embeds invisible information into the image or not, might be effective. In fact, a GAN detector[7] based on pixel co-occurrence matrices has been proposed with reference to steganalysis methods. In this study, we propose a method for detecting diffusion-generated images by sparse coding based on our previous work[8] on steganalysis, especially on universal steganalysis independent of embedding method. The outline of the proposed method is as follows: First, the image is divided into small patch

regions, and all the patch regions are reconstructed using the basis image. During the reconstruction, sparse coding is used to obtain sparse coefficients that contain many zeros, and features are calculated. In our proposed method, we propose two different features. By using a discriminator with each of these features as input, we expect to discriminate between diffusion-generated and real images using a small number of data. We also evaluate the robustness of the discriminator against JPEG compression.

In the following sections of this paper, we describe the proposed method in Section 2, conduct experiments and discuss the results in Section 3, and summarize in Section 4.

II. Proposed Method for Detecting Diffusion-Generated Images

In the proposed method for detecting diffusion-generated images, sparse coding is performed on patch images, features are calculated based on the calculated sparse coefficients, and a simple discriminator is used to detect diffusion-generated images. As a specific process of this method, the input image is a three-channel full-color RGB image, and the gray-scale image of the G component, which has the greatest influence on the gray-scale component, is targeted for processing, and is divided into 8×8 patch images by sliding the image one pixel at a time. For all the patch images, sparse coding is performed using 64 basis images and sparse coefficients are calculated. The features are computed based on these coefficients, and a discriminator using the features as input is used to discriminate between diffusion-generated images and natural images. We propose the following two different features as input to the discriminator.

- Sparse Coding Coefficient Statistics (SCCS)
- Sparse Coding Basis Statistics (SCBS)

For the discriminator, we used SVM, which is a popular binary classification model. The following subsections describe the detailed processing of the proposed method.

A. Sparse Coding

Sparse coding is a method of representing input data as a linear sum of a small number of bases. Fig. 1 shows a schematic diagram of patch reconstruction for sparse coding. In Fig. 1, for a patch image to be processed, sparse coding coefficients are calculated for multiple basis images of the same patch size, and a reconstructed image that approximates the input patch image is computed by calculating a linear sum of the basis images and sparse coding coefficients. In this study, the patch size is 8×8, and the patch images are extracted by shifting the G component of the image one pixel at a time. For each patch image, sparse coding is performed using 64 basis images to obtain the coefficients.

B. Sparse Coding Coefficient Statistics (SCCS)

SCCS features are computed from the coefficients based on the results of the two processes, the following six features are computed.

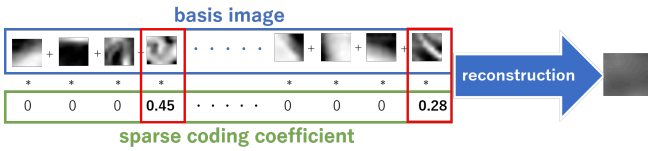


Fig. 1: Schematic of Sparse Coding

- Reconstruction error before and after sparse code shrinkage
- Average of noise percentage before and after sparse code shrinkage
- Kurtosis of the modified histogram
- Skewness of modified histogram
- Mean of modified histogram
- Standard deviation of modified histogram

In the following subsections, the above features are explained for each process.

1) Feature Extraction Process Using Sparse Code Shrinkage: Sparse code shrinkage is a method for removing Gaussian noise from non-Gaussian distributed data. The process of denoising by sparse code shrinkage is described below. First, let c_{ij} $(i = 1, 2, \ldots, PN;\ j = 1, 2, \ldots, 64)$ denote the coefficient values arranged in one dimension, where the first argument i is the patch image number, PN is the number of patch images, and the second argument j is the basis image number. The threshold value val for removing noise is calculated as follows:

$$val = \frac{\sqrt{2} \cdot \sigma^2}{d_{scale}} \tag{1}$$

where the noise variance σ is the absolute mean of the deviations of the coefficients and d_{scale} is the scale parameter, computed as the square root of the mean square of the coefficients. The coefficients after sparse code shrinkage, denoted as sc_{ij}, are calculated based on the thresholded coefficients c'_{ij}, where c'_{ij} are obtained by subtracting the threshold value from the absolute value of the coefficients c_{ij}. The calculation is defined as:

$$sc_{ij} = \begin{cases} \mathrm{sgn}(c_{ij}) \cdot c'_{ij} & \text{if } c'_{ij} > 0 \\ 0 & \text{if } c'_{ij} \leq 0 \end{cases} \tag{2}$$

where $\mathrm{sgn}(c_{ij})$ obtains the sign of the coefficients c_{ij} before taking the absolute value. The coefficients sc_{ij} after sparse code shrinkage computed in this way are used to compute the following two features:

- Reconstruction error before and after sparse code shrinkage
- Average of noise percentage before and after sparse code shrinkage

For the reconstruction error before and after sparse code shrinkage, we calculate the mean squared error (MSE) of pixel values between the shrinkage reconstruction image computed using sc_{ij}, and the original reconstruction image computed using c_{ij}. This MSE serves as the first feature. Furthermore,

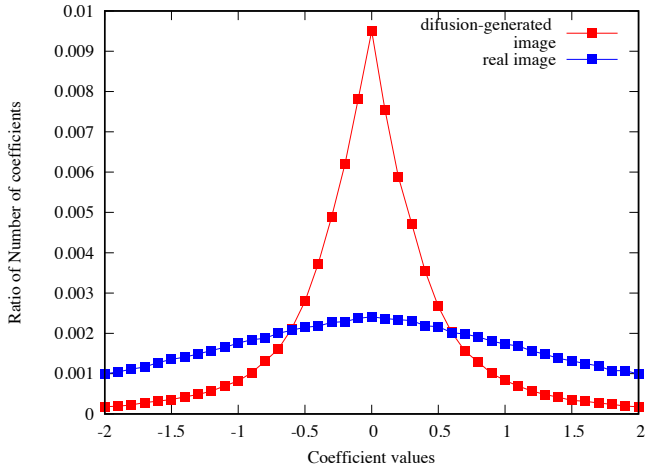


Fig. 2: Examples of modified histograms of diffusion-generated and real images

the second feature, the average noise percentage, is computed using the following formula:

$$\frac{\sum_{i=1}^{PN}\sum_{j=1}^{64}\left|\frac{sc_{ij}-c_{ij}}{c_{ij}}\right|}{N} \tag{3}$$

where N is the number of the coefficients c_{ij}. These two features aim to extract characteristics of residual Gaussian noise on diffusion-generated images.

2) Feature Extraction Process from Modified Histograms: We introduce a modified histogram of sparse coding coefficients. In this paper, a modified histogram is calculated in the following steps. Firstly, lower and upper limits are set to -20 as the lower limit and +20 as the upper limit to count the number of coefficients with, so we can count the number of sparse coding coefficients between -20 and +20. This range is divided into 4000 bins having 0.01 interval. The bin value is normalized. For example, a bin value between 0 and 0.01 represents the percentage of the number of coefficients whose values lie within that range. The value of the bin between 0 and 0.01 minus the number of zero coefficients is the new value of the bin between 0 and 0.01. This correction is made to take into account the fact that an extremely large number of zeros in the coefficients can adversely affect the difference in the feature values between the diffusion-generated image and the real images.

Fig. 2 shows examples of the modified histograms of the diffusion-generated image and the real image. In order to effectively capture the differences in the modified histograms in Fig. 2, we use the following four features: kurtosis, skewness, mean, and standard deviation.

Based on the computed modified histogram, for each n bin value $x_1, x_2, ..., x_n$, the mean $\bar{x}$ and standard deviation std are computed. The kurtosis K and the skewness S are also calculated. By using these features, we consider extracting differences in the overall statistics of the coefficients obtained by sparse coding in diffusion-generated images and real images.

C. Sparse Coding Basis Statistics (SCBS)

In SCBS features, the feature value consists of 64 values computed for each basis image. Specifically, it is calculated using the following formula:

$$\frac{\sum_{i=1}^{PN}|c_{i,j}|}{PN} \quad (j = 1, 2, \ldots, 64) \tag{4}$$

This computation yields a feature value where each value represents the sum of absolute coefficients for each basis image. By using these features, we consider extracting the difference in the coefficient statistics for each basis image between diffusion-generated and real images.

III. Experiments

The following experiments were conducted to evaluate the discrimination accuracy of SVM with SCCS and SCBS features as input features and their robustness to JPEG compression, respectively. First, we conducted experiments on the basis image and sparsity, which are considered to affect the discrimination accuracy of our method, and compared the accuracy of each. Then, we evaluated the robustness of the SVM trained on uncompressed data against JPEG compression by using test data with Q-values of 90 and 50, respectively.

A. Dataset

In this experiment, we prepared eight datasets of six diffusion-generated and two real images. As diffusion models, the following three models downloaded from Hugging Face[9] were used to generate the images and create the datasets.

- DeepFloyd/IF-II-L-v1.0
  DeepFloyd is licensed under the DeepFloyd License, Copyright (c) Stability AI Ltd. All Rights Reserved.
- emirianJR/epiCRealism[10]
- stabilityai/stable-diffusion2-1[11]

Using these three models, we generated 1000 images each for the category “dog” and “non-dog” from the same prompts. We also collected 1000 real images of the category “dog” from the Open Image Dataset V7[12] using FiftyOne, and 1000 real images including various subjects taken in RAW mode. TABLE I lists the datasets used in this experiment. Fig. 3 shows examples of images generated by each diffusion-generated model using the same prompt as input. In the experiments, the aspect ratio was fixed and the diffusion-generated images were resized to 256×256 and the real images were resized to 256 on the longest side before processing.

B. Experiments on Different Dictionaries

In this experiment, we evaluate the effect of using dictionaries consisting of different basis images on the accuracy of the proposed method. Specifically, we use dictionaries trained on two different datasets and compute the coefficients by sparse coding for each dataset. Based on the coefficients calculated here, SCCS and SCBS features are obtained, and each SVM is trained and evaluated by a 10-fold cross-validation. In the following subsections, we describe the specific ways in which the experiment is performed.

TABLE I: Dataset List

	Diffusion models	Size	Class	Number of data	Identifier
Diffusion-generated images	DeepFloyd IF	256×256	dog	1000	IFd
		256×256	non-dog	1000	IFnd
	epiCRealism	512×512	dog	1000	ECd
		512×512	non-dog	1000	ECnd
	Stable Diffusion V2-1	768×768	dog	1000	SDd
		768×768	non-dog	1000	SDnd
Real images	-	3264×2448 or 2448×3264	dog	1000	NLd
	-	long side: 1024	non-dog	1000	NLnd



Fig. 3: Images generated with "A vintage train chugs along a scenic route, nostalgic and scenic" as a prompt. From left, generated by DeepFloyd IF, epiCRealism, and Stable Diffusion V2-1.

1) Dictionaries: In this experiment, two dictionaries are prepared. Let us denote them as dictionary A (dicA) and dictionary B (dicB), respectively. DicA was trained using IFd, a dataset of diffusion-generated images. DicB is trained on NLnd, a dataset of real images. The method of creating the dictionaries is as follows. First, 100 images are randomly selected from the dataset. For each selected image, 2000 patch images of 8×8 are extracted at random locations. Thus, 64 basis images were trained from a total of 200,000 patch images. In the specific program, spams.trainDL in python were used, the parameter of the optimization relation was set to lambda1 = 0.015, and the number of training cycles was set to 100.

2) Calculation of Sparse Coding Coefficients: Sparse coding is performed using dicA and dicB, and the coefficients are computed for each. For the specific execution, spams.lasso in python was used and the parameter of the optimization relation was set to lambda1 = 0.015. Sparse coding was performed on each patch of the image by shifting the image by one pixel, and the coefficients were calculated.

3) Discriminator Learning: For each dataset, SCCS and SCBS features are computed from the coefficients obtained using each of dicA and dicB separately. For all 12 pairs of diffusion-generated images and real images, we trained SVMs using the standardized SCCS and SCBS features as inputs, respectively. The SVMs were trained and evaluated by 10-fold cross-validation. That is, the average accuracy of SVMs trained 10 times on a dataset of 2000 data sets, with 1800 training data and 200 test data randomly changed. TABLE II and III show the results for dicA and dicB using SVM, respectively. Although there is no significant difference in accuracy between the different dictionaries, there is a slight improvement in accuracy using dicA trained with IFd for all datasets. Based on these results, we trained the dictionaries using IFd in the subsequent experiments.

TABLE II: Evaluation accuracy of discriminators of SCCS features in different dictionaries

	dictionary	IFd	IFnd	ECd	ECnd	SDd	SDnd
NLd	dicA	**0.958**	**0.847**	**0.867**	**0.697**	**0.834**	**0.824**
	dicB	0.956	0.834	0.850	0.672	0.819	0.819
NLnd	dicA	**0.988**	**0.951**	**0.940**	**0.834**	**0.837**	**0.696**
	dicB	0.984	0.937	0.932	0.809	0.830	0.692

TABLE III: Evaluation accuracy of discriminators of SCBS features in different dictionaries

	dictionary	IFd	IFnd	ECd	ECnd	SDd	SDnd
NLd	dicA	**0.971**	**0.964**	**0.939**	**0.903**	**0.850**	**0.869**
	dicB	0.964	0.955	0.912	0.877	0.842	0.856
NLnd	dicA	**0.991**	**0.984**	**0.972**	**0.953**	**0.886**	**0.759**
	dicB	0.985	0.979	0.964	0.920	0.878	0.719

C. Experiments on Sparsity

In order to investigate the effect of sparsity of sparse coding coefficients on identification accuracy, we made the following changes compared to III-B.

- Increased number of training data for the dictionary
- Increased number of dictionary training iterations
- Change sparsity parameters

We newly used IFd and changed the parameter lambda1 = 0.2 during training to improve sparsity compared to III-B. The number of training cycles was increased to 2000 to learn dictionary C (dicC). The number of patches was increased fivefold to 1,000,000 patches from the 200,000 patches used in dicA. In addition, the calculation of the sparse coding coefficients was modified to improve the sparsity of the coefficients by setting lambda1 = 0.2. Table IV and table V show the results of training SVMs with SCCS and SCBS features as inputs, respectively. The accuracies shown in their tables were obtained by 10-fold cross-validation. Table IV shows that improving the sparsity of the coefficients in the SCCS features decreased the overall accuracy. Table V shows that improving the sparsity of the coefficients in the SCBS features improves the overall accuracy. This suggests that increasing the sparsity of the coefficients decreases the statistical difference of the overall coefficients between the diffusion-generated images

TABLE IV: Evaluation accuracy of discriminators of SCCS features at different sparsity

	lambda1	IFd	IFnd	ECd	ECnd	SDd	SDnd
NLd	0.015	0.958	**0.847**	**0.867**	0.697	**0.834**	**0.824**
	0.2	**0.963**	0.846	**0.867**	**0.733**	0.763	0.796
NLnd	0.015	**0.988**	**0.951**	**0.940**	**0.834**	**0.837**	**0.696**
	0.2	0.981	0.891	0.908	0.752	0.776	0.691

TABLE V: Evaluation accuracy of discriminators of SCBS features at different sparsity

	lambda1	IFd	IFnd	ECd	ECnd	SDd	SDnd
NLd	0.015	0.971	0.964	**0.939**	0.903	0.850	0.869
	0.2	**0.978**	**0.967**	0.934	**0.906**	**0.874**	**0.900**
NLnd	0.015	**0.991**	**0.984**	0.972	**0.953**	0.886	0.759
	0.2	0.989	0.983	**0.976**	0.929	**0.926**	**0.800**

and real images, while increasing the statistical difference of the coefficients for each basis image.

D. Evaluation Experiments on JPEG Compression

In this experiment, we evaluated the robustness of JPEG compression in the proposed method. For both SCCS and SCBS features, we trained SVM by computing features using dicC under the same conditions as in III-C. Then, we applied JPEG compression to 200 pieces of test data used for evaluating the accuracy of SVM with Q-values of 90 and 50, and evaluated SVM using each JPEG compressed test data. In other words, we conducted a realistic evaluation experiment on unknown images with JPEG compression. In order to obtain the actual accuracy, we computed the average accuracy of 10 times while randomly changing the training and test data. The results are shown in TABLE VI and VII. The results in TABLE VI and VII show that both features are sufficiently robust to JPEG compression. In particular, the SCCS features maintain their accuracy even with higher JPEG compression.

E. Discussion

As shown in Table VI, the SVM with SCCS features as input does not lose accuracy when comparing the results for the uncompressed test data and the JPEG-compressed test data with Q value = 50. Therefore, it is hypothesized that the distribution of coefficients obtained by sparse coding, i.e., the modified histogram, is less affected by high-frequency components.

To investigate the influence of high-frequency components in the distribution of coefficients obtained by sparse coding, we conducted comparison experiments between high-frequency images containing many high-frequency components and JPEG-compressed high-frequency images in which the information on high-frequency components was reduced by applying JPEG compression to the high-frequency images. Firstly, 100 high-frequency images of 256×256 were collected for the NLnd dataset. Specifically, for all 1000 images in NLnd, a region of 256×256 was cut out by sliding the

TABLE VI: Evaluation accuracy of JPEG compression in discriminators of SCCS features

	Q-value	IFd	IFnd	ECd	ECnd	SDd	SDnd
NLd	100	0.962	0.842	0.858	0.752	0.766	**0.797**
	90	0.953	0.832	0.860	0.748	0.778	0.791
	50	**0.967**	**0.897**	**0.887**	**0.801**	**0.782**	0.788
NLnd	100	0.975	0.893	**0.905**	0.743	0.783	0.698
	90	0.972	0.899	0.902	**0.749**	**0.788**	**0.703**
	50	**0.976**	**0.911**	0.895	0.715	0.782	0.701

TABLE VII: Evaluation accuracy of JPEG compression in discriminators of SCBS features

	Q-value	IFd	IFnd	ECd	ECnd	SDd	SDnd
NLd	100	**0.983**	**0.954**	**0.936**	**0.909**	**0.874**	**0.898**
	90	0.972	0.947	0.932	0.908	0.867	0.895
	50	0.967	0.947	0.919	0.899	0.782	0.877
NLnd	100	0.989	0.989	**0.979**	0.918	**0.931**	0.807
	90	**0.990**	**0.990**	**0.979**	**0.924**	0.925	**0.819**
	50	0.989	0.985	0.976	0.918	0.876	0.778

image by 256 pixels in height and width, and the portion of the image that was less than 256×256 at the edge was ignored. In other words, since the image size of NLnd is 3264px on the long side and 2448px on the short side, 108 images can be collected from a single image, or 108,000 256×256 images in total. In addition, 100 JPEG-compressed images were created for these 100 images, with Q value = 50. The average of the mean square error of the pixel values of the uncompressed and JPEG-compressed images for all 100 images was calculated to be 58.4. This value indicates that considerable image degradation has occurred due to JPEG compression, i.e., the 100 uncompressed images collected are high-frequency images that contain many high-frequency components. Sparse coding was performed on these 200 images using dicC under the same conditions as in III-D to obtain the coefficients. Using these coefficients, a modified histogram was created to see the distribution of the coefficients. In this experiment, Bhattacharyya distance was used as a measure for comparing the similarity of histograms. Bhattacharyya distance D_B is defined as

$$D_B(p,q) = -\ln\left(\sum_i \sqrt{p(i)q(i)}\right) \tag{5}$$

where $p(i)$ and $q(i)$ denote the probability values of the i-th bin of two histograms, respectively. This distance quantitatively evaluates the degree of overlap of the histograms; the greater the overlap, the smaller the value. In other words, a value closer to 0 indicates that the histograms are in agreement. For 100 pairs of uncompressed and JPEG-compressed high-frequency images, the mean Bhattacharyya distance of the histograms is 1.62×10^{-2} and the variance is 1.41×10^{-5}. To show that the histograms are visually close, the histogram computed from the high-frequency image used in this experiment is shown in Fig. 4. Based on these results, it can be said that the distribution of sparse coding coefficients is fairly insensitive

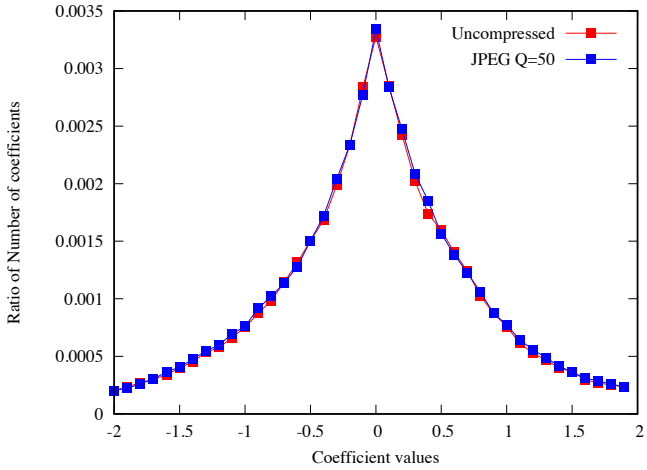


Fig. 4: Modified histograms of the uncompressed and JPEG-compressed high-frequency images

to high-frequency components.

To confirm the same thing in the 6-dimensional feature space in the SCCS features, we conducted an experiment using SVM with SCCS features as input to discriminate between uncompressed images and JPEG compressed images for the 200 images mentioned earlier. The average accuracy obtained from the 10-fold cross-validation was 0.515, confirming that the SCCS features are difficult to discriminate. In other words, as with the distribution of coefficients in the modified histogram, the SCCS features are not easily affected by high-frequency components.

From the previous experiments, it is confirmed that the sensitivity of SVM with SCCS features as input is very low to changes in high-frequency components. This means that the method using SCCS features would be able to maintain the discrimination accuracy in the robustness experiments against JPEG compression.

IV. Conclusion

In this paper, we proposed a new method for detecting diffusion-generated images from real images using features extracted by sparse coding. One of the most notable features of this method is that it is sufficiently robust to JPEG compression with a small amount of training data. In particular, high robustness was confirmed for the SCCS features. Based on these results, we conducted additional experiments and confirmed that the distribution of sparse coding coefficients was not easily affected by high-frequency components, and that high robustness with respect to JPEG compression was obtained by computing features using the distribution of sparse coding coefficients.

For future prospects, we believe that new features are needed that can achieve higher accuracy than the current ones on a wide range of data sets while maintaining high robustness against image degradation such as JPEG compression by using the distribution of sparse coding coefficients.

References

[1] Q. Bammey, "Synthbuster: Towards Detection of Diffusion Model Generated Images," *IEEE Open Journal of Signal Processing*, 2023.

[2] Z. Wang, J. Bao, W. Zhou, *et al.*, *DIRE for Diffusion-Generated Image Detection*, 2023. arXiv: 2303.09295v1 [cs.CV].

[3] Y. Luo, J. Du, K. Yan, and S. Ding, "LaRE^2: Latent Reconstruction Error Based Method for Diffusion-Generated Image Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 17 006–17 015.

[4] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, *Raising the Bar of AI-generated Image Detection with CLIP*, 2024. arXiv: 2312.00195 [cs.CV].

[5] S. A. Khan and D.-T. Dang-Nguyen, "CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection," in *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 2024, pp. 1006–1015.

[6] M. Zhu, H. Chen, Q. Yan, *et al.*, *GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image*, 2023. arXiv: 2306.08571 [cs.CV].

[7] L. Nataraj, T. M. Mohammed, S. Chandrasekaran, *et al.*, *Detecting GAN generated Fake Images using Co-occurrence Matrices*, 2019. arXiv: 1903 . 06836 [cs.CV].

[8] M. Niimi and H. Noda, "An application of Sparse Code Shrinkage to image steganalysis based on supervised learning," in *2011 18th IEEE International Conference on Image Processing*, IEEE, 2011, pp. 1941–1944.

[9] *Hugging Face*. [Online]. Available: https://huggingface.co/.

[10] C. O. RAIL-M, *CreativeML Open RAIL-M license*, Hugging Face. [Online]. Available: https://huggingface.co/spaces/CompVis/stable-diffusion-license.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 10 684–10 695.

[12] *Open Images Dataset V7*. [Online]. Available: %5Curl%7Bhttps://storage.googleapis.com/openimages/web/index.html%7D.

Generation of Target Speech with Speaker Individuality Based on Accent Conversion for English Pronunciation Learning

Rei Hamakawa* and Michiharu Niimi†
* Graduate School of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: hamakawa.rei112@mail.kyutech.jp
† Faculty of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: niimi@ai.kyutech.ac.jp

***Abstract*—This paper first provides an overview of the English pronunciation learning support tool. The tool aims to use "accent-modified speech that retains the learner's voice quality" as the "target speech." We propose a new conversion model based on conventional methods for this accent conversion. Specifically, we improve the conventional LSTM-based DNN model for accent conversion by adopting a transformer-based model. Our experiments investigated the model's ability to handle the unique katakana pronunciation characteristic of Japanese speakers. The results confirmed the effectiveness of the proposed conversion method, although challenges remain, such as the scarcity of Japanese speech data and the need to improve the accuracy of speaker identity retention.**

I. Introduction

According to a report by EF Education First[1], Japan ranks 87th out of 113 non-native English-speaking countries and regions, and 15th out of 23 in Asia, in other words, Japan is classified as a "low proficiency" level. Therefore, the reform of English education is an urgent issue. The new curriculum guidelines were introduced in 2020 for elementary schools, 2021 for junior high schools, and 2022 for high schools. These guidelines aim to develop communication skills through a balanced learning approach to listening, reading, speaking, and writing in English.

However, it is extremely challenging to provide appropriate instruction to each student due to the insufficient English proficiency of English teachers and the shortage of Assistant Language Teachers (ALTs). From several reports with respect to English teacher quality, it is exceedingly difficult to provide appropriate instruction in speaking to each student. Pronunciation instruction is particularly important for Japanese students, as there are significant differences between Japanese and English pronunciation. In pronunciation instruction, simply listening to the target audio makes it difficult for beginner learners of English to recognize the differences between their own pronunciation and the target pronunciation. If we can provide ideal audio that makes the differences between the learner's pronunciation and the target pronunciation easily recognizable, and visually represent these differences using phonetic symbols or other means, it can lead to more effective learning. Furthermore, developing this as a PC-based system (English pronunciation learning support tool) will enable individual learning.

The concept of a "golden speaker," which refers to a speaker with a voice similar to second-language (L2) learners, has been applied in computer-assisted pronunciation training (CAPT) for L2 learners[2][3]. Because the "golden speaker" has the same voice quality as them, it serves as an ideal target for pronunciation training compared to native speakers' voices. Nagano et al.'s [4] study explored the potential of using modified learner speeches for pronunciation training. In this study, Japanese learners were divided into two groups: one trained to imitate the speech of a standard English speaker, and the other trained to imitate their own speeches pre-modified to match the prosody of the standard English speaker. The results showed that the latter group's post-training speech was rated as closer to native-like than that of the first group. Hence, we propose using accent-converted speech that retains the learner's voice quality as the target speech in pronunciation instruction. We expect that this approach allows learners to hear the correct pronunciation in their own voice quality, which makes it easier for them to recognize the differences between their own pronunciation and the target pronunciation.

Based on this concept, we're trying to develop an English pronunciation learning support tool that generates golden speaker's speech with the L2 learner's voice quality and the correct pronunciation as target speech, shows pronunciation differences visually, and provides feedback. The golden speaker's speech with the learner's voice quality and the correct pronunciation is generated using accent conversion technology, a type of speech processing technology. The input to the accent converter is the L2 learners' spoken English, and this technology adjusts the learners' accent to the native accent while preserving the speaker's voice quality. Traditional accent converters use an LSTM-based DNN model, but we improve it by adopting a transformer-based model. While DNNs require a large amount of training data, Japanese data have not

traditionally been included as processing targets. This paper collects a small amount of Japanese English pronunciation data and conducts experiments.

First, in section 2, we describe the proposed tool. Then, in section 3, we discuss the generation of model speech with the learner's voice quality. Section 4 presents the experimental results of accent conversion using Japanese data, and finally, section 5 concludes the paper.

II. Proposed English Speech Learning Support Tool

In this section, we outline the concept of an English speech learning support tool designed to address the issues identified in the pronunciation learning process.

A. The Process of Learning Pronunciation

To acquire native English pronunciation, it is necessary to work on narrowing the gap between one's pronunciation and that of native English speakers. This requires first learning the fundamental knowledge of pronunciation, such as phonetic symbols and phonological changes, and then engaging in actual pronunciation training. In other words, we assume that the cycle: recording own pronunciation, comparing it with native audio, and then receiving the feedback is needed, and this cycle represents what we consider the ideal pronunciation training.

B. Current Issues

Typically, when learning pronunciation, learners first listen to the recorded pronunciation that comes with their study materials, imitate it, and practice. They progress from words to phrases and then to sentences. However, there are phonological changes such as linking, reduction, and flapping, leading to pronunciations that often differ from the phonetic symbols provided. Additionally, the recorded pronunciations in learning materials are usually voiced by a single male or female native speaker for each word, sentence, or passage. Therefore, if the voice quality of the English learner differs significantly from that of the recorded speaker, it becomes difficult to recognize the differences in pronunciation. Moreover, learners cannot receive feedback on how their pronunciation differs from the ideal pronunciation, which makes it difficult for many Japanese learners to overcome the typical Japanese accent, known as the katakana accent in Japan.

C. Overview of the Support Tool

In our support tool, users first press the record button, read out a word or sentence of their choice, and records it. The system then converts the recorded speech to match a native accent while preserving the user's voice quality. This generates a speech that makes it easier to recognize pronunciation differences. They can play back and compare their own speech with the converted speech. Furthermore, the phonetic symbols of the their recorded speech and the generated speech are extracted and displayed. This provides visual feedback on the pronunciation differences. The differences between the extracted phonetic symbols of the their pronunciation and the target pronunciation are highlighted in red. This allows beginner learners to recognize the differences between their own pronunciation and target pronunciation more easily, As a result of this, pronunciation learning becomes more efficient even for self-study.

Using phonetic symbols to show differences is only one effective method in pronunciation learning. Therefore, future developments of this system include implementing features to visualize differences in characteristics beyond phonemes, such as prosody.

III. Generation of Model Speech with the Learner's Voice Quality

Using accent conversion technology with DNN models, we generate target speech with the learner's voice quality to make pronunciation differences easily recognizable.

A. Conventional Method

Waris Quamer et al. [5] proposed a method that does not require reference speech for accent conversion. This method is based on Tacotron2[6] model that generates speech from text. Fig. 1 shows an overview of this system.

The system consists of six models: acoustic model, accent encoder, seq2seq model (translator), speaker encoder, seq2seq model (synthesizer), and vocoder. The flow of Waris Quamer et al.'s accent conversion method is as follows:

1) First, the acoustic model generates a phonetic-posteriorgram (PPG) that represents the linguistic content independent of the speaker. In Waris Quamer's method, the acoustic model was implemented using Kaldi[7]. Typically, the PPG is high-dimensional, but in this method, bottleneck features from the layer prior to the final softmax layer of the acoustic model are used instead of the PPG to extract low-dimensional linguistic features. The PPG has dimensions of time frames $\times$ 256.
2) The accent encoder generates a fixed-dimension embedding vector, which represents the accent from the native speech. This model was trained to maximize the cosine similarity between speech samples with the same accent. The accent embedding is a 256-dimensional feature which captures a speaker's accent.
3) The PPG and accent embedding are fed into the seq2seq model (translator), which adjusts the PPG based on the accent embedding to convert L2 PPG into L1 PPG.
4) The speaker encoder generates a fixed-dimension embedding vector which represents the speaker identity. This model is trained as speaker-verification and maximize the cosine similarity between utterances from the same speaker. The speaker embedding is a 256-dimensional feature which capture a speaker's identity.
5) The converted L1 PPG and speaker embedding are fed into the seq2seq model (synthesizer), which adjusts based on the speaker embedding to generate a mel-spectrogram with L1 accent while preserving L2 speaker's voice quality.

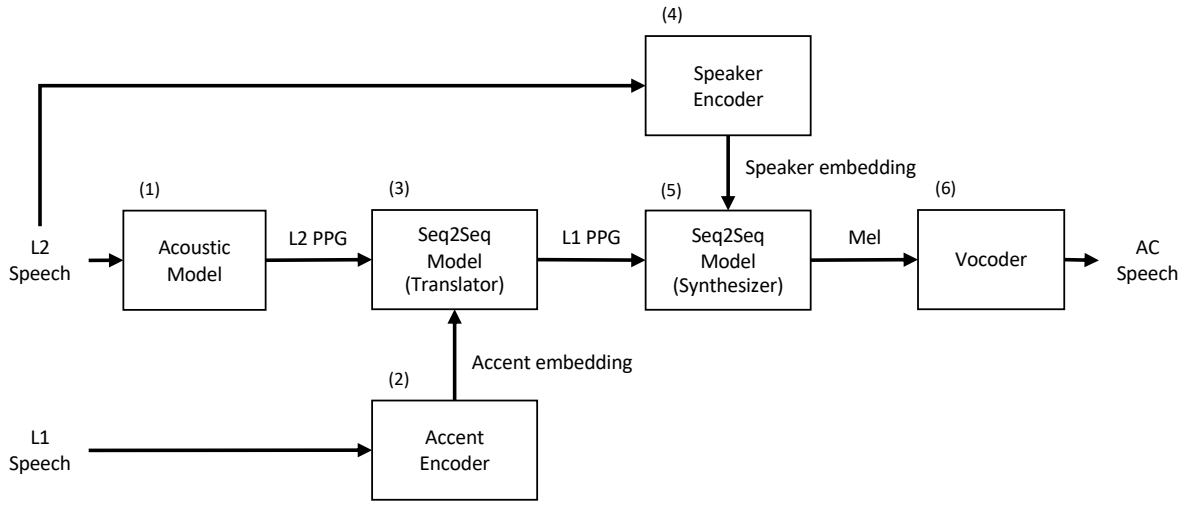


Fig. 1. Overview of conventional system

6) The generated mel-spectrogram is fed into the vocoder to generate the speech with L1 accent and L2 speaker's voice quality.

B. Proposed Method

In the conventional method, the translator and synthesizer are LSTM-based models inspired by Tacotron2 whose architecture uses a location-sensitive attention mechanism, which is prone to fail with long utterances. Waris Quamer et al.[5] suggest that another direction for future research is to make the model robust for synthesizing long utterances.

Additionally, since the synthesizer is trained on data from 24 speakers and generates mel-spectrograms from PPGs, it produces high-precision outputs for the voice qualities of those 24 speakers. However, through several experiments, we found that it does not sufficiently retain the voice quality when generating mel-spectrograms for speakers not included in the training data. In the experiments, training the Tacotron2-based model using two NVIDIA Tesla V100 GPUs took a long time due to the use of LSTMs in both the encoder and decoder. LSTMs require sequential processing because the output at each time step depends on the output from the previous time step. This dependency makes parallel processing impossible. It causes long training time, therefore, fine-tuning with the conventional method is impractical. Based on the above considerations, we identified the following issues in prior research:

- Synthesis of long utterances
- Handling of speakers not included in the training data
  - Fine-tuning could be a solution, but their model makes this very time-consuming.

To address these issues, we propose replacing the translator and synthesizer with transformer-based models. The transformer uses self-attention mechanisms that allow it to retain and process information from earlier parts of long sequential data. This capability makes it robust for long utterance synthesis. Additionally, self-attention mechanisms enable parallel processing. They significantly reduce training time compared to Tacotron2 and make fine-tuning more feasible. We replace the Tacotron2-based translator and synthesizer models with transformer-based models, pre-train on data from numerous speakers, and fine-tune with a small amount of data from a single L2 learner. This approach aims to generate personalized "golden speaker" speech and solve the problems of the conventional model. The other models remain unchanged from the conventional method, and we use pre-trained models for experiments.

Fig. 2 shows the transformer-based translator model. Unlike the standard Transformer, which takes a sequence of words as input and converts each word to a vector, the input to the proposed model is already vectorized PPG, which is processed directly. The differences from a standard transformer are summarized below:

- Adding Tacotron2's pre-net to the input side of the encoder and decoder
- Concatenating the encoder's output with the accent embedding before inputting it to the decoder
- Adding linear projections to generate L1 PPG and predict stop tokens
- Adding a post-net, composed of five CNN layers, to the output side

The pre-net from Tacotron2 functions as a bottleneck. It compresses and extracts the input information, which contributes to learning efficiency. To convert the PPG into the target pronunciation based on the accent embedding, the encoder's output is concatenated with the accent embedding and input into the decoder. The stop token controls the output termination timing. The post-net improves PPG quality by predicting and adding residuals to the predicted PPG. The structure of the encoder and decoder follows the standard

transformer design. Multihead Attention consists of multiple attention layers which allows simultaneous focus on multiple parts. The Masked Multihead Attention in the decoder functions similarly but operates with masked information. The Feed Forward Network (FFN) is a layer that performs linear transformations.

The synthesizer also has a transformer-based structure similar to the translator. It takes PPG and speaker embedding as inputs and outputting a mel-spectrogram with dimensions of time frames $\times$ 80 at the final output layer.

Since the transformer-based model uses self-attention mechanisms, it can process long sequential data in parallel. This reduces training time compared to the LSTM-based Tacotron2 and makes fine-tuning easier. We aim for high-precision accent conversion by pre-training the translator and synthesizer on English speech data from many speakers, then fine-tuning with a small amount of data from one speaker. In the translator training, the model trained without Japanese speaker data is fine-tuned with a small amount of Japanese speaker data to adapt to the unique katakana pronunciation of Japanese speakers. In synthesizer training, fine-tuning is performed using one speaker's English utterances to retain the user's voice quality sufficiently. Although it is ideal to fine-tune using English utterances, since English beginners may not be able to speak English well, we also attempt fine-tuning using only Japanese utterances to synthesize the "golden speaker" speech.

IV. Experiment

A. Tranning Data

For the pre-training of the translator, we use the CMU ARCTIC [8] and L2-ARCTIC [9] datasets. The CMU ARCTIC database is constructed as a phonetically balanced single-speaker database of American English. L2-ARCTIC is a non-native English speech corpus designed for research in speech conversion, accent conversion, and mispronunciation detection. This corpus includes English read speech from 24 non-native speakers whose first languages (L1) are Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese. In previous experiments using conventional methods, Japanese speakers were not included in the training data, and it is anticipated that these methods would not accommodate the unique katakana pronunciation characteristic of Japanese speakers. Therefore, in this study, to apply accent conversion to the distinctive katakana pronunciation of Japanese speakers, we collect and use the same reading sentences from a Japanese speaker as fine-tuning data for training.

For the pre-training of the synthesizer, we use the train-other-500 subset of the Librispeech dataset[10]. Additionally, to sufficiently maintain speaker characteristics and synthesize the speech of a "golden speaker," we collect data from two Japanese speakers for fine-tuning. The data from each speaker is used independently for fine-tuning. We conduct experiments using only the English utterances and only the Japanese utterances recorded from each speaker.

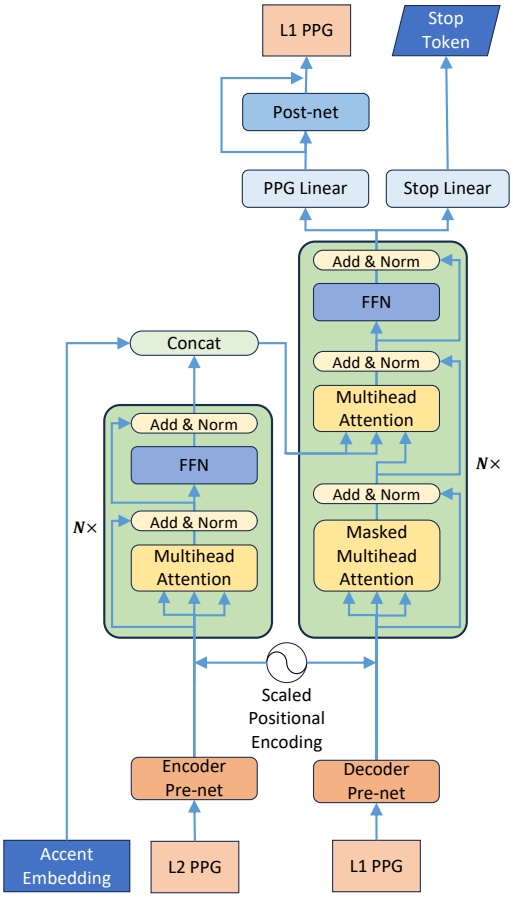


Fig. 2. The proposed model

TABLE I
TRANSLATOR EXECUTION GPU ENVIRONMENT SPECIFICATIONS

Component	Specification
GPU	NVIDIA GeForce RTX 4070 Ti
GPU Memory	12G

B. Experimental Environment and Model Parameters

Table I and Table II show the GPU of experimental environments for the translator and synthesizer, respectively. Due to the memory capacity limitations of the GPUs used, we set the number of layersN of the encoder and decoder to 1 and the number of heads to 2, as illustrated in Fig. 2. The same configuration is applied to the synthesizer.

C. Training

1) Pre-Training with Data Excluding Japanese Speakers: In the translator experiments, pre-training is conducted using the CMU ARCTIC and L2-ARCTIC datasets, excluding Japanese speaker data. The input data includes the CMU ARCTIC speakers SLT, CLB, and RMS, as well as the non-native speakers from L2-ARCTIC (excluding NJS, YKWK, TXHC, and ZHAA). Unlike conventional methods that used only a

TABLE II
SYNTHESIZER EXECUTION GPU ENVIRONMENT SPECIFICATIONS

Component	Specification
GPU	NVIDIA GeForce RTX 4090
GPU Memory	24G

single native speaker's data from CMU ARCTIC as the target pronunciation for accent conversion, this study additionally incorporates the data of three native speakers used as inputs into the target pronunciation set. This expansion increases the combinations of input and target pronunciations. As a result of this, the variety of the training data is enhanced. The Adam optimizer is employed for training, with a batch size of 32 and 200 epochs.

2) Fine-Tuning with Japanese Speaker Data: The model trained with data excluding Japanese speakers is fine-tuned using data from a single Japanese speaker. The Adam optimizer is employed for fine-tuning, with a batch size of 32 and 100 epochs.

3) Pre-training of the Synthesizer: In the synthesizer experiments, we initially use approximately 30 utterances per speaker from the Librispeech train-other-500 subset for pre-training. The synthesizer is trained independently. Prior to this, PPG and speaker embeddings are generated using the acoustic model and speaker encoder, respectively. These are then input into the Synthesizer, which adjusts speaker characteristics based on the Speaker Embedding and predicts the mel-spectrogram from the PPG and speaker embedding. The Adam optimizer is employed for training, with a batch size of 24 and 200 epochs.

4) Fine-Tuning of the Synthesizer: The model pre-trained with Librispeech data is fine-tuned using data from Japanese speakers. The Adam optimizer is employed, and fine-tuning is performed over 400 epochs. Experiments are conducted using 25, 50, and 75 utterances of English speech and 25, 50, and 75 utterances of Japanese speech to investigate the impact of the amount of data used for fine-tuning on the accuracy of maintaining speaker's voice quality. The Adam optimizer is employed for these experiments, with a batch size of 24 and 400 epochs.

D. Evaluation

1) Evaluation Method for the Translator: To assess the extent to which Japanese read-aloud speech can be converted to a native-like accent, we focus on the number of correctly accent-converted words within a single sentence, that is, the number of correctly accent-converted words in a sentence is divided by the total number of words in the sentence. For n sentences, the average of these ratios is taken as the evaluation score. Note that the evaluation does not include metrics for audio quality or fluency.

2) Evaluation Data for the Translator: For evaluation data, we use 50 sentences read aloud by two non-native speakers whose native languages are Arabic and Mandarin, and two non-native speakers whose native language is Japanese.

TABLE III
EVALUATION RESULTS

Native Language	Fine-tuning	
	Before	After
Arabic	0.90	0.79
Mandarin	0.89	0.83
Japanese (speaker1)	0.63	0.81
Japanese (speaker2)	0.65	0.77

3) Evaluation Results for the Translator: Table III shows the evaluation results of the model before and after fine-tuning. Before fine-tuning, the model could convert about 90% of the words for speakers whose native languages are Arabic and Mandarin, but it could convert fewer than 65% of the words for speakers whose native language is Japanese. After fine-tuning, approximately 80% of the words were converted for speakers of all native languages, including Arabic, Mandarin, and Japanese. Fine-tuning to adapt to the unique katakana pronunciation of Japanese speakers reduced the conversion accuracy for speakers of Arabic and Mandarin. Inappropriate conversions included significantly different pronunciations of words after conversion, words that could not be recognized as words, and omissions in the output.

Additionally, we evaluated the pre-fine-tuning model using only 19 sentences composed of 11 to 13 words from the evaluation sentences. The results are shown in Table IV. The 50 sentences used for evaluation consisted of a maximum of 13 words. Comparing Table III and Table IV, it can be said that the evaluation scores for speakers whose native languages are Arabic and Mandarin are not affected by the number of words.

4) Evaluation Results for the Synthesizer: Currently, fine-tuning of the synthesizer has only been conducted on data from two speakers, so quantitative and qualitative evaluations of speaker identity retention accuracy and other metrics have not been performed. Of the two speakers tested, one produced speech that seemed to better retain speaker identity when fine-tuned with both English and Japanese utterances compared to the conventional model. Additionally, it was observed that the number of data samples appeared to affect the audio quality.

For the other speaker, the fine-tuned model did not exhibit superior speaker identity retention compared to the conventional model. No significant impact of fine-tuning with either English or Japanese utterances was observed; both sets of data resulted in speech with a similar level of speaker identity retention. Moreover, the time required for fine-tuning the synthesizer in the experimental environment shown in Table II is presented in Table V. Even when fine-tuning with 75 data samples over 400 epochs, the process was completed in approximately 6 minutes. This demonstrates that fine-tuning a transformer-based model is feasible in practice.

E. Discussion

In the pre-fine-tuning model for the translator, only 63% of the words in the data from speakers whose native language is

TABLE IV
EVALUATION RESULTS FOR LONG SENTENCES

Native Language	Before Fine-tuning
Arabic	0.92
Mandarin	0.89
Japanese (speaker1)	0.52
Japanese (speaker2)	0.59

TABLE V
TIME REQUIRED FOR SYNTHESIZER FINE-TUNING

Number of Data	Time [minutes]
25	4.52
50	5.54
75	6.37

Japanese were appropriately converted. This is likely due to the significant differences in pronunciation characteristics between the speakers in the training data and the unique katakana English pronunciation of Japanese speakers. This differences is thought not to lead to expected effective conversions. After fine-tuning, this issue was mitigated, with approximately 80% of the words being appropriately converted. However, since the fine-tuning data included only one Japanese speaker, and pronunciations can vary even among Japanese speakers, some words were still not converted correctly. To achieve a level of accuracy usable by Japanese beginners in English, it is essential to increase the training data to comprehensively cover the unique katakana pronunciations of Japanese speakers.

The transformer-based model proposed in this study did not show a decrease in evaluation scores for longer sentences compared to the overall score of the 50 sentences. This result indicates that the proposed method's conversion accuracy is not affected by sentence length. This addresses the issue reported in conventional LSTM-based models where longer sentences are more prone to conversion failures. In the pre-fine-tuning model, the evaluation scores for sentences composed of 11 to 13 words showed a decline for Japanese speakers, which is an acceptable result given that the training data did not include Japanese speakers.

V. CONCLUSION

This paper first presented an overview of the English pronunciation learning support tool we aim to develop. The tool converts the learner's pronunciation into a target pronunciation. The target pronunciation is designed to retain the learner's voice quality while having the correct accent. For accent conversion, we proposed a transformer-based DNN model based on conventional methods. This study confirmed that fine-tuning with data from one Japanese speaker improved conversion accuracy for the unique katakana English pronunciations of Japanese speakers. Furthermore, switching to a transformer model helped address the issue found in conventional LSTM-based models, in which long sentences often resulted in conversion failures.

However, the conversion accuracy is still not at a satisfactory level, likely due to insufficient training data. Although we have not yet evaluated the retention of speaker characteristics by the synthesizer, it is expected that the characteristics of the speaker's voice, such as pitch, may influence retention.

Future challenges include collecting a large amount of English reading data from speakers whose native language is Japanese, conducting experiments to improve the accuracy of speaker identity retention, extracting phonetic symbols from English speeches, and building the actual English speech learning support tool.

REFERENCES

[1] *The world's largest ranking of countries and regions by english skills*, https://www.ef.com/wwen/epi/, (Accessed on 07/02/2024).

[2] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, no. 10, pp. 920–932, 2009, Spoken Language Technology for Education.

[3] S. Ding, C. Liberatore, S. Sonsaat, *et al.*, "Golden speaker builder – an interactive tool for pronunciation training," *Speech Communication*, vol. 115, pp. 51–66, 2019, ISSN: 0167-6393.

[4] K. Nagano and K. Ozawa, "English speech training using voice conversion," in *Proc. First International Conference on Spoken Language Processing (ICSLP 1990)*, 1990, pp. 1169–1172.

[5] W. Quamer, A. Das, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Zero-Shot Foreign Accent Conversion without a Native Reference," in *Proc. Interspeech 2022*, 2022, pp. 4920–4924.

[6] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[7] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[8] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Workshop on Speech Synthesis (SSW 5)*, 2004, pp. 223–224.

[9] G. Zhao, S. Sonsaat, A. Silpachai, *et al.*, "L2-arctic: A non-native english speech corpus," in *Proc. Interspeech*, 2018, pp. 2783–2787.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

Proposal of Blind Extractable Additive Video Watermarking Method

Nao Harada, Rinka Kawano, Masaki Kawamura †
Graduate School of Sciences and Technology for Innovation, Yamaguchi University
† E-mail: m.kawamura@m.ieice.org

***Abstract*—To improve the quality of stego video, it is necessary to reduce the total amount of information embedded in the video as a watermark. This paper proposes two techniques. The first technique is distributed watermarking. A watermark is embedded in some of the blocks. The same watermark is embedded in multiple frames. By time averaging the extracted watermarks, error can be reduced. The second technique is detection signal embedding. In order to detect stego frames, a detection signal is introduced. This allows the watermark to be embedded in any frame. In addition, resistance to frame removal attacks can be provided. We embedded a watermark in videos and evaluated its performance using normalized correlations of the extracted watermarks. We also evaluated the quality of the stego video. The results showed that compared with the DCT-DWT method, the performance of the proposed method was better under different types of attacks, including compression attacks and frame removal attacks. In addition, the image quality was high with a PSNR of over 41 dB.**

I. Introduction

Digital watermarking is a technique for the covert embedding of information into digital content such as still images, videos, and music. Depending on their purpose, watermarking methods can be divided into robust watermarking [1] and fragile watermarking [2]. Robust watermarking is used to protect copyrights and is resistant to compression and content modification. Fragile watermarking is for detecting content tampering and is therefore not resistant to compression or attacks. Watermarking methods include non-blind watermarking [3] and blind watermarking [4]. The former requires the original content to extract the watermark. The latter does not. Since non-blind methods are not practical in most cases where it is difficult to obtain the original image or where there is no permission to access the original image, blind methods are preferred.

In this paper, we focus on robust video watermarking. Similar to watermarking still images, there are two embedding domains for video frames: the spatial domain [5] and the frequency domain [6], [7]. In general, it is known that embedding in the frequency domain is more resistant to attacks than embedding in the spatial domain. Sang *et al.* [7] proposed a robust DCT-DWT method that embeds the watermark in a domain combining the discrete cosine (DCT) and discrete wavelet (DWT). However, their method was not practical because it was not blind. In addition, video embedding techniques also include additive embedding [6], [7] and quantization index modulation (QIM) [8]. Video is computationally intensive because it contains a lot of information. Therefore, it is desirable to be able to embed with a simple process such as additive embedding. However, existing methods using additive embedding are mostly non-blind. The method of Yokota and Kawamura [6] was able to extract additively embedded watermarks using blind source separation (BSS). During the extraction, stego frames can be detected from signals separated by BSS using the correlation between signals. However, the stego frames and the non-embedded frames were adjacent to each other. As a result, the video image flickered during playback. Therefore, it was easy to detect that the watermark existed. The method was also vulnerable to the frame removal attack.

We propose two techniques that improve the quality of stego video. The first one is to reduce the number of watermarks embedded in a single frame. The proposed method embeds one watermark per frame, while Yokota and Kawamura's method [6] embeds six or more of the same watermark in a frame. In this technique, the frame is divided into blocks. A watermark is embedded in some of the blocks to suppress flickering. Because the extracted watermark contains errors, the same watermark is embedded in multiple frames. By time averaging, a watermark with fewer errors can be achieved. The other technique is to introduce a detection signal to find stego frames. By embedding a detection signal in stego frames, the frames can be detected even under frame removal attacks. In addition, the frames can be selected at arbitrary rather than fixed intervals (e.g., one every three frames). These two techniques can reduce the total amount of information embedded as a watermark in a video. They also eliminate the need for complex BSS processing. The watermark can be extracted by simply taking the difference between two adjacent frames of the video.

To demonstrate the effectiveness of the proposed method, the accuracy of the extracted watermarks and the quality of the stego frames are evaluated by computer simulations. We will also compare the proposed method with the DCT-DWT method proposed by Sang *et al.* [7] in terms of filtering and frame removal attack results. In addition, the video is compressed using MPEG4 and its compression resistance is evaluated.

The organization of this paper is as follows: Section II describes the embedding process of the proposed method, Section III presents the experimental results, and Section IV concludes.

II. Proposed Method

The size of a video frame is $L_w \times L_h$, and the total number of frames is M. m frames that are not adjacent to each other are randomly selected as stego frames. In this section,

we describe the procedure for generating a stego frame by additively embedding a watermark in the μ-th frame. Fig. 1 shows an overview of the proposed method. Using the similarity between consecutive frames, the watermark is extracted by computing the difference between the μ-th and $(\mu-1)$-th frames.

A. Embedding method

The main points of the proposed embedding process are as follows.

- The watermark is embedded in P blocks in each frame.
- L different patterns are available for the selection of the P blocks. The index of the pattern is encoded in the form of a spreading code.
- The generated spreading codes are embedded in the frame as a detection signal.
- Both the watermark and the detection signal are embedded using additive embedding.

1) Embedding watermark: In the proposed method, the frame is decomposed into YCbCr-components. The watermark is embedded in the Y-component. This approach is common because compression techniques such as JPEG and MPEG tend to reduce the resolution of the Cb and Cr-components, while the Y-component typically keeps high resolution [9]. Let Y^μ be the Y-component of the μ-th frame. Let $\tilde{Y}^\mu$ be the Y-component normalized so that the maximum luminance value Y^μ is 1. In other words, the luminance value at the coordinate (x, y) is normalized by

$$\tilde{Y}^\mu(x,y) = Y^\mu(x,y)/255. \tag{1}$$

The frame $\tilde{Y}^\mu$ is divided into blocks of $h \times h$ pixels. P blocks are randomly selected as embedded blocks. In this paper, there are L distinct patterns for the combinations of selected blocks. In other words, by specifying the index of the pattern, the selected blocks can be determined. The index of the pattern is also embedded in the frame as a detection signal. The seed value of the random number is kept as the key for extraction.

The watermark $\boldsymbol{w}$ is a one-dimensional binary sequence. That is,

$$\boldsymbol{w} = (w_1, w_2, \cdots, w_e), \tag{2}$$

where $w_i \in \{+1, -1\}$, $i = 1, 2, \cdots, e$. A discrete cosine transform (DCT) is performed on each embedding block. To form the DCT coefficient vector $\boldsymbol{d}^\mu$, the Q coefficients are selected from the intermediate frequency components in a zigzag scan order. Note that the number of DCT coefficients, Q, and the number of embedded blocks, P, are determined such that $Q \times P = e$. That is, the DCT coefficient vector $\boldsymbol{d}^\mu$ at the μ-th frame is given by

$$\boldsymbol{d}^\mu = (d_1^\mu, d_2^\mu, \cdots, d_e^\mu). \tag{3}$$

Let $\tilde{\boldsymbol{d}}^\mu$ be the DCT coefficient vector after the additive embedding process, as given by

$$\tilde{\boldsymbol{d}}^\mu = \left(\tilde{d}_1^\mu, \tilde{d}_2^\mu, \cdots, \tilde{d}_e^\mu\right), \tag{4}$$

where

$$\tilde{d}_i^\mu = d_i^\mu + \alpha w_i,\ i = 1, 2, \cdots, e, \tag{5}$$

where $\alpha > 0$ is the embedding intensity of the watermark. After embedding, an inverse discrete cosine transform (IDCT) is performed on each block. A normalized stego frame $\tilde{Y}_s^\mu$ is generated by combining all blocks, including non-embedded blocks. This frame is then restored to the original 256 levels of luminance to create the stego frame Y_s^μ. The value of Y_s^μ may overflow due to the embedding of the watermark. Therefore, we introduce a boundary process given by

$$\chi = \lfloor 255 \times \tilde{Y}_s^\mu(x,y) \rfloor, \tag{6}$$

$$Y_s^\mu(x,y) = \begin{cases} 255, & 255 < \chi \\ \chi, & 0 \le \chi \le 255 \\ 0, & \chi < 0 \end{cases}, \tag{7}$$

where $\lfloor \cdot \rfloor$ is a floor function.

2) Embedding detection signal: The index $\ell = 1, 2, 3, \cdots, L$ identifies the blocks selected during watermark embedding. In other words, L spreading codes $\boldsymbol{\xi}^\ell$ are generated as the detection signals. The spreading codes are stored as secret keys. For error correction during detection, the length of the detection signal $\boldsymbol{\xi}^\ell$ is k bits as given by

$$\boldsymbol{\xi}^\ell = \left(\xi_1^\ell, \xi_2^\ell, \cdots, \xi_k^\ell\right), \tag{8}$$

where $\xi_j^\ell \in \{+1, -1\}$ and $L < k$. The spreading codes are generated to satisfy

$$E\left[\boldsymbol{\xi}^\ell \cdot \boldsymbol{\xi}^\lambda\right] = k\ (\ell = \lambda), \tag{9}$$

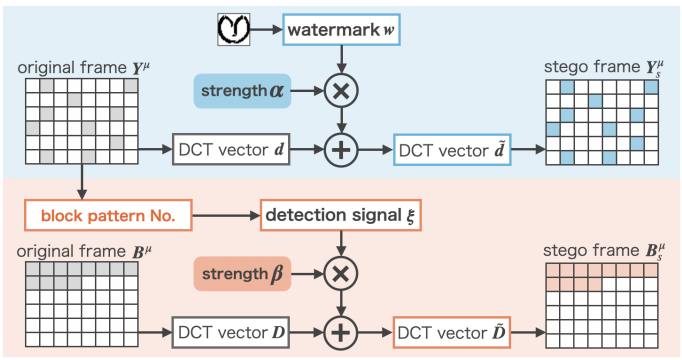
$$E\left[\boldsymbol{\xi}^\ell \cdot \boldsymbol{\xi}^\lambda\right] \simeq 0\ (\ell \neq \lambda). \tag{10}$$

The detection signal is embedded in the Cb-component frame, which represents a different color space than the Y-component frame used for watermark embedding. Human vision is generally known to be more sensitive to red (Cr-component) and less sensitive to blue (Cb-component). Consequently, this method embeds the detection signal in the Cb-component frame to minimize visual discomfort [10]. As the detection signal, one of the L spreading codes is embedded in the Cb-component of the μ-th frame, B^μ. Using (1) with Y^μ replaced by B^μ, the Cb-component frame B^μ is normalized so that its maximum value is 1. The frame $\tilde{B}^\mu$ is divided into blocks of $h \times h$ pixels. k blocks are selected from the frame in raster scan order. On each selected block, the DCT is performed. A coefficient is selected from the intermediate frequency component of the block. This sequence of DCT coefficients is represented by the vector $\boldsymbol{D}^\mu$ given by

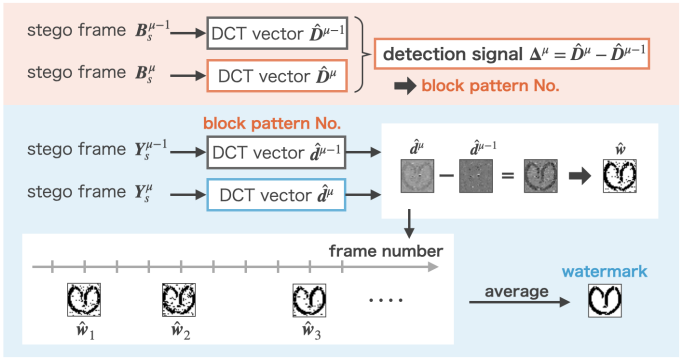
$$\boldsymbol{D}^\mu = (D_1^\mu, D_2^\mu, \cdots, D_k^\mu). \tag{11}$$

One of the selected detection signals $\boldsymbol{\xi}^\ell$ is additively embedded into the DCT coefficients, one bit at a time. Let $\widetilde{\boldsymbol{D}}^\mu$ be the embedded DCT coefficients, as given by

$$\widetilde{\boldsymbol{D}}^\mu = \left(\widetilde{D}_1^\mu, \widetilde{D}_2^\mu, \cdots, \widetilde{D}_k^\mu\right), \tag{12}$$



(a) Embedding process



(b) Extraction process

Fig. 1. Overview of proposed method

where

$$\tilde{D}_j^\mu = D_j^\mu + \beta \xi_j^\ell,\ j = 1, 2, \cdots, k, \tag{13}$$

where $\beta > 0$ is the embedded intensity of the detection signal.

After the embedding process, the IDCT process is performed on each block. A frame $\tilde{B}_s^\mu$ is generated by concatenating all blocks, including non-embedded blocks. Using (6) and (7), the values of the frame $\tilde{B}_s^\mu$ are recovered to their original luminance levels. The stego frame B_s^μ is generated. A color frame is generated by combining the Y, Cr, and Cb-components with embedded watermarks or detection signals. All frames, including non-embedded frames, are lined up to generate stego video.

B. Extracting method

The main points of the proposed extraction process are as follows.

- As in the process of embedding the detection signal, the Cb-component of the frames is divided into blocks, the DCT is performed, and the μ-th DCT coefficient vector $\widehat{\boldsymbol{D}}^\mu$ is generated.
- From the time difference of vectors $\widehat{\boldsymbol{D}}^\mu$ and $\widehat{\boldsymbol{D}}^{\mu-1}$, the frame embedding the detection signal is detected.
- The detection signal is extracted from the difference in the DCT coefficients.
- The correlations between the extracted detection signal and the L spreading codes are calculated, and the label with the largest correlation is found to be the pattern used in that frame.
- From the Y-components of the frames, the watermark is extracted from the blocks indicated by the label patterns.

1) Extracting detection signal: First, the detection signal is extracted from the video. As in the process of embedding the detection signal, the DCT coefficient vector $\widehat{\boldsymbol{D}}^\mu$ of the μ-th frame B_s^μ is generated, as expressed in

$$\widehat{\boldsymbol{D}}^\mu = \left(\widehat{D}_1^\mu, \widehat{D}_2^\mu, \cdots, \widehat{D}_k^\mu\right). \tag{14}$$

To detect a stego frame from the whole video, the detection signal is first detected. To determine the presence of the detection signal, the difference vector between the μ-th and the $(\mu-1)$-th DCT coefficient vectors is defined as

$$\boldsymbol{\Delta}^\mu = \widehat{\boldsymbol{D}}^\mu - \widehat{\boldsymbol{D}}^{\mu-1}. \tag{15}$$

Then, the intensity of the difference vector is defined as

$$\|\boldsymbol{\Delta}^\mu\|_1 = \sum_{j=1}^{k} \left|\widehat{D}_j^\mu - \widehat{D}_j^{\mu-1}\right|. \tag{16}$$

If the video is not attacked by anything, then $\widehat{\boldsymbol{D}}^\mu = \boldsymbol{D}^\mu$. Assuming that the difference between the μ-th vector and the $(\mu-1)$-th vector is sufficiently small, i.e., $\left|D_j^\mu - D_j^{\mu-1}\right| \simeq 0$, then from (13), the intensity of the difference would be

$$\|\boldsymbol{\Delta}^\mu\|_1 = \sum_{j=1}^{k} |\beta \xi_j| \simeq \beta k. \tag{17}$$

It is determined that the detection signal is embedded in the μ-th frame if the intensity $\|\boldsymbol{\Delta}^\mu\|_1$ is greater than a threshold γ.

Now that the frame with the embedded detection signal is known, the next step is the extraction of the detection signal. By finding the correlation between the difference vector and the L spreading codes, we can find the detection signal with the highest correlation. This index represents the selected blocks used for embedding. In other words, the detected index $\hat{\ell}$ is obtained by

$$\hat{\ell} = \arg\max_{\ell} E\left[\boldsymbol{\xi}^\ell \cdot \boldsymbol{\Delta}^\mu\right]. \tag{18}$$

Hence, the detection signal is $\boldsymbol{\xi}^{\hat{\ell}}$.

2) Extracting watermark: The detection signal has been extracted. Accordingly, the watermarked frame can be identified as Y_s^μ. The watermark is extracted from the block specified by the detected index $\hat{\ell}$. The Y-components Y_s^μ and $Y_s^{\mu-1}$ of two consecutive stego frames are divided into blocks, respectively. Then, the DCT is performed on each block. The coefficient vector of the DCT is given by

$$\hat{\boldsymbol{d}}^\mu = \left(\hat{d}_1^\mu, \hat{d}_2^\mu, \cdots, \hat{d}_e^\mu\right). \tag{19}$$

We can assume that the coefficient vector $d^{\mu-1} \simeq d^\mu$ holds because the two successive stego frames are similar. From (5),

TABLE I
EVALUATION VIDEOS (CIF FORMAT)

Video	No. of frames, M	Frame size [px]
football_a	360	352×288
tempete	260	352×288



Fig. 2. Example of logo image used as watermark (32×32 pixels)

the estimated watermark can be calculated from the difference in the coefficient vectors as

$$\hat{w}_i^{\mu} = \mathrm{sgn}\left(\hat{d}_i^{\mu} - \hat{d}_i^{\mu-1}\right). \tag{20}$$

C. Averaged watermark

Although a watermark can be extracted from a single frame, it contains many errors. Since the same watermark is embedded in multiple frames, multiple watermarks extracted from a video can be averaged to generate an averaged watermark. Let Ω be the set of frame numbers that have been detected as stego frames. Hence, the averaged watermark is given by

$$\bar{w}_i = \mathrm{sgn}\left(\sum_{\mu\in\Omega} \hat{w}_i^{\mu}\right),\ i = 1, 2, \cdots, e. \tag{21}$$

III. EXPERIMENT

A. Experimental conditions

As in the DCT-DWT method [7], the two evaluation videos [11] shown in Table I were used for comparison with the DCT-DWT method [7]. We also evaluated four other videos. All of them showed similar results. Only the two results reported in [7] were shown for comparison. The original videos were in the CIF format with 352×288 pixels. These videos were compressed to the MPEG4 format in lossless mode using FFmpeg [12]. The total number of frames, M, varied from video to video, but the number of frames used for embedding was set to $m = 60$ frames. Fig. 2 shows an example of a logo image used as a watermark. Its size was $e = 32 \times 32$ pixels. $L = 8$ types of spreading codes were generated as detection signals. Each frame was divided into blocks with a size of 8×8 pixels ($h = 8$). The DCT was performed on $P = 128$ blocks. The eight DCT coefficients ($Q = 8$) from the 29th to the 36th were selected from each block in a zigzag scan order. An arbitrary binary sequence can be used as a watermark. Since the comparison method [7] uses a logo image, we use a binary logo image as a watermark. The watermark was embedded into the DCT coefficient vector.

The performance of the proposed method was evaluated in terms of the normalized correlation (NC) of the estimated logo image. The NC between two images I^1 and I^2 is defined by

$$\mathrm{NC} = \frac{\sum_{i=1}^{e}\left(I_i^1 - \bar{I}^1\right)\left(I_i^2 - \bar{I}^2\right)}{\sqrt{\sum_{i=1}^{e}\left(I_i^1 - \bar{I}^1\right)^2}\sqrt{\sum_{i=1}^{e}\left(I_i^2 - \bar{I}^2\right)^2}}, \tag{22}$$

where $\bar{I}^1$ and $\bar{I}^2$ are the average pixel values of images I^1 and I^2, respectively. In addition, the stego video has to be generated in such a way that the watermark can be extracted and the degradation of the video quality is minimized. The image quality of a stego frame was evaluated in terms of the peak signal-to-noise ratio (PSNR) given by

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{255^2}{\mathrm{MSE}}\right)[\mathrm{dB}], \tag{23}$$

where MSE is the mean squared error between two color frames $F^1(x, y)$ and $F^2(x, y)$, given by

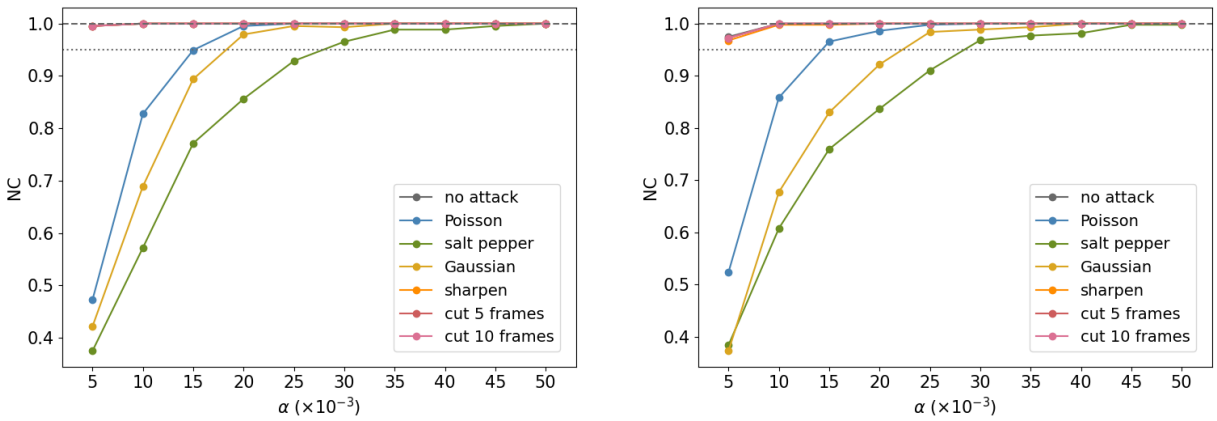
$$\mathrm{MSE} = \frac{1}{3L_hL_w}\sum_{x=1}^{L_h}\sum_{y=1}^{L_w}\sum_{c\in\Omega_c}\left\{F_c^1(x, y) - F_c^2(x, y)\right\}^2, \tag{24}$$

where Ω_c is the set of color channels $\{\mathrm{Y}, \mathrm{Cb}, \mathrm{Cr}\}$. Since the PSNR of non-embedded frames diverges, the quality of a stego video is expressed as the average PSNR of stego frames excluding non-embedded frames.

B. Evaluating robustness against attacks

The robustness of the stego videos when using the proposed method was evaluated. The bit length of the detection signal was $k = 16$ bits. The DCT was performed on $k = 16$ blocks. The 24th coefficients were selected from each block in a zigzag scan order. The embedding strength and threshold for the detection signal were set to $\beta = 0.2$ and $\gamma = 2.0$, respectively. The embedding strength for the watermark was set to $\alpha = 0.005, 0.010, 0.015, \cdots, 0.050$. The generated stego videos were saved in a lossless format. Then, attacks, i.e., Poisson filter, salt and pepper noise (noise density: 2.0%), Gaussian noise (SD $\sigma = 0.05$), sharpening, cutting frames (5 frames), cutting frames (10 frames) were applied to the stego videos. For comparison with the DCT-DWT method [7], the same attacks were used for evaluation. The cutting frames attack implies the frame removal attack. Five or ten frames were randomly removed.

An averaged watermark $\bar{\boldsymbol{w}}$ was extracted from the attacked videos. Its NC value was evaluated. Fig. 3 shows the NC value for the embedding strength α. (a) and (b) represent the NC for *football_a* and *tempete*, respectively. The attacks are represented by each of the colors shown in the legend. NC values of 0.95 and 1.0 are represented by the dotted and dashed lines, respectively. For both videos, when $\alpha = 0.030$ or higher, the watermark was extracted with an NC value of 0.95 or higher. With respect to the image quality of the frame, the minimum embedding strength α with an NC greater than or equal to 0.95 was defined as the optimal value. Accordingly, the optimal embedding strength was $\alpha = 0.030$. The method proposed by Yokota and Kawamura [6] is unable to sufficiently extract the watermark when subjected to frame removal attacks, resulting in significant degradation even after averaging. In contrast, the proposed method uses a detection signal that is resistant to such attacks, allowing accurate watermark extraction.



(a) football_a (b) tempete

Fig. 3. NC of average watermark under various attacks

The NC values of the proposed method and the DCT-DWT method [7] were compared for two evaluation videos, football_a and tempete. The proposed method was evaluated on these stego videos with the optimal embedding strength $\alpha = 0.030$. Table II shows the NC values of the averaged watermark $\bar{\boldsymbol{w}}$ of both methods. The NC values of the proposed method were equal to or better than those of the DCT-DWT method.

Next, the average PSNR of the stego frames for the proposed method and the DCT-DWT method are shown in Table III. We found that the image quality of the proposed method was better than that of the DCT-DWT method. Considering the fact that the DCT-DWT method is a non-blind method [7] and looking at the results of the NC values and the PSNR, the proposed method is superior to the DCT-DWT method. Furthermore, when the optimal embedding strength is set to $\alpha = 0.030$, the proposed method shows no flickering during the stego video playback.

C. Evaluating compression resistance

The compression resistance of the proposed method was evaluated. The bit length of the detection signal was set to $k = 44$ bits because it is difficult to find the detection signal due to compression. The embedding strength β and the threshold γ of the detection signal were set to $\beta = 0.3$ and $\gamma = 2.5$, respectively. The embedding strength for the watermark was set to $\alpha = 0.01, 0.02, \cdots, 0.10$. To compress the videos at a constant quality, a constant rate factor of $\mathrm{crf} = 18$ was specified during compression.

As shown in Fig. 4, we evaluated the NC value of the averaged watermark $\bar{\boldsymbol{w}}$. The horizontal axis is the embedding strength of the watermark, α. The vertical axis is the NC after compression with $\mathrm{crf} = 18$. The watermark was extracted almost correctly when $\alpha \geq 0.05$. Next, the quality of the stego videos was evaluated. The optimal embedding strength for the compression attack was $\alpha = 0.05$. Since the PSNR is more than 41 dB, we can see that the image quality of the proposed method is good. Furthermore, when the optimal embedding strength is set to $\alpha = 0.05$, the proposed method shows no flickering during the stego video playback.

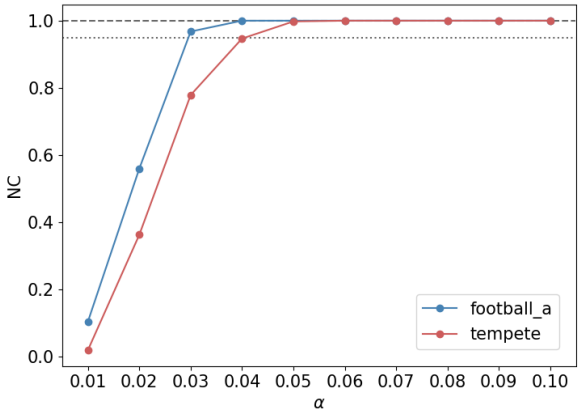


Fig. 4. NC value of averaged watermark after compression

IV. Conclusion

In video watermarking, the watermark should be embedded at high speed because of the large amount of data. Thus, an additional embedding is desirable. However, there has been no blind decoding method other than the method using BSS [6]. A blind extraction method using BSS was proposed by Yokota and Kawamura [6]. Their method was vulnerable to removal attacks because it required embedding watermarks at fixed intervals.

In this paper, we proposed a method that can perform blind decoding while using additive embedding. Two techniques were introduced: distributed watermark embedding and detection signal embedding. Stego frames are divided into blocks, and watermarks are embedded in a distributed manner to improve the quality of a stego video. The same watermark is embedded in several frames. An averaged watermark is created by averaging the watermarks extracted from these frames. Stego frames are selected by using eight different patterns. Each of the selection patterns is represented by an index. The index

TABLE II
NC VALUES OF AVERAGED WATERMARK

	football_a		tempete	
	Proposed Method	DCT-DWT Method	Proposed Method	DCT-DWT Method
non-attack	1.0000	0.9982	1.0000	0.9985
Poisson filter	1.0000	0.9780	1.0000	0.9843
salt-pepper filter (0.02)	0.9654	0.9655	0.9681	0.9644
Gaussian filter (0.05)	0.9931	0.9333	0.9885	0.9258
sharpening	1.0000	0.9985	1.0000	0.9978
cutting frames (5)	1.0000	0.9670	1.0000	0.9663
cutting frames (10)	1.0000	0.8951	1.0000	0.8881

TABLE III
COMPARISON OF AVERAGE PSNR OF STEGO FRAMES AT STRENGTH $\alpha = 0.030$.

Video	Method	PSNR [dB]
foorball_a	Proposed Method	46.5816
	DCT-DWT Method	39.3984
tempete	Proposed Method	46.8538
	DCT-DWT Method	39.6583

TABLE IV
AVERAGE PSNR OF STEGO FRAMES AT STRENGTH $\alpha = 0.05$.

Video	PSNR[dB]
football_a	41.3555
tempete	42.0173

is encoded by a spread spectrum code and embedded as a detection signal. Stego frames can be easily detected with the detection signal. Resistance to frame removal attacks is also provided. As a result of implementing these techniques, it was found that the average PSNR of stego frames could be higher than 41 dB, even when using a large embedding strength ($\alpha = 0.05$) for compression tolerance. Furthermore, by using the average of the extracted watermarks, a watermark with small errors could be extracted.

We compared the proposed method with the DCT-DWT method [7]. Comparing the two methods in terms of NC values against non-geometric attacks, the NC values of the proposed method were larger than those of the DCT-DWT method. We also evaluated the compression resistance of the proposed method. When the embedding was $\alpha = 0.05$ or more, an NC value over 0.95 could be achieved.

V. ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP20K11973 and JP24K15106.

REFERENCES

[1] A. Sverdlov, S. Dexter, A. M. Eskicioglu, "Robust DCT-SVD domain image watermarking for copyright protection: Embedding data in all frequencies," 2005 13th European Signal Processing Conference, Antalya, Turkey, pp. 1-4, 2005

[2] X. Zhang, S. Wang, "Statistical Fragile Watermarking Capable of Locating Individual Tampered Pixels," IEEE Signal Processing Letters, vol. 14, no. 10, pp. 727-730, 2007

[3] C. Pradhan, S. Rath, A. K. Bisoi, "Non Blind Digital Watermarking Technique Using DWT and Cross Chaos," Procedia Technology, vol. 6, pp. 897-904, 2012

[4] I. Hong, I. Kim, S. S. Han, "A blind watermarking technique using wavelet transform," 2001 IEEE International Symposium on Industrial Electronics Proceedings, Korea, vol. 3, pp. 1946-1950, 2001

[5] P. T. Yu, H. H. Tsai, J. S. Lin, "Digital watermarking based on neural networks for color images," Signal Processing, Vol. 81, Issue 3, pp. 663-671, 2001

[6] A. Yokota, M. Kawamura, "BSS-Based Extraction For Additive Video Watermarking," APSIPA ASC 2021, pp. 1640-1646, 2021

[7] J. Sang, Q. Liu, C. L. Song," Robust video watermarking using a hybrid DCT-DWT approach," Journal of Electronic Science Technology vol. 18, issue 2, 2020

[8] N. I. Yassin, N. M. Salem, M. I. El Adawy, "QIM blind video watermarking scheme based on Wavelet transform and principal component analysis," Alexandria Engineering Journal, vol. 53, Issue 4, pp. 833-842, 2014

[9] Podpora, Michal, G. P. Korbas, and A. Kawala-Janik, "YUV vs RGB-Choosing a Color Space for Human-Machine Interaction," FedCSIS (Position Papers), pp. 29-34, 2014

[10] Roorda, Austin, and D. R. Williams,"The arrangement of the three cone classes in the living human eye," Nature, vol. 397, no. 6719, pp. 520-522, 1999

[11] Xiph.org Video Test Media, https://media.xiph.org/video/derf/, (Viewed 2024/5/20)

[12] FFmpeg, https://ffmpeg.org/, (Viewed 2023/5/15)

Transfer-Based Adversarial Attack Against Multimodal Models by Exploiting Perturbed Attention Region

Raffaele Disabato*, AprilPyone MaungMaung†, Huy H. Nguyen† and Isao Echizen†‡
* University of Bologna, Bologna, Italy
† National Institute of Informatics, Tokyo, Japan
‡ University of Tokyo, Tokyo, Japan

***Abstract*—Multimodal models such as GPT-4 and Claude 3 have shown remarkable performance with vision capabilities, enabling exciting new applications with multimodal interactions. With new opportunities, new security concerns come along. Previous studies showed that attackers can force multimodal models to generate unwanted output by manipulating image modality. However, images manipulated by previous attacks are noisy and do not work across different multimodal models. To this end, we propose a new adversarial attack against multimodal models that is stealthy and effective in attacking multiple models. Specifically, we reinforce the previous common weakness attack with multiple surrogate vision transformers to attain strength and utilize an attention mask to focus on essential areas in the image. By doing so, the proposed attack does not add noise to the whole area of the image while maintaining attack strength. Experiment results show that the proposed attack is stealthy (noise patterns in manipulated images by the proposed method are invisible) and can attack two open-source multimodal models, LLaVA 1.5 and MiniGPT-4, in gray-box settings assuming surrogate models are similar to the ones used by targeted victim multimodal models.**

I. Introduction

In recent years, vision-language models (VLMs) have garnered enormous attention and acclaim for their remarkable versatility across a spectrum of tasks, ranging from Visual Question Answering (VQA) [1], [2] to image captioning [3], [4] and beyond [5], [6]. This surge in popularity stems from their ability to seamlessly integrate information from diverse modalities, offering unprecedented performance in understanding and generating content. However, amidst the fervor surrounding their capabilities, a pressing concern emerges, that is, the vulnerability of these models to adversarial attacks.

In this paper, we delve into a concerning scenario where a user encounters an image online (*e.g.*, on social media) that has been manipulated by a malicious actor and unknowingly interacts with a VLM using the image. For example, a user can ask a VLM about a food recipe by showing a food image, and the VLM may output harmful/toxic content. The manipulated image is a carefully crafted adversarial example where imperceptible noise is injected to mislead VLMs like MiniGPT-4 [7] or LLaVA [8]. As the output of VLMs is compromised, VLMs may potentially generate harmful content that will lead to spreading misinformation or amplifying toxic narratives. Consequently, safeguarding the integrity and trustworthiness of VLMs against such threats is imperative, not only to preserve the reliability of VLMs but to also uphold the societal well-being reliant upon their outputs.

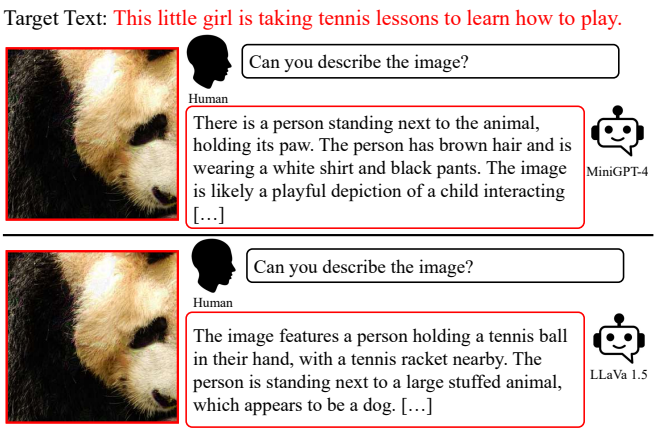


Fig. 1. Example of adversarial example by the proposed attack that reflects target text on LLaVa 1.5 and MiniGPT-4. The adversarial example is generated with perturbation budget ϵ of $4/255$ under maximum norm ℓ_∞.

Towards the robustness of VLMs, researchers have proposed attacks such as AttackVLM [9] and the common weakness attack (CWA) [10]. These previous attacks add adversarial noise to the whole image. Therefore, the noise produced by these attacks becomes visible as the strength of the attacks is increased. In addition, AttackVLM utilizes only one surrogate model and works only on one multimodal model. To this end, we propose a new adversarial attack reinforcing CWA [10] by drawing inspiration from the methodologies outlined for AttackVLM [9]. An example of adversarial example by the proposed attack on different VLMs are shown in Fig. 1. The proposed attack leverages the attention flow of vision transformers used in VLMs to achieve stealthiness while preserving the efficacy of the attack. We make the following contributions in this paper.

- We propose a novel adversarial attack against VLMs that exploits perturbed attention regions reinforced by

AttackVLM [9] and CWA [10] to achieve strength and stealthiness. Consequently, the adversarial examples that the proposed attack produces can fool multiple VLMs (MiniGPT-4 [7] and LLaVA 1.5 [8]) under gray-box settings.

- We conduct experiments to verify the effectiveness of the proposed attack.

This paper aims to elucidate the inherent vulnerabilities shared across multiple VLMs by synthesizing insights from diverse attack methodologies, underscoring the urgency of fortifying their defenses against adversarial manipulation.

II. Related Work

In this section, we briefly review vision language models (VLMs) and existing adversarial attacks on them.

A. Vison-Language Models (VLMs)

Multimodal models are current artificial intelligence (AI) frontiers. VLMs are multimodal models that can process and understand textual and visual information [1], [7], [8], [11], [12]. Most of these models utilize a pre-trained large language model (LLM) and a large vision encoder such as CLIP. During training, the vision encoder remains fixed. Only a projection layer or cross-attention that bridges vision and language is trained. Among the many VLMs, we focus on LLaVA 1.5 [8] and MiniGPT-4 [7] for their remarkable performance and open-source nature to evaluate the proposed attack in this paper.

B. Adversarial Attacks on VLMs

Generally, adversarial examples are carefully perturbed input data (usually imperceptible) to machine learning models that cause the models to make erroneous predictions [13]–[15]. Adversarial example generation (adversarial attacks) can be categorized into two major settings, white-box and black-box, according to the knowledge of the targeted victim model available to the attacker. The white-box threat model assumes the attacker has full knowledge of the model, including its weights and training data. In contrast, the attacker does not know the model in black-box settings. When partial knowledge of the model is available, adversarial example generation is often categorized as gray-box instead of black-box. Depending on the attacker's goal, adversarial attacks that generate adversarial examples can also be classified as targeted or non-targeted attacks. Targeted attacks force the model to output the attacker's intended output, and non-targeted attacks cause the model to make erroneous predictions without specifications.

In the multimodal domain, researchers have also investigated adversarial examples against VLMs. Nonetheless, compared with adversarial examples in the unimodal domain, adversarial examples against multimodal models are less explored. An early work showed that multimodal attacks are stronger than unimodal attacks in gray-box settings and demonstrated the attacks against visual question answering (VQA) and image captioning tasks [16]. Another work proposed an adversarial attack that forces a generative model to produce toxic and harmful outputs [17]. Similar to adversarial examples for image classification models, this study demonstrated that adversarial examples can be crafted against the OpenFlamingo model [18]. More recent works include AttackVLM [9] that introduces attacking strategies to VLMs, AttackBard [19] that investigates the adversarial robustness of Google's Bard, a multimodal language model with image processing capabilities, and the common weakness attack (CWA) [10], which is a powerful transfer-based attack for various models.

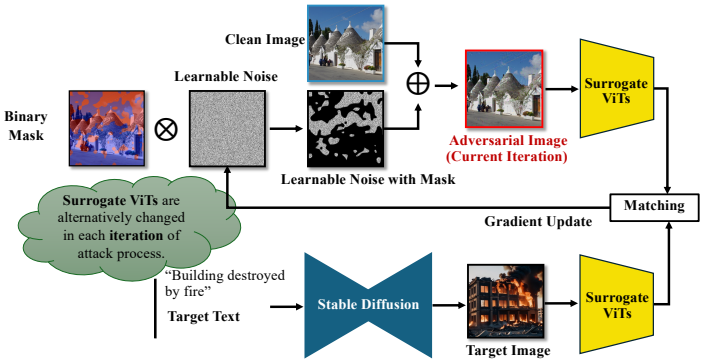


Fig. 2. Overview of proposed method.

In this paper, we focus on gray-box adversarial attacks assuming that surrogate models are highly similar to those used in victim multimodal models. Such gray-box settings are practical in real-world scenarios because open-source foundation models are publicly available online. Specifically, we focus on a transfer-based gray-box attack similar to CWA [10] against two victim VLMs: LLaVA 1.5 [8] and MiniGPT-4 [7].

III. Method

The proposed attack is a targeted attack against VLMs. Given a clean image $\mathbf{x}_{\text{cln}} \in [0,1]^{c\times h\times w}$ with channel c, height h, and width w and a targeted text P, our goal is to create an adversarial example $\mathbf{x}$ that forces a VLM p_θ parameterized by θ that takes $\mathbf{x}$ and a query q to output a response that reflects the targeted text P (*i.e.*, $p_\theta(\mathbf{x}, q) \approx P$). For example, q may be "*describe the image*" or "*what is unusual about the image*" or any content-related question, and p_θ outputs the response related to P instead of the actual content of $\mathbf{x}_{\text{cln}}$. To achieve this, by taking inspiration from AttackVLM [9], we first convert the targeted text P to a targeted image $\mathbf{x}_{\text{tgt}}$ by utilizing existing text-to-image models such as Stable Diffusion [20]. Then, we match the features of $\mathbf{x}_{\text{tgt}}$ with those of $\mathbf{x}$ as in AttackVLM, leveraging a common weakness attack (CWA) with an attention mask for stealthiness. Fig. 2 depicts an overview of the proposed attack.

There are three major components in the proposed attack: image-image feature matching from AttackVLM [9], the common weakness attack (CWA) [10], and targeting important regions in an image with an attention mask. Next, we provide some background on these components and present the proposed attack.

A. Image-Image Feature Matching

Inspired by previous works, the authors of the paper on AttackVLM [9] pointed out that VLMs might not be reliable for optimizing similarity across different modalities. Therefore, they introduced matching image-image similarity (same modalities) by utilizing a public text-to-image model (e.g., Stable Diffusion [20]) to generate a targeted image first. Then, they optimized the following objective,

$$\arg\max_{\mathbf{x}} \; s(\mathbf{x})^\top s(\mathbf{x}_{\text{tgt}}) \text{ s.t. } \|\mathbf{x}_{\text{cln}} - \mathbf{x}\|_\infty \leq \epsilon, \qquad (1)$$

where $s(\cdot)$ is an image encoder (surrogate model). By doing so, VLMs misinterpret adversarial images generated by AttackVLM as the targeted text. We adopt this objective as the loss function in the proposed attack.

B. Common Weakness Attack

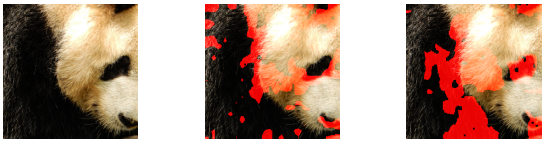
Informed by the previous work [21], the authors for the common weakness attack (CWA) [10] indicate that optimizing adversarial examples adheres to Empirical Risk Minimization (ERM) principles, and a restricted number of training models could result in significant generalization errors. Therefore, they proposed optimizing the loss landscape's flatness and the closeness between the local optima of different models to generate more transferable adversarial examples. CWA also incorporates the momentum iterative (MI) [22] attack and introduces MI-CWA to improve the attack success rate further. We base the proposed attack on the MI-CWA with two major modifications: (1) a different loss function (image-image feature matching) for targeting VLMs and (2) an attention mask for focusing on important regions only to reduce the perturbation.

C. Attention Masks

Although CWA [10] is an effective state-of-the-art transferrable adversarial attack, it perturbs all pixels in an image, resulting in visible noise artifacts. Therefore, we consider focusing only on important regions of an image (*i.e.*, attention) to reduce the impact of the noise added to adversarial images. We explore two attention masks: rollout attention mask (RAM) [23] and layer-wise relevance propagation attention mask (LRP-AM) [24]. An example of the different attention masks is shown in Fig. 3.

RAM [23]. Rollout attention was originally proposed to quantify the attention flow in the Transformer for the natural language processing domain. However, this technique was also used for vision transformers (ViTs) to visualize attention in images. Therefore, we utilize this attention mask to focus on important regions of an image in the proposed attack. Specifically, we use the image encoder components of ViT-L/336 [25] and EVA-CLIP-giant/224 [26] to generate attention masks for clean images. Then, we merge all masks from all surrogate models by using the OR logic operation.

LRP-AM [24]. LRP is an explainability technique applicable to deep neural networks. It functions by propagating prediction backward through a neural network, using a predefined set of propagation rules designed for this purpose. As LRP provides pixel-level relevance and highlights important regions that influence VLMs, the mask produced by LRP (LRP-AM) offers superior performance in our experiments. Unlike RAM, we use ViT-L/336 [25] only to analyze each pixel's relevance in clean images directly.



(a) Clean (b) LRP-AM (c) RAM

Fig. 3. Attention mask visualization. LRP-AM shows less noisy attention.

D. Proposed Attack

We build the proposed attack on top of the MI-CWA [10]. Algorithm 1 details the procedure of the proposed attack. First, we need to generate a targeted image $\mathbf{x}_{\text{tgt}}$ from a given targeted text P using existing public text-to-image models. We utilized Stable Diffusion [20] in our experiments. Since we are attacking VLMs, we substitute the loss function in MI-CWA with image-image feature matching loss (Eq. 1), as described in Section III-A. We generate a binary mask based on the attention mask described in Section III-C to reduce the noise added to adversarial examples. We utilize two surrogate models in the proposed attack (*i.e.*, $\mathcal{S} = \{$ViT-L/336 [25], EVA-CLIP-giant/224 [26]$\}$). Given these two surrogate models with attention masks, we run momentum-based optimization as in MI-CWA [10]. The proposed attack can generate stealthier adversarial examples against LLaVA 1.5 [8] and MiniGPT-4 [7] in transfer-based gray-box settings.

IV. Experiments

A. Settings

Datasets. We used two well-known datasets: a development dataset used in a competition on adversarial attacks and defense for NeurIPS 2017 (hereafter referred to as NIPS17) [27] and ImageNet [28]. NIPS17 contains 1000 images that are labelled with ImageNet labels. We utilized 200 images from NIPS17. ImageNet contains 1.28 million images, and we used 1000 images from its validation set to represent a broad range of visual content. We randomly sampled MS-COCO captions [29] for targeted text descriptions, ensuring a diverse set of image concepts encompassing a wide range of objects, scenes, and activities.

Target Image Generation. We generated targeted images from targeted texts by using Stable Diffusion [20] 2.1. We used Pseudo Linear Multi-Step (PLMS) sampling for 50 steps with a classifier-free guidance scale value of 7.5. The targeted images had a dimension of 512×512.

Victim and Surrogate Models. Our experiments considered two open-source VLMs, LLaVA 1.5 [8] and MiniGPT-4 [7],

Algorithm 1 Modified MI-CWA algorithm (Proposed)

Require: clean image $\mathbf{x}_{\text{cln}}$, targeted image $\mathbf{x}_{\text{tgt}}$, perturbation budget ϵ, iterations T, loss function L, surrogate models $\mathcal{S} = \{s_i\}_{i=1}^{n}$, step sizes β and α

Initialize $\boldsymbol{m} \leftarrow 0$, inner momentum $\hat{\boldsymbol{m}} \leftarrow 0$, $\mathbf{x}_0 \leftarrow \mathbf{x}_{\text{cln}}$
Generate a binary mask bm
for all $i \leftarrow 1, \ldots, n$ **do**
  Calculate target features $F_{\text{tgt}}^{i} \leftarrow s_i(\mathbf{x}_{\text{tgt}})$
end for
for all $t \leftarrow 0, \ldots, T-1$ **do**
  $\mathbf{x}_t^0 \leftarrow \mathbf{x}_0$
  for all $i \leftarrow 1, \ldots, n$ **do**
    Calculate adversarial features $F_{\text{adv}}^{i-1} \leftarrow s_i(\mathbf{x}_t^{i-1})$
    Calculate $\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{x}} L(F_{\text{tgt}}^{i-1}, F_{\text{adv}}^{i-1})$
    Update inner momentum by $\hat{\boldsymbol{m}} \leftarrow \mu \cdot \hat{\boldsymbol{m}} + \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$
    Update $\boldsymbol{x}_t^i$ by $\boldsymbol{x}_t^i \leftarrow \text{clip}_{\boldsymbol{x}_{cln},\epsilon}(\boldsymbol{x}_t^{i-1} + \beta \cdot \boldsymbol{m} \cdot bm)$
  end for
  Calculate update $\boldsymbol{g} \leftarrow \boldsymbol{x}_t^n - \boldsymbol{x}_t$
  Update momentum $\boldsymbol{m} \leftarrow \mu \cdot \boldsymbol{m} + \boldsymbol{g}$
  Update $\boldsymbol{x}_{t+1}$ by $\boldsymbol{x}_{t+1} \leftarrow \text{clip}_{\boldsymbol{x}_{cln},\epsilon}(\boldsymbol{x}_t + \alpha \cdot \text{sign}(\boldsymbol{m}))$
end for
return $\boldsymbol{x}_T$

as victims for evaluation. We used two surrogate models, the image encoder components of ViT-L/336 [25] and EVA-CLIP-giant/224 [26], to generate attention masks and extract image embeddings in the proposed attack.

Evaluation Method. Sentence T5 Large [30], a state-of-the-art Sentence Transformer model, was chosen to evaluate the semantic content of the captions. This selection leverages Sentence T5 Large's ability to handle full-length descriptions, capturing the nuances of the semantic relationships compared with traditional text encoders with truncation limitations.

Implementation Details. We implemented the proposed attack on top of the publicly available MI-CWA code[1]. We utilized feature matching loss from the AttackVLM official code repository[2] and attention masks from the public code base[3]. The parameters were an inner step size value of 250 and an outer step size value of 2, and we ran the attack for 500 steps. The perturbation budget ϵ had a value of $4/255$ under maximum norm ℓ_∞. For comparison, we ran AttackVLM image-image feature matching (FM-ii) [9] with a step size of 1 and a perturbation budget ϵ of $4/255$ under ℓ_∞. We used the query *"Describe this image"* when querying VLMs for both clean and adversarial examples.

B. Results

To verify the effectiveness of the proposed attack, we calculated the Sentence BERT score between the targeted texts and generated texts mentioned in Section IV-A to measure the attack strength. In addition, to assess visual quality while maintaining attack strength, we utilized three metrics: the structural similarity index measure (SSIM), L2 distance, and a state-of-the-art perceptual quality metric, TOPIQ [31]. We also compared the performance of the proposed attack with two state-of-the-art attacks against VLMs: CWA [10] and AttackVLM (FM-ii) [9].

[1]https://github.com/huanranchen/AdversarialAttacks/
[2]https://github.com/yunqing-me/AttackVLM
[3]https://github.com/jacobgil/vit-explain

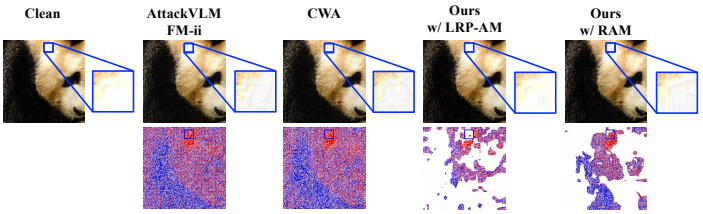


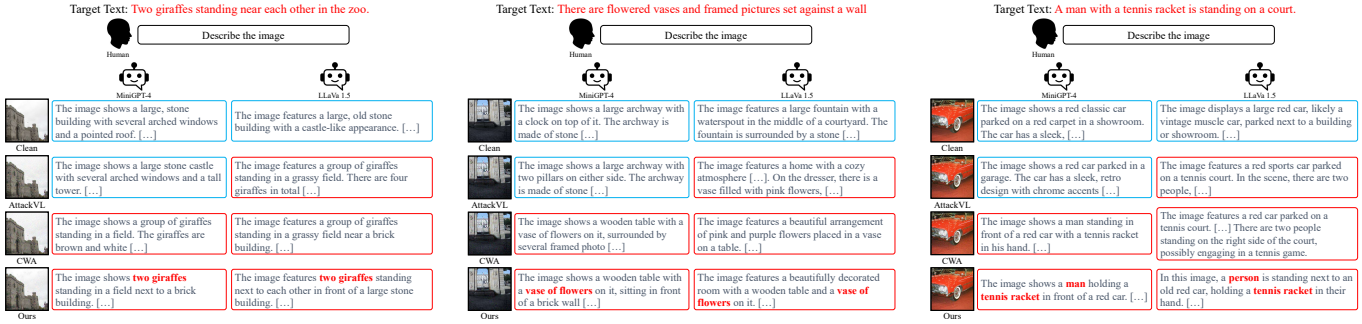
Fig. 4. Examples of adversarial examples generated by AttackVLM [9], CWA [10], and proposed attack with their corresponding noise maps.

TABLE I
SENTENCE BERT SCORE (↑) BETWEEN TARGETED TEXTS AND VICTIM MODEL GENERATED TEXTS. BEST RESULTS ARE IN BOLD, AND SECOND BEST RESULTS ARE UNDERLINED.

	NIPS17 (200 images)		ImageNet (1000 images)	
Attack	LLaVA 1.5	MiniGPT-4	LLaVA 1.5	MiniGPT-4
Clean	0.66	0.67	0.66	0.67
CWA [10]	<u>0.79</u>	**0.78**	**0.79**	**0.78**
AttackVLM (FM-ii) [9]	**0.80**	0.67	**0.79**	0.67
Ours (w/ RAM)	0.73	0.72	0.72	0.71
Ours (w/ LRP-AM)	0.75	<u>0.75</u>	<u>0.74</u>	<u>0.74</u>

Sentence BERT Score. Table I summarizes the Sentence BERT score for the two datasets: NIPS17 and ImageNet (random 1000 validation images). The generated adversarial examples were tested against the two VLMs: LLaVA 1.5 [8] and MiniGPT-4 [7]. AttackVLM achieved the highest score for LLaVA 1.5, but it did not work on MiniGPT-4. The reason is that AttackVLM is designed to use one surrogate model. In contrast, CWA and the proposed attack utilize multiple surrogate models. Consequently, CWA and the proposed attack could be applied to both VLMs. Although the proposed attack did not achieve the highest score due to noise reduction, the attack performance was comparable. Fig. 5 shows the output of adversarial examples by victim VLMs.

Visual Quality. Table II presents visual quality results for different metrics in comparison with the state-of-the-art methods. As expected, the proposed method achieved the highest visual quality performance across different metrics. The attention masks used in the proposed attack effectively reduced noise artifacts. Fig. 4 shows adversarial examples with their corresponding noise maps by different methods. From the figure, adversarial examples by the proposed attack contained less noise, which was confirmed in an objective visual quality evaluation. It is worth noting that the TOPIQ score for ImageNet (1000 images) was lower than that of NIPS17. One possibility is that the size of ImageNet images varies, and we did not consider the aspect ratio while resizing the images, which might have affected the perceptual visual quality.



(a) Example 1 (b) Example 2 (c) Example 3

Fig. 5. Responses of LLaVa 1.5 and MiniGPT-4 on clean images and adversarial examples by the proposed method, AttackVLM [9], and CWA [10].

TABLE II
VISUAL QUALITY EVALUATION OF GENERATED ADVERSARIAL EXAMPLES. BEST RESULTS ARE IN BOLD, AND SECOND BEST RESULTS ARE UNDERLINED.

	NIPS17 (200 images)			ImageNet (1000 images)		
Attack	SSIM (↑)	L2 (↓)	TOPIQ (↑)	SSIM (↑)	L2 (↓)	TOPIQ (↑)
CWA [10]	0.79	17.22	0.65	0.79	19.29	0.35
AttackVLM (FM-ii) [9]	0.58	19.15	0.66	0.77	22.12	0.36
Ours (w/ RAM)	**0.93**	**9.5**	0.71	0.91	**10.39**	**0.38**
Ours (w/ LRP-AM)	**0.93**	10.47	**0.76**	**0.92**	11.81	0.37

V. DISCUSSION

We demonstrated a transfer-based adversarial attack against VLMs with comparable attack strength and less noise. Such stealthier adversarial examples may be proliferated and distributed online by malicious actors. Honest users may feed these adversarial examples to VLMs such as LLaVA 1.5 or MiniGPT-4. As a result, VLMs may be compromised, potentially generating harmful content, spreading misinformation, or amplifying toxic narratives.

Responsible Artificial Intelligence (AI) Development. As VLMs are still under development, paying attention to safety early in development is indispensable. The ability to generate high-quality adversarial examples emphasizes the importance of responsible AI development. VLMs are increasingly integrated into various applications, so it is crucial to consider the potential for misuse. Our work highlights the need for robust security measures and responsible development practices to ensure that VLMs are not weaponized to spread misinformation or manipulate public discourse.

Limitations. While the proposed attack can generate finer adversarial examples against VLMs, it still has limitations. This paper evaluated the proposed attack against LLaVA 1.5 and MiniGPT-4. We shall investigate other VLMs for further evaluation. In addition, the proposed attack in its current form only deploys two surrogate models, so we shall further investigate and explore more surrogate models to improve transferability in our future work. We shall also conduct an in-depth analysis of the relationship between surrogate models and VLMs.

VI. CONCLUSION

This paper proposes a transfer-based adversarial attack against vision-language models (VLMs) using multiple surrogate models. Unlike the previous methods, the proposed attack utilizes attention masks to reduce the noise added to adversarial examples. Experiments confirmed that the proposed attack is comparable to the previous attack methods in terms of attack strength and is superior in terms of visual quality. We hope that our work encourages model developers towards safe AI system development.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grants JP21H04907, 23K19983, and JP24H00732, by JST CREST Grants JPMJCR18A6 and JPMJCR20D3 including AIP challenge program, by JST AIP Acceleration Grant JPMJCR24U3, by JST K Program Grant JPMJKP24C2 Japan, and by the project for the development and demonstration of countermeasures against disinformation and misinformation on the Internet with the Ministry of Internal Affairs and Communications of Japan.

REFERENCES

[1] J.-B. Alayrac, J. Donahue, P. Luc, *et al.*, "Flamingo: A visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[2] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 030–18 040.

[3] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, PMLR, 2023, pp. 19 730–19 742.

[4] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 200–212, 2021.

[5] F. Bao, S. Nie, K. Xue, *et al.*, "One transformer fits all distributions in multi-modal diffusion at scale," in *International Conference on Machine Learning*, PMLR, 2023, pp. 1692–1717.

[6] A. Q. Nichol, P. Dhariwal, A. Ramesh, *et al.*, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 17–23 Jul 2022, pp. 16 784–16 804.

[7] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[8] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[9] Y. Zhao, T. Pang, C. Du, *et al.*, "On evaluating adversarial robustness of large vision-language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[10] H. Chen, Y. Zhang, Y. Dong, X. Yang, H. Su, and J. Zhu, "Rethinking model ensemble in transfer-based adversarial attacks," *arXiv preprint arXiv:2303.09105*, 2023.

[11] A. Awadalla, I. Gao, J. Gardner, *et al.*, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[12] J. Y. Koh, R. Salakhutdinov, and D. Fried, "Grounding language models to images for multimodal inputs and outputs," in *International Conference on Machine Learning*, PMLR, 2023, pp. 17 283–17 300.

[13] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[14] B. Biggio, I. Corona, D. Maiorca, *et al.*, "Evasion attacks against machine learning at test time," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, Springer, 2013, pp. 387–402.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[16] I. Evtimov, R. Howes, B. Dolhansky, H. Firooz, and C. C. Ferrer, "Adversarial evaluation of multimodal models under realistic gray box assumption," *arXiv preprint arXiv:2011.12902*, 2020.

[17] X. Qi, K. Huang, A. Panda, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak large language models," *arXiv preprint arXiv:2306.13213*, 2023.

[18] C. Schlarmann and M. Hein, "On the adversarial robustness of multi-modal foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3677–3685.

[19] Y. Dong, H. Chen, J. Chen, *et al.*, "How robust is google's bard to adversarial image attacks?" *arXiv preprint arXiv:2309.11751*, 2023.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[21] H. Huang, Z. Chen, H. Chen, Y. Wang, and K. Zhang, "T-sea: Transfer-based self-ensemble attack on object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 514–20 523.

[22] Y. Dong, F. Liao, T. Pang, *et al.*, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185–9193.

[23] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," *arXiv preprint arXiv:2005.00928*, 2020.

[24] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[26] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.

[27] A. Kurakin, I. Goodfellow, S. Bengio, *et al.*, "Adversarial attacks and defences competition," in *The NIPS'17 Competition: Building Intelligent Systems*, Springer, 2018, pp. 195–231.

[28] O. Russakovsky, J. Deng, H. Su, *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[29] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[30] J. Ni, G. H. Abrego, N. Constant, *et al.*, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," *arXiv preprint arXiv:2108.08877*, 2021.

[31] C. Chen, J. Mo, J. Hou, *et al.*, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *arXiv preprint arXiv:2308.03060*, 2023.

A Permutation-based Reversible Data Hiding Method with Zero Visual Distortion

Wendi Zhu†, KokSheik Wong†, Minoru Kuribayashi‡
†School of Information Technology, Monash University Malaysia, Malaysia.
‡Center for Data-driven Science and Artificial Intelligence, Tohoku University, Japan.
{wzhu0045@student., wong.koksheik@}monash.edu, kminoru@tohoku.ac.jp

***Abstract*—Recently, several approaches have been designed to hide data in portable document format (PDF) files. These approaches have demonstrated their advantages in different application scenarios, including copyright verification, covert communication/steganography, and content forensics. However, they often suffer from visual distortion or lack universal applicability. In this work, we propose a reversible and transparent method that exploits the coding properties of text objects (i.e., substrings) in a PDF-compliant document to embed data. In particular, the position information of the substrings is adjusted to hide data, where each unique permutation of the substrings encodes a bit sequence. Subsequently, the distance of each substring from the left margin is corrected so that the processed PDF has the exact layout or appearance of the original PDF, hence completely preserving the quality of the original PDF file. In the best-case scenario, to hide one bit of data, 5.88 bits of the PDF file are required, i.e., $1 : 5.88$. In addition, this method can be deployed in tandem with conventional data hiding methods to hide more data and to hide data in different ways.**

I. Introduction

Data hiding is a process of concealing data in a host such as text, image, audio, and even DNA sequences to serve specific purposes [1], [2]. In recent years, much effort has been put into data hiding research for images and videos, but research into data hiding algorithms for text document has received relatively less attention. However, as generative artificial intelligence (AI) improves, more users turn to large language models (LLM) for various creative projects. It is essential to acknowledge that the rapid advancement of LLM also serves as a continuous reminder of the dangers associated with fake news and hence the need to safeguard the copyrights and integrity of textual contents [3]. Among the digital file formats for text, PDF is the most adopted format due to its extensive list of features, including preserving layout (look-and-feel), providing universal accessibility, supporting interactive features, to name a few [4]. More importantly, PDF documents remain ubiquitous in the current digital landscape, where online working and learning are becoming increasingly popular. In addition, despite the proliferation of these technologies, we have not significantly advanced in creating more universal file types in addition to the PDF. However, the platforms for accessing PDF content are rapidly expanding. As the digital ecosystem evolves, ensuring the integrity and safety of content within PDFs becomes paramount, not just for safeguarding intellectual property but also for protecting personal and sensitive information.

As the PDF document standard evolves, many innovative and clever methods have been developed for hiding data within PDF files. However, almost all conventional data hiding methods fail to maintain visual consistency after the data is hidden. It is due to the exploitation of visible text features in PDF files, including font type, size, and color, to hide data. Our method distinguishes itself from the conventional methods by ensuring that the document's appearance remains identical to that of the original document, i.e., complete quality preservation (CQP) [5], [6]. We leverage the inherent characteristics of the PDF format to facilitate data hiding, where only invisible features are manipulated to hide data. Our work makes the following contribution: (a) exploiting the rendering order of substrings within the 'TJ' operators to hide data, which is the first of its kind; (b) achieving complete quality preservation in the processed PDF file; (c) achieving reversibility by removing the inserted data, and; (d) enabling integration with conventional methods to hide more data and hide data in different ways.

II. Related Work

This section establishes the rationale of our study by providing an overview of the PDF file structure, examining conventional distortionary and distortion-free data hiding methods, identifying research gaps and defining quality-preserving techniques.

A. Overview of PDF file structure

Portable document format (PDF) has four integral segments: *Header*, *Body*, *Cross-Reference Table (CRT)*, and *Trailer* (Fig. 1). The *Header*, which only includes the PDF ISO standard version, starts with '%PDF-' followed by '1.4' or '2.0', indicating the PDF build standard. The *Body* contains critical content, including various operators that facilitate the organization and presentation of the text. For instance, the 'Tf' operator indicates the font type used, and the 'TJ' operator controls the text content rendering (Fig. 1). The *CRT* facilitates access to indirect objects within a PDF file, while the *Trailer* assists PDF processors in quickly locating essential elements.

†This research is supported by the Fundamental Research Grant Scheme: *Complete Quality Preservation and Reversible Data Hiding - from formulation to application* (Grant No. FRGS/1/2023/ICT07/MUSM/02/1).

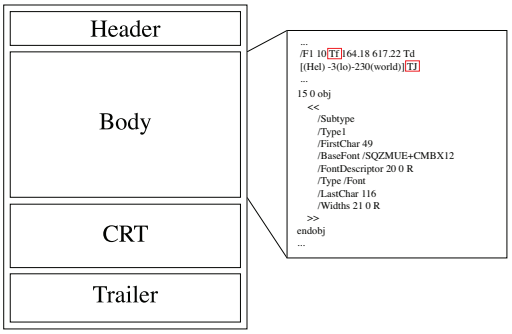


Fig. 1. PDF Structure (reproduced from [5])

The key advantage of PDFs is their portability, ensuring consistent document appearance across diverse platforms. Each PDF file complies with ISO guidelines [7] and supports various media forms [8]. PDFs are also more secure than formats such as Microsoft Word, leading to increased focus on enhancing their security features [9]. Important PDFs are often encrypted with user and owner passwords, where the user password allows viewing, while the owner password sets restrictions on copying and editing [5]. These characteristics make PDFs the preferred choice for handling important documents.

B. Synonym-based data hiding method

Researchers have explored synonym-based methods to hide data in text documents, replacing words like 'teaching' with 'tutoring' to encode data such as '01', while keeping 'teaching' to encode '00' [10]. However, synonym substitution can lead to significant content distortion and cannot conceal a large payload, as achieving larger payloads requires significantly longer text. Xiang et al. proposed using a Word Indexing Compression (WIC) algorithm to compress the cipher text before hiding, thereby enhancing the feasibility of using synonyms for data hiding [11].

C. Image-based data hiding method

Due to its portability, researchers continue to study data hiding methods in PDF files by converting them into images and modifying the content at the pixel level. Li et al. modified the text boundaries to embed data, causing slight page distortion imperceptible to humans, but it suffers from having a limited capacity, i.e., only 2 bits can be embedded per text line [12]. Khosravi et al. devised a method that adjusts the details of multiple character spacing to hide data [13]. Their approach leveraged different text editors' varying settings for correcting space, to embed hidden data in the document without notable yet inevitable distortion.

D. Open space data hiding method

Besides synonym-based methods, researchers also attempted to manipulate the open space due to its transparency. UniSpaCh by Por et al. [14] exploited multiple Unicode space characters to encode the data in the files. However, a statistical analysis could detect irregularities in using Unicode space characters, which is rare in an ordinary document.

Many open space methods cannot achieve complete quality preserving (CQP) [14], but some are distortion-free and more transparent than the earlier methods. One such method manipulates non-breaking spaces ('A0' in hexadecimal) in PDF editors [15]. This approach substitutes the original ASCII code space '20' by 'A0' to encode '1' in binary or leaves it unchanged (i.e., remains as '20') to encode '0'. However, embedding data n in the spaces between characters allows higher capacity but leads to file size expansion. Furthermore, embedding data in spaces between words will decrease the capacity, and some editors may display the 'A0' characters visibly.

E. Data hiding method based on PDF's feature

Some researchers have dealt with the PDF structure standard. Kuribayashi et al. introduced a technique that utilizes the 'TJ' operator to manipulate the encoding of the space information [5], [16]. An integer I is split into $(I-m_i)+(m_i)$, where m_i is the decimal equivalent of the bit sequence to be embedded. Since the space value I remains unchanged, Kuribayashi et al.'s method completely preserved the visual consistency. However, this method introduced new pairs of parentheses '()' followed by the numbers m_i. This feature leads to a diminished embedding efficiency and a surge in redundancy.

A data hiding method is considered CQP if the original and modified documents (i.e., embedded with data) are identical. The methods mentioned above mainly distort the visual / layout of the processed document due to hiding data. Therefore, there is a gap in exploring more CQP approaches. Before our current work, CQP data hiding research mostly manipulated the whitespace between substrings to conceal information. We will contribute to providing a new permutation-based method with no visual distortion by adjusting the rendering order of the text showing operators.

III. Methodology

We propose a method to alter the order of substrings, then subsequently restore the text width and spacing by adjusting their glyphs. The framework consists of two main parts: (1) hiding data, and; (2) extracting the hidden data and restoring the original PDF file. The following subsections detail the subprocesses involved in each part of the framework.

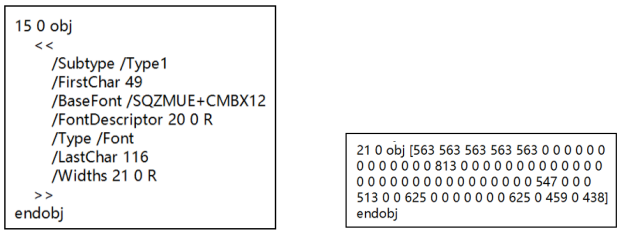
A. Embedding data

1) Encrypting the data: Without loss of generality, we assume that the data m to be hidden is encoded in binary. We divide m into non-overlapping segments m_j, where each segment can assume a different length k_j, which is defined later in Eq. (1). Each segment m_j is then encrypted using a stream cipher to obtained the encrypted segment n_j.

2) Determining substrings' order: We embed each encrypted segment n_j by forcing a group of selected substrings to appear in a specific order. To facilitate the presentation, suppose that we have a PDF file that encodes only the phrase "Hello world". "Hello world" can be encoded using three

TABLE I
ALL POSSIBLE PERMUTATIONS OF THREE SUBSTRINGS. HERE, *Lehmer Code* IS USED FOR ASSOCIATING EACH PERMUTATION WITH A 2-BIT SEQUENCE.

Permutation	2-bit
u_1, u_2, u_3	00
u_1, u_3, u_2	01
u_2, u_1, u_3	10
u_2, u_3, u_1	11
u_3, u_1, u_2	-
u_3, u_2, u_1	-


15 0 obj
<<
/Subtype /Type1
/FirstChar 49
/BaseFont /SQZMUE+CMBX12
/FontDescriptor 20 0 R
/Type /Font
/LastChar 116
/Widths 21 0 R
>>
endobj
21 0 obj [563 563 563 563 563 0 0 0 0 0 0 0 0 0 0 0 0 813 0 0 0 0 0 0 0 0 0 0 0 0

$w_{\text{“world”}} = 2346$. Subsequently, the original distance from the left margin is computed for each substring:

$$\begin{aligned} d_{\text{“Hel”}} &= 0; \\ d_{\text{“lo”}} &= d_{\text{“Hel”}} + w_{\text{“Hel”}} + s_{\text{“lo”}} \\ &= 0 + 1351 + 3 = 1354; \\ d_{\text{“world”}} &= d_{\text{“lo”}} + w_{\text{“lo”}} + s_{\text{“world”}} \\ &= 1354 + 757 + 230 = 2341. \end{aligned} \tag{5}$$

Suppose our substrings are permuted as follows to encode the data segment ‘10’:

$$[-s_{\text{“lo”}}\ (\texttt{lo})\ -s_{\text{“Hel”}}\ (\texttt{Hel})\ -s_{\text{“world”}}\ (\texttt{world})]\ \text{TJ}. \tag{6}$$

By referring to the distance values d_i stipulated in Eq. (5), the space values $s_{i'}$ are adjusted as follows:

$$\begin{aligned} s_{\text{“lo”}} &= d_{\text{“lo”}} = 1354; \\ s_{\text{“Hel”}} &= d_{\text{“Hel”}} - d_{\text{“lo”}} - w_{\text{“lo”}} \\ &= 0 - 1354 - 757 = -2111; \\ s_{\text{“world”}} &= d_{\text{“world”}} - d_{\text{“Hel”}} - w_{\text{“Hel”}} \\ &= 2341 - 0 - 1351 = 990. \end{aligned} \tag{7}$$

To complete the correction step, the ‘TJ’ tag is updated to:

$$[-1354\ (\texttt{lo})\ 2111\ (\texttt{Hel})\ -990\ (\texttt{world})]\ \text{TJ}. \tag{8}$$

With the aforementioned updates, the text “Hello world” will appear exactly the same as in the original PDF file. In essence, instead of the original rendering sequence, the offset for the substring ‘lo’ (i.e., $s_{\text{“lo”}}$) is first computed, and the substring ‘lo’ is rendered. Subsequently, the offset for the substring ‘Hel’ is computed, and the substring ‘Hel’ is rendered. The same process is applied to handle ‘world’.

5) Setting permissions: According to PDF standards, assigning both owner and user passwords is common for protecting sensitive documents. The user password allows viewing and, if permitted by the owner, content copying. We restricted document access to viewing only (which is a commonly implemented setting, e.g., credit card statement), ensuring that our raw data remains unexplored and the no-distortion cover remains innocuous.

B. Extracting data

Contrary to the embedding process, we must open the processed and protected PDF file using the password and perform the extraction operation. Since each ‘TJ’ operator is independent, we only need to iterate through all the ‘TJ’ operators from the beginning of the PDF document instead of extracting them in reverse order.

First, the number of substrings on the jth line N_j is determined. Next, the original ordering of the substrings is extracted from the processed text (with hidden data) by comparing the definite distance of each substring from the left margin, i.e., the entity d_i. Specifically, d_i is determined by rearranging the terms as follows:

$$d_i = s_{i'} + d_{i'-1} + w_{i'-1}. \tag{9}$$

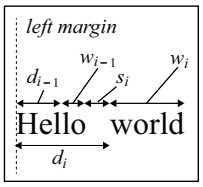


Fig. 3. Illustration of the semantic for d_i, d_{i-1}, w_{i-1} and s_i. Here, ‘world’ is the ith substring.

Here, a smaller d_i implies that the associated substring s_i is closer to the left margin and hence the ith sub-string would appear earlier in the original order, and vice versa.

For illustration purposes, we can examine the example captured by Eq. (7), where the number of substrings on the jth line is $N_j = 3$. By comparing the distances $d_{\text{“world”}}$, $d_{\text{“Hel”}}$, and $d_{\text{“lo”}}$, we can determine their original positions as 3rd, 1st, and 2nd, respectively. This is deduced from the fact that $d_{\text{“Hel”}} < d_{\text{“lo”}} < d_{\text{“world”}}$. Consequently, this ordering is represented by the permutation (u_3, u_1, u_2). By referencing the mapping rule in Table I, we can extract the hidden data, i.e., ‘01’. This process is repeated for the remaining rows for further extraction until the entire document is processed. The extracted encrypted data segment n_j is decrypted to obtain the original data segment m_j. The data segments m_j are subsequently concatenated to form the complete data m.

IV. EXPERIMENTS

This section evaluates and validates the proposed method in terms of transparency, capacity, and robustness. The proposed data hiding method is implemented in Python3. The first five chapters of Genesis in the Old Testament of the Bible (generated by Latex using TeX Live 2016) and “The Romance of Three Kingdoms” (https://threekingdoms.com/), generated by Microsoft Word 2016 (Version 16.0.17328.20184), are utilized as cover PDF file to be embedded for experiment purposes. Randomly generated bit sequence is embedded into the PDF for experiment purposes.

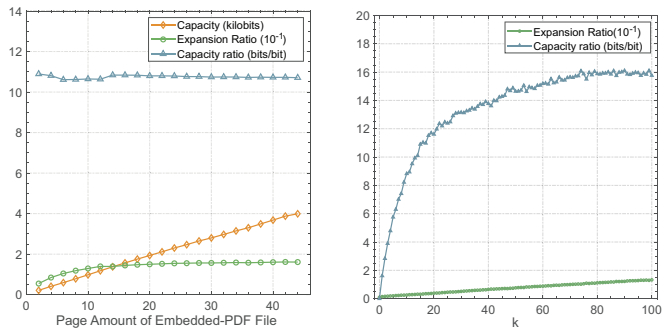
A. Transparency

Transparency refers to the level of similarity between the processed PDF and its original counterpart, where higher similarity implying higher transparency, and vice versa. We exported both the original and processed PDF documents as images, then quantified the differences between them using the SSIM image quality metric. The results consistently showed an SSIM of 1.0, indicating that our data hiding method does not introduce any visual distortions (i.e., CQP) and thus maintains the highest possible level of transparency.

B. Capacity

This section examines how the performance of the proposed method is influenced by the number of pages in the PDF file and capping the number of bits to be embedded using each ‘TJ’ line. The metrics adopted to measure the rate of change in file size ϕ and the capacity ratio η are defined as:

$$\text{Expansion Ratio, } \phi = \frac{\text{Increment of File Size}}{\text{Original File Size}}, \tag{10}$$



(a) Capacity/ϕ/η vs number pages (b) Capacity/ϕ vs parameter k

Fig. 4. Performance evaluation for various number of pages and parameter values for k.

$$\text{Capacity Ratio, } \eta = \frac{\text{Size of hidden data (bits)}}{\text{File size increment (bytes)} \times 8}. \quad (11)$$

The results in Fig. 4(a) show that capacity (orange diamond) increases linearly with the number of pages, using all text rows for data hiding. The capacity ratio η (blue triangle) remains constant at ~ 0.17 data bits per PDF bit. On the other hand, the file size expansion ratio ϕ (green circle) gradually increases with the number of pages but plateaus when the number of pages exceeds 30. Our method favors shorter PDFs, with $\phi = 5.5\%$ for hiding $21,339$ bits in a two-page PDF. However, for PDFs with ≥ 30 pages, ϕ remains nearly constant. For instance, $\phi = 15.7\%$ for concealing $280,165$ bits in a 30-page PDF and rises slightly to 16.1% for $399,256$ bits in a 44-page PDF. This minimal increase in ϕ is likely due to the fundamental base size of PDF documents, where added data represents a smaller percentage of the total file size as the document lengthens, slowing the rate of ϕ increase.

Figure 4(b) examines how the parameter k (bit length embedded into each 'TJ' line) affects the performance of our data hiding method. Adjusting k helps control the file size expansion ratio ϕ by capping the payload per line between 0 and 100 bits. As k increases, so does ϕ, and vice versa. For example, with $k = 29$ bits, $\phi = 4.9\%$. Thus, selecting a suitable k value depends on the acceptable level of ϕ.

C. Robustness and Security

To understand practical applications, we evaluated the Bit Error Rate (BER) for data extraction after applying standard lossless compression (ZIP/Flate) on PDFs, embedding at maximum rate on all pages. Our experiments confirmed that data embedded in PDFs created by Latex or Microsoft Word remained unchanged after compression and decompression, achieving a BER of 0.

As for security, our proposed method offers the first-of-its-kind feature, i.e., allowing data extraction from the copied-and-pasted texts. Imagine a scenario where an attacker, during a business meeting or critical event, attempts to rapidly copy the contents of a PDF into her document. She will find that the text appears disordered and chaotic upon pasting, because the substrings are deliberately reordered to hide data. This unexpected strategy effectively shields the file from rapid copy-paste actions. If the copy-paste operation is indeed performed, the unique order can be examined to determine the hidden data, which can contain data such as the owner, document source, etc. More importantly, our proposed method allows two ways of data extraction. The first way is the typical data extraction pathway using the processed PDF file, and the second way is to examine the pasted texts. This mode of data extraction will be further explored in our future work.

V. DISCUSSIONS

A. Comparison with conventional methods

To demonstrate the advantages of the our proposed method, we conduct a comparative analysis, examining the performance of both past representative methods and those within the same category as ours.

Li et al.'s method [12] converts a PDF host document to an image, then modifies minor section of the edge of the characters. However, they can only embed 2 bits of data in a single line, hence its capacity is very low for practical use. On the other hand, Lee & Tsai's classic method [15] exploits the special character 'A0' to mark spaces [15], where they proposed two ways: replace the '20' code by 'A0', which leads to no file size increment but suffers from low capacity, or; to add additional 'A0' space between the characters which brings high capacity but at the expense of increasing the file size.

As for the method using the 'TJ' operator, Kuribayashi et al. utilized space lengths in the 'TJ' operators by considering them as a one-dimensional feature, making subtle distortion [16]. In a different work, Kuribayashi & Wong [5] introduce no-distortion after hiding data in a PDF file, but their work suffers from large ϕ, for example, hiding 3,802 bits leads to $\phi = 12\%$.

More importantly, our proposed method manipulates the source code of the PDF so that it can be used on top of other conventional methods. For instance, since Kuribayashi et al.'s solution (i.e., StealthPDF) introduces '()' to split the space value [5], it will increase the substrings in the 'TJ' operator, but with more substrings, we can hide more data. As an illustration, when setting $k = 29$ as narrated in Section IV-B, we can continue to embed data in a PDF processed by StealthPDF. Here, we could embed $4,275$ bits, which is a 14-bit improvement over our method alone and a total of $42,435$ bits when combined with StealthPDF. On the other hand, if we combine our method with Jiang et al.'s method [18], we can add 4,275 bits directly to their capacity because their targeted embedding location is the commentary section of the dictionary, which does not affect our embedding process.

B. Limitation and future work

Although our method offers excellent transparency, there are mutual checks and balances on the performance of data hiding, as stated by Deljavan Amiri et al. [19]. Specifically, we are limited in the embedding capacity by the number of 'TJ' operators in the PDF file. In addition, since our approach relies on the 'TJ' operators, it is not available in the face of 'Tj' operators. In the future, we want to investigate more

fonts and other PDF editors' editing behaviour to verify the general applicability of the proposed method. In addition, for the documents encoded using the 'Tj' operator, we will also consider using other examined operators as reported in the literature.

VI. Conclusions

In this work, we proposed a permutation-based method to hide data in PDF files containing text. We employ a novel technique that modifies the order of substrings to encode the bit sequence. Subsequently, we adjust the glyphs of the permuted substrings to restore text width and spacing to render the exact same appearance as the original PDF, hence completely preserving the quality of original host PDF file. At the receiver's end, the hidden data can be extracted, and the original PDF file can be restored when the correct passwords and keys are in possession. Our proposed method stands out for its ability to embed data discreetly without compromising the visual integrity of the document, bolstered by encryption and access control to enhance security. It demonstrates a balanced performance between data hiding capacity and minimal file size overhead, where it requires $2,873$ kB in the PDF file to embed over $21,000$ bits of payload. Our method's compatibility with the conventional data hiding methods expands its utility and security dimensions, making it a versatile tool for various applications, including copyright tracking, secure communication, and digital forensics.

Future research should expand the capabilities of text and multimedia operators in PDFs to fully utilize their potential as dynamic content containers. Additionally, we plan to explore ways to exploit the permuted or semi-randomized text produced by the copy-and-paste operation from our processed PDFs to achieve specific applications.

References

[1] D. Artz, "Digital steganography: Hiding data within data," *IEEE Internet computing*, vol. 5, no. 3, pp. 75–80, 2001.

[2] H.-J. Shiu, K.-L. Ng, J.-F. Fang, R. C. Lee, and C.-H. Huang, "Data hiding methods based upon dna sequences," *Information Sciences*, vol. 180, no. 11, pp. 2196–2208, 2010.

[3] A. Jo, "The promise and peril of generative ai," *Nature*, vol. 614, no. 1, pp. 214–216, 2023.

[4] A. Cakir, "Usability and accessibility of portable document format," *Behaviour & Information Technology*, vol. 35, no. 4, pp. 324–334, 2016.

[5] M. Kuribayashi and K. Wong, "StealthPDF: Data hiding method for PDF file with no visual degradation," *Journal of Information Security and Applications*, vol. 61, p. 102 875, 2021.

[6] K. Wong, M. N. M. Nazeeb, and J.-L. Dugelay, "Complete quality preserving data hiding in animated gif with reversibility and scalable capacity functionalities," in *Digital Forensics and Watermarking*, Cham: Springer International Publishing, 2021, pp. 125–135, ISBN: 978-3-030-69449-4.

[7] J. T. Hackos, "International standards for information development and content management," *IEEE Trans. Professional Communication*, vol. 59, no. 1, pp. 24–36, 2016.

[8] B. A. de Boer, A. T. Soufan, J. Hagoort, *et al.*, "The interactive presentation of 3d information obtained from reconstructed datasets and 3d placement of single histological sections with the 3d portable document format," *Development*, vol. 138, no. 1, pp. 159–167, 2011.

[9] A. Castiglione, A. De Santis, and C. Soriente, "Security and privacy issues in the portable document format," *Journal of Systems and Software*, vol. 83, no. 10, pp. 1813–1822, 2010.

[10] F. Peng, X. Li, and B. Yang, "Improved pvo-based reversible data hiding," *Digital Signal Processing*, vol. 25, pp. 255–265, 2014.

[11] L. Xiang, W. Wu, X. Li, and C. Yang, "A linguistic steganography based on word indexing compression and candidate selection," *Multimedia Tools and Applications*, vol. 77, no. 21, pp. 28 969–28 989, 2018.

[12] L. Li, H.-J. Zhang, J.-L. Meng, and Z.-M. Lu, "Robust pdf watermarking against print–scan attack," *Sensors*, vol. 23, no. 17, p. 7365, 2023.

[13] B. Khosravi, B. Khosravi, B. Khosravi, and K. Nazarkardeh, "A new method for pdf steganography in justified texts," *Journal of information security and applications*, vol. 45, pp. 61–70, 2019.

[14] L. Y. Por, K. Wong, and K. O. Chee, "Unispach: A text-based data hiding method using unicode space characters," *Journal of Systems and Software*, vol. 85, no. 5, pp. 1075–1082, 2012.

[15] I.-S. Lee and W.-H. Tsai, "A new approach to covert communication via pdf files," *Signal processing*, vol. 90, no. 2, pp. 557–565, 2010.

[16] M. Kuribayashi, T. Fukushima, and N. Funabiki, "Robust and secure data hiding for PDF text document," *IEICE Trans. Information and Systems*, vol. 102, no. 1, pp. 41–47, 2019.

[17] D. Lehmer, "Teaching combinatorial tricks to a computer," in *Symposia in Applied Mathematics*, vol. 10, American Mathematics Society, 1960, pp. 179–193.

[18] Z. Jiang, H. Wang, and S. Han, "A robust pdf watermarking scheme with versatility and compatibility," *Multimedia Tools and Applications*, pp. 1–27, 2024.

[19] M. Deljavan Amiri, M. Meghdadi, and A. Amiri, "HVS-based scalable image watermarking: A novel approach to image copyright protection against scalable compression," *Multimedia Tools and Applications*, vol. 78, no. 6, pp. 7097–7124, 2019.

Generation of Photo Slideshow with Song based on Closeness between Concept of Lyrics and That of Images

Mei HASHIMOTO* and Michiharu NIIMI†
* Graduate School of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: hashimoto.mei582@mail.kyutech.jp
† Faculty of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: niimi@ai.kyutech.ac.jp

***Abstract*—This paper proposes a method that allows users to easily convert a large number of still images into movies by displaying photos in sync with memories or favorite songs, which offers a new media viewing method for users to look back at the vast number of photos in their photo folders. We assume here that users select a song and have images stored in local PC. One method of synchronization involves selecting images based on elapsed time of song. However, since lyrics convey meaning, it can be better to display the images that align with both lyrics' meaning and images' meaning matched. In order to do this, we need the space which shares the concept of both lyrics and images, and select colors as the concept. Displayed images are determined based on the distance between points that are mapped from words and images to color space. Specifically, the color concepts from words are extracted using the Color Image Scale, and that from images are determined as the most frequent colors by cluster analysis. Experimental results show that it is possible to easily create movies where images with colors that match the mood of the lyrics are displayed.**

I. Introduction

In recent years, the household penetration rate of mobile devices has exceeded 97%, and their evolution has been remarkable. In particular, smartphones have rapidly spread and are equipped with increasingly advanced features. This enables us to capture high-quality photos of everyday moments as memories and easily share them with friends. The emergence of large-capacity storage and cloud services allows for worry-free photo-taking without concern for storage limitations. Therefore, the number of photos in photo folders continues to increase. However, we rarely look back at these accumulated photos.

Additionally, the use of movie-sharing sites and social media platforms such as YouTube, TikTok, and Instagram has increased rapidly, which provides excellent ways for easily sharing information with other users. Currently, it is common to see users who post dance movies synchronized with popular songs or musics or lyric movies that align photos with the meaning and timing of the lyrics of memorable songs. Combining songs and visuals that match in impression can enhance and emphasize each other's impact. Many users may be interested in content that synchronizes with music and lyrics with natural. Therefore, the ability to easily create visual content that matches music is considered highly valuable. Such synchronized visual content includes still images, therefore, it is one possible approach to utilize the accumulated photos.

There are several tools that users manually synchronize music and photos to create movies. For instance, in Clipchamp of Microsoft, users select BGM from available options, choose image files, and manually set/edit display timing and sequence to create movies. Additionally, there are automated movie creation services such as iPhone's "For You" and "Google Photos." "For You" enhances enjoyment of photos and movies in the iPhone's standard photo library by automatically organizing them based on facial detection and image analysis. Using its "Memories" feature, it automatically picks photos and movies for specific dates, locations, people, or events and creates slideshow movies combined with BGM. Google Photos features an "Auto Creations" function offering many themes like growth records, annual highlights, and dog movies, which automatically select appropriate photos stored within Google Photos to create 10 to 20-second slideshow movies accompanied by themed BGM. However, these tools do not generate movies synchronized with music lyrics.

Therefore, in this study, we propose a method to automatically display photos that match favorite musics with lyrics. It can be a new viewing of photos method for users to look back at a large collection of photos in their photo folders. We aim to develop a method that can easily generate slideshow movies where images switch in sync with lyrics automatically from a set of still images when given a piece of music.

Recently, there are researches which have focused on selecting images based on lyric information. Ishizaki et al.[1] extract keywords (overall impression words) from lyrics and further derive search queries for images from specific lyric lines. They use these search keywords to select similar images available on the web. They leverage the tags associated with web images. This method assumes scraping lyric information from lyric-sharing sites where time information is not attached

to lyrics, thus not enabling synchronized display of lyrics and images as they appear. Moreover, it relies on tagged images. Because tags are not typically put to photo collections in local folder, this approach is unsuitable for our purpose.

Furthermore, there is a research of concept acquisition extracting "meaning" from different media types such as audio and images. Ohishi et al.[6] propose a new framework for extracting targeted speaker conversations from mixed audio using semantic clues. By mapping semantically related images and audio to a shared embedding space, they effectively extract desired audio content.

For the purpose of automatic synchronization between lyrics and images in our study, it is essential to compute the correlation between lyrics of a song and images in a suitable feature space, which is denoted as shared concept space.

For instance, Niiho et al.[5] construct a comprehensive impression word space using canonical correlation analysis based on image and music feature vectors and their impression words. By mapping an unknown image's feature vector into this space, they recommend a music that closely match its impression from a pre-learned set of music.

There is also research converting extracted features from lyrics into visual information. Midorikawa et al.[3] directly correlate two music features: the average of tempo and the ratio of high-frequency to image saturation and brightness. In addition, Murata et al.[4] convert lyrics into color vectors using the Color Image Scale[2], which is denoted as CIS. This scale comprehensively expresses human-standard sensations from colors by using words that express emotion and sensibility adequately, which are referred to as "impression words", CIS describes the relationship of colors and impression words. Several impression words corresponds to one color. They define music color vectors as the color vectors extracted from lyrics using CIS, then "color changes" are generated for the entire music due to lyrics consisting of many word strings. They recommend similar music based on this similarity metric based on the color changes.

In this paper, as the first step of our research, we select a color space as the shared concept space. Specifically, we extract color information from lyrics using CIS and designate the most frequent color from images as their color information. Based on the similarity of this color information, we determine which images should display. Thus, we compare the similarity between lyrics and images in the color space. Firstly, for given a piece of music, we extract lyric data annotated with time information automatically. Next, from the lyric data, we extract impression words and convert them into colors based on CIS. Using this information and matching it with the most frequent color of images, we select images that match the lyrics.

In Section II, we describe the proposed method, then, in Section III, we presents the experiments conducted with our method and their results, and discuss it. Finally, in Section IV, we summarize and discuss future challenges and prospects.

II. Proposed Method

A. System Overview

We have a piece of music with lyrics, that is a song, and a collection of digital photos (Digital Image) taken with digital devices such as smart phones. The music prepared by the user is referred to as M, and the collection of digital photos is denoted as $D = \{d_i | i = 1, 2, \ldots, n\}$($D$ consists of n digital photos).

First, keywords are extracted from M. These are represented as $kw_j (j = 1, 2, \ldots, m)$(where m keywords are extracted from the music M). The time (singing time) when each kw_j is sung from the beginning of the song is denoted as st_j. That is, kw_m is sung after st_m elapsed time from the beginning of the song.

If we denote the function that maps specific keywords to their concept space as $C_{KeyWord}$, and the function that maps image data to their concept space as $C_{DigitalImage}$, for each kw_i,

$$\underset{i}{\text{argmin}} \quad \text{distance}(C_{KeyWord}(kw_j), C_{DigitalImage}(d_i)) \quad (1)$$

we find i that satisfies the above equation. Here, distance(x,y) represents the distance between x and y. By performing this process for all kw_j, each kw_j is associated with the one image d_i that is closest in the concept space.

Utilizing the information of st_j, we can arrange d_i on the time axis. When we make a movie from the set of images with M as the BGM, images that correspond to the concepts of the lyrics would be displayed following lyrics' meanings. This is the image display system that we propose in this study, which aims to synchronize the display of images with the concepts of the lyrics by leveraging the proximity of concepts between the lyrics and the images.

We set the concept for songs, which are actually lyrics, and images, and construct an image display system based on the similarity in the concept space. Fig.1 illustrates the outline of our proposed system. Similarly to Murata et al.[4], the shared concept space, which is the common concept to be extracted from both songs and images, is set as "color," The color concept from the lyrics is generated based on CIS, and the color concept from the images is defined as the most frequent color in the respective image.

Each process is explained in detail below.

1) $C_{KeyWord}$: The function $C_{KeyWord}$, which maps specific keywords extracted from music lyrics into a shared concept space, consists of three components: music_spleeter, whisper and music_concept. In music_spleeter, we use a library called spleeter, which can split the sounds of each instrument in the music using machine learning, to separate them into Vocal, Piano, Bass, Drums, and other. Then, vocal audio files are input into the Whisper speech recognition model. As a result of this, we can get lyrics from the song. The lyrics is consisting of word strings which are complete sentence and/or phrase making a certain meaning, and we call it a quasi-sentence. From a lyrics, we can extract multiple quasi-sentences. In addition, the Whisper output the starting

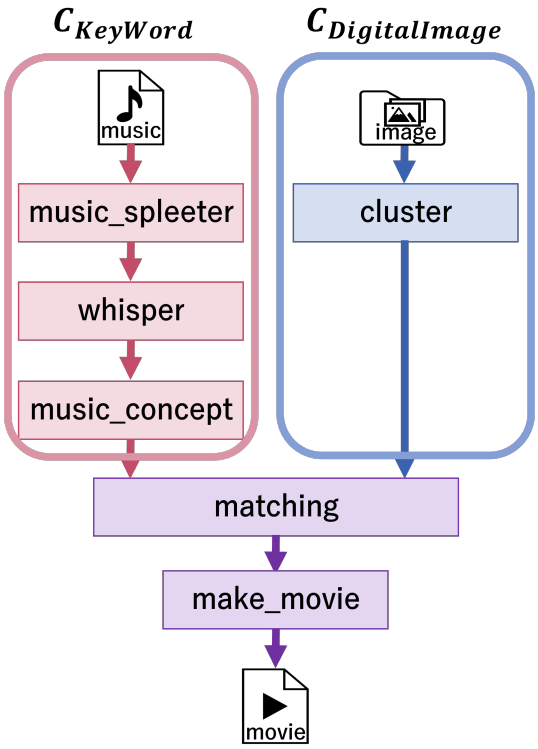


Fig. 1: Overview of the Image Display System

time of each quasi-sentence. For each quasi-sentence, we employ morphological analysis with mecab to extract nouns, verbs, and adjectives which are regarded as keywords for the quasi-sentence. Finally, a quasi-sentence is mapped into a impression word of CIS. This is done by the followings in detail. A quasi-sentence includes multiple keywords, and there are multiple impression words we pre-defined to use our system. We calculate a similarity between all keywords and all impression words. Based on this calculation, we extract a pair of keyword and impression word that minimizes the distance in shared concept space, call the keyword of the pair "a representative word" of quasi-sentence.

2) $C_{DigitalImage}$: The function $C_{DigitalImage}$ mapping image data to the concept space performs clustering using the K-means method. The number of clusters is set to five, and the cluster center of the cluster with the largest number of data points is extracted as the most frequent color of the image data. This color is then mapped to the concept space to set the concept.

3) matching: After extracting the concepts from the lyrics and the images, we link each keyword to the image that minimizes the distance in the shared concept space between the lyrics and the images by following equation (1). This process is performed for all extracted lyric concepts. Using the timestamps of the lyrics, the image data linked to the lyrics are arranged accordingly on the time axis.

4) make_movie: Based on the arranged image data, a movie is created by displaying images along with the music M. To minimize the discrepancy between the appearance of words and the display of images, fade-in and fade-out image processing is applied. The crossfadein and crossfadeout functions from the MoviePy library are used for the fade-in and fade-out effects. The current image is faded out over one second while the next image is faded in simultaneously. To ensure that images can be properly appreciated and are not difficult to see, images with a display time of three seconds or less will not undergo fade-in and fade-out processing.

TABLE I: Used songs

Melody	Song title/artist	Performance time
bright song	shinning star / shiho	4m42s
sad song	Diamond Dust-tribute to Frozen- / aruku	4m18s
refreshing song	Success! / TOYro	1m24s

III. Experiment

A. Experimental condition

In this experiment, we used three free music tracks : bright song, sad song, and refreshing song, as shown in Table I. We also collected 524 images taken by students in my laboratory. Additionally, we selected 27 categories of impression words. The relationship between the impression words and colors in the shared concept space is shown in Table II.

B. Results

Using the proposed system, we were able to extract concepts from both the music and the images, and create movies using images from the image folder. The results of this experiment are presented for each structure based on the overview of the system in Section II-A. Additionally, the movies created were shared within the member of my laboratory, and evaluated it. We also describe this evaluation.

1) Song concept extraction: In music_spleeter, we were able to separate the song into Vocal, Piano, Bass, Drums, and other components. The Vocal audio files were transcribed into lyrics using Whisper, and morphological analysis was performed using mecab, then mapped to the shared concept space. For the multiple keywords obtained from the lyrics, the most similar representative word was mapped to the shared concept space to obtain the concept (color). Table III, Table IV and Table V show the results of the concept extraction for the chorus parts of the three songs used in this experiment.

2) Image concept extraction: For 524 images, cluster analysis was performed using the K-means method to extract the most frequent color representing the concept of each image. Table VI shows the images subjected to concept extraction, the most frequent colors obtained through cluster analysis, and the concepts (colors) obtained by mapping these most frequent colors to the shared concept space.

3) Concept matching: By following the proposed steps, we were able to obtain the concepts from both quasi-sentences of the song and the images. Using the obtained concepts, we linked the images with quasi-sentences that distributes

TABLE II: Impression words and their corresponding colors [2]

Impression words	Color (R,G,B)
cute, innocent, childlike, charming, pretty	(239,143,184)
happy, joyful, cheerful, energetic, casual	(255,88,80)
splendid, bustling, vivid, flashy	(192,0,112)
active, dynamic, bold, exciting, passionate	(0,140,113)
dynamic,energetic,intense,forceful	(16,0,32)
gorgeous, glamorous, sexy, rich, luxurious	(132,48,143)
wild, powerful, tough, robust, manly	(0,117,62)
lovely, sweet, fresh	(255,212,200)
mild, familiar, gentle, pleasant	(255,213,159)
natural, honest, relaxed, simple, unhurried	(198,205,156)
graceful, feminine, emotional, delicate, beautiful	(219,149,173)
classic, nostalgic, antique, traditional, meticulous	(104,0,31)
strong, solid, heavy	(0,60,60)
modest, delicate, feminine, elegant	(197,184,199)
elegant, humble, dignified	(165,129,145)
authentic, dignified, refined, robust	(85,0,53)
romantic, fantastical, pale, pure, neat	(225,236,255)
juicy, reserved, plain, peaceful	(217,253,255)
lively, healthy, fresh, safe	(218,248,109)
quiet, subtle, reserved, nonchalant, simple	(124,152,156)
smart, stylish, intellectual, calm, subdued	(0,50,117)
dandy, gentlemanly, masculine, serious, steady	(0,71,91)
formal, noble, sacred, solemn	(107,107,98)
brisk, clean, pure, clear, shipshape	(138,176,223)
refreshing, youthful, young, sporty	(91,189,206)
modern, speedy, innovative, agile, intelligent	(54,119,133)
mechanical, precise, rational, artificial	(161,161,148)

with near distance in shared concept space. We were then able to arrange the appearance times of the lyrics and the corresponding image file names as data on the timeline.

4) Make movie: Following the timeline data obtained by the above Section III-B3 (concept matching), we were able to create a movie synchronized with the music by displaying the images in the order aligned with the music as the BGM (Background Music). For the created movie, the keywords from the lyrics, impression words, and the images displayed at those times are shown in Table VII, VIII and IX based on III-B1) Music Concept Extraction.

5) Subjective evaluation: The created movie was shared within the member of my laboratory, and evaluated with subjectively. The summarized results are presented below.

- Regarding synchronization between the music and images, there are instances where images aligned well with the concept and mood of the lyrics. But there are also parts where the relationship between lyrics and images is unclear.
- To improve the quality of the movie and clarify the relationship between music and images, it would be beneficial to display lyrics when images corresponding to the concept of the lyrics are shown.
- Although the purpose of this system is different, it my be interesting using AI-generated images based on impression words and lyrics.

TABLE III: Concept of song:shinning star

Keyword	Representative word	Impression word	Color (R,G,B)
wind, sound, confine	sound	mild	(255,213,159)
dream, sleep, illusion	dream	cute	(239,143,184)
world	world	nature	(198,205,156)

TABLE IV: Concept of song:Diamond Dust-tribute to Frozen-

Keyword	Representative word	Impression word	Color (R,G,B)
sun, see, scenery, forget, white, diamond	scenery	romantic	(225,236,255)
living, proof, give	proof	safety	(218,248,109)
beside, stay	beside	antique	(104,0,31)

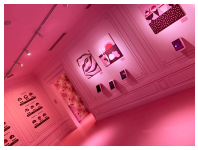
C. Discussion

Using the proposed model, we synchronized image display with music based on the proximity of lyrical and image concepts. We were able to display images that matched the mood of the lyrics in the music. In this study, Despite setting only color as a concept, there were several instances where subjects matching lyrical words were displayed. Specifically, in Table IV, the keyword “scenery” was assigned a light blue concept. When selecting images that matched the concept of the lyrics from the image folder at random, a photo depicting the sky, sea, and a brightly shining white sun was coincidentally chosen. This means that we can select images that perfectly matched the lyrics “day, scenery, forget, white diamond.” However,

TABLE V: Concept of song:Success!

Keyword	Representative word	Impression word	Color (R,G,B)
tomorrow, extend, grasp	tomorrow	gentle	(255,213,159)
future, hand	hand	strong	(0,60,60)
match, time, suddenly	match	clear	(138,176,223)

TABLE VI: Concept of image

Image	Image label	Most frequent color (R,G,B)	Concept (R,G,B)
	pink room	(182,78,100)	(192,0,112)
	five-storied pagoda	(159,44,19)	(104,0,31)
	strawberry	(106,120,81)	(107,107,98)
	fireworks	(7,9,7)	(0,60,60)
	baby blue eyes	(73,166,227)	(91,189,206)

challenges arise when there are prominent subjects in the image or when keywords like “nature” have multiple meanings. This leads to the selection of images that do not align with our intended imagery. For instance, photos containing family, friends, lovers, or pets in the image folder may not match well with lyrical words appearing in the lyrics. Additionally, the concept varies depending on whether it is mapped to the shared concept space as “nature” in general or as “natural” in a specific sense. Therefore, it is necessary to set keywords considering not only individual words but also their context.

Furthermore, considering insights gleaned from comments on the created movie, potential improvements are discussed below.

Based on the data files storing the conceptual data of the songs and images created in this experiment, it appears that the concepts for both songs and images were appropriately set. However, there are three potential reasons for the unclear relationship between quasi-sentences and images in certain parts of the created movie.

(1) It may be insufficient for extracting image concepts. In this study, the most frequent color in the image is used as the feature for image representation. Therefore, as seen in photos of strawberries and fireworks in Table VI, the extracted color represents the background color rather than the color of the strawberries or fireworks that we want to focus on. This misalignment can lead to attention being drawn away from our intended image and impressions. In the future, incorporating methods that specify attention positions could enable the creation of movies more synchronized with the music.

(2) It may be insufficient for matching the concepts of lyrics and images. In this paper, the system is set to select the image with the closest associated impression word even though the distance is far in the shared concept space. This means that the selected image may be close to other impression words. Therefore, it may cause a sense of incongruity by displaying an image with a different concept. In the future, it is necessary to increase the number of images to ensure that there is always an image that matches the concept.

(3) It would be insufficient the number of impression words and colors in the shared concept space. In this study, we have set 27 categories of impression words based on CIS. By increasing the number of impression words and colors, it would be expected higher accurate classification in shared concept space. Furthermore, CIS links the impression words of emotion and sensibility to colors based on a standard human perception of color. Therefore, since adjectives are set as impression words rather than nouns or verbs, there may be issues in similarity word

TABLE VII: Keyword and displayed image:shinning star

Displayed image	Keyword	Impression word
	wind	mild
	dream	cute
	world	nature

TABLE VIII: Keyword and displayed image:Diamond Dust-tribute to Frozen-

Displayed image	Keyword	Impression word
	scenery	romantic
	proof	safety
	beside	stay

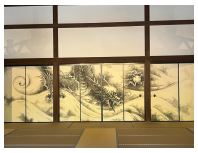
calculations using Word2Vec.

Additionally, as mentioned in the comments, it may enhance our proposed system by displaying lyrics on the images.

IV. Conclusion

In this paper, we proposed a method to automatically generate slideshow synchronized with song by using concepts for lyrics and images, and utilizing their proximity. The experimental results demonstrated that by extracting and matching concepts between lyrics and images, we confirmed that slideshows synchronized with music are generated, which displays images that matched the lyrics at the time of their occurrence. However, there are several areas for improvement, such as refining the shared concept space and the method of image concept extraction, to achieve better synchronization with music and higher slideshow quality. In addition, it is also important future task to find methods for quantitative evaluation.

TABLE IX: Keyword and displayed image:Success!

Displayed image	Keyword	Impression word
	tomorrow	gentle
	hand	strong
	match	clear

References

[1] Hiromi Ishizaki et al. “Music Slideshow Generation Based on Web Image Retrieval with Queries Constructed from Lyrics (in Japanese)”. In: *IPSJ Journal* 54.4 (2013), pp. 1263–1274.

[2] Shigenobu Kobayashi. *Color Image Scale*. Distributed in the U.S. by Kodansha America, 1992.

[3] Nozomi Midorikawa, Toru kano, and Yuriko Takeshima. “Music Visualization Based on Lyrics and Acoustic Features (in Japanese)”. In: *ITE Technical Report* 41.12 (2017), pp. 271–272.

[4] Ken Murata et al. “A method for recommending songs based on time series similarity based on the story development of lyrics (in Japanese)”. In: *Proceedings of 84th National Convention of IPSJ*. 2022, pp. 525–526.

[5] Ryutaro Niho and Yasuyuki Saito. “A study of automatic recommendation system of music matching image impesslon”. In: *ITE Technical Report Vol.37, No.7*. 2013, pp. 23–26.

[6] Yasunori Ohishi et al. “Conceptbeam: Concept driven target speech extraction”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 4252–4260.

Disposable-key-based image encryption for collaborative learning of Vision Transformer

Rei Aso*, Sakaya Shiota† and Hitoshi Kiya‡
* Tokyo Metropolitan University, Tokyo
E-mail: aso-rei@ed.tmu.ac.jp
† Tokyo Metropolitan University, Tokyo
E-mail: sayaka@tmu.ac.jp
‡ Tokyo Metropolitan University, Tokyo
E-mail: kiya@tmu.ac.jp

***Abstract*—We propose a novel method for securely training the vision transformer (ViT) with sensitive data shared from multiple clients similar to privacy-preserving federated learning. In the proposed method, training images are independently encrypted by each client where encryption keys can be prepared by each client, and ViT is trained by using these encrypted images for the first time. The method allows clients not only to dispose of the keys but to also reduce the communication costs between a central server and the clients. In image classification experiments, we verify the effectiveness of the proposed method on the CIFAR-10 dataset in terms of classification accuracy and the use of restricted random permutation matrices.**

I. Introduction

Deep neural networks (DNNs) have been deployed in various applications. Training a high-quality DNN model requires a huge amount of training data, but model training raises privacy concerns, especially when dealing with sensitive or personal information [1]. To collect a huge amount of training data, in centralized machine learning, each client prepares information on local data and sends it to a central server for model training. Federated learning (FL) [2], [3] is a learning approach that allows multiple clients to collaboratively train a shared model while keeping their data decentralized and private. In FL, clients participate in the training process using their local data, without sharing their raw data. However, the iterative learning process in FL generates some issues in terms of learning costs for clients and communication costs in a distributed environment [4], [5]. To overcome these issues, we propose a novel approach in which multiple clients efficiently collaborate to train a high-quality model.

In privacy-preserving collaborative learning, each client collects data and securely sends the data or an update obtained from the data to a central server for model training. In FL, in each round, each client independently computes an update to the current model based on its local data and communicates this update to a central server so that the client-side updates are aggregated to compute a new global model. Accordingly, conventional collaborative learning including FL requires both resources for computing the update in each client and communication efficiency [1]. In this paper, we propose a novel method for privacy-preserving collaborative learning to reduce these costs. The method is carried out by using encrypted data, called learnable encryption [6] where data is independently encrypted by each client. The encrypted data is sent to a central server for model training one time only, and a global model is trained by using the encrypted data from clients, so the method allows us not only to reduce the communication costs but to also avoid the cost for computing an update at all clients. We make the following contributions in this paper.

(a) We propose a novel method using learnable encryption for privacy-preserving collaborative learning that can reduce both communication costs and computation sources of clients.
(b) We propose the use of restricted random permutation matrices to reduce the influence of data encrypted with independent keys.

In the proposed method, different random permutation matrices used as encryption keys are applied to each piece of data by each client, but this causes the performance of models to degrade. To improve this issue, we also consider the use of restricted random permutation matrices. In experiments, the effectiveness of the method is verified in an image classification task under the use of the vision transformer (ViT) [7].

II. Related Work

The proposed method will be discussed under the use of ViT and learnable image encryption, which are summarized here.

A. Vision transformer

ViT [7] is commonly used in image classification tasks and is known to provide a high classification performance. As shown in Fig. 1, in ViT, an input image $x \in \mathbb{R}^{h \times w \times c}$ is divided into N patches with a size of $p \times p$, where h, w, and c are the height, width, and number of channels of the image. Also, N is given as hw/p^2. Afterward, each patch is flattened into $x_p^i = [x_p^i(1), x_p^i(2), \ldots, x_p^i(L)]$. Finally, a sequence of embedded patches is given as

$$z_0 = [x_{class}; x_p^1\mathbf{E}; x_p^2\mathbf{E}; \ldots x_p^i\mathbf{E}; \ldots x_p^N\mathbf{E}] + \mathbf{E_{pos}}, \quad (1)$$

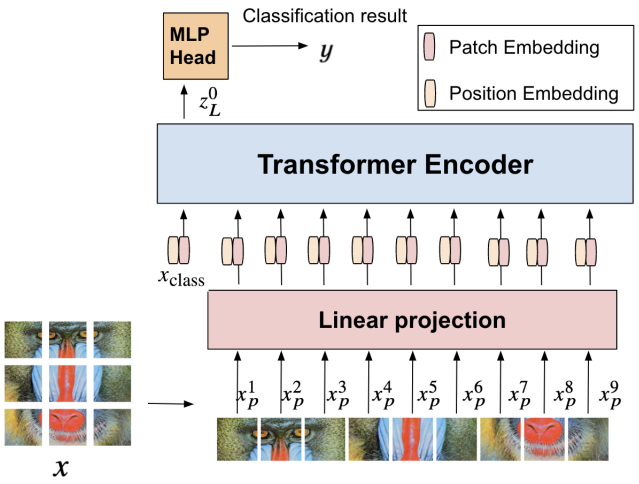


Fig. 1: Overview of ViT

where

$$\begin{aligned}\mathbf{E_{pos}} =&((e^0_{pos})(e^1_{pos})\dots(e^i_{pos})\dots(e^N_{pos})),\\ L =&p^2c, x_{class}\in\mathbb{R}^D,\ x^i_p\in\mathbb{R}^L,\ e^i_{pos}\in\mathbb{R}^D,\\ \mathbf{E}\in&\mathbb{R}^{L\times D},\ \mathbf{E_{pos}}\in\mathbb{R}^{(N+1)\times D}.\end{aligned}$$

x_{class} is a class token, $\mathbf{E}$ is an embedding (patch embedding) that linearly maps each patch to dimensions D, $\mathbf{E_{pos}}$ is an embedding (position embedding) that gives position information to patches in the image, e^0_{pos} is the position information of a class token, e^i_{pos} is the position information of each patch, and z_0 is a sequence of embedded patches. Afterward, z_0 is input into the transformer encoder. The encoder outputs only the class token, which is a vector of condensed information on the entire image, and it is used for classification.

Previous studies have indicated that when models are trained with images encrypted by one client, the performance of the models is degraded compared with models trained with plain images [8], [9]. If multiple clients collaboratively train models by using images encrypted with independent keys, the performance of the models will be more degraded than that of one client. Accordingly, we propose a novel method that can improve the performance degradation even when multiple clients collaboratively train a model by using images encrypted with independent keys.

B. Learnable image encryption

Various image transformation methods that use a secret key, often referred to as perceptual image encryption, have been studied so far for many applications [6]. In this paper, we focus on learnable images transformed with a secret key, which have been studied for deep learning. Learnable encryption enables us to directly apply encrypted data to a model as training and testing data. Encrypted images have no visual information on plain images in general, so privacy-preserving learning can be carried out by using visually protected images. In addition, the use of a secret key allows us to embed unique features controlled with the key into images. Adversarial defenses [10]–[12] and access control [13], [14] are carried out with encrypted data using these unique features.

A block-wise learnable image encryption method (LE) with an adaptation layer was introduced [15] as the first learnable image encryption method, and then another encryption method, a pixel-wise encryption (PE) method that does not use any adaptation layer, was proposed [16]. However, both encryption methods are not robust enough against ciphertext-only attacks, as reported in [17], [18]. To enhance the security of encryption, LE was extended to an extended learnable image encryption method (ELE) by adding a block scrambling (permutation) step and a pixel encryption operation with multiple keys [19]. However, ELE still has inferior accuracy compared with using plain images, even when an additional adaptation network is used to reduce the influence of the encryption. Recently, block-wise encryption was also pointed out to have a high similarity to isotropic networks such as ViT and ConvMixer [9], [20], and the similarity enables us to reduce performance degradation. However, no conventional learnable encryption methods have been designed for privacy-preserving model training with multiple clients.

III. Proposed Method

A. Overview

Fig. 2 shows an overview of the proposed method, where datasets that M clients have are shared to train a global model. In the framework, a pretrained model is fine-tuned by using the shared datasets. Below is the procedure for model training and testing.

1 Each client prepares keys for image encryption and carries out image encryption with the keys.
2 Each client sends the encrypted images to a central server.
3 A global model is fine-tuned by using the encrypted images on the server.
4 Each client receives the trained model from the server.
5 Each client encrypts a test image by using a key and inputs it to the model to obtain an estimation result.

In this framework, the server cannot get both visual information of images and the keys. In addition, each client can use a key that is different from other clients' keys. Clients may also change keys for each image.

B. Image encryption with random permutation matrices

ViT has a high similarity to block-wise encryption [8], so the proposed method is carried out using block-wise encryption, which consists of block scrambling (block permutation) and pixel permutation. Below is the procedure for block scrambling, in which $\mathbf{E_{bs}}$ is a random permutation matrix for block

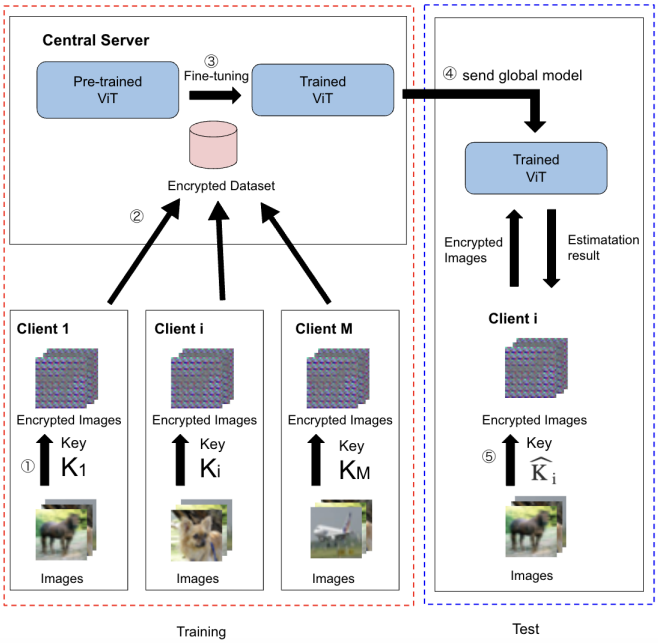


Fig. 2: Framework of proposed method

permutation, given by

$$\mathbf{E_{bs}} = \begin{pmatrix} E_{bs}(1,1) & \dots & E_{bs}(1,j) & \dots & E_{bs}(1,N) \\ \vdots & & \vdots & & \vdots \\ E_{bs}(i,1) & \dots & E_{bs}(i,j) & \dots & E_{bs}(i,N) \\ \vdots & & \vdots & & \vdots \\ E_{bs}(N,1) & \dots & E_{bs}(N,j) & \dots & E_{bs}(N,N) \end{pmatrix}. \quad (2)$$

(a) Divide an image $x \in \mathbb{R}^{h \times w \times c}$ into N non-overlapped blocks with a size of $p \times p$ such that $\mathbf{B} = \{B_1, B_2, ..., B_N\}^\top$, $B_i \in \mathbb{R}^{p^2 c}$, where p is also equal to the patch size of ViT.

(b) Generate a random integer sequence with a length of N by using a key as

$$l = [l(1), \cdots l(i), \cdots l(N)] \quad (3)$$

where

$$l(i) \in \{1, 2, \cdots, N\},$$
$$l(i) \neq l(j) \text{ if } i \neq j, \ i, j \in \{1, 2, \cdots, N\}.$$

(c) Define $\mathbf{E_{bs}}$ in Eq.(2) as

$$E_{bs}(i,j) = \begin{cases} 0 & (l(j) \neq i) \\ 1 & (l(j) = i) \end{cases}. \quad (4)$$

(d) Define a vector b as

$$b = \{1, 2, \cdots, N\}, \quad (5)$$

and transform it with $\mathbf{E_{bs}}$ as

$$\hat{b} = b\mathbf{E_{bs}} \quad (6)$$
$$= \{\hat{b}(1), .., \hat{b}(i).., \hat{b}(N)\}, \ \hat{b}(i) \in \{1, 2, .., N\}.$$

(e) Give permutated blocks $\hat{\mathbf{B}} = \{\hat{B}_1, \hat{B}_2, ..., \hat{B}_N\}^\top$ as

$$\hat{B}_i = B_{\hat{b}(i)}. \quad (7)$$

In the proposed method, each client carries out the above procedure with their keys. Next, the procedure for pixel permutation is explained. In the method, R, G, and B values in each block are randomly permutated. A random permutation matrix for pixel permutation $\mathbf{E_{ps}}$, which is applied to each block, is expressed as

$$\mathbf{E_{ps}} = \begin{pmatrix} E_{ps}(1,1) & \dots & E_{ps}(1,j) & \dots & E_{ps}(1,L) \\ \vdots & & \vdots & & \vdots \\ E_{ps}(i,1) & \dots & E_{ps}(i,j) & \dots & E_{ps}(i,L) \\ \vdots & & \vdots & & \vdots \\ E_{ps}(L,1) & \dots & E_{ps}(L,j) & \dots & E_{ps}(L,L) \end{pmatrix}, \quad (8)$$

where $L = p^2 C$.

Below is the procedure for the permutation.

(a) Generate a random integer sequence with a length of L by using a key as

$$u = [u(1), \cdots u(i), \cdots u(L)]. \quad (9)$$

where

$$u(i) \in \{1, 2, \cdots, L\},$$
$$u(i) \neq u(j) \text{ if } i \neq j, \ i, j \in \{1, 2, \cdots, L\}.$$

(b) Define $\mathbf{E_{ps}}$ in Eq.(8) as

$$E_{ps}(i,j) = \begin{cases} 0 & (u(i) \neq j) \\ 1 & (u(i) = j) \end{cases}. \quad (10)$$

(c) Vectorize the elements of each block B_i as $x_p^i \in \mathbb{R}^L$, and transform it with $\mathbf{E_{ps}}$ with

$$x'^i_p = x_p^i \mathbf{E_{ps}}. \quad (11)$$

(d) Concatenate the transformed vectors into an encrypted image block $\hat{B}_i$.

C. Use of restricted random permutation matrices

A permutation matrix is a square binary matrix that has exactly one entry of 1 in each row and each column with all other entries 0. Every permutation matrix is orthogonal, with its inverse equal to its transpose. As describe above, two permutation matrices, $\mathbf{E_{bs}}$ and $\mathbf{E_{ps}}$, are used for image encryption in the proposed method.

Three types of permutation matrices are considered. The performance of trained models can be managed by the type of permutation matrix. For $N = 5$, an example with $\mathbf{E_{bs}}$ is given below, where * indicates a fixed element.

(A) Identity matrix ($N_{bs} = N$):

$$\mathbf{E_{bs}} = \begin{pmatrix} 1^* & 0 & 0 & 0 & 0 \\ 0 & 1^* & 0 & 0 & 0 \\ 0 & 0 & 1^* & 0 & 0 \\ 0 & 0 & 0 & 1^* & 0 \\ 0 & 0 & 0 & 0 & 1^* \end{pmatrix}. \quad (12)$$

If E_{bs} is the identity matrix of $N \times N$, in which all its diagonal elements equal 1, and 0 everywhere else, no

permutation is carried out. In this type, the number of fixed diagonal elements N_{bs} is given as $N_{bs} = N$.

(B) Restricted random permutation matrix ($0 < N_{bs} < N$): $N_{bs} < N$ diagonal elements are fixed at a value of 1, where the positions of the fixed elements are randomly selected. When using $0 < N_{bs} < N$, $\mathbf{E_{bs}}$ is called a restricted random permutation matrix. For $N_{bs} = 2$,

$$\mathbf{E_{bs}} = \begin{pmatrix} 1^* & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1^* & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{13}$$

(C) Unrestricted random permutation matrix ($N_{bs} = 0$): The positions of N elements are randomly selected as

$$\mathbf{E_{bs}} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{14}$$

The above equations correspond to block scrambling (permutation) operations in each image. For $N_{bs} = N$, Eq. (13) is reduced to Eq. (12), and Eq. (13) is reduced to type (C) for $N_{bs} = 0$. Similarly, $\mathbf{E_{ps}}$ is also classified into three types, so $N_{ps} < L$ diagonal elements are fixed as in Eq. (13). The transformation with $\mathbf{E_{ps}}$ is called pixel shuffling or pixel permutation. Fig. 3 shows an example of encrypted images, where $N = 196$ and $L = 768$ were used.

D. Properties of proposed method

The properties of the proposed method are summarized below.

- Each client can share their images encrypted with their keys to train a global model.
- The performance degradation of models caused by encrypted images can be reduced by using the embedding structure of ViT and restricted permutation matrixes.
- Transmission between clients and the central server is carried out to share encrypted images only one time during model training, so the communication cost of the proposed method is lower than conventional methods such as federated learning.

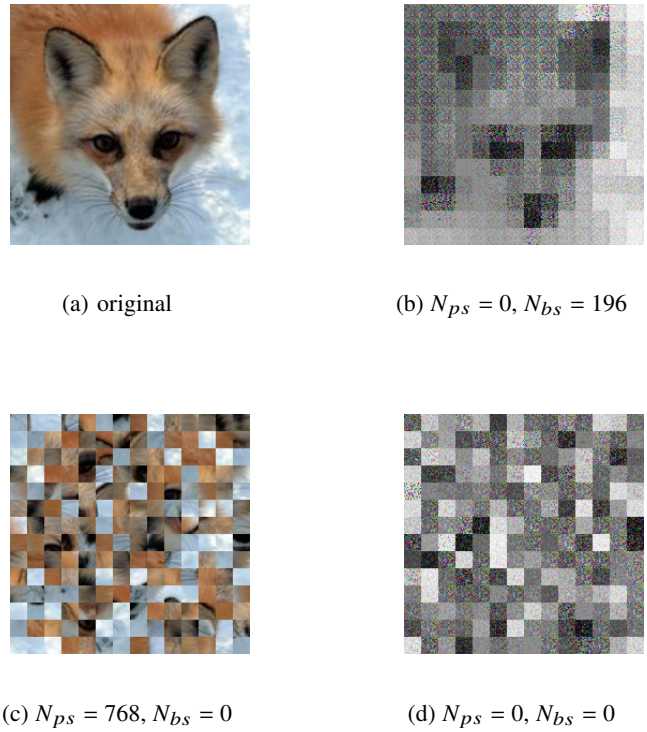
In addition to the above properties, each client can use different keys for each image. Accordingly, the keys used for image encryption are not required to be managed carefully.

TABLE I: Classification accuracy(%)

	w/o encryption	common key	different key ($N_{bs} = 0, N_{ps} = 0$)
acc(%)	97.68	88.1	82.71

IV. Experiment Result

The effectiveness of the proposed method was verified in experiments.



(a) original (b) $N_{ps} = 0$, $N_{bs} = 196$

(c) $N_{ps} = 768$, $N_{bs} = 0$ (d) $N_{ps} = 0$, $N_{bs} = 0$

Fig. 3: Example of encrypted images

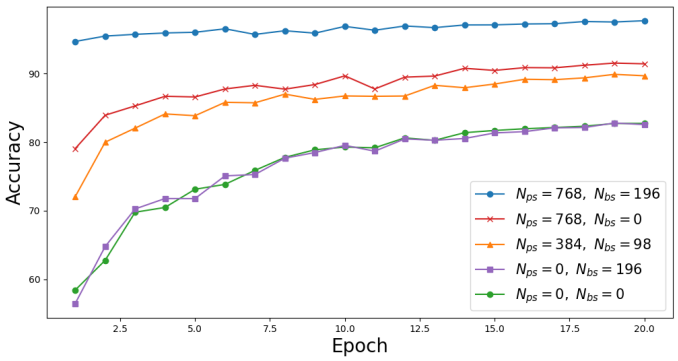


Fig. 4: Learning curve of training process

A. Setup

In experiments, a model was prepared that was pre-trained trained with Image-Net, "vit_base_patch16_224," from the timm library, and the CIFAR10 dataset, which consists of 50,000 training and 10,000 test color images with a size of 32×32, was used to fine-tune the pre-trained model. Each client was given 10,000 randomly selected images as training data without duplicates, where images were resized from $32 \times 32 \times 3$ to $224 \times$

TABLE II: Accuracy of using restricted matrix

Image (N_{bs}, N_{ps})	Accuracy (%)
Baseline (plain)	97.02
Proposed (0, 768)	91.39
Proposed (147, 576)	90.69
Proposed (49, 576)	90.66
Proposed (98, 384)	89.65
Proposed (147, 192)	88.33
Proposed (49, 192)	88.19
Proposed (196, 0)	82.53
Proposed (0, 0)	82.71

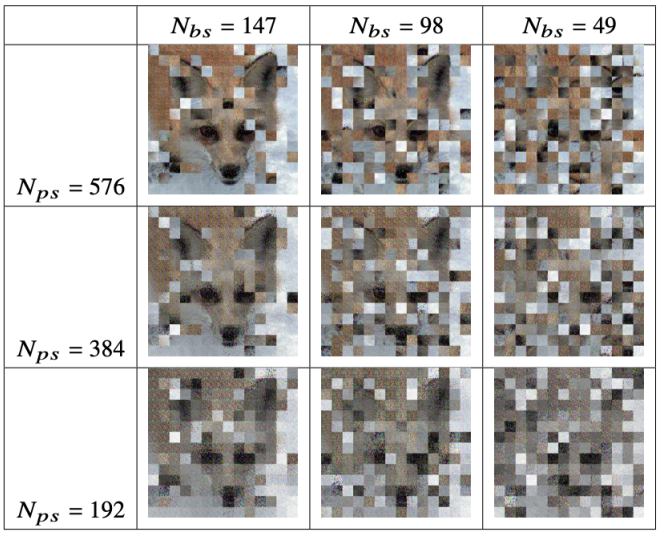


Fig. 5: Example of encrypted images using restricted random permutation matrix

224×3 to fit the size of images to that of ViT. Fine-tuning was carried out under the use of a batch size of 64, a learning rate of 0.0001, a momentum of 0.9, and a weight decay of 0.0005 using the stochastic gradient descent (SGD) algorithm for 20 epochs. A cross-entropy loss function was used as the loss function. We evaluated the classification accuracy by inputting the 10,000 test images to the final global model. In the setting of ViT, the patch size p in patch embedding was set to 16, the number of split patches in an input image was $N = 196$, and the dimensionality of output feature vectors was $D = 384$ and $L = 768$. In addition, independent keys were assigned to all images every epoch.

B. Classification accuracy

In Table I, the classification accuracy of the global model using the proposed method is compared with that of the baseline method without encryption to confirm the influence of encryption conditions, where common key indicates that all clients used the same keys, and different key means that independent keys generated with $N_{bs} = N_{ps} = 0$, which were to use unrestricted random permutation matrices, were applied to all images. From the table, even when different keys were applied to images, the proposed method still allowed us to maintain a high classification accuracy. However, the accuracy of the models decreased due to the influence of image encryption. Table II shows the result of using restricted random permutation matrices. Fig. 4 shows the learning curves during training. From these results, the use of restricted random permutation matrices was demonstrated to be effective in improving the accuracy of the models. Some restricted conditions outperformed the method using a common key.

C. Visibility of encrypted images

Fig. 5 shows an example of encrypted images under restricted conditions. From the figure, a less restricted condition gave encrypted images with higher visual information than the original ones. In contrast, with a less restricted condition, a model could be trained with a higher quality in general. Accordingly, restricted permutation matrices should be selected carefully on the basis of classification accuracy and visual protection.

D. Security analysis

In collaborative learning, the privacy of all local data has to be protected. In this framework, both the server that trains the global model and other clients are assumed to be untrusted, so they may try to restore original data from encrypted data. The objective of an attacker is to restore visual information from encrypted images. We assume that the attacker has access to encrypted images and the encryption algorithm but does not possess the secret keys. Accordingly, the attacker can only carryout ciphertext-only attacks (COAs) using encrypted images. Block-wise image encryption methods have been studied in terms of robustness against various COAs including state-of-the-art ones [18], [21]. The proposed method allows clients to assign independent keys to each image. Therefore, when applying independent keys to images, encrypted images are more robust against state-of-the-art attacks than images encrypted with a common key.

V. Conclusion

In this paper, we proposed a novel method for collaborative learning with ViT. The method, which uses encrypted images, allows clients not only to assign different keys for encryption but to also use restricted random matrices to improve the accuracy of models. Compared with conventional methods such as FL, the communication costs and computation resources of clients can be reduced by using our method. In experiments, the effectiveness of the method was demonstrated in terms of accuracy and visual protection.

Ackowledgment

This study was partially supported by JSPS KAKENHI (Grant Number JP21H01327) and JST CREST (Grant Number JPMJCR20D3).

References

[1] S. Park and J. C. Ye, "Multi-task distributed learning using vision transformer with random patch permutation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, pp. 2091–2105, 2023. DOI: 10.1109/TMI.2022.3218783.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54, 2017, pp. 1273–1282.

[3] K. Rajkumar, A. Goswami, K. Lakshmanan, and R. Gupta, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3922–3924, 2022. DOI: 10.1109/TIFS.2022.3214717.

[4] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106 775, 2021.

[5] H. Sawada, S. Imaizumi, and H. Kiya, *Enhancing security using random binary weights in privacy-preserving federated learning*, 2024. arXiv: 2409.19988 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2409.19988.

[6] H. Kiya, A. P. Maung Maung, Y. Kinoshita, S. Imaizumi, and S. Shiota, "An overview of compressible and learnable image transformation with secret key and its applications," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, e11, 2022.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7*, OpenReview.net, 2021.

[8] M. AprilPyone and H. Kiya, "Privacy-preserving image classification using an isotropic network," *IEEE MultiMedia*, vol. 29, no. 2, pp. 23–33, 2022. DOI: 10.1109/MMUL.2022.3168441.

[9] Z. Qi, A. MaungMaung, Y. Kinoshita, and H. Kiya, "Privacy-preserving image classification using vision transformer," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 543–547. DOI: 10.23919/EUSIPCO55093.2022.9909972.

[10] M. AprilPyone and H. Kiya, "Block-wise image transformation with secret key for adversarially robust defense," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2709–2723, 2021.

[11] R. Iijima, S. Shiota, and H. Kiya, "A random ensemble of encrypted vision transformers for adversarially robust defense," *IEEE Access*, vol. 12, pp. 69 206–69 216, 2024. DOI: 10.1109/ACCESS.2024.3400958.

[12] M. Maung, A. Pyone, and H. Kiya, "Encryption inspired adversarial defense for visual classification," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1681–1685. DOI: 10.1109/ICIP40778.2020.9190904.

[13] H. Kiya, R. Iijima, A. MaungMaung, and Y. Kinoshita, "Image and model transformation with secret key for vision transformer," *IEICE Transactions on Information and Systems*, vol. E106-D, no. 1, pp. 2–11, 2023. DOI: 10.1587/transinf.2022MUI0001.

[14] A. Maungmaung and H. Kiya, "A protection method of trained cnn model with a secret key from unauthorized access," *APSIPA Transactions on Signal and Information Processing*, vol. 10, e10, 2021. DOI: 10.1017/ATSIP.2021.9.

[15] M. Tanaka, "Learnable image encryption," in *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2018, pp. 1–2. DOI: 10.1109/ICCE-China.2018.8448772.

[16] W. Sirichotedumrong, Y. Kinoshita, and H. Kiya, "Pixel-based image encryption without key management for privacy-preserving deep neural networks," *IEEE Access*, vol. 7, pp. 177 844–177 855, 2019. DOI: 10.1109/ACCESS.2019.2959017.

[17] A. H. Chang and B. M. Case, *Attacks on image encryption schemes for privacy-preserving deep neural networks*, 2020. arXiv: 2004.13263 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2004.13263.

[18] H. Ito, Y. Kinoshita, M. Aprilpyone, and H. Kiya, "Image to perturbation: An image transformation network for generating visually protected images for privacy-preserving deep neural networks," *IEEE Access*, vol. 9, pp. 64 629–64 638, 2021. DOI: 10.1109/ACCESS.2021.3074968.

[19] K. Madono, M. Tanaka, M. Onishi, and T. Ogawa, *Block-wise scrambled image recognition using adaptation network*, 2020. arXiv: 2001.07761 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2001.07761.

[20] H. Kiya, T. Nagamori, S. Imaizumi, and S. Shiota, "Privacy-preserving semantic segmentation using vision transformer," *Journal of Imaging*, vol. 8, no. 9, 2022, ISSN: 2313-433X. DOI: 10.3390/jimaging8090233. [Online]. Available: https://www.mdpi.com/2313-433X/8/9/233.

[21] T. Chuman and H. Kiya, "A jigsaw puzzle solver-based attack on image encryption using vision transformer for privacy-preserving dnns," *Information*, vol. 14, no. 6, 2023, ISSN: 2078-2489. DOI: 10.3390/info14060311. [Online]. Available: https://www.mdpi.com/2078-2489/14/6/311.

Significance of Lower Frequency Regions for Audio Deepfake Detection

Arth J. Shah, and Hemant A. Patil
Speech Research Lab, DA-IICT, Gandhinagar, India
E-mail: {202101154, hemant_patil}@daiict.ac.in

***Abstract*—Deepfake audios are created using deep learning methods. Audio security, due to deepfake audio creation, has been a severe issue in recent days. Many techniques have been explored to detect attacks by fake audio generators. Similar to Spoofed Speech Detection (SSD), audio deepfake systems also rely on characteristics of live speech to detect whether the audio is deepfake or real, however, they aim to fool humans instead of voice biometrics system. As the attackers are free to mount various types of audio attacks, Audio Deepfake Detection (ADD) plays a vital role in defending against fake AI/deep learning-generated audio attacks. For this study, we explored various lower frequency-based acoustic features, such as Generalized Morse Wavelet (GMW)-based features in combination with Mel spectrogram-based features, using Convolutional Neural Network (CNN)-based classifier for ADD task. For performance comparison, we employed the traditional spectrogram. Further, feature-level data fusion (instead of score-level) of proposed GMW-based with Mel spectrogram feature gave 1.93% improvement in overall accuracy.**

***Index Terms* – Audio Deepfake, Morse Wavelet, Lower Frequency Resolution, Mel Spectrogram, Speech Intelligibility.**

I. Introduction

Deepfake is a content, where the original recording of audio, video, and photo has been replaced by computer-generated fake signal using deep learning algorithm. Deepfake was first enlightened by a user named "Deepfakes" on Reddit in 2017 [1]. Due to advances in technology, the activities of criminals have increased several folds w.r.t. deepfake attacks. With the increase in the availability of facilities, deepfakes have been created more critically, which makes them harder to detect. In recent years, numerous such attacks based on deepfake have been made [2], [3], which drew the attention of researchers. For such an advanced generated deepfake, we need a strong detection system that can classify audio more accurately. The existing baseline approaches used acoustic features for the Audio Deepfake Detection (ADD) task, whose success motivated us to explore other acoustic features for the ADD task. Moving through the characteristics of audio, many acoustic features have been successfully used for various speech security-based applications. Many times, speech signals are analyzed using spectrogram of speech signals. We investigate significance of low frequency-based acoustic features, in particular, Generalized Morse Wavelet (GMW)-based features for ADD. To that effect, we estimated Continuous Wavelet Transform (CWT), having GMW as mother wavelet function [4]. We used Morse Wavelet to form a scalogram before testing them on the CNN classifier. Motivated by encouraging results by Mel spectrogram features, we explored other lower frequency-based features, in particular, GMW-based features for ADD task. Relative significance of lower frequency region than full band (w.r.t. Shannon sampling paradigm) is also illustrated in this study.

Both basic types of data fusion strategies, namely, feature-level fusion and score-level fusion were examined in order to investigate possible *complementary* information captured by individual systems of both the fusions. After release of the Fake or Real (FoR) dataset, much progress has been made on machine learning-based approaches, and also a few deep learning-based approaches have been explored. In [5], the authors explore machine learning-based approach, such as Long Short-Term Memory (LSTM), and VGG-16 for ADD task on FoR dataset. In [6], authors proposed the use of baseline Mel Frequency Cepstral Coefficients (MFCC), with a variety of models, such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost) for ADD task. However, the feature-based approach still remains unexplored. In this study, authors propose GMW-based features fused with Mel spectrogram for ADD task. In part, this overcomes the practical problem of choosing the appropriate mother wavelet function because generalized Morse Wavelet function is a super-family of different analytic wavelets [7]. Also unlike classic analytical wavelet (which is approximately analytic and thus, resulting in spectral leakage at lower frequency), such as Morlet, Morse Wavelet shows exact systematic behaviour (i.e., no spectral leakage).

The Morse Wavelet-based scalogram and Mel-spectrogram both are known to have a good resolution in the lower frequency regions. However, the scalogram as shown in Fig. 1 (a) has much higher frequency resolution for frequencies below 0.1 kHz (i.e., 100 Hz) and it can be observed from the scalogram that there is hardly any significant energy in that region. So the benefit of using Morse scalogram because of its excellent frequency resolution property in that region is not well utilized for ADD task. However, for the Mel spectrogram, even though the frequency bins are spaced in a Mel scale, the Mel spectrogram is able to emphasize on frequency regions beyond 100 Hz as well. In Fig. 1 (b), we can note that the harmonic structure for spectrogram is less clear than Mel spectrogram and Morse Wavelet spectrograms. Similar observations can be made in Fig. 1 (c) (Panel I), has

abundantly blurred pitch source harmonics (i.e., kF_0, $k \in Z$) structure, as compared to Fig. 1 (c) (Panel II), which has sharp harmonic structure, resulting into clear classification while using Mel spectrogram for ADD task. To the best of authors' knowledge, this is the first study that employes and explores wavelet-based features for ADD task.

The remaining paper is organized as follows: Section II provides the details about lower frequency-based features. Section III presents information about experimental setup used for this study. Section IV presents experimental results and the related discussion. Section V provides summary, conclusion and the future potential research directions for this study.

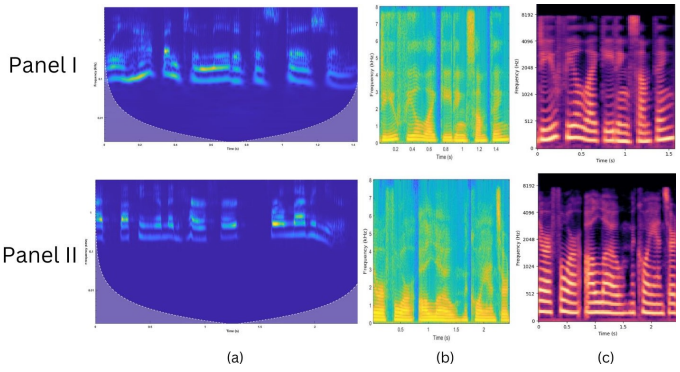


Fig. 1: Panel I (fake signal), and Panel II (real signal): (a) Morse Wavelet spectrogram, (b) spectrogram, and (c) Mel spectrogram, respectively.

II. Lower Frequency-Based Approach For Feature Extraction

A. Mel Spectrogram

Due to mathematical structure of Mel (melody) scale, Mel spectrogram of a speech is known to emphasize lower frequency region more as compared to higher frequency region [8]. The spectrogram of an speech signal ($m(t_0)$) can be defined as magnitude square of inner product. In particular, i.e.,

$$|< \mathrm{m}(t_0), g_{u,\zeta}(t_0) >|^2 = \left| \int_{-\infty}^{\infty} m(t_0) g(t_0 - u) e^{-j\zeta t_0} dt_0 \right|^2, \tag{1}$$

where $g_{u,\zeta}(t_0)$ is called time-frequency atoms. We extracted Mel spectrograms by 30 *ms* analysis window length, 20 *ms* window overlap, 32 number of sub-band filters, and taking normalized window. For such scalogram-based features, we were motivated to use Morse Wavelet features following its success to various audio-based tasks [9].

B. Continuous Wavelet Transform (CWT)

After the success of MFCC and Mel spectrogram features [5], [6], it is important to investigate the significance of low frequency regions for the ADD. The Mel frequency spectrogram and MFCC features follow the Mel scale, which follows the perceptual scale of hearing by humans. The Mel scale given by :

$$Mel_Scale = 1127 \ln \left(1 + \frac{f}{700}\right), \tag{2}$$

where f is in Hz and hence, has higher frequency resolution in the lower frequency regions, in comparison high frequency regions. To further investigate the significance of the low frequency regions for the ADD, for this study, we propose the use of CWT-based time-frequency representation, popular as the *scalogram*. CWT-based representations are well known for *Multi-Resolution Analysis (MRA)*. In order to achieve this, we used a CWT-based method that finally produced high frequency resolution in low frequency limits. For a signal $m(t)$, CWT is given by :

$$\begin{aligned} W_m(a,b) &=< m(t_0), \psi_{a_0,b_0}(t_0) >, \\ &= \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} m(t_0) \psi^* \left(\frac{t_0 - b_0}{a_0}\right) dt_0, \end{aligned} \tag{3}$$

where $< \cdot, \cdot >$ is inner product operation, which is used to calculate wavelet coefficients, and * represents complex conjugate operation. In Eq. (3), parameter a_0 is known as the scale parameter, whereas parameter b_0 is known as translation parameter. A scalogram is a graphic representation of energy of CWT coefficients (i.e., $| < m(t_0), \psi_{a_0,b_0}(t_0) > |^2$). In this study, we computed CWT coefficients by taking into account the Morse Wavelet. The word wavelet means a short duration wave. Alternatively, not every short wave is called a wavelet. A wavelet $\psi(t_0) \in L^2(R)$ (i.e., Hilbert space for signals with finite energy and Lebesgue integral functions) weakly satisfies the condition, represented by :

$$C_\psi = \int_0^{\infty} \frac{|\Psi(\omega_0)|^2}{\omega_0} d\omega_0 < \infty, \tag{4}$$

In Eq. (4), $\Psi(\omega_0)$ represents Fourier transform of mother wavelet function, $\psi(t_0)$. This proves $\Psi(\omega_0)$ drops its value to zero as fast as $\omega_0 \to 0$. Furthermore, to validate Eq. (4), we are required to impose $\Psi(0) = 0$, which is equivalent to :

$$\Psi(\omega_0)|_{\omega_0=0} \equiv \int_{-\infty}^{\infty} \psi(t_0) dt_0 = 0. \tag{5}$$

Further, for task of energy normalization, we consider wavelets satisfying $\|\psi(t_0)\| = 1$, i.e., $\int_{-\infty}^{\infty} |\psi(t_0)|^2 dt_0 = 1$.

C. Proposed GMW-Based Features

Systematic wavelets exist in several types, such as complex Shannon, log-normal, Gaussian derivative, Cauchy, and Morlet wavelets, and are mostly noted in the frequency-domain [10]–[13]. GMW represent the family of systematic wavelets, which addresses the problem of choosing a specific type of analytic wavelet [7], [14], [15]. GMWs represent a set of eigenvectors w.r.t. a time-frequency localization operator, whose details can be studied from [16]. From the set of eigenvectors, the largest eigenvector is considered, which makes the 0th GMW. The largest eigenvector is known to enable *optimal energy concentration* properties [16], [17]. To that effect, we represent

the 0th GMW as the Morse Wavelet. It is an analytic wavelet and is expressed mathematically as [7], [14], [17]:

$$\Psi_{\beta,\gamma}(\omega_0) = \int_{-\infty}^{+\infty} \psi_{\beta,\gamma}(t_0) e^{-j\omega_0 t_0} dt_0, \tag{6}$$

where the parameters β and γ let degree of freedom additionally and make GMWs form a *family* of analytic wavelets [15]. This degree of freedom is known to establish the wide range of behaviour of the GMWs depending on the parameters β and γ. To that effect, Eq. (6) can be said to be a general approach to construct an *exactly* analytic wavelet [15]. Furthermore, Eq. (6) ensures that the wavelet satisfies the admissibility condition, which require the wavelet to have finite energy and zero-mean, i.e.,

$$\int_{-\infty}^{+\infty} |\psi_{\beta,\gamma}(t_0)|^2 dt_0 = 1. \tag{7}$$

For analysis purposes, the dimensionless derivatives of the Morse Wavelet are considered [15]. In particular, its 2nd order derivative is denoted by $P^2_{\beta,\gamma}$, and is called the *wavelet duration* or *inverse bandwidth*. The significance of the 2nd-order derivative is discussed in detail in the Appendix. The quantity $P^2_{\beta,\gamma}/2\pi$ sets the number of oscillations in the wavelet, where $P^2_{\beta,\gamma}$ is given as [15]:

$$P^2_{\beta,\gamma} \equiv -\frac{\omega_0^2 \beta,\gamma \Psi''_{\beta,\gamma}(\omega_{0\beta,\gamma})}{\Psi_{\beta,\gamma}(\omega_{0\beta,\gamma})} = \beta\gamma. \tag{8}$$

Furthermore, when $P^2_{\beta,\gamma}$ is fixed, and γ is varied, the wavelet behaviour corresponds to different families of wavelets. Hence, the parameter γ is also termed as the *family*. Further, the parameter γ is also known to be related to area of Heisenberg's box in the time-frequency plane [7]. In particular, $\gamma = 3$ transforms to the *'Airy family'*, where the Morse Wavelet behaves similarly to the conventional Morlet wavelet. Interestingly, the conventional Morlet wavelet has significance in the human perception process for both hearing and vision [18]. Human hearing mechanism is historically known to be basis for the developing various state-of-the-art features in speech-related pattern recognition problems. In particular, the well known MFCC features are derived by Mel frequency wrapping, which is truely motivated by the Weber's law of human perception, also developed using the perception mechanism of human hearing [19]. Likewise, the Mel scale is also known as the perceptual scale of hearing. This motivated us to explore the perceptual scales and hence, the low frequency regions for the ADD. In this paper, we consider the Morse Wavelet with $\gamma = 3$, instead of the conventional Morlet wavelet, because the coventional Morlet wavelet can be observed to defer from the analyticity under special cases, however, the Morlet-like wavelet obtained by taking $\gamma = 3$ in Morse Wavelet, shows *strictly* analytic behaviour, as shown in the Fig. 2. In particular, the Morlet wavelet, as observed in Fig. 2 (b), shows a *spectral leakage* in the negative frequencies as shown by its Wigner Ville distribution. On the contrary, as observed in Fig. 2 (d), upon analysis of the Morse Wavelet at $\gamma = 3$, it can be observed that there is no spectral leakage in the negative

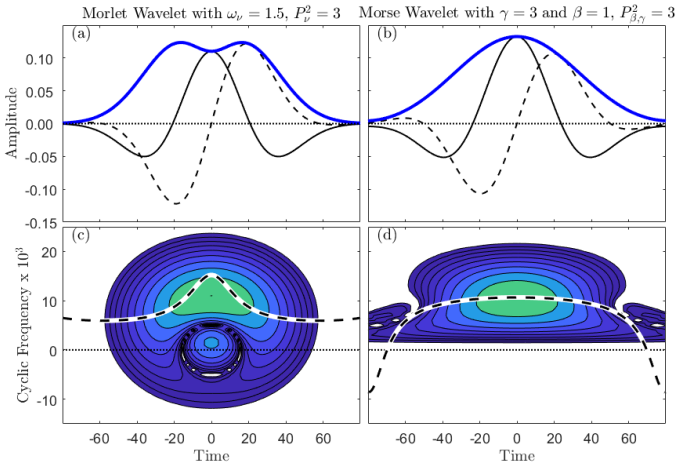


Fig. 2: Analysis of spectral leakage (a) Morlet wavelet, *vs.* (b) Morse Wavelet, and (c), (d) their respective Wigner-Ville distributions. After [15], [19].

frequency regions of the GMW. This indicates that strict analytic properties are exhibited by the Morse Wavelet, which is a characteristic of wavelets that satisfy the Cauchy-Riemann equations of differentiablity. This means that the wavelet function $\Psi(\omega)$ equals zero for $\omega < 0$. The Morlet wavelet, on the other hand, does not exhibit strict analytic properties, as its wavelet function $\Psi(\omega)$ has non-zero values for $\omega < 0$. This makes the Morse Wavelet a preferred choice in applications, where the presence of negative frequency components can lead to inaccuracies in the analysis, and in particular, the ADD task discussed in this paper. We believe that this remarkable exact analytic properties of Morse Wavelets (i.e., $\psi(\omega) = 0$ for $\omega < 0$) may help to efficiently capture class-specific attributes (e.g. real *vs.* deepfake) of given signal that are present near lower frequency regions, whereas for approximately analytic wavelets (such as Morlets), the energy in their lower frequency region may experience blur primarily due to spectral leakage of energy in the negative frequencies [19]. Fig. 3 represents the flowchart of process carried out while classification of fake and real audios.

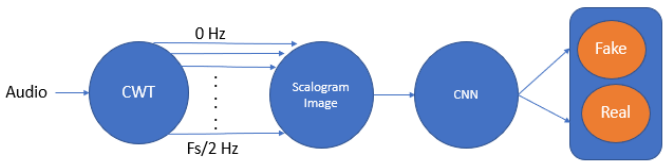


Fig. 3: Functional block diagram of proposed ADD system.

III. Experimental Setup

A. Dataset Used

1) Fake or Real (FoR) Dataset: It is one of the most well known, statistically meaningful, and freely available dataset for ADD tasks[20]. It contains around 87,000 fake utterances as well as 111,000 real utterances, which were generated by more than 1200 speakers. Further, FoR has more number of training files as compared to ASVSpoof 2019 dataset [21] and hence,

FoR is used in this paper. The large amount of data makes the model train properly, thereby leading to proper evaluation of model. The main reason to use the FoR dataset instead of others is that it includes speech produced by state-of-the-art voice synthesis algorithms. A high number of files in the dataset can also classify points into different classes without overfitting. FoR dataset have been released in 4 versions. FoR-original is the version, which has original collected utterances. FoR-norm is the version of dataset that has samples which were collected after performing certain set of prepossessing on the original audio. FoR-2second contains the samples truncated to 2 second and at last for-rerecorded consist of all utterance which were recorded again from the original audio with gadgets like speaker and microphone. Released in the early half of 2021, it makes it one of the most recently released datasets. The authors in [22], used ASVspoof dataset which consists of total 125,000 samples, compared to which FoR consist of 195,541 utterances of total. We firmly believe that, more the number of data samples used for training dataset, more accurately the model gets trained. As a result to which authors chooses FoR dataset over ASVspoof dataset. Because the *FoR-norm* version of the FoR dataset is normalized and preprocessed, the authors have been using it for all experiments in this manuscript's study. This is because we feel it to be the best option for more accurate and thorough model training. The gender ratio in the FoR-norm version is clearly more balanced than in the FoR-original dataset, which is another incentive to utilize it. It has been divided in training (77.73%), testing (6.68%), and validation (15.58%) sets, to make up the dataset. The sampling rate of speech data is 16 kHz.

B. Pattern Classifier Used

The CNN classifier used in this work consists of a Fully-Connected (FC) layer after two consecutive 2D-convolutional layers. The CNN is provided with an input feature dimension of 512 × 512. Using a 7 x 7 kernel, the first Convolutional Neural Layer (CNL) contains 3 input channels ($I_{channels}$) and 8 output channels ($O_{channels}$). The second CNL has a 3 × 3 kernel, 8 $I_{channels}$, and 16 $O_{channels}$. The last CNL has a 3 x 3 kernel, 16 $I_{channels}$, and 32 $O_{channels}$. Every CNL has padding and stride values of 1. A max-pool layer with a 3 × 3 kernel and a stride of 3 is used to reduce dimensions. The activation function is the Rectified Linear Unit (ReLU). The choice for optimizer was Adam. The Binary cross-entropy Loss Function (BceLF) is chosen as the loss metric. This architecture and configuration are designed for tasks involving input images of size 512 × 512, demonstrating a multi-layered approach for hierarchical feature extraction and classification. Learning rate was defined as 0.001.

C. Performance Metrics

1) Precision: It is implemented to measure how much accurately the model works on positive samples, which also predicts how many positive samples are correctly classified.

2) Recall: Alternet to Precision recall works somewhat similar to precision. The main difference is, Recall measures the percentage of actual positive cases that were accurately anticipated to be positive.

3) F1-Score: It balances Precision and Recall scores. It can simply be defined as harmonic mean of precision and recall.

IV. Experimental Results

This Section examines and proposes ML / DL-based experiments performed for measuring the accurate working of proposed feature vector in combination with pattern classifier. Table I presents the comparison of results w.r.t. several features used on CNN classifier. The Mel spectrogram, which capture information about how frequency region gave us an significant result of 95.07 % accuracy, which is 17.24 % more than spectrogram. This motivated us to explore other features having good resolution in lower frequency region. The proposed GMW features gave 79.56 % accuracy, which is 1.73 % higher than spectrogram. Since recall is higher in Morse Wavelet and spectrogram, it indicates that false negative samples are much less in comparison with true negative. Since Mel spectrogram gave high precision, the number of false positive samples are less compared to the Morse Wavelet and spectrogram. Low precision in Morse Wavelet and spectrogram indicates that these two features are not able to predict real speech properly, whereas Mel spectrogram is not able to predict fake speech precisely as compared with real speech.

TABLE I: Results *w.r.t.* Different Explored Features and CNN Classifier

Features	Accuracy (in %)	Precision	Recall	F1-Score
Spectrogram	77.83	68.85	99.77	76.94
Morse Wavelet	**79.56**	74.52	88.38	79.46
Mel Spectrogram	**95.07**	97.52	92.27	95.06

A. Results using Feature-Level Fusion

We first extracted two feature sets and fused them by concatenation like a stack, which resulted in a new feature vector of dimension 1024 x 512 pixels per frame. Table II represents results *w.r.t.* feature-level fusion. By performing feature-level fusion of Mel Spectrogram and Morse Wavelet, we achieved an accuracy of 97.00 %, which is 1.93 % higher than Mel spectrogram alone. By performing feature-level fusion, we are also able to get Equal Error Rate (EER) as low as 2.99 %. On the other hand, the fusion of Mel spectrogram and Spectrogram resulted in an accuracy of 90.22 %.

TABLE II: Feature-Level Fusion on CNN Classifier

Feature 1	Feature 2	Accuracy (in %)	EER
Morse Wavelet	Spectrogram	79.08	20.91
Mel Spectrogram	Spectrogram	90.22	9.75
Morse Wavelet	Mel Spectrogram	**97.00**	**2.99**

B. Results using Score-Level Fusion

After increase in ≈ 3 % accuracy by performing feature-level fusion, the authors explored the potential of score-level fusion for two features. In this type of fusion, we fused the scores gained by each model while testing, which were taken in

ratio of α and $(1-\alpha)$, where $\alpha \in [0,1]$, i.e., $\alpha S_{Morse} + (1+\alpha)S_{spectrogram}$, where S_{Morse}, and $S_{spectrogram}$ are scores for wavelet, and spectrogram, repressively. Fig. 4 shows that Mel Spectrogram and spectrogram when fused at the score-level gives a bit less accuracy than the accuracy gained by fusing Morse Wavelet and spectrogram. However, score-fusion of Morse Wavelet and spectrogram keeps on performing poorly and thus, employing different type of fusion further for ADD. This findings is in agreement with other studies in pattern recognition literature, where various data fusion strategies, such as, signal-level, feature-level, classifier-level, and score level may give different inferences w.r.t.complementary information captured by various individual pattern recognition systems. In particular, in our study, we found that feature-level fusion captures better complementary information then score-level fusion.

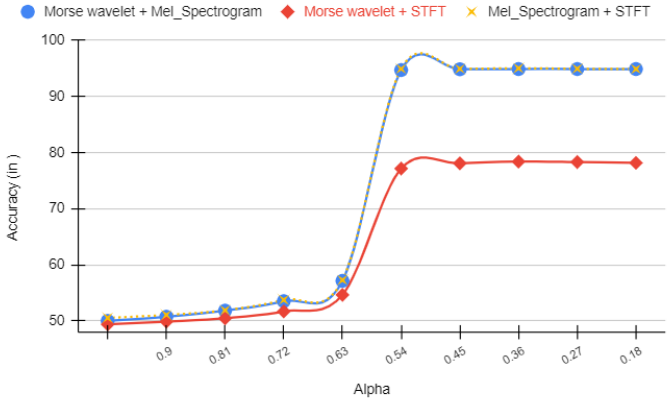


Fig. 4: Score-Level Fusion of Features.

C. Discussion

Mel spectrogram as well as GMW features emphasize lower frequency region, however, both are not giving the same results, which is due to the difference in relative joint time-frequency resolution captured by spectrogram *vs.* CWT as per Heisenberg's uncertainty principle in signal processing framework [23]. The amount of frequency resolution of CWT is comparatively very high than Mel spectrogram. In order to analyze their relative feature discrimination power, we perform feature-level fusion of both the features [24]. Even though Morse Wavelet is not performing good, giving us an accuracy of 79.56 %, when fused with the Mel spectrogram, which is initially giving accuracy of 95.07 %, the accuracy jump to 97 %, indicating that the Mel spectrogram is also able to capture characteristics of a few higher frequency regions, which on the other hand, Morse Wavelet is unable to capture alone. In order to detect deepfake attack, we need frequency region from all possible sub-bands, i.e., from 0 to $Fs/2$, however, the emphasis should be laid on the lower frequency region as captured by the Mel spectrogram. On the other hand, spectrogram (which employ Windowed Fourier Transform (WFT) or Short-Time Fourier Trasnform (STFT)) has *constant* resolution in each frequency band (in the entire time-frequency panel), i.e., it captures information from all frequency range. Spectrogram not emphasizing to lower frequency region particularly, resulting in low accuracy performance than other lower frequency resolution-based features. Fusion of other features with spectrogram also leads to lower accuracy than individual lower frequency resolution-based feature sets.

V. Summary and Conclusions

In this study, we employed features based on low frequency regions, namely, GMW and Mel spectrogram for ADD task. In particular, we employed Morse Wavelet and Mel spectrogram with CNN as pattern classifier, which gave relatively better results. Further, feature-level fusion, we are able to obtain even more higher accuracy than the individual features. However, the proposed approach requires high time and computational storage. Furthermore, many other wavelet-based approaches based on lower frequency resolution can be explored for ADD task. Future plans includes experiments to explore more low frequency resolution-based features, as well as other pattern classifiers for ADD task. Historically, the discrete time implementation of CWT involves discrete wavelet bases $\psi_{j,k}(t_0) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t_0 - 2^j \cdot k}{2^j}\right)$ that are related to multi-resolution analysis (MRA) and its implementation via Quadrature Mirror Filterbank (QMF) involving cascade of lowpass and highpass filters relating several properties of wavelets, such as vanishing moments, smoothness, regularity, and support size [25]. Developing this theoretical as well as practical viewpoint for GMW remains an open research problem. Our findings indicate that lower frequency content in deepfake is discriminating in nature for ADD, which may be attributed to Instantaneous Frequencies (IFs) in lower frequency region. Thus, exploring Morse Wavelets for IF estimation in the classic Delprat's IF estimation using ridge analysis remains another challenging open research problem.

Appendix

The kth order moment of the demodulated Morse Wavelet by the peak frequency ω_ψ can be written as [19] :

$$m_{k;\psi} = \int_{-\infty}^{+\infty} t^k \psi(t) e^{-i\omega_\psi t} dt. \tag{9}$$

As per the relationship between moments in time-domain and derivatives in frequency-domain,

$$\frac{m_{k;\psi}}{m_{0;\psi}} = i^k \frac{\psi^{(k)}(\omega_\psi)}{\tilde{\Psi}(\omega_\psi)}, \tag{10}$$

where $\tilde{\Psi}(\omega_\psi)$ is the wavelet's *dimensionless derivatives* at $\omega_0 = \omega_\psi$, and defined as [15], [17]:

$$\tilde{\Psi}_k(\omega_0) \equiv \omega_0^k \frac{\Psi^{(k)}(\omega_0)}{\Psi(\omega_0)}, \tag{11}$$

where k as superscript denotes the kth order derivative, and $\tilde{\Psi}(\omega_\psi)$ is nothing but $\tilde{\Psi}_k(\omega_0)$ evaluated at $\omega_0 = \omega_\psi$. Therefore,

$$\frac{m_{k;\psi}}{m_{0;\psi}} = i^k \frac{\psi^{(k)}(\omega_\psi)}{\tilde{\Psi}(\omega_\psi)} = i^k \frac{\tilde{\Psi} k(\omega_\psi)}{\omega_0^k{}_\psi}. \tag{12}$$

Since the wavelet satisfies the weak admissibility condition, its mean is zero ,i.e., $m_{1;\psi} = 0$. Therefore, the higher-order demodulate moments can be normalized by the 2nd moment and hence, the dimensionless measure using the 2nd moment is defined as

$$P_{\beta,\gamma} = \omega_\psi \sqrt{\frac{m_{2;\psi}}{m_{0;\psi}}} = \sqrt{|\tilde{\Psi}_2(\omega_\psi)|}, \tag{13}$$

where $\tilde{\Psi}_2(\omega_\psi) = -\beta\gamma$ (as represented in [15]).

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Priyanka Gupta, LNMIIT, Jaipur for her help during this research work. Further, they acknowledge partial support from Ministry of Electronics and Information Technology (MeitY), Grant ID: 11(1)2022-HCC.(TDIL).

REFERENCES

[1] S. Maddocks, "'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political'deep fakes," *Porn Studies*, vol. 7, no. 4, pp. 415–423, 2020.

[2] A. R. Javed, Z. Jalil, W. Zehra, T. R. Gadekallu, D. Y. Suh, and M. J. Piran, "A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions," *Engineering Applications of Artificial Intelligence*, vol. 106, pp. 104–456, 2021.

[3] A. Ahmed, A. R. Javed, Z. Jalil, G. Srivastava, and T. R. Gadekallu, "Privacy of web browsers: A challenge in digital forensics," in *Genetic and Evolutionary Computing: Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computing, October 21-23, 2021, Jilin, China 14*, Springer, 2022, pp. 493–504.

[4] P. Gupta, S. Gupta, and H. Patil, "Voice liveness detection using bump wavelet with CNN," in *9th International Conference on Pattern Recognition and Machine Intelligence*, 2021.

[5] A. Hamza, A. R. R. Javed, F. Iqbal, *et al.*, "Deepfake audio detection via MFCC features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022.

[6] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian Journal for Science and Engineering*, pp. 1–12, 2021.

[7] J. M. Lilly and S. C. Olhede, "Generalized morse wavelets as a superfamily of analytic wavelets," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 6036–6041, 2012.

[8] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, Vol 3, 2022.

[9] E. A. Martinez-Ríos, R. Bustamante-Bello, S. Navarro-Tuch, and H. Perez-Meana, "Applications of the generalized morse wavelets: A review," *IEEE Access*, Vol 3, 2022.

[10] M. Holschneider, "Wavelets : An Analysis Tool," *Oxford Science Publications*, 1995.

[11] S. G. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 2nd Edition, 1999.

[12] H. Knutsson, C.-F. Westin, and G. Granlund, "Local multiscale frequency and bandwidth estimation," in *Proceedings of* 1st *International Conference on Image Processing (ICIP)*, vol. 1, 13-16 Nov. 1994, pp. 36–40.

[13] X. Zhu and J. Kim, "Application of analytic wavelet transform to analysis of highly impulsive noises," *Journal of Sound and Vibration*, vol. 294, no. 4-5, pp. 841–855, 2006.

[14] S. C. Olhede and A. T. Walden, "Generalized Morse Wavelets," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2661–2670, 2002.

[15] Lilly, Jonathan M. and Olhede, Sofia C., "Higher-order properties of analytic wavelets," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 146–160, 2008.

[16] I. Daubechies and T. Paul, "Time-frequency localization operators-a geometric phase space approach: II. The use of dilations," *Inverse Problems*, vol. 4, no. 3, p. 661, 1988.

[17] J. M. Lilly and S. C. Olhede, "On the analytic wavelet transform," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 4135–4156, 2010.

[18] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[19] P. Gupta and H. A. Patil, "Morse wavelet transform-based features for voice liveness detection," *Computer Speech & Language*, vol. 84, p. 101 571, 2024.

[20] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, IEEE, 2019, pp. 1–10.

[21] M. Todisco, X. Wang, V. Vestman, *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[22] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved deepfake detection using whisper features," *arXiv preprint arXiv:2306.01428*, 2023.

[23] P. Busch, T. Heinonen, and P. Lahti, "Heisenberg's uncertainty principle," *Physics Reports*, vol. 452, no. 6, pp. 155–176, 2007.

[24] R. A. Rasool, "Feature-level vs. score-level fusion in the human identification system," *Applied Computational Intelligence and Soft Computing*, vol. 2021, pp. 1–10, 2021.

[25] S. G. Mallat, "Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$," *Transactions of the American Mathematical Society*, vol. 315, no. 1, pp. 69–87, 1989.

Fine-Grained Privacy-Preserving Image Retrieval in Cloud Environment

Jing Liang*, Libo Wang† and Peiya Li‡
* College of Cyber Security, Jinan University, Guangzhou, China
E-mail: liangjing99@stu.jnu.edu.cn
† School of Cyberspace Security, Hainan University, Haikou, China
E-mail: wanglibo@hainanu.edu.cn
‡ College of Cyber Security, Jinan University, Guangzhou, China
E-mail: lpy0303@jnu.edu.cn

***Abstract*—With the advancement of cloud computing, an increasing number of resource-constrained data owners prefer to store their images in the cloud. Considering security and privacy, images should be encrypted before upload. Current privacy-preserving image retrieval techniques still face challenges such as high access control cost, insufficient granularity and low retrieval efficiency. Therefore, in this article we design an efficient fine-grained privacy-preserving image retrieval scheme based on polynomial access policies. Our scheme first employs transfer learning techniques to extract image features using pre-trained deep learning models to enhance retrieval accuracy. Then an encrypted hierarchical K-means structure is established to cluster these features that improves retrieval efficiency without compromising data privacy. For fine-grained access control, a role-based polynomial access control technique is designed to enable precise permission management. Our scheme combines access control policies with index trees to greatly improve access speed in ciphertext domains. The practicality of the proposed scheme is demonstrated through experimental evaluations of its security and performance.**

I. Introduction

With the burgeoning advancements in mobile cloud technology, data proprietors with constrained resources predominantly upload images to cloud servers for storage [1]. The search for a specific image from an extensive dataset has attracted growing attention in numerous practical fields, such as disease detection and diagnosis [2] and online shopping [3], etc, wherein similar images can be conveniently retrieved utilizing cloud servers. Nevertheless, underlying security concerns continue to impede the universal adoption of image outsourcing to cloud servers. Direct uploads of images to cloud servers can readily compromise personal information, including facial attributes and identity details [4]. The previously mentioned requirements are conveniently complied with Searchable Encryption (SE) technology [5]. Unlike traditional encryption schemes, which cannot guarantee the availability of encrypted images or cannot be used on devices with limited resources, SE can protect image data privacy while allowing operations like image retrieval on cloud servers [6].

However, the application of most privacy-preserving image retrieval schemes in daily-use scenarios has some problems that cannot be ignored, including suboptimal accuracy, inefficient retrieval, and a lack of fine-grained access control. First, low retrieval precision means that the retrieval scheme cannot effectively retrieve the information that the user really wants, which is unacceptable to the user in practical engineering applications [7]. Taking medical images with rich details as an example, high-precision retrieval can help doctors make more accurate diagnoses in the face of complex and changing lesions. Secondly, the needs of efficient image retrieval services are reinforced by the increase of large image databases among all kinds of areas [1]. Designing an efficient image retrieval scheme is crucial as it reduces the service response time. Finally, coarse-grained access control ensures that outsourced images are only accessible to users possessing the corresponding search key, allowing authorized users to seamlessly access the entire outsourced image repository [8], [9]. Given the numerous security risks associated with unauthorized access, it is imperative to implement fine-grained access control techniques to regulate image access for each user.

Currently, there is plenty of research being done on privacy-preserving image retrieval techniques. Lu et al. [10] proposed the first encrypted multimedia database retrieval scheme, but the visual word-based feature extraction method leads to low retrieval accuracy. Based on Color Layout Descriptor (CLD) and Edge Histogram Descriptor (EHD), Xia et al. [1] proposed a privacy-preserving content-based image retrieval scheme, but obtained low retrieval accuracy as these two features can only roughly describe the image content. To improve retrieval accuracy, [11] and [12] employed a pre-trained Convolutional Neural Network (CNN) model to extract feature vectors, which further improves the retrieval accuracy as CNN model imitates the structure of the visual cortex and captures complex and abstract image information hierarchically. In terms of retrieval performance, since scheme [13] that used Homomorphic Encryption (HE) technology to encrypt feature vectors leads to low retrieval efficiency, Li et al. [14] designed a sub-simhash index structure to imporve efficiency. Besides, Qin et al. [15] presented an efficient and privacy-preserving scheme with high retrieval efficiency by using Locality Sensitivity Hashing (LSH). In [16], hierarchical index trees based on

‡Corresponding author

K-means clustering were constructed, which outperform the linear index [17] in terms of retrieval efficiency. Most of the above privacy-preserving image retrieval schemes provide coarse-grained access control to encrypted images, which still poses a threat to the security of sensitive images. Khan et al. [18] designed an access control based on attribute-level encryption, but complex access policies and a large number of user attributes lack flexibility and dynamism. Tong et al. [9] used polynomial-based access policy to realize access control, but linear retrieval reduces the access speed.

To address the aforementioned issues, we propose an efficient fine-grained privacy-preserving image retrieval scheme. We employ transfer learning techniques to extract image features using pre-trained deep learning models, then design an index tree of the feature vectors using K-means clustering algorithm. Moreover, the access structure of each image is specified by constructing role polynomials. If the user is unable to access a particular image, the corresponding computation is much larger than a given threshold. After that, images and indexes are encrypted and uploaded to the cloud server. We combine the index tree with an access control policy to speed up access while realizing fine-grained access control. In summary, our contributions are as follows:

1) High retrieval accuracy. We use transfer learning and pre-trained deep learning models to extract image features, which represent image content more accurately than manual methods. This approach achieves higher retrieval accuracy compared to traditional schemes using local and global features.

2) Fine-grained access control. We employ a polynomial-based access policy for fine-grained control, instead of complex attribution policies. Each image's access rights are defined by a role-based polynomial. If a user can't access an image, the calculation result will exceed the set threshold.

3) Fast access speed. Instead of linear retrieval, we use hierarchical K-means algorithm to categorize all image features and construct an index tree, which assigns similar image features in the same branching structure. It is worth noticing that the structure achieves efficient retrieval while realizing fine-grained access control.

The rest of the paper is organized as follows. In Section 2, our proposed scheme and its detailed procedure, threat models are presented. Section 3 analyzes security and experimental results. We conclude this paper in Section 4.

II. Proposed Scheme

A. System Model

Our scheme contains three entities, namely image owner, cloud server and image users, which are shown in Fig. 1. The image owner is in charge of the keys generation and distribution. Besides, the image owner needs to construct an index, encrypt the image and index before uploading them to the cloud server. The image users undertake to generate valid query trapdoor by using the keys distributed by image owner. After receiving the retrieval results from the cloud server, the image users can decrypt them by utilizing his/her secret key.

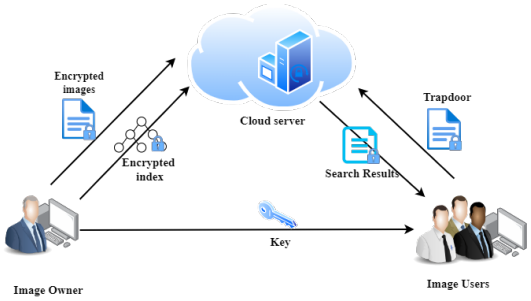


Fig. 1. System model of our scheme.

The cloud server with unlimited resources is responsible for providing the storage and computing services for the image owner and retrieval service for image users.

B. Threat Model

In our system, the cloud server is considered to be honest but curious. That is, the cloud server will follow our protocol to perform the service provider role correctly, but it may tamper or forge sensitive information from encrypted images, index, or trapdoor. The image owner and user are always trusted. Based on available information to the cloud server, we consider the following threat models.

Known Ciphertext Attack Model. The cloud server only has access to all encrypted images and corresponding feature vectors, encrypted index tree and trapdoor. In particular, we aim at preventing the cloud server from learning valuable information related to image content, encrypted index tree, trapdoor and encrypted search results.

Known Background Attack Model. In this stronger threat model, the cloud server possesses all information consistent with the *Known Ciphertext Attack Model*. Moreover, the cloud server will obtain additional information, such as some plaintexts, which would be used to infer more plaintexts from inner products calculating results.

C. Detailed Constructions

Our scheme construction consists 5 algorithms: **Key_Gen**, **Index_Buld**, **Data_Enc**, **Query_Gen** and **Image_Retri**. In the **Key_Gen** algorithm, the image owner first generates keys to set up the system. Then, by using the **Index_Buld** algorithm, the image owner builds an efficient search index tree for image search. For privacy protection purpose, the image owner utilizes the **Data_Enc** algorithm to encrypt all images in the database, the corresponding features as well as the index tree, which will be outsourced to the cloud after encryption. The **Query_Gen** algorithm allows image users to generate a secure image search request. On receiving users' request, the cloud server runs the **Image_Retri** algorithm to retrieve the ciphertexts of relevant images. We now introduce the detail of each algorithm in our scheme.

Key_Gen: With the security parameter λ, the image owner generates the image encryption key k_m and the feature vector encryption key $k_f = \{\gamma, M, M^{-1}\}$. Specifically, $\gamma \in \mathbb{Z}_{p_1}$ is a

random number and public, M is a $(d+\alpha+2l+2)\times(d+\alpha+2l+2)$-dimensional random invertible matrix, M^{-1} is inverse of M, p_1, p_2 are two integers that define the range of number, $p_1 \gg p_2$. The image owner publishs $\{p_1, p_2\}$ and sends k_m, k_f to search users.

Index_Buld: In this stage, the image owner needs to generate the index tree according to the feature vectors. Then, for all non-leaf nodes of index tree, the owner creates a polynomial for it based on its children's user roles. We will describe these steps one by one.

1) Feature Extraction. For each image m_i in the image set $\mathcal{M} = \{m_1, ..., m_n\}$, the image owner employs the pre-trained network RegNet [19] to extract a feature vector and compresses it into a d-dimensional bit vector by using PCA [9].

2) Hierarchical K-means Algorithm. The image owner devides all the feature vector set $\mathcal{F} = \{f_1, f_2, ...f_n\}$ into K different clusters by K-means. Specifically, K random vectors are chosen as the centroid vectors; then, each vector computes their distance to the centroid vectors and joins in the nearest cluster; finally, the centroid vectors will be updated as the mean value of new vectors in the current cluster. The second and third steps will be executed iteratively until a fixed number of times or the centroid vectors become stable.

After initial partitioning, the vectors in each subspace may still be too many. In this case, to ensure the number of vectors in each subspace less than the fixed number K, the image owner will execute K-means on the corresponding subspace where excessive vectors exist. The partitioning will not end until each subspace meets the requirements. After the index building, each feature vector will become one leaf node of the index tree.

3) Fine-Grained Access Control. The polynomial-based access policy implements fine-grained access control for different authorized users, where each authorizer is the one allowed to access the specific image associated with it. Specifically, we assume there are ρ roles in the access role set $\mathcal{A} = \{a_1, a_2, ..., a_\rho\}$ and given a positive integer E. For each image $m_i \in \mathcal{M}$, we construct a role polynomial as in

$$g_i(x) = E \cdot \prod_{a_t \in \mathcal{A}} (x - a_t)^2 = \sum_{j=0}^{2\rho} a'_{ij} x^j, \tag{1}$$

where a'_{ij} is the coefficient of the role polynomial $g_i(x)$ of image m_i. To decide whether certain search user can access the image m_i, it just needs to check whether certain search user's access role x_1 is a root of polynomial $g_i(x)$. If $g_i(x_1) = 0$, it means that x_1 is a root of $g_i(x)$ and certain search user is permitted to access this image. Otherwise, $g_i(x_1) \neq 0$ means x_1 is not the root of $g_i(x)$ and $g_i(x_1)$ will be larger than E, which seriously affect the results of similarity calculation.

With the cofficients of each image's polynomial, the image owner extends the corresponding feature vectors $f_i = (f_{i1}, f_{i2}, ..., f_{id})$ to $f'_i = (f_i, \|f_i\|, a'_{i,0}, a'_{i,1}, ..., a'_{i,2\rho}, \theta)$ in the index tree, where $\|f_i\|$ represents the modules length of vector f_i, $a'_{i,j} (j = 0, 1, ..., 2\rho)$ represents the coefficients of the access control polynomial, $\theta \in \mathbb{Z}_{p_2}^{\alpha}$ is a random vector. For all non-leaf nodes of the index tree, the image owner creates a polynomial for it based on its children's user roles from the bottom up. Specifically, if a user can search a node's all child with his/her role, he/she should also be permitted to search this node. In this case, roots of a non-leaf node's polynomial should contain all roots of its children's polynomial. Finally, the image owner extends all nodes vectors, indicating that the index tree has been built.

Data_Enc: In order to protect the privacy of images from the cloud, the image owner will encrypt all images, feature vectors associated with leaf and non-leaf nodes of the index tree. The image owner utilizes the conventional encryption method AES and the key k_m to encrypt image set $\mathcal{M} = \{m_1, m_2, ..., m_n\}$, then outputs encrypted image set $\mathcal{C} = \{c_1, c_2, ..., c_n\}$. As our design has the same encryption process for non-leaf nodes and leaf nodes f'_i, we show the encryption of f'_i for expression simplicity as below.

Given an index vector f'_i, the image owner encrypts f'_i as

$$\hat{f}_i = (\gamma \cdot f'_i + \varepsilon_i) \cdot M, \tag{2}$$

where $\varepsilon_i \in \mathbb{Z}_{p_2}^{d+\alpha+2\rho+2}$ is a random integer noise vector, $2|max(\varepsilon_i)| \ll \gamma, |max(\varepsilon_i)|$ denotes the maximum absolute value of ε_i's elements.

After encrypting feature vectors, the image owner replaces all feature vectors and non-leaf nodes assigned to nodes of the index tree with their corresponding ciphertexts. Then the encrypted images and index tree are uploaded to the cloud server for storage and retrieval.

Query_Gen: To perform an encrypted image search, the user generates a query trapdoor for the queried image m_q. Specifically, the user first extracts the feature vector $f_q = (f_{q,1}, f_{q,2}, ..., f_{q,d})$ from m_q and randomly selects a vector $\beta \in \mathbb{Z}_{p_2}^{\alpha}$. Given the user's access role $\widetilde{x}$, f_q will be extended as $f'_q = (-2r_q f_q, 1, 1, \widetilde{x}, \widetilde{x}^2, ..., \widetilde{x}^{2\rho}, \beta)$, where $r_q \in \mathbb{Z}_{p_2}$ and positive in each query. After that, f'_q is encrypted as

$$\hat{f}_q = M^{-1} \cdot (\gamma \cdot f_q'^{\top} + \varepsilon_q^{\top}), \tag{3}$$

where $\varepsilon_q \in \mathbb{Z}_{p_2}^{d+\alpha+2\rho+2}$ is a random integer noise vector that is different in each query, $f_q'^{\top}$ and $\varepsilon_q^{\top}$ are column vectors of f'_q and ε_q, respectively. Finally, the encrypted trapdoor will be formally denoted as $\mathcal{TD} = \{\hat{f}_q\}$ and sent to the cloud for retrieving similar images.

Image_Retri: On receiving the search request from users, the cloud server searches the k most similar images through the root node to leaf nodes of the encrypted index tree. At each level, the cloud finds the entry point of the next level by finding the node with a vector that has the distance to the trapdoor as

$$\begin{aligned} D_{i,q} &= \lceil \frac{\hat{f}_i \cdot \hat{f}_q}{\gamma^2} \rfloor_{p_1} \\ &= r_q(\|f_i - f_q\|^2 - \|f_q\|^2) + \sum_{i=1}^{\alpha} \theta_i \beta_i + g_i(x), \end{aligned} \tag{4}$$

where $\lceil \rfloor_{p_1}$ is the nearest integer with modulus p_1. Denote U as the upper bound of $D_{i,q}$. If the access role $\widetilde{x}$ of the user

meets the polynomial-based access policy, then $g_i(\widetilde{x}) = 0$ and $D_{i,q} \leq U$. Otherwise, $g_i(\widetilde{x})$ would be a great number so that $D_{i,q} \gg U$, which indicates the user is not allowed to access the node. The cloud server can find out the node that has the smallest D as the entry node of the next level. When the cloud server arrives the leaf node, it will get the identity of encrypted image that is the most similar to the queried image and add it to the result list $\mathcal{R}$. If the size of $\mathcal{R} < k$, the cloud server will go back to the sibling or parent node of this leaf node to find the next nearest neighbor node until k similar images are available.

III. Security and Performance Analysis

To illustrate that our scheme is secure and feasible in practice, we present the detailed security and performance analysis separately.

A. Security Analysis

In this section, we analysis the security from image privacy, index and trapdoor privacy and query unlinkability.

1) Image Privacy. In our scheme, each image $m_i \in \mathcal{M}(i \in [1, n])$ is encrypted by the conventional encryption method AES with key k_m. As long as the key k_m is secure, the content of the image set can be well protected.

2) Index and Query Privacy. We first define the Learning With Errors (LWE) problem [20], then prove the index and query privacy under *known ciphertext attack model* and *known background attack model* given in Section II-B.

Define 1 (LWE problem): Given polynomial many samples of $(x_i \in \mathbb{Z}_{p_1}^{d+\alpha+2\rho+2})$ with $y_i = f \cdot x_i^\top + \varepsilon_i$, where the error term $\varepsilon_i \in \mathbb{Z}_{p_1}$ is drawn from some probability distribution, it is computationally infeasible to recover the vector f with non-negligible probability.

Theorem 1: Our scheme can protect the index and query privacy under the known ciphertext attack model.

Proof: In our scheme, each extended feature f_i' in the index tree is encrypted as $\hat{f}_i = (\gamma \cdot f_i' + \varepsilon_i) \cdot M$.

Since f_i' and ε_i are two vectors, we can convert $\hat{f}_i$ into $d + \alpha + 2\rho + 2$ dot products of $d + \alpha + 2\rho + 2$-dimensional vectors as $\hat{f}_i(k) = \gamma \cdot f_i' \cdot M(k) + \varepsilon_i \cdot M(k)$, where $1 \leq k \leq d+\alpha+2\rho+2$, $\hat{f}_i(k)$ is the kth element of $\hat{f}_i$, and $M(k)$ is the kth column of M. We denote $\gamma \cdot M(k)$ as $\widetilde{M}(k)$ and $\varepsilon_i \cdot M(k)$ as $\widetilde{\varepsilon_i}$, then we have $d + \alpha + 2\rho + 2$ samples $(\widetilde{M}(k),\ \hat{f}_i(k))$ with $\hat{f}_i(k) = f_i' \cdot \widetilde{M}(k) + \widetilde{\varepsilon_i}$.

Therefore, recovering f_i' from the encryption vector $\hat{f}_i(k)$ becomes the LWE problem given in Definition 1. For the polynomial-time adversary, it is computationally infeasible to solve the LWE problem. Moreover, matrix M and $\widetilde{M}$ are not available to the adversary, thus recovering f_i' from $\hat{f}_i$ more difficult than solving the LWE problem. Similarly, non-leaf nodes in the index tree are also protected.

With regard to query f_q and its ciphertext $\hat{f}_q$, we also can convert its encryption to $d + \alpha + 2\rho + 2$ dot products as $\hat{f}_q(k) = \widetilde{M^{-1}}(k) \cdot f_q'^\top + \widetilde{\varepsilon_q^\top}$, where $\widetilde{M^{-1}}(k) = \gamma \cdot M^{-1}(k)$, $M^{-1}(k)$ is the kth row of M^{-1} and $\widetilde{\varepsilon_q^\top} = M^{-1}(k) \cdot \varepsilon_q^\top$. Thus, recovering $\hat{f}_q$ is reduced to solve the LWE problem by given $d+\alpha+2\rho+2$ samples $(\widetilde{M^{-1}}(k), \hat{f}_q(k))$. In addition, M^{-1} is also not available to the adversary since M is the secret key and only known to user. Therefor, if the LWE problem is hard, recovering index and query vectors from their corresponding ciphertexts is computationally infeasible for the polynomial-time adversary. That is, the privacy of index and query is protected well under the *known ciphertext attack model*.

Theorem 2: Our scheme can protect the index and query privacy under the known background attack model.

Proof: In the *known background attack model*, the adversary may attempt a linear analysis attack to recover the plaintexts from the results of inner product calculations. Specifically, given a query vector f_q and any two index vectors f_a, f_b, the adversary can construct an equation as $D_{a,b} = r_q(\|f_a - f_q\|^2 - \|f_b - f_q\|^2)$. The linear analysis attack is that if the adversary can obtain more than $2d$ query vectors, it can construct $2d$ for $2d$ unknown index vectors (f_a, f_b) and recover f_a, f_b from them. However, due to the existence of the random number r_q, such an attack is not feasible in our scheme. r_q in each query is different. Therefore, our scheme can protect the privacy of index and query under the *known background attack model*.

3) Query Unlinkability. We prove that our scheme can achieve query unlinkability by proving the following theorem.

Theorem 3: Our scheme can guarantee that the adversary cannot distinguish any two queries that are generated from the same query image.

Proof: Assuming that the two feature vectors f_{q_1}, f_{q_2} are extracted from the query image m_q, the user randomly selects different r_{q_1}, r_{q_2} and β_1, β_2 to extend f_{q_1}, f_{q_2} into two different f'_{q_1}, f'_{q_2}, respectively. Then, the user chooses two different $\varepsilon_{q_1}, \varepsilon_{q_2}$ to encrypt f'_{q_1}, f'_{q_2} with $\hat{f_{q_1}} = M^{-1} \cdot (\gamma \cdot f_{q_1}'^\top + \varepsilon_{q_1}^\top)$, $\hat{f_{q_2}} = M^{-1} \cdot (\gamma \cdot f_{q_2}'^\top + \varepsilon_{q_2}^\top)$, respectively. After these several operations, the ciphertexts $\hat{f_{q_1}}$ and $\hat{f_{q_2}}$ will become completely different. Therefore, according to $\hat{f_{q_1}}$ and $\hat{f_{q_2}}$, the adversary cannot distinguish $\hat{f_{q_1}}$ and $\hat{f_{q_2}}$ are generated from the same query image m_q. That is, our scheme can guarantee the query unlinkability.

B. Performance Analysis

The performance of our scheme construction is evaluated from two aspects of accuracy and efficiency. We perform a series of experimental evaluations using Python 3.11 on the Windows 11 operating system with Intel Core (TM) i7-12700H CPU. Through transfer learning, we extract global average pooling layer of RegNet [19], pre-trained on the ImageNet dataset. All the images are taken from the Corel-10K dataset, which contains 100 categories of images and each category contains 100 similar images.

1) Accuracy. The measure of accuracy evaluation is precision at top-k (P@k), defined as $P@k = \frac{k_c}{k}$. Here, k is the number of returned similar images, and k_c is the number of images in the returned k similar images that belong to the same category as the query image. Since the top-k results are

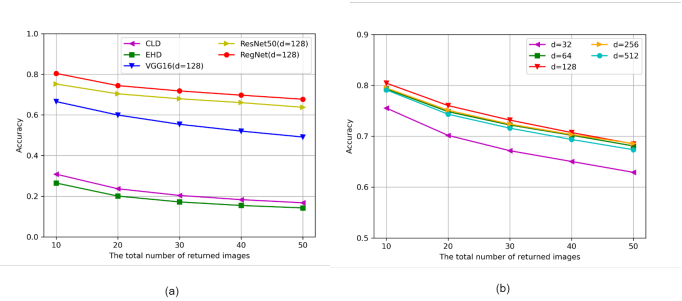


Fig. 2. Accuracy under various situations.

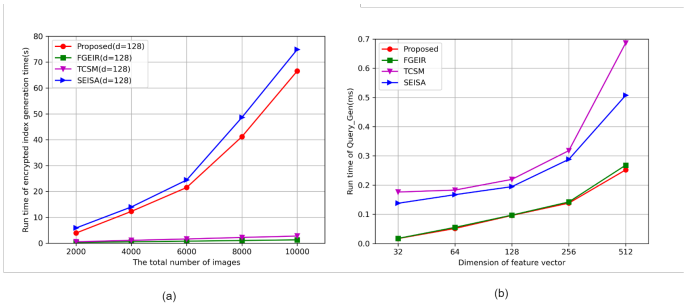


Fig. 3. Encrypted index generation and Query_Gen costs of various schemes.

determined by euclidean distance between f_i and f_q during the search process, the key factors affecting the retrieval accuracy are the feature extraction method and dimensional reduction method.

Fig. 2(a) and (b) show the impact of these factors on retrieval accuracy. In Fig. 2(a), when the total number of images is 10000 and the total number of returned images k varies from 10 to 50, the retrieval accuracy generated by the feature vectors extracted from CLD and EHD and the three pretrained models VGG16, ResNet50, and RegNet decreases. This is because the total number of images does not increase with the increase of k. We can see that the feature vector extracted by the CNN method performs better than those of CLD and EHD. In addition, the retrieval accuracy of the RegNet model is better than those of VGG16 and ResNet50 models overall. Therefore, in Fig. 2(b) we test the features extracted by the RegNet model with different dimensions. The PCA technology can be used to reduce 512 dimension feature vectors to 32, 64, 128 and 256 dimensions. In Fig. 2(b), when the total number of images is 10000, the retrieval accuracy does not increase as the dimension increases. It can be seen that the retrieval accuracy is relatively high when $d = 128$.

Further, in Table 1, we test the effect of polynomial coefficients over retrieval accuracy, where $d = 128$. We notice that there is almost no difference in retrieval accuracy for the same k, which indicates that polynomial coefficients in our scheme do not affect the search accuracy.

2) Efficiency. We test the execution time of algorithms with FGEIR [8], TCSM [12] and SEISA [21]. In Fig. 3(a), in order to control variables, We draw the encrypted index generation time which includes index generation and indexes encryption and excludes the feature extraction time by varying the number of images. The more the total number of images are, the more time it takes to generate the encrypted index

TABLE I
ACCURACY UNDER CORRECT ROLES

k	10	20	30	40	50
Yes[1]	0.8044	0.7603	0.7317	0.7072	0.6849
No[2]	0.8045	0.761	0.7326	0.7094	0.6904

[1] "Yes" indicates that polynomial coefficients are added.
[2] "No" indicates that polynomial coefficients are not added.

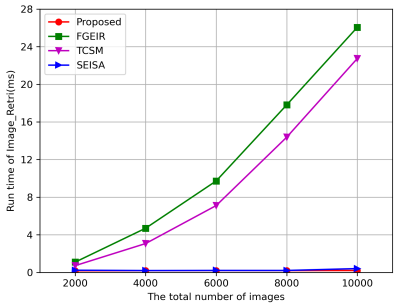


Fig. 4. Run time of Image_Retri.

when setting $d = 128, k = 20$. Since FGEIR and TCSM are linear search and not need to build index tree, that will be less time in generating encrypted index. And because the index construction is a one-time operation, it will not be generated repeatedly due to the updates. Therefore, the encrypt overhead of our scheme is acceptable in practice.

Fig. 3(b) plots the curves between trapdoor generation cost and feature vector dimension, it can be seen that computation costs increase with the increase of dimensions. However, our scheme is less expensive since we only use one matrix to encrypt the feature vector and there is no split operation. In particular, the two curve of our scheme and FGEIR almost coincide beacause the query trapdoor generated by our scheme and FGEIR are almost the same, but FGEIR needs to re-encrypt the query before uploading to the cloud server.

In Fig. 4, we display the image retrieval time in *Image_Retri*, which is almost constant with the total number of images as $k = 20$ in our scheme. The time required for our scheme, FGEIR, TCSM and SEISA to retrieve a image is 0.19 ms, 26.05 ms, 22.73 ms and 0.4 ms when the total number of image is 10000. In our scheme, the cloud server does not need to traverse the entire tree completely, which results in a shorter retrieval time than FGEIR and TCSM. In addition, our scheme and SEISA use different encryption schemes to protect the index, which can speed up the retrieval while ensuring security. In other words, when the user searches in the database of massive images, our solution improves the access speed and gives the user a better search experience.

IV. Conclusions

In this paper, we propose an efficient fine-grained privacy-preserving image retrieval scheme. Specifically, to improve the retrieval accuracy, we first obtain a better image representation than global or local features through transfer learning. Then, we construct a hierarchical index tree by using the K-means algorithm to enhance the retrieval efficiency. The access control based on image permission is achieved by implementing a polynomial access strategy. By integrating indexes with access policies, we can accelerate access to ciphertext domains in cloud environments. Finally, the experiments present that compared to other access control schemes, our scheme allows for fast access while ensuring the security of outsourced data.

Acknowledgment

This work was supported in part by the National Key R&D Program of China under Grant 2022YFB3103500, in part by the National Natural Science Foundation of China under Grants 62302195, and in part by the GuangDong Basic and Applied Basic Research Foundation under Grant 2022A1515011960.

References

[1] Z. Xia, N. N. Xiong, A. V. Vasilakos, and X. Sun, "Epcbir: An efficient and privacy-preserving content-based image retrieval scheme in cloud computing," *Information Sciences*, vol. 387, pp. 195–204, 2017.

[2] D. Li, Q. Lü, X. Liao, T. Xiang, J. Wu, and J. Le, "Avpmir: Adaptive verifiable privacy-preserving medical image retrieval," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–15, 2024. DOI: 10.1109/TDSC.2024.3355223.

[3] A. Dagan, I. Guy, and S. Novgorodov, "Shop by image: Characterizing visual search in e-commerce," *Information Retrieval Journal*, vol. 26, no. 1, p. 2, 2023.

[4] J. Anju and R. Shreelekshmi, "A faster secure content-based image retrieval using clustering for cloud," *Expert Systems with Applications*, vol. 189, p. 116 070, 2022.

[5] D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in *Proceeding 2000 IEEE symposium on security and privacy. S&P 2000*, IEEE, 2000, pp. 44–55.

[6] C. Mao, Z. Shen, K. Chen, Y. Liu, Q. Meng, and F. Wang, "Dcirm: Dynamic and controllable image retrieval scheme in multi-owner multi-user settings," *IEEE Transactions on Services Computing*, 2024.

[7] T. Yang, J. Ma, Y. Miao, *et al.*, "Mu-teir: Traceable encrypted image retrieval in the multi-user setting," *IEEE Transactions on Services Computing*, vol. 16, no. 2, pp. 1282–1295, 2022.

[8] Y.-H. Chen and M.-C. Huang, "Fine-grained encrypted image retrieval in cloud environment," *Mathematics*, vol. 12, no. 1, p. 114, 2023.

[9] Q. Tong, Y. Miao, L. Chen, *et al.*, "Vfirm: Verifiable fine-grained encrypted image retrieval in multi-owner multi-user settings," *IEEE Transactions on Services Computing*, vol. 15, no. 6, pp. 3606–3619, 2021.

[10] W. Lu, A. Swaminathan, A. L. Varna, and M. Wu, "Enabling search over encrypted multimedia databases," in *Media Forensics and Security*, SPIE, vol. 7254, 2009, pp. 404–414.

[11] Y. Li, J. Ma, Y. Miao, *et al.*, "Dvrei: Dynamic verifiable retrieval over encrypted images," *IEEE Transactions on Computers*, vol. 71, no. 8, pp. 1755–1769, 2021.

[12] Y. Li, J. Ma, Y. Miao, *et al.*, "Traceable and controllable encrypted cloud image search in multi-user settings," *IEEE Transactions on Cloud Computing*, vol. 10, no. 4, pp. 2936–2948, 2020.

[13] J. S. Li, I. H. Liu, C. J. Tsai, Z. Y. Su, and C. G. Liu, "Secure content-based image retrieval in the cloud with key confidentiality," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2020.

[14] M. Li, M. Zhang, Q. Wang, S. S. M. Chow, and C. Lit, "Instantcryptogram: Secure image retrieval service," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018.

[15] J. Qin, H. Li, X. Xiang, *et al.*, "An encrypted image retrieval method based on harris corner optimization and lsh in cloud computing," *IEEE Access*, pp. 24 626–24 633, 2019.

[16] L. Yingying, J. Ma, M. Yinbin, Y. Wang, and K. K. R. Choo, "Similarity search for encrypted images in secure cloud computing," *IEEE Transactions on Cloud Computing*, vol. PP, no. 99, pp. 1–1, 2020.

[17] Y. Li, J. Ma, Y. Miao, L. Liu, and K. K. R. Choo, "Secure and verifiable multi-key image search in cloud-assisted edge computing," *IEEE Transactions on Industrial Informatics*, vol. PP, no. 99, pp. 1–1, 2020.

[18] S. Khan, H. Abbas, and W. Iqbal, "Verifiable privacy-preserving image retrieval in multi-owner multi-user settings," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[19] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 428–10 436.

[20] Z. Brakerski, C. Gentry, and S. Halevi, "Packed ciphertexts in lwe-based homomorphic encryption," in *Public-Key Cryptography–PKC 2013: 16th International Conference on Practice and Theory in Public-Key Cryptography, Nara, Japan, February 26–March 1, 2013. Proceedings 16*, Springer, 2013, pp. 1–13.

[21] J. Yuan, S. Yu, and L. Guo, "Seisa: Secure and efficient encrypted image search with access control," in *2015 IEEE conference on computer communications (INFOCOM)*, IEEE, 2015, pp. 2083–2091.

EAViT: External Attention Vision Transformer for Audio Classification

Aquib Iqbal[†*], Abid Hasan Zim[‡*], Md Asaduzzaman Tonmoy[§], Limengnan Zhou[¶], Asad Malik[‖], Minoru Kuribayashi[††]

[†] Department of Computer Science, University of Massachusetts, Amherst, MA, USA, E-mail: aquibiqbal@umass.edu

[‡] Department of Mechanical Engineering, Aligarh Muslim University, India, E-mail: abid@zhcet.ac.in

[§] Department of Electrical Engineering, Aligarh Muslim University, India, E-mail: asaduzzaman.tonmoy@outlook.com

[¶] School of Electronic and Information Engineering, UESTC, China, E-mail: dreamzlmn@foxmail.com

[‖] School of Information Technology, Monash University Malaysia, Malaysia, E-mail: asad.malik@monash.edu

[††] Center for Data-driven Science and Artificial Intelligence, Tohoku University, Japan, E-mail: kminoru@tohoku.ac.jp

***Abstract*—This study presents the External Attention Vision Transformer (EAViT) model, a novel approach designed to improve audio classification accuracy in response to the growing need for precise classification systems, particularly for recommendation engines and user personalization in applications such as music streaming. Accurate audio classification is crucial for organizing vast audio libraries into coherent categories, enabling users to find and interact with their preferred audio content more effectively. In this study, we utilize the GTZAN dataset, which comprises 1,000 music excerpts spanning ten diverse genres. Each 30-second audio clip is segmented into 3-second excerpts to enhance dataset robustness and mitigate overfitting risks, allowing for more granular feature analysis. The EAViT model integrates multi-head external attention (MEA) mechanisms into the Vision Transformer (ViT) framework, effectively capturing long-range dependencies and potential correlations between samples. This external attention (EA) mechanism employs learnable memory units that enhance the network's capacity to process complex audio features efficiently. The study demonstrates that EAViT achieves a remarkable overall accuracy of 93.99%, surpassing state-of-the-art models.**

I. Introduction

The increasing availability of digital music resources has led to rapid advancements in multimedia technology. As a result, consumers have progressively shifted towards accessing music through online streaming platforms. AI methods are particularly effective in addressing the complex task of meeting diverse and intricate music retrieval needs from extensive music collections. For instance, consumers may desire to listen to a specific song within a particular genre that conveys a certain emotional tone. In these instances, accurate music labelling is crucial to ensure the precise music is delivered [1], [2]. Additionally, numerous subscription and recommendation systems require the inclusion of music genre preferences to provide customers with more specific and tailored content. Music is a highly diverse art form that encompasses various elements, including rhythm, melody, and harmony. Music media systems use textual labels to categorize or retrieve music, since comprehending music in its exact form requires substantial previous knowledge. Hence, the classification of music genres is an essential component of music information retrieval [3].

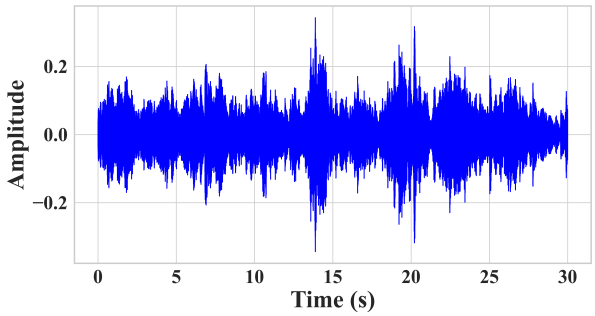


Fig. 1. Raw waveform representation of an audio sample.

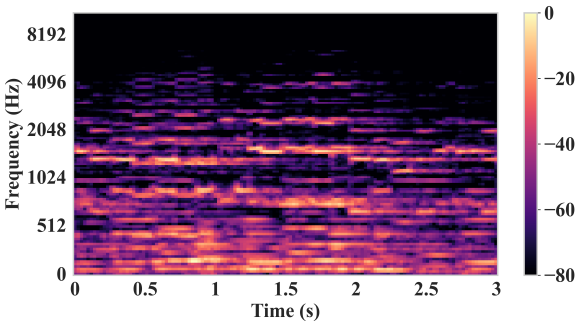


Fig. 2. Mel spectrogram of an audio sample.

Over the past decade, traditional machine learning classifiers have been employed for this purpose. However, these classifiers possess a shallow structure, limiting their capacity to effectively learn and interpret music data. Manual extraction of musical elements has shown limited applicability and lacks robustness. To address the intricate aspects of music, innovative deep-learning approaches have been adopted in recent years. These advanced techniques offer improved accuracy

*Equal contributions.

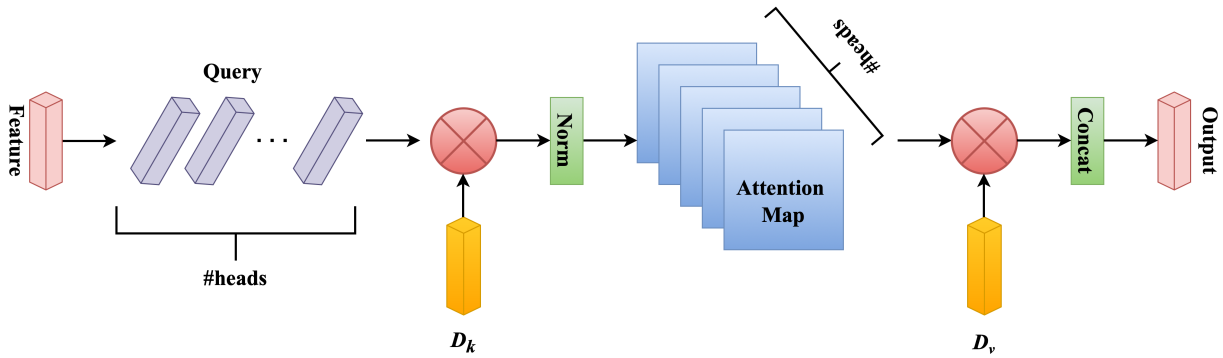


Fig. 3. Multi-head external-attention (MEA).

and efficiency in music genre classification and retrieval [4], [5]. The most recent advancements in audio classification performance are generated using distinct resource-intensive methodologies [6].

Over the years, a large number of classification methods have been created to categorize audio-based data [7]. A study conducted video classification and audio event identification using a convolutional neural network (CNN) on an extensive dataset. By utilizing subsets of varying sizes and initializing convolutional layers with a gammatone filter bank, which is modelled after the human ear, the study examined the impact of these methods on the system's accuracy. This approach allowed for a thorough investigation of model training and performance [8]. Researchers have employed Long Short-Term Memory (LSTM) networks to achieve promising results for the MGC task [9]. Another study introduced a hybrid architecture combining CNNs and RNNs to capture the spatiotemporal characteristics of music [10]. Additionally, a study proposed an attention mechanism-based Bi-LSTM architecture to exploit the varying importance of temporal frames, further enhancing the accuracy and performance of music classification tasks [11]. A research study compared a CNN and a TDSN for environmental sound classification using spectrogram images. The results revealed that the CNN achieved superior performance on the ESC-10 dataset, with an accuracy of 77%, compared to 56% for the TDSN [12]. Additionally, a lightweight CNN utilizing MFCC for environmental sound classification on the UrbanSound8k dataset achieved a notable accuracy of 95.59%, showcasing competitive performance against more complex models [13]. A further investigation concentrated on classifying music across three distinct datasets (ISMIR 2004, LMD, and ethnic African music). The findings revealed that CNN outperformed previous classifiers by achieving an impressive accuracy rate of 92% [14].

Recently introduced transformer-based models offer substantial advancements in AI applications, exhibiting notable performance enhancements compared to current models [15], [16]. Recent advancements include the Causal Audio Transformer (CAT), which employs MRMF extraction and an acoustic attention block for enhanced audio modelling. The proposed causal module mitigates overfitting, facilitates knowledge transfer, and improves interpretability. It demonstrated superior performance on datasets such as ESC50, AudioSet, and UrbanSound8K, and are adaptable to various other transformer-based models [17]. Another study applied a transformer model to music genre classification, converting audio to log-amplitude Mel spectrograms, and achieved promising results on the GTZAN dataset, indicating the potential of transformers in this domain [18].

In this study, EAViT with MEA is proposed to classify music genres in the GTZAN dataset. The proposed approach includes the introduction of an innovative EA mechanism to enhance the classification accuracy of ViT-based models when applied to audio data. EAViT integrates the strengths of traditional vision transformers with specialized attention mechanisms tailored for audio signal processing, achieving superior performance over conventional models. Extensive experiments were conducted on the GTZAN dataset to benchmark our model against state-of-the-art methods, demonstrating significant improvements in classification accuracy. Our contributions in this paper are summarized as follows:

- The EAViT model is proposed for the task of music genre classification, leveraging the integration of MEA mechanisms within its framework.
- A 30-second raw audio clip is segmented into 3-second portions. Each segment is transformed into the frequency domain via Short-Time Fourier Transform (STFT), converted to decibel units to produce a spectrogram, and further processed with a mel filter bank to create a mel spectrogram, which is then saved as an image.
- The proposed models are compared with state-of-the-art methods from existing literature.

The rest of the paper is structured as follows: Section II outlines the methodology, Section III explains the proposed method, Section IV presents the experiments and results, and Section V contains the conclusion.

II. Methodology

A. Data

GTZAN dataset was used which comprises of 1,000 audio excerpts spanning 10 different genres: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, and Rock. Each

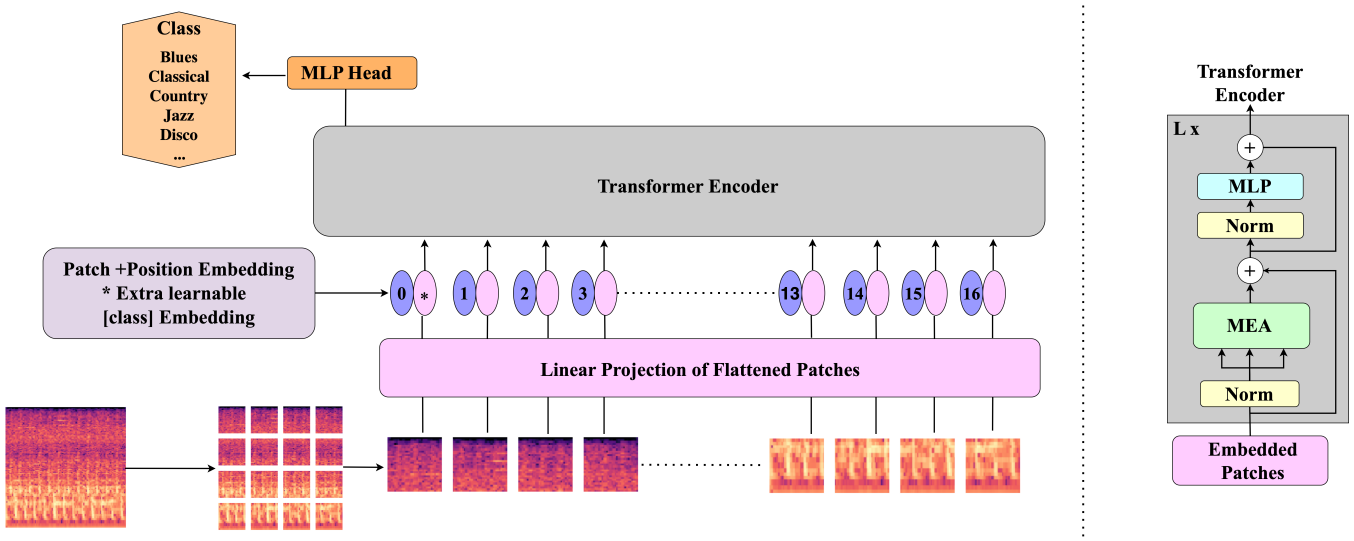


Fig. 4. Model Overview: EAViT.

genre includes 100 excerpts, each lasting approximately 30 seconds and stored as 22,050 Hz, 16-bit, mono audio files [19].

We segmented each 30-second audio excerpt from the GTZAN dataset into 3-second clips, significantly expanding the dataset and mitigating the risk of overfitting. This segmentation allows for a more granular analysis of audio features, enhancing the dataset's robustness. To illustrate the transformation process, we included both the raw audio signals and the corresponding mel spectrogram Fig. 1 and Fig. 2.

Mel Spectrogram represents sound in the time-frequency domain, using the mel scale on the y-axis and the decibel scale to indicate colour intensity. This spectrogram is created by applying a filter bank to the frequency-domain representation of time-windowed audio signals, providing a perception-aligned view of the sound. The conversion process involves transforming each 3-second audio segment into the frequency domain using the STFT. The resulting frequency data is then converted into decibel units to form a standard spectrogram. By applying a mel filter bank, we transform this spectrogram into a mel spectrogram, which is then saved as an image file.

B. Multi-head external-attention (MEA)

In this study, we incorporated the MEA mechanism into the transformer encoder of the ViT. To improve the network's capabilities, the EA mechanism uses two learnable memory units, $M_k \in \mathbb{R}^{d\times S}$ and $M_v \in \mathbb{R}^{d\times S}$, which serve as key and value components. Importantly, M_k and M_v function as memories for all samples in the training set and are independent of the input. Specifically, the flow process of the EA module is defined by Eq.(1) and (2), as follows:

$$A = (\alpha)_{n,m} = \text{Norm}(F_{\text{in}} M_k^T), \tag{1}$$

$$F_{\text{out}} = AM_v, \tag{2}$$

where F_{in} represents the input feature, F_{out} represents the output feature, and Norm$(\cdot)$ denotes the double normalization operator. MEA can be written as:

$$h_i = \text{ExternalAttention}(F_i, M_k, M_v), \tag{3}$$

$$F_{\text{out}} = \text{MultiHead}(F, M_k, M_v) \tag{4}$$

$$= \text{Concat}(h_1, \ldots, h_H)W_o, \tag{5}$$

where h_i represents the i-th head, H denotes the total number of heads, and W_o is a linear transformation matrix ensuring consistent input and output dimensions. $M_k \in \mathbb{R}^{d\times S}$ and $M_v \in \mathbb{R}^{d\times S}$ serve as the shared memory units across different heads [20]. Here Fig. 3 shows the structure MEA.

III. Proposed Method

A. External Attention Vision Transformer (EAViT)

The conventional ViT utilizes multi-head self-attention, which enhances the feature at each position by computing a weighted sum of features based on pair-wise affinities across all positions. This process effectively captures long-range dependencies within a single sample. However, self-attention suffers from quadratic complexity and overlooks potential correlations between different samples. In contrast, EA leverages two small, learnable, shared memories, which can be efficiently implemented using two cascaded linear layers and two normalization layers. In this study, we have utilized the EAViT model, which incorporates MEA within the transformer encoder.

The ViT is specifically designed for image classification tasks by directly applying the transformer architecture to sequences of image patches [15]. It closely adheres to the original transformer design. In order to process 2D images, the input tensor $x \in R^{H\times W\times C}$ is transformed into a set of flattened 2D patches denoted as $x_P \in R^{N\times(P^2\cdot C)}$, where C

represents the number of channels. The main image has a resolution of (H, W), while each image patch has a (P, P) resolution. Consequently, the actual sequence length for the transformer is given by $N = HW/P^2$. The output of a learnable linear projection (as expressed in Eq.(6)) called patch embeddings, associates each vectorized patch with the model dimension D, as the transformer maintains consistent widths across all its layers.

$$\mathrm{z}_0 = \left[\mathrm{x}_{\text{class}}\,; \mathrm{x}_p^1\mathbf{E}\,; \mathrm{x}_p^2\mathbf{E}\,; \ldots\,; \mathrm{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos}, \tag{6}$$

where $\mathbf{E} \in \mathbb{R}^{\left(P^2 \cdot C\right) \times D}$ and $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$.

Researchers introduce a trainable embedding into the sequence of embedded patches, akin to BERT's $[class]$ token, represented as $(z_0^0 = x_{\text{class}})$. This embedding's state at the output of the Transformer encoder, denoted as (z_L^0), is used as the image representation y (refer to Eq.(9)). During both the pre-training and fine-tuning stages, a classification head is connected to z_L^0. In the pre-training phase, a Multi-Layer Perceptron (MLP) with a single hidden layer guides the classification head. In contrast, a single linear layer is utilized during the fine-tuning phase [21]. Here Fig. 4 shows the overview of the model.

$$\mathrm{z}'_\ell = \mathrm{MEA}\left(\mathrm{LN}\left(\mathrm{z}_{\ell-1}\right)\right) + \mathrm{z}_{\ell-1}, \quad \ell = 1 \ldots L \tag{7}$$

$$\mathrm{z}_\ell = \mathrm{MLP}\left(\mathrm{LN}\left(\mathrm{z}'_\ell\right)\right) + \mathrm{z}'_\ell, \quad \ell = 1 \ldots L \tag{8}$$

$$\mathrm{y} = \mathrm{LN}\left(\mathrm{z}_L^0\right) \tag{9}$$

To retain positional information, position embeddings are added to the patch embeddings. As more complex 2D-aware position embeddings did not significantly enhance performance, the researcher opted for traditional learnable 1D position embeddings. The generated sequence of embedding vectors is then fed into the encoder. The Transformer encoder comprises layers of MEA and MLP blocks, as described in Eq.(7) and (8). Each block is preceded by Layernorm (LN) and followed by residual connections.

In our proposed transformer-based methodology for the classification of 2D Mel spectrogram images, critical hyperparameters are meticulously selected to enhance model efficacy. The input images, resized to 256×256 pixels, are segmented into 16 patches. The learning rate is established at 0.001, accompanied by a weight decay of 0.0001 to mitigate overfitting, and the batch size is set at 256 to facilitate efficient training iterations. The transformer architecture incorporates 16 layers, each with a projection dimension of 32 and 8 attention heads. The MLP classification head comprises two hidden layers with [2048, 1024] units. These hyperparameters collectively define the model's capacity, attentiveness, and overall classification performance.

**The model has been reimplemented.

TABLE I
PROPOSED MODELS COMPARISON WITH SOTA METHODS FROM LITERATURE.

Model	Accuracy
ViT**	91.79%
RNNCA [22]	93.1%
1D-CNN with BiRNN and attention mechanism [23]	91.99%
SA-SLnO with optimization [24]	85.63%
EAViT (our)	**93.99%**

TABLE II
EVALUATION OF CLASS-SPECIFIC CLASSIFICATION RESULTS OF THE PROPOSED EAVIT MODEL

Class	Precision	Recall	F1-Score
Blues	0.94	0.96	0.95
Classical	0.99	0.97	0.98
Country	0.89	0.93	0.91
Disco	0.95	0.97	0.96
Hiphop	0.93	0.90	0.91
Jazz	0.92	0.94	0.93
Metal	0.96	0.93	0.95
Pop	0.99	0.92	0.96
Reggae	0.93	0.92	0.93
Rock	0.90	0.94	0.92

IV. EXPERIMENTS AND RESULTS

A. Evaluation Matrices

To evaluate the performance of the models, we employed key performance metrics, including accuracy, precision, recall, and the F1-score. Accuracy represents the ratio of correct predictions to the total number of predictions. Precision is calculated as the ratio of true positive predictions to the sum of true positive and false-positive predictions across all classes. Recall is the ratio of true positive predictions to the sum of true positive and false-negative predictions across all classes. The F1-Score is the harmonic mean of precision and recall, balancing the consideration of both false positive and false negative predictions. The mathematical expressions for these metrics are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{12}$$

$$\text{F1-score} = 2 \times \left(\frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}\right) \tag{13}$$

Where TP denotes True Positives, TN denotes True Negatives, FN denotes False Negatives, and FP denotes False Positives.

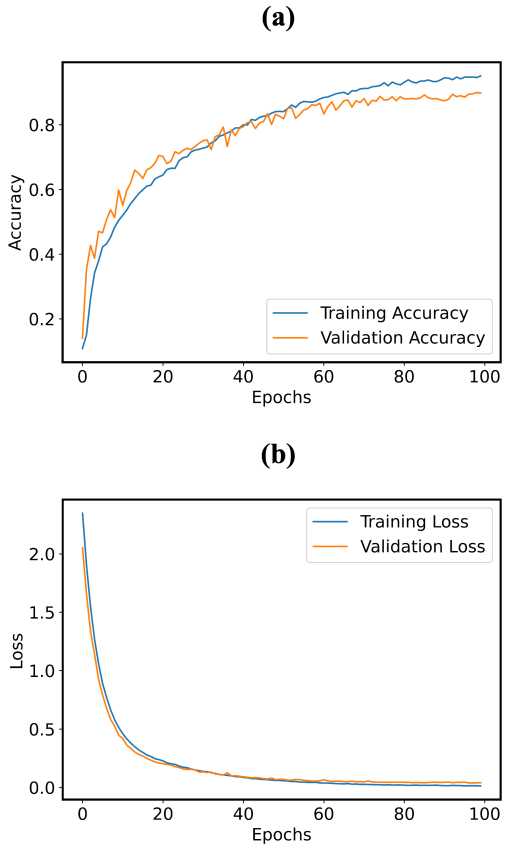


Fig. 5. proposed models: (a) Training and validation accuracy over epochs. (b) Training and validation loss over epochs.

B. Environment

The experiment was conducted using Python 3.11 and TensorFlow 2.16.1. The hardware employed for both training and testing included an AMD Ryzen 9 5900X processor, 64GB of RAM, and an Nvidia GeForce RTX 3090 GPU with 24GB of VRAM.

C. Analysis

In this study, the EAViT model was employed to classify ten different types of music genres. The GTZAN dataset was used in this study. EAViT achieved an impressive overall accuracy of 93.99%. The results underscore the exceptional performance of the EAViT model in music genre classification. Here, Fig. 5 (a) illustrates the training and validation accuracy, and Fig. 5 (b) presents the training and validation loss over a span of 100 epochs respectively. The graphs indicate that the model demonstrates optimal fitting to the dataset. The confusion matrix depicted in Fig. 6 provides a comprehensive overview of the model's performance, illustrating the models classification accuracy.

To demonstrate the robustness and efficacy of the proposed model in this study, we conducted a comparative analysis with state-of-the-art (SOTA) models from the literature, including ViT [15], Recurrent Neural Networks with Channel

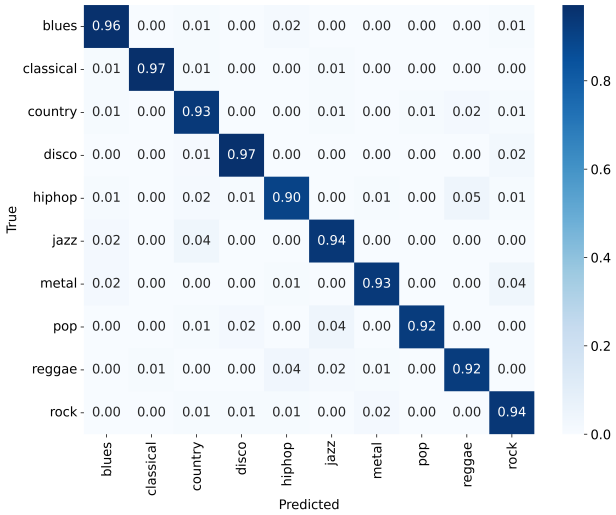


Fig. 6. Confusion matrix of the model.

Attention Mechanism (RNNCA) [22], 1D-CNN with BiRNN and attention mechanism [23], and SA-SLnO with optimization [24]. Table I unequivocally shows that the results of our proposed method surpass the performance of these SOTA models. EAViT surpasses the ViT by 2.37%, RNNCA by 0.95%, and the 1D-CNN with BiRNN and attention mechanism by 2.15%. Furthermore, in comparison to SA-SLnO with optimization model, EAViT achieves a remarkable 9.31% increase in classification accuracy. Here, Table II presents the evaluation metrics, including Precision, Recall, and F1-score, for each class of the proposed EAViT model. These findings underscore the robustness of the proposed model as it demonstrates consistent performance across various classes.

V. Conclusion

In this study, we introduced the External Attention Vision Transformer (EAViT) model, designed to enhance audio classification accuracy by integrating MEA mechanisms within the ViT framework. The EAViT model effectively captures long-range dependencies and correlations between different audio samples, addressing the limitations of traditional self-attention mechanisms. Through extensive experimentation on the GTZAN dataset, comprising 1,000 audio excerpts across ten genres, EAViT achieved a notable accuracy of 93.99%, outperforming state-of-the-art models such. Key metrics, including precision, recall, and F1-score, further demonstrated the model's robustness and reliability. The segmentation of 30-second audio clips into 3-second excerpts enhanced dataset robustness and allowed for detailed feature analysis. These results underscore the potential of the EAViT model to significantly advance audio classification, improving user experiences in various audio-related applications, including music streaming and environmental sound recognition. Future work could explore EAViT's application to other audio classification tasks and datasets, highlighting its versatility and efficacy.

References

[1] A. Elbir and N. Aydin, "Music genre classification and music recommendation by using deep learning," *Electronics Letters*, vol. 56, no. 12, pp. 627–629, 2020.

[2] Y. Li, W. Cao, W. Xie, J. Li, and E. Benetos, "Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes," *IEEE Transactions on Multimedia*, vol. 26, pp. 1346–1360, 2023.

[3] Y. Singh and A. Biswas, "Robustness of musical features on deep learning models for music genre classification," *Expert Systems with Applications*, vol. 199, p. 116 879, 2022.

[4] S. K. Prabhakar and S.-W. Lee, "Holistic approaches to music genre classification using efficient transfer and deep learning techniques," *Expert Systems with Applications*, vol. 211, p. 118 636, 2023.

[5] A. Sterling, J. Wilson, S. Lowe, and M. C. Lin, "Isnn: Impact sound neural network for audio-visual object classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 555–572.

[6] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Learning attentive representations for environmental sound classification," *IEEE Access*, vol. 7, pp. 130 327–130 339, 2019.

[7] S. Hershey, S. Chaudhuri, D. P. Ellis, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE, 2017, pp. 131–135.

[8] B. Mishachandar and S. Vairamuthu, "Diverse ocean noise classification using deep learning," *Applied Acoustics*, vol. 181, p. 108 141, 2021.

[9] J. Dai, S. Liang, W. Xue, C. Ni, and W. Liu, "Long short-term memory recurrent neural network based segment features for music genre classification," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2016, pp. 1–5.

[10] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 2392–2396.

[11] Y. Yu, S. Luo, S. Liu, H. Qiao, Y. Liu, and L. Feng, "Deep attention based music genre classification," *Neurocomputing*, vol. 372, pp. 84–91, 2020.

[12] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019.

[13] Y. A. Al-Hattab, H. F. Zaki, and A. A. Shafie, "Rethinking environmental sound classification using convolutional neural networks: Optimized parameter tuning of single feature extraction," *Neural Computing and Applications*, vol. 33, no. 21, pp. 14 495–14 506, 2021.

[14] Y. M. Costa, L. S. Oliveira, and C. N. Silla Jr, "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied soft computing*, vol. 52, pp. 28–38, 2017.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] A. H. Zim, A. Iqbal, A. Malik, Z. Dong, and H. Wu, "Tcnformer: Temporal convolutional network former for short-term wind speed forecasting," *arXiv preprint arXiv:2408.15737*, 2024.

[17] X. Liu, H. Lu, J. Yuan, and X. Li, "Cat: Causal audio transformer for audio classification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[18] Y. Zhuang, Y. Chen, and J. Zheng, "Music genre classification with transformer classifier," in *Proceedings of the 2020 4th international conference on digital signal processing*, 2020, pp. 155–159.

[19] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[20] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5436–5447, 2022.

[21] A. H. Zim, A. Ashraf, A. Iqbal, A. Malik, and M. Kuribayashi, "A vision transformer-based approach to bearing fault classification via vibration signals," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 1321–1326.

[22] J. Gan, "Music feature classification based on recurrent neural networks with channel attention mechanism," *Mobile Information Systems*, vol. 2021, no. 1, p. 7 629 994, 2021.

[23] K. Zhang, "Music style classification algorithm based on music feature extraction and deep neural network," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, p. 9 298 654, 2021.

[24] B. Kumaraswamy and P. Poonacha, "Deep convolutional neural network for musical genre classification via new self adaptive sea lion optimization," *Applied Soft Computing*, vol. 108, p. 107 446, 2021.

Successive Speaker Relative Transfer Function Estimation Through Relative Transfer Matrix in Noisy Reverberant Environments

Wageesha N. Manamperi, Thushara D. Abhayapala
The Australian National University, Canberra, Australia
E-mail: {*wageesha.manamperi, thushara.abhayapala*}*@anu.edu.au*

***Abstract*—Estimating the Relative Transfer Functions (ReTFs) of multiple speakers in a noisy reverberant environment are beneficial to signal processing applications of acoustic scene analysis, denoising, dereverberation, signal enhancement, and separation. Whilst most audio applications exploit the covariance matrix structure of multichannel recordings, estimating ReTF in the presence of simultaneously active multiple speakers is still considered to be a challenging problem. In this paper, we propose a novel method for estimating the ReTF using the relative transfer matrix (ReTM) of multiple microphones for multi-talker scenarios, which is suitable for noisy reverberant rooms. The method is based on the noise-only ReTM, and calculates the covariance matrices of (i) the first speaker and noise, and (ii) both speakers and noise at two microphone groups to reconstruct the ReTF of the second speaker. We demonstrate the ReTF estimation accuracy using numerical simulation of two speakers and two noise sources in a reverberant environment. The proposed method offers an accurate estimation with a low Hermitian angle. Additionally, the proposed algorithm is shown to better extract the voice of the successive speaker from the noisy microphone recordings over various SNR levels using a minimum variance distortionless response beamformer with improved noise reduction performance.**

I. Introduction

A relative transfer function (ReTF) is a spatial feature that describes the acoustic channel between two microphones in response to a single sound source [1]. Knowledge of desired sound source ReTF is essential in many spatial signal processing applications, such as beamforming, source localization, speech enhancement, source separation, and acoustic echo cancellation [1]–[5]. As a result, ReTF estimation of the target speaker in one or more interfering speakers within a noisy reverbarant room is an active problem in audio signal processing.

Many approaches have been developed for estimating the ReTF of a single active sound source, such as least-squares [1], [6], [7], covariance subtraction [8], [9] and covariance whitening [10]–[12], as well as manifold learning [13]. Common to these approaches is the use of signals' power spectral density (PSD) matrices calculated during the (i) noise-only, and (ii) speech plus noise, time segments for estimating the ReTF. Despite this, multiple and concurrent speakers scenarios are challenging problems as it could instead provide joint subspace spanning of the ReTF of all speakers [10].

There exist some techniques that generalize the ReTF for multiple simultaneous sound sources [14], [15]. The recently proposed Relative Transfer Matrix (ReTM) [15], where receivers allocate into two multi-microphone groups, is seen to be well performed when applied to speech enhancement [16], [17], and speaker separation [18] in a multi-source noisy reverberant environment.

Several ReTF estimation methods in a multiple and concurrent speaker scenario have been proposed, including (i) an expectation maximization algorithm that assumes only single source active at a given time-frequency bin and noise PSD is time-invariant [19]; (ii) a joint diagonalization method with simultaneous confirmatory factor analysis [20]; (iii) a more robust joint diagonalization based on linear algebraic concepts [21]; (iv) an orthogonal Procrustes problem with a priori information of the ReTFs [22]; and (v) blind oblique projection method [23]. Common drawbacks to these approaches are assuming W-disjoint orthogonality conditions, and a sufficiently high signal-to-noise ratio (SNR).

In [24], Gode and Doclo introduce a covariance blocking and whitening method to estimate the ReTF of the successive speaker in a noisy reverberant environment from two speakers activate successively. Alternatively, this paper proposes a novel method for estimating the ReTF of the second speaker by utilizing some of the properties of the covariance matrices imposed by the ReTM within noisy reverberant rooms, thus neither the estimate of the noise covariance matrix nor the ReTF of the first speaker requires unlike in [24]. The derivation of this method is based on the sole assumption that first speaker and multiple noise sources do simultaneously active in a given time segment prior to the time segment of the dual-speaker plus all noise sources as in [24]. Similar to the ReTM, we derive the proposed method by dividing multiple microphones into two groups and calculating the covariance matrices between them for both the single-speaker and dual-speaker segments. In the sequel, we first formulate the problem and introduce the ReTM. Secondly, in Section III, we present the new ReTF estimation method. Then in Section IV, we verify this ReTF estimation algorithm of the second speaker via a simulation study and show increased similarity with the ground truth ReTF

Thanks to Australian Research Council (ARC) Discovery Project Grant DP200100693 for funding.

vector in both free-field and high reverberant environments at low SNR levels. Finally, in Section V, we further evaluate the noise reduction performance in terms of SNR improvement when employing the proposed method in a minimum variance distortionless response (MVDR) beamformer.

II. Problem Definition

In this section, we formulate the problem of estimating the Relative Transfer Function (ReTF) of the second speaker in a noisy mixture of two speakers that activate successively. We first present the system model for the specific scenario, and then review the Relative Transfer Matrix (ReTM), used by the proposed ReTF estimation algorithm.

A. System Model

Consider a reverberant environment with two speakers that are successively activated and $\mathcal{L}$ background noise sources. Let there be Q arbitrary distributed microphones. In the short time Fourier transform (STFT) domain, we express the received signal as $y_q(f,t), q = 1, \cdots, Q$, and speech signal at each speaker as S_1 and S_2, and noise source signals as $N_\ell(f,t), \ell = 1, \cdots, \mathcal{L}$. The received signals in matrix form as

$$\mathbf{y}(f,t) = \mathbf{h}_1(f)S_1(f,t) + \mathbf{h}_2(f)S_2(f,t) + \mathbf{H}_N(f)\mathbf{N}(f,t), \tag{1}$$

where $\mathbf{y}(f,t) = [y_1(f,t), \ldots, y_Q(f,t)]^T$ and $\{\cdot\}^T$ is the matrix transpose, and $\mathbf{h}_1(f)$, and $\mathbf{h}_2(f)$ are the $[Q \times 1]$ vectors of the relative transfer functions (ReTFs) with respect to a reference microphone of the first and second speaker, respectively, $\mathbf{N}(f,t) = [N_1(f,t), \ldots, N_\mathcal{L}(f,t)]^T$ and $\mathbf{H}_N(f)$ is the Q by $\mathcal{L}$ noise sources transfer function matrix, which we define next.

The problem discussed in this paper is to estimate the ReTF vector of the successive speaker, $\mathbf{h}_2(f)$ from the mixture of multichannel recordings, $\mathbf{y}(f,t)$ assuming a specific scenario of three time segments with known segment boundaries as (i) multiple noise sources-only segment ($\mathcal{T}_1$), (ii) first speaker and multiple noise sources ($\mathcal{T}_2$), and (iii) first and second speakers and multiple noise sources ($\mathcal{T}_3$) as shown in Fig. 1

B. The Relative Transfer Matrix

Consider two groups of microphones $\{A\}$ and $\{B\}$ assigned with Q_A and Q_B microphones, respectively ($Q = Q_A + Q_B$). $\mathbf{y}_A(f,t)$, and $\mathbf{y}_B(f,t)$ are denoted as the vector of microphone signals for each group, therefore, the received signals at each microphone group are given as

$$\mathbf{y}_A(f,t) = \mathbf{h}_{A1}(f)S_1(f,t) + \mathbf{h}_{A2}(f)S_2(f,t) + \mathbf{H}_{AN}(f)\mathbf{N}(f,t), \tag{2}$$

$$\mathbf{y}_B(f,t) = \mathbf{h}_{B1}(f)S_1(f,t) + \mathbf{h}_{B2}(f)S_2(f,t) + \mathbf{H}_{BN}(f)\mathbf{N}(f,t), \tag{3}$$

where $\mathbf{y}_A(f,t) = [y_A^{(1)}(f,t), \cdots, y_A^{(Q_A)}(f,t)]^T$, $\mathbf{h}_{A1}(f) = [h_{A1}^{(1)}(f), \ldots, h_{A1}^{(Q_A)}(f)]^T$, $\mathbf{h}_{A2}(f) = [h_{A2}^{(1)}(f), \ldots, h_{A2}^{(Q_A)}(f)]^T$, and $\mathbf{H}_{AN}(f)$ is a $[Q_A \times \mathcal{L}]$ matrix with elements defined by the acoustic transfer functions. The vectors $\mathbf{y}_B(f,t)$, $\mathbf{h}_{B1}(f)$, $\mathbf{h}_{B2}(f)$, and $\mathbf{H}_{BN}(f)$ are similarly defined.

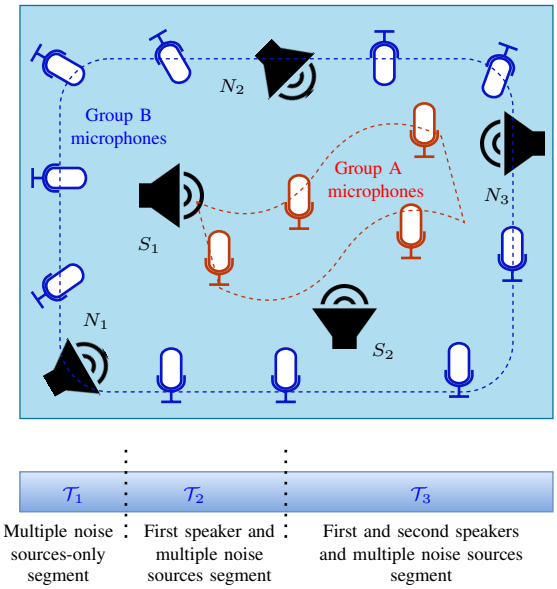


Fig. 1. Illustration of the problem setup.

The relative transfer matrix (ReTM) of noise sources, $\boldsymbol{\mathcal{R}}_{AB}^{(N)}(f)$, is defined as in [15]

$$\boldsymbol{\mathcal{R}}_{AB}^{(N)}(f) \triangleq \mathbf{H}_{AN}(f)\mathbf{H}_{BN}(f)^\dagger, \tag{4}$$

where $(\cdot)^\dagger$ is Moore-Penrose inverse, assuming the validity, i.e., $Q_B \geq \mathcal{L}$. Thus, at the first time segment $\mathcal{T}_1$, we can relate the received signal at group $\{A\}$ and $\{B\}$ using

$$\mathbf{y}_A^{(\mathcal{T}_1)}(f,t) = \boldsymbol{\mathcal{R}}_{AB}^{(N)}(f)\mathbf{y}_B^{(\mathcal{T}_1)}(f,t).$$

Note that ReTM is defined by the spatial properties of the sound sources such that it is independent of the sound source signals. In applications, the ReTM is constant for a stationary environment.

Assuming a stationary acoustic scenario, the next section shows how to estimate the ReTF of the successive speaker from the given multichannel recordings.

III. ReTF Estimation with ReTM

In this section, we propose a method for successive speaker ReTF estimation that utilizes some of the properties of the covariance matrices imposed by the ReTM.

A. Noise Sources ReTM

Here, we define the noise sources-only ReTM that will help for deriving our proposed method. Considering the background noise-only signals at $\mathcal{T}_1$ time segment, we can write the received signals as

$$\mathbf{y}_A^{(\mathcal{T}_1)} = \mathbf{H}_{AN}\mathbf{N}, \tag{5}$$

$$\mathbf{y}_B^{(\mathcal{T}_1)} = \mathbf{H}_{BN}\mathbf{N}. \tag{6}$$

The ReTM of the background noise sources can be estimated using the covariance matrices-based approach [15]. The background noise covariance matrices of microphone groups $\{A\}$ and $\{B\}$ are given as

$$\begin{aligned}\boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_1)}(f) &\triangleq E\{\mathbf{y}_A^{(\mathcal{T}_1)}(f,t)\mathbf{y}_A^{(\mathcal{T}_1)^*}(f,t)\},\\ \boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_1)}(f) &\triangleq E\{\mathbf{y}_B^{(\mathcal{T}_1)}(f,t)\mathbf{y}_A^{(\mathcal{T}_1)^*}(f,t)\},\end{aligned} \tag{7}$$

where $E\{\cdot\}$ denotes the expectation which can be obtained by averaging across the time frames, and $[\cdot]^*$ denotes the conjugate transpose. With further simplifications as in [15], we approximate

$$\boldsymbol{\mathcal{R}}_{AB}^{(N)}(f) \approx \boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_1)}(f)\left(\boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_1)}(f)\right)^{\dagger}. \tag{8}$$

For convenience, we omit the dependency of time (t) and frequency (f) in the rest of this section.

B. Successive Speaker ReTF using Noise ReTM

Following the ReTM, our proposed ReTF estimation method is also based on covariance matrices between two microphone groups. In the second time segment, $\mathcal{T}_2$, the received signals at microphone groups $\{A\}$ and $\{B\}$ are given as

$$\mathbf{y}_A^{(\mathcal{T}_2)} = [\mathbf{h}_{A1}\ \mathbf{H}_{AN}][S_1\ \mathbf{N}]^T, \tag{9}$$

$$\mathbf{y}_B^{(\mathcal{T}_2)} = [\mathbf{h}_{B1}\ \mathbf{H}_{BN}][S_1\ \mathbf{N}]^T. \tag{10}$$

When all sources are active in the third time segment, $\mathcal{T}_3$, we obtain the received signals at microphone groups $\{A\}$ and $\{B\}$ as

$$\mathbf{y}_A^{(\mathcal{T}_3)} = [\mathbf{h}_{A1}\ \mathbf{h}_{A2}\ \mathbf{H}_{AN}][S_1\ S_2\ \mathbf{N}]^T, \tag{11}$$

$$\mathbf{y}_B^{(\mathcal{T}_3)} = [\mathbf{h}_{B1}\ \mathbf{h}_{B2}\ \mathbf{H}_{BN}][S_1\ S_2\ \mathbf{N}]^T. \tag{12}$$

Assuming all signal components in both (11) and (12) to be uncorrelated, the noisy covariance matrices for time segment $\mathcal{T}_3$ of microphone groups $\{A\}$ and $\{B\}$, given as

$$\begin{aligned}&\boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_3)} \triangleq E\{\mathbf{y}_A^{(\mathcal{T}_3)}\mathbf{y}_A^{(\mathcal{T}_3)^H}\}\\ &= [\mathbf{h}_{A1}\ \mathbf{h}_{A2}\ \mathbf{H}_{AN}]\\ &\begin{bmatrix}E\{|S_1|^2\} & 0 & \mathbf{0}_{Q_A\times Q_A}\\ 0 & E\{|S_2|^2\} & \mathbf{0}_{Q_A\times Q_A}\\ \mathbf{0}_{Q_A\times Q_A} & \mathbf{0}_{Q_A\times Q_A} & \boldsymbol{\mathcal{P}}_{NN}\end{bmatrix}\begin{bmatrix}\mathbf{h}_{A1}^*\\ \mathbf{h}_{A2}^*\\ \mathbf{H}_{AN}^H\end{bmatrix}\\ &= \mathbf{h}_{A1}E\{|S_1|^2\}\mathbf{h}_{A1}^* + \mathbf{h}_{A2}E\{|S_2|^2\}\mathbf{h}_{A2}^* + \mathbf{H}_{AN}\boldsymbol{\mathcal{P}}_{NN}\mathbf{H}_{AN}^*\end{aligned} \tag{13}$$

where $\boldsymbol{\mathcal{P}}_{NN} \triangleq E\{\mathbf{N}\mathbf{N}^*\}$.

Similarly, we can write

$$\begin{aligned}\boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_3)} = \mathbf{h}_{B1}E\{|S_1|^2\}\mathbf{h}_{A1}^* + \mathbf{h}_{B2}E\{|S_2|^2\}\mathbf{h}_{A2}^*\\ +\mathbf{H}_{BN}\boldsymbol{\mathcal{P}}_{NN}\mathbf{H}_{AN}^*.\end{aligned} \tag{14}$$

Likewise, the covariance matrices for the time segment $\mathcal{T}_2$ of microphone groups $\{A\}$ and $\{B\}$, can be written as

$$\boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_2)} = \mathbf{h}_{A1}E\{|S_1|^2\}\mathbf{h}_{A1}^* + \mathbf{H}_{AN}\boldsymbol{\mathcal{P}}_{NN}\mathbf{H}_{AN}^*, \tag{15}$$

$$\boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_2)} = \mathbf{h}_{B1}E\{|S_1|^2\}\mathbf{h}_{A1}^* + \mathbf{H}_{BN}\boldsymbol{\mathcal{P}}_{NN}\mathbf{H}_{AN}^*. \tag{16}$$

We subtract (15) from (13) to remove both the first speaker and the background noise sources from the noisy mixture at group $\{A\}$. We then express the covariance matrix of the second speaker at group $\{A\}$ as

$$\begin{aligned}&\boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_3)} - \boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_2)} = \mathbf{h}_{A2}E\{|S_2|^2\}\mathbf{h}_{A2}^*,\\ &= E\{|S_2|^2\}\mathbf{h}_{A2}\mathbf{h}_{A2}^*,\\ &= E\{|S_2|^2\}\underbrace{\begin{bmatrix}h_{A2}^{(1)}h_{A2}^{(1)*} & \dots & h_{A2}^{(1)}h_{A2}^{(Q_A)*}\\ \vdots & \ddots & \vdots\\ h_{A2}^{(Q_A)}h_{A2}^{(Q_A)*} & \dots & h_{A2}^{(Q_A)}h_{A2}^{(Q_A)*}\end{bmatrix}}_{\mathbf{R}_{AA}},\end{aligned} \tag{17}$$

where $\mathbf{R}_{AA}$ is a $[Q_A \times Q_A]$ matrix of transfer function coefficients of the second speaker between channels at microphone group $\{A\}$.

Similarly, we subtract (16) from (14) as

$$\begin{aligned}&\boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_3)} - \boldsymbol{\mathcal{P}}_{BA}^{(\mathcal{T}_2)} = E\{|S_2|^2\}\mathbf{h}_{B2}\mathbf{h}_{A2}^*,\\ &= E\{|S_2|^2\}\underbrace{\begin{bmatrix}h_{B2}^{(1)}h_{A2}^{(1)*} & \dots & h_{B2}^{(1)}h_{A2}^{(Q_A)*}\\ \vdots & \ddots & \vdots\\ h_{B2}^{(Q_A)}h_{A2}^{(Q_A)*} & \dots & h_{B2}^{(Q_A)}h_{A2}^{(Q_A)*}\end{bmatrix}}_{\mathbf{R}_{BA}},\end{aligned} \tag{18}$$

where $\mathbf{R}_{BA}$ is a $[Q_B \times Q_A]$ matrix of transfer function coefficients of the second speaker between microphone groups $\{B\}$ and $\{A\}$.

Consider microphone channels n, and m at microphone group $\{A\}$ and $\{B\}$, respectively. We propose to divide $\{n, m\}$ elements of (17) by (18) to obtain ReTF $\{n, m\}$ pair with respect to the second speaker as

$$\begin{aligned}\hat{h}_2^{\{n,m\}} &= \frac{E\{|S_2|^2\}h_{A2}^{(n)}h_{A2}^{(n)*}}{E\{|S_2|^2\}h_{B2}^{(m)}h_{A2}^{(n)*}}\\ &= \frac{h_{A2}^{(n)}}{h_{B2}^{(m)}}.\end{aligned} \tag{19}$$

From (19), we obtain ReTF vector of the second speaker $\hat{\boldsymbol{h}}_2$ for $n \in 1, \dots, Q_A$. Therefore, the ReTF vector of the second speaker with respect to group $\{A\}$ as the reference channel at microphone group $\{B\}$ can be extracted from the covariance matrix addition.

Note that similarly, the ReTF of the first speaker, $\boldsymbol{h}_1$ can be obtained considering the covariance matrices at the time segments of $\mathcal{T}_1$ and $\mathcal{T}_2$ as in (17), and (18) and then followed by covariance matrices division in (19). However, in this paper, we cover neither the method nor the results of the ReTF of the first speaker from the noisy mixture.

IV. Experiments

This section presents experimental results for the proposed ReTF estimation algorithm using simulated recordings compared to the ground truth ReTFs.

A. Experiment Methodology

We utilize an open-source toolbox [25] to model the room impulse response (RIR) from the sound sources to irregularly distributed microphones in a $6 \times 7 \times 3$ m rectangular room with reverberation coefficients (β) of $\{0, 0.8, 1\}$. The ideal room with no reflection (free-space) would have $\beta = 0$, and $\beta = 1$ for the total reflection. We note that for a traditional implementation of a real room, the reverberation coefficients should be $0 < \beta < 1$. We consider two speech sources, two background noise sources, and 15 microphones. Two speaker locations are: Speaker 1: (3 m, 4.5 m, 1.2 m), Speaker 2: (3 m, 2.5 m, 1.2 m) and two background noise sources locations are: Source 1: (2 m, 0.9 m, 1.8 m), Source 2: (1 m, 6 m, 2.5 m) with respect to the origin position in the left corner of the room. We convolve the speech sources RIRs with both male (Speaker 1) and female (Speaker 2) speech utterances from the TIMIT dataset [26] and noise sources RIRs with vacuum noise, and music signal. The received signals are down-sampled to 16 kHz and ranged from 5 to -10 dB SNR of background noise and added with 40 dB SNR of white Gaussian noise at each microphone. Here, we calculate the background SNR by averaging SNR at each receiver over all 15 receivers. Note that the SNR is defined with respect to all sources in the mixture. The recordings are short-time-Fourier-transformed with an 8192-point window size. We assign $Q_A = 5$ and $Q_B = 10$ number of receivers to groups $\{A\}$ and $\{B\}$, respectively.

We use a distance measure based on the Hermitian angle [9] between the ground truth ReTF vector $\boldsymbol{h}_2$ and the estimated ReTF vector $\hat{\boldsymbol{h}}_2$ of the successive speaker as

$$\Theta(\mathbf{h}_2, \hat{\boldsymbol{h}}_2) = \arccos\Big(\frac{|\mathbf{h}_2^* \hat{\boldsymbol{h}}_2|}{||\mathbf{h}_2||_2 ||\hat{\boldsymbol{h}}_2||_2}\Big),$$

where $\arccos(\cdot)$ denotes inverse of the cosine function, $|\cdot|$ denotes the absolute value, and $\|\cdot\|_2$ denotes the ℓ_2 norm.

B. Results and Discussion

Fig. 2 shows the real and imaginary parts of the ground truth and estimated ReTFs of the successive speaker with respect to the second, and the first channels in the groups $\{A\}$ and $\{B\}$, respectively, for $\beta = 1$ at SNR level of 0 dB. We observe that many ReTF coefficients of both ground truth $\mathbf{h}_2$ and estimated $\hat{\boldsymbol{h}}_2$ are almost the same over the frequency up to 8 kHz. We discuss the results of these similarities between the ReTF vectors using the Hermitian angle in Table I.

Table I provides the distance measure results. We first observe that the proposed successive speaker ReTF estimation method is seen to have a lower Hermitian angle with the increase of the SNR level for all three reverberation coefficients. This suggests that the proposed method performs better at high SNR levels.

Second, we observe that the higher distance measure results in increased reverberation coefficients. Further, the Hermitian angle had a lower variation in the angles for all three reverberation coefficients, however, is seen to obtain higher variation for the lower SNR conditions. This suggests that the proposed method is able to accurately estimate the ReTF of the successive speaker from the noisy multichannel recording of two speakers' mixture.

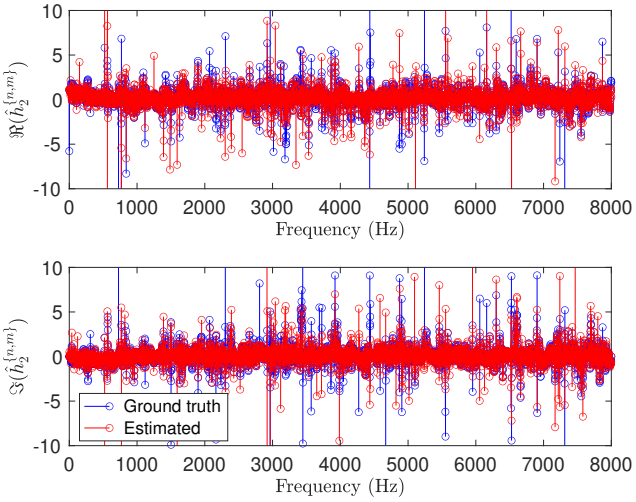


Fig. 2. The real and imaginary parts of the successive speaker ReTF obtained using the proposed method at microphone channel $n = 2, m = 1$ over the frequency at SNR level of 0 dB for $\beta = 1$.

TABLE I
DISTANCE MEASURE: HERMITIAN ANGLE FOR VARIOUS SNR LEVELS

SNR level	Reverberation coefficients		
	$\beta = 0$	$\beta = 0.8$	$\beta = 1$
5 dB	6.75°	8.95°	9.21°
0 dB	8.06°	11.94°	13.35°
-5 dB	10.55°	17.65°	17.86°
-10 dB	16.86°	27.68°	26.33°

V. Applications in Speech Enhancement

In this section, we employ the minimum variance distortionless response (MVDR) beamformer for noise suppression and examine the performance in terms of the SNR improvement to evaluate the estimated ReTF vectors of the successive speaker in Section IV.

The MVDR beamformer requires an estimate of the noise-only covariance matrix that is equivalent to the received signals at $\mathcal{T}_1$ time segment in (7), and the estimate of the ReTFs of the second speaker, $\hat{\boldsymbol{h}}_2$. The weights of the MVDR beamformer can be expressed as in [27]

$$\mathbf{w} = \frac{\boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_1)^{-1}} \hat{\boldsymbol{h}}_2}{\hat{\boldsymbol{h}}_2^* \boldsymbol{\mathcal{P}}_{AA}^{(\mathcal{T}_1)^{-1}} \hat{\boldsymbol{h}}_2}, \tag{20}$$

where $(\cdot)^{-1}$ denotes the inversion operator, and $\mathbf{w}$ is the Q_A by 1 filter vector for a given frequency bin. The beamformer output is given by [28]

$$z(f,t) = \mathbf{w}^*(f)\mathbf{y}_A(f,t). \tag{21}$$

TABLE II
SNR IMPROVEMENT, ΔSNR FOR VARIOUS SNR LEVELS

SNR level	Reverberation coefficients		
	$\beta = 0$	$\beta = 0.8$	$\beta = 1$
5 dB	15.28 dB	12.16 dB	12.51 dB
0 dB	15.02 dB	15.42 dB	16.46 dB
-5 dB	14.46 dB	18.72 dB	18.88 dB
-10 dB	14.57 dB	21.64 dB	21.55 dB

We define the SNR improvement between the input and output of the beamformer as

$$\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in,avg}},$$

where SNR_{out} is the filtered output SNR, and $\text{SNR}_{\text{in,avg}}$ is the average input SNR among all microphones in group $\{A\}$.

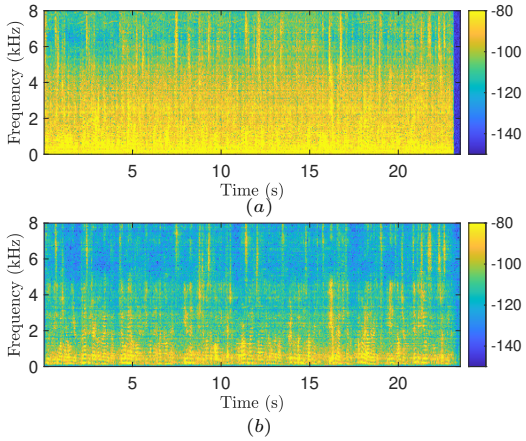


Fig. 3. Spectrogram plots of (a) microphone recordings at microphone channel 1 of Q_A, and (b) the extracted speech signals of the second speaker at 0 dB SNR level for $\beta = 0.8$ ($T_{60} = 500$ ms).

The spectrogram plots of the mixture and desired speech signals at the SNR level of 0 dB for $\beta = 0.8$ are shown in Fig. 3. Immediately we observe that the noise is removed and the voice of speaker 2 is shown to be successfully extracted at the beamformer output. We share the audio files of both input and output at the beamformer for $\beta = 0.8$ at all SNR levels[1].

Table II depicts the SNR improvement at the MVDR beamformer for the proposed ReTF estimation method as a function of the reverberation coefficient for $\beta = \{0, 0.8, 1\}$ over various SNR levels. We observe that in both reverberant environments, the SNR improvements are gradually increasing with decreased SNR levels. The SNR improvement at the beamformer obtains almost the same value with increased reverberation. A similar result with slight degradation is obtained for free-field scenario with an increased SNR level. As expected, the beamformer's output derived using the proposed ReTF estimation method is seen to be enhanced the successive speaker from the noisy mixture, illustrated by the high SNR improvement of 12 dB or above in both free-space and reverberant environments. The results confirm that the proposed ReTF method accurately enhances the voice of the second speaker in dual-speaker noisy reverberant rooms.

[1]https://github.com/wnilmini/Successive_Speaker_ReTF_Estimation

VI. CONCLUSION

We have proposed a ReTF estimation method for the second speaker in noisy and reverberant environments with two speech sources and multiple noise sources where the speech sources activate successively. The method utilizes the properties of the covariance matrices between two multichannel microphone groups imposed by the ReTM. Simulation analysis showed that the proposed method achieves a low Hermitian angle in both free-space and reverberant environments with differing SNR level. The MVDR beamformer which derives via the estimated ReTFs, is seen to have enhanced the voice of the successive speaker from the noisy microphone recordings with decreasing SNR levels. The accuracy of the ReTF estimation was increased within low reverberant environments, but SNR improvement was shown to improve with increasing reflection coefficient. In the future, we plan to extend this algorithm to the more challenging multiple-speaker scenarios exploiting the ReTM.

REFERENCES

[1] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[2] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 17, no. 4, pp. 546–555, Mar. 2009.

[3] L. Birnie, P. Samarasinghe, T. Abhayapala, and D. Grixti-Cheng, "Noise retf estimation and removal for low snr speech enhancement," in *'IEEE Workshop Mach. Learning Signal Process.'*, Oct. 2021, pp. 1–6.

[4] W. Manamperi, P. N. Samarasinghe, T. D. Abhayapala, and J. Zhang, "GMM based multi-stage Wiener filtering for low SNR speech enhancement," in *Proc. IEEE Int. Workshop on Acoust. Signal Enhancement*, Sep. 2022, pp. 1–5.

[5] E. Warsitz, A. Krueger, and R. Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Mar. 2008, pp. 73–76.

[6] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. on Signal Process.*, vol. 44, no. 8, pp. 2055–2063, Aug. 1996.

[7] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. on Speech and Audio Process.*, vol. 12, no. 5, pp. 451–459, Aug. 2004.

[8] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 22, no. 4, pp. 785–799, Feb. 2014.

[9] R. Varzandeh, M. Taseska, and E. A. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Joint Workshop Hands-free Speech Comm. and Microphone Arrays*, Mar. 2017, pp. 11–15.

[10] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Jun. 2009.

[11] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jun. 2007.

[12] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *Proc. Eur. Signal Process. Conf.*, Sep. 2018, pp. 2499–2503.

[13] A. Brendel, J. Zeitler, and W. Kellermann, "Manifold learning-supported estimation of relative transfer functions for spatial filtering," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2022, pp. 8792–8796.

[14] A. Deleforge, S. Gannot, and W. Kellermann, "Towards a generalization of relative transfer functions to more than one source," in *Proc. Eur. Signal Process. Conf.*, Aug. 2015, pp. 419–423.

[15] T. D. Abhayapala, L. Birnie, M. Kumar, D. Grixti-Cheng, and P. N. Samarasinghe, "Generalizing the relative transfer function to a matrix for multiple sources and multichannel microphones," in *Proc. Eur. Signal Process. Conf.*, Sep. 2023, pp. 336–340.

[16] M. Kumar and et al., "Speech denoising in multi-noise source environments using multiple microphone devices via relative transfer matrix," in *Eur. Signal Process. Conf. (in press)*, Sep. 2024, pp. 336–340.

[17] W. N. Manamperi and T. D. Abhayapala, "Relative transfer matrix for drone audition applications: Source enhancement," in *Asia-Pacific Signal and Inf. Process. Assoc. Annu. Summit and Conf. (in press)*, Dec. 2024.

[18] W. N. Manamperi and T. D. Abhayapala, "Blind separation of multiple speakers in noisy reverberant environments using relative transfer matrix," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (unpublished)*, 2024.

[19] B. Schwartz, S. Gannot, and E. A. Habets, "Two model-based EM algorithms for blind source separation in noisy environments," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 25, no. 11, pp. 2209–2222, Aug. 2017.

[20] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 27, no. 7, pp. 1136–1150, Apr. 2019.

[21] C. Li, J. Martinez, and R. C. Hendriks, "Low complex accurate multi-source RTF estimation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, May 2022, pp. 4953–4957.

[22] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early psd estimation and recursive RETF update in reverberant environments by means of the orthogonal procrustes problem," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 28, pp. 755–769, Jan. 2020.

[23] D. Cherkassky and S. Gannot, "Successive relative transfer function identification using blind oblique projection," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 28, pp. 474–486, Dec. 2019.

[24] H. Gode and S. Doclo, "Covariance blocking and whitening method for successive relative transfer function vector estimation in multi-speaker scenarios," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2023, pp. 1–5.

[25] E. A. Habets, *Room impulse response (RIR) generator*, [Online]. Available: https://www.audiolabserlangen.de/fau/professor/habets/software/rir-generator, 2006.

[26] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[27] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[28] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 25, no. 4, pp. 692–730, Jan. 2017.

Heavy-tailed Distributions-Based Online Semi-blind Source Separation for Nonlinear Echo Cancellation

Liyuan Zhang*, Xianrui Wang*†, Yichen Yang*†, Tetsuya Ueda*, Shoji Makino* and Jingdong Chen†
* Waseda University, Japan
† Northwestern Polytechnical University, Xi'an, China

***Abstract*—Recently proposed semi-blind source separation (SBSS) based acoustic echo cancellation (AEC) algorithms have attracted significant research interest due to their ability to track dynamic acoustic environments in the presence of near-end signals. Two source models are considered in these existing algorithms, i.e., the spherical generalized super-Gaussian distribution and the circular super-Gaussian distribution with a low-rank spectrogram model. In this paper, we aim to further enhance AEC performance by leveraging more flexible source models. Several novel algorithms are subsequently proposed. Simulations demonstrate the superiority of proposed algorithms in various situations.**

I. Introduction

Acoustic echo cancellation (AEC) plays an indispensable role in real-time full-duplex communication [1]–[3]. The performance of conventional adaptive filters based algorithms generally suffers dramatic degradation in double-talk situations, i.e., when far-end signal and near-end signal coexist [1], [4], [5]. To address this problem, researchers reformulated AEC as a semi-blind source separation (SBSS) problem [6], [7] and subsequently derived several SBSS-AEC algorithms [8]–[13].

Existing SBSS-AEC algorithms can be roughly divided into two categories based on the source model assumption: the auxiliary function-based independent vector analysis (AuxIVA) [14] which considers spherical super-Gaussian distribution [10], [11], and the Itakura-Saito divergence-based independent low-rank matrix analysis (IS-ILRMA) [15], [16], which considers circular super-Gaussian distribution and non-negative matrix factorization (NMF) spectrogram model [9].

Although these algorithms have achieved satisfying AEC performance, there is still room for further improvement by considering more generalized and flexible source models. In this paper, we adopt two heavy-tailed distributions: the complex generalized super-Gaussian distribution (GGD) and Student's t-distribution. Subsequently, we derive three novel algorithms, namely, GGD-ILRMA, t-AuxIVA and t-ILRMA based SBSS-AEC. To update NMF parameters in a real-time manner, we apply recursive approximation and semi-supervised NMF (SSNMF) techniques [17], [18]. The performance of the proposed algorithms is studied under complex real-time circumstances. Several simulations validate the efficacy and superiority of the proposed algorithms.

The remainder of this paper is organized as follows: In Section II, a bilinear signal model and the problem formulation of SBSS-AEC are introduced. Next, on the basis of conventional GGD-AuxIVA and IS-ILRMA based SBSS-AEC [9]–[11] described in Section III-A, we propose GGD-ILRMA-based SBSS-AEC algorithm utilizing alternating iterative projection (AIP) as update rules and recursive approximation for online implementation in Section III-B. In Section III-C, all the mentioned GGD-based SBSS-AEC algorithms are extended to their heavy-tailed counterparts with Student's t-distributions, where t-AuxIVA and t-ILRMA based SBSS-AEC algorithms are proposed. The feasibility of SSNMF is discussed in Section III-D. Simulations and results considering two scenarios, i.e., speech and music, are shown and analyzed in Section IV.

II. Signal Model and Problem Formulation

Considering a full-duplex communication scenario, a loudspeaker with unknown non-linearity is used to play the far-end signal. The output of the loudspeaker is convolved with acoustic impulse response (AIR) to generate the echo. The superimposition of the near-end signal and the echo is captured by the microphone and then transmitted back to the far end. We use odd power series expansion technique to approximate the non-linearity of the loudspeaker as described in [8], [19], [20]. In double-talk situation, the observed microphone signal in short-time Fourier transform (STFT) domain can be denoted as [9]–[11]

$$\begin{aligned} Y_{i,j} &= E_{i,j} + S_{i,j} \\ &= \sum_{n=1}^{N} \sum_{l=1}^{L} a_n H_{i,j,l} X_{n,i,j-l+1} + S_{i,j}, \end{aligned} \tag{1}$$

where $i = 1, \ldots, I$ and $j = 1, \ldots, J$ are frequency-bin and time-frame indices, respectively, I and J are the number of frequency bins and time frames, respectively, N is the expansion order, a_n is the n-th order odd power series expansion coefficient, L is the convolutive transfer function (CTF) filter length, $H_{i,j,l}$ is the CTF filter coefficient [9], [21], and $X_{n,i,j}$, $E_{i,j}$ and $S_{i,j}$ are the STFTs of nth-order far-end signal, nonlinear echo and near-end signal, respectively.

The nonlinear echo can be reformulated as a bilinear form [10]

$$E_{i,j} = \mathbf{h}_{i,j}^T \mathbf{X}_{i,j} \mathbf{a}, \tag{2}$$

where

$$\mathbf{h}_{i,j} = [H_{i,j,1} \quad H_{i,j,2} \quad \cdots \quad H_{i,j,L}]^T, \tag{3}$$

$$\mathbf{a} = [a_1 \quad a_2 \quad \cdots \quad a_N]^T, \tag{4}$$

$$\mathbf{X}_{i,j} = [\tilde{\mathbf{x}}_{i,j} \quad \tilde{\mathbf{x}}_{i,j-1} \quad \cdots \quad \tilde{\mathbf{x}}_{i,j-L+1}]^T, \tag{5}$$

$$\tilde{\mathbf{x}}_{i,j} = [X_{1,i,j} \quad X_{2,i,j} \quad \cdots \quad X_{N,i,j}]^T, \tag{6}$$

$\mathbf{X}_{i,j} \in \mathbb{C}^{L\times N}$ is a matrix constructed by far-end signal series $X_{n,i,j}$ and $(\cdot)^T$ denotes the transpose operation. Using (2), the estimation of a_n and $H_{i,j,l}$ can be interpreted as two sub-separate SBSS problems. After defining a new far-end reference vector

$$\mathbf{x}_{i,j} = \mathbf{X}_{i,j}\mathbf{a}, \tag{7}$$

(2) is rewritten as

$$E_{i,j} = \mathbf{h}_{i,j}^T \mathbf{x}_{i,j}. \tag{8}$$

Then, the mixing process in (1) can be rewritten with a vector form

$$\tilde{\mathbf{y}}_{i,j} = \mathbf{H}_{i,j}\tilde{\mathbf{s}}_{i,j}, \tag{9}$$

with

$$\tilde{\mathbf{y}}_{i,j} = \begin{bmatrix} Y_{i,j} & \mathbf{x}_{i,j}^T \end{bmatrix}^T, \tag{10}$$

$$\tilde{\mathbf{s}}_{i,j} = \begin{bmatrix} S_{i,j} & \mathbf{x}_{i,j}^T \end{bmatrix}^T, \tag{11}$$

$$\mathbf{H}_{i,j} = \begin{bmatrix} 1 & \mathbf{h}_{i,j}^T \\ \mathbf{0}_{L\times 1} & \mathbf{I}_L \end{bmatrix}, \tag{12}$$

where $\mathbf{H}_{i,j} \in \mathbb{C}^{(L+1)\times(L+1)}$ is the mixing matrix, $\mathbf{0}_{L\times 1}$ is a column vector of length L with all elements equal to 0, $\mathbf{I}_L$ is an identity matrix of size $L \times L$. Assuming that $\mathbf{H}_{i,j}$ is non-singular, to extract the near-end signal, the demixing matrix $\hat{\mathbf{W}}_{i,j} \in \mathbb{C}^{(L+1)\times(L+1)}$ is obtained as the inverse of $\mathbf{H}_{i,j}$, which is

$$\hat{\mathbf{W}}_{i,j} = \begin{bmatrix} 1 & -\hat{\mathbf{h}}_{i,j}^T \\ \mathbf{0}_{L\times 1} & \mathbf{I}_L \end{bmatrix}, \tag{13}$$

where $\hat{\mathbf{h}}_{i,j}^T$ is a column vector with L parameters to be estimated. With this structure, the near-end signal is extracted as

$$\hat{S}_{i,j} = \hat{\mathbf{w}}_{i,j}^H \tilde{\mathbf{y}}_{i,j}, \tag{14}$$

where the extraction filter $\hat{\mathbf{w}}_{i,j} = [1 \quad -\hat{\mathbf{h}}_{i,j}^T]^H$ is the first row of (13) and $(\cdot)^H$ denotes the Hermitan transpose. Therefore, the target of SBSS-AEC in double-talk situation is transformed into estimating $\hat{\mathbf{w}}_{i,j}$ to extract the near-end signal. Note that (2) can also be written as $E_{i,j} = \mathbf{a}^T \mathbf{x}_{i,j}^\star$, where $\mathbf{x}_{i,j}^\star = \mathbf{X}_{i,j}^T \mathbf{h}_{i,j}$. Similar derivation as (9)–(14) is omitted for simplification [10].

III. Generalization of Online SBSS-AEC Algorithms

A. Conventional method

In SBSS-AEC, the near-end source can be extracted by exploiting mutual independence between near-end and reference signals. Therefore, since the determinant of (13) equals one, the following recursive negative log-likelihood function is derived [11], [17]

$$\mathcal{L}_j = -\frac{1}{\sum_{j'=1}^{j}(\eta)^{j-j'}} \sum_{j'=1}^{j} (\eta)^{j-j'} \log p\left(\mathbf{s}_{j'}\right), \tag{15}$$

where $\eta \in (0,1)$ is a forgetting factor and

$$\mathbf{s}_j = [S_{1,j} \quad S_{2,j} \quad \ldots \quad S_{I,j}]^T. \tag{16}$$

In GGD-AuxIVA-based SBSS-AEC, the near-end signal is modeled with a spherical complex GGD [9]–[11], [22]

$$p_{\mathrm{GGD}}(\mathbf{s}_j) \propto \exp\left[-\left(\frac{\|\mathbf{s}_j\|_2}{\gamma}\right)^{\beta}\right], \tag{17}$$

where $\|\cdot\|_2$ stands for ℓ_2 norm, γ and β stand for positive scale and shape parameters, respectively. We assume that $0 < \beta \leq 2$ to satisfy the precondition of majorization-minimization (MM) method [22], [23].

Here, a_n and $H_{i,j,l}$ are updated alternatingly. When we fix a_n, the following cost function for CTF coefficients $\hat{\mathbf{h}}_{i,j}$ can be obtained using MM method

$$\mathcal{L}_j^{\mathbf{h},+} = \sum_{i=1}^{I} \hat{\mathbf{w}}_{i,j}^H \mathbf{V}_{i,j} \hat{\mathbf{w}}_{i,j}. \tag{18}$$

To further reduce computational cost, we employ AIP update rules [10], where the auxiliary matrix $\mathbf{V}_{i,j} \in \mathbb{C}^{(L+1)\times(L+1)}$ is partitioned as

$$\mathbf{V}_{i,j} = \begin{bmatrix} \sigma_{y,i,j}^2 & \mathbf{g}_{i,j}^H \\ \mathbf{g}_{i,j} & \mathbf{C}_{i,j} \end{bmatrix}, \tag{19}$$

$\mathbf{g}_{i,j}$ and $\mathbf{C}_{i,j}$ are recursively updated based on (18)

$$\sigma_{y,i,j}^2 = \eta\sigma_{y,i,j-1}^2 + (1-\eta)\varphi_{\mathrm{GGD}}(\sigma_{s,j})\left|Y_{i,j}\right|^2, \tag{20}$$

$$\mathbf{g}_{i,j} = \eta\mathbf{g}_{i,j-1} + (1-\eta)\varphi_{\mathrm{GGD}}(\sigma_{s,j}) Y_{i,j}^* \mathbf{x}_{i,j}, \tag{21}$$

$$\mathbf{C}_{i,j} = \eta\mathbf{C}_{i,j-1} + (1-\eta)\varphi_{\mathrm{GGD}}(\sigma_{s,j}) \mathbf{x}_{i,j}\mathbf{x}_{i,j}^H, \tag{22}$$

where $(\cdot)^*$ denotes conjucate operation, $\varphi_{\mathrm{GGD}}(\sigma_{s,j})$ is the score function

$$\varphi_{\mathrm{GGD}}(\sigma_{s,j}) = \sigma_{s,j}^{\beta-2}, \tag{23}$$

and $\sigma_{s,j} = \|\hat{\mathbf{s}}_j\|_2$ is the auxiliary variable. Next, the CTF filter is updated as [10]

$$\hat{\mathbf{h}}_{i,j} = \left(\mathbf{C}_{i,j}^{-1}\mathbf{g}_{i,j}\right)^*. \tag{24}$$

When we fix $H_{i,j,l}$, the update of expansion coefficient $\hat{\mathbf{a}}$ can be derived similarly as (18)–(24) [10]. After updating all the parameters, the near-end signal can be extracted using (14).

While GGD-AuxIVA-based SBSS-AEC is effective in most scenarios, as shown in (17), the multivariate source model of AuxIVA is inflexible and cannot cope with specific harmonic structures of each source due to higher-order correlations between frequency bins [24]. As another state-of-the-art BSS algorithm, ILRMA is more generalized and effective in modeling sources by using the low-rank time-frequency NMF structure, especially in music scenarios [15]. Nevertheless, the ILRMA-based SBSS-AEC method in [9] assumes IS-NMF as the generative model, the probabilistic distribution of which is deficient compared with GGD-NMF. Therefore, in this paper, we first improve the performance of GGD-AuxIVA-based and IS-ILRMA-based SBSS-AEC by using GGD-NMF as the source generative model and propose a new GGD-

ILRMA based SBSS-AEC algorithm.

B. Proposed GGD-ILRMA based SBSS-AEC

In GGD-ILRMA-based SBSS-AEC, the source generative model follows a time-frequency-wise isotropic complex GGD

$$p_{\mathrm{GGD}}(S_{i,j}) = \frac{\beta}{2\pi r_{i,j}^2 \Gamma(\frac{2}{\beta})} \exp\left[-\left(\frac{|S_{i,j}|}{r_{i,j}} \right)^{\beta} \right], \tag{25}$$

$$r_{i,j}^{p} = \sum_{k=1}^{K} t_{i,k} v_{k,j}, \tag{26}$$

where $t_{i,k} \geq 0$ and $v_{k,j} \geq 0$ are the nonnegative basis and activation elements of the basis matrix $T_n \in \mathbb{R}_{\geq 0}^{I \times K}$ and the activation matrix $V_n \in \mathbb{R}_{\geq 0}^{K \times J}$, repectively, $k = 1, \ldots, K$ is the integral index of the basis, K is the number of NMF bases, $r_{i,j}$ is a time-frequency-varying variance corresponding to the low-rank source model, $p > 0$ is the parameter that defines NMF domain and $\Gamma(\cdot)$ is the Gamma function [25]. Using MM method, the score function is derived as

$$\varphi_{\mathrm{GGD}}(r_{i,j}) = \frac{\beta}{2} \left| \hat{S}_{i,j} \right|^{\beta-2} r_{i,j}^{-\beta}. \tag{27}$$

In ILRMA, the update rules for $\hat{\mathbf{w}}_{i,j}$ are equivalent to those in AIP-based AuxIVA [10], [15]. Regarding the source model, we adopt the update rules in literature [25] and adapt the update of basis elements from offline to online scenarios using recursive approximation [9], [17]. Information from the previous frames is utilized to update the scaling factor of the basis elements in the current frame:

$$t_{i,k}^{\mathrm{GGD}} \leftarrow t_{i,k}^{\mathrm{GGD}} \left[\frac{\beta \sum_{j'=1}^{j} \alpha_{\mathrm{GGD}}^{j-j'} \left| \hat{S}_{i,j'} \right|^{\beta} r_{i,j'}^{-(p+\beta)} v_{k,j'}^{\mathrm{GGD}}}{2 \sum_{j'=1}^{j} \alpha_{\mathrm{GGD}}^{j-j'} r_{i,j'}^{-p} v_{k,j'}^{\mathrm{GGD}}} \right]^{\frac{p}{p+\beta}}, \tag{28}$$

$$v_{k,j}^{\mathrm{GGD}} \leftarrow v_{k,j}^{\mathrm{GGD}} \left[\frac{\beta \sum_{i=1}^{I} \left| \hat{S}_{i,j} \right|^{\beta} r_{i,j}^{-(p+\beta)} t_{i,k}^{\mathrm{GGD}}}{2 \sum_{i=1}^{I} r_{i,j}^{-p} t_{i,k}^{\mathrm{GGD}}} \right]^{\frac{p}{p+\beta}}, \tag{29}$$

where $\alpha_{\mathrm{GGD}} \in (0, 1)$ is a forgetting factor for NMF variables. With an appropriate setup of α_{GGD}, the convergence speed of the NMF model can be increased during the early stages.

C. Proposed Student's t-distribution-based SBSS-AEC

Student's t-distribution is also a popular heavy-tailed distribution model, which is effective in modeling audio sources due to its heavier tail and controllable degrees of freedom parameter [26]–[28]. In this paper, we extend GGD-AuxIVA-based and GGD-ILRMA-based SBSS-AEC using Student's t-distributions, and propose t-AuxIVA-based and t-ILRMA-based SBSS-AEC.

For t-AuxIVA-based SBSS-AEC, we assume a spherical complex Student's t-distribution with zero-mean as p.d.f. of the near-end signal [26], [27]:

$$p_{\mathrm{t}}(\mathbf{s}_j) \propto \left(1 + \frac{2}{v} \frac{\|\mathbf{s}_j\|_2^2}{\gamma_j} \right)^{-\frac{I+v}{2}}, \tag{30}$$

where $v > 0$ is the degrees of freedom parameter and γ_j is the uniform variance over frequency bins in the j-th time frame. Using (15) and (30), the following score function is derived:

$$\varphi_{\mathrm{t}}(\sigma_{s,j}) = \left(1 + \frac{I}{v} \right) \left(\frac{\gamma_j}{2} + \frac{\sigma_{s,j}^2}{v} \right)^{-1}. \tag{31}$$

Since the cost function is identical to (18) when we fix a_n, $\hat{\mathbf{w}}_{i,j}$ can be updated with the AIP algorithm (19)–(22).

As to t-ILRMA-based SBSS-AEC, the isotropic complex Student's t-distribution is assumed as the following source generative model:

$$p_{\mathrm{t}}(S_{i,j}) = \frac{1}{\pi r_{i,j}^2} \left(1 + \frac{2}{v} \frac{|S_{i,j}|^2}{r_{i,j}^2} \right)^{-\frac{2+v}{2}} \tag{32}$$

where $r_{i,j}$ is defined as (26). Using MM algorithm, the score function is derived as:

$$\varphi_{\mathrm{t}}(r_{i,j}) = \left(1 + \frac{2}{v} \right) \left(r_{i,j}^2 + \frac{2 \left| \hat{S}_{i,j} \right|^2}{v} \right)^{-1}. \tag{33}$$

Using (33) and AIP, $\hat{\mathbf{w}}_{i,j}$ is updated in a manner similar to GGD-ILRMA-based SBSS AEC. Analogously, the update rules of NMF matrices are derived using the recursive approximation technique [17], [25], [28]:

$$t_{i,k}^{\mathrm{t}} \leftarrow t_{i,k}^{\mathrm{t}} \left[\frac{\sum_{j'=1}^{j} \alpha_{\mathrm{t}}^{j-j'} \left| \hat{S}_{i,j'} \right|^2 c_{i,j'}^{\mathrm{t}} r_{i,j'}^{-p} v_{k,j'}^{\mathrm{t}}}{\sum_{j'=1}^{j} \alpha_{\mathrm{t}}^{j-j'} r_{i,j'}^{-p} v_{k,j'}^{\mathrm{t}}} \right]^{\frac{p}{p+2}}, \tag{34}$$

$$v_{k,j}^{\mathrm{t}} \leftarrow v_{k,j}^{\mathrm{t}} \left[\frac{\sum_{i=1}^{I} \left| \hat{S}_{i,j} \right|^2 c_{i,j'}^{\mathrm{t}} r_{i,j}^{-p} t_{i,k}^{\mathrm{t}}}{\sum_{i=1}^{I} r_{i,j}^{-p} t_{i,k}^{\mathrm{t}}} \right]^{\frac{p}{p+2}}, \tag{35}$$

where

$$c_{i,j}^{\mathrm{t}} = \left(\frac{v}{v+2} r_{i,j}^2 + \frac{2}{v+2} \left| \hat{S}_{i,j} \right|^2 \right)^{-1} \tag{36}$$

and $\alpha_{\mathrm{t}} \in (0, 1)$ is a forgetting factor for NMF variables.

D. Semi-Supervised NMF (SSNMF)

In AEC, little attention is paid to online ILRMA-based SBSS algorithms due to the challenges of implementing NMF in real-time [15]. Nonetheless, in AEC applications, it is often feasible to obtain a registered near-end voice dataset of a certain length beforehand. Therefore, the SSNMF [18] technique can address this issue. In SSNMF, the basis matrix is pre-trained on the registered speech dataset before real-time NMF processing. During online processing, this basis matrix is fixed, and only the time-dependent activation matrix is updated. This approach alleviates the clustering problem typically associated with NMF, as there is only one source and one reference. The SSNMF technique offers a promising alternative to the previously mentioned recursive approximation technique.

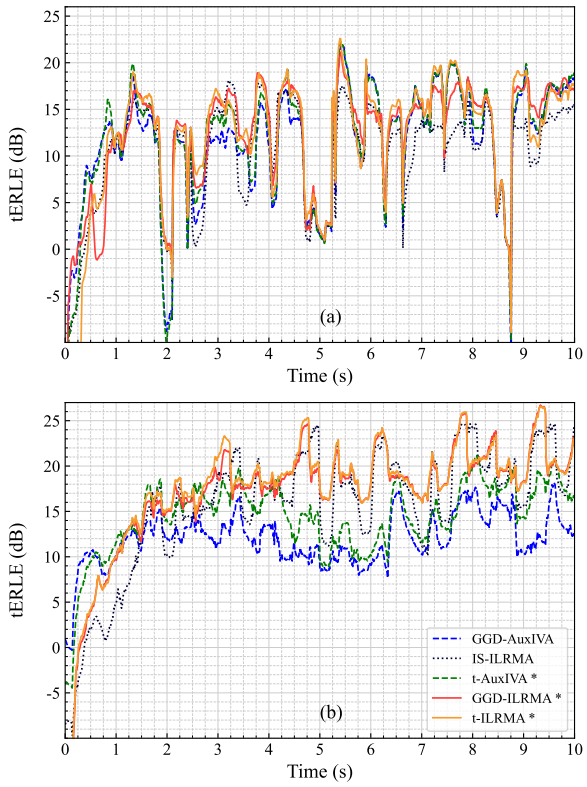


Fig. 1. Comparison of tERLE performance in different double-talk cases. (a) Speech (b) Music.

TABLE I
PARAMETER SETUP

Scenario	Speech				Music			
Algorithm	β	v	p	η	β	v	p	η
IS-ILRMA	-	-	-	0.992	-	-	-	0.990
GGD-AuxIVA	0.4	-	-	0.970	1	-	-	0.970
GGD-ILRMA	0.4	-	1	0.985	1	-	2	0.980
t-AuxIVA	-	10	-	0.970	-	30	-	0.970
t-ILRMA	-	10	1	0.985	-	30	2	0.980

IV. SIMULATIONS AND RESULTS

A. Experimental Setup

To maintain consistency, we use the same AIR setup as in [10], where the reverberation time, T_{60}, is approximately 300 ms. We consider the hard-clipping function [9], [20] to simulate loudspeaker distortions. We prepare two double-talk scenarios using speech and music signals, respectively. The signal-to-echo ratio (SER) is set to be 0 dB. In double-talk speech scenario, two 10-second-long male speech signals from the CMU ARCTIC dataset [29] are selected as the far-end and near-end signals, respectively. For SSNMF, we select random near-end registered speech signals to generate a 40-second pre-trained dataset. During the pre-training process, basis matrices are trained under 30 iterations with offline ILMRA algorithms. In double-talk music scenario, we use the same music signals as in the paper [9].

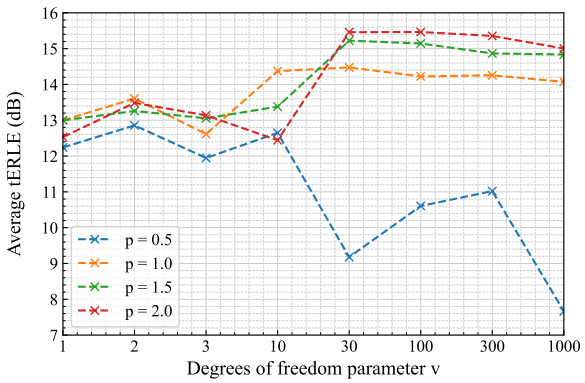


Fig. 2. Average tERLE of 10-group double-talk Speech cases based on t-ILRMA-based SBSS-AEC

TABLE II
DOUBLE-TALK SPEECH SOUND QUALITY

Algorithms	PESQ	STOI
IS-ILRMA	1.502	0.921
GGD-AuxIVA	1.637	0.954
GGD-ILRMA (proposed)	1.679	0.953
t-AuxIVA (proposed)	1.635	0.955
t-ILRMA (proposed)	**1.792**	**0.960**

The sampling rate of all signals is 16 kHz. In short-time analysis, a von Hann window with a length of 1024 samples and a 75% overlap between consecutive frames is used. The nonlinear series expansion order N and CTF filter length L are set to 3 and 5, respectively. The number of basis matrices K is set to 8 for speech signals and 4 for music signals. The AIP initialization setups are consistent to that in [10]. The forgetting factors of the NMF basis matrix, α_{GGD} and α_{t}, are set below 0.02. The setup of the other parameters is shown in Table I.

In this section, we compare the AEC performance of proposed heavy-tailed distribution-based SBSS-AEC algorithms with IS-ILRMA-based and GGD-AuxIVA-based SBSS-AEC [9], [10]. We use true echo return loss enhancement (tERLE) [8] as performance metric. Additionally, perceptual evaluation of speech quality (PESQ) [30] and short time objective intelligibility (STOI) [31] are used as performance evaluation.

B. Simulation Results

First, we compare the performance in double-talk speech scenario. Fig. 1 (a) shows the tERLE performance of AuxIVA-based, SSNMF-ILRMA-based with GGD or Student's t-distribution, and IS-ILRMA-based online SBSS-AEC algorithms. It is evident that Student's t-distribution-based algorithms achieve comparable performance, and SSNMF-ILRMA-based methods outperform AuxIVA-based methods in the last nine seconds. All heavy-tailed distributions based algorithms

perform better than IS-ILRMA based counterpart under optimal parameter settings, demonstrating robustness and superior performance.

To illustrate the flexibility of heavy-tailed distribution-based algorithms more clearly, we present an example of parameter adjustment for t-ILRMA SBSS-AEC for 10-group double-talk speech cases from the CMU ARCTIC dataset. Figure 2 shows that when NMF domain parameter p is set to 2 and the degrees of freedom parameter v is set to 30, relevant algorithm can achieve best results. These curves highlight the process of modifying the source model. Table II shows the average PESQ and STOI performance of 10 double-talk speech cases with optimal parameter settings where proposed methods achieve better or comparable sound quality compared to baseline methods.

In double-talk music scenario, as displayed in Fig. 1 (b), it is quite obvious that ILRMA-based SBSS-AEC outperforms AuxIVA-based algorithms. Besides, all the proposed online ILRMA-based SBSS-AEC algorithms converge much faster than IS-ILRMA-based algorithms in the fisrt two seconds. This case demonstrates the superiority of ILRMA-based online SBSS-AEC algorithms in complex situations.

V. Conclusions

To further enhance the performance of SBSS-AEC algorithms, in this paper, we adopted two heavy-tailed distributions, i.e., the complex generalized super-Gaussian distribution (GGD) and Student's t-distribution. Based on them, we derived three novel algorithms, namely, GGD-ILRMA, t-AuxIVA and t-ILRMA based SBSS-AEC. Besides, online update rules for NMF were introduced. Simulations validated the effectiveness and superiority of proposed algorithms.

Acknowledgment

This work was supported by JSPS KAKENHI 23H03423.

References

[1] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in network and acoustic echo cancellation*. Springer, 2001.

[2] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1988, pp. 2570–2573.

[3] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: A practical approach*. John Wiley & Sons, 2005.

[4] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, no. 6, pp. 1140–1156, 2006.

[5] T. Gansler, S. L. Gay, M. M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Trans. Speech and Audio Process.*, vol. 8, no. 6, pp. 656–663, 2000.

[6] J. Gunther, "Learning echo paths during continuous double-talk using semi-blind source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 646–660, 2011.

[7] S. Makino, *Audio source separation*. Springer, 2018.

[8] G. Cheng, L. Liao, H. Chen, and J. Lu, "Semi-blind source separation for nonlinear acoustic echo cancellation," *IEEE Signal Processing Letters*, vol. 28, pp. 474–478, 2021.

[9] G. Cheng, L. Liao, K. Chen, Y. Hu, C. Zhu, and J. Lu, "Semi-blind source separation using convolutive transfer function for nonlinear acoustic echo cancellation," *Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 88–95, 2023.

[10] X. Wang, Y. Yang, A. Brendel, T. Ueda, S. Makino, J. Benesty, *et al.*, "On semi-blind source separation-based approaches to nonlinear echo cancellation based on bilinear alternating optimization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2973–2987, 2024.

[11] K. Lu, X. Wang, T. Ueda, S. Makino, and J. Chen, "A computationally efficient semi-blind source separation approach for nonlinear echo cancellation based on an element-wise iterative source steering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 756–760.

[12] X. Wang, K. Mo, Y. Yang, L. Zhang, S. Makino, and J. Chen, "A cascaded semi-blind source separation method for joint acoustic echo cancellation, interference suppression, and noise reduction," in *Proc. IEEE Int. Workshop on Acoust. Sig. Enhance.*, 2024, pp. 369–373.

[13] Y. Yang, X. Wang, A. Brendel, W. Zhang, J. Benesty, S. Makino, *et al.*, "A data-reuse semi-blind source separation approach for nonlinear acoustic echo cancellation," in *Proc. IEEE Int. Workshop on Acoust. Sig. Enhance.*, 2024, pp. 379–383.

[14] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoust.*, 2011, pp. 189–192.

[15] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[16] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[17] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. Joint Workshop on Hands-free Speech Commun. and Mic. Arrays*, 2014, pp. 107–111.

[18] T. Wang, F. Yang, R. Zhu, and J. Yang, "Real-time independent vector analysis using semi-supervised nonnegative matrix factorization as a source model," in *Proc. Interspeech*, 2021, pp. 1842–1846.

[19] M. Schrammen, S. Kühl, S. Markovich-Golan, and P. Jax, "Efficient nonlinear acoustic echo cancellation by dual-stage multi-channel Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 975–979.

[20] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2065–2079, 2012.

[21] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, 2009.

[22] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–4.

[23] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[24] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2006.

[25] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, *et al.*, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, pp. 1–25, 2018.

[26] Y. Liang, G. Chen, S. Naqvi, and J. A. Chambers, "Independent vector analysis with multivariate Student's t-distribution source prior for speech separation," *Electronics Letters*, vol. 49, no. 16, pp. 1035–1036, 2013.

[27] J. Harris, B. Rivet, S. M. Naqvi, J. A. Chambers, and C. Jutten, "Real-time independent vector analysis with Student's t source prior for convolutive speech mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 1856–1860.

[28] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, and N. Ono, "Independent low-rank matrix analysis based on complex Student's t-distribution for blind audio source separation," in *Proc. IEEE Int. Workshop Machine Learning for Sig. Process.*, 2017, pp. 1–6.

[29] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. ISCA Workshop on Speech Synthesis*, 2004, pp. 223–224.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4214–4217.

A Single-Input/Binaural-Output Perceptual Rendering Based Speech Separation Method in Noisy Environments

Tianqin Zheng[†], Hanchen Pei[*], Ningning Pan[‡], Jilu Jin[†], Gongping Huang[*], Jingdong Chen[†], and Jacob Benesty[§]
[*]School of Electronic Information, Wuhan University, Wuhan, Hubei 430072, China
[†]CIAIC, Northwestern Polytechnical University, Xi'an, Shaanxi, China
[‡]RIIS, Southwestern University of Finance and Economics, Chengdu, Sichuan 610074, China
[§]INRS-EMT, University of Quebec, Montreal, Canada

***Abstract*—In this paper, we address the challenge of single-channel speech separation in noisy environments, where two active speakers and background noise are present in the observed signal. We propose using a dual path recursive neural network (DPRNN) to estimate the desired binaural signals from the single-channel noisy input. When the estimated binaural signal is played through headsets, listeners perceive the two speakers as originating from opposite directions, with the background noise coming from a separate direction. Additionally, the background noise is perceived to be further away from the two speakers, resulting in an improved signal-to-noise ratio (SNR). Research in psychoacoustics indicates that spatial unmasking in the perceptual domain enhances speech intelligibility in complex auditory scenes. This hypothesis is supported by both subjective and objective evaluations, including a significant 26% improvement in modified rhyme test (MRT) scores reported in this paper.**

***Index Terms*—Source separation, binaural hearing, speech enhancement, speech intelligibility.**

I. Introduction

Single-channel speech separation (SCSS), which focuses on isolating speech signals from two competing speakers using a single microphone observation, for enhancing both speech quality and intelligibility, consistently poses a challenge within the realms of source separation and speech enhancement [1], [2], [3]. Numerous approaches have been explored to address the SCSS problem, such as deep clustering methods [4], time-domain end-to-end methods [5], [6], and frequency-domain mask-based techniques [7]. With the development of more well-designed and compounded neural networks, the performance of speech separation is reported to be further improved [8], [9], [10]. However, the existence of acoustic background noise makes the SCSS problem more complicated and degrades the performance of existing methods by a large margin [11], [12]. Recently, many robust SCSS approaches have been introduced to mitigate the impact of additive noise. For example, [13] demonstrates the relative inseparability of noise and introduces a training objective inspired by SI-SDR. This approach exploits the inseparability of noise to implicitly segment the signal and reduce noise separation errors, thereby facilitating the training of more efficient separation systems using noisy oracle sources. Similarly, [14] proposes a novel network that unifies speech enhancement and separation using gradient modulation to improve noise robustness. Despite these advancements, these methods still encounter issues such as speech distortion and residual noise.

Despite these efforts and advances in solving the problem of SCSS in noisy environments, the majority of existing methods focus on producing a monaural output for each target speaker, which ignores the benefits of the human binaural auditory perception. Research on psychoacoustics has shown that (normal) human hearing is intrinsically capable of localizing sound sources and suppressing unwanted noise, known as the cocktail party effect, suggesting that properly spatializing sound sources in the perceptual domain helps in speech separation. More specifically, three typical scenarios of binaural presentations were investigated [15], [16], i.e., homophasic, heterophasic, and antiphasic. Among these methods, antiphasic presentation, where speech and noise are perceived from opposite directions, was demonstrated to offer the highest speech intelligibility [15], [17], [18], [19]. Based on the aforementioned findings, a single-input/binaural-output (SIBO) antiphasic noise reduction method [20] and a multiple-input/binaural-output (MIBO) antiphasic target speaker extraction method [21] were proposed, where significant improvement in speech intelligibility was reported compared to the corresponding monaural methods, especially in environments with low signal-to-noise ratios (SNRs) and low signal-to-interference ratios (SIRs).

In this paper, we adopt a dual-path recurrent neural network (DPRNN) structure [8], which includes an encoder, a rendering network, and a decoder, to produce a binaural output from a single-channel noisy observation. When the binaural signal is played back to listeners through headsets, it creates the perception of two competing speakers from opposite directions (e.g., the right and left sides of the head), while the background noise appears to be coming from behind. Additionally, the background noise is perceived to be further away from the two speakers, leading to an improved SNR. Simulations and experiments are conducted to evaluate the performance of the

978-616-590-477-3 ©2024 APSIPA APSIPA ASC 2024

proposed method. Both objective and subjective test results indicate that this method effectively improves speech quality and intelligibility.

II. Signal Model and Problem Formulation

We consider the scenario where a single microphone captures a mixture of two competing speech signals along with additional background noise. The observed signal can be formulated as

$$y(n) = s_1(n) + s_2(n) + v(n), \tag{1}$$

where $s_1(n)$ and $s_2(n)$ are the clean speech signals of the two speakers, $v(n)$ denotes the additive noise, and n is the discrete time index, which will be omitted hereinafter for readability.

This work aims to produce a binaural signal (for the left and right ears) from y utilizing a deep neural network (DNN). When the estimated binaural signal is presented to listeners through headsets, the two competing speakers and background noise will be perceived from three distinct directions/zones in the perceptual domain. Thus, the training target of the network can be formulated as

$$y_{\mathrm{L}} = \sum_{i=1}^{2} h_{\mathrm{L}}^{i} * s_i + h_{\mathrm{L}}^{v} * v, \tag{2}$$

$$y_{\mathrm{R}} = \sum_{i=1}^{2} h_{\mathrm{R}}^{i} * s_i + h_{\mathrm{R}}^{v} * v, \tag{3}$$

where $*$ denotes the linear convolution operation, h_{L}^{i}, $i = 1, 2$, and h_{L}^{v} denote the binaural room impulse responses (BRIRs) from the desired rendering locations of s_i, $i = 1, 2$, and v to the left ear, respectively, and h_{R}^{i}, $i = 1, 2$, and h_{R}^{v} denote the BRIRs from the desired locations to the right ear correspondingly. The BRIRs control the perceived locations of each source. In this work, we aim to render s_1 and s_2 to opposite directions, i.e., to the left and right side of the head, respectively, (or vice versa), while rendering v to the backside of the head.

III. Method Derivation

A. Model Architecture

As shown in Fig. 1(a), we adopt the DPRNN [8] as the backbone network to estimate the binaural signal in an end-to-end manner. The network consists of three modules: an encoder, a rendering network, and a decoder.

1) Encoder: The encoder is a 1-dimensional (1D) convolution layer with kernel size 16 and stride 8, followed by a rectified linear unit (ReLU) activation function. The input noisy speech sequence is then mapped into a latent feature space with dimension $C = 256$, thus giving a representation $\mathbf{A} \in \mathbb{R}^{C \times L}$, where L is the sequence length after the convolution operation.

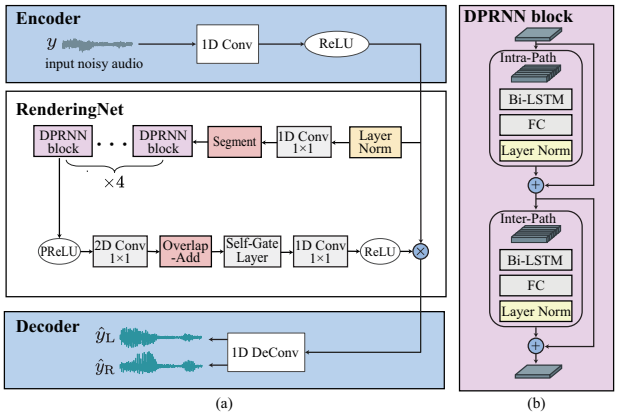


Fig. 1. Model architecture: (a) flowchart of the rendering based DPRNN-SIBO and (b) details of the DPRNN block.

2) Rendering Network: The rendering network is the essential module, which begins with a layer normalization operation to stabilize the training process. Then, it is followed by a special 1D convolution layer with a kernel size and stride of 1, denoted as a 1×1 convolution, which serves as a bottleneck to further fuse the information. The processed feature is then segmented into short frames of length $K = 250$ with a frame shift of $P = 125$. Thus, the feature dimension is changed to $\mathbb{R}^{C \times K \times T}$, where T is the number of frames.

The core part of the rendering network consists of 4 repetitions of a DPRNN block, illustrated in Fig. 1(b). *Each DPRNN block contains an intra-path and an inter-path module. Both modules include a bidirectional LSTM (Bi-LSTM) layer, a fully connected (FC) layer and the layer normalization operation. In addition, skip connections are adopted for both modules.* The 4 repeats of DPRNN block fully capture and exploit both local and long-term dependencies of the feature in and across short time frames. A PReLU function with one parameter follows and introduces a nonlinear transform. Then, a 2D convolution layer (kernel size 1) is employed to expand the feature dimension from C to $2 \times C$ to produce stacked two gain vectors related to the left and right ear outputs.

After the overlap-add operation, each gain vector is transformed back to size $\mathbb{R}^{C \times L}$, which is then passed through a self-gate module (composed of a hyperbolic tangent and a sigmoid activation function), a 1×1 1D convolution layer, and a ReLU function. Finally, the rendering network yields the binaural transfer functions, which are then elementwise multiplied in with the noisy representation $\mathbf{A}$ to produce estimates of the binaural signal in latent space.

3) Decoder: The decoder is a 1D deconvolution layer to reconstruct binaural estimates, i.e., $\hat{y}_{\mathrm{L}}$ and $\hat{y}_{\mathrm{R}}$, from the latent-space binaural estimates, which is a reverse process of the encoder.

B. Rendering Permutation Invariant Training Objective

The loss function is formulated using the scale-invariant SNR (SI-SNR) [22]. Specifically, for the left and right ear channels, the loss function is defined as follows:

$$\mathcal{L}\left(\hat{y}_{\mathrm{L}} \mid y_{\mathrm{L}}\right) = 10\log_{10}\frac{E\left(|y_{\mathrm{L}} - \hat{y}_{\mathrm{L}}|^2\right)}{E\left(|y_{\mathrm{L}}|^2\right)}, \tag{4}$$

$$\mathcal{L}\left(\hat{y}_{\mathrm{R}} \mid y_{\mathrm{R}}\right) = 10\log_{10}\frac{E\left(|y_{\mathrm{R}} - \hat{y}_{\mathrm{R}}|^2\right)}{E\left(|y_{\mathrm{R}}|^2\right)}, \tag{5}$$

where $E(\cdot)$ denotes mathematical expectation. Considering the binaural output, the loss function should be

$$\mathcal{L}\left(\hat{y}_{\mathrm{L}},\ \hat{y}_{\mathrm{R}}\right) = \mathcal{L}\left(\hat{y}_{\mathrm{L}} \mid y_{\mathrm{L}}\right) + \mathcal{L}\left(\hat{y}_{\mathrm{R}} \mid y_{\mathrm{R}}\right). \tag{6}$$

Note that rendering s_1 and s_2 to the left and right sides, respectively, or vice versa, are both acceptable for training. Thus, there are two possible permutations. To handle this, we exploit a direction permutation invariant loss during training:

$$\mathcal{L} = \min\left\{\mathcal{L}(\hat{y}_{\mathrm{L}_1},\ \hat{y}_{\mathrm{R}_1}),\ \mathcal{L}(\hat{y}_{\mathrm{L}_2},\ \hat{y}_{\mathrm{R}_2})\right\}, \tag{7}$$

where $\mathcal{L}(\hat{y}_{\mathrm{L}_1},\ \hat{y}_{\mathrm{R}_1}) = \mathcal{L}\left(\hat{y}_{\mathrm{L}} \mid y_{\mathrm{L}_1}\right) + \mathcal{L}\left(\hat{y}_{\mathrm{R}} \mid y_{\mathrm{R}_1}\right)$ denotes the loss when rendering s_1 and s_2 to left and right sides, respectively. In contrast, $\mathcal{L}(\hat{y}_{\mathrm{L}_2},\ \hat{y}_{\mathrm{R}_2})$ denotes the permutation case, i.e., the right and left side, respectively.

IV. Simulations

A. Training Setup

1) Dataset: We employ the WHAM! dataset [11] for both training and development purposes, which is constructed by incorporating noise from the WHAM! noise dataset [11] into the standard Wall Street Journal (WSJ0) two-speaker mixture dataset [4]. During the process of blending the two speech signals, the longer signal is adjusted to the length of the shorter one by truncating it in a 'minimum' mode, with a sampling rate set at 8 kHz. Our objective is to spatially position the two speech signals to the left and right sides, each at a distance of 1 meter from the listener's head. Concurrently, the additional noise is spatially rendered to the rear of the head, at varying distances of 1 meter, 2 meters, and 4 meters.

The location of each source is controlled by the corresponding BRIRs taken from the dataset [23]. The training targets are then generated by (2) and (3).

2) Training Configuration: For comparison, we exploit the original DPRNN (referred to as DPRNN-base in the following) with the same network structure as the baseline model. Unlike the proposed model (named as DPRNN-SIBO hereinafter), DPRNN-base achieves source separation by directly estimating the waveform of s_1 and s_2.

To train the models, we use the adaptive moment (Adam) estimation optimizer [24]. The initial learning rate is set to 10^{-3}. Given the case when training loss does not decrease in 3 consecutive epochs, the learning rate would accordingly half its value.

B. Simulation Results

1) Directional Perception Validation: To assess the effectiveness of the DPRNN-SIBO's rendering outcomes, we have synthesized 20 distinct test mixtures derived from the WHAM! dataset. It is important to note that these test signals were not present in either the training or development datasets. Additionally, we have established an input SNR of 0 dB, which is calculated as the ratio of the power of the mixed speech signals to that of the accompanying noise.

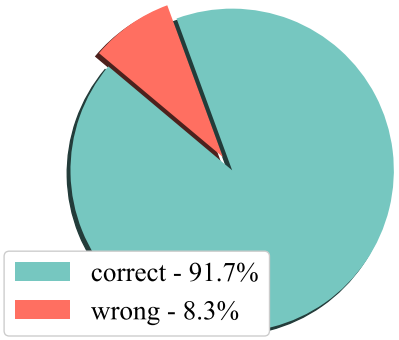


Fig. 2. Percentage of correct answers in the direction perception test. Correct answers include either "speaker 1 from left, speaker 2 from right, noise from back" or "speaker 2 from left, speaker 1 from right, noise from back." Any other choices are considered incorrect.

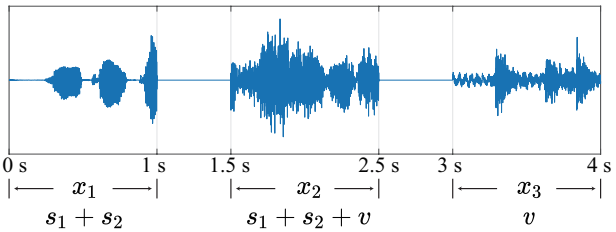


Fig. 3. Schematic diagram of observed signal construction, composed of five segments: x_1, x_2, and x_3 are the mixed signal, observed signal, and noise segment, respectively, with two silent segments of 0.5 seconds in the middle.

We invited five participants without hearing loss to listen to the binaural outputs of DPRNN-SIBO and listeners were required to select the direction of two speakers and noise. They are instructed to choose one from the following options for each source: 1) left, 2) right, and 3) back. Since we train DPRNN-SIBO with direction permutation invariant loss, both choices, i.e., "speaker 1 from left, speaker 2 from right, noise from back" and "speaker 2 from left, speaker 1 from right, noise from back" are considered to be correct. The findings are depicted in Fig. 2. It is evident that the direction of speakers and noise were properly chosen by most participants, leading to a 91.7% accuracy rate, which proves that DPRNN-SIBO can render sources to desired directions.

2) Impact of Rendering Distance: To evaluate the impact of rendering distance of the additive noise, we constructed 20 sets of observation signals as illustrated in Fig. 3. Both speech signals and noise are the same as those used in Section IV-B1. As shown in Fig. 3, the observed signal is 4-second long, consisting of five segments: a 1-second segment of 2-mixed speech signal, denoted as $x_1 = s_1 + s_2$, 0.5-second of silence, another 1-second of 2-mixed speech and noise, denoted as $x_2 = s_1 + s_2 + v$, another 0.5-second of silence, and 1-second

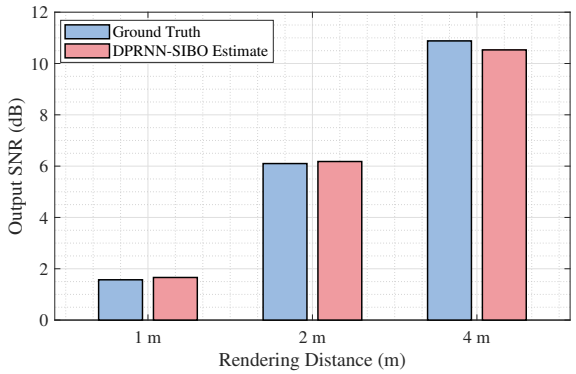


Fig. 4. Output SNR of the binaural signals estimated by DPRNN-SIBO and the ground truth SNR (dB).

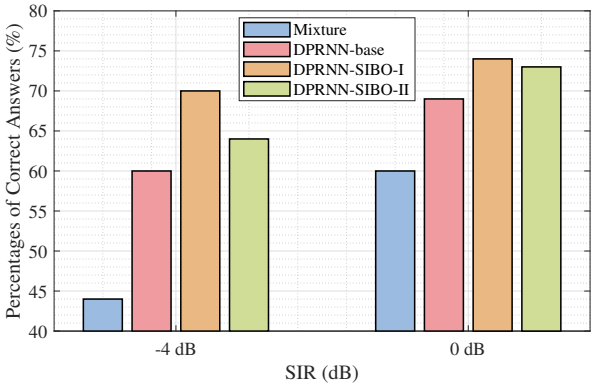


Fig. 5. Percentages of correct answers for input mixture, estimation of DPRNN-base, DPRNN-SIBO-I, and DPRNN-SIBO-II in the MRT test.

of additive noise, denoted as $x_3 = v$. Note that segment x_1, x_3 is utilized to calculate SNR and segment x_2 is used to make the observed signal similar to that in the training set. The input SNR is set to be 0 dB.

With the proposed DPRNN-SIBO, we can obtain the left ear and right ear outputs for x_1 and x_3, which are denoted as $\{\hat{x}_{\mathrm{L},1}, \hat{x}_{\mathrm{R},1}\}$, and $\{\hat{x}_{\mathrm{L},3}, \hat{x}_{\mathrm{R},3}\}$, respectively. Then the output SNRs of left and right ears are defined as follows:

$$\mathrm{SNR_L} = 10\log_{10}\frac{E\left(|\hat{x}_{\mathrm{L},1}|^2\right)}{E\left(|\hat{x}_{\mathrm{L},3}|^2\right)}, \tag{8}$$

$$\mathrm{SNR_R} = 10\log_{10}\frac{E\left(|\hat{x}_{\mathrm{R},1}|^2\right)}{E\left(|\hat{x}_{\mathrm{R},3}|^2\right)}. \tag{9}$$

Here, we define the binaural SNR (biSNR) as follows:

$$\mathrm{biSNR} = \frac{1}{2}(\mathrm{SNR_L} + \mathrm{SNR_R}). \tag{10}$$

Figure 4 lists the biSNRs achieved by DPRNN-SIBO and the ground-truth biSNRs, which are computed using (10) with the ground-truth signals formulated by (2) and (3). It can be shown that the biSNR of the output of DPRNN-SIBO closely approximates the ground-truth biSNR, proving that DPRNN-SIBO can render the additive noise to the desired distance.

3) MRT Tests: To evaluate the intelligibility improvement achieved by DPRNN-SIBO, modified rhyme tests (MRTs) [25] were conducted. We selected 48 carrier utterances from MRT database in the same way as our previous work [20], [21]. Then, the MRT speech signals were mixed with speech signals randomly selected from WSJ0 test dataset and additive noise from WHAM! noise dataset. The input SIRs were set to -4 dB and 0 dB and SNRs were randomly selected from -6 dB to 3 dB. DPRNN-SIBO and DPRNN-base are adopted to render or separate the speech. We consider two configurations of DPRNN-SIBO, denoted as DPRNN-SIBO-I and DPRNN-SIBO-II, both of which were designed to render the 2 speech signals to 1 m away from the listener to the left and right direction, while rendering the additive noise to 1 m and 4 m away from the listener to the backward direction, respectively. To evaluate speech intelligibility, participants were required to choose the word they hear in the MRT carrier sentence "please select the word –" from all the testing signals. More correct answers provided by the participants indicate higher speech intelligibility achieved by the tested method.

Figure 5 shows the percentage of correct answers of MRT from DPRNN-base and the proposed two methods, DPRNN-SIBO-I and DPRNN-SIBO-II. We can see that all tested methods are capable of increasing speech intelligibility by a large margin as compared to the MRT results of the mixed signal. Among them, DPRNN-base increases the accuracy by 9% and 16% in the 0 dB and -4 dB experiments while DPRNN-SIBO-I increases the accuracy by 14% and 26%, respectively. Also, both configurations of the proposed method outperform DPRNN-base in both input SIR scenarios. Specifically, DPRNN-SIBO-I achieves 10% increase than DPRNN-base in -4 dB, and 5% increase in 0 dB. Moreover, we note that DPRNN-SIBO-I consistently performs better than DPRNN-SIBO-II. This can be attributed to the fact that DPRNN-SIBO-II pushes the noise source to a distance of 4 m for more noise suppression, which results in more distortion of the speech signal. This also explains the lower percentage of correct answers for DPRNN-base.

V. Conclusions

This paper presented a deep learning-based single-input/binaural-output rendering method to improve speech intelligibility in complex acoustic environments, where speech signals from two competing speakers and background noise are captured by a single microphone. A rendering network based on DPRNN was trained to take the single-channel mixture as input and generate a binaural signal. When delivered through headphones, this binaural signal allows listeners to perceive each speaker from opposite directions while placing the noise further away in the back direction. The proposed rendering method effectively reduces interference from competing speakers and improves speech intelligibility, especially in low-SIR conditions. MRT results show a notable 26% improvement over the mixed speech signal, demonstrating the effectiveness of the proposed method.

REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Springer Science & Business Media, 2006.

[2] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, Oct. 2018.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, pp. 31–35, 2016.

[5] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE ICASSP*, pp. 696–700, IEEE, 2018.

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1256–1266, Aug. 2019.

[7] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[8] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE ICASSP*, pp. 46–50, IEEE, 2020.

[9] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

[10] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE ICASSP*, pp. 21–25, IEEE, 2021.

[11] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[12] S. Zhao and B. Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *Proc. IEEE ICASSP*, pp. 1–5, IEEE, 2023.

[13] M. Maciejewski, J. Shi, S. Watanabe, and S. Khudanpur, "Training noisy single-channel speech separation with noisy oracle sources: A large gap and a small step," in *Proc. IEEE ICASSP*, pp. 5774–5778, 2021.

[14] Y. Hu, C. Chen, H. Zou, X. Zhong, and E. S. Chng, "Unifying speech enhancement and separation with gradient modulation for end-to-end noise-robust speech separation," in *Proc. IEEE ICASSP*, pp. 1–5, 2023.

[15] J. Licklider, "The influence of interaural phase relations upon the masking of speech by white noise," *J. Acoust. Soc. Am.*, vol. 20, pp. 150–159, March. 1948.

[16] L. A. Jeffress, "A place theory of sound localization.," *J. Comp. Physiol. Psychol.*, vol. 41, p. 35, Feb. 1948.

[17] Y. Wang, J. Chen, J. Benesty, J. Jin, and G. Huang, "Binaural heterophasic superdirective beamforming," *Sensors*, vol. 21, no. 1, p. 74, 2020.

[18] J. Jin, J. Chen, J. Benesty, Y. Wang, and G. Huang, "Heterophasic binaural differential beamforming for speech intelligibility improvement," *IEEE Trans. Veh. Technol.*, vol. 69, pp. 13497–13509, Nov. 2020.

[19] J. Benesty, G. Huang, J. Chen, and N. Pan, *Microphone Arrays*, vol. 22. Berlin, Germany: Springer-Verlag, 2023.

[20] N. Pan, Y. Wang, J. Chen, and J. Benesty, "A single-input/binaural-output antiphasic speech enhancement method for speech intelligibility improvement," *IEEE Signal Process. Lett.*, vol. 28, pp. 1445–1449, 2021.

[21] X. Wang, N. Pan, J. Benesty, and J. Chen, "On multiple-input/binaural-output antiphasic speaker signal extraction," in *Proc. IEEE ICASSP*, pp. 1–5, IEEE, 2023.

[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE ICASSP*, pp. 626–630, IEEE, 2019.

[23] B. I. Bacila and H. Lee, "360 binaural room impulse response (BRIR) database for 6dof spatial perception research," in *Audio Engineering Society Convention 146*, Audio Engineering Society, 2019.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] S. ANSI, "3.2, methods for measuring the intelligibility of speech over communication systems," *American National Standards Institute*, pp. 3–2, 1989.

Real-Time Noise Estimation for Lombard-Effect Speech Synthesis in Human–Avatar Dialogue Systems

Yuto Ishikawa*, Osamu Take*, Tomohiko Nakamura†, Norihiro Takamune*,
Yuki Saito*, Shinnosuke Takamichi*‡, and Hiroshi Saruwatari*
* The University of Tokyo, Tokyo, Japan
† The National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
‡ Keio University, Kanagawa, Japan

***Abstract*—To facilitate smooth communication between a human (client) and an avatar/robot, it is essential to generate avatar/robot speech that reduces the listening effort for the client in noisy acoustic environments. In this paper, we propose a framework for the following two purposes by leveraging the estimated noise using our previously proposed real-time blind speech extraction method: (i) real-time adjustment so that the power of the avatar/robot speech relative to that of the estimated noise is constant and sufficiently large, and (ii) mimicking the Lombard effect by simply applying filters to the avatar/robot speech. The proposed framework can also smoothly connect speeches generated from voice conversion (avatar) and text-to-speech (robot) techniques. Subjective evaluations demonstrated that the proposed framework achieves natural synthesis in noisy environments.**

I. Introduction

Systems that allow human-like avatars and robots to interact with humans (clients) have been researched and developed for decades [1]. One essential component for realizing such systems is an autonomous dialogue system for robots [2]. By utilizing advances in technologies such as automatic speech recognition (ASR) [3], dialogue control [4], response generation [5], and text-to-speech (TTS) [6], robots can respond to clients' queries flexibly. These advances in automatic dialogue are expected to reduce human resources by replacing roles that only humans can take conventionally. However, in practical applications, irregular cases that an autonomous robot cannot deal with may exist (e.g., speech recognition errors and out-of-domain dialogue scenes). In such cases, the robot can be switched to an avatar, and a human operator can take over the response to handle client's questions. In this way, by having an automatic dialogue in normal cases and by having a human operator intervention only when necessary, a single operator can concurrently control multiple avatars. As a result, the number of avatars that a single operator can control increases, leading to a reduction in human resources. Additionally, in this case, the consistency between the avatar and robot speakers can be maintained by converting the operator's voice to the robot's speaker voice using a voice conversion (VC) technique [7]. On the other hand, since the power of the voice output by VC is correlated with that of the input voice spoken by the operator, there can be a large difference in power between the automatic speech (TTS) and the converted speech of the operator (VC). As a result, clients may feel discomfort owing to the discontinuity of speech that they hear. In the following, the robot's voice generated by a TTS technique, the voice converted from the operator's speech using a VC technique, and the avatar's/robot's voice that the clients hear are called *TTS voice*, *VC voice*, and *synthesized voice*, respectively. In this paper, we address a practical situation where an avatar/robot interacts with a single client in real-world noisy environments.

In such cases, the background noise makes a client's and synthesized voices obscure. Consequently, the client and the avatar's operator have difficulty listening to each other's speech, and the accuracy of ASR for the robot is also degraded. To extract the client's speech from the audio signals observed by a microphone array embedded in the avatar/robot, we previously proposed a real-time speech extraction framework [8]. This is expected to make it easier for the operator to understand the client's speech and improve the accuracy of ASR in noisy environments. On the other hand, since noise information is discarded through speech extraction, the avatar's operator and the robot cannot consider the background noise in the client's environment and cannot always generate a voice that the client can hear easily in noisy environments. As a result, the synthesized voice remains hidden in noise, which hinders smooth communication.

In this paper, we propose a framework for generating natural and audible speech for the client. In this framework, we utilize the property of a linear demixing filter, that is, it can accurately separate noise signals in a situation where a single directional target speaker exists in diffuse noise [9]. By utilizing this property, we estimate the client's background noise in real time using the multichannel speech extraction method proposed in [8] and adjust the power of the synthesized voice to be audible with less listening effort according to the estimated noise level. This adjustment results in the continuity of the power of the synthesized voice when switching between the TTS and VC

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 (for framework development) and Tateisi Science and Technology Foundation (for practical experiment).

voices. Additionally, to further improve the intelligibility of the synthesized voice for the client in noisy environments, we partly simulate the Lombard effect [10], [11], which is the involuntary change in human speech features in noisy environments, on the synthesized voice. Subjective evaluations showed that the synthesized voice generated by our proposed framework sounds natural in noisy environments compared with that generated without processing or by simply applying a constant gain.

II. Related works

A. Real-time speech extraction framework

In [8], we previously proposed the real-time speech extraction framework based on independent low-rank matrix analysis (ILRMA) [12] and rank-constrained spatial covariance matrix estimation (RCSCME) [13]. In ILRMA, the short-time Fourier transform (STFT) is utilized to process the observed audio signals in the time-frequency domain. Let us define the STFTs of the observed, source, and separated signals as $\boldsymbol{x}_{ij} \in \mathbb{C}^M$, $\boldsymbol{s}_{ij} \in \mathbb{C}^N$, and $\boldsymbol{y}_{ij} \in \mathbb{C}^N$, respectively. Here, $i \in \{1, ..., I\}$, $j \in \{1, ..., J\}$, $m \in \{1, ..., M\}$, and $n \in \{1, ..., N\}$ are the indices of the frequency bins, time frames, microphones, and sources, respectively. If each source is a point source and the reverberation time is sufficiently shorter than the STFT window length, the observed signal is approximately represented as $\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}$. Here, $\boldsymbol{A}_i = (\boldsymbol{a}_{i1}, ..., \boldsymbol{a}_{iN}) \in \mathbb{C}^{M \times N}$ is the mixing matrix and $\boldsymbol{a}_{in}$ is the steering vector for the nth source to the microphone array. In ILRMA, under a determined condition ($M = N$), we estimate the time-invariant demixing matrix $\boldsymbol{W}_i \in \mathbb{C}^{N \times M}$ and subsequently separate the observed signals into each source signal as $\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}$. Here, to resolve the scale ambiguity in each frequency bin of the separated signals, the frequency-wise scale of each separated signal is aligned with that of the observed signal at the reference microphone channel using projection back (PB) [14].

However, when we apply ILRMA to a mixture that consists of a directional target speech and diffuse noise, one separated signal corresponding to the target speech includes diffuse noise components [15]. On the other hand, the other separated signals contain only diffuse noise and suppress the target speech with high accuracy [9]. RCSCME utilizes these properties of the separated signals and efficiently estimates the parameters for a multichannel Wiener filter. RCSCME uses parameters estimated by ILRMA to reduce the number of to-be-estimated parameters and can achieve rapid processing. However, it is difficult to extend the speech extraction method consisting of ILRMA and RCSCME for real-time processing in a straightforward manner because the update algorithm for ILRMA is computationally costly owing to numerous matrix operations. Here, we exploit two facts: 1) the ILRMA output required in RCSCME is only the time-invariant demixing matrix and 2) RCSCME itself is computationally efficient and can be executed within the shift length of the STFT. On the basis of these facts, in [8], a framework that performs ILRMA and RCSCME in parallel by introducing the blockwise batch algorithm [16], [17] has been proposed. This framework can execute RCSCME at the shift length intervals and output the extracted target speech with low latency (less than 100 ms, excluding input/output latency). On the other hand, since it is difficult to perform ILRMA within the shift length, it estimates and outputs the demixing matrix at longer time intervals.

Since ILRMA is a fully blind method, it is necessary to select the channel containing the target speech from the separated signals without any prior information. However, by introducing a regularizer based on spatial prior information into ILRMA, we can induce the separated signal at a specific channel to contain the target speech. As one such regularizer, in [8], ILRMA with null-based spatial regularization (NSR-ILRMA) has been proposed. The spatial regularization term is calculated using a steering vector for the target speech, $\hat{\boldsymbol{a}}_i$, derived from prior information about the client's position against the avatar/robot. The cost function of NSR-ILRMA is expressed as

$$\begin{aligned}\mathcal{T} =& \frac{1}{J} \sum_{i,j,n} \left(\frac{|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{r_{ijn}} + \log r_{ijn} \right) - \sum_i \log |\det \boldsymbol{W}_i|^2 \\ &+ \sum_{i,n} \mu_{in} (1 - \delta_{nn^{(\mathrm{t})}}) |\boldsymbol{w}_{in^{(\mathrm{t})}} \hat{\boldsymbol{a}}_i|^2 + \mathrm{const.}, \end{aligned} \tag{1}$$

where $r_{ijn} > 0$ denotes the time-varying variance of the separated signal y_{ijn}, const. represents the term independent of $\boldsymbol{w}_{in}$ and r_{ijn}, and $\cdot^{\mathsf{H}}$ is the Hermitian transpose. Here, $n^{(\mathrm{t})}$ is a prespecified channel index corresponding to the target speech. This regularizer induces $\boldsymbol{w}_{in}$ $(n \neq n^{(\mathrm{t})})$ to form the null in the direction of the target speech, and thus, $\boldsymbol{w}_{in^{(\mathrm{t})}}$ seems to correspond to the target speech. The process of minimizing (1) is detailed in [8]. In this case, we can choose $n^{(\mathrm{t})}$ for the channel selection.

B. Lombard effect

The Lombard effect is the involuntary modification of the characteristics of human speech produced in noisy environments to enhance its audibility and intelligibility for listeners [10]. It has been reported that the Lombard speech, which refers to speech produced under the Lombard effect, exhibits changes in various characteristics, such as its power, spectral characteristics, and fundamental frequency [11]. In [11], the spectra of speech produced in a markedly noisy environment were compared with those produced in a quiet environment. This comparison showed an average increase in the power of approximately 5 dB in the low-frequency band around 200 Hz and in the high-frequency band above 5000 Hz, whereas an average increase in the power of approximately 15 dB is observed in the frequency band around 500 to 4000 Hz. It is also reported that relative to quiet environments, the mean vocal level (the power of speech) increased by approximately 14.5 dB.

III. Proposed method

A. Motivation

A simple approach to improving audibility in noisy environments is by applying a constant gain to the synthesized

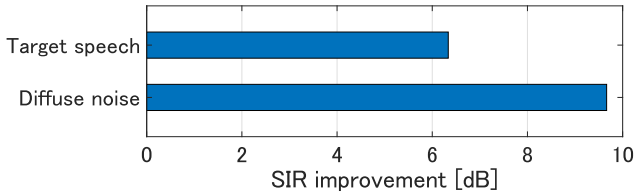


Fig. 1. SIR improvements of the target speech (above) and diffuse noise (below) estimated by proposed method with offline NSR-ILRMA.

voice. However, since noise is unstationary, a fixed gain can be an inappropriate setting. For example, if the gain is small, the synthesized voice can remain inaudible in very noisy environments, and conversely, if the gain is large, the synthesized voice can become uncomfortably loud in quiet environments. Therefore, we must adaptively adjust the gain to the noise intensity in real time.

To address this issue, we propose a method to estimate the power of time-varying noise in real time and adjust that of the synthesized voice to ensure that the power of the synthesized voice is sufficiently higher than that of noise at a constant rate. This approach is expected to generate an audible voice suitable for various noisy environments. Moreover, even in a special case where TTS and VC voices switch, this processing makes the power of the synthesized voice almost constant as long as the noise power does not change abruptly, enabling a smooth connection of the synthesized voice. If the observed signals are directly used to calculate the noise power without any processing, the client's speech might be erroneously considered as background noise. Therefore, the power of synthesized voice increases in correlation with not only the power of the background noise but also that of the client's speech. As a result, even if the environment is consistently quiet, the synthesized voice will become louder when the client speaks loudly. Therefore, a method is required to suppress the client's speech and estimate only the noise power in real time.

B. Proposed real-time noise estimation procedure

In this section, we aim to remove the target speech and estimate the noise accurately in real time. First, we focus on the characteristics of source separation methods based on the linear demixing filter, such as ILRMA. In [9], it is reported that the linear demixing filter can accurately separate noise signals in a situation where a single directional target speaker exists in diffuse noise. In this situation, the observed signal $\boldsymbol{x}_{ij}$ can be represented as $\boldsymbol{x}_{ij} = \tilde{\boldsymbol{a}}_i \tilde{s}_{ij} + \boldsymbol{u}_{ij}$, where $\tilde{\boldsymbol{a}}_i$ and $\tilde{s}_{ij}$ denote the steering vector and the source signal of the target speech, respectively, and $\boldsymbol{u}_{ij}$ denotes the source image of diffuse noise. By performing the real-time speech extraction method with NSR-ILRMA, we can obtain the linear demixing matrix $\boldsymbol{W}_i$ and the index corresponding to the target speech $n^{(\mathrm{t})}$ in real time. Here, the scale of $\boldsymbol{W}_i$ is modified by PB as

$$\boldsymbol{W}_i \leftarrow \mathrm{diag}\big(\boldsymbol{e}_{m_{\mathrm{ref}}}^{\mathsf{H}} \boldsymbol{W}_i^{-1}\big) \boldsymbol{W}_i, \tag{2}$$

where $\mathrm{diag}(d_1, ..., d_N)$ denotes the operator that outputs an $N \times N$ diagonal matrix whose diagonal elements are given by the vector $(d_1, ..., d_N)$, m_{ref} is the index of the reference microphone channel, and $\boldsymbol{e}_m \in \mathbb{C}^M$ is a one-hot vector with 1 only in the mth element. By utilizing these parameters, we can define the estimated noise signal $\hat{u}_{ij}$ as

$$\hat{u}_{ij} = \sum_{n \neq n^{(\mathrm{t})}} \boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}. \tag{3}$$

Here, we focus on the facts that the demixing filters $\boldsymbol{w}_{in}$ $(n \neq n^{(\mathrm{t})})$ can form the null of the steering vector for the target speech $\tilde{\boldsymbol{a}}_i$ and then $\boldsymbol{w}_{in}^{\mathsf{H}} \tilde{\boldsymbol{a}}_i \approx 0$ $(n \neq n^{(\mathrm{t})})$, resulting in $\hat{u}_{ij} = \sum_{n \neq n^{(\mathrm{t})}} \boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{u}_{ij}$. We note that the noise power is reduced compared with the true value by the noise component leaking in the target speech direction, $\boldsymbol{w}_{in^{(\mathrm{t})}}^{\mathsf{H}} \boldsymbol{u}_{ij}$. We experimentally verified this characteristic, and Fig. 1 shows the source-to-interference ratio (SIR) [18] improvements for the target speech and the diffuse noise when offline NSR-ILRMA is applied to mixture signals consisting of a single directional target speech and diffuse noise. The experimental conditions for Fig. 1 are as follows: We used a 4.2-s-long female speech signal from the JSUT dataset [19] as a dry source. Then the dry source was convolved with the recorded impulse response and its signal length was adjusted to 5.12 s by padding with the zero value. The recording conditions for both the diffuse noise and impulse response and the settings for STFT were the same as those described in Section IV-A. The target speech was mixed with the diffuse noise so that the input signal-to-noise ratio (SNR) became 0 dB for the entire signal at a reference microphone channel except for the silent intervals. We applied NSR-ILRMA to the mixture signal, and an estimated noise signal was then obtained by calculating (3). As demonstrated in Fig. 1, we empirically confirmed that NSR-ILRMA can separate the background noise with higher accuracy than in the case of the target speech under the experimental conditions used in this study. This result is consistent with [9], [13]. Moreover, it is expected that we can accurately estimate the diffuse noise in real time using the real-time speech extraction method.

Using the real-time estimated noise, we attempt to generate audible speech by applying a time-varying gain to the synthesized voice so that the ratio between the power of the output speech and noise estimated in real time is constant. Furthermore, by imitating the spectral energy changes described in Section II-B, we simulated the Lombard effect on the synthesized voice. These modifications will enable the generation of more natural-sounding speech across diverse noisy environments.

C. Proposed framework

Fig. 2 illustrates our proposed framework. In the following, we describe the details of the processing part highlighted in red frame.

1) SNR modification using real-time noise estimation: First, we consider modifying the vocal level of the synthesized voice. For the client to hear the synthesized voice clearly in noisy environments, we should calculate the gain to achieve a specified output SNR. This requires obtaining the power of both the synthesized voice and noise.

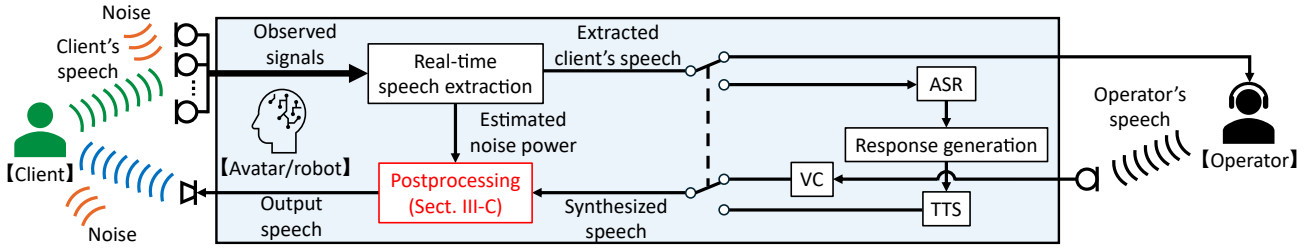


Fig. 2. Schematic of proposed framework for dialogue system embedded in avatar/robot.

For the power of the synthesized voice, sample speeches are prepared for both TTS and VC voices, and their powers are precalculated as references. That is, when a $T^{(l)}$-sample-long segment of TTS or VC voices $s^{(l)}_{\mathrm{sample}}[t]$ ($l \in \{\mathrm{TTS}, \mathrm{VC}\}$, $t = 1, ..., T^{(l)}$) is given beforehand, we define the average power (the power per unit time) of the synthesized voice $S^{(l)}$ as

$$S^{(l)} = \frac{1}{T^{(l)}} \sum_{t=1}^{T^{(l)}} \left(s^{(l)}_{\mathrm{sample}}[t] \right)^2 . \tag{4}$$

On the other hand, for the noise power, to suppress the client's speech and adapt to the time-varying noisy environment, we use the noise signal estimated at the shift length intervals using the real-time speech extraction framework and (3). Here, to suppress the rapid changes in noise power, the exponential smoothing method is introduced and noise information history is also considered. The noise signal estimated by the real-time speech extraction framework is denoted as $u[t]$. This noise signal $u[t]$ is estimated at every shift length ΔT. For the integer k, the noise power estimate U_k for the time $t = (k-1)\Delta T+1, ..., k\Delta T$ is defined using the estimated noise signal at that time interval as

$$U_k = (1-\lambda)U_{k-1} + \frac{\lambda}{\Delta T} \sum_{t=(k-1)\Delta T+1}^{k\Delta T} (u[t])^2 , \tag{5}$$

where $\lambda \in (0, 1]$ denotes a forgetting factor and U_0 is set to 0.

Finally, when the output SNR is set to R, the time-varying gain $\alpha^{(l)}_k$ for the synthesized voice is calculated as

$$\alpha^{(l)}_k = 10^{\frac{R}{20}} \sqrt{\frac{U_k}{S^{(l)}}}. \tag{6}$$

2) Simple imitation of Lombard effect: Next, to imitate the increased energy in the relatively high-frequency band observed in the Lombard speech, we consider boosting the energy in the 500–4000 Hz range by 5–10 dB compared with the other frequency bands. In this paper, we apply a bandpass filter that allows the 500–4000 Hz frequency band to pass through and an all-pass filter with a similar phase response to the bandpass filter to the input TTS/VC voice in parallel. Subsequently, the filtered signals are summed. That is, if the transfer functions of the bandpass and the all-pass filters in the z-domain are represented as $h_{\mathrm{B}}(z)$ and $h_{\mathrm{A}}(z)$, respectively, the transfer function of our proposed filter is expressed as $h_{\mathrm{B}}(z) + h_{\mathrm{A}}(z)$.

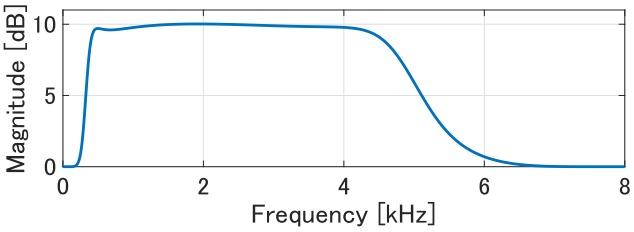


Fig. 3. Magnitude frequency response for proposed filter.

The bandpass filter $h_{\mathrm{B}}(z)$ is designed as follows: 5th-order Butterworth type, a passband of 500–4000 Hz, a passband ripple of 3 dB, stopband frequencies of 50 and 7000 Hz, and a stopband attenuation of 60 dB. The cutoff frequencies of this filter become about 376 and 4720 Hz. Then, the 5th-order infinite impulse response all-pass filter $h_{\mathrm{A}}(z)$ is designed by searching for parameters using a greedy algorithm to minimize the mean squared error between phase responses of this filter and the bandpass filter. Here, the gain of the bandpass filter relative to that of the all-pass filter is determined to increase the energy in the passband by 10 dB. The magnitude frequency response of the proposed filter $h_{\mathrm{B}}(z) + h_{\mathrm{A}}(z)$ is shown in Fig. 3.

Since this filtering affects the power of the entire output speech, the processing should be conducted in the following order: first, perform the filtering described in this section, and then perform the SNR modification described in Section III-C1.

IV. Experiments

In our experiments, we consider three application scenarios where a human and an avatar/robot are talking under a diffuse-noise condition: (i) listening to only the TTS voice, (ii) listening to only the VC voice, and (iii) listening to the TTS voice followed by the VC voice almost continuously with different powers. We evaluate the effectiveness of the proposed method by subjective evaluation.

A. Experimental conditions

Diffuse noise and impulse responses were recorded at the Ito International Research Center, The University of Tokyo. A circular microphone array with a radius of 3.25 cm and composed of four omnidirectional microphones was placed at a height of 1 m from the floor. During the diffuse noise recording, 10 participants were seated 2–4 meters apart from the microphone array and engaged in conversations with others

TABLE I

PREFERENCE SCORES ON NATURALNESS OF OUTPUT VOICE. VALUES ENCLOSED IN PARENTHESES DENOTE p-VALUES OF STUDENT'S t-TEST. **BOLD** INDICATES SIGNIFICANTLY (p-VALUE < 0.025) BETTER SCORES.

Case	Noise condition	Number of evaluators	Baseline / Naive	Baseline / Proposed	Naive / Proposed
(i)	-5 dB	26	0.115 / **0.885** ($< 10^{-10}$)	0.154 / **0.846** ($< 10^{-10}$)	0.558 / 0.442 (9.70×10^{-2})
	0 dB	28	0.098 / **0.902** ($< 10^{-10}$)	0.089 / **0.911** ($< 10^{-10}$)	0.268 / **0.732** ($< 10^{-10}$)
	5 dB	25	0.070 / **0.930** ($< 10^{-10}$)	0.040 / **0.960** ($< 10^{-10}$)	0.100 / **0.900** ($< 10^{-10}$)
(ii)	-5 dB	25	0.110 / **0.890** ($< 10^{-10}$)	0.030 / **0.970** ($< 10^{-10}$)	0.090 / **0.910** ($< 10^{-10}$)
	0 dB	25	0.110 / **0.890** ($< 10^{-10}$)	0.060 / **0.940** ($< 10^{-10}$)	0.040 / **0.960** ($< 10^{-10}$)
	5 dB	25	0.050 / **0.950** ($< 10^{-10}$)	0.020 / **0.980** ($< 10^{-10}$)	0.020 / **0.980** ($< 10^{-10}$)
(iii)	-5 dB	24	0.083 / **0.917** ($< 10^{-10}$)	0.073 / **0.927** ($< 10^{-10}$)	0.094 / **0.906** ($< 10^{-10}$)
	0 dB	25	0.090 / **0.910** ($< 10^{-10}$)	0.030 / **0.970** ($< 10^{-10}$)	0.060 / **0.940** ($< 10^{-10}$)
	5 dB	26	0.058 / **0.942** ($< 10^{-10}$)	0.067 / **0.933** ($< 10^{-10}$)	0.038 / **0.962** ($< 10^{-10}$)

around them or read a text provided to them beforehand. Simultaneously, music was played from loudspeakers embedded in the ceiling. The impulse responses were recorded under the following conditions: the height of the target speaker was 1.1 m, the horizontal distance between the target speaker and the microphone array was 1 m, and the reverberation time T_{60} was around 750 ms.

Next, we describe the processes for synthesizing the TTS and VC voices. For both, the synthesized speech speaker was one female speaker from the SaSLaW corpus [20]. For the TTS voice, FastSpeech 2 [21] was used as the deep neural network model to output the mel-spectrogram of the speech, and the trained HiFi-GAN [22] was used as the vocoder to generate the speech waveform. The details of TTS models are based on [20]. Then, FastSpeech 2 was pretrained on the JSUT dataset [19] and fine-tuned on single-speaker recordings of SaSLaW. For the VC voice, RVC[1] trained on the synthesized speech speaker's data was used, and the input speaker was one female speaker from the JVS corpus [19].

In this paper, we simulate a situation where the VC voice is quieter and more difficult to hear than the TTS voice in different noisy environments. To achieve this, the VC voices were adjusted so that the average power of each VC voice became -5 dB relative to that of each TTS voice. Here, the TTS voices used were adjusted beforehand so that their average power became constant. Then, we adjusted the noise signal so that the average power of the noise signal at the reference channel became -5, 0, and 5 dB relative to that of the TTS voice, thereby creating three types of noise condition. We note that, as a result, the SNRs of the TTS voice were 5, 0, and -5 dB, and the SNRs of the VC voice were 0, -5, and -10 dB, in the same order as above. Then, for experiments (i) and (ii), to obtain clean synthesized speech signals, two utterances were selected from either TTS or VC voices and concatenated with 5-s-long silent intervals in between. The silent interval before the first utterance was used to calculate the noise power sufficiently and the silent interval between the utterances was used to simulate a conversation between a human and an avatar/robot. Similarly, for experiment (iii), a clean synthesized speech signal was created by concatenating a 5-s-long silent interval, one TTS utterance, a 1-s-long silent interval, and one VC utterance in this order. The 1-s-long interval between the TTS and VC utterances simulates a scene where the operator intervenes suddenly, whereas the 5-s-long silent interval before the TTS utterances serves the same purpose as in (i) and (ii). In total, there were nine experimental conditions: Three experimental scenarios and three noise mixing conditions. For each experimental condition, 10 clean synthesized speech signals were created. The clean synthesized speech signals created in this way are referred to as *standard speech signals*.

Next, we describe how to generate the input signals for the real-time speech extraction framework in each noisy environment. To simulate client utterances in noisy environments, we used single-female-speaker utterances from the JSUT dataset [19] as client's dry sources. These utterances were within 5 s in length. After convolving the dry sources with the impulse responses, the resulting signals were mixed with the noise signals after power adjustment with TTS voices. In the mixing process, the average power of each convolved signal was adjusted to be 0 dB relative to the power of the noise, and the timing of each convolved signal was aligned with the 5-s-long silent intervals inserted before two TTS or VC voices in the experiments (i) and (ii), and before the first TTS voice in the experiment (iii). The sampling rate of all signals was set to 16 kHz.

For comparison, the following three methods were used.

- **Baseline**: the standard speech signals are output without any additional processing.
- **Naive**: signals applied at a constant gain of $+10$ dB to the standard speech signals are output.
- **Proposed**: signals adjusted by the proposed real-time framework with an output SNR of $R = 15$ dB for each noise power estimate are output.

For the real-time speech extraction framework, NSR-ILRMA was used in the ILRMA part described in [8], and the other conditions were also the same as those in [8]. STFT was performed using a 64-ms-long Hann window with a shift length of 32 ms. The sample speeches $s^{(\mathrm{TTS})}_{\mathrm{sample}}[t]$ and $s^{(\mathrm{VC})}_{\mathrm{sample}}[t]$

[1] https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI

used for SNR modification were generated by selecting one utterance from TTS and VC voices, respectively, and applying the filtering process described in Section III-C2. For the forgetting factor λ, we set it to 0.5 by experimentally choosing it from some values in $(0, 1]$.

For subjective evaluation, we generated evaluation speech signals to simulate listening under each noise condition by mixing the output speech signal for each compared method with the corresponding noise signal. Then, two of the three evaluation speech signals were selected as an evaluation pair, and AB preference tests (listening tests) were conducted for each pair to determine which evaluation speech signal sounded more natural in a noisy environment. Each evaluator listened to and evaluated a total of 12 evaluation pairs containing the same number of evaluation pairs for each combination of the compared method.

B. Subjective evaluation

For each experiment and each noise mixing condition, 24–28 evaluators were recruited and 96–112 responses were collected via the crowdsourcing platform "Lancers"[2]. The significant differences in preference scores were evaluated using Student's two-tailed t-test with a significance level of 5%. The results are shown in Table I. Under almost all experimental conditions, the scores were significantly higher for Naive than for Baseline and for Proposed than for Naive. We also consider the following to be the reason why there was no significant difference between Naive and Proposed only in the case (i) with a noise condition of -5 dB. First, under this condition, the ratio between the power of the TTS voice and noise signal was 5 dB. Therefore, the synthesized voice was adjusted so that the resulting SNR became 15 dB for Naive. On the other hand, Proposed also functions so that the average power of the synthesized voice became 15 dB relative to the noise power estimate under this experimental condition. However, since noise components remain in the separated signal corresponding to the target speech by ILRMA, the noise power estimate is slightly lower than the actual noise power. Then, in the gain calculation (6), U_k becomes smaller and the gain also becomes smaller, resulting in a slightly lower SNR than the ideal output SNR of $R = 15$ dB. Thus, Naive has slightly higher power than Proposed, making it easier and slightly more natural for the client to hear in a noisy environment. This results in both methods showing a competitive performance.

V. Conclusion

In this paper, for the dialogue system embedded in the avatar/robot, we proposed a real-time framework to modify the synthesized speech so that it is easy for clients to hear. We estimated the noise power in real time by utilizing the real-time speech extraction framework based on ILRMA and RCSCME. In the postprocessing part, we performed (i) power adjustment using estimated noise power in real time to make the synthesized speech audible and (ii) filter processing to partially imitate the Lombard effect. Subjective evaluation experiments confirmed that the proposed framework can generate a more natural speech in noisy environments than the simple approaches.

[2]https://www.lancers.jp/

References

[1] O. Mubin, M. I. Ahmad, S. Kaur, W. Shi, and A. Khan, "Social robots in public spaces: A meta-review," in *Proc. ICSR*, 2018, pp. 213–220.

[2] T. Kawahara, "Spoken dialogue system for a human-like conversational robot ERICA," in *Proc. IWSDS*, 2018, pp. 65–75.

[3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. ASLP*, vol. 22, no. 4, pp. 745–777, 2014.

[4] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *SIGKDD Explor. Newsl.*, vol. 19, no. 2, pp. 25–35, 2017.

[5] C. Dong, Y. Li, H. Gong, *et al.*, "A survey of natural language generation," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–38, 2022.

[6] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv*, 2021.

[7] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. ASLP*, vol. 29, pp. 132–157, 2020.

[8] Y. Ishikawa, K. Konaka, T. Nakamura, N. Takamune, and H. Saruwatari, "Real-time speech extraction using spatially regularized independent low-rank matrix analysis and rank-constrained spatial covariance matrix estimation," in *Proc. HSCMA*, 2024.

[9] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. ASLP*, vol. 17, no. 4, pp. 650–664, 2009.

[10] E. Lombard, "Le signe de l'elevation de la voix," *Annales Maladies de L'Oreille et du Larynx*, vol. 37, pp. 101–119, 1911.

[11] A. L. Pittman and T. L. Wiley, "Recognition of speech produced in noise," *JSLHR*, vol. 44, no. 3, pp. 487–496, 2001.

[12] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, pp. 1626–1641, 2016.

[13] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, "Blind speech extraction based on rank-constrained spatial covariance matrix estimation with multivariate generalized Gaussian distribution," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 1948–1963, 2020.

[14] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[15] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP JASP*, vol. 2003, no. 11, pp. 1–10, 2003.

[16] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," *IEICE Trans. Fundamentals*, vol. E87-A, pp. 1941–1948, 2004.

[17] Y. Mori, H. Saruwatari, T. Takatani, *et al.*, "Blind separation of acoustic signals combining SIMO-model-based independent component analysis and binary masking," *EURASIP JASP*, vol. 2006, no. 034970, pp. 1–17, 2006.

[18] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] S. Takamichi, R. Sonobe, K. Mitsui, *et al.*, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *AST*, vol. 41, no. 5, pp. 761–768, 2020.

[20] O. Take, S. Takamichi, K. Seki, Y. Bando, and H. Saruwatari, "SaSLaW: Dialogue speech corpus with audio-visual egocentric information toward environment-adaptive dialogue speech synthesis," in *Proc. INTERSPEECH*, 2024.

[21] Y. Ren, C. Hu, X. Tan, *et al.*, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.

[22] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020, pp. 17 022–17 033.

Inference Efficient Source Separation Using Input-dependent Convolutions

Shogo Seki and Li Li
CyberAgent, Inc., Japan
{seki_shogo,li_li}@cyberagent.co.jp

***Abstract*—This study proposes a large-capacity inference-efficient source separation model that considers input mixtures. Convolutional neural networks (CNNs) are fundamental for building robust separation models, and CNN-based separation models are attractive because of their lower computational complexity compared to recent competitive separation models. A premise of these models is that a single kernel is applied to the input of each convolution, different input mixtures such as those with speakers of the same or different genders are not considered, resulting in suboptimal separation performance. To overcome this problem, the proposed method employs an input-dependent convolution called conditionally parameterized convolution (CondConv) in CNN-based separation models. CondConv contains multiple kernels and generates an aggregated single kernel based on the input. This can increase the network capacity while maintaining the computational complexity during inference as a standard convolution, thereby enabling separation models to account for the input mixtures. Through experimental evaluations in a speech separation task, the proposed input-dependent convolution approach consistently improved several CNN-based separation models while maintaining a negligible increase in inference.**

I. Introduction

Source separation is a technique for separating the underlying source signals in an observed mixture signal and has various practical applications. With the continuous growth of deep neural networks (DNNs), significant progress has been made in single-channel source separation, in which the observed signals are obtained from a single microphone.

As one of the most successful separation models, Conv-TasNet [1] is a well-known approach that designs an encoder-separator-decoder framework and combines it with convolutional neural networks (CNNs), to achieve source separation in the time domain. Other successful separation models use dual-path modules as an efficient way to model long-form signals [2], and band-split architectures provide full-and sub-band modeling of signals [3]. These separation models were initially based on recurrent neural networks (RNNs), but have since been developed by introducing more sophisticated architectures, such as Transformers [4] and state-space models (SSMs) [5], [6], thereby achieving state-of-the-art performances [7]–[9]. Although outperforming the novel competitive separation models is no longer easy, CNN-based separation models [10], [11], including Conv-TasNet, remain attractive because of their lower computational complexity in developing lightweight and low-latency systems.

Convolutional layers play a pivotal role in CNN-based separation models, in which a single kernel is applied to individual inputs. However, a wide variety of source mixtures can be used as the input of separation models. For example, considering two speakers, mixed speech can be obtained from speaker pairs of the same gender (intra-gender) or different genders (inter-gender). In intra-gender speaker mixtures, the speakers in the mixture signal have closely overlapping pitch ranges, and convolution kernels are required to distinguish the differences. However, these kernels are inappropriate for inter-gender speaker mixtures because the components can be separated based on a single speaker. The converse can also occur. In such situations, convolutional layers cannot handle different types of input mixtures separately, failing to transform hidden representations, which can lead to suboptimal separation model performance.

To overcome this limitation, we propose the use of conditionally parameterized convolution (CondConv) [12], an input-dependent convolution approach [12]–[14] for CNN-based separation models. CondConv consists of multiple kernels and a routing module to generate linear combination coefficients based on the input. An aggregated single kernel is then generated as a linear combination of the multiple kernels and applied to the input. This enables the convolutional layers to use different kernels for each input, and the separation models are expected to effectively separate individual input mixtures. Furthermore, because CondConv uses a simple routing module and aggregates kernels rather than inputs, the computational cost is almost equivalent to that of a standard convolution, and a more extensive network capacity is retained.

II. Preliminaries: CNN-based separation models

To formulate the single-channel source separation problem addressed in this study, we provide an overview of the convolutional encoder-separator-decoder architecture. Because the proposed method is based on CNN-based separation models, we used Conv-TasNet to provide a simple description.

A. Problem Formulation

Let us suppose that a single microphone receives a mixed signal consisting of J source signals. Let $s_j[n]$ and $x[n]$ be the waveform samples of the j-th source signal and the observed signal, respectively, where $n \in \{1, \dots, N\}$ is the sample index in the time domain. The observed signal is expressed as:

$$x[n] = \sum\nolimits_{j=1}^{J} s_j[n]. \tag{1}$$

The single-channel source separation problem is the task of separating each source signal $s_j[n], j \in 1, \ldots, J$ given a single-channel mixture signal $x[n]$.

B. Encoder-Separator-Decoder Architecture

Before feeding into a convolutional encoder, the observed mixture signal is first divided into overlapped segments of length L, denoted as $\mathbf{x}(t) \in \mathbb{R}^L$, where $t \in \{1, \ldots, T\}$ represents the frame index, and T is the total number of segments.

Each mixture segment $\mathbf{x}(t)$ is encoded by a one-dimensional convolutional layer with a learnable weight $\mathbf{B} \in \mathbb{R}^{D \times L}$:

$$\mathbf{h}(t) = \mathcal{H}(\mathbf{B}\mathbf{x}(t)), \tag{2}$$

where $\mathcal{H}(\cdot)$ represents an arbitrary activation function such as a rectified linear unit (ReLU) [15].

Encoded representation $\mathbf{h}(t) \in \mathbb{R}^D$ is then fed into a separator network, such as a temporal convolutional network (TCN) [16] in Conv-TasNet, to estimate masks $\mathbf{m}_j(t) \in \mathbb{R}^D$. The TCN consists of repeated stacks of dilated convolutional blocks with exponentially increasing dilation factors. Each block is composed of three-layer convolutional networks, where the first and last layers are pointwise convolutions and the middle layer is a depthwise convolution with dilation[1]. Given an input $\mathbf{e}_{\text{input}}$, each block outputs the transformed representation $\mathbf{e}_{\text{output}}$:

$$\mathbf{e}_{\text{output}} = (\texttt{PConv} \circ g \circ \texttt{DConv} \circ g \circ \texttt{PConv})(\mathbf{e}_{\text{input}}), \tag{3}$$

where $\circ$ represents the composition of functions, and $g(\cdot)$ represents a composite function of parametric ReLU (PReLU) [17] and layer normalization [18]: $g(\mathbf{e}) = (\texttt{LayerNorm} \circ \texttt{PReLU})(\mathbf{x})$.

The latent representations of each source signal are predicted by: $\mathbf{v}_j(t) = \mathbf{m}_j(t) \odot \mathbf{h}(t)$, where $\odot$ represents element-wise multiplication. The estimated representation $\mathbf{v}_j(t) \in \mathbb{R}^D$ is decoded into a source segment with a one-dimensional transposed convolutional layer with a parameter $\mathbf{U} \in \mathbb{R}^{L \times D}$:

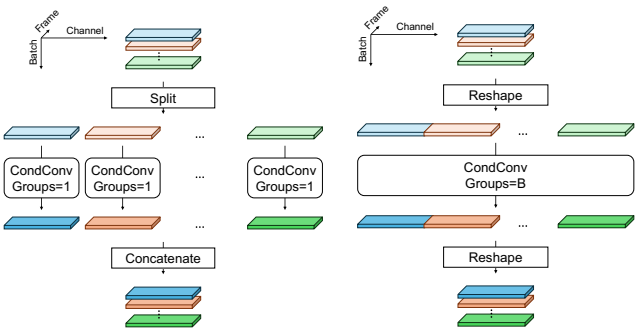
$$\hat{\mathbf{s}}_j(t) = \mathbf{U}\mathbf{v}_j(t), \tag{4}$$

Finally, separated signals are obtained by applying overlap-add to waveform segments.

III. Proposed Method

In the proposed method, CondConv was adapted to the conventional CNN-based separation models, further presenting an accelerating approach for GPU computing.

[1]Although the original Conv-TasNet contains a skip connection branch in each block [1], we omit it for simplicity. Indeed, we validated that the skip connection did not significantly improve performance in our preliminary experiments.



(a) Naïve convolution. (b) Group convolution.

Fig. 1: Comparison of algorithms using (a) naïve convolution and (b) group convolution.

A. Conditionally Parameterized Convolution (CondConv)

CondConv differs from standard convolution in that it contains multiple kernels and an additional routing module. Let x and $W_k, k \in \{1, \ldots, K\}$ be the input and k-th kernel of a CondConv, respectively. CondConv aggregates each kernel W_k to generate a single kernel through a linear combination:

$$f(x; W_1, \ldots, W_K) = \left(\sum\nolimits_k \alpha_k(x) W_k\right) * x, \tag{5}$$

where $\alpha_k(\cdot)$ is an input-dependent scalar weight computed from the routing module. K is the number of experts, and CondConv is equivalent to a mixture of experts (MoE) formulated as follows:

$$\left(\sum\nolimits_k \alpha_k(x) W_k\right) * x = \sum\nolimits_k \alpha_k(x) \left(W_k * x\right). \tag{6}$$

This implies that CondConv has the same capacity as the MoE model.

For the routing module, we designed a network composed of global average pooling, dropout, and fully-connected layers, obtaining the routing weights $\alpha_1(x), \ldots, \alpha_K(x)$ through a sigmoid function:

$$\alpha_1(x), \ldots, \alpha_K(x) = (\texttt{Sigmoid} \circ r)(x), \tag{7}$$

$$r(x) = (\texttt{FC} \circ \texttt{DropOut} \circ \texttt{GlobalAveragePool})(x). \tag{8}$$

Because the standard convolutional layers in CNN-based models can be replaced with CondConvs, the convolutional encoder in eq. (2) can be rewritten as follows:

$$\mathbf{h}(t) = \mathcal{H}\left(\sum\nolimits_k \left[\beta_k(\mathbf{x}(t))\mathbf{B}_k\right] \mathbf{x}(t)\right), \tag{9}$$

where, $\beta_k(\cdot)$ and $\mathbf{B}_k$ represent the k-th routing weight and learnable weight, respectively. Similarly, the convolutional decoder in eq. (4) can be reformulated as follows:

$$\hat{\mathbf{s}}_j(t) = \sum\nolimits_k \left[\gamma_k(\mathbf{x}(t))\mathbf{U}_k\right] \mathbf{v}_j(t), \tag{10}$$

where, $\gamma_k(\cdot)$ represents the k-th routing weight, and $\mathbf{U}_k$ denotes the k-th learnable weight of the convolutional decoder. For the separator, the convolutional layers can be replaced in a similar same manner.

B. Implementation

Notably, the routing weights in CondConv depend on the input samples, necessitating repeated convolution operations for minibatch data during training. One way to perform CondConv is to apply convolutions sample-by-sample; however, this was found to slow down training. To achieve fast GPU computing, a modified algorithm was developed using a group convolution.

Fig. 1 compares the algorithms using naïve and group convolutions. In the naïve convolution approach, the CondConv operation must be repeated using the same number of batch sizes, which is inefficient for GPU computing and results in bottlenecks during training.

Focusing on the fact that each CondConv operation in Fig. 1 (a) is independent of another, these operations can be regarded as a single convolution with minibatch size as groups (Fig. 1 (b)) [19], [20]. This implementation is beneficial for GPU acceleration, enabling fast training.

IV. Experimental Evaluation

A. Settings

The proposed approach was experimentally evaluated under a speaker separation scenario in which the task was to separate out two sources from single-channel mixture signals. We used the WSJ0-2mix dataset [21], which is a popular dataset for benchmarking monaural speaker separation systems. WSJ0-2mix uses samples from the WSJ0 corpus, comprising 20000, 5000, and 3000 two-speaker mixtures in the training, validation, and test sets, respectively. The speakers in the training and validation sets are disjoint from those in the test set. For each mixture, two utterances are overlapped, and mixed with at random signal-to-noise ratios (SNRs) in the range of $[-5, 5]$ dB. The sampling rate of each mixture was 8 kHz.

Selecting the CNN-based separation models: Conv-TasNet [1], SuDoRM-RF [10], and its lightweight variants (SuDoRM-RF++ and SuDoRM-RF++GC) [11] as the baseline models, we developed the proposed method by replacing the standard convolutional layers in these models with CondConv layers. SuDoRM-RF and its variants are efficient CNN-based separation models, where each separator uses stacks of U-net-like network architectures [22]. For the Conv-TasNet model, we used an `espnet` implementation[2] and training recipe. For the other models, we used the official implementation[3]. For the CondConv layers, the number of experts was set to four and the dropout rate was set to 0.2. All the models were trained using four-second segments and the Adam optmizer [23], with an initial learning rate of 0.001. The negative scale-invariant signal-to-distortion ratio (SI-SDR) [24] was used as the loss function, and utterance-wise permutation invariant training (PIT) was conducted to solve the permutation problems [25].

SI-SDR improved (SI-SDRi) was used as the metric to evaluate the separation performance. We also investigated the multiply-accumulate operations (MACs) for one-second inputs to estimate the computational efficiency and used the `torchprofile`[4].

TABLE I: Comparative studies for (a) modules, (b) network sizes, and (c) layers; the best separation performance is in bold.

(a) Comparison of encoder (Enc.), separator (Sep.), and decoder (Dec.) modules with (✓) or without (✗) CondConv.

Enc.	Sep.	Dec.	# Params [M]	# MACs [G/s]	SI-SDRi [dB]
✗	✗	✗	8.71	7.02	15.63
✗	✗	✓	8.73	7.02	16.21
✗	✓	✗	34.78	7.03	15.90
✗	✓	✓	34.80	7.03	16.07
✓	✗	✗	8.73	7.02	15.69
✓	✗	✓	8.74	7.02	16.29
✓	✓	✗	34.80	7.03	16.40
✓	✓	✓	34.81	7.03	**16.82**

(b) Comparison of separation models using standard convolutions (Conv) with large network size and using CondConv.

Method	# Params [M]	# MACs [G/s]	SI-SDRi [dB]
Conv	8.71	7.02	15.63
Conv 4x	34.84	28.07	15.85
CondConv	34.81	7.03	**16.82**

(c) Comparison of separation models using various CondConvs, where the values in parentheses represent the numbers of experts.

Method	# Params [M]	# MACs [G/s]	SI-SDRi [dB]
Conv	8.71	7.02	15.63
CondConv (1)	8.75	7.02	15.78
CondConv (2)	17.44	7.02	16.00
CondConv (4)	34.81	7.03	**16.82**
CondConv (8)	69.56	7.04	16.28
CondConv (16)	139.05	7.05	16.12

B. Results

First, the proposed CondConv approach using Conv-TasNet was investiated. TABLE I presents a comparison of the Conv-TasNet models using standard convolutions and CondConvs.

1) Comparison of module-level replacements: In our comprehensive evaluation, we examined the encoder, separator, and decoder by replacing the convolutional layers with CondConvs. Although replacing the encoder with CondConv had little effect, replacing the separator and decoder with CondConv was effective. Moreover, replacing all the modules was the most effective. Accordingly, we apply CondConv to all convolutional layers in the following evaluation.

2) Comparison with models of similar size: To clarify the performance compared with models of the same size, we compared the proposed CondConv approach with a large Conv-TasNet model. The conventional large Conv-TasNet performs better than small Conv-TasNet, but the computational cost is linearly increased. The proposed CondConv approach achieves not only better separation performance than large Conv-TasNet but also maintains computational costs similar to those of small Conv-TasNet.

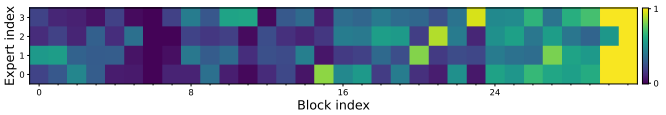
[2]https://github.com/espnet/espnet

[3]https://github.com/etzinis/sudo_rm_rf

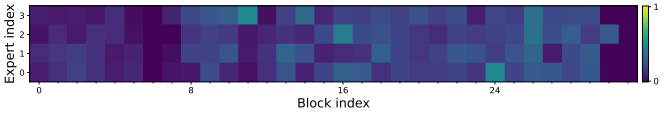
[4]https://github.com/zhijian-liu/torchprofile

TABLE II: SI-SDRi [dB] for mixtures of intra-gender and inter-gender speaker pairs. The values in brackets represent the improvement over baseline (Conv).

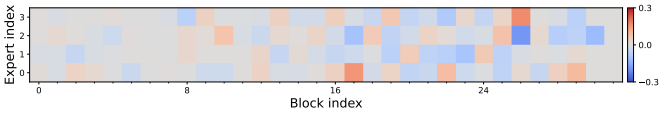
Method	Intra-gender	Inter-gender	Total
Conv	14.54 (±0.00)	16.69 (±0.00)	15.63 (±0.00)
CondConv	**16.05** (+1.51)	**17.56** (+0.87)	**16.82** (+1.19)



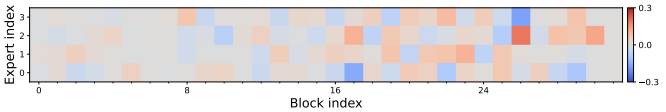
(a) Mean



(b) Standard deviation



(c) Mean difference of intra-gender speaker pairs from total pairs.



(d) Mean difference of inter-gender speaker pairs from total pairs.

Fig. 2: Visualization of routing weight statistics of the last CondConv layers at each TCN block: (a) mean, (b) standard deviation and mean differences of (c) intra-gender and (d) inter-gender speaker pairs from total pairs.

3) Comparison with the number of experts: We compared CondConvs with different numbers of experts, where the number of experts was set as $\{1, 2, 4, 8, 16\}$. Separation performance tended to improve as the number of experts increased. On the other hand, too many experts did not necessarily provide significant improvement. This may be because the network capacity becomes too large, complicating training. CondConv with four experts achieved the best results, and we used the same number of experts for subsequent experiments.

The improvement in separation performance provided by the proposed approach is analyzed next. TABLE II compares the separation performances of different types of two-speaker pairs. The proposed approach consistently improves performance. When comparing different speaker pairs, the proposed method contributes more to improving performance in intra-gender cases.

A visualization of the routing weights for each TCN block in Conv-TasNet using CondConv is shown in Fig. 2, where the horizontal axis represents the block index and the vertical axis represents the expert index. In the shallow block, the routing weights have a small standard deviation, and similar weights were used by all experts, which is consistent with the previous finding that replacing the encoder with CondConv results in little improvement. The routing weights tended to vary in deeper blocks.

TABLE III: Comparison with other separation models

Method	# Params [M]	# MACs [G/s]	SI-SDRi [dB]
DPRNN [2]	2.6	42.2	18.8
SepFromer [26]	26.0	59.5	20.4
TF-GridNet [7]	14.5	231.1	23.5
SPMamba [8]	6.1	N/A	†22.5
DPMamba [9]	59.8	N/A	24.4
Conv-TasNet			
Conv	8.71	7.02	15.63
CondConv	34.81	7.03	**16.82**
SuDoRM-RF			
Conv	2.66	10.60	16.70
CondConv	10.27	10.61	**16.80**
SuDoRM-RF++			
Conv	2.72	9.11	16.29
CondConv	10.59	9.12	**16.33**
SuDoRM-RF++GC			
Conv	0.30	3.33	13.19
CondConv	1.14	3.33	**14.27**

†`https://github.com/JusperLee/SPMamba`

We compared the routing weights for mixtures of intra-gender and inter-gender speaker pairs (Fig. 2 (c) and (d)). As described above, the routing weights in shallow blocks show similar activation patterns in both intra- and inter-gender cases. Meanwhile, the routing weights in the deeper blocks become active and vary depending on the expert. Furthermore, some of the experts behave complementarily depending on the type of speaker pairs. This demonstrates that the proposed CondConv approach considers input mixtures.

Finally, the proposed approach was evaluated for different CNN-based methods and compared it with novel competitive separation models; the results are shown in TABLE III. Recent state-of-the-art separation models achieve sufficient separation performances. In contrast, CNN-based separation models incur at least four times lower computational costs than these powerful models. We also found that the proposed CondConv approach consistently achieved better performance in all CNN-based models although the performance improvements depended on the separation models. Hence, the proposed approach has significant potential to improve existing separation models.

V. Conclusions

This study presented an input-dependent convolution approach for CNN-based source separation models. The proposed method employs CondConv and replaces the standard convolutional layers in the separation models, enabling the models to increase the network capacity while maintaining the computational complexity during inference. In the experimental evaluation of a two-speaker separation task on the WSJ0-2mix, the proposed method consistently improved the performance of various separation models, demonstrating that the input types of mixtures are considered during separation.

[1] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[2] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 46–50.

[3] Y. Luo and J. Yu, "Music Source Separation With Band-Split RNN," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.

[4] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is All You Need," in *Advances in Neural Information Processing Systems*, 2017.

[5] A. Gu, K. Goel, and C. Re, "Efficiently Modeling Long Sequences with Structured State Spaces," in *International Conference on Learning Representations*, 2022.

[6] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[7] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating Full- and Sub-Band Modeling for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[8] K. Li and G. Chen, "SPMamba: State-space model is all you need in speech separation," *arXiv preprint arXiv:2404.02063*, 2024.

[9] X. Jiang, C. Han, and N. Mesgarani, "Dual-path Mamba: Short and Long-term Bidirectional Selective Structured State Space Models for Speech Separation," *arXiv preprint arXiv:2403.18257*, 2024.

[10] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -RF: Efficient Networks for Universal Audio Source Separation," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2020, pp. 1–6.

[11] E. Tzinis, Z. Wang, X. Jiang, and P. Smaragdis, "Compute and Memory Efficient Universal Sound Source Separation," *Journal of Signal Processing Systems*, vol. 94, no. 2, pp. 245–259, 2022.

[12] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "CondConv: Conditionally Parameterized Convolutions for Efficient Inference," in *Advances in Neural Information Processing Systems*, 2019.

[13] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic Convolution: Attention Over Convolution Kernels," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 030–11 039.

[14] C. Kosma, G. Nikolentzos, and M. Vazirgiannis, "Time-Parameterized Convolutional Neural Networks for Irregularly Sampled Time Series," *arXiv preprint arXiv:2308.03210*, 2023.

[15] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[16] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[18] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[19] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5987–5995.

[20] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[21] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 31–35.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[23] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[24] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 626–630.

[25] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[26] C. Subakan, M. Ravanelli, S. Cornell, F. Grondin, and M. Bronzi, "Exploring self-attention mechanisms for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2169–2180, 2023.

Relative Transfer Matrix for Drone Audition Applications: Source Enhancement

Wageesha N. Manamperi, Thushara D. Abhayapala
The Australian National University, Canberra, Australia
E-mail: {*wageesha.manamperi, thushara.abhayapala*}*@anu.edu.au*

***Abstract*—The Relative Transfer Matrix (ReTM), recently introduced as a generalization of the relative transfer function for multiple sources and multiple microphones, shows promising performance when applied to speech denoising and speaker separation in a noisy reverberant room. This work utilizes the ReTM to propose a novel framework for drone noise suppression. Difficulties in noise cancellation for drone audition applications arise due to its self-generated noise that causes an adverse noisy environment of low Signal-to-Drone Noise Ratio (SDNR) levels. In this paper, we divide the drone on-board microphones into two multichannel groups to approximately estimate the drone noise from one group to the other with known drone noise ReTM for denoising. We demonstrate the ReTM spatial mapping ability in both indoor and outdoor experiments using hovering drone on-board microphone recordings with low average magnitude spectrum error. Finally, we validate the method in a real-life environment for source signal enhancement over different $\overline{\text{SDNR}}$ conditions and offer both improved speech intelligibility and signal-to-distortion ratio.**

I. Introduction

Making drones hear is a fundamental requirement of drone audition functions, such as sound source localization [1]–[5], source separation [6], source tracking [7], [8], and signal enhancement [9]–[13]. However, (i) its self-generated noise, and (ii) on-board microphones being closer to the drone noise sources compared to the desired sound source on the ground, cause a highly adverse noisy environment with an extremely low signal-to-drone noise ratio ($\overline{\text{SDNR}}$) (defined as the power ratio between the source signal and the drone noise) level [14]. As a result, drone noise reduction is in itself an active problem in acoustic signal processing that enables drone audition applications, e.g., search and rescue missions.

Many early drone noise reduction approaches are based on the beamforming [9], [15]–[19]. Other popular algorithms are spatial filtering-based methods [20], [21], blind source separation [20], [22], [23], spherical sector harmonics-based methods [24], as well as supervised approaches including template-based approaches [25], correlation matrix-based methods [26], and reference-based methods [27], [28]. Recently favored are machine learning-based techniques, the well-known form of supervised approaches, for drone noise suppression [29]–[32]. For example, a denoising autoencoder based on fully convolutional neural networks to cancel drone noise was discussed in [32]. While these methods suppress the drone noise in low $\overline{\text{SDNR}}$ levels, they suffer from difficulties in obtaining an accurate estimate of the drone noise signals from on-board multichannel recordings due to the contribution of the noise source signals. This paper utilizes the recently proposed Relative Transfer Matrix (ReTM) [33], which is defined as the generalization of the relative transfer function for multiple simultaneous sound sources and multiple microphones, to propose a signal-independent solution that solely depends on the drone noise sources position in a static acoustic environment for drone noise cancellation.

ReTM is a new spatial feature containing relevant cues of the multiple sound sources, and independent of the emitted source signals, is defined by allocating the receivers into two multi-microphone groups [33]. Related works [34]–[36] highlight the promising performances of the ReTM methods when applied to speech enhancement, and speaker separation in a multi-source noisy reverberant environment.

In this paper, we present a novel algorithm for drone noise reduction at very low $\overline{\text{SDNR}}$ levels using the ReTM. The derivation of this method is similar to [34] in a room scenario that exploits covariance matrices of the multiple noise sources-only signals by dividing the receivers into two multi-microphone groups based on the ReTM presented in [33]. Compared to [34], the proposed algorithm is (i) applied for drone noise estimation, (ii) analyzed the estimation accuracy in terms of the average magnitude spectrum error using two drone acoustic datasets, i.e., DREGON [37] and AVQ dataset [11] in indoor and outdoor environments, respectively, and (iii) presented the speech denoising performance at extreme $\overline{\text{SDNR}}$ conditions (< -10 dB). We verify the proposed method's applicability for source enhancement via a real-life experiment conducted outdoors, and show increased speech intelligibility and signal-to-distortion ratio (SDR) over differing $\overline{\text{SDNR}}$ level.

II. Problem Formulation

Let there be Q microphones mounted on a drone capturing the sound produced by a target speech source. The signal received by the qth microphone due to the source in time domain is

$$m_q(n) = d_q(n) * s(n) + \sum_{\ell=1}^{\mathcal{L}} h_{q,\ell}(n) * v_\ell(n), \qquad (1)$$

where $d_q(n)$ is the drone-related impulse response from the sound source to the qth microphone, $s(n)$ is the source signal, $v_\ell(n)$ is the drone noise due to the ℓth motor, $h_{q,\ell}(n)$ is the

Thanks to Australian Research Council (ARC) Discovery Project Grant DP200100693 for funding.

impulse response from the ℓth motor to the qth microphone and $\mathcal{L}$ is the number of motors on the drone and '$*$' is the convolution operator.

In the short-time Fourier transform (STFT) domain, (1) can be written as

$$M_q(f,t) = d_q(f)S(f,t) + \sum_{\ell=1}^{\mathcal{L}} H_{q,\ell}(f)V_\ell(f,t), \quad (2)$$

where $M_q(f,t)$, $S(f,t)$, and $V_\ell(f,t)$ are the STFTs of $m_q(n)$, $s(n)$ and $v_\ell(n)$, respectively, $d_q(f)$ is the drone-related transfer function [38] for the qth microphone, $H_{q,\ell}(f)$ is the noise transfer function and t and f are the frame and frequency indices, respectively with $t \in \{1,\dots,T\}$ and $f \in \{1,\dots,F\}$ where T and F are the number of time frames and frequency bins, respectively. We simplify (2) further and can be expressed in matrix form as

$$\mathbf{M}(f,t) = \mathbf{d}(f)S(f,t) + \mathbf{H}(f)\mathbf{V}(f,t), \quad (3)$$

where $\mathbf{M}(f,t) = [M_1(f,t),\dots,M_Q(f,t)]^T$ and $\{\cdot\}^T$ is the matrix transpose, and $\mathbf{d}(f)$, is the $[Q \times 1]$ vector of drone-related transfer functions, $\mathbf{H}(f)$ is the Q by $\mathcal{L}$ drone noise sources transfer function matrix, and $\mathbf{V}(f,t)$ are the $\mathcal{L}$ drone noise source signals that similarly defined.

The problem discussed in this paper is to suppress the drone noise by estimating the relative transfer matrix (ReTM) of the $\mathcal{L}$ drone noise sources from the received microphone signals, which we explain next.

III. Drone Noise Estimation and Reduction with Relative Transfer Matrix

This section uses the ReTM to propose a drone noise reduction algorithm. We first briefly review the ReTM presented in [33], which divides the multiple microphones into two groups to relate the signal received at the first microphone group to the second. We then blindly estimate the drone noise sources ReTM using covariance matrices. Finally, we utilize the drone noise ReTM to map the drone noise from one group to the other for denoising.

A. Drone Noise Sources Relative Transfer Matrix

Consider two microphone groups, $\{A\}$ and $\{B\}$ with Q_A and Q_B microphones, respectively ($Q = Q_A + Q_B$). Let $\mathbf{M}_A(f,t) = [M_1^{(A)}(f,t),\cdots,M_{Q_A}^{(A)}(f,t)]^T$, and $\mathbf{M}_B(f,t) = [M_1^{(B)}(f,t),\cdots,M_{Q_B}^{(B)}(f,t)]^T$ denote the vector of microphone signals for each group. The received signals at each microphone group in matrix form can be written as

$$\mathbf{M}_A(f,t) = \mathbf{d}_A(f)S(f,t) + \mathbf{H}_A(f)\mathbf{V}(f,t), \quad (4)$$

$$\mathbf{M}_B(f,t) = \mathbf{d}_B(f)S(f,t) + \mathbf{H}_B(f)\mathbf{V}(f,t). \quad (5)$$

The drone noise sources ReTM, $\boldsymbol{\mathcal{R}}_{AB}(f)$, can be defined as in [33]

$$\boldsymbol{\mathcal{R}}_{AB}(f) \triangleq \mathbf{H}_A(f)\mathbf{H}_B(f)^\dagger, \quad (6)$$

where $(\cdot)^\dagger$ is Moore-Penrose inverse, assuming the validity, i.e., $Q_B \geq \mathcal{L}$.

Consider the received signals by microphone groups when the speech is inactive ($S(f,t) = 0$), then the drone noise-only signals can be obtained from the received signals. Following the ReTM mapping as in [33], we can relate the drone noise signals received by microphone group $\{A\}$ with ReTM as

$$\mathbf{M}_A^{(v)}(f,t) = \boldsymbol{\mathcal{R}}_{AB}(f)\mathbf{M}_B^{(v)}(f,t). \quad (7)$$

Note that $\boldsymbol{\mathcal{R}}_{AB}(f)$ is defined by the spatial properties of the noise sources such that it is independent of the sound emitted by the drone motors and propellers. In drone audition applications, the $\boldsymbol{\mathcal{R}}_{AB}(f)$ is constant for a drone hovering scenario.

We note that it is typically straightforward to estimate the ReTM in (6) by assuming that segments of recording containing the drone noise-only signal when the drone flying in the hovering manoeuvre above the ground prior to the speech source being active, which we present in next subsection.

B. Blind Estimation of the ReTM from Received Signals

Here, we process the received signals in (4), and (5) when the speech source is inactive to estimate the drone noise sources ReTM that will be the building block of our proposed drone noise reduction algorithm.

Formatting the ReTM estimation using covariance matrices of

$$\begin{aligned} \boldsymbol{\mathcal{P}}_{AA}^{(v)}(f) &\triangleq E\{\mathbf{M}_A^{(v)}(f,t)\mathbf{M}_A^{(v)^*}(f,t)\}, \\ \boldsymbol{\mathcal{P}}_{BA}^{(v)}(f) &\triangleq E\{\mathbf{M}_B^{(v)}(f,t)\mathbf{M}_A^{(v)^*}(f,t)\}, \end{aligned} \quad (8)$$

where $E\{\cdot\}$ denotes the expectation which can be obtained by averaging across the time frames, and then with further simplifications as in [33], we approximate

$$\boldsymbol{\mathcal{R}}_{AB}(f) \approx \boldsymbol{\mathcal{P}}_{AA}^{(v)}(f)\left(\boldsymbol{\mathcal{P}}_{BA}^{(v)}(f)\right)^\dagger. \quad (9)$$

We observe from (9) that the drone noise sources ReTM ($\boldsymbol{\mathcal{R}}_{AB}(f)$) only relies on the estimation of the covariance matrices of the drone noise-only signals (no active speech), and does not require the number of noise sources in the mixture. The common approach is to obtain the noise-only segments using a Voice Activity Detection (VAD) algorithm, however, due to drones being loud these algorithms are not recommended. In practice, the drone noise-only segments can be isolated by initiating to activate the speech source after the drone broadcasts a specific signal, such as a tone. Although blind estimation of ReTM is used in this paper, the proposed method can also be implemented by pre-trained or semi-blind estimation.

The next section proposes how to suppress the drone noise to enhance the target speech captured by a hovering drone.

C. Drone Noise Suppression using Drone Noise Sources ReTM

Assuming that the drone noise sources ReTM, $\boldsymbol{\mathcal{R}}_{AB}(f)$ is accurately estimated, thus, we can approximately calculate the drone noise signals at microphone group $\{A\}$ by following the ReTM mapping in (7), we approximate that

$$\boldsymbol{\mathcal{R}}_{AB}(f)\mathbf{M}_B^{(v)}(f,t) \approx \mathbf{H}_A(f)\mathbf{V}(f,t). \quad (10)$$

We can subtract (10) from the received signals/noisy speech at the microphone group $\{A\}$ to remove the drone noise from the group $\{A\}$ microphone signals as

$$\mathbf{M}_A(f,t) - \boldsymbol{\mathcal{R}}_{AB}(f)\mathbf{M}_B(f,t) \approx \mathbf{d}_A(f)S(f,t), \qquad (11)$$

thus, enhancing the target speech. However, in practice, the approximation of (10) is not accurate. To make the exposition concise, we omit the dependency of time (t) and frequency (f) in the rest of this section.

However, we accomplish a 'distorted' version of the target speech signal S by expanding the left side of (11), such that

$$\begin{aligned}
\hat{\mathbf{S}} &= \mathbf{M}_A - \boldsymbol{\mathcal{R}}_{AB}\mathbf{M}_B, \\
&= \mathbf{d}_A S + \mathbf{H}_A\mathbf{V} - \boldsymbol{\mathcal{R}}_{AB}(\mathbf{d}_B S + \mathbf{H}_B\mathbf{V}), \\
&= \mathbf{d}_A S + \mathbf{H}_A\mathbf{V} - \boldsymbol{\mathcal{R}}_{AB}\mathbf{d}_B S - \underbrace{\boldsymbol{\mathcal{R}}_{AB}\mathbf{H}_B}_{\mathbf{H}_A}\mathbf{V}, \\
&= \mathbf{d}_A S + \mathbf{H}_A\mathbf{V} - \boldsymbol{\mathcal{R}}_{AB}\mathbf{d}_B S - \mathbf{H}_A\mathbf{V}, \\
&= \mathbf{d}_A S - \boldsymbol{\mathcal{R}}_{AB}\mathbf{d}_B S, \\
&= \underbrace{[\mathbf{d}_A - \boldsymbol{\mathcal{R}}_{AB}\mathbf{d}_B]}_{\text{distortion}} S, \qquad (12)
\end{aligned}$$

where $\hat{\mathbf{S}}$ is a $Q_A \times 1$ vector consists Q_A copies of estimated target speech signal S. We obtain a complete suppression of the drone noise from the received signals in (12), along with an accurate estimate of drone noise source ReTM, $\boldsymbol{\mathcal{R}}_{AB}$. However, we note that this is a 'distorted' version of the target speech signal S in terms of both the drone noise sources transfer matrix and transfer function of the speech source, but in future, the ReTM may be exploited using multiple copies of denoised speech to do further enhancement.

IV. Experiments

In this section, we illustrate the proposed method's performance using real-life multichannel recordings of a hovering drone from two publicly available drone acoustic datasets; (i) DREGON [37], and (ii) AVQ [11]. The experimental procedures and results are described. First, the DREGON dataset was used to measure the drone noise estimation performance in an indoor environment. Second, the AVQ dataset was examined to analyze the algorithm performance in an outdoor environment.

The proposed method's performance is analyzed using a qualitative metric in terms of the average magnitude spectrum error of the first microphone in group $\{A\}$ as in [33]

$$\epsilon_1(f) = \underset{t}{\text{mean}}\, 10\log_{10}\frac{|\hat{M}_1^{(A)}(f,t) - M_1^{(A)}(f,t)|^2}{|M_1^{(A)}(f,t)|^2},$$

where $\hat{M}_1^{(A)}(f,t)$ denotes the estimated drone noise signal at the first microphone of the group $\{A\}$ calculated using (7) with the drone noise sources ReTM, which we will refer to as *'remote'* signals from here on. We note that microphones are randomly grouped as group $\{A\}$ and $\{B\}$, such that no microphone picks twice [33]. But in the future, we discuss the microphone group selection to provide a complete analysis of the performance variation.

A. Indoor Environments

1) Experimental Recordings: We used *hovering* audio signals ('DREGON_hovering_nosource_room2.wav.') of the '*In Flight Noise-Only Recordings*' from the DREGON dataset in [37] conducted in a $10 \times 10 \times 2.5$ m rectangular room ($T_{60} = 150$ ms). These recordings were made with a flying drone without any speech sources. The DREGON dataset has a constellation of 8 microphones in a cubic-shaped structure placed below the motor-propeller plane of the drone. We assigned 4 microphones on the bottom (channels $\{0, 2, 4, 6\}$) to group $\{A\}$, and 4 microphones on the top (channels $\{1, 3, 5, 7\}$) to group $\{B\}$. The audio recordings were processed directly in the short-time Fourier domain for a 2^{18} window size, 44.1 kHz sampling, and 10 second duration that was long enough to satisfy the multiplicative transfer function [39].

2) Results: We discuss the magnitude spectrum error of the drone noise estimation using the ReTM in Fig. 1 and reconstruction of the drone noise signals at microphone group $\{A\}$ in Fig. 2. Fig. 1 shows the magnitude spectrum of the measured signal $M_1^{(A)}$, and remote signal $\hat{M}_1^{(A)}$ for a single time frame as well as their error averaged over all time frames. We observe that there is no difference between $M_1^{(A)}$ and $\hat{M}_1^{(A)}$, which is equivalent to being a perfect mapping. This is further confirmed by a very low error observed throughout the frequency band of $20 - 16$k Hz. The results suggest that the drone noise sources ReTM exactly map the drone noise source's spatial properties and accurately estimate the drone noise signals at the microphone channel 0.

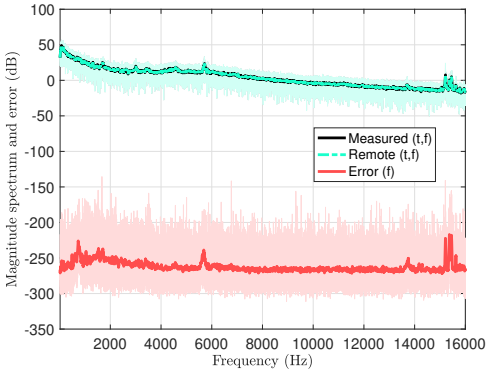


Fig. 1. Magnitude spectrum and error averaged over time for in flight noise-only recordings from the DREGON dataset [37].

We present both time domain and spectrogram plots of the measured and remote signals in Fig. 2. Strong similarity is once again shown for hovering recordings confirming the accuracy of the estimated drone noise sources ReTM. We share a link to the audio files on GitHub[1].

B. Outdoor Environments

1) Experimental Recordings: We utilized an AVQ dataset [11] recorded outdoors from a fixed drone position on a

[1]https://github.com/wnilmini/ReTM_Source_Enhancement/tree/main/Drone%20noise%20Estimation

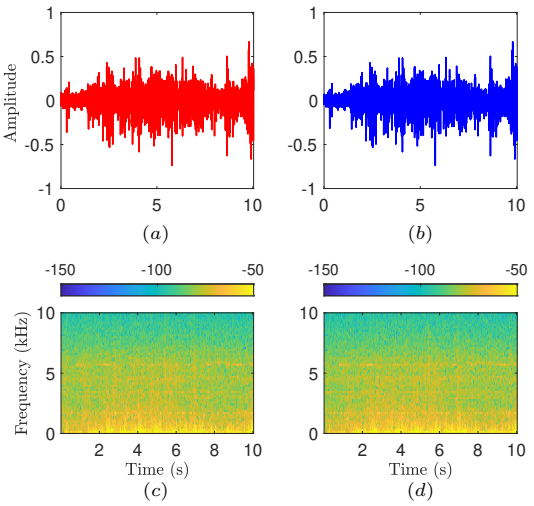


Fig. 2. Measured and remote drone noise signals at channel 0 at microphone group $\{A\}$: (a,c) measured signals, (b,d) remote signals, (a,b) time domain plots, and (c,d) spectrogram plots.

tripod at the height of 1.8 m with two static speech sources (scenario: 'subset 1'). We selected microphone recordings from the *'sequence 2'*, drone noise-only signals with a constant speed of 100% for ReTM mapping. The AVQ dataset has 8 microphones in a circular array located 15 cm above the body of the drone. We assigned 4 channels $\{4, 8, 1, 6\}$ to group $\{A\}$, and 4 channels $\{3, 5, 7, 2\}$ to group $\{B\}$. The 30 second long audio recordings were processed directly in the short-time Fourier domain for a 2^{20} window size, 44.1 kHz sampling rate.

2) Results: Fig. 3 displays the magnitude spectrum error of the drone noise estimation using the ReTM for the outdoor environment. A similar result is also obtained for the AVQ dataset as the DREGON dataset despite being outdoors, supporting that the drone noise sources ReTM is modeled correctly. As intended, both time domain and spectrogram plots of the remote signals in Fig. 4 are also seen to be an accurate estimate of the measured signals.

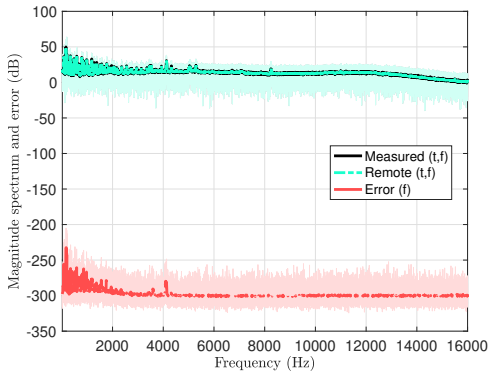


Fig. 3. Magnitude spectrum and error averaged over time for in flight noise-only recordings from the AVQ dataset [11].

These results suggest that the drone noise at microphone group $\{A\}$ can be accurately estimated using the drone noise at microphone group $\{B\}$ with the drone noise sources ReTM in both the indoor and outdoor environments. This method can be easily generalized to different types of drones using the drone noise-only recordings obtained when the drone is stationary e.g., hovering manoeuvre. In the next section, ReTM-based drone noise reduction algorithm will perform for a target speech enhancement.

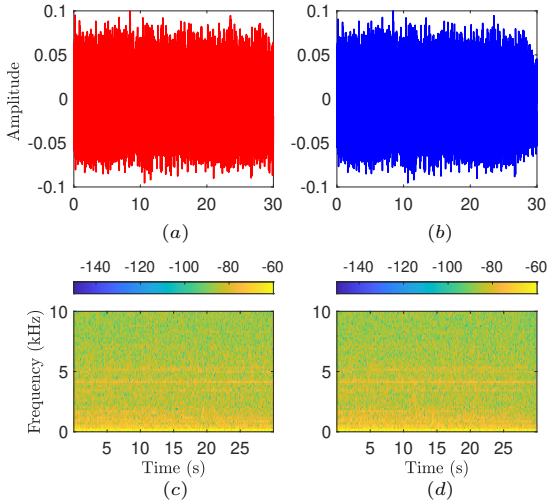


Fig. 4. Measured and remote drone noise signals at channel 4 at microphone group $\{A\}$: (a,c) measured signals, (b,d) remote signals, (a,b) time domain plots, and (c,d) spectrogram plots.

V. Application in Source Signal Enhancement

This section briefly examines the proposed drone noise reduction algorithm in an outdoor environment in the presence of a speech source at very low $\overline{\text{SDNR}}$ conditions (< -10 dB). Our intention is to analyze the speech enhancement performance over different $\overline{\text{SDNR}}$ levels in terms of signal distortion and speech intelligibility. We evaluate the quality of the enhanced speech using objective measures, (i) Signal-to-Distortion Ratio (SDR), using the BSS-Eval toolbox [40] and (ii) the Short-Time Objective Intelligibility score (STOI) [41]. We want to iterate that we do not consider comparing the proposed algorithm with baseline methods in this paper as our main focus here is to quantify the signal distortion introduced by the ReTM, in such a case, we expect that the proposed method can be performed lower STOI value than the state-of-the-art methods as in [10].

We used the AVQ dataset in [11] for the evaluation. The multichannel time-domain recordings are produced by adding the drone noise signals recorded at the microphone array separately (discussed in Section IV-B) with speech-only recordings from the '*sequence 4*' of the same subset. Then, the speech-only recordings are scaled to simulate different $\overline{\text{SDNR}}$ levels from -30 dB to -15 dB with an increment of 5 dB. We also used the same parameter settings and microphone group selection as in Section IV-B.

Fig. 5 shows the time-domain and spectrogram plots of the recorded and enhanced signals at the channel 4 (remote channel) of the microphone group $\{A\}$ at $\overline{\text{SDNR}}$ level of -20 dB. Immediately we observe that the proposed method significantly reduced the strong drone noise in the mixture signal (Fig. 5b & d). Additionally, a clear preservation of the speech spectrum is observed at the lower frequencies in Fig. 5d. We share a link to the enhanced signals[2].

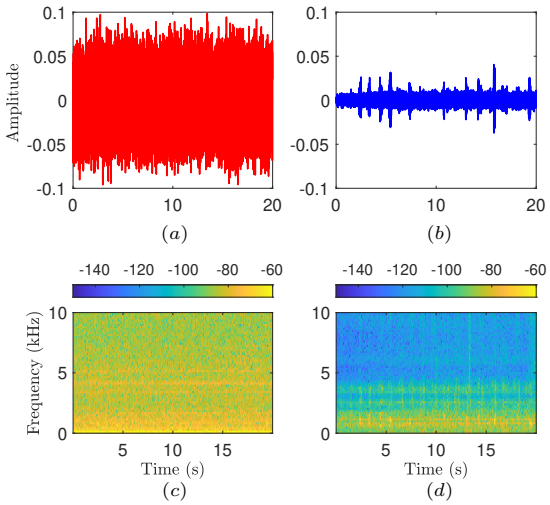


Fig. 5. Recorded and enhanced signals at channel 4 at microphone group $\{A\}$ at $\overline{\text{SDNR}}$ level of -20 dB: (a,c) recorded signals, (b,d) enhanced signals, (a,b) time domain plots, and (c,d) spectrogram plots.

TABLE I
SIGNAL ENHANCEMENT PERFORMANCE EVALUATION USING IMPROVEMENTS IN SDR AND INPUT/OUTPUT STOI FOR VARIOUS $\overline{\text{SDNR}}$ LEVELS

$\overline{\text{SDNR}}$ level	SDR (dB)	STOI (%)	
		Input	Output
-15 dB	2.04	20.55	47.86
-20 dB	-5.19	19.44	43.08
-25 dB	-6.88	17.14	40.15
-30 dB	-8.73	15.75	36.43

We will now examine the speech enhancement performance with the proposed drone noise reduction algorithm over -30 dB and -15 dB $\overline{\text{SDNR}}$ range. Table I depicts the improvements in SDR and input/output STOI scores as a function of $\overline{\text{SDNR}}$ levels. We observe that both the SDR and STOI performance have degraded with worsening $\overline{\text{SDNR}}$ condition. However, the STOI improvement is nearly double compared to the input scores over all $\overline{\text{SDNR}}$ levels. Again, SDR is slightly better with higher $\overline{\text{SDNR}}$ levels. Here we reiterate that this is due to distortion introduced with the drone noise sources ReTM in (12). In contrast, from the informal listening, we find that the enhanced speech signals are able to greatly increase the understandability of the speech content compared to the recorded signals. The results confirm that the proposed method better enhance the source signal in extreme noisy environments although both STOI and SDR are gradually degrading with decreased $\overline{\text{SDNR}}$ level.

[2]https://github.com/wnilmini/ReTM_Source_Enhancement/tree/main/Source%20Enhancement

VI. CONCLUSION

This paper proposed a novel algorithm to reduce the drone noise for drone audition applications using ReTM. We separate the multiple microphones mounted on a drone into two multi-microphone groups to estimate the drone noise from one group to the second. Extensive experimental studies using both indoor and outdoor real-life datasets confirmed the accuracy of the proposed method. This was shown by the very lower magnitude spectrum error obtained for the full frequency band. STOI and SDR values were shown to be improved for speech enhancement at very low $\overline{\text{SDNR}}$ levels in outdoor environments. In the future, the ReTM may be exploited using multiple copies of enhanced speech together filtering approaches to minimize the spectral distortion further.

REFERENCES

[1] M. Wakabayashi, H. G. Okuno, and M. Kumon, "Drone audition listening from the sky estimates multiple sound source positions by integrating sound source localization and data association," *Adv. Robot.*, vol. 34, no. 11, pp. 744–755, Jun. 2020.

[2] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and K. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, Sep. 2014, pp. 1902–1907.

[3] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 30, pp. 508–519, Jan. 2022.

[4] K. Hoshiba, K. Washizaki, M. Wakabayashi, *et al.*, "Design of uav-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, p. 2535, 2017.

[5] B. Yen and Y. Hioka, "Noise power spectral density scaled snr response estimation with restricted range search for sound source localisation using unmanned aerial vehicles," *EURASIP J. on Audio, Speech, and Music Process.*, no. 1, pp. 1–26, Dec. 2020.

[6] T. Morito, O. Sugiyama, R. Kojima, and K. Nakadai, "Partially shared deep neural network in sound source separation and identification using a uav-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, Oct. 2016, pp. 1299–1304.

[7] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Sound source tracking by drones with microphone arrays," in *Proc. IEEE/SICE Int. Symp. on Syst. Integr.*, Jan. 2020, pp. 796–801.

[8] L. Wang, R. Sanchez-Matilla, and A. Cavallaro, "Tracking a moving sound source from a multi-rotor drone," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, Oct. 2018, pp. 2511–2516.

[9] Y. Hioka, M. Kingan, G. Schmid, and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," in *Proc. IEEE Int. Workshop on Acoust. Signal Enhancement*, Sep. 2016, pp. 1–5.

[10] W. N. Manamperi, T. D. Abhayapala, P. N. Samarasinghe, and J. A. Zhang, "Drone audition: Audio signal enhancement from drone embedded microphones using multichannel wiener filtering and gaussian-mixture based post-filtering," *Appl. Acoust.*, vol. 216, p. 109 818, Jan. 2024.

[11] L. Wang, R. Sanchez-Matilla, and A. Cavallaro, "Audio-visual sensing from a quadcopter: Dataset and baselines for source localization and sound enhancement," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, Nov. 2019, pp. 5320–5325.

[12] W. N. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Improved GMM wiener filtering approach for signal enhancement," in *Proc. INTER-NOISE and NOISE-CON Congr. and Conf.*, Aug. 2023, pp. 3584–3594.

[13] B. Yen, Y. Hioka, G. Schmid, and B. Mace, "Multi-sensory sound source enhancement for unmanned aerial vehicle recordings," *Appl. Acoust.*, vol. 189, no. 108590, Feb. 2022.

[14] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Estimating drone motor related acoustic transfer function: A preliminary investigation," in *Proc. Asia-Pacific Signal and Inf. Process. Assoc. Annu. Summit and Conf.*, Dec. 2020, pp. 156–160.

[15] T. Ishiki and M. Kumon, "A microphone array configuration for an auditory quadrotor helicopter system," in *Proc. IEEE Int. Symp. on Safety, Security, and Rescue Robot.*, Oct. 2014, pp. 1–6.

[16] T. Ishiki and M. Kumon, "Design model of microphone arrays for multirotor helicopters," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, Sep. 2015, pp. 6143–6148.

[17] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Beamforming-based acoustic source localization and enhancement for multirotor UAVs," in *Proc. Eur. Signal Process. Conf.*, Sep. 2018, pp. 987–991.

[18] P. Misra, A. A. Kumar, P. Mohapatra, and P. Balamuralidhar, "Aerial drones with location-sensitive ears," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 154–160, Jul. 2018.

[19] Y. Hioka, M. Kingan, G. Schmid, R. McKay, and K. A. Stol, "Design of an unmanned aerial vehicle mounted system for quiet audio recording," *Appl. Acoust.*, vol. 155, no. 7, pp. 423–427, Dec. 2019.

[20] L. Wang and A. Cavallaro, "Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles," *IEEE Sensors J.*, vol. 17, no. 8, pp. 2447–2455, Feb. 2017.

[21] L. Wang and A. Cavallaro, "Acoustic sensing from a multirotor drone," *IEEE Sensors J.*, vol. 18, no. 11, pp. 4570–4582, Apr. 2018, ISSN: 1530-437X.

[22] L. Wang and A. Cavallaro, "Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles," in *Proc. IEEE Int. Conf. on Adv. Video and Signal Based Surveillance*, Aug. 2016, pp. 152–158.

[23] L. Wang and A. Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 28, pp. 2523–2537, Aug. 2020.

[24] H. Bi, F. Ma, T. D. Abhayapala, and P. N. Samarasinghe, "Spherical sector harmonics based directional drone noise reduction," in *Proc. IEEE Int. Workshop on Acoust. Signal Enhancement*, Sep. 2022, pp. 1–5.

[25] P. Marmaroli, X. Falourd, and H. Lissek, "A UAV motor denoising technique to improve localization of surrounding noisy aircrafts: Proof of concept for anti-collision systems," in *Proc. Acoust.*, Apr. 2012, pp. 1–6.

[26] K. Furukawa, K. Okutani, K. Nagira, *et al.*, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, Nov. 2013, pp. 3943–3948.

[27] S. Yoon, S. Park, Y. Eom, and S. Yoo, "Advanced sound capturing method with adaptive noise reduction system for broadcasting multicopters," in *Proc. IEEE Int. Conf. Consumer Electron.*, Jan. 2015, pp. 26–29.

[28] S. Yoon, S. Park, and S. Yoo, "Two-stage adaptive noise reduction system for broadcasting multicopters," in *Proc. IEEE Int. Conf. Consumer Electron.*, Jan. 2016, pp. 219–222.

[29] T. Morito, O. Sugiyama, R. Kojima, and K. Nakadai, "Partially shared deep neural network in sound source separation and identification using a uav-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, Oct. 2016, pp. 1299–1304.

[30] Z. W. Tan, A. H. Nguyen, and A. W. Khong, "An efficient dilated convolutional neural network for UAV noise reduction at low input SNR," in *Proc. Asia-Pacific Signal and Inf. Process. Assoc. Annu. Summit and Conf.*, Nov. 2019, pp. 1885–1892.

[31] L. Wang and A. Cavallaro, "Deep learning assisted time-frequency processing for speech enhancement on drones," *IEEE Trans. on Emerging Topics in Computational Intelligence*, vol. 5, no. 6, pp. 871–881, Aug. 2020.

[32] C. Chun, K. M. Jeon, T. Kim, and W. Choi, "Drone noise reduction using deep convolutional autoencoder for UAV acoustic sensor networks," in *Proc. IEEE Int. Conf. on Mobile Ad Hoc and Sensor Syst. Workshops*, Nov. 2019, pp. 168–169.

[33] T. D. Abhayapala, L. Birnie, M. Kumar, D. Grixti-Cheng, and P. N. Samarasinghe, "Generalizing the relative transfer function to a matrix for multiple sources and multichannel microphones," in *Proc. Eur. Signal Process. Conf.*, Sep. 2023, pp. 336–340.

[34] M. Kumar and et al., "Speech denoising in multi-noise source environments using multiple microphone devices via relative transfer matrix," in *Eur. Signal Process. Conf. (in press)*, Sep. 2024, pp. 336–340.

[35] W. N. Manamperi and T. D. Abhayapala, "Successive speaker relative transfer function estimation through relative transfer matrix in noisy reverberant environments," in *Asia-Pacific Signal and Inf. Process. Assoc. Annu. Summit and Conf. (in press)*, Dec. 2024.

[36] W. N. Manamperi and T. D. Abhayapala, "Blind separation of multiple speakers in noisy reverberant environments using relative transfer matrix," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (unpublished)*, 2024.

[37] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: Dataset and methods for uav-embedded sound source localization," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, Oct. 2018, pp. 1–8.

[38] W. N. Manamperi, T. D. Abhayapala, L. I. Birnie, J. A. Zhang, and P. N. Samarasinghe, "Drone audition: On measurements and modeling of drone-related transfer functions," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process. (unpublished)*,

[39] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, Apr. 2007.

[40] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.

[41] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Feb. 2011.

Beamforming informed independent low-rank matrix analysis for sound source enhancement in unmanned aerial vehicles

Jin Xuan Teh*, Norihiro Takamune†, Hiroshi Saruwatari†, Benjamin Yen‡, Michael Kingan* and Yusuke Hioka*
* Acoustics and Vibration Research Centre, University of Auckland, Auckland, New Zealand
† The University of Tokyo, Tokyo, Japan
‡ Tokyo Institute of Technology, Tokyo, Japan

***Abstract*—This study proposes the integration of informed, supervised, and blind sound source enhancement approaches for unmanned aerial vehicle (UAV) applications. The proposed method incorporates a beamformer, representing the informed approach, a pre-recorded noise database for the supervised approach, and independent low-rank matrix analysis (ILRMA) for the blind approach. This method aims to improve sound source enhancement performance while addressing the permutation ambiguity problem of the output channels inherent to ILRMA. The method leverages the fixed spatial relationship between the UAV's propellers and microphones to capture spatial information of the noise generated the propellers. This is achieved by deriving a noise covariance matrix from pre-recorded propeller noise signals and incorporating it into a general eigenvalue beamformer to effectively suppress these noises. The filter weights of the beamformer are then used to inform the spatially regularised ILRMA, guiding the algorithm with additional spatial information of the sound sources. Experimental results demonstrate significant performance improvements, including a 13 dB increase in the source-to-distortion ratio and a 0.21 point increase in the short-time objective intelligibility score. The proposed method outperforms both the original ILRMA and the beamformer in most scenarios and effectively addresses the global permutation ambiguity problem.**

I. Introduction

Unmanned aerial vehicles (UAVs) are increasingly used in applications such as surveillance, filming, rescue operations, and wildlife monitoring. However, the disruptive noise generated by their rotors and propellers, known as 'ego-noise' significantly hinders their effectiveness in auditory applications [1]. This noise degrades the quality of audio recordings captured by onboard microphones, reducing their practical utility. High-quality audio recordings are essential in various domains. For example, environmental monitoring relies on clear audio to track natural events, and wildlife research depends on high-quality recordings to study animal behaviour and habitats [2]. In emergency response scenarios, precise audio data are crucial for effective decision-making, highlighting the potential of UAVs in outdoor search and rescue missions [3].

Addressing the challenge of ego-noise in UAV applications has led to the development of several approaches to enhance sound sources. These approaches can be categorised into three main types: informed, supervised, and blind. The informed approach utilises spatial cues to enhance the target signal, with beamforming being a widely used method [4]. The supervised approach leverages additional information, such as pre-recorded noise data, rotor speed, or additional microphones to capture noise-only signals. These techniques are often used in conjunction with machine learning techniques to estimate noise characteristics for effective noise suppression [5], [6]. The blind approach employs blind source separation (BSS) algorithms, which do not rely on prior information and separate sound sources based on their statistical properties, with independent low-rank matrix analysis (ILRMA) being a notable example [7].

Given the dominant ego-noise signal in UAV applications, achieving good performance in sound source enhancement is challenging. Researchers have increasingly combined different approaches to develop frameworks that maximise performance. Yen et al. [5] combined informed and supervised approaches, using rotor state information to estimate the noise covariance matrix with machine learning methods, which then informed a minimum variance distortionless respons (MVDR) beamformer to further suppress UAV ego-noise. Similarly, our previous work [8] used pre-recorded noise signals to inform a beamformer and estimate the power spectral density (PSD) of the target source. For integration with the blind approach, Lin et al. [9] combined BSS with spatial information of the target source using time-spatial filtering. These studies demonstrate the potential for superior performance through the integration of different approaches.

This study explores integrating informed, supervised, and blind approaches to improve sound source enhancement performance for UAV applications. The proposed framework combines a beamformer (informed approach), a pre-recorded noise database (supervised approach), and ILRMA (blind approach), leveraging each of their strengths while addressing their limitations. This integration aims to guide ILRMA to achieve enhanced performance under real-world conditions. The framework leverages the stationary nature of spatial characteristics between the microphone and UAV noise to estimate a noise spatial covariance matrix. It also uses a target spatial covariance matrix estimated from a range of potential target positions to design a beamformer weight to spatially regularise ILRMA.

II. Background

The proposed method integrates a beamformer framework from a previous study [8] with ILRMA [7]. This section presents the mathematical formulation of the UAV sound source enhancement problem (Section II-A), followed by an overview of the fundamental components, which include the beamforming framework (Section II-B) and ILRMA (Section II-C). The formulations provided here form the basis for the proposed method discussed in Section III.

A. Problem setup

Consider a problem where U independent source signals comprise one target source and $(U-1)$ spatially coherent noise sources are considered. Each target and noise source is assumed to be independent; thus, the sources are considered mutually uncorrelated. These sound sources are captured by an M-channel microphone array under the assumption that the system is determined $(M = U)$. The sound propagation from each source to the microphone array is modelled using a linear and convolutive mixture model. To analyse these sound signals in the time-frequency domain, the short-time Fourier transform (STFT) is applied to the observed signals. This effectively converts the convolutive mixture model into an instantaneous mixture model within each frequency bin. The resulting STFT of the target source and the noise sources are denoted as s_{ij} and $n_{ij,u}$, respectively, where $i = 1, ..., I$ indexes the frequency bins, $j = 1, ..., J$ indexes the time frames, and $u = 1, ..., U$ indexes the sources. Additionally, the STFT of ambient noise, represented as $\mathcal{V}_{ij} \in \mathbb{C}^M$, encompasses both background noise and inherent microphone self-noise. These signals make up the signals observed by the microphone array $x_{ij,m}$, which are expressed in vector form as follows:

$$\begin{aligned}\mathbf{x}_{ij,m} &= [x_{ij,1}, \ldots, x_{ij,M}]^{\mathsf{T}} \\ &= \mathbf{a}_{i,1} s_{ij} + \sum_{u=2}^{U} \mathbf{a}_{i,u} n_{ij,u} + \mathcal{V}_{ij}.\end{aligned} \tag{1}$$

Here, $m = 1, ..., M$ indicates the microphone index, and ^{T} denotes a transpose. The steering vector [10] for each source, $\mathbf{a}_{i,u}$, is described as an array of transfer functions $a_{i,u,m}$ from the source u to the microphone m and is defined as follows:

$$\mathbf{a}_{i,u} = [a_{i,u,1}, \ldots, a_{i,u,M}]^{\mathsf{T}}. \tag{2}$$

B. Beamforming framework

Beamforming is a spatial filtering technique that isolates target signals while attenuating noise signals. The output of a beamformer can be expressed as follows:

$$\mathbf{y}_{ij} = \mathbf{w}_i^{\mathsf{H}} \mathbf{x}_{ij}. \tag{3}$$

Here, ^{H} denotes a Hermitian transpose. The vector $\mathbf{w}_i = [w_{i,1}, \ldots, w_{i,M}]^{\mathsf{T}}$ represents the beamformer's filter weights, $w_{i,m}$, applied to the signals observed by each microphone m. In the absence of ambient noise sources $\mathcal{V}_{ij}$, the optimal filter weight can be found by maximising the signal-to-noise ratio (SNR) [11], which is expressed as follows:

$$\xi_i = \frac{{\mathbf{w}_i}^{\mathsf{H}} \mathbf{R}_{i,\mathrm{s}} \mathbf{w}_i}{{\mathbf{w}_i}^{\mathsf{H}} \mathbf{R}_{i,\mathrm{n}} \mathbf{w}_i}, \tag{4}$$

where $\mathbf{R}_{i,\mathrm{s}}$ and $\mathbf{R}_{i,\mathrm{n}}$ represent the spatial covariance matrices of frequency i for the target and noise signal, respectively. The solution to this optimisation problem, known as generalised eigenvalue (GEV) beamforming [12], [13], is given by:

$$\mathbf{w}^{\mathsf{GEV}} = \mathcal{P}\{\mathbf{R}_{\mathrm{n}}^{-1} \mathbf{R}_{\mathrm{s}}\}. \tag{5}$$

For brevity, i is omitted unless otherwise specified hereafter. Here, $\mathcal{P}\{\cdot\}$ denotes the operator that extracts the eigenvector corresponding to the largest eigenvalue through generalised eigenvalue decomposition [14]. To achieve a distortionless magnitude response, blind analytical normalisation (BAN) compensation factor was proposed [13]:

$$\psi = \frac{\sqrt{{\mathbf{w}^{\mathsf{GEV}}}^{\mathsf{H}} \mathbf{R}_{\mathrm{n}} \mathbf{R}_{\mathrm{n}} \mathbf{w}^{\mathsf{GEV}}}}{{\mathbf{w}^{\mathsf{GEV}}}^{\mathsf{H}} \mathbf{R}_{\mathrm{n}} \mathbf{w}^{\mathsf{GEV}}}, \tag{6}$$

where, ψ corrects the distortions in magnitude inherent to the GEV beamformer. Consequently, applying ψ allows the adjusted $\mathbf{w}^{\mathsf{GEV}}$ to reassemble the MVDR's distortionless magnitude response, albeit with phase distortions. The resulting beamformer is denoted as GEV beamformer with BAN compensation factor (GEVB) and is expressed as: $\mathbf{w}^{\mathsf{GEVB}} = \psi \mathbf{w}^{\mathsf{GEV}}$.

C. ILRMA

In BSS, the goal is to recover source signals from observed signals. Let the STFT of the source signals in each time-frequency bin be described as, $\mathbf{h}_{ij} = [h_{ij,1}, \ldots, h_{ij,U}]^{\mathsf{T}}$, where the order of sources in $\mathbf{h}_{ij}$ is unknown, reflecting the "blind" nature of BSS. This stands in contrast to the distinguishable sources described in (1). The mixing system is modelled as $\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{h}_{ij}$, where $\mathbf{A}_i$ is the mixing matrix, and each component $\mathbf{a}_{i,u}$ within $\mathbf{A}_i$ is a steering vector for each source. The separation process can be written as:

$$\mathbf{y}_{ij} = {\mathbf{W}^{\mathsf{BSS}}}_i \mathbf{x}_{ij}, \tag{7}$$

where $\mathbf{y}_{ij} = [y_{ij,1}, \ldots, y_{ij,U}]^{\mathsf{T}}$ are the separated signals, and ${\mathbf{W}^{\mathsf{BSS}}}_i = \left[{\mathbf{w}^{\mathsf{BSS}}}_{i,1}, \ldots, {\mathbf{w}^{\mathsf{BSS}}}_{i,U}\right]^{\mathsf{H}}$ is the demixing matrix. In a determined problem where $\mathbf{A}_i$ is invertible, ideal ${\mathbf{W}^{\mathsf{BSS}}}_i$ is defined as $\mathbf{A}_i^{-1}$.

In ILRMA, the separated signal is assumed to follow a time-varying complex Gaussian distribution with a zero mean and variance $l_{ij,u}$, corresponding to $y_{ij,u}$ [7]. $l_{ij,u}$, is arranged in a two-dimensional array $\mathbf{L}_u \in \mathbb{R}_+^{I \times J}$ representing the spectral model for the u-th source [15]. Similar to non-negative matrix factorization (NMF), $\mathbf{L}_u$ can be represented as $\mathbf{L}_u = \mathbf{T}_u \mathbf{V}_u$, where $\mathbf{T}_u \in \mathbb{R}_+^{I \times K}$ and $\mathbf{V}_u \in \mathbb{R}_+^{K \times J}$ are non-negative matrices with K NMF bases. The cost function for ILRMA, expressing the negative log-likelihood function, is given as:

$$\mathcal{J} = \sum_{ij,u} \left[\frac{\left| \mathbf{w}^{\mathsf{BSS}\mathsf{H}}{}_{i,u} \mathbf{x}_{ij} \right|^2}{l_{ij,u}} + \log l_{ij,u} \right] - 2J \sum_i \log \left| \det \mathbf{W}^{\mathsf{BSS}}{}_i \right|. \tag{8}$$

Here, constant terms have been omitted for simplicity.

To further improve the separation accuracy and optimisation stability, a method called spatially regularised ILRMA (SR-ILRMA) have been proposed [16]. It incorporates a regularisation term into the ILRMA framework, thereby resolving the global permutation ambiguity problem while improving the convergence of the cost function during iterative optimisation [16]. The global permutation ambiguity problem refers to the uncertainty regarding which output channel corresponds to the target speech, a challenge that arises from the random channel allocation in the original ILRMA method. By resolving this ambiguity, SR-ILRMA enhances the reliability of source separation. The SR-ILRMA cost function, denoted by $\mathcal{J}_R$, extends the original ILRMA cost function $\mathcal{J}$ by incorporating a regularisation term. This term is scaled by a weight parameter γ_u and quantifies the difference between the current estimate $\mathbf{w}^{\mathsf{BSS}}{}_{i,u}$ and a supervisor matrix $\widehat{\mathbf{w}}^{\mathsf{BSS}}_{i,u}$, as follows:

$$\mathcal{J}_{\mathrm{R}} = \mathcal{J} + \sum_{i,u} \gamma_u \| \mathbf{w}^{\mathsf{BSS}}{}_{i,u} - \widehat{\mathbf{w}}^{\mathsf{BSS}}_{i,u} \|^2. \tag{9}$$

The update rule and detailed description for minimising the SR-ILRMA cost function, $\mathcal{J}_R$, can be found in [16].

III. Proposed Method

In UAV audition, extracting a clear target signal is challenging due to the low SNR [5]. To address this, a novel framework is proposed that combines several established approaches to maximise enhancement performance. First, informed and supervised approaches are integrated, using a pre-recorded noise database and a GEVB beamformer to create beamformer with practical utility. The filter weights of this beamformer are then employed to inform ILRMA through spatial regularisation, improving source enhancement and addressing the global permutation ambiguity problem inherent to ILRMA.

The proposed framework is illustrated in Figure 1, with the input and pre-processing elements depicted in grey. The spatial covariance matrices, marked in green, are discussed in Section III-A. The beamformer-informed ILRMA component, highlighted in blue, is detailed in Section III-B.

A. Spatial covariance matrix

The noise spatial covariance matrix, $\mathbf{R}_{\mathrm{n}}$, is crucial for both the beamformer and the spatially regularised ILRMA. Assuming that the spatial characteristics of the UAV ego-noise are stationary due to the fixed position of the microphone array [5], $\mathbf{R}_{\mathrm{n}}$ can be estimated using a sampled noise covariance matrix, here i is omitted for simplicity:

$$\mathbf{R}_{\mathrm{n}} = \frac{1}{JP} \sum_{j=1}^{J} \sum_{p=1}^{P} \mathbf{x}_{j,p} \mathbf{x}^{\mathsf{H}}_{j,p}, \tag{10}$$

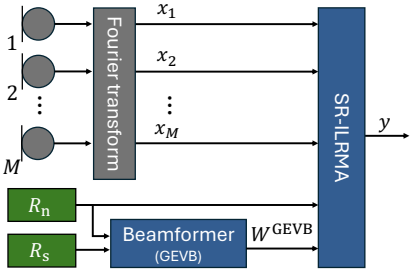


Fig. 1: Block diagram of the framework. For clarity, the frequency and frame indices have been omitted.

where $\mathbf{x}_p$ is the p-th noise sample signal, with $p = 1, ..., P$ indicating the noise sample index. By leveraging the fixed spatial relationship between the microphone array and the UAV noise sources, $\mathbf{R}_{\mathrm{n}}$ can be pre-calculated using pre-recorded rotor noise data. This method eliminates the need for complex methods, such as voice activity detection (VAD) or machine learning techniques, which are commonly employed to estimate the spatial covariance matrix in sound source enhancement for UAVs. This simplifies implementation and enhances practicality.

The target spatial covariance matrix, $\mathbf{R}_{\mathrm{s}}$, is used in the beamformer. It is derived by considering the target source's varying positions relative to the microphone array during UAV operations. Figure 2 illustrates the relative position of the target source and the UAV from both top and side views. The top view assumes the azimuth angle relative to the UAV direction is zero, indicating the target source is directly ahead. The side view shows the elevation angle varying within a predefined range. This assumption enhances the framework's robustness in real-world scenarios where the precise elevation angle is often unknown. $\mathbf{R}_{\mathrm{s}}$ is calculated as:

$$\mathbf{R}_{\mathrm{s}} = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{a}_q \mathbf{a}^{\mathsf{H}}_q. \tag{11}$$

Here, $\mathbf{a}_q$ denotes the steering vector for various potential positions of the target source relative to the microphone array, with $q = 1, ..., Q$ indicating the target position index. These steering vectors are pre-measured.

B. Beamformer informed ILRMA

Beamformer informed ILRMA aims to enhance separation performance and address the global permutation ambiguity in ILRMA. The proposed method uses beamformer filter weights to inform ILRMA by embedding spatial information of sound sources into the demixing process and dedicating one output channel specifically for the target source. This strategy ensures consistent allocation of the target source to a designated channel, reducing uncertainty in standard ILRMA outputs.

The beamformer in this framework employs the GEVB beamformer described in Section II-B and uses the spatial covariance matrix detailed in Section III-A to calculate the filter

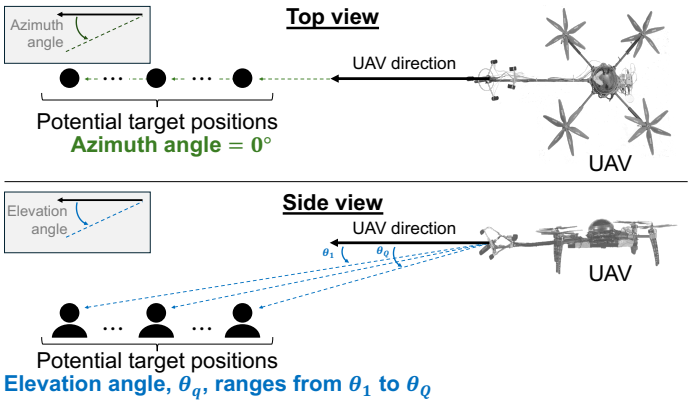


Fig. 2: Target source positions: azimuth (top view) and elevation (side view).

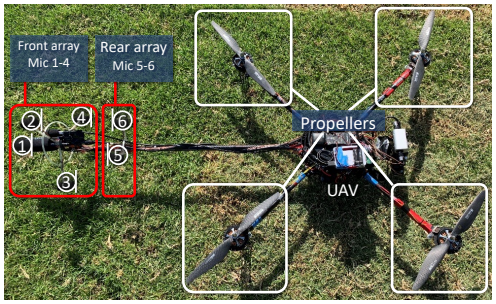


Fig. 3: UAV setup (from [5]). Red boxes highlight microphone arrays, white circles show microphone numbers, and white boxes highlight propellers.

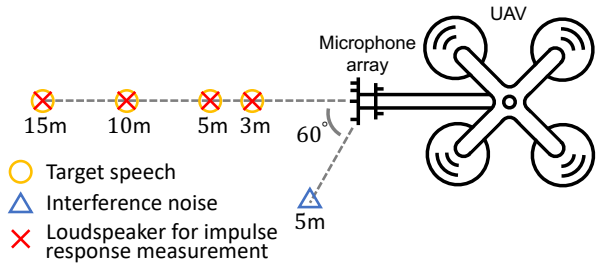


Fig. 4: Illustration of the relative positions.

weight $\mathbf{w}^{\text{GEVB}}$. This approach has shown robust performance across various target positions in a previous study [8].

Subsequently, ILRMA is informed by the beamformer through supervisor matrix, $\widehat{\mathbf{W}}_i^{\text{BSS}}$, derived via a multi-step process using $\mathbf{R}_{\text{n}}$ and $\mathbf{w}^{\text{GEVB}}$. This process begins with the eigenvalue decomposition of $\mathbf{R}_{\text{n}}$ into its eigenvectors and eigenvalues sorted in descending order. The first $U-1$ eigenvectors, corresponding to the largest eigenvalues, are selected to capture the most significant noise components. These eigenvectors and the filter weights $\mathbf{w}^{\text{GEVB}}$ are then used to construct $\widehat{\mathbf{W}}_i^{\text{BSS}}$. Specifically, $\widehat{\mathbf{W}}_i^{\text{BSS}}$ incorporates the first $U-1$ eigenvectors for the initial $U-1$ channels and $\mathbf{w}^{\text{GEVB}}$ for the U-th channel. This configuration provides spatial information on the dominant noise components and the target source while dedicating the final output channel specifically for the target source. Finally, $\widehat{\mathbf{W}}_i^{\text{BSS}}$ is used in the cost function for SR-ILRMA.

IV. Experimental Results

The proposed method was evaluated in real-world scenarios with varying target speech positions and sound levels. Performance was measured using objective metrics such as the source-to-distortion ratio (SDR) [17] and short-time objective intelligibility (STOI) [18]. Additionally, SDR improvement (SDRi) and STOI improvement (STOIi) were assessed based on the differences between output and input performance metrics.

A. Experiment

The proposed framework was evaluated using UAV recordings detailed in a previous study [5]. The system's microphone array includes six microphones: four in the front sub-array (Microphone 1 is a shotgun, Microphones 2, 3, and 4 are cardioids) and two cardioid microphones in the rear sub-array (Microphones 5 and 6) as shown in Figure 3. The database included the followings:

- **Target speech:** ten sentences from the Centre for Speech Technology Research VKTS Corpus [19], played at 60 dBA and 80 dBA (measured at one metre from the loudspeaker), and recorded at distances of 3 m, 5 m, 10 m, and 15 m, yielding 80 sets of target speech.
- **UAV noise:** 64 recordings of hovering UAV ego-noise without the target signal. Five sets were used for performance evaluation, and 59 sets for estimating the noise covariance matrix.
- **Interference noise:** background music and traffic noise recorded at 60 dBA from a distance of 5 m at a 60-degree angle to the front of the microphone array (Figure 4).
- **Impulse response:** measured at 3 m, 5 m, 10 m, and 15 m to generate the estimated spatial covariance matrix for the target source.

Recordings were conducted at Harry James Taupaki Reserve in Auckland, New Zealand, capturing ambient noises including wind, bird and insect calls [5]. Figure 4 shows the positions of the target signal, interference noise, and the loudspeaker for measuring impulse responses. These recordings enabled a comprehensive evaluation of the proposed method.

B. Results and discussion

1) Addressing global permutation ambiguity problem: The proposed method's effectiveness in correctly allocating the target output channel across various noise levels was evaluated. Accuracy was measured as the percentage of instances where the target output channel was correctly identified. Figure 5 compares the accuracy of the proposed method, beamformer-informed ILRMA (iILRMA), with the original ILRMA at different input SNRs. The original ILRMA assumes that speech signals exhibit higher kurtosis than noise signals [20] and uses kurtosis of the output signals to estimate the speech channel

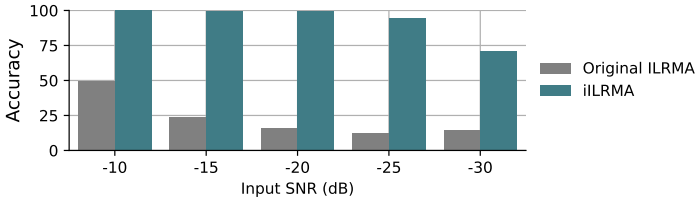


Fig. 5: Accuracy of channel allocation at varying input SNR.

[21]. The target channel is correctly identified if the channel with the highest kurtosis matches the channel with the highest SDR. In contrast, iILRMA deems the target channel correctly identified if the allocated target channel has the highest SDR among all channels. Results show that the accuracy of the original ILRMA decreases as SNR decreases, from 50% at -10 dB to 24% at -15 dB. Conversely, iILRMA achieves near-perfect accuracy at input SNRs from 0 to -20 dB, with accuracy declining to 95% at -25 dB and 71% at -30 dB. This demonstrates significant improvement over the original ILRMA and identified a critical threshold around -25 dB for the proposed method's effectiveness.

2) Performance comparison: The source enhancement performance of the proposed method is compared against the original ILRMA and GEVB beamformer. Figure 6 presents the SDRi and STOIi of various enhancement methods. The results indicate that iILRMA outperforms the original ILRMA in both SDRi and STOIi metrics. When compared to the GEVB beamformer, iILRMA shows better SDRi from -10 to -20 dB input SNR but performs worse at -25 and -30 dB input SNR. This decline in performance at lower SNRs is attributed to the reduced accuracy observed at -25 dB, as shown in Figure 5. Additionally, the beamformer's design, characterised by a distortionless magnitude response, may contribute to this outcome. For STOIi, iILRMA exhibits superior performance from -10 dB to -25 dB and comparable performance at -30 dB compared to the beamformer. The diminishing improvement in STOIi at lower SNRs can be attributed to iILRMA's reduced channel accuracy at -25 dB and below, as discussed in Section IV-B1. These results demonstrate that iILRMA provides better source enhancement performance in most scenarios.

3) Real world performance: The proposed method was evaluated under real-world conditions, unlike the scenarios with fixed input SNRs discussed in Sections IV-B1 and IV-B2. The evaluation involved varying sound levels and target speech positions to assess its robustness and performance. The results are shown in Figure 7. At a target sound level of 80 dBA, iILRMA demonstrated significant performance improvements, with SDR and STOI improvements of 13 dB and 0.21, respectively, at a target distance of 3 metres. However, at a target sound level of 60 dBA, the proposed method showed suboptimal performance, indicating that enhancing sound sources under such challenging conditions remains an area for future research. Audio samples are accessible at: https://github.com/JinXuanTeh/audio-sample-BF-ILRMA

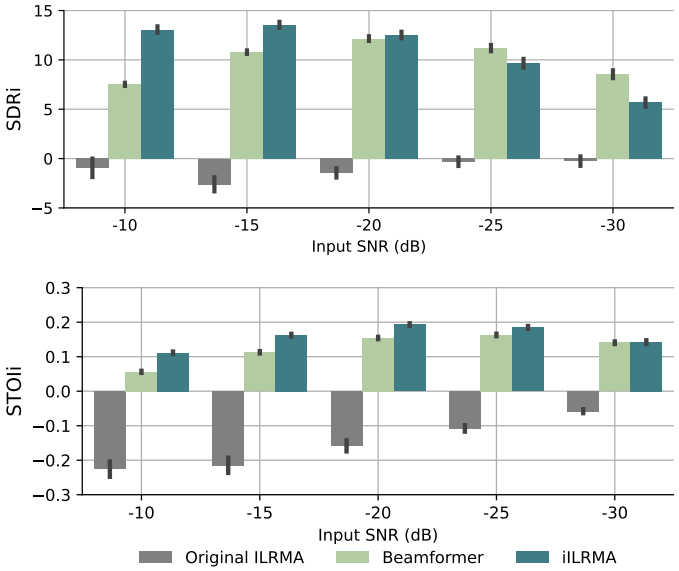


Fig. 6: Source enhancement performance of different methods with 95% confidence intervals.

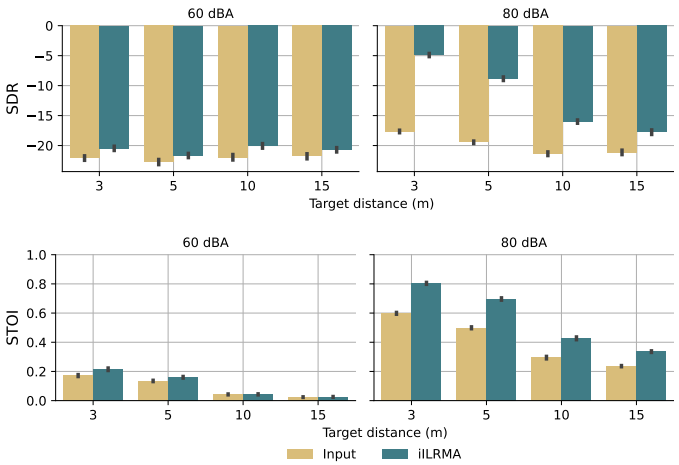


Fig. 7: Source enhancement performance in real-world conditions with 95% confidence intervals.

V. Conclusion

This study explored the integration of informed, supervised, and blind approaches to enhance sound source performance in UAV applications. The proposed framework combines a beamformer (informed approach), a pre-recorded noise database (supervised approach), and ILRMA (blind approach). Experimental results demonstrated that the proposed method outperforms both the original ILRMA and the beamformer in most scenarios. iILRMA showed substantial gains in SDR and STOI metrics, achieving an SDR improvement of 13 dB and a STOI improvement of 0.21. Additionally, the proposed method effectively addresses the global permutation ambiguity problem, ensuring more accurate channel allocation across various SNR levels. Future research should focus on enhancing the framework's performance under extremely low SNR conditions and exploring the integration of additional sound

source enhancement algorithms to further improve robustness and effectiveness in diverse environments. This comprehensive approach demonstrates the potential for superior performance through the integration of different methods.

ACKNOWLEDGMENT

This research was partially funded by the Royal Society of New Zealand Catalyst Seeding programme: New Zealand – Japan Joint Research Projects (JSP-UOA1901-JR), the Kajima Foundation's Support Program for International Joint Research Activities (2024-kyodoshin-05) and the Acoustics and Vibration Research Centre at the University of Auckland.

REFERENCES

[1] R. McKay and M. J. Kingan, "Multirotor unmanned aerial system propeller noise caused by unsteady blade motion," in *25th AIAA/CEAS Aeroacoustics Conference*. 2019.

[2] D. P. Nowacek, F. Christiansen, L. Bejder, J. A. Goldbogen, and A. S. Friedlaender, "Studying cetacean behaviour: New technological approaches and conservation applications," en, *Animal Behaviour*, vol. 120, pp. 235–244, 2016.

[3] K. Nakadai, M. Kumon, H. G. Okuno, *et al.*, "Development of microphone-array-embedded uav for search and rescue task," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 5985–5990.

[4] Y. Hioka, M. Kingan, G. Schmid, and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2016.

[5] B. Yen, Y. Hioka, G. Schmid, and B. Mace, "Multisensory sound source enhancement for unmanned aerial vehicle recordings," en, *Applied Acoustics*, vol. 189, p. 108 590, 2022.

[6] W. N. Manamperi, T. D. Abhayapala, P. N. Samarasinghe, and J. Zhang, "Drone audition: Audio signal enhancement from drone embedded microphones using multichannel wiener filtering and gaussian-mixture based post-filtering," en, *Applied Acoustics*, vol. 216, p. 109 818, 2024.

[7] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[8] J. X. Teh, B. Yen, M. Kingan, and Y. Hioka, "Source enhancement with different mvdr beamformer designs for unmanned aerial vehicle audition," 2024.

[9] L. Wang and A. Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2523–2537, 2020.

[10] M. Brandstein and D. Ward, *Microphone Arrays*. Springer Berlin Heidelberg, 2001.

[11] J. Li and P. Stoica, *Robust adaptive beamforming* (Wiley Series in Telecommunications and Signal Processing ; v.88), eng. Hoboken, NJ: John Wiley, 2006.

[12] H. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory* (Detection, Estimation, and Modulation Theory). Wiley, 2004.

[13] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[14] G. Golub and C. Van Loan, *Matrix Computations* (Johns Hopkins Studies in the Mathematical Sciences). Johns Hopkins University Press, 2013.

[15] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," en, *APSIPA Transactions on Signal and Information Processing*, vol. 8, no. 1, 2019.

[16] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 746–750, 2018.

[17] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, *First stereo audio source separation evaluation campaign: data, algorithms and results*. Springer Berlin Heidelberg, 2007, pp. 552–559.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[19] J. Yamagishi, V. Christophe, and M. Kirsten, *CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)*, Sound, Data retrieved from Edinburgh DataShare, Nov. 2019.

[20] G. Li, M. E. Lutman, S. Wang, and S. Bleeck, "Relationship between speech recognition in noise and sparseness," en, *International Journal of Audiology*, vol. 51, no. 2, pp. 75–82, 2011.

[21] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, "Blind speech extraction based on rank-constrained spatial covariance matrix estimation with multivariate generalized Gaussian distribution," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1948–1963, 2020.

SMoLnet-T: An Efficient Complex-spectral Mapping Speech Enhancement Approach with Frame-wise CNN and Spectral Combination Transformer for Drone Audition

Zhi-Wei Tan* and Andy W. H. Khong†
* Nanyang Technological University, School of Electrical and Electronic Engineering, Singapore
† Nanyang Technological University, School of Electrical and Electronic Engineering and Lee Kong Chian School of Medicine, Singapore
E-mail: {zhiwei.tan, andykhong}@ntu.edu.sg

***Abstract*—Speech enhancement for drone audition applications is challenging due to the low SNR with large spectra feature overlap and limited computing resources. We propose SMoLnet-T, a complex spectral mapping approach with frame-wise CNN and newly-formulated spectral combination transformers. SMoLnet-T incorporates dilated CNN to extract spectral maps of high-frequency resolution for its transformers. This allows it to focus on a higher level of abstraction and determine the combination of spectral maps is crucial for enhancement across a large temporal context. Experiment results with noise recorded from a hovering drone highlight the efficacy of SMoLnet-T over DPTNet with significantly lower computational requirements and speech distortion while achieving improved speech intelligibility with SNR < -23 dB.**

***Index Terms*—Convolution neural network, deep learning, transformer, drone audition, speech enhancement**

I. Introduction

Drone audition has profound applications in search and rescue, surveillance, and package delivery. These applications require enabling (acoustic) technologies such as source localization [1, 2], sound classification [3], noise suppression [4], and tracking [5]. Unlike conventional microphone array systems for indoor environments [6], speech enhancement for unmanned aerial vehicles (UAVs) is challenging due to the ego-noise generated during flight [7]. Operating under such an adverse low signal-to-noise ratio (SNR) environment limits the enhancement performance when the spectra feature of speech and drone noise overlap. Furthermore, the limited onboard computing resources necessitate the algorithm to be "light-weight".

Deep learning approaches for speech enhancement under low SNR conditions have been gaining attention in recent years [8–10]. One such approach involves the estimation of rotor noise power spectral density by exploiting UAV rotor characteristics and signals from the received microphones [8]. More recently, a small model on low SNR network (SMoLnet) has been proposed for single-channel drone noise reduction [9]. Employing exponentially dilated convolutions [11, 12], the convolutional neural network (CNN) within SMoLnet requires fewer learnable parameters while achieving significant noise reduction and improved speech intelligibility comparable to larger models. An independent study [13] that focuses on single-channel speech enhancement showed that the resource-efficient SMoLnet achieves reasonable speech quality (in terms of perceptual evaluation of speech quality (PESQ) [14] and scale-invariant signal-to-distortion ratio (SI-SDR) [15]) compared to the more extensive single-channel dual-path transformer network (DPTNet) [16] and the deep complex U-net (DCUNet) [17] at very low SNR. However, due to the shorter temporal context of SMoLnet, it suffers from lower speech intelligibility performance in terms of the extended short-time objective intelligibility (ESTOI) measure [18].

To leverage longer temporal context, recent deep learning approaches incorporate transformers [19]. In particular, unlike CNNs and recurrent neural networks (RNNs), the transformer architecture exploits significant temporal contexts more effectively via its self-attention mechanism and has found applications in speech processing models [16, 20]. Of particular interest is the time-domain DPTNet [16], which employs the intra- and inter-transformer for the exploitation of local and sub-global temporal information, respectively. It is also useful to note that separating the time-domain signal into local frames and global chunks reduces the computational complexity required to process all the frames.

As an extension, the time-domain Sepformer [21] alleviates the need for the RNN within DPTNet by employing multiple transformers with fixed positional encoding [19]. This facilitates modeling of the temporal positioning information resulting in a highly-parallelized model that does not require step-wise processing by RNN. The model, however,

This research was conducted under project WP6 within the Delta-NTU Corporate Lab with funding support from A*STAR under its IAF-ICP programme (Grant no: I2201E0013) and Delta Electronics Inc.

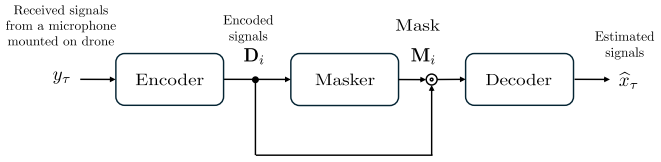


Fig. 1. Masking-based approaches for speech processing algorithms [16, 20, 21, 23]

requires a significantly larger number of learnable parameters to achieve similar performance to that of DPTNet. More recently, a magnitude-domain approach [20], where short-time Fourier transform (STFT) [22] features serve as input to the transformer, has been proposed. This approach performs similarly to the Sepformer with lower computational cost by achieving a trade-off between several global chunks and local frames. Apart from enhancing the magnitude (while leaving the phase unprocessed), the use of multiple transformers require considerable computational resources.

We propose the SMoLnet-T, consisting a frame-wise CNN with newly formulated spectral-attention transformers. The proposed approach is based on complex spectral mapping, which extracts high-resolution spectral feature maps through exponentially increasing frequency-dilated convolutions. These spectral maps, along with their associated frequency position information, are retained for the subsequent spectral combination transformers. In contrast to existing transformer-based approaches such as DPTNet that focuses on alternating between the local or sub-global information, SMoLnet-T focuses on a higher level of abstraction by determining which combination of high-resolution spectral maps generated by the frame-wise CNN is suitable for enhancement in the global aspect. This avoids the need for masking, which, in general, extracts features ineffectively.

II. Problem formulation and literature review

The received signal of a single-channel microphone mounted on a UAV can be expressed as

$$y_\tau = x_\tau + v_\tau, \tag{1}$$

where $x_\tau, v_\tau \in \mathbb{R}$ are the (clean) speech and noise signals, respectively, $1 \leq \tau \leq \mathcal{T}$ is the sample index with $\mathcal{T}$ being the number of samples. Due to the significant computation associated with the high-fidelity speech signal, transformer-based approaches [16, 20, 21], in general, focus on reducing the processing of large $\mathcal{T}$ samples. The general framework for the use of these transformer networks is depicted in Fig. 1.

The DPTNet [16] employs an encoder

$$\underline{\mathbf{D}}^{[1]} = \mathfrak{f}_{\mathsf{seg}_2}\Big(\mathfrak{f}_{\mathsf{relu}}\Big(\mathfrak{f}_{\mathsf{conv}}\big(\mathfrak{f}_{\mathsf{seg}_1}\left(\mathbf{y}\right)\big)\Big)\Big), \tag{2}$$

where $\mathbf{y} = [y_1, \ldots, y_\mathcal{T}]^\mathsf{T}$ with $(\cdot)^\mathsf{T}$ being the transpose operator. Here, $\mathfrak{f}_{\mathsf{seg}_1} : \mathbb{R}^{1\times\mathcal{T}} \to \mathbb{R}^{L\times I}$ segments the received signal into I overlapping vectors each with L samples. The convolution function $\mathfrak{f}_{\mathsf{conv}} : \mathbb{R}^{L\times I} \to \mathbb{R}^{N\times I}$ subsequently filters these samples with N number of filters each of length L. The function $\mathfrak{f}_{\mathsf{relu}}$ is the rectified linear unit (ReLU) [24] activation unit, and $\mathfrak{f}_{\mathsf{seg}_2} : \mathbb{R}^{N\times I} \to \mathbb{R}^{N\times K\times P}$ further segments the activated features to P overlapping chunks each of length K. It is important to note that $N < L \leq \mathcal{T}$.

With $\mathbf{D}_p^{[1]} \in \mathbb{R}^{N\times K}$ being the pth matrix of $\underline{\mathbf{D}}^{[1]}$, DPTNet employs B consecutive dual-path transformers such that [16]

$$\widetilde{\mathbf{D}}_p^{[b]} = \mathfrak{f}_{\mathsf{intra-t}}\left(\mathbf{D}_p^{[b]}\right), \tag{3}$$

$$\mathbf{D}_k^{[b+1]} = \mathfrak{f}_{\mathsf{inter-t}}\left(\widetilde{\mathbf{D}}_k^{[b]}\right), \tag{4}$$

where $1 \leq b \leq B$ is the index of the dual-path transformer, $\mathbf{D}_p^{[b]} \in \mathbb{R}^{N\times K}$ is the encoded signal, and $\widetilde{\mathbf{D}}_p^{[b]} \in \mathbb{R}^{N\times K}$ and $\widetilde{\mathbf{D}}_k^{[b]} \in \mathbb{R}^{N\times P}$ are the intermediate outputs. Here, $1 \leq p \leq P$ and $1 \leq k \leq K$, are, respectively, the local and global indices with P being the number of chunks and K being the length of chunks. In (4), $\mathbf{D}_k^{[b+1]}$ is the encoded signal for the subsequent dual-path transformers.

The variables $\mathfrak{f}_{\mathsf{intra-t}}^{[b]}$ and $\mathfrak{f}_{\mathsf{inter-t}}^{[b]}$ in (3) and (4) are the bth dual-path transformers, where each computes a self-attention mechanism given by

$$\mathbf{Q}_{i,h} = \mathbf{D}_i\mathbf{W}_h^{\mathsf{Q}}, \quad \mathbf{K}_{i,h} = \mathbf{D}_i\mathbf{W}_h^{\mathsf{K}}, \quad \mathbf{V}_{i,h} = \mathbf{D}_i\mathbf{W}_h^{\mathsf{V}}.$$

We have omitted b for clarity and that $h \in [1, H]$ denotes the head index with H being the number of heads. The variable

$$i = \begin{cases} p, & \text{for } \mathfrak{f}_{\mathsf{intra-t}}; \\ k, & \text{for } \mathfrak{f}_{\mathsf{inter-t}}, \end{cases} \tag{5}$$

denotes the embedding index. The query $\mathbf{Q}_{i,h}$, keys $\mathbf{K}_{i,h}$, and values $\mathbf{V}_{i,h}$ are, respectively, computed using weights

$$\mathbf{W}_h^{\mathsf{Q}}, \mathbf{W}_h^{\mathsf{K}}, \mathbf{W}_h^{\mathsf{V}} \in \begin{cases} \mathbb{R}^{K\times\frac{D}{H}}, & \text{for } \mathfrak{f}_{\mathsf{intra-t}}; \\ \mathbb{R}^{P\times\frac{D}{H}}, & \text{for } \mathfrak{f}_{\mathsf{inter-t}}. \end{cases} \tag{6}$$

Thereafter, the multi-head attention is achieved via

$$\mathbf{A}_i = \mathfrak{f}_{\mathsf{cat}}\Big([\mathbf{A}_{i,1}, \ldots, \mathbf{A}_{i,h}, \ldots, \mathbf{A}_{i,H}]\Big)\mathbf{W}^{\mathsf{O}}, \tag{7}$$

where $\mathfrak{f}_{\mathsf{cat}}$ concatenates the single-head attention $\mathbf{A}_{i,h} \in \mathbb{R}^{N\times\frac{D}{H}}$ along the second dimension and $\mathbf{W}^{\mathsf{O}} \in \mathbb{R}^{D\times D}$ denotes output weights to achieve $\mathbf{A}_i \in \mathbb{R}^{N\times D}$. Here,

$$\mathbf{A}_{i,h} = \mathfrak{f}_{\mathsf{softmax}}\left(\frac{\mathbf{Q}_{i,h}\mathbf{K}_{i,h}^\mathsf{T}}{\sqrt{D}}\right)\mathbf{V}_{i,h}. \tag{8}$$

With (7), the DPTNet employs the improved transformer [25] described by

$$\dot{\mathbf{D}}_i = \mathfrak{f}_{\mathsf{ln}_1}\left(\mathbf{A}_i + \mathbf{D}_i\right), \tag{9}$$

$$\ddot{\mathbf{D}}_i = \mathfrak{f}_{\mathsf{relu}}\left(\mathfrak{f}_{\mathsf{RNN}}\left(\dot{\mathbf{D}}_i\right)\right)\mathbf{W}_2 + \mathbf{b}_2, \tag{10}$$

$$\dddot{\mathbf{D}}_i = \mathfrak{f}_{\mathsf{ln}_2}\left(\dot{\mathbf{D}}_i + \ddot{\mathbf{D}}_i\right), \tag{11}$$

where $\mathfrak{f}_{\mathsf{ln}_1}$ and $\mathfrak{f}_{\mathsf{ln}_2}$ denote layer normalizations, $\mathbf{W}_2 \in \mathbb{R}^{D\times D}$ and $\mathbf{b}_2 \in \mathbb{R}^{1\times D}$ denote the weight and the bias, respectively. The variables $\dot{\mathbf{D}}_i$ and $\ddot{\mathbf{D}}_i$ are the intermediate outputs of the

improved transformer, and the output $\dddot{\mathbf{D}}_i$ is used as the input of (3) or (4) such that

$$\ddot{\mathbf{D}}_i = \begin{cases} \widetilde{\mathbf{D}}_i, & \text{if subsequent transformer is } \mathfrak{f}_{\mathsf{inter-t}}; \\ \mathbf{D}_i, & \text{if subsequent transformer is } \mathfrak{f}_{\mathsf{intra-t}}. \end{cases} \tag{12}$$

The function $\mathfrak{f}_{\mathsf{RNN}}$ in (10) is an RNN that learns the order information of the input. We note that the improved transformer differs from the transformer [19] in two ways—a feed-forward network is employed instead of $\mathfrak{f}_{\mathsf{RNN}}$ and that a fixed position encoding is added to $\mathbf{D}_i$.

If there are no subsequent transformer (i.e., $b = B$) in (12), the estimated signal is decoded using

$$\mathbf{M} = \mathfrak{f}_{\mathsf{ola}}\Big(\mathfrak{f}_{\mathsf{conv}}\left(\ddot{\mathbf{D}}_i\right)\Big), \tag{13}$$

$$\widehat{\mathbf{x}} = \mathfrak{f}_{\mathsf{deconv}}\left(\mathbf{D}_p^{[1]} \odot \mathbf{M}\right), \tag{14}$$

where $\mathbf{M}$ is the mask, and $\mathfrak{f}_{\mathsf{deconv}}$, $\mathfrak{f}_{\mathsf{ola}}$, and $\odot$ are transpose convolution, overlap-add, and Hadamard product operators. With reference to (8), it is important to note that the sequential information due to the attention mechanism is on the first dimension, and its computational requirement scales quadratically with this dimension. For DPTNet, this dimension has N elements determined by $\mathfrak{f}_{\mathsf{conv}}$.

In contrast to DPTNet, the Sepformer [21] alleviates the need of RNN in (10) by incorporating sinusoidal positional encoding [19] to $\mathbf{D}_p^{[b]}$ in (4) and $\widetilde{\mathbf{D}}_k^{[b]}$ in (3), respectively, via

$$\mathbf{D}_p^{'[b]} = \mathbf{D}_p^{[b]} + \mathbf{E}^{N\times K}, \qquad \widetilde{\mathbf{D}}_k^{'[b]} = \widetilde{\mathbf{D}}_k^{[b]} + \mathbf{E}^{N\times P}, \tag{15}$$

where the nth row and kth column of $\mathbf{E}^{N\times K}$ is expressed as

$$e_{n,k} = \begin{cases} \sin\left(\frac{k}{10000^{2n/N}}\right), & \text{if } n \text{ is even;} \\ \cos\left(\frac{k}{10000^{2n/N}}\right), & \text{if } n \text{ is odd,} \end{cases} \tag{16}$$

and similarly for $\mathbf{E}_{N\times P}$, i.e.,

$$e_{n,p} = \begin{cases} \sin\left(\frac{p}{10000^{2n/N}}\right), & \text{if } n \text{ is even;} \\ \cos\left(\frac{p}{10000^{2n/N}}\right), & \text{if } n \text{ is odd.} \end{cases} \tag{17}$$

Instead of employing a learned encoder-decoder in (2) and (14) for DPTNet and SepFormer, the magnitude-based SepFormer (Mag-SepFormer) [20] operates on the spectral-temporal domain by re-formulating (1) as

$$y_{t,f} = x_{t,f} + v_{t,f}, \tag{18}$$

where $1 \le t \le T$ is the time index with T being the number of time frames, while $1 \le f \le F$ is the frequency index with F being the number of the frequency bins. As opposed to [26] where speech signals are modeled by an autoregressive model, we define $\mathbf{y}_f = [y_{1,f}, \ldots, y_{T,f}]^{\mathsf{T}}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_F] \in \mathbb{R}^{T\times F}$. The Mag-SepFormer encodes the received signals as

$$\mathbf{D}^{\mathsf{mag}} = \left|\mathbf{Y}^{\mathsf{T}}\right|, \tag{19}$$

where $|\cdot|$ is the element-wise magnitude operator. The estimated signal is then given by

$$\mathbf{M}^{\mathsf{mag}} = \mathfrak{f}_{\mathsf{relu}} \circ \mathfrak{f}_{\mathsf{conv-b}} \circ \mathfrak{f}_{\mathsf{ola}} \circ \mathfrak{f}_{\mathsf{conv}} \circ \mathfrak{f}_{\mathsf{p-relu}} \circ \mathfrak{f}_{\mathsf{sep}}\left(\mathbf{D}^{\mathsf{mag}}\right), \tag{20}$$

$$\mathbf{X}^{\mathsf{T}} = \mathbf{D}^{\mathsf{mag}} \odot \mathbf{M}^{\mathsf{mag}} \odot \exp\left(\jmath\angle\mathbf{Y}^{\mathsf{T}}\right), \tag{21}$$

where $\circ$, $\mathfrak{f}_{\mathsf{sep}}$, and $\mathfrak{f}_{\mathsf{p-relu}}$ are, respectively, the function composite, the SepFormer, and parametric ReLU [27] operations. The function $\mathfrak{f}_{\mathsf{c-branch}}(\cdot) = \mathfrak{f}_{\mathsf{sigmoid}}\big(\mathfrak{f}_{\mathsf{conv}}(\cdot)\big) + \mathfrak{f}_{\mathsf{tanh}}\big(\mathfrak{f}_{\mathsf{conv}}(\cdot)\big)$ with $\mathfrak{f}_{\mathsf{sigmoid}}$ and $\mathfrak{f}_{\mathsf{tanh}}$ being the sigmoid and hyperbolic tangent operations, respectively. Here, $\exp(\jmath\angle\mathbf{Y}^{\mathsf{T}})$ defines the noisy (unprocessed) phase information with $\exp(\cdot)$ being the exponential function, $\jmath = \sqrt{-1}$, and $\angle$ being the associated angle. The signal $\mathbf{X} \in \mathbb{C}^{T\times F}$ in (21) is then decoded back to time-domain via the inverse short-time Fourier transform.

It is worth noting that a time-domain signal with a high sampling rate and long duration results in $P >> K$ for the time-domain dual-path transformers and that computational complexity associated with $\mathfrak{f}_{\mathsf{inter-t}}$ in (4) is significantly larger than $\mathfrak{f}_{\mathsf{intra-t}}$ in (3). In contrast to the time-domain DPTNet and Sepformer, Mag-Sepformer employs a larger frame length $F > K$ to reduce the number of chunks (i.e., $P > T$) and that $P > F$ is set to strike a balance between the computation complexity of $\mathfrak{f}_{\mathsf{inter-t}}$ and $\mathfrak{f}_{\mathsf{intra-t}}$.

While aforementioned approaches, in general, employ dual-path transformers to reduce computational complexity, these approaches still rely on multiple transformers (i.e., $B > 2$). Furthermore, they employ either features extracted from only a single layer of convolution layer in (2) or the unprocessed magnitude spectrum in (19). The use of such encoded features are limiting since they contain largely noisy signals which are subsequently masked via (13) or (20), respectively.

III. The Proposed SMoLnet-T model

With reference to Fig. 2, the proposed small model on low SNR network with post transformer (SMoLnet-T) consists of a frame-wise CNN and a spectral combination transformer. The frame-wise CNN is based on the first ten layers of the SMoLnet, where it extracts high-resolution spectral feature maps. Since the spectral order in each frame is maintained, the subsequent transformer focuses on the crucial extracted feature maps based on the overall temporal information.

A. Frame-wise SMoLnet

With $y_{t,f}$ defined in (18), and that $\mathbf{y}_{\Re,t}$ and $\mathbf{y}_{\Im,t} \in \mathbb{R}^{F\times 1}$ are the real and imaginary component of $\mathbf{y}_t = [y_{t,1}, \ldots, y_{t,F}]^{\mathsf{T}}$, respectively, the proposed SMoLnet-T with $L = \log_2(F) + 1$ convolution layers is expressed as

$$\mathbf{Z}_t^{[l]} = \mathfrak{f}_{\mathsf{CNN}}^{[l]}\left(\mathbf{Z}_t^{[l-1]}\right). \tag{22}$$

Here, $1 \le l \le L$ is the layer index and $\mathbf{Z}_t^{[l]} = \left[\mathbf{z}_{t,1}^{[l]}, \ldots, \mathbf{z}_{t,F}^{[l]}\right] \in \mathbb{R}^{C\times F}$ consists the C spectral feature with

$$\mathbf{Z}_t^{[0]} = [\mathbf{y}_{\Re,t}, \mathbf{y}_{\Im,t}]^{\mathsf{T}} \tag{23}$$

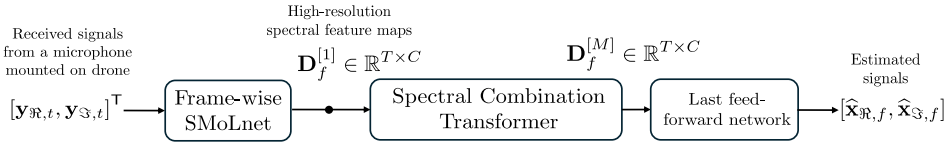


Fig. 2. Proposed direct-mapping approach with frame-wise SMoLnet and spectral combination transformer.

being the complex spectral input. The lth convolution layer in (22) is given by

$$\mathfrak{f}_{\mathsf{CNN}}^{[l]} = \mathfrak{f}_{\mathsf{relu}}^{[l]} \circ \mathfrak{f}_{\mathsf{ln}}^{[l]} \circ \mathfrak{f}_{\mathsf{di-conv}}^{[l]} \circ \mathfrak{f}_{\mathsf{pad}}^{[l]} \tag{24}$$

with $\mathfrak{f}_{\mathsf{relu}}^{[l]}, \mathfrak{f}_{\mathsf{ln}}^{[l]}, \mathfrak{f}_{\mathsf{di-conv}}^{[l]}$, and $\mathfrak{f}_{\mathsf{pad}}^{[l]}$ being the ReLU activation, layer normalization [28], convolution, and zero padding functions, respectively. Here, filter $\mathfrak{f}_{\mathsf{di-conv}}^{[l]}$ is of length $k = 3$, and a dilation rate $\mathcal{D}(l) = 2^{l-1}$ is set across the frequency dimension to achieve a receptive field larger than F. In addition, $\mathbf{Z}_t^{[l]}$ has been padded on both ends such that $\mathfrak{f}_{\mathsf{pad}}^{[l]}\left(\mathbf{Z}_t^{[l]}\right) \in \mathbb{R}^{C\times(F+\mathcal{D}(l)+1)}$. Hence, the fth node of the lth layer is achieved by leveraging the corresponding node of the $l-1$th layer. This process facilitates the positional consistency of the frequency information and maintains the high-resolution frequency information throughout. In contrast to employing a transformer that relies on adding sinusoidal encoding to the input or RNN to maintain positional consistency, the proposed CNN ensures position consistency while efficiently exploiting global high-resolution frequency information.

B. Spectral combination transformer

Since high-frequency spectral features have been achieved in $\mathbf{Z}_t^L \in \mathbb{R}^{C\times F}$ in (22), subsequent transformer layers are formulated to focus on the temporal aspect. Given $\mathbf{D}_f^{[1]} = \left[\mathbf{z}_{1,f}^{[L]}, \ldots, \mathbf{z}_{t,f}^{[L]}, \ldots, \mathbf{z}_{T,f}^{[L]}\right]^{\mathsf{T}} \in \mathbb{R}^{T\times C}$ with $\mathbf{z}_{t,f}^{[L]}$ from (22), the proposed SMoLnet-T employs M consecutive transformers

$$\mathbf{D}_f^{[m+1]} = \mathfrak{f}_{\mathsf{SCT}}^{[m]}\left(\mathbf{D}_f^{[m]}\right), \tag{25}$$

where the mth transformer is computed from

$$\mathfrak{f}_{\mathsf{SCT}}^{[m]} = \mathfrak{f}_{\mathsf{ln}_2}^{[m]} \circ \mathfrak{f}_{\mathsf{ffn}_2}^{[m]} \circ \mathfrak{f}_{\mathsf{ffn}_1}^{[m]} \circ \mathfrak{f}_{\mathsf{ln}_1}^{[m]} \circ \mathfrak{f}_{\mathsf{mha}}^{[m]} \circ \mathfrak{f}_{\mathsf{pe}}^{[m]}. \tag{26}$$

For succinctness in notation, we omit m henceforth. Similar to Sepformer [21] and Mag-Sepformer [20], we avoid the need for RNN in DPTNet via position-encoded spectral maps

$$\mathbf{D}'_f = \mathfrak{f}_{\mathsf{pe}}\left(\mathbf{D}_f\right) = \mathbf{D}_f + \mathbf{E}^{T\times C}, \tag{27}$$

where the tth row and cth column of the fixed position encoding $\mathbf{E}^{T\times C}$ is given by [19]

$$e_{t,c} = \begin{cases} \sin\left(\frac{t}{10000^{2c/C}}\right), & \text{if } c \text{ is even;} \\ \cos\left(\frac{t}{10000^{2c/C}}\right), & \text{if } c \text{ is odd.} \end{cases} \tag{28}$$

Compared with RNN and CNN models, where the sequence order is inherent, a transformer-based approach requires positional encoding such as (28) to achieve such information.

With $\mathbf{D}'_f$, the multi-head attention is given by

$$\mathbf{A}_f = \mathfrak{f}_{\mathsf{mha}}\left(\mathbf{D}'_f\right) = \mathfrak{f}_{\mathsf{cat}}\left([\mathbf{A}_{f,1}, \ldots, \mathbf{A}_{f,H}]\right)\mathbf{W}^{\mathsf{O}}, \tag{29}$$

where $\mathbf{W}^{\mathsf{O}} \in \mathbb{R}^{\frac{C}{H}\times C}$ is the output weights and

$$\mathbf{A}_{f,h} = \mathfrak{f}_{\mathsf{softmax}}\left(\frac{\mathbf{Q}_{f,h}\mathbf{K}_{f,h}^{\mathsf{T}}}{\sqrt{C}}\right)\mathbf{V}_{f,h} \tag{30}$$

is the single-head self-attention. Defining weights $\mathbf{W}_h^{\mathsf{Q}}, \mathbf{W}_h^{\mathsf{K}}, \mathbf{W}_h^{\mathsf{V}} \in \mathbb{R}^{C\times\frac{C}{H}}$ for the query, key, and value, respectively, we have

$$\mathbf{Q}_{f,h} = \mathbf{D}'_f\mathbf{W}_h^{\mathsf{Q}}, \mathbf{K}_{f,h} = \mathbf{D}'_f\mathbf{W}_h^{\mathsf{K}}, \mathbf{V}_{f,h} = \mathbf{D}'_f\mathbf{W}_h^{\mathsf{V}}. \tag{31}$$

Here, $0 \le h \le H$ denotes the head index with H being the number of heads. The hth head of the proposed attention mechanism in (30), therefore, focuses on the important feature map (out of the C features maps) in $\mathbf{D}_f^{[m]}$ for every t. Hence, $\mathbf{A}_f^{[m]}$ encapsulates the top H important feature maps to focus on, resulting in enhanced network training by highlighting the important spectral features within the temporal period T. The attention mechanism in (29) is translatable across frequencies since the proposed CNN in (22) maintains the frequency position consistency for the feature maps.

With $\mathbf{A}_f$ in (29) and $\mathbf{D}'_f$ in (27), the transformer estimates the feature maps for its subsequent transformer via

$$\dot{\mathbf{D}}_f = \mathfrak{f}_{\mathsf{ln}_1}\left(\mathbf{A}_f + \mathbf{D}'_f\right), \tag{32}$$

$$\ddot{\mathbf{D}}_f = \mathfrak{f}_{\mathsf{ffn}_1}\left(\dot{\mathbf{D}}_f\right) = \mathfrak{f}_{\mathsf{relu}}\left(\dot{\mathbf{D}}_f\mathbf{W}_1 + \mathbf{b}_1\right), \tag{33}$$

$$\dddot{\mathbf{D}}_f = \mathfrak{f}_{\mathsf{ffn}_2}\left(\ddot{\mathbf{D}}_f\right) = \mathfrak{f}_{\mathsf{relu}}\left(\ddot{\mathbf{D}}_f\mathbf{W}_2 + \mathbf{b}_2\right), \tag{34}$$

$$\ddddot{\mathbf{D}}_f = \mathfrak{f}_{\mathsf{ln}_2}\left(\dot{\mathbf{D}}_f + \dddot{\mathbf{D}}_f\right), \tag{35}$$

where $\mathfrak{f}_{\mathsf{ffn}_1}$ and $\mathfrak{f}_{\mathsf{ffn}_2}$ are feed-forward networks corresponding to weights $\mathbf{W}_1 \in \mathbb{R}^{C\times 4C}$, $\mathbf{W}_2 \in \mathbb{R}^{4C\times C}$, and biases $\mathbf{b}_1 \in \mathbb{R}^{C\times 1}$ and $\mathbf{b}_2 \in \mathbb{R}^{4C\times 1}$, respectively. Here, $\ddddot{\mathbf{D}}_f$ serves as the input to the subsequent transformer (i.e, $\mathbf{D}_f^{[m+1]}$ in (25)).

With $\mathbf{D}_f^{[M]}$ being the output of the last transformer in (35), the estimated speech is given by

$$[\widehat{\mathbf{x}}_{\Re,f}, \widehat{\mathbf{x}}_{\Im,f}] = \mathbf{D}_f^{[M]}\mathbf{W}_{\mathsf{last}} + \mathbf{b}_{\mathsf{last}}, \tag{36}$$

$$\widehat{\mathbf{x}} = \widehat{\mathbf{x}}_{\Re,f} + \jmath\widehat{\mathbf{x}}_{\Im,f}, \tag{37}$$

where $\mathbf{W}_{\mathsf{last}} \in \mathbb{R}^{C\times 2}$ and $\mathbf{b}_{\mathsf{last}} \in \mathbb{R}^{1\times 2}$ are, respectively, the weights and bias of the last feed-forward neural network. Here, $\widehat{\mathbf{x}}_{\Re,f}$ and $\widehat{\mathbf{x}}_{\Im,f}$ are the real and imaginary components of the estimated signal, respectively. In contrast to the mask [16] in (13), the proposed transformer performs a direct mapping from the frame-wise spectral maps (i.e. $\mathbf{Z}_t^L \in \mathbb{R}^{C\times F}$ in (22)) to the output. This allows the overall

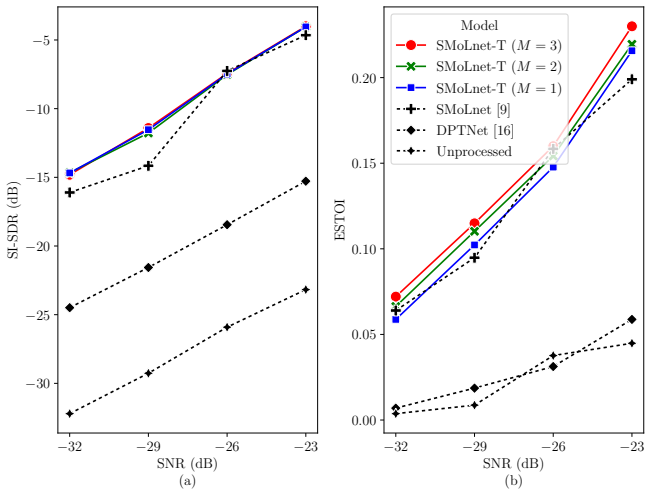


Fig. 3. Speech enhancement performance in terms of (a) SI-SDR [15] and (b) ESTOI [18].

transformer network to decide, at a higher level of abstraction, which subset of high-resolution spectral maps generated by the frame-wise SMoLnet is crucial at t. This is achieved by the consistency in position and the high-resolution frequency information retained by the proposed frame-wise SMoLnet.

IV. Experiment Results

We evaluate our proposed model using a noise dataset collected from a hovering drone, and the noise signals are synthetically mixed with speech utterances from WSJ0-84 [29]. An hour of noise was used, and out of the 8714 speech utterances, 6382, 1166, and 1052 were allocated for training, validation, and testing, respectively. The speakers and SNRs for training, validation, and testing were varied to validate the model's generalization for out-of-training speakers and SNRs. More specifically, for training, validation, and testing, the SNRs (in decibels) are drawn from $\{-30, -27, -24, -21, -18\}$, $\{-31, -28, -25, -22, -19\}$, and $\{-32, -29, -26, -23\}$, respectively. A batch size of two was selected to train the proposed SMoLnet-T and the baselines DPTnet [16] and SMoLnet [9], which were recently evaluated in [13] as two of the superior models for single-channel drone auditions. SMoLnet-T requires low GPU memory usage, and a larger batch size and length can be selected. However, for a fair comparison with the baselines, which require a significant amount of GPU memory during training, we fixed the batch size to two. For SMoLnet-T and SMoLnet, we set $C = 64$, a Hamming window of length 2048 with 50% overlap, resulting in $F = 1025$ and $T = 123$. For SMoLnet-T, $M \in \{1, 2, 3\}$ was used to highlight the importance of the temporal context achieved via the spectral combination transformer. An Adam optimizer [30] was used with a learning rate of 0.001 and the model with the best validation result over 100 epochs was selected for testing.

We evaluate the speech enhancement performance in terms of speech intelligibility via ESTOI [18] and speech distortion via SI-SDR [15]. As shown in Fig. 3, all models exhibited improved ESTOI and SI-SDR over the received (unprocessed) signal. SMoLnet-T generally achieves lower speech distortion than the baselines for all tested SNRs. In particular, for out-of-training SNR of -32 dB, SMoLnet-T achieves significantly lower speech distortion with an SI-SDR improvement of 2.3 and 9.3 dB over SMoLnet and DPTNet, respectively. Although the SI-SDR is similar across the number of transformers M in SMoLnet-T, ESTOI increases with M. With $M = 3$, SMoLnet-T achieves an ESTOI improvement of 0.02 over SMoLnet at an SNR of -23 dB. The improvement in both speech intelligibility and speech distortion by SMoLnet-T highlights the efficacy of the proposed spectral combination transformer in leveraging temporal context over the CNN approach in SMoLnet. Furthermore, the overall higher performance for SMoLnet-T and SMoLnet compared to DPTnet highlights the viability of direct mapping over masking under low SNR scenarios.

TABLE I
Computational resources requirement for a batch size of 2 each with 64 k samples on the Nvidia RTX6000 Ada.

Method	Train time (mins/epoch)	Model size	GPU mem. (GB)
SMoLnet-T ($M = 1$)	2.8	192 k	4.7
SMoLnet-T ($M = 2$)	4.3	254 k	5.6
SMoLnet-T ($M = 3$)	6.0	317 k	6.7
SMoLnet [9]	1.6	249 k	2.8
DPTnet [16]	40.0	2.6 M	40.0

We evaluate the required computation resources, including the training time per epoch, model size in terms of the number of learnable parameters, and the amount of GPU memory usage used during training. The Nvidia RTX6000 Ada, which has 48 GB of GPU available memory, is used. As shown in Table I, the proposed SMoLnet-T requires significantly lower computational resources than DPTNet. More specifically, SMoLnet-T with $M = 1$ requires seven times lower GPU memory to train and is 14 times faster for an epoch. These results highlight that SMoLnet-T can be trained with longer sequences, requires lower GPU resources, and achieves faster training iterations. Although the computation needs are higher than SMoLnet, SMoLnet-T achieves a larger context length ($T = 123$) compared to the nine frames context in SMoLnet, which is beneficial for improving speech intelligibility and reducing speech distortion.

V. Conclusion

We proposed the SMoLnet-T, which incorporates a frame-wise CNN with newly formulated spectral-attention transformers. These transformers achieve higher efficacy over CNN in SMoLnet by leveraging high-resolution frequency spectral maps from its frame-wise SMoLnet and longer temporal context. Experiment results demonstrate that SMoLnet-T achieves improved speech intelligibility and lower speech distortion under low SNR over DPTNet with reduced computational requirements.

REFERENCES

[1] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, pp. 508–519, 2022.

[2] M. Wakabayashi, H. G. Okuno, and M. Kumon, "Multiple sound source position estimation by drone audition based on data association between sound source localization and identifications," *IEEE Robot. Autom. Letters*, vol. 5, no. 2, pp. 782–789, 2020.

[3] T. Morito, O. Sugiyama, R. Kojima, and K. Nakadai, "Partially shared deep neural network in sound source separation and identification using a UAV-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robot. Syst.*, 2016, pp. 1299–1304.

[4] C. Premachandra and Y. Kunisada, "GAN based audio noise suppression for victim detection at disaster sites with UAV," *IEEE Trans. Services Comput.*, vol. 17, no. 1, pp. 183–193, 2024.

[5] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Sound source tracking by drones with microphone arrays," in *Proc. IEEE/SICE Int. Symp. Syst. Integration*, 2020, pp. 796–801.

[6] W. Zhang, A. W. H. Khong, and P. A. Naylor, "Adaptive inverse filtering of room acoustics," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 2008, pp. 788–792.

[7] B. Yen, Y. Li, and Y. Hioka, "Rotor noise-aware noise covariance matrix estimation for unmanned aerial vehicle audition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2491–2506, 2023.

[8] B. Yen, Y. Hioka, and B. Mace, "Improving power spectral density estimation of unmanned aerial vehicle rotor noise by learning from non-acoustic information," in *Proc. Int. Work. Acoust. Signal Enhanc.*, 2018, pp. 545–549.

[9] Z.-W. Tan, A. H. T. Nguyen, and A. W. H. Khong, "An efficient dilated convolutional neural network for UAV noise reduction at low input SNR," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 1885–1892.

[10] Z.-W. Tan, A. H. T. Nguyen, Y. Liu, and A. W. H. Khong, "Multichannel noise reduction using dilated multichannel U-net and pre-trained single-channel network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 266–270.

[11] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent.*, 2016, p. 1–13.

[12] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv Prepr. arXiv1609.03499*, pp. 1–15, 2016.

[13] D. Mukhutdinov, A. Alex, A. Cavallaro, and L. Wang, "Deep learning models for single-channel speech enhancement on drones," *IEEE Access*, vol. 11, pp. 22 993–23 007, 2023.

[14] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU, ITU-T Rec. P. 862, 2000.

[15] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[16] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, 2020.

[17] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–19.

[18] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Info. Process. Syst.*, 2017, pp. 5998–6008.

[20] D. de Oliveira, T. Peer, and T. Gerkmann, "Efficient transformer-based speech enhancement using long frames and STFT magnitudes," in *Proc. Interspeech*, 2022, pp. 2948–2952.

[21] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 21–25.

[22] S. H. Nawab and T. F. Quatieri, "Short-time Fourier transform," in *Adv. Top. Signal Process.* Upper Saddle River, New Jersey: Prentice-Hall, 1987, pp. 289–337.

[23] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[24] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[25] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. Waibel, "Self-attentional acoustic models," in *Proc. Interspeech*, 2018, pp. 3723–3727.

[26] B. Liu, V. G. Reju, and A. W. H. Khong, "A linear source recovery method for underdetermined mixtures of uncorrelated ar-model signals without sparseness," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 4947–4958, 2014.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," *in Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1026–1034, 2015.

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[29] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," *Linguist. Data Consortium, Philadelphia*, 2007.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 127–142.

Drone audition: implementation of an indoor multi-drone system for sound source tracking

Benjamin Yen*† and Kazuhiro Nakadai *
* Nakadai Lab, Institute of Science Tokyo, Tokyo, Japan

***Abstract*—In this study, we introduce an indoor multi-drone system implementation for tracking sound sources. With the ease of algorithm development and rapid prototyping in mind, the proposed system provides a user-friendly alternative over large-scale drone swarms by utilising a number of mini-sized Crazyflie drones, designed to operate under the open-source Robotics Operating System 2 (ROS 2) based Python environment. Equipped with a microphone array, sound source tracking is realised by triangulating the spectral response from the Multiple Signal and Classification (MUSIC) algorithm combined with particle filtering. We evaluated the system using real-life demonstrations, and results show that the system successfully tracks a single moving sound source in real-time.**

I. Introduction

In many applications, acquiring and relaying information in a time-efficient manner is crucial. Drones, with the amount of development in recent years, have shown to be an ideal platform for acquiring a broad range of information [1]. Furthermore, it has been shown to be especially useful when navigating through difficult-to-reach environments [2], which is commonly encountered in search and rescue [3] and military surveillance [4]. Amongst the range of sensing technologies available for drones, vision (i.e. via a video camera) is most commonly used. However, there are many situations where the effectiveness of vision information can be severely impeded, including low lighting, excessive vegetation and even collapsed buildings. On the other hand, audio information is immune to many of such conditions and, therefore, has the potential to be a viable alternative [5], [6]. For this reason, drone audition, where drones are used as mobile acoustic sensors to perceive their surrounding environment and perform relevant tasks using audible sound, has gained increasing attention [7].

Among existing drone audition-related studies, many of them are carried out in areas such as sound source separation [8]–[11] and sound source localisation [12]–[15]. For sound source localisation (and tracking), which is the main focus of this study, most existing work uses a microphone array [16] that is mounted on the drone to obtain spatial information of the sound source(s) that is unavailable with single microphone approaches. In addition, due to the high levels of rotor noise, which is one of the ongoing challenges in drone audition [10], the use of microphone arrays helps suppress such noise influences. For example, several studies utilised a noise correlation matrix (NCM) to reduce drone noise influences and, in turn, improve localisation performance based on the MUltiple SIgnal Classification (MUSIC) algorithm. Here, the NCM estimation ranges from simple frame averaging of the input signal's spectrogram [12], [17] to taking UAV input states and following a Gaussian process regression [18]. Other methods opted to directly reduce drone noise at the microphone inputs by utilising a Wiener postfilter (WF) [15], [19], beamforming [20], and DNN-based methods [14], [21].

As shown, most studies in sound source localisation for drones focused on single-drone (or single microphone array) approaches. While it allows the *direction* of the sound source to be tracked, it is not capable of tracking the exact *location* of the target in terms of 3D coordinates, which is often desired in practical scenarios. To address this, authors from [22], [23] made use of multiple microphone arrays as multiple drones with a single array attached. By triangulating the localisation outputs of each drone, we obtain the exact location of the target sound source. This includes triangulation of the target source directions from each drone that is processed through a Gaussian sum filter [22] or by integrating the MUSIC spectrum from each drone to be used for particle filtering, known as Particle Filtering with Integrated MUSIC (PAFIM) [23]. Recently, a method that allows the drones to autonomously navigate themselves such that the microphone arrays are in the optimal position for localising the sound source has also been proposed [24]. A unique aspect of these studies is that the sound source's location can be tracked continuously, whether the sound source is fixed in location or moving.

While these recent studies in multi-drone sound source tracking have shown promising results, most are evaluated under simulation. Several factors hinder the real-life implementation of such a system. First, due to the multi-drone nature, a system that allows time-synchronised inputs from multiple drones is required for the triangulation process to be carried out effectively. Implementing such a system outdoors is challenging. Furthermore, it is often difficult to gain access to an environment that allows the development of a system, which, on top of the implementation aspect, also includes extensive testing. For example, in Japan, airspace is bounded by various restrictions, and thus, many public spaces prohibit the presence of drones [25]. However, for the development and evaluation of multi-drone drone audition-related research, it may not always be necessary to make use of a fully outdoor system. Instead, an indoor system that is free of such restrictions, where the drone's location can be readily tracked within a confined space, is a perfectly viable alternative for

†B. Yen is an International Research Fellow of the Japan Society for the Promotion of Science (JSPS) at the Nadakai Lab, Institute of Science Tokyo.

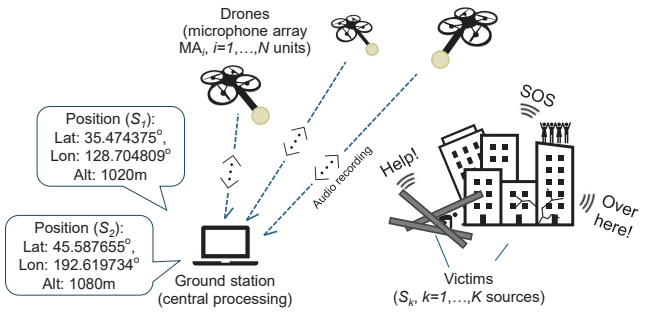


Fig. 1. Overview of sound source tracking using multiple drones.

rapid development and real-life evaluation of such algorithms.

Fortunately, such a system can be achieved with the *Crazyflie* nano drone [26]. The small-sized Crazyflie drone can be equipped with a microphone array (known as the *audio deck*) developed specifically for the drone [27] to perform drone audition-related applications. Therefore, this study proposes an indoor multi-drone system utilising the Crazyflie to track static or moving sound sources in real-time. By utilising the *audio deck* and a motion-capture-based *lighthouse positioning system* for location and orientation information of each drone, multi-drone 3D sound source tracking can be achieved. We note that while this study focuses on sound source tracking, such a system is also adaptable for the development of sound source separation or enhancement-related applications. To demonstrate the system's real-time capabilities, we provide experimental evaluation via video demonstration. In summary, the main contributions of this study are:

1) the implementation of an indoor multi-drone audition system.
2) the practical realisation of a multi-drone sound source tracking based on the indoor system.

As such, this study focuses on the implementation of multi-drone sound source tracking, rather than theoretical development. Due to the unique nature of drone audition using multiple drones in an indoor setting, in this study, we describe the design challenges and solutions addressing these challenges.

II. Problem Setup

Fig. 1 illustrates an overview of this study. The problem assumes N drones, each mounted with a M sensor microphone array, resulting in $\mathrm{MA}_1, \ldots, \mathrm{MA}_N$ microphone arrays. Each ith microphone array receives K target sources $S_k(\omega, f)$, *spatially coherent* rotor noise from U drone rotors ($N_u(\omega, f)$) and *spatially incoherent* ambient noise, such as air conditioning $\mathbf{v}(\omega, f)$, in an indoor environment. For the multi-drone scenario, noise from neighbouring drones $N_{j\in N(j\neq i)}(\omega, f)$ (i.e., all drones apart from the ith drone) are also accounted for [28]. Assuming that the drones are usually a distance apart from each other, we assume that these drone noises are point sources propagating from the jth drone to the ith microphone array. ω and f denote the angular frequency and the frame index, respectively. The short-time Fourier transform (STFT) of the ith microphone array's (MA_i) input signals are expressed in vector form as

$$\begin{aligned}\mathbf{x}_i(\omega, f) &:= \left[X_{i,1}(\omega, f), \quad \cdots, \quad X_{i,M}(\omega, f)\right]^T \\ &= \sum_{k=1}^{K} \mathbf{a}_{i,S_k}(\omega, \phi_{i,S_k}, \theta_{i,S_k}) S_k(\omega, f) \\ &+ \sum_{u=1}^{U} \mathbf{a}_{i,u}(\omega, \phi_{i,u}, \theta_{i,u}) N_u(\omega, f) \\ &+ \sum_{j\in N(j\neq i)} \mathbf{a}_{i,j}(\omega, \phi_{i,j}, \theta_{i,j}) N_j(\omega, f) + \mathbf{v}(\omega, f), \quad (1)\end{aligned}$$

where ^{T} denotes the transpose. $[\phi_{i,S_k}, \theta_{i,S_k}]$ and $[\theta_{i,u}, \phi_{i,u}]$ are the spherical directions (i.e. azimuth and elevation, respectively) radiated from the kth target and the uth rotor, respectively, to MA_i in the microphone array's own body coordinates. $\mathbf{a}_i(\omega, \phi_i, \theta_i)$ is the steering vectors between an arbitrary source and each microphone m in MA_i. Assuming overdetermined cases (i.e. $M \geq U + K + N - 1$ for each ith drone/microphone array) the system is tasked to track the location of the K target source signals [22] from the $N \times M$-channel noisy recordings.

In addition to the microphone input signals, assuming $x(f), y(f)$, and $z(f)$ respectively indicate the location in 3D Cartesian coordinates, and $\phi(f), \theta(f)$ and $\psi(f)$ respectively indicate yaw, pitch, and roll angles, the location of each microphone array MA_i at frame f is described as

$$\boldsymbol{m}_i(f) = [\boldsymbol{m}_{i,xyz}(f)^T, \boldsymbol{m}_{i,\phi\theta\psi}(f)^T]^T, \quad (2)$$

$$\boldsymbol{m}_{i,xyz}(f) = [x_i(f), y_i(f), z_i(f)]^T, \quad (3)$$

$$\boldsymbol{m}_{i,\phi\theta\psi}(f) = [\phi_i(f), \theta_i(f), \psi_i(f)]^T. \quad (4)$$

Finally, the location of the kth sound source is defined as

$$\boldsymbol{e}_k(f) = [x_k(f), y_k(f), z_k(f)]^T. \quad (5)$$

III. System and Methodology

Fig. 2 shows a flowchart of the proposed multi-drone sound source tracking framework. Prior to introducing the proposed indoor multi-drone platform, including its hardware, we first present the method driving the multi-drone sound source tracking. The method is based on PAFIM [23], and the following sections describe each component of the sound source tracking system.

A. MUSIC

PAFIM, as is apparent from the name, uses the MUSIC algorithm. MUSIC performs localisation by first performing Eigenvalue decomposition on the microphone array signal's correlation matrix $\mathbf{R}(\omega, t)$, calculated as

$$\begin{aligned}\mathbf{R}(\omega, t) &= \frac{1}{T_R} \sum_{\tau=f}^{f+T_R-1} \mathbf{x}(\omega, \tau)\mathbf{x}^H(\omega, \tau) \\ &= \mathbf{E}(\omega, t)\mathbf{\Lambda}(\omega, t)\mathbf{E}^{-1}(\omega, t), \quad (6)\end{aligned}$$

where T_R is the number of frames, and ^{H} is the Hermitian conjugate. By assuming orthogonality between the target sound

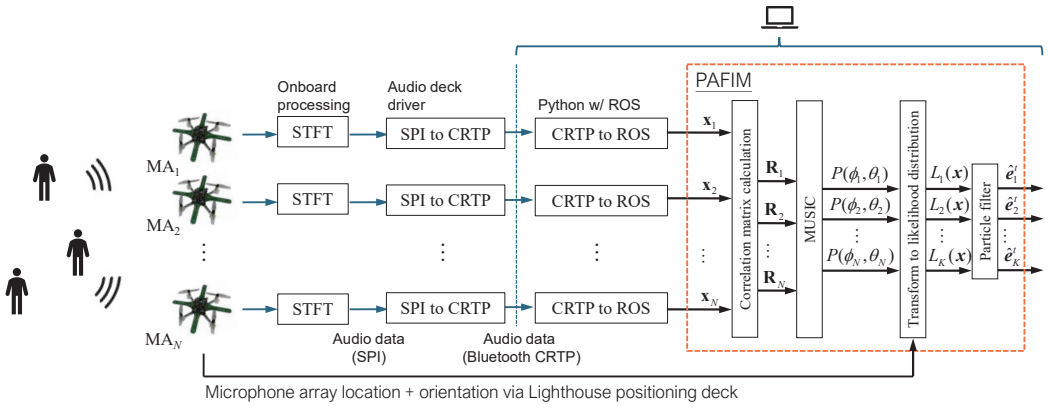


Fig. 2. Flowchart of indoor multi-drone system with sound source tracking using PAFIM.

source and the noise subspace, the spectrum indicating the likelihood of the target source's direction (at each time step t) is calculated using eigenvectors $\mathbf{E}_N(\omega, t) \in \mathbf{E}(\omega, t)$ which corresponds to the eigenvalues in the noise subspace of $\mathbf{\Lambda}(\omega, t)$ as

$$P(\phi, \theta, t) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} \frac{\|\mathbf{a}^H(\omega, \phi, \theta)\mathbf{a}(\omega, \phi, \theta)\|}{\|\mathbf{a}^H(\omega, \phi, \theta)\mathbf{E}_N(\omega, t)\|}. \tag{7}$$

Here, $\mathbf{a}(\omega, \phi, \theta)$ is the steering vector of a sound signal arriving from direction (ϕ, θ). ω_L and ω_H indicate the minimum and maximum frequency bins, respectively.

B. Particle Filtering with MUSIC

Assuming an arbitrary location $\boldsymbol{x}$, its direction relative to drone location $\boldsymbol{m}_i$ is defined as (ϕ_i, θ_i). Using this, once the MUSIC spectrum $P_i(\phi, \theta)$ for each i-th drone is calculated, a location likelihood $L(\boldsymbol{x})$ that triangulates $P_i(\phi, \theta)$ to $\boldsymbol{x}$ can be calculated as

$$L(\boldsymbol{x}) = \sum_i P_i(\tilde{\phi}_i^{\text{round}}, \tilde{\theta}_i^{\text{round}}), \tag{8}$$

$$\tilde{\phi}_i^{\text{round}} = \text{round}(\tilde{\phi}_i), \qquad \tilde{\theta}_i^{\text{round}} = \text{round}(\tilde{\theta}_i), \tag{9}$$

$$\begin{bmatrix} \cos\tilde{\phi}_i \cos\tilde{\phi}_i \\ \sin\tilde{\phi}_i \cos\tilde{\phi}_i \\ \sin\tilde{\phi}_i \end{bmatrix} = \boldsymbol{R}_n^{-1} \begin{bmatrix} \cos\phi_i \cos\phi_i \\ \sin\phi_i \cos\phi_i \\ \sin\phi_i \end{bmatrix}, \tag{10}$$

where $\text{round}(\cdot)$ rounds the direction with respect to the resolution of transfer function $\boldsymbol{a}(\omega, \phi_S, \theta_S)$. $\boldsymbol{R}_i$ is the rotation matrix that converts between the i-th drone's body coordinates $\boldsymbol{m}_{i,\phi\theta\psi}$ and world coordinates.

Following, $L(\boldsymbol{x})$ can be used for calculating the weighting factor in a particle filter setting, assuming a distribution of potential locations $\boldsymbol{x}$. For each time step t, I_P number of particles is utilised. We define the state $\boldsymbol{x}_t^{i_\text{P}} = [x_t^{i_\text{P}}, y_t^{i_\text{P}}, z_t^{i_\text{P}} \dot{x}_t^{i_\text{P}}, \dot{y}_t^{i_\text{P}}, \dot{z}_t^{i_\text{P}}]^T$, which includes the 3-dimentional location and velocity of the particle i_P, and $w_t^{i_\text{P}}$ denotes the weight of the particle. We then employ an excitation-damping model as the prediction model

$$\boldsymbol{x}_t^{i_\text{P}} = \boldsymbol{F}\boldsymbol{x}_{t-1}^{i_\text{P}} + \boldsymbol{H}\boldsymbol{v}, \tag{11}$$

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{I} & T\boldsymbol{I} \\ \boldsymbol{O} & a\boldsymbol{I} \end{bmatrix}, \qquad \boldsymbol{H} = \begin{bmatrix} \boldsymbol{O} \\ b\boldsymbol{I} \end{bmatrix}, \tag{12}$$

$$\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{13}$$

where $\boldsymbol{I} \in \mathbb{R}^{3\times3}$ and $\boldsymbol{O} \in \mathbb{R}^{3\times3}$ denote the identity matrix and zero matrix, respectively. Parameter a controls the ratio of velocity to carry on from the previous time step $t-1$ and parameter b controls the level of excitation of particles to introduce in time step t. As mentioned, the location likelihood $L(\boldsymbol{x}_{t,pos}^{i_\text{P}})$ is used to calculate the weighting of each particle, defined as

$$w_t^{i_\text{P}} = w_{t-1}^{i_\text{P}} \frac{L(\boldsymbol{x}_{t,pos}^{i_\text{P}})}{\sum_i L(\boldsymbol{x}_{t,pos}^{i_\text{P}})}, \tag{14}$$

where $\boldsymbol{x}_{t,pos}^{i_\text{P}} = [x_t^{i_\text{P}}, y_t^{i_\text{P}}, z_t^{i_\text{P}}]^T$. Finally, using the prediction model and the weightings obtained, the sound source location at time step t is estimated as

$$\boldsymbol{e}(t) = \sum_{i_\text{P}=1} w_t^{i_\text{P}} \boldsymbol{x}_t^{i_\text{P}}. \tag{15}$$

Further details regarding PAFIM, including its initialisation can be found in [23].

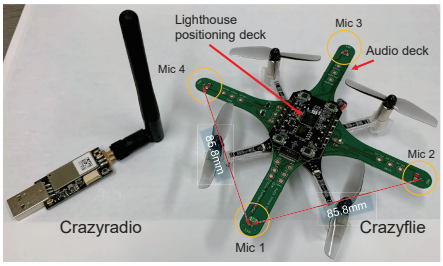


Fig. 3. Crazyradio (left) and Crazyflie with the audio deck and lighthouse positioning deck attached (right).

IV. System Implementation

This section introduces the proposed implementation of an indoor multi-drone sound source tracking system. We first

introduce the Crazyflie drone, including any sensor and hardware extensions used. Following, we introduce the relevant equipment and their arrangement in the indoor space for real-time 3D sound source tracking in a stable and robust manner.

A. Crazyflie Drone

The Crazyflie is a 92×92×29mm (motor-to-motor) small-sized nano drone specialised for research development (see Fig. 3). One of its features is that the drones operate using the Robotics Operation System 2 (ROS 2), which is a set of open-source software libraries and tools that allows the development of algorithms and systems for robot-related applications [29]. This allows the development of real-time applications with time-synchronised data input from multiple sensors and subsequent processing.

Another feature of the Crazyflie drone is the *extension decks*, which allow various types of sensors to be easily equipped for the Crazyflie. Some examples of such extension decks include the greyscale camera with the *AI-deck* or the *Multi-ranger Deck* to give the drones obstacle avoidance capabilities. In this study, we utilise the *Lighthouse positioning deck*, which, operating along with the *Lighthouse V2 base station* motion capture camera, provides the Crazyflie drones with real-time location (to the nearest 2-4cm with 0.3mm precision [30]) and orientation information. As mentioned in Section I, although there are no official microphone array extension decks available for Crazyflie, authors in [27] have developed a custom-designed open-source microphone array deck (known as the *Audio deck*). For this study, the audio deck is manufactured as a 4mm thick PCB equipped with four micro-electromechanical system (MEMS) microphones arranged in a 85.8×85.8mm square (see Fig. 3). To help offload some processing for the ground station PC, the audio deck is equipped with an STM32F446 microcontroller, which performs STFT on the recorded audio signals prior to data transmission.

B. Indoor Multi-Drone Sound Source Tracking System

Fig. 4 shows an overview of the indoor multi-drone system platform. The system consists of a swarm of Crazyflie drones. Each drone communicates with the central ground station (i.e. laptop PC) via the Crazyradio (see Fig. 3). This includes transmission of the audio, position and orientation data from the Crazyflie to the ground station and controller command from the ground station to each of the drones for automated drone control. As mentioned in Section IV-A, the Lighthouse positioning deck requires the Lighthouse V2 base station in order for real-time position and orientation data to be provided. As such, four base stations are used, each positioned at a corner of the experiment environment (see Fig. 4).

As shown in Fig. 2, data transmission from the Crazyflie to the ground station is carried out by first passing the microphone signals via serial peripheral interface (SPI) protocol, where data is sent to the STM32F446 microcontroller. Following, the data is then set to the ground station via the Crazyradio with a custom Bluetooth protocol known as the Crazy RealTime Protocol (CRTP) [27]. Due to the data size of the audio data from the Crazyflie's microphone array in STFT form, each drone requires pairing with a single Crazyradio. Consequently, this increases the risk of interference between different radio signals. To combat this issue, the Crazyradios are separated as far away as possible (as shown in Fig. 4).

TABLE I
AUDIO RECORDING AND PAFIM PARAMETERS.

Variable	Value
Sampling rate	48 kHz
Audio resolution	16 bit
STFT length	2048
ω_L	200 Hz
ω_H	4000 Hz
a	0.9
b	0.3
I_{P}	500

V. EVALUATION: INDOOR DEMONSTRATION

In this section, we evaluate the sound source tracking performance of the indoor multi-drone system as a proof of concept through demonstration, using the PAFIM method with array placement planning as the tracking algorithm. The experimental settings for the demonstration are described in Section V-A, with some discussions presented in Section V-B.

A. Experimental Settings

Fig. 4 shows a top view of the experimental settings, including the initial locations of the target sound source and the drones. The setting consists of a single target sound source emitting male speech, with four drones placed around the sound source. In this initial study, we evaluate the feasibility of the indoor sound source tracking system. Therefore, the target source was emitted loud enough for MUSIC to recognise the target as the primary sound source (i.e. the input signal-to-noise ratio (SNR) ≥ 5 dB). The target sound source moved around the room randomly for around $1-3$ minutes. The Crazyflie drones were flying at a fixed height of 1.2 m. Details regarding the audio recording and PAFIM-related parameters are shown in Table I. We note that despite the STFT parameter settings, for time-synchronisation between all sensors, all inputs were sampled at 10 Hz during processing. A total of five demonstration runs were carried out.

For the ground station, we specifically use the ROS 2 Galactic with Python version 3.8.10 under the Ubuntu 20.04 Focal Rossa operating system.

B. Results and Discussion

For this preliminary study, we evaluate the system through video demonstrations. Here, we discuss the performance using the following demonstration video*. Overall, PAFIM is able to successfully localise and track the target sound source accurately. In the case where the target sound source is continuously moving, whether quick or slow, the system is able to track and update the correct target source location. There are occasions where the estimated location deviates from the actual location. However, as long as the target sound source

*https://tinyurl.com/3mvzbf3n

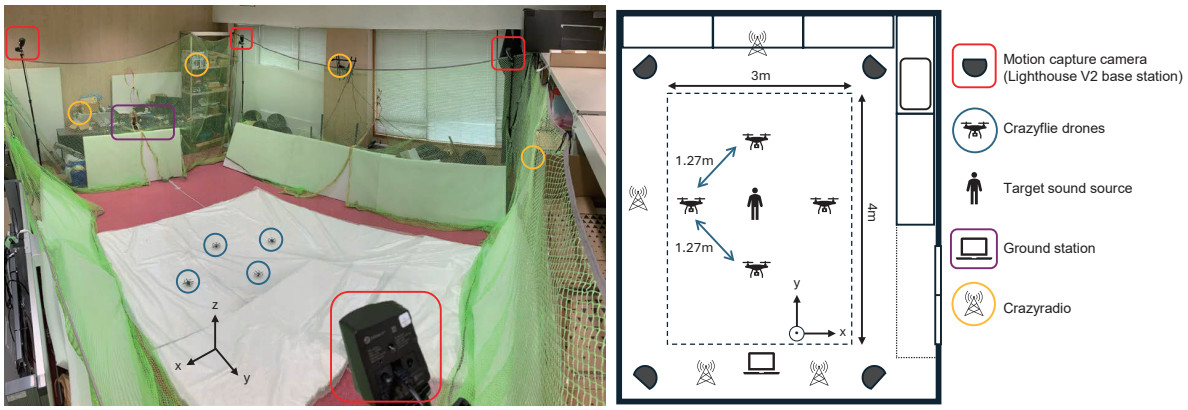


Fig. 4. Indoor multi-drone sound source tracking system (left) and experimental setting (right). Note that while the ground station computer is not shown in the figure, the location is highlighted by the purple box regardless.

continues to make sound, PAFIM will eventually converge to the correct location.

Due to the nature of the particle filter, after initialisation, it requires a few iterations for the particles to converge towards the target sound source position. As such, a minor lag occurs between the target source sounds making a sound and PAFIM updating the new target location. When the target sound source is moving, it also takes a few time frames to catch up to the target location. However, the lag is only apparent if the target source moves quickly. The current system is configured such that if it cannot detect the presence of the target sound source after five time frames, the particle filter will be reset. Therefore, as apparent in the demonstration, when the target stops making sound, the particle filter will reinitialise upon detecting the target sound again.

Despite the performance of the PAFIM algorithm, there are some factors that require addressing. As apparent from the video, the sound source tracking is only carried out in the xy plane, and the altitude of the target sound source is not estimated. This is due to the planar nature of the microphone array on the Crazyflies, where it is relatively difficult to distinguish the elevation direction of the target sound, resulting in altitude estimations with insufficient accuracy. Another issue includes the lack of explicit drone noise suppression for sound source tracking. Therefore, in this initial study, this is overcome by ensuring the target source being relatively loud. Addressing these issues, including a redesign of the microphone array to accommodate 3D sound source tracking and implementation of drone noise reduction algorithms, remains as future work.

The remaining demonstration videos, each showing similar levels of performance, can be found here[†].

VI. Conclusion

This study introduces a practical implementation of an indoor multi-drone system based on the Crazyflie drone and ROS 2 for real-time sound source location tracking. By equipping each Crazyflie with a custom-designed microphone array and the sensors for positioning and orientation information, a system that tracks audible sound sources is realised. Evaluation through experimental demonstrations showed the system had successfully tracked a single moving sound source in real-time. Future work includes addressing the lack of a drone noise suppression algorithm and enabling tracking of the target sound source in all three spatial dimensions.

[†]https://tinyurl.com/23cz9e5x

Acknowledgment

This research was supported by the commissioned research fund provided by F-REI (JPFR23010102) and JSPS KAKENHI through Grant No. 22F22769 and Grant No. 22KF0141. The authors would like to extend their thanks to Frederike Dümbgen and Adrien Hoffet for their valuable advice and input in replicating the audio deck for the Crazyflie.

References

[1] A. Koubaa and A. Azar, Eds., *Unmanned Aerial Systems*. Elsevier, 2021.

[2] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "Dregon: Dataset and methods for uav-embedded sound source localization," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 1–8.

[3] S. Grogan, R. Pellerin, and M. Gamache, "The use of unmanned aerial vehicles and drones in search and rescue operations – a survey," in *Proceedings of the PROLOG*, Sep. 2018, pp. 1–13.

[4] M. A. Ma'sum, M. K. Arrofi, G. Jati, *et al.*, "Simulation of intelligent unmanned aerial vehicle (UAV) For military surveillance," in *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Sep. 2013, pp. 161–166.

[5] K. Hoshiba, O. Sugiyama, A. Nagamine, R. Kojima, M. Kumon, and K. Nakadai, "Design and assessment of sound source localization system with a uav-embedded microphone array," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 154–167, 2017.

[6] S. V. Sibanyoni, D. T. Ramotsoela, B. J. Silva, and G. P. Hancke, "A 2-d acoustic source localization system for drones in search and rescue missions," *IEEE Sensors Journal*, vol. 19, no. 1, pp. 332–341, 2018.

[7] K. Nakadai and H. G. Okuno, "Robot audition and computational auditory scene analysis," *Advanced Intelligent Systems*, vol. 2, no. 9, p. 2 000 050, 2020.

[8] Z.-W. Tan, A. H.-T. Nguyen, and A. W.-H. Khong, "An efficient dilated convolutional neural network for UAV noise reduction at low input SNR," in *2019 Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA)*, Nov. 2019, pp. 1885–1892.

[9] L. Wang and A. Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2523–2537, 2020.

[10] Y. Hioka, M. Kingan, G. Schmid, and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2016, pp. 1–5.

[11] B. Yen, Y. Li, and Y. Hioka, "Rotor noise-aware noise covariance matrix estimation for unmanned aerial vehicle audition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[12] K. Nakadai, M. Kumon, H. G. Okuno, *et al.*, "Development of microphone-array-embedded UAV for search and rescue task," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 5985–5990.

[13] K. Yamada, M. Kumon, and T. Furukawa, "Belief-driven control policy of a drone with microphones for multiple sound source search," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov. 2019, pp. 5326–5332.

[14] T. Spadini, G. S. Imai Aldeia, G. Barreto, *et al.*, "On the application of SEGAN for the attenuation of the ego-noise in the speech sound source localization problem," in *2019 Workshop on Communication Networks and Power Systems (WCNPS)*, Oct. 2019, pp. 1–4.

[15] B. Yen and Y. Hioka, "Noise power spectral density scaled snr response estimation with restricted range search for sound source localisation using unmanned aerial vehicles," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–26, 2020.

[16] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications* (Digital Signal Processing). Springer, 2001. (visited on 06/26/2017).

[17] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2012, pp. 3288–3293.

[18] K. Furukawa, K. Okutani, K. Nagira, *et al.*, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov. 2013, pp. 3943–3948.

[19] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 508–519, 2022.

[20] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Acoustic source localization from multirotor UAVs," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 10, pp. 8618–8628, 2019.

[21] L. Wang and A. Cavallaro, "Deep-learning-assisted sound source localization from a flying drone," *IEEE Sensors Journal*, vol. 22, no. 21, pp. 20 828–20 838, 2022.

[22] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Sound source tracking by drones with microphone arrays," in *2020 IEEE/SICE International Symposium on System Integration (SII)*, IEEE, 2020, pp. 796–801.

[23] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Assessment of sound source tracking using multiple drones equipped with multiple microphone arrays," *International journal of environmental research and public health*, vol. 18, no. 17, p. 9039, 2021.

[24] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Placement planning for sound source tracking in active drone audition," *Drones*, vol. 7, no. 7, 2023.

[25] T. Ministry of Land Infrastructure and Tourism. "Flight rules for unmanned aircraft (drones and model aircraft, etc.)" (2023), [Online]. Available: https://www.mlit.go.jp/en/koku/uas.html (visited on 07/18/2024).

[26] W. Giernacki, M. Skwierczyński, W. Witwicki, P. Wroński, and P. Kozierski, "Crazyflie 2.0 quadrotor as a platform for research and education in robotics and control engineering," in *2017 22nd International Conference on Methods and Models in Automation and Robotics (MMAR)*, 2017, pp. 37–42.

[27] F. Dümbgen, A. Hoffet, M. Kolundžija, A. Scholefield, and M. Vetterli, "Blind as a bat: Audible echolocation on small robots," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1271–1278, 2023.

[28] B. Yen, T. Yamada, K. Itoyama, and K. Nakadai, "A performance assessment on rotor noise-informed active multidrone sound source tracking methods," *Drones*, vol. 8, no. 6, p. 266, 2024.

[29] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, eabm6074, 2022.

[30] A. Taffanel, B. Rousselot, J. Danielsson, *et al.*, "Lighthouse positioning system: Dataset, accuracy, and precision for UAV research," *arXiv preprint arXiv:2104.11523*, 2021.

Implementation of a Robot Operation System-based network for sound source localization using multiple drones

Takumi Yamamoto*, Kotaro Hoshiba*, Benjamin Yen*†, Kazuhiro Nakadai*
* Institute of Science Tokyo, Japan
E-mail: yamamoto.t.by@m.titech.ac.jp Tel/Fax: +81-03-5734-3094

***Abstract*—This paper presents a communication framework for drone audition systems using multiple drones. A robust communication system is required to ensure a smooth operation of the drone audition system used for search and rescue operations in disaster-stricken areas. Such a communication system must be capable of recovering swiftly from communication disconnections and ensuring low-latency communication even when using multiple drones. To address these issues, we propose a distributed network using the Robot Operation System and evaluate its effectiveness.**

I. Introduction

Prompt search and rescue of victims in disaster-stricken areas are important tasks [1], and to date, various disaster robots have been developed for search and rescue operations [2]. In particular, search operations using drones, which can move quickly regardless of ground conditions, have attracted increasing attention and are actively studied. Many of these techniques are vision-based [3], [4]. However, the performance of vision-based methods is limited under poor lighting conditions or when victims are buried in rubble, occluding their visibility. Unfortunately, such conditions are common in disaster scenarios.

To solve this problem, drone audition technology, which detect victims by human-related sounds such as voices and emergency whistles using drone-embedded microphone array, is being researched and developed [5]–[13]. Such systems require robust communication systems between the drone and the ground station. In particular, communication robustness is of utmost importance when multiple drones are used.

This paper addresses the construction of a communication system for drone audition using multiple drones, with a focus on seamless communication with the onboard system of each drone (including control and sensor systems), as well as robustness against external adversarial influences.

II. Requirements for communication system in drone audition system

A major field of study in drone audition technology involves localizing sound sources using a single drone [7]–[11]. However, sound source localization using single drone (or single microphone array) can only provide the direction of the sound source. To obtain the three-dimensional position of a sound source with a single drone, it is necessary to perform triangulation for multiple localization results from different drone locations, or find the intersection between the estimated sound source direction and the ground surface assuming that the source exists on the ground surface. These methods have the following problems: the former method requires a long time to estimate the position of the sound source, and the latter method is not possible if the sound source is not located on the ground surface, such as on the upper floors of a building. To solve these problems, the method of localizing sound sources in three dimensions using multiple drones has been studied [14].

Most existing studies in multi-drone sound source localization assume all drones communicate with a central ground station (e.g. laptop PC). Naturally, such a system needs a communication network to send and receive data between each drone and the ground station. Three-dimensional sound source localization using multiple drones requires low-latency, time-synchronized integration of various information, such as the three-dimensional position and orientation of each drone and the spatial spectrum obtained by processing acoustic signals recorded by each drone-embedded microphone array. The obtained information should be in a format that can be adapted to various platforms, as it also needs to be integrated with the drone's flight systems for status monitoring and flight control purposes.

The use of ROS (Robot Operating System) as a communication network is one of the solutions to satisfy those requirements. The conventionally used ROS1 runs a single ROS Master as a host machine on the network, and ROS node (a process that performs calculations and publishes/subscribes data on ROS) running on each client machine publishes and subscribes data using ROS topics (a bus for a node to publish/subscribe data) through a ROS Master to communicate and channel information between client machines. However, once the communication between the ROS Master and the ROS node is disconnected, the ROS node terminates its operation and is difficult to recover. In the event of a disaster, the increase in communication network traffic due to emergency communications and physical damage to base stations may cause the

†B. Yen is an International Research Fellow of the Japan Society for the Promotion of Science at Institute of Science Tokyo.

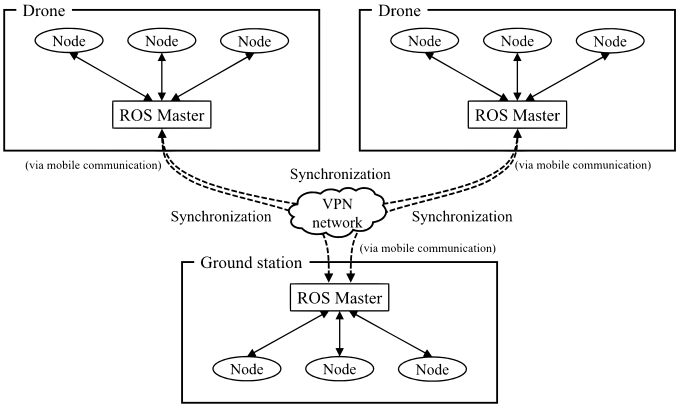


Fig. 1. Schematic diagram of the constructed network.

communication bandwidth to decrease, resulting in unstable communication. Therefore, the communication system needs to be self-recoverable even if communication is disconnected.

ROS2, a newer version of ROS, uses a communication protocol called DDS (Data Distribution Service), which enables communication without a ROS Master, so that the ROS nodes will not be terminated if the connection is disconnected [15]. However, the drone used in this research only supports ROS1. Therefore, it is necessary to construct a system that can recover even if communication is intermittently disconnected using ROS1 is needed. In addition, network traffic reduction is needed to integrate information across multiple drones with low latency.

In this paper, we propose a communication system to achieve sound source localization using multiple drones with low latency and robustness against communication disconnections. To ensure robustness against environments susceptible to unstable communication, a distributed network of independent ROS sub-networks on each drone and the ground station is constructed, and each ROS Master synchronizes all nodes sequentially. The constructed network is expected to recover from intermittent disconnections, including the ability to autonomously reconnect again after disconnection.

III. Construction of ROS network

This section presents the constructed ROS network. The constructed ROS network is a distributed ROS1 network with a design focus on robustness against unstable communication environments. ROS1 is adopted because the network needs to be connected to an existing ROS1 system for drone control as mentioned in Section II.

A schematic diagram of the constructed network is shown in Fig. 1. The network has independent ROS1 networks running ROS Master on each drone and ground station. These are virtually placed on the same local network using VPN (Virtual Private Network), and the topics are shared across all ROS networks by synchronizing each ROS Master via mobile communications. This framework allows each ROS network to continue operating independently even if the

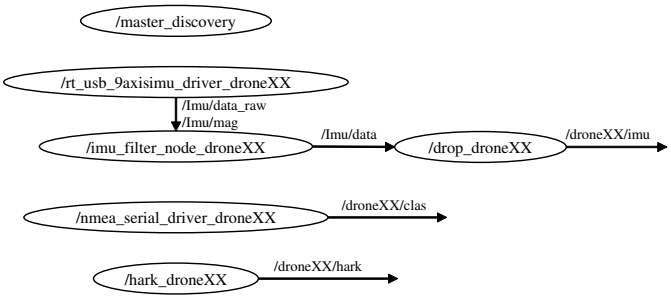


Fig. 2. ROS network for drones.

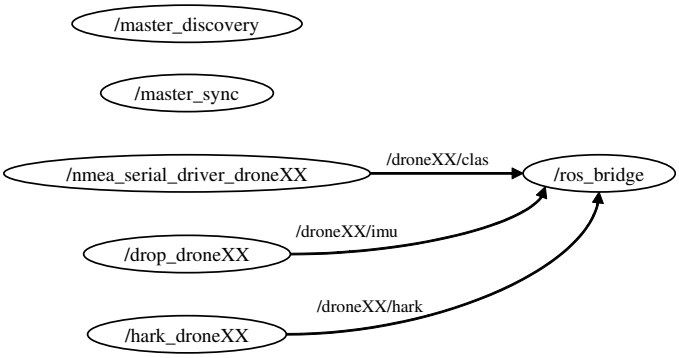


Fig. 3. ROS network for the ground station.

mobile communication is disconnected such that ROS topic synchronization can be restored when the communication is reconnected. To support various platforms, topics published in the ROS1 network are also published to ROS2 through ros1_bridge, described below. In addition, to reduce network traffic, the communication overhead is reduced by combining spatial spectra calculated by the sound source localization method and the directional information of sound sources into a single ROS message; a ROS-specific data structure used to transmit information (via ROS topics) between different ROS nodes, including client machines.

Fig. 2 and 3 show the ROS networks for the drones and ground station. The ellipses represent nodes, and the arrows represent topics. Note that each node independently publishes or subscribes to messages at an uniformed refresh rate for each. In other words, there is no interaction between nodes in publication or subscription. Details of the nodes and topics are presented as follows. The ID of the drone is inserted to "XX".

/master_discovery
: Node included in fkie_multimaster package [16]. This node searches ROS Masters running in the local network and obtains their status and information on connected nodes and services.

/master_sync
: Node included in fkie_multimaster package [16]. Using the information obtained by /master_discovery, this node synchronizes and shares the nodes and

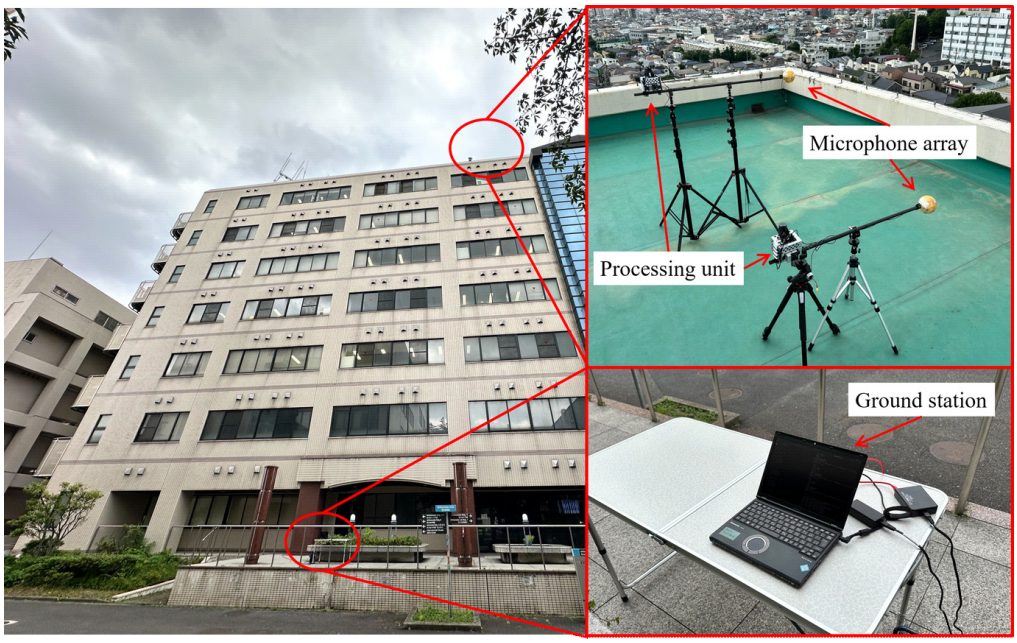


Fig. 4. Experimental setting.

services connected to each ROS Master.

/rt_usb_9axisimu_driver_droneXX

Node included in rt_usb_9axisimu_driver package [17]. This node obtains raw data of linear acceleration, angular velocity (/Imu/data_raw), and raw data of magnetic field (/Imu_mag) at the sampling rate of 100 Hz from the IMU (Inertial Measurement Unit).

/imu_filter_node_droneXX

Node included in imu_tools package [18]. This node calculates the orientation from the raw data of linear acceleration, angular velocity (/Imu/data_raw), and raw data of magnetic field (/Imu_mag) with low calculation cost using Madgwick's algorithm [19], and integrates it with the raw data (/Imu/data).

/drop_droneXX

Node included in the standard package, topic_tools. To reduce the traffic, this node drops the publishing cycle from 100 Hz to 2 Hz and republishes as /droneXX/imu.

/nmea_serial_driver_droneXX

Node included in nmea_navsat_driver package [20]. This node obtains information (/droneXX/clas) including latitude, longitude, and altitude in HAE (Height Above Ellipsoid) from the GNSS (Global Navigation Satellite System) receiver with CLAS (Centimeter Level Augmentation Service) at a sampling rate of 2 Hz.

/hark_droneXX

Acoustic signals recorded by the drone-embedded microphone array are processed using open-sourced robot audition software HARK [21]. This node publishes the processed results of sound source localization (/droneXX/hark) including spatial spectra calculated by MUSIC (MUltiple SIgnal Classification), sound source directions, sound power, and source ID at a sampling rate of 2 Hz.

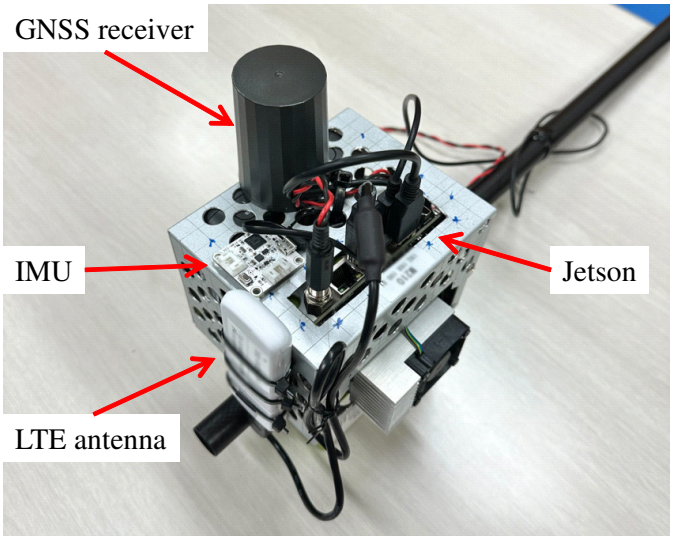


Fig. 5. Processing unit.

/ros_bridge

Node included in ros1_bridge package [22]. This node bridges topics and services from ROS1 to ROS2 and from ROS2 to ROS1.

TABLE I
AVERAGE BANDWIDTH FOR EACH TOPIC

Topic	Bandwidth
/drone01/imu	630 B/s
/drone01/clas	232 B/s
/drone01/hark	10,880 B/s

IV. EXPERIMENT

To evaluate the effectiveness of the constructed ROS network, communication experiments were performed.

A. Experimental procedure

To reproduce the communication status mimicking drone flight, the systems embedded on the drone were placed away from the ground station, and communication between each system and the ground station was performed. Fig. 4 shows the experimental setting. In this experiment, two drone-embedded systems consisting of a microphone array and a processing unit were placed on the roof of the building, and the ground station was placed on the ground floor. As shown in Fig. 5, the processing unit consisted of a Jetson Orin NX, IMU sensor, GNSS receiver, and an LTE (Long Term Evolution) antenna. The processing unit and the ground station were connected to the VPN via LTE and WiFi of the building, respectively. Each processing unit published the topics, /droneXX/imu, /droneXX/clas, /droneXX/hark (XX=01 or 02), and the ground station subscribed to them.

First, the latency from the processing units publishing messages to the ground station subscribing them, and the average bandwidth used for each topic were measured. Second, to evaluate the robustness against communication disconnections, the status of the ROS network when the communication is disconnected and reconnected was verified. Each message was continuously published via their relevant ROS topic from each processing unit on the rooftop for approximately 7 minutes, and the ground station was subscribed to each of them throughout the experiment. During publication and subscription, each processing unit's LTE network was intentionally disconnected 12 times. At the same time, the subscription status at the ground station was monitored to observe its response to the disconnections.

B. Results

First, we present the results of the measured ROS topic communication latency. Table I shows the average bandwidth used for each topic, measured over several minutes of communication. For /drone01/hark, since it contains a spatial spectrum of 1368 data points, its bandwidth is approximately 20 to 50 times larger than other topics. The total bandwidth of all topics was approximately 11.7 kB/s for one drone-embedded system and 23.4 kB/s for two drone-embedded systems. Even considering the reduction of communication speed due to VPN, it was found to be sufficient for low-latency communication.

The results of the measured latency are shown in Fig. 6. The latency of most messages was approximately 0.1 s, and

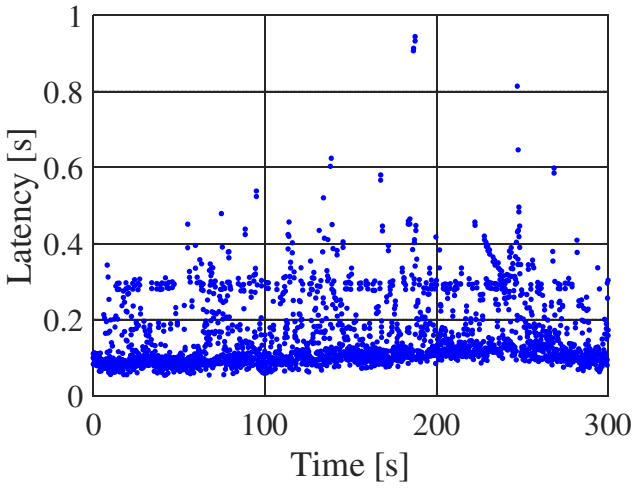


Fig. 6. Latency from publication to subscription of messages.

the average latency was approximately 0.16 s. Since the system performs sound source localization every 0.5 s, the latency was found to be acceptable.

Next, we present results of the monitored subscription statuses at the ground station when the communication is disconnected and reconnected. Fig. 7 shows the status of topic subscription. The times when /drone01/imu, /drone01/clas and /drone01/hark were subscribed are plotted in blue, red, and black, respectively. The dotted magenta lines represent times when the communication was disconnected. Even if the communication was disconnected, it was found that topic could be subscribed after the communication was reconnected. The periods for recovery were mostly within 20 s. Although it takes a dozen seconds to recover message subscriptions because the system needs time to establish communication, the constructed network is proven to be a recoverable ROS network.

Overall, results show that the ROS network constructed in this paper is practical and capable of communicating with low latency while ensuring robustness against unstable communication environments.

C. Demonstration

Using the constructed ROS network, sound source localization in an actual outdoor environment was demonstrated. In this demonstration, two drones with the microphone array and processing unit were used as shown in Fig. 8. The position and orientation of each drone and obtained spatial spectra published via the ROS network are subscribed at the ground station. Using this information, the three-dimensional position of the sound source was estimated. Result of the demonstration is shown in Fig. 9. Fig. 9(b) and (c) show the obtained spatial spectra plotted with the axes given in Fig. 9. The sound power arriving from each direction is plotted in a colormap. The components on the left side of spatial spectra indicate the drone's noise, and the components on the right side indicate the sound source. Fig. 9(d) shows the positions of the drones

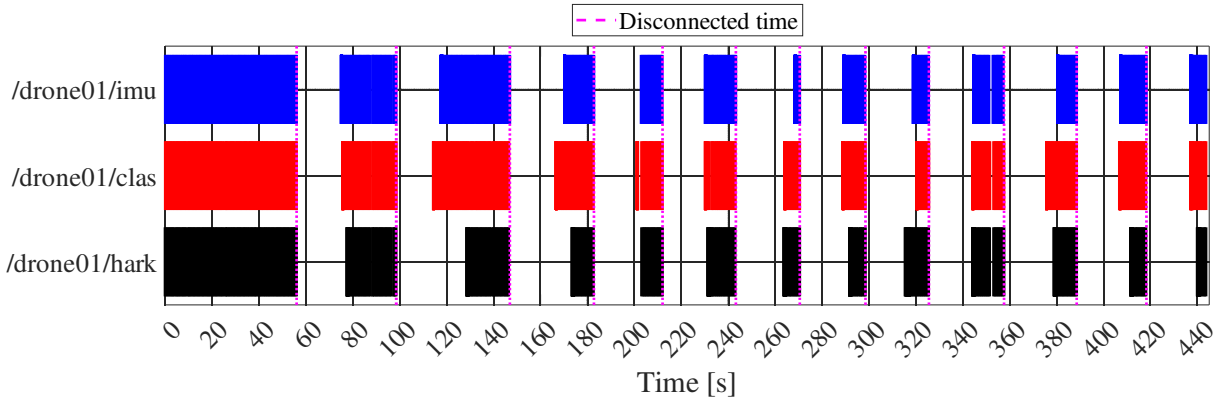


Fig. 7. Message subscription status at the ground station when the network is disconnected.

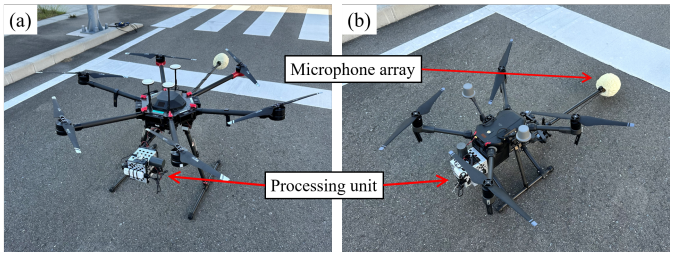


Fig. 8. Drones with the microphone array and processing unit used in the demonstration. (a) DJI Matrice 600 Pro, (b) DJI Matrice 210 RTK V2.

and the estimated sound source plotted on the point cloud of the field. As shown, by using the constructed ROS network, it was possible to obtain information from the drones located far from the ground station and estimate the position of the sound source in real time.

V. Conclusion

To construct a network for transmitting and receiving data between drones and ground stations for sound source localization using multiple drones, a distributed network using ROS that can integrate and synchronize information compatible with various platforms in a low latency is proposed in this paper. To enable operation even in unstable communication environments, each drone and ground station has an independent ROS network, and the ROS masters running on each network are sequentially synchronized so that the system can recover even if the communication is reconnected after disconnection. Results from evaluation experiments show that the constructed ROS network is a practical system that communicates with low latency while being highly robust against unstable communication environments.

Acknowledgment

This work was supported by JSPS KAKENHI Grant No. JP22K14218 and F-REI No. JPFR23010102.

References

[1] V. Viswan and M. L. Madhav, “Mission-critical management using media independent handover,” *International Journal of Computer Applications in Engineering Sciences*, vol. III, no. I, pp. 32–36, 2013.

[2] S. Tadokoro, “Disaster robotics - results from the impact tough robotics challenge,” *Springer International Publishing*, 2019.

[3] A. Quan, C. Herrmann, and H. Soliman, “Project vulture: A prototype for using drones in search and rescue operations,” *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pp. 619–624, 2019.

[4] S. Sambolek and M. Ivasic-Kos, “Automatic person detection in search and rescue operations using deep cnn detectors,” *IEEE Access*, vol. 9, no. I, pp. 37 905–37 922, 2021.

[5] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, “Dregon: Dataset and methods for uav-embedded sound source localization,” *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–8, 2018.

[6] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, “Drone audition: Sound source localization using on-board microphones,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 508–519, 2022.

[7] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, “Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter,” *Proc. of IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, pp. 3288–3293, 2012.

[8] T. Ohata, K. Nakamura, T. Mizumoto, T. Tezuka, and K. Nakadai, “Improvement in outdoor sound source detection using a quadrotor-embedded microphone array,” *Proc. of IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, pp. 1902–1907, 2014.

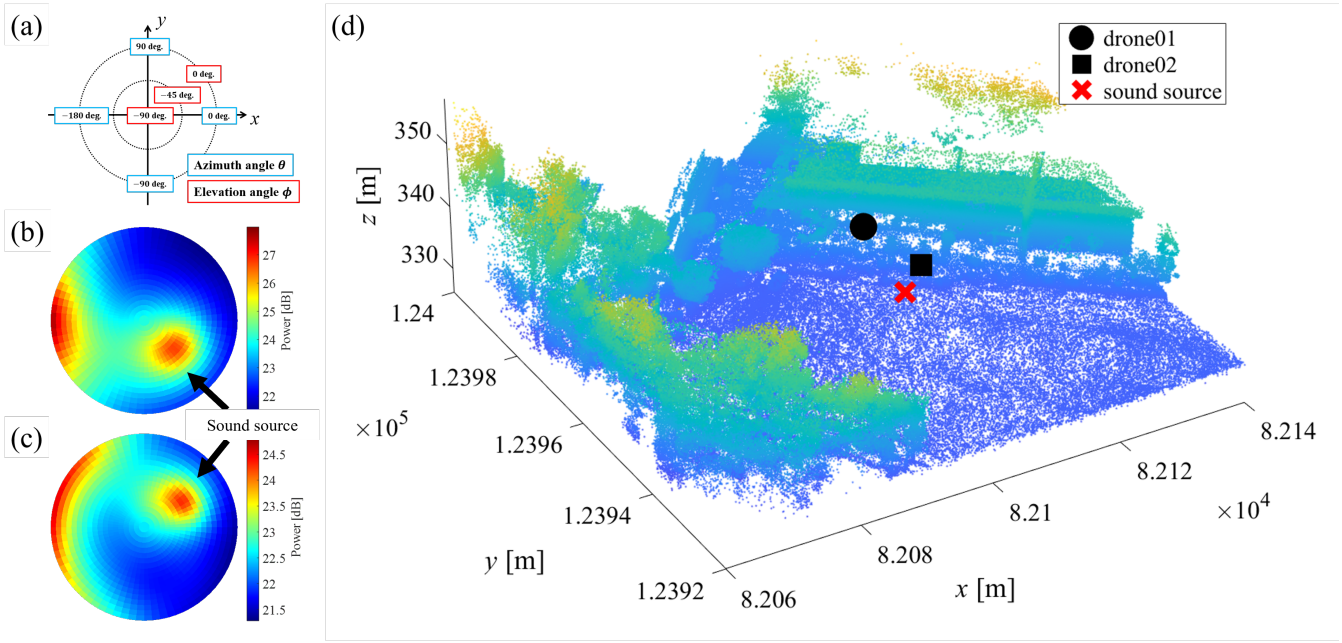


Fig. 9. Results of the demonstration. (a) Setting of the coordinate system of the azimuth and elevation angle for plotting the spatial spectrum, (b) spatial spectrum obtained by drone01, (c) spatial spectrum obtained by drone02, (d) visualization result of the position of the drones and estimated sound source.

[9] K. Hoshiba, K. Washizaki, M. Wakabayashi, *et al.*, "Design of uav-embedded microphone array system for sound source localization in outdoor environments," *International Journal of Computer Applications in Engineering Sciences*, vol. 17, no. 11, pp. 1–16, 2017.

[10] K. Hoshiba, K. Nakadai, M. Kumon, and H. G. Okuno, "Assessment of music-based noise-robust sound source localization with active frequency range filtering," *J. of Robotics and Mechatronics*, vol. 30, no. 3, pp. 426–435, 2018.

[11] K. Hoshiba, I. Komatsuzaki, and N. Iwatsuki, "Proposal of practical sound source localization method using histogram and frequency information of spatial spectrum for drone audition," *Drones*, vol. 8, no. 4, 159, 2024.

[12] T. Yamada *et al.*, "Placement planning for sound source tracking in active drone audition," *Drones*, vol. 7, no. 7, 2023.

[13] B. Yen *et al.*, "Performance evaluation of sound source localisation and tracking methods using multiple drones," *Proc. InterNoise 2023*, pp. 1926–1937, 2023.

[14] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Assessment of sound source tracking using multiple drones equipped with multiple microphone arrays," *International Journal of Environmental Research and Public Health*, vol. 18, no. 17, 9039, 2021.

[15] Y. Maruyama, S. Kato, and T. Azumi, "Exploring the performance of ros2," *In Proceedings of the 13th International Conference on Embedded Software*, pp. 1–10, 2016.

[16] GitHub, *Fkie/multimaster_fkie*. [Online]. Available: https://github.com/fkie/multimaster_fkie.

[17] GitHub, *Rt-net/rt_usb_9axisimu_driver*. [Online]. Available: https://github.com/rt-net/rt_usb_9axisimu_driver.

[18] GitHub, *Ccnyroboticslab/imu_tools*. [Online]. Available: https://github.com/CCNYRoboticsLab/imu_tools.

[19] S. O. H. Madgwick, A. J. L. Harrison, and R. Vaidyanathan, "Estimation of imu and marg orientation using a gradient descent algorithm," *2011 IEEE International Conference on Rehabilitation Robotics*, pp. 1–7, 2011.

[20] GitHub, *Ros-drivers/nmea_navsat_driver*. [Online]. Available: https://github.com/ros-drivers/nmea_navsat_driver.

[21] *Hark*. [Online]. Available: https://hark.jp/.

[22] GitHub, *Ros2/ros1_bridge*. [Online]. Available: https://github.com/ros2/ros1_bridge.

Drone audition: dataset and methods for ground surface material classification using drone noise in outdoor environment

Tsubasa Yano*, Benjamin Yen*† and Kazuhiro Nakadai*
* Tokyo Institute of Technology
E-mail: {yano, benjamin, nakadai}@ra.sc.e.titech.ac.jp

***Abstract*—This paper considers a technique to estimate the material of ground surfaces directly below the drone based on the rotor noise generated by the drone. It is anticipated that this technique will enable us to assess the damage in areas that are not easily accessible by humans. In this study, drone noise was recorded while flying over multiple materials, such as asphalt, soil, and water, in a real environment with external noise, such as wind. Following, a machine learning-based model was trained using this dataset to classify the various materials using the recorded audio. Results indicate that while the classification was successful, there is a need for a classification method that is more robust to external noise.**

I. Introduction

In recent years, natural disasters such as earthquakes and landslides caused by torrential rains have become more frequent and often cause significant damage. In Japan, Typhoon No. 19 in 2019 caused more than 950 landslides in 20 prefectures, mainly in eastern Japan [1], and torrential rain that hit Kumamoto Prefecture in July 2020 caused several national roads to be closed due to overflowing rivers and landslides. Furthermore, the 2024 Noto Peninsula earthquake that occurred in January of this year caused more than 90 sections of prefectural roads to be closed to traffic [2]. Natural disasters caused by extreme weather and other factors are on the rise, not only in Japan but around the world [3], and countermeasures and responses are required.

One solution to these problems is the use of highly manoeuvrable drones. Drones can be used to quickly and inexpensively survey areas that are difficult to access directly by humans.

In this paper, we make use of the fact that the acoustic signal recorded by the microphone mounted on the drone includes not only noise directly from the drone's rotors, but also rotor noise that is reflected once on the ground surface. As such, we can address the problem of ground surface material estimation using this reflected sound with deep learning. Since there is no dataset for estimating the ground surface material outdoors, the creation of such a dataset will be conducted as well. In addition, since there is concern about the effects of external noise caused by strong winds in an outdoor environment, we also propose a custom-made windshield for the microphone array used in this study and investigate its effect against various forms of windshield materials.

II. Related work

The majority of studies conducted to assess the disaster situation by installing sensors on drones utilize cameras and microphones as sensors [4]. Although cameras are effective sensors and have been employed in disaster-stricken areas, their use is limited under conditions of poor visibility, such as at night. Microphones, however, are immune to such limitations and are currently being investigated as a potential solution, often referred to as "drone audition" [5]–[10].

Most drone audition approaches involve suppressing drone rotor noise in order to detect voices coming from victims. In other words, research has been conducted with the objective of suppressing drone rotor noise. On the other hand, this paper takes a different approach to drone audition research in that it actively utilizes drone rotor noise not as an object to be suppressed but as useful information for understanding the material properties of the ground surface. If ground surface material estimation through reflected drone rotor noise is feasible, as illustrated in Fig. 1, it opens the possibility to detect landslides or road collapses by identifying changes in ground surface material or features. This is expected to be beneficial in comprehending the extent of the damage. To the best of our knowledge, few such studies have been conducted, with the exception of ours. Additionally, while our previous study [11] estimates the ground surface material using noise recorded while changing the ground surface material, it has yet to show that it is effective in a real environment. This is due to the data being recorded in an anechoic chamber, and thus noise other than drone rotor noise (e.g. wind noise) was not considered. Furthermore, the amount of data measured was extremely small (about one minute for each condition).

A potential application that utilizes research on the estimation of material properties from acoustic signals includes the sound impact inspection of bridge structures. Generally, in such inspections, skilled workers estimate the characteristics and degree of deterioration of bridge structures from the sound generated when the bridge is struck. However, recent years have seen a growing number of studies conducted to learn the skills of skilled inspectors using machine learning methods

†B. Yen is an International Research Fellow of the Japan Society for the Promotion of Science (JSPS) at the Nadakai Lab, Tokyo Institute of Technology.



Fig. 1: Necessity of ground surface material estimation.

to obtain equivalent performance [12]. Some of these studies have reported the use of drones equipped with a percussion inspection function [13]. However, these studies were conducted on concrete, and no research has been conducted from the viewpoint of discriminating various materials from the reflected sound. Furthermore, drone rotor noise is treated as unnecessary background noise even when a drone with a percussion inspection function is used, and there is a lack of perspective on actively using drone noise, as in this paper.

III. Dataset construction

A. Guidelines for dataset design

The factors that affect the recorded drone rotor noise include not only the material of the ground surface but also the drone's altitude, the rotational speed of each motor, and the geometry of the ground surface. In this paper, as a preliminary step in the application to outdoor environments, we decided to create a dataset based on the following guidelines.

1) All parameters are fixed except for the ground surface material and the drone's altitude.
2) The ground surface material is selected based on the actual disaster situation.
3) For the drone altitude, two types of altitude are set: one is the same height as that often used in actual disaster rescue scenes, and the other is lower than that but still high enough to reach above the human head. (This is because it is anticipated that the attenuation level of reflected drone noise, when flown too high, will make material estimation difficult.)
4) To secure the amount of data, the recording time for each material is as long as possible within the limits of the drone's battery.

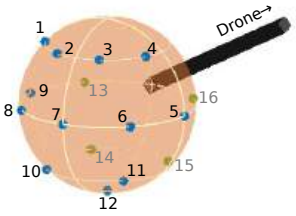
B. Dataset creation method

As shown in Fig. 2, a 16-channel spherical microphone array is installed a distance forward of the drone, and a Raspberry Pi for recording and storage is installed at the rear. The installation position is adjusted with consideration of the center of gravity of the drone. In this state, when the drone is in operation, the radiated noise from the rotors is recorded by all channels of the microphone array for various types of ground surface materials.

The materials selected for the ground surface were soil, asphalt, and water. Two types of water surface were selected: 0.7 m deep (hereafter, "shallow water") and 5 m deep (hereafter, "deep water"), for a total of four materials. All surfaces were nearly horizontal, without any slopes or large



(a) Drone with equipments.



(b) Mic. array channel arrangement.

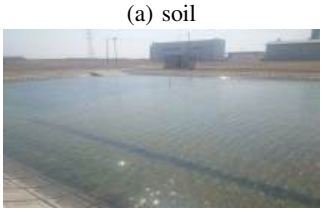
Fig. 2: Drone and microphone array for recording.



(a) soil



(b) asphalt



(c) shallow water (depth: 0.7 m)



(d) deep water (depth: 5 m)

Fig. 3: Ground surface materials.

irregularities. The surface of each material is shown in the Fig. 3. Two drone altitudes, 5 m and 10 m, were selected in accordance with the previously established guidelines.

Audio recording was first conducted at an altitude of 5 m above each material for 8 minutes while hovering. Upon completion, the drone ascended to an altitude of 10 m and recorded for another 8 minutes while hovering. For safety reasons, if the drone's low battery alarm sounded during the recording, the recording process was terminated and the drone was immediately returned for landing. In fact, during data collection for audio at an altitude of 10 m in shallow water, the recording was stopped after 6 minutes and 20 seconds due to the drone's low battery alarm sounding midway through the recording. This was probably because the drone's battery consumption was higher than expected due to the strong winds. Eventually, we were able to record approximately 1 h of data. The other conditions are shown in TABLE I.

C. Analysis of recorded data

To ascertain whether the recorded drone noise exhibits any anomalies, we conducted a comprehensive examination of

TABLE I: Dataset preparation specifications.

Aircraft in use	DJI Matrice 210
Number of array channels	16 (Sphereical)
Sampling rate	16 kHz

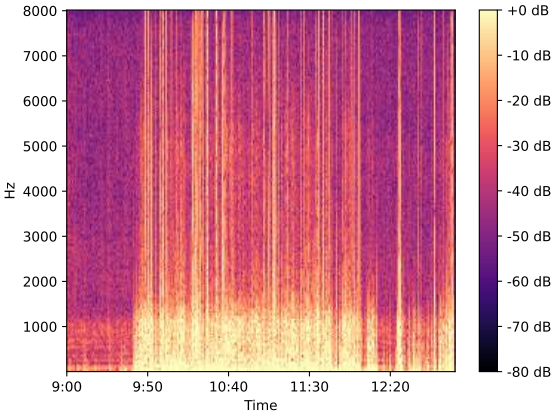


Fig. 4: Occurrences of sound clipping.

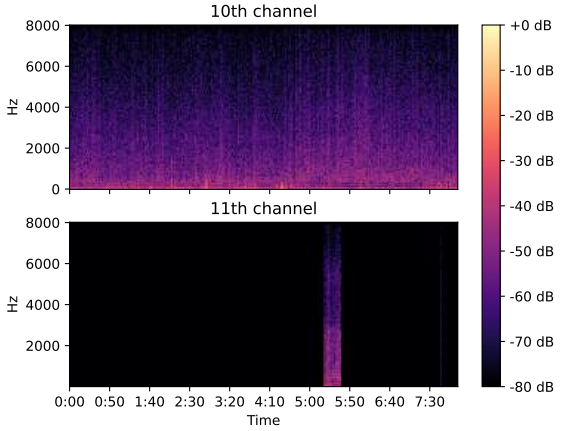


Fig. 5: Interval in which data value is zero.

the spectrogram and waveforms, complemented by a detailed auditory analysis. This process led to the identification of two defects.

- Sound clipping is frequently observed in the recorded data.
- There are intervals in which the recorded data value is zero or very close to zero.

Fig. 4 shows a spectrogram of the recording of the 14th channel (located at the rear side, see Fig. 2(b)) recorded over shallow water, with brighter colors indicating stronger intensity. It can be seen that impulsive signals are frequently recorded in the low and high frequency bands. The sound clipping caused by such impulsive noise occurs at a common time in many channels. This indicates that the cause of the clipping was strong winds during the recording. When such clipping occurs, the reflected sound signal cannot be picked up properly, which may result in degraded estimation performance. In contrast, a spectrogram from recordings of the 10th and 11th channels recorded above the soil in Fig. 5 illustrates a phenomenon in which the value of the recorded data becomes zero. As shown, especially in the 11th channel, nearly 90% of the recordings have a value of zero. The length of time and frequency of this phenomenon differ from channel to channel and recording to recording. After consulting with the microphone manufacturer, it was found that strong winds can cause the diaphragm of the MEMS microphones used to be stuck to the inner wall of the microphone, resulting in extremely low sensitivity.

As mentioned in Sec. III-B, it was windy at the time of recording, with an average wind speed of 4˜7m/s and a maximum of 17 m/s observed with the anemometer on hand. As a consequence of the aforementioned issues, a significant proportion of the dataset is unusable for ground surface material estimation. Consequently, only the usable parts of the recorded noise was extracted for the formation of the dataset. The specific method is described in Sec. V-A.



(a) Windshield components. (b) Attached to microphone array.

Fig. 6: Proposed windshield.

IV. Windshield installation on microphone array

As a consequense of the findings from Sec. III-C, in order to robustly perform ground surface material estimation in an outdoor environment, it is critical to address the presence of external noise, including strong winds. One effective approach to achieve this is by attaching a windshield to the microphone array. Materials with porous structures are renowned for their exceptional sound absorption capabilities, with polyurethane being a particularly prevalent choice in civil construction and transportation applications [14]. We propose the use of a windshield made of porous polyurethane foam, as illustrated in Fig. 6. This windshield is capable of enclosing the entire microphone array through the combination of two hemispherical parts. The material utilized for this windshield contains numerous air bubbles within the windshield due to its inherent ability to expand 100 times in volume when combined with water. Consequently, it is remarkably lightweight at 21 grams in total, making it an ideal choice for use in drones with tight payload restrictions, and it is also anticipated to exhibit high sound absorption properties.

In this paper, we verify the performance of the windshield through experiments to determine whether ground surface material estimation is possible using drone noise recored with the proposed windshield attached to the microphone array. Details regarding the methodology and results of the experiment can be found in Sec. V-B.

V. Evaluation

A. Ground surface material classification

The feasibility of the created dataset was evaluated by proposing a machine learning-based drone noise to ground surface material classifier.

1) Data to be used: It is necessary to eliminate from the recorded data any intervals that are unsuitable for the purposes of training and estimation. Specifically, the following steps were taken:

- The material was limited to three types: soil, asphalt, and shallow water.
- Only data collected at an altitude of 5 meters were used.
- For the shallow water recordings, the sections with severe clipping (the initial one minute and a half and the last 10 seconds) were removed and 6 minutes 17 seconds of the eight minutes were extracted.
- Data with intervals which data value is zero among 16 channels were removed (found in channels 1, 4, 9, 10, 12, 13, 14, and 15).

Hereafter, shallow water is simply referred to as "water".

2) Networks and experimental conditions: In conducting deep learning-based classification, two networks were employed as classifiers: a four-layer convolutional neural network (CNN) depicted in Fig. 7 and ResNet18 [15] utilized in previous study [11]. To create the input data for the networks, each recording was divided into three different sections (training, validation, and test), and converted into the time-frequency domain using short-time Fourier transform (STFT) under the conditions outlined in TABLE II. Subsequently, the spectrograms were extracted with a data length of 256 frames and a shift length of 128 frames, and the real and imaginary components of each extracted spectrogram were utilized as input. In this experiment, eight channels were utilized, resulting in the acquisition of 1,094 input features with a length of $16\times256\times256$ for training, and 106 for both the validation and testing. The labels of the training data were three-dimensional one-hot vectors corresponding to the ground surface materials.

TABLE III illustrates the length of each section in each recording, and TABLE IV illustrates the training conditions for the CNN and ResNet18. The epoch with the optimal performance was selected for evaluation within the range where overlearning does not occur based on the learning curve. The evaluation was conducted using the following accuracy measure.

$$\text{Accuracy}(\%) = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}} \times 100$$

Additionally, the confusion matrices were computed and analyzed for trends by material.

3) Results: TABLE V illustrates the appropriate response rate for each data section. The accuracy of the test data was 70.8% for the CNN using the model at the conclusion of 24 epochs and 82.1% for ResNet18 using the model at the conclusion of 38 epochs. The confusion matrix heatmaps in these cases are shown in Fig. 8.

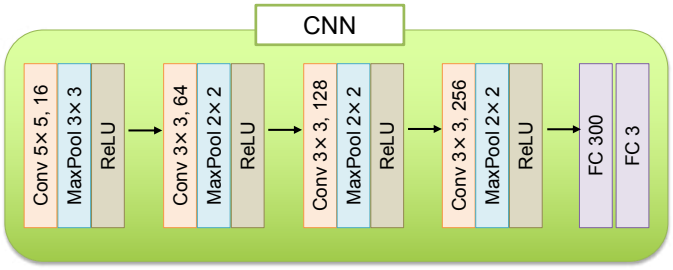


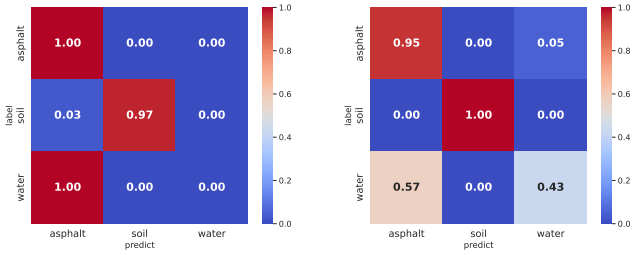
Fig. 7: CNN architecture used in this experiment.

TABLE II: STFT parameters

STFT frame length	512
STFT hop length	128
STFT window function	Hanning window

4) Discussions: As shown in TABLE V, the accuracy for the CNN on the test data is as high as 70%, while the accuracy on the validation data is only 38%. In light of the fact that the task in this paper is 3-class classification task and the accuracy is only 33%, it would be premature to accept these results at face value. An analysis of the cause is therefore necessary. The next step is to shuffle the data and perform cross-validation in order to ascertain the source of the bias. Conversely, ResNet18 yielded an accuracy exceeding 80% for all data sections, suggesting that the learning process for ground surface material estimation was effective. It is postulated that the superior learning capabilities of ResNet18 relative to CNN are attributable to its greater number of layers, which facilitate the suppression of noise other than reflected sound.

The heatmaps of the confusion matrices indicate that while both networks demonstrate high reproducibility for asphalt and soil, the reproducibility for water is relatively low. In particular, the CNN did not have any data point that could be estimated as water, indicating that it was not possible to distinguish between water and asphalt. This is likely due to the dynamic nature of the wind generated by the drone's propeller, which causes the water surface to be in a state of flux, and the irregularity of the



(a) CNN (b) ResNet18

Fig. 8: Confusion matrices for ground surface material classification.

TABLE III: Time length of each section (seconds)

Section	Training	Validation	Test
Asphalt	405	40	40
Soil	407	40	40
Water	313	32	32

TABLE IV: Training parameters

Loss function	Cross entropy
Optimizer	Adam
Learning rate	10^{-4}
Epochs	100

characteristics of the reflected sound. However, this remains a hypothesis at this point, and future analysis will be conducted in conjunction with the performance differences among the networks.

B. Windshield effects

1) Experimental conditions: A running car was used to record the sound in order to reproduce a high wind environment. Specifically, a jig with a microphone array attached was held out of the window from a moving car to record the target sounds (whistle sound and voice). The microphone array was kept at approximately 1 m from the car window while controlling it so that it did not rotate. To assess the efficacy of the proposed windshield, the following four types of microphone arrays were utilized.

1) No windshield was attached.
2) The windshield proposed in this paper was attached.
3) A 10 mm thick polyurethane (PU) foam material was affixed to cover the few channels located on the windward side.
4) A 10 mm thick polyethylene (PE) foam material was affixed to cover the few channels located on the windward side.

The recordings were made on a circuit of approximately 350 m per lap, with three to six laps run for each condition. To measure the wind speed, the anemometer was placed outside the window along with the jig. The wind speed during the run ranged from 12 to 20m/s, with no significant differences between each condition. The performance of the windshield is evaluated by comparing the spectrum and the rate of clipping occurrence per unit time for 13th channel, located on the windward side, of each recording.

2) Results and Discussions: Fig. 10 shows a spectrum of the recordings from each windshield configuration. It can be observed that the proposed windshield exhibits a sound insulation effect of approximately 10 to 20 dB across a wide range of frequencies, with the highest level of sound insulation overall compared to other foams. In particular, the sound insulation performance is markedly elevated in the high-frequency range above 3000 Hz. Given that the target sound in this experiment is situated between 100 and 3000 Hz and that wind noise is replete with high-frequency components, it is anticipated that the proposed windshield will be effective in preventing sound clipping due to strong winds.

TABLE V: Accuracy (in % of correct material estimation) for each method and dataset.

Network	Training	Validation	Test
CNN	68.1%	36.8%	70.8%
ResNet18	100%	70.8%	82.1%

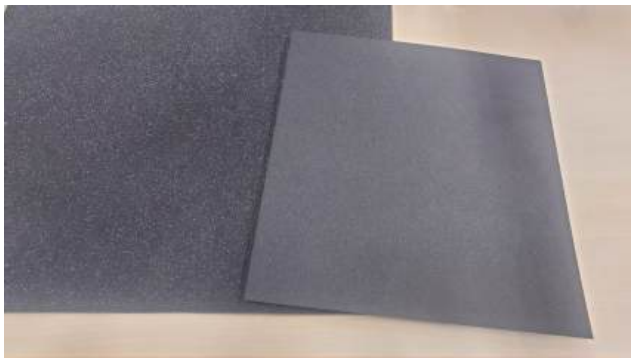


Fig. 9: PU foam (left in the figure) and PE foam (right in the figure)

TABLE VI depicts the number of observation points per second, wherein the absolute value of the recorded sound observation exceeded a specified threshold (th). While other foam materials did not prevent sound clipping, the proposed windshield does not have a single data point that suggests sound clipping, and the absolute value of the maximum observed value is approximately 0.67, indicating that the windshield can withstand even stronger winds. Therefore, together with the spectrum shown earlier, it can be shown that the proposed windshield is extremely effective against strong winds.

VI. Conclusion

In this paper, we presented a robust ground surface estimation method from drone noise reflections. To this end, we

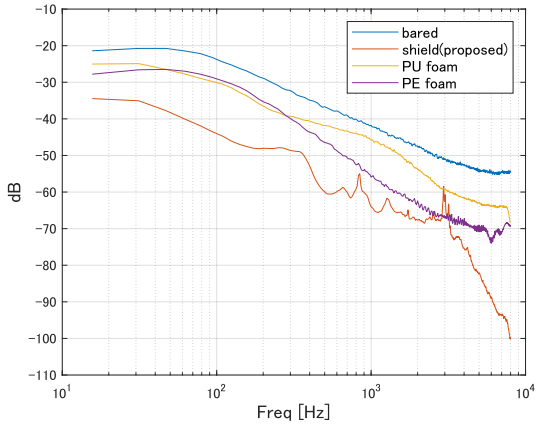


Fig. 10: Spectrum of recordings from each windshield configuration.

TABLE VI: the observation points where the absolute value $\geq th$ per second

threshold	bared	shield(proposed)	PU foam	PE foam
$th = 0.99$	16.59	0	0.358	0.682
$th = 0.9$	41.73	0	1.433	2.317

created a dataset of rotor noise and four types of ground surface materials (soil, asphalt, and two types of water at different depths) in an outdoor environment. Based on this dataset, we developed a convolutional neural network (CNN) and ResNet18-based ground surface material estimation method for the three types of materials except for deep water. The efficacy of the developed method was evaluated using the generated datasets, and a 70.8% correct response rate was observed on ResNet18. Moreover, we proposed a novel windshield for the microphone array as a means of mitigating external noise, and demonstrated its efficacy in preventing sound clipping caused by strong winds. The forthcoming stages of this project will involve the expansion of the dataset, data recording using the proposed windshield, an analysis of the ease of identification due to differences in networks, and real-time processing.

ACKNOWLEDGMENT

This work was supported by KAKENHI JP22F22769 and JP22KF0141, and F-REI (JPFR23010102).

REFERENCES

[1] Ministry of Land, Infrastructure, Transport and Tourism, *Overview of landslides caused by typhoon no. 19 of 2019 ver2.1 (in Japanese)*, https://www.mlit.go.jp/river/sabo/jirei/r1dosha/r1typhoon19_gaiyou191224r.pdf, Dec. 2019.

[2] I. prefecture, *Information on the 2024 noto peninsula earthquake, information on damage, etc. (12th report) (in Japanese)*, https://www.pref.ishikawa.lg.jp/saigai/documents/202401041500higaihou.pdf, Jan. 2024.

[3] W. M. Organization, *Wmo atlas of mortality and economic losses from weather, climate and water extremes (1970–2019)*, https://library.wmo.int/idurl/4/57564, 2021.

[4] K. Matsui, K. Hasegawa, J. Ohya, Y. Kato, and M. Yokozawa, "Study on estimation method of landslide location and scale by segmentation of rgbd images acquired by the camera attached to a drone," *Research Report: Computer Vision and Image Media (CVIM)*, no. 16, pp. 1–6, 2023.

[5] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3288–3293, 2012. DOI: 10.1109/IROS.2012.6385994.

[6] K. Hoshiba, K. Washizaki, M. Wakabayashi, *et al.*, "Design of uav-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, 2017, 2535.

[7] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Placement planning for sound source tracking in active drone audition," *Drones*, vol. 7, no. 7, 2023. DOI: 10.3390/drones7070405.

[8] M. Kumon, H. G. Okuno, and S. Tajima, "Alternating drive-and-glide flight navigation of a kiteplane for sound source position estimation," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2114–2120, 2021.

[9] B. Yen, T. Yamada, K. Itoyama, and K. Nakadai, "A performance assessment on rotor noise-informed active multidrone sound source tracking methods," *Drones*, vol. 8, no. 6, 2024.

[10] M. C. L. Wang and A. G. Rossberg, "Drone audition for bioacoustic monitoring," *Methods in Ecology and Evolution*, vol. 14, no. 12, pp. 3068–3082, 2023.

[11] T. Yano, K. Nishida, K. Itoyama, and K. Nakadai, "Ground surface material estimation by using drone rotor noise," *SICE SI*, 2023.

[12] H. Emoto, Y. Baba, H. Asano, and Y. Nagase, "Comparision of ai method on hammering sounds at concrete bridge," *Artificial Intelligence and Data Science*, vol. 1, no. J1, pp. 514–521, 2020. DOI: 10.11532/jsceiii.1.J1_514.

[13] F. Uehan, "Inspection techniques of concrete bridge members using uav," *Concrete Journal*, vol. 57, no. 9, pp. 699–704, 2019. DOI: 10.3151/coj.57.9_699.

[14] S. Dong, Y. Duan, X. Chen, *et al.*, "Recent advances in preparation and structure of polyurethane porous materials for sound absorbing application," *Macromolecular Rapid Communications*, p. 2 400 108, 2024.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

Swarm Active Audition System with Robots and Drones for a Search and Rescue Task

Kazuhiro Nakadai*, Makoto Kumon†, Yoko Sasaki‡, Kotaro Hoshiba*, and Benjamin Yen*
* Tokyo Institute of Technology
E-mail: {nakadai@ra.sc, hoshiba@rmsv.mech, benjamin@ra.sc}.e.titech.ac.jp
† Kumamoto University
E-mail: kumon@gpo.kumamoto-u.ac.jp
‡ The National Institute of Advanced Industrial Science and Technology (AIST)
E-mail: y-sasaki@aist.go.jp

***Abstract*—This paper addresses the design of system based on drone audition and its necessity to search for rescuers. We propose a swarm active audition system (SAAS) that employs robots and drones as an efficacious means of identifying individuals in need of rescue during the initial stages of search and rescue operations. The concept and system architecture of SAAS is presented, and the effectiveness of drone audition functionality is demonstrated on a case study basis. Furthermore, interviews will be conducted with relevant personnel involved in search and rescue activities at disaster sites to ascertain the necessity of SAAS in the field. The necessity and lack of SAAS functions will be discussed in light of the analysis of these interviews.**

I. Introduction

In the context of disaster relief, the search for individuals in need of rescue represents the primary and most crucial objective, given that a failure to rescue a person in need within 72 hours of a disaster's occurrence has been linked to a significantly reduced probability of survival. In disaster areas, the disruption to the traffic network is such that even emergency vehicles are unable to pass through with ease, thereby rendering drone searches an effective solution. While mounting a camera on a drone and conducting a search based on the camera information is an effective approach, its effectiveness is limited when individuals are obscured by debris or when visibility is constrained, such as in foggy weather.

Considering this situation, "drone audition" has started as an extension of robot audition [1], in which a drone equipped with a microphone array is used to search for rescuers in need. Okutani et al. [2] introduced a method of multiple signal classification based on generalized singular value decomposition (GSVD-MUSIC), a technique developed for robot audition [3]. In another early study, Basiri et al. proposed a 4 ch microphone array that can localize sound sources in both azimuth and elevation [4]. Ohata et al. proposed iGSVD-MUSIC, an augmented variant of GSVD that incorporates an incremental estimation of the dynamically-changing drone noise [5]. Additionally, Furukawa et al. proposed a further noise estimation method based on a Gaussian process, with the objective of improving sound source localization from the sky [6].

In 2014, the Japanese Cabinet Office launched the ImPACT (Impulsing Paradigm Change through Disruptive Technologies Program), which included the Tough Robotics Challenge (TRC), led by Professor Tadokoro of Tohoku University. The field of drone audition has made notable advancements in Japan with the backing of this initiative. The project encompassed extensive research into drone audition, particularly in the areas of sound source localization and tracking. This was based on the premise that the primary objective of drone audition is to locate individuals in disaster zones. Once the fundamental offline technology had been established, attention was directed towards the development of a technology that would be robust in the real environment and operational online. This involved the design of an accessible microphone array [7], the implementation of 3D positional sound source localization [8], and the handling of multiple sound sources [9]. At last, the online demonstration has been achieved [10], [11].

With regard to groups engaged in drone audition research on source localization and tracking outside of Japan or ImPACT TRC during this period, for instance, Wang et al. [12] reported the use of time-frequency masks for source localization, a technique frequently employed in acoustic signal processing. Lauzon et al. [13] reported a study in which ground-based microphones were used to localize drones, as opposed to drone-mounted microphones, for the purpose of sound source localization. However, there were few reports of drone audition studies, in part due to the absense of a representative drone audition dataset, such as the DREGON dataset [14].

Sound source separation and speech enhancement have been studied more widely. The majority of studies were based on signal processing [15]–[20], with a few employing deep learning techniques [21]. In ImPACT TRC, research has been conducted on the integration of sound source separation and classification as a subsequent step following sound source localization and tracking [22]–[24]. Many of the methods proposed in these studies employ a combination of signal processing and deep learning [22], [24], or solely deep learning [23], in part due to the advent of deep learning.

In recent years, there have been ambitious attempts to perform automatic speech recognition for indoor use [25].

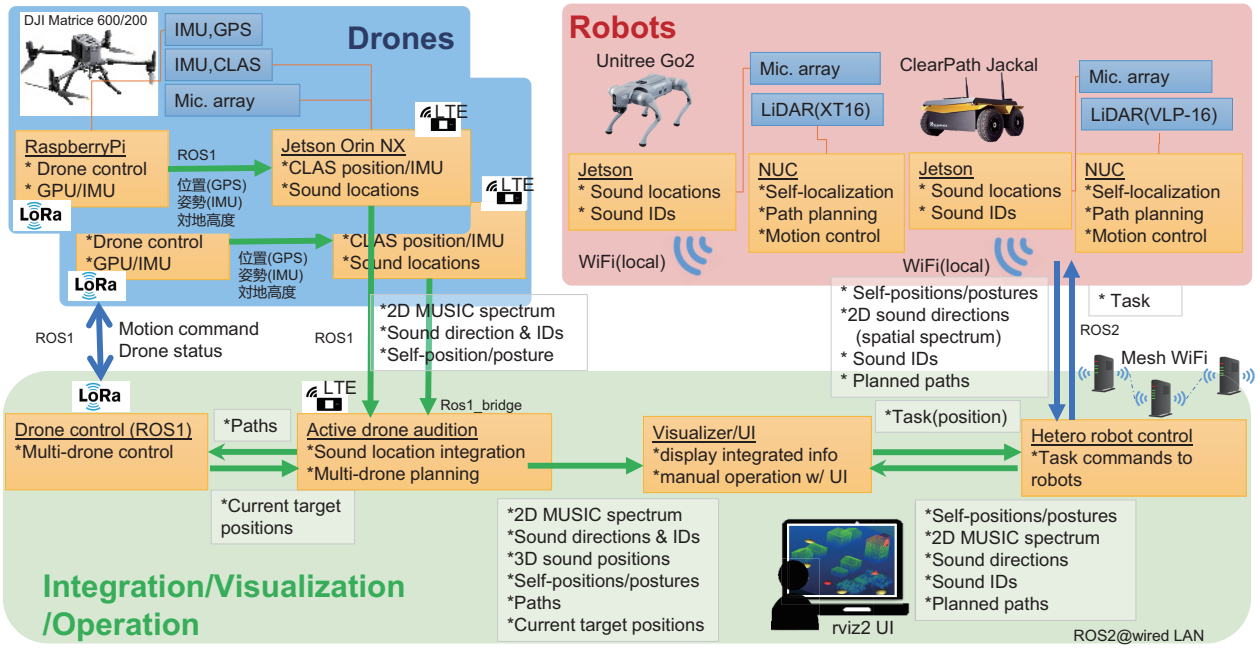


Fig. 1. The Architecure of Swarm Active Audition System

Additionally, studies have been conducted to explore the potential of employing the stop-perceive-act strategy in robot audition [1]. This study presents a unique and practically effective approach that attempts to address the noise issue by utilizing a kiteplane that can glide with its rotors stopped and listen to sounds during quiet periods when the rotors are stopped [26].

For sound source localization and tracking, a method for estimating the 3D position of a sound source has been reported. However, this method is limited by two assumptions: first, that the source is on the ground [10], [11]; and second, that multiple observations are required to localize sound sources [9], [26]. The aforementioned assumption is overly simplistic and therefore unsuitable for application in search and rescue situations, given that a person in need of rescue may be located on an upper floor of a building. The latter approach presents two significant challenges. Firstly, the time required to locate a person in need of rescue is considerable. Secondly, the search area narrows considerably with the application of this approach.

To address these challenges, research has been conducted employing the use of multiple drones simultaneously [27]–[29]. The use of multiple drones allow for the application of triangulation to achieve 3D localization without the aforementioned assumptions. Yamada et al. identified a limitation of the method, namely that the source directions obtained from multiple observations do not necessarily intersect at a single point. They proposed a solution, which involved the use of a Gaussian filter [27]. This was enhanced by replacing it with particle filtering with integrated MUSIC (PAFIM) [28], and by introducing Wiener filtering [29]. Additionally, they proposed the concept of "active drone audition," in which multiple drones proactively undertake optimal positioning to enhance target sound source localization [28]. However, these studies evaluated the efficacy of the proposed methods exclusively through numerical simulations using audio signals synthesized based on the recorded audio data.

In light of the aforementioned limitations, we propose the swarm active audition system (SAAS), which employs a heterogeneous fleet of drones and robots to facilitate human intervention, as an extension and realization of the PAFIM-based sound source localization and tracking method. Our objective is not merely to ascertain the functionality of SAAS in a real-world setting; We are also interested in its potential utility in a real-world disaster scenario. We aim to ensure that our research is not merely theoretical but has tangible, practical applications, this paper aims to contribute to the following discussions:

- The proposal of SAAS for the search and rescue of individuals in need in a disaster setting.
- An overview of the advancements made in SAAS, with a particular emphasis on its drone audition functionality.
- A survey of the system requirements necessary for effective operation in actual disaster areas, and an examination of the extent to which the proposed SAAS system meets those requirements.

II. Swarm Active Audition System

Figure 1 illustrates the architecture of the swarm active audition system (SAAS). It consists of three principal components: **Drones**, **Robots** and Integration/Visualization/Operation (**IVO**).

Drones ensures the *high speed* and *wide area* of the search by enabling 3D sound source localization through the use of multiple drones. This component receives PAFIM results implemented in the active drone audition module and optimally positions the drones. In addition to sound source localization

and tracking, each drone is also responsible for sound source classification in conjunction with sound source separation.

Robots ensures the *accuracy* of sound source localization and the feasibility of *runnability*. **Drones** is capable of traversing distances at high speeds, yet their search patterns are relatively coarse. In contrast, **Robots** moves at slower speeds, employing a stop-perceive-act strategy to focus intensively on areas where rescuers are likely to be located with high *accuracy*. In addition, a wheeled robot and a quadruped robot are employed together, with the optimal robot selected according to the area to be searched to enhance *runnability*. Furthermore, **Robots** play a pivotal role in updating the 3D map in real-time using onboard LiDAR [30], and in performing path planning on the updated map through a graph search-based method [31].

IVO guarantees *system robustness* and *user friendliness*. For *system robustness*, three measures must be taken to achieve low latency and robust data communication. First, a combination of multiple wireless communication paths is employed. Specifically, for communication between **Drones** and **IVO**, multiple communication methods, including WiFi, LTE, and LoRa, are employed in conjunction to ensure the robustness of communication and connection. In order to facilitate communication between **Robots** and **IVO**, Mesh WiFi is adopted to ensure transparent communication over a broad area. For stable communication within **IVO**, wired connections are employed. Furthermore. the benefits of ROS2 are maximized through the use of distributed modules across multiple PCs. Second, data compression is performed with consideration of the bandwidth limitations inherent to each communication path. For example, there is no requirement for **IVO** to convey precise information. In the event that the objective is solely for display purposes, the quantity of information may be diminished to the extent that it does not impede the user's comprehension. Furthermore, discretization does not necessarily need to be performed in an even manner. Intelligent data compression should be performed by widening the range of discretization in the range where information degradation is not a problem, and by narrowing the range of discretization in the range where information degradation is a problem. Finally, ROS2 is selected due to its stability for large-scale systems through the utilization of the data distribution service (DDS). However, initially, ROS1 will be used for communication within **Drones** and between **Drones** and **IVO**, due to the constraints of the embedded system of drones.

To enhance *user friendliness*, the user interface (UI) will be designed to facilitate comprehension of the situation at a glance for users who are not familiar with the system. They will be achieved by eliminating unnecessary sound sources, such as drone noise, and by employing color-coding to differentiate between the various sound sources displayed on the 3D map according to their respective types. Futhermore, the sound sources are displayed as ellipsoids that take the confidence level into account, thereby providing information on the sound sources that should be prioritized. Additionally, UI allows the human operator to intervene with the system at any time as needed based on such information, for example, to specify areas to be searched intensively.



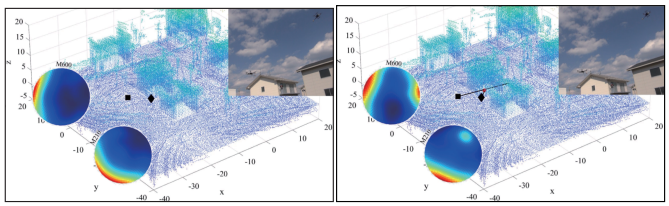
Fig. 2. DJI Matrice 600 Pro(left) and DJI Matrice 210 RTK V(right). Each has a 16 ch spherical microphone array (white) at the end of the pole which is covered by windshield made of special urethane with light weight.



Fig. 3. Clearpath JACKAL(left) and Unitree Go2(right). Each has a 16 ch cylindrical microphone array and LiDAR.



Fig. 4. Generated 3D Point Cloud for Fukushima RTF town area field



a) Two drones are hovering (black square/diamond)

b) A sound source at 2F is successfully detected (red dot)

Fig. 5. Sound Source Localization with Multiple Drones

The construction of SAAS is scheduled for completion by November 2024, along with the online demonstration. The following demonstration will be taken into consideration. The two drones depicted in Fig. 2 and the two robots depicted in Fig. 3 are utilized. (1) The two drones are capable of performing 3D multiple sound source localization broadly over a wide area. (2) Each type of sound source detected is displayed on a 3D point cloud map as an ellipsoid of a different color, with the diameter corresponding to the error margin. (3) The wheeled and legged robots initiate high-precision sound localization and autonomously move to the location of the sound source to identify the individual in need of rescue.

A. System Performance

At the time of writing this paper, the system construction is still in progress. However, the implementation of each component has been completed, and the communication between the three components has been verified. This section presents illustrative examples of the operation of each component on a case study basis.

Figures 2 and 3 illustrate the drones and robots used in the test. A variety of drones and robots with disparate characteristics were selected for inclusion in the study. These included the DJI Matrice 600 Pro (six propellers), the DJI Matrice 210 RTK V (four propellers), the Clearpath JACKAL (four wheels), and the Unitree Go2 (four legs). Figure 4 depicts a 3D point cloud map of the town area at Fukushima Robot Test Field (RTF), which serves as the test site. This 3D point cloud map was generated in advance using the two robots used in **Robots**. Despite the town area comprising several blocks, including residential and debris areas, and an intricate internal structure, the map was successfully generated.

The objective of the test was to ascertain the ability of the two drones to detect calls for assistance from the second floor of House A, which is included in the target scenario. Figure 5 shows the situation of the test. As illustrated in Figure 5a), the two drones are hovering around 10 m from the sound source. The drones are shown as photos in the upper right view and also as a black square and diamond on the 3D point cloud map. The black square and diamond correspond to the DJI Matrice 600 Pro and 210 RTK V, respectively. The two circles on the left represent sound intensity maps, with red indicating high sound pressure levels from the corresponding direction, and blue indicating the opposite. The red areas within the circles in this figure are caused by drone propeller noise during hovering.

In Figure 5b), the individual on the second floor initiated a call for help. The two circles exhibited red and light blue areas, which were caused by the calls. At the same time, two black lines emerged from the black square and diamond, indicating the direction of the sound source as observed from the drone's perspective. In general, the two lines do not intersect in 3D space. However, PAFIM was able to successfully identify and estimate the position of the sound source, marked as a red dot, based on particle filtering.

III. Survey from rescue workers at disaster site

Drone audition has been studied in a seed-driven manner, with the interest in the effectiveness of acoustic signal processing and machine learning technologies in harsh environments. This strongly influences the statement that the application of drone audition technology is to search for people in need of rescue. It remains unclear whether drone audition technology and SAAS are genuinely beneficial or necessary in such contexts. This section presents the identified needs within the field.

We, thus, investigated the expectations for drone audition technology and its segregation from other technologies through interviews with the results of drone utilization at disaster sites to date and expectations for drones at disaster sites.

TABLE I
Interview Items (in excerpts)

Class	Content
Overview	Purpose of using drones in S&R operations Cases where drone audition would have contributed to S&R? Provide info on the status of drone operations in the field
Data	Who will use it for what and how? Challenges in camera and thermography? Cases already using auditory technology?
Env.	Provide the surrounding environment in the field Challenges in data communication? Any other sound sources?
Use case	Any use of drones in peacetime? Any chances for drone audition except for S&R

The four parties interviewed were two prefectural fire departments and two companies that engage in drone operation. All four were involved in rescue operations following the Noto Peninsula earthquake in January 2024.

Table I enumerates the selected interview items, which were designed by considering issues at disaster sites, the environment at disaster sites, the needs of robot audition technology, and comparison with other technologies. The interviews were conducted in approximately one hour per party, allowing for a comprehensive examination of the specific details, rather than a mere affirmation or negation.

a) Issues at disaster sites: Only one of the four parties had prior experience in searching for individuals in need of rescue. This is due to the fact that the request for rescue was made more than two weeks after the disaster occurred, and the timing for searching for those in need of rescue was missed. In practice, most the drone applications are mainly for creating orthoimages to assess the situation. The party with the experience indicated that the search for rescuers was conducted manually, rather than by drone. The first priority in the search for rescuers is the safety of the search party members of the search party, and it takes time to ensure their safety before conducting the search.

b) Environment at disaster site: While environmental conditions may vary considerably from one location to another, the following examples are representative of the typical circumstances encountered.

- In order to generate ortho imaging, the drone is often required to undertake a flight of between 10 and 20 minutes in duration, at a speed of between 2 and 3 m/s and an altitude of between 50 and 100 m.
- There are a variety of sound sources at the site, and the major loud sources include wind, people, heavy machinery, helicopters.
- In the case of a fire, the sounds of flames and water spray are dominant enough to drown out the drone noise.

c) Needs of robot audition technology: Positive comments are as follows:

- It would be an efficacious strategy to utilize the drone audition in conjunction with the drone-equipped loudspeaker to call out to individuals.
- In the context of the transportation of medical supplies, it is essential to ascertain whether the intended recipient has indeed received the items in question. As an extension of

the drone audition, it would be advantageous if mutual communication via verbal means were possible.
- In mountain rescue contexts, it is not uncommon for rescuers to be difficult to detect victims, even when equipped with cameras or infrared cameras, due to the presence of trees and other obstructions. The ability to detect the voice of a person in need of rescue would therefore be a valuable addition to current technology. Furthermore, the potential for mutual communication by voice would enhance the efficacy of this system.
- Ultimately, the rescue of individuals from collapsed buildings relies on human hearing. Given that a person cannot commence a search until they have ensured their own safety, it would be beneficial if drone audition could be employed to augment the search process.

There was a negative option that rescuers may have difficulty speaking loudly, given that they are often in a debilitated state. It seems unlikely that a low volume voice can be detected by drone audition.

d) Comparison with other technologies: In practice, the search process heavily relies on the human senses, including hearing. Nevertheless, they are willing to adopt new technologies that are expected to be effective. In fact, imaging and electromagnetic wave search devices have been employed to search for rescuers. However, all of these devices have their respective advantages and disadvantages, including the potential for delays and a restricted search area. It is notable that no single method has been definitively identified as the optimal solution.

IV. Discussions

This section presents the findings of the survey and an analysis of the extent to which the proposed SAAS aligns with user requirements. It is also necessary to determine whether the functions provided by SAAS are essential or whether there are functions currently lacking in SAAS.

A. Discussion on the survey results

In the context of disaster site operations, insights were gleaned from a number of valuable and informative sources. With regard to the issues and comparisons presented, it is evident that the use of drones for the purpose of searching for rescuers is not yet a common practice, regardless of the type of sensors employed. Moreover, even when searching for individuals in need of assistance, the search is still conducted manually. This is largely attributable to structural deficiencies in search and rescue operations and to the awareness for the search activities. Furthermore, the sensing devices currently available for search and rescue operations are still in their infancy, necessitating the urgent establishment and deployment of technologies for locating individuals in need of rescue. It is similarly important to familiarise people with the technology in question, thereby enabling them to utilise it for available tasks during normal times.

In the context of disaster site environments, it is important to recognise that there are a multitude of sound sources present, beyond those originating from targets. This indicates that, in addition to sound source localization and tracking, sound source classification is an effective approach. In comparison to orthoimagery, drone audition searches for rescuers are likely to necessitate the implementation of lower altitudes. To guarantee the security of personnel operating drones at such altitudes, it is imperative to enhance automated control systems and technologies designed to assist drone operators, including dynamic mapping and path planning.

With regard to the requirements of drone audition technology, there is a consensus among the relevant parties that the drone audition function would be beneficial in a disaster scenario. Furthermore, the participants expressed greater enthusiasm for the potential of the drone audition function to facilitate communication between the drone and rescue personnel. Conversely, there is a concern that the drone may not be effective when the volume of the voice from the rescuers is low.

B. Discussions on the SAAS functions

In light of the aforementioned discussion, an examination will be conducted to ascertain whether the current SAAS comprises the requisite functions for those working at disaster sites, and whether any functions are absent.

The initial step is to examine the function of drone audition. The principal functions of **Drones** are 3D sound localization, high-speed and wide-area search, and sound source classification. The initial two functions will be invaluable in the field for two key reasons: One factor is that the circumstances at the site of a search and rescue operation are known to be variable, and the location of the rescuers cannot be assumed to be known in advance. The second reason pertains to the fact that, given the current reliance on human resources for search operations, the safety of the searchers takes precedence over the search itself. The third function is also essential for an efficient search, given the presence of a multitude of potent sound sources at the site. Conversely, it is postulated that the issue of detecting low voices can be resolved by flying the drone at lower altitudes and utilizing **Robots** as previously introduced in SAAS. It may be necessary in the future to introduce a stop-perceive-act strategy for **Drones** as well.

Robots has two main functions; enhancement of the runnability using heteogeneous robots and dyamic map generation with path planning. The first function is a suitable option, as no assumption about the circumstances can be put in disaster site. The second function is advantageous in situations where the specific circumstances of the disaster site are not immediately apparent. Moreover, it is beneficial for the secure operation of drones at low altitude, which may be necessary for drone audition tasks. One concern is how well the robot can detect low-volume that cannot be detected by the drone. A detailed evaluation will be necessary.

IVO is designed to enhance its robustness through the integration of diverse wireless communication channels, the implementation of data compression techniques that account for the constraints of communication bandwidth, and the

incorporation of ROS2. No opinions were identified that would impede the implementation of these measures. Since a variety of wireless channels are in use at disaster sites, the ability to secure multiple wireless communication paths is considered to be advantageous. As we felt at the time of the interviews, the people working at the disaster site are basically not experts in engineering technology. It is therefore essential to provide a UI that allows the users to grasp the situation at a glance. Several requests were received for mutual voice communication with the rescuers. This is not currently planned for SAAS and is an item that should be seriously considered for UI implementation in the future.

V. Conclusion

In this paper, we propose a swarm active audition system (SAAS) in which multiple drones and robots can work together to search for people in need of rescue, and demonstrate its effectiveness, focusing on sound source localization by multiple drones. In addition, to verify whether SAAS is useful in actual disaster sites, we interviewed parties who have engaged in disaster site work and discussed its usefulness by contrasting the results of the interviews and the functions of SAAS. As a result, the SAAS functions are all in line with user requests although a concern if detecting low volume voices raised, which can be solved by total SAAS not only by drone audition. In addition, there were strong requests for a voice communication function as a function desired in the future. We will consider introducing this function in the future.

Acknowledgment

This work was supported by KAKENHIJP22F22769 and JP22KF0141, and F-REI (JPFR23010102).

References

[1] K. Nakadai *et al.*, "Active audition for humanoid," in *AAAI*, AAAI, 2000, pp. 832–839.

[2] K. Okutani *et al.*, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *IROS*, IEEE, 2012, pp. 3288–3293.

[3] K. Nakamura *et al.*, "Real-time super-resolution sound source localization for robots," in *IROS*, IEEE, 2012, pp. 694–699.

[4] M. Basiri *et al.*, "Robust acoustic source localization of emergency signals from micro air vehicles," in *IEEE/RSJ IROS*, 2012, pp. 4737–4742.

[5] T. Ohata *et al.*, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *IROS*, 2014, pp. 1902–1907.

[6] K. Furukawa *et al.*, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *IROS*, IEEE, 2013, pp. 3943–3948.

[7] T. Ishiki and M. Kumon, "Design model of microphone arrays for multirotor helicopters," in *IROS*, 2015, pp. 6143–6148.

[8] K. Washizaki *et al.*, "Position estimation of sound source on ground by multirotor helicopter with microphone array," in *IROS*, 2016, pp. 1980–1985.

[9] K. Yamada *et al.*, "Belief-driven control policy of a drone with microphones for multiple sound source search," in *IROS*, 2019, pp. 5326–5332.

[10] K. Nakadai *et al.*, "Development, deployment and applications of robot audition open source software HARK," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 16–25, 2017.

[11] K. Hoshiba *et al.*, "Design and assessment of sound source localization system with a UAV-embedded microphone array," *JRM*, vol. 29, no. 1, pp. 154–167, 2017.

[12] L. Wang and A. Cavallaro, "Time-frequency processing for sound source localization from a micro aerial vehicle," in *ICASSP*, 2017, pp. 496–500.

[13] J.-S. Lauzon *et al.*, "Localization of RW-UAVs using particle filtering over distributed microphone arrays," in *IROS*, 2017, pp. 2479–2484.

[14] M. Strauss *et al.*, "DREGON: Dataset and Methods for UAV-Embedded Sound Source Localization," in *IROS*, IEEE, 2018, pp. 5735–5742.

[15] Y. Hioka *et al.*, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," in *IWAENC*, IEEE, 2016, pp. 1–5.

[16] S. Yoon *et al.*, "Two-stage adaptive noise reduction system for broadcasting multicopters," in *2016 IEEE International Conference on Consumer Electronics (ICCE)*, 2016, pp. 219–222.

[17] L. Wang and A. Cavallaro, "Acoustic sensing from a multi-rotor drone," *IEEE Sensors Journal*, vol. 18, no. 11, pp. 4570–4582, 2018.

[18] Z.-W. Tan *et al.*, "An efficient dilated convolutional neural network for UAV noise reduction at low input SNR," in *APSIPA ASC*, 2019, pp. 1885–1892.

[19] B. Yen *et al.*, "Multi-sensory sound source enhancement for unmanned aerial vehicle recordings," *Applied Acoustics*, vol. 189, p. 108 590, 2022.

[20] W. N. Manamperi *et al.*, "Drone audition: Audio signal enhancement from drone embedded microphones using multi-channel Wiener filtering and Gaussian-mixture based post-filtering," *Applied Acoustics*, vol. 216, p. 109 818, 2024.

[21] L. Wang and A. Cavallaro, "Deep learning assisted time-frequency processing for speech enhancement on drones," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 6, pp. 871–881, 2021.

[22] S. Uemura *et al.*, "Outdoor acoustic event identification using sound source separation and deep learning with a quadrotor-embedded microphone array," in *The 6th International Conference on Advanced Mechatronics*, 2015.

[23] T. Morito *et al.*, "Partially shared deep neural network in sound source separation and identification using a UAV-embedded microphone array," in *IROS*, 2016, pp. 1299–1304.

[24] O. Sugiyama *et al.*, "Outdoor acoustic event identification with DNN using a quadrotor-embedded microphone array," *JRM*, vol. 29, no. 1, pp. 188–197, 2017.

[25] T. Fuhrman *et al.*, "An interactive indoor drone assistant," in *IROS*, 2019, pp. 6052–6057.

[26] M. Kumon *et al.*, "Alternating drive-and-glide flight navigation of a kiteplane for sound source position estimation," in *IROS*, 2021, pp. 2114–2120.

[27] T. Yamada *et al.*, "Assessment of sound source tracking using multiple drones equipped with multiple microphone arrays," *International Journal of Environmental Research and Public Health*, vol. 18, no. 17, 2021, ISSN: 1660-4601.

[28] T. Yamada *et al.*, "Placement planning for sound source tracking in active drone audition," *Drones*, vol. 7, no. 7, 2023.

[29] B. Yen *et al.*, "A performance assessment on rotor noise-informed active multidrone sound source tracking methods," *Drones*, vol. 8, no. 6, 2024.

[30] S. Niijima *et al.*, "Semi-automatic town-scale 3D mapping using building information from publicly available maps," *IEEE Access*, vol. 10, pp. 32 244–32 254, 2022.

[31] Y. Sato *et al.*, "STP4: Spatio temporal path planning based on pedestrian trajectory prediction in dense crowds," in *RiTA 2022: The 10th International Conference on Robot Intelligence Technology and Applications*, 2022.

A Noisy Context Optimization Approach for Chinese Spelling Correction

Guangwei Zhang[1], Yongping Xiong[1], Ruifan Li[2]*
[1]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
[2]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
E-mail: {gwzhang, ypxiong, rfli}@bupt.edu.cn

***Abstract*—The task of Chinese Spelling Correction (CSC) aims to detect and correct Chinese spelling errors. Recently, BERT-based models have dominated the research on CSC. These methods suffer from unsatisfactory performance when dealing with noisy contexts since incorrect characters are often influenced by other typos. In this paper, we propose A Noisy Context Optimization approach for Chinese Spelling Correction (NCO-Spell). Specifically, in the pre-training stage, the multi-character masking strategy is proposed to construct multi-typo texts, enabling the model more robust in noisy environments. Furthermore, we increase the size of the confusion set by dynamically updating. Second, in the inference stage, an iterative algorithm is incorporated to correct the wrong characters one by one, and each iteration will gradually reduce the noise. Extensive experiments and detailed analyses on a widely used benchmark demonstrate that NCO-Spell is effective.**

I. Introduction

Chinese contains more than 10,000 characters, of which 3,500 are common characters [1]. Usually, spelling errors could occur. These spelling errors are typically caused by human writing, automatic speech recognition, or optical character recognition systems. As shown in Table I, the first case is that the sentence contains only one typo, and its semantics are similar. The second case contains multiple consecutive different typos, and they are pronounced similarly.

Chinese spelling correction (CSC) task has been proposed. The mainstream methods is using language models [2], [3]. Recently, BERT [4] has been proven to be effective in various applications and has become the fundamental model for CSC tasks, achieving the best performance [5]–[8]. However, BERT is a masking model that masks characters based on a specific probability. For multi-typo texts, the context of each character contains at least one typo, which results in noisy information. BERT could not effectively deal with this situation.

Basically, Chinese spelling errors are mainly caused by the misuse of phonetically or visually similar characters. Therefore, it is of importance to incorporate the similarity knowledge of characters. To this end, some works utilize confusion sets to incorporate this knowledge [9], [10]. However, confusion sets are usually generated by heuristic rules or human annotations, which have limited coverage and the similarity knowledge of characters is not fully exploited.

In this paper, we propose NCO-Spell, a **N**oisy **C**ontext **O**ptimization Approach for CSC task. Firstly, we propose a multi-character-based masking strategy. This strategy simulates multiple adjacent misspellings in the dataset and increases the masking of dynamic confusion sets. We update the uncorrected characters in the training process into the confusion set and obtain a dynamic confusion set. Secondly, we propose a method of iterative inference in the inference stage. Only the wrong character with the highest probability will be corrected each time. The corrected result is used for the next iteration to gradually reduce the influence of noise. Thirdly, we conduct experiments on the widely used benchmark dataset, i.e., SIGHAN15 [3]. The results show that our approach outperforms all compared methods.

The major contributions are highlighted as follows. 1) We propose a masking strategy based on multi-characters, which makes the model more robust to noisy contexts. 2) We propose an iterative inference method to gradually reduce the impact of noise. 3) We conduct extensive experiments to show our method outperforms state-of-the-art baselines.

TABLE I: Examples of Chinese spelling errors. Misspelling characters are marked in red. The correct characters are marked in blue.

Type	Input	Translation
Wrong	他没学过汉子。	He has never learned a man.
Correct	他没学过汉字。	He has never learned Chinese characters.
Wrong	下次我们一起持分号码?	Shall we hold the number together next time?
Correct	下次我们一起吃饭好吗?	Shall we have dinner together next time?

II. Related Work

CSC is an important and challenging task in natural language processing. It mainly needs to detect typos based on the judgment of the semantics and correct these typos with a full understanding of the context. Most of the early works used statistics and rules for detection and correction and then adopted the credibility of the language model for judgment [3], [11]. These error correction techniques calculate the sentence composition probabilities through local context information and failed to consider the global context information.

Recently, BERT [4] has been proposed, which is a bidirectional language model based on transformer encoders. More and more works use BERT-like models to directly map each character in a sentence to the correct character for CSC task. For example, considering that Chinese typos are mostly

*Corresponding author.

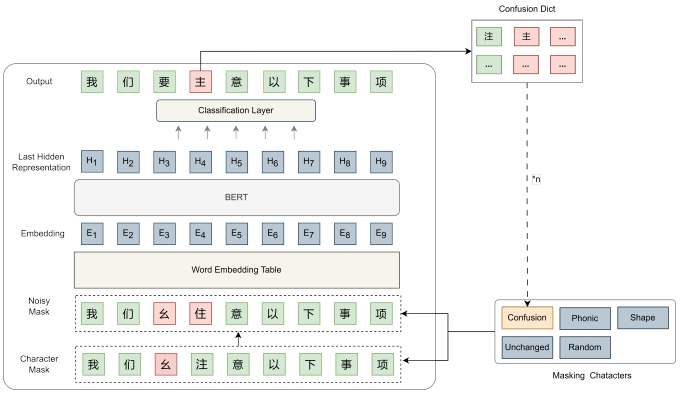


Fig. 1: Framework of NCO-Spell. Take the sentence ”我们要注意以下事项” as an example. On the left, this component presents the masking strategy, which forces the model to construct a noisy context for the character ”幺”. On the right, this component presents the update module for the confusion set, which illustrates how our approach adds new error-prone characters ”注-主” into the confusion set.

similar in pronunciation or shape, Hong et al. [12] improved the masking method and added a confidence-similarity filter in the decoding stage. Cheng et al. [5] constructed graphs for the relationship between the phonetic and grapheme of characters to participate in character classification to fuse similar information of characters. Liu et al. [6] used similar characters for masking in the pre-training stage, integrates phonic and stroke information into the output, and learns more character information. Huang et al. [13] fused phonetics, glyphs, and character information through a gating system in the fine-tuning stage to obtain more character information. Furthermore, Li et al. [14] refined the knowledge representations of pre-trained language models, and guides the model to avoid predicting these common characters through an error-driven way. Li et al. [15] proposed the LEAD framework, which renders the CSC model to learn heterogeneous knowledge from the dictionary in terms of phonetics, vision, and meaning. However, only relying on the previous random masking, BERT could not effectively deal with this situation.

In addition, some works are proposed based on confusion set [3], [5], [6], [10]. Wang et al. [10] proposed a pointer network model guided by a confusion set for Chinese spelling correction. Cheng et al. [5] used a confusion set in the final character classification. Liu et al. [6] used a fixed-size phoneme and grapheme confusion set to obtain character information. The performance of these models heavily relies on the quality of the confusion set.

III. Methodology

In this section, we describe the proposed approach in detail.

A. Task Formulation

Chinese Spelling Correction (CSC) task can be formalized as a sequence labeling problem. Given a text sequence of n Chinese characters $X = (x_1, x_2, ..., x_n)$, the goal is to output $Y = (y_1, y_2, ..., y_n)$, where X represents the original text containing some error characters, and Y represents the correct text after correction.

TABLE II: Probability Distribution of Different Masking Strategies.

Strategy	Random	Unchanged	Phonic	Shape	Confusion
PLOME	10%	15%	60%	15%	-
NCO-Spell	10%	15%	50%	15%	10%

B. Multi-character Mask

The framework of our proposed NCO-Spell is shown in Fig. 1. The input of this module is the sequence of embeddings $E = e_1, e_2, ..., e_n$, where e_i denotes the embedding of character x_i in a given text $X = (x_1, x_2, ..., x_n)$. Then, E is fed into the transformer encoder and the encoder generates hidden representation matrix $H = (h_1, h_2, ..., h_n)$ for H, where $h_i \in R^{768}$ is the representation of x_i. Finally, the final output of this module is fed into a one-layer classification network with softmax normalization.

In order to construct the noisy context, after masking the selected characters, we mask the other characters with a certain probability in the sentence to simulate the situation of multi-typo text. When selecting the source of mask characters, we added dynamic confusion set to enhance the model's learning of error-prone characters. In the following subsections, we will introduce the masking strategy and the dynamic confusion set in the masked content.

Masking Strategy. To make the model robust to the noisy context, we propose a multi-character masking strategy for the pre-training corpus. This method contains two key points: 1) Masking Characters. Because the frequency of phonetic errors is about twice that of visual errors [16], so in the masking strategy, the proportion of characters that are phonetically related to the original character is large, and we increase the selection of the dynamic confusion set. Finally set the corresponding selection probability of each strategy as shown in Table II. 2) Masking Position. The contextual disturbance mainly occurs in noisy environments around misspelled characters, so the distance between these masked characters and the original typo is defined as the ”masking distance”. In Fig. 1, the masking distance is 1. If there are no characters to mask in the original sentence, the original text will not be masked in any positions.

Dynamic Confusion Set. The initial confusion set is constructed by Wang et al. [9]. When using a confusion set, the performance of the model generally depends heavily on the quality of the confusion set, and it is difficult to find the latest obfuscated characters that are beyond the existing field. As the training iterations progress, the model will continue to generate new errors, and characters will be associated with more error-prone characters, so the size of the confusion set also needs to be continuously updated. The dynamic update of the confusion set includes two points: **1**) Selection of Samples. During the

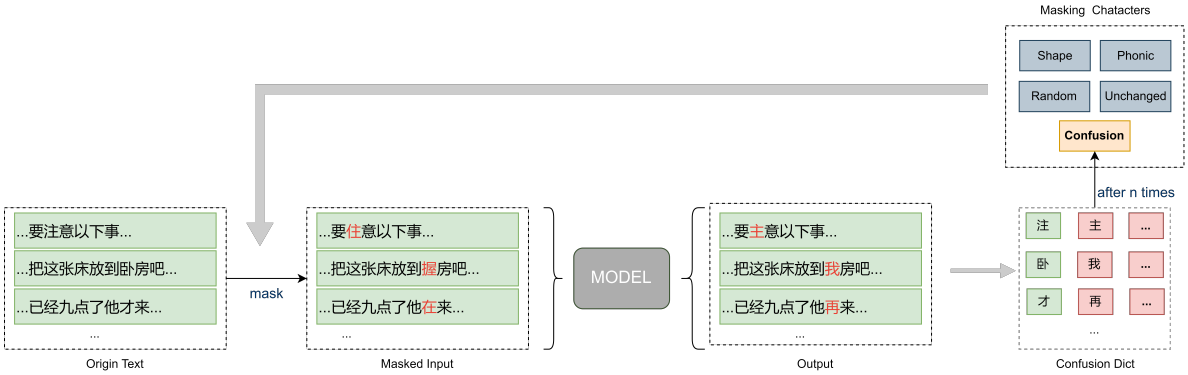


Fig. 2: Illustration of Updating Strategy on the Confusion Set. These are some sentences where the correction fails. For example, "注" is masked to "住" and as the input. After being corrected by the approach, "主" is obtained, which is the uncorrected character. Then our approach temporarily stores correct and uncorrected in the confusion dictionary and updates them into the confusion set after n times of training.

training process of the model, there will be correction failures, including changing correct characters to other characters or incorrect characters not being corrected successfully. The result is characters that the model believes fit the context but are actually incorrect, they are the model's confusing characters. Such characters will be updated into the confusion set as new confusion set samples. **2**) Dynamic Update Method. After obtaining new samples of the confusion set, temporarily store the new samples of the confusion set for several training times in a dictionary, and update the data in the dictionary into the confusion set after this training process. In order to ensure the quality of the confusion set, the confusion character pairs to be updated will be added to the confusion set only after the loss is reduced to a certain threshold. In the subsequent training process, the corresponding characters in the confusion set will be randomly used for masking. The confusion set update strategy is shown in Fig. 2.

Through dynamically updating the confusion set, we obtain more error-prone characters. The initial confusion set contains 4,922 characters, and each character corresponds to 7.8 confusion characters on average. In contrast, the dynamic confusion set contains 6,685 characters, and each character corresponds to 116 confusion characters on average.

C. Iterative Inference

In order to further illustrate the problem of the current model in multi-typo text, we use NCO-Spell to conduct experiments on some sentences containing multi-typos, and the correction results are shown in Table III.

It can be found that affected by the noise of the context, it is often impossible to completely correct all typos, especially when there are continuous errors, the errors in the middle position are affected by the noise, so they cannot be detected. There are a total of 1100 sentences in the evaluation set of SIGHAN15, of which 123 sentences contain multi-typos, accounting for about 11%. Therefore, it is necessary to optimize the error correction for multi-typo texts.

TABLE III: Examples of NCO-Spell Results for Multi-typos.

Type	Text	Translation
Origin	他说「请座，请座」。	He said "Please take a seat, please take a seat".
Result	他说「请座，请坐」。	He said "Please take a seat, please sit down".
Origin	有吗吗、弟弟还有我。	Is there any, My brother and me.
Result	有妈吗、弟弟还有我。	Do you have a mother, brother and me?
Origin	下次我们一起持分号码?	Shall we hold the number together next time?
Result	下次我们一起吃分号吗?	Shall we have a semicolon next time?

The iterative inference method aims to reduce the noise effect in the sentence by correcting only the character with the highest probability in each iteration, putting the corrected sentence as input into the next iteration for correction, and so on until there is no need to correct characters. In this way, on the one hand, the influence of the noisy context can be gradually reduced, on the other hand, the recall rate can be improved, all wrong characters can be corrected as much as possible, and each corrected character is the character with the highest confidence, and each iteration obtains an optimal solution of the sentence. Iterative Inference is model-agnostic and it can be used in other CSC methods to achieve further improvements. The flow chart of the inference algorithm is shown in Fig. 3.

IV. Experiments

We present details for pre-training NCO-Spell and the fine-tuning results on the most widely used benchmark dataset.

A. Pre-training Settings

Dataset. We use ChineseNlpCorpus[1][2] as the original pre-training data. We decompose the information into sentences using periods, question marks, exclamation points, and other punctuation marks indicating the end, and after the screening, 110.7 million effective sentences are formed.

[1]https://github.com/SophonPlus/ChineseNlpCorpus
[2]https://github.com/brightmart/nlp_chinese_corpus

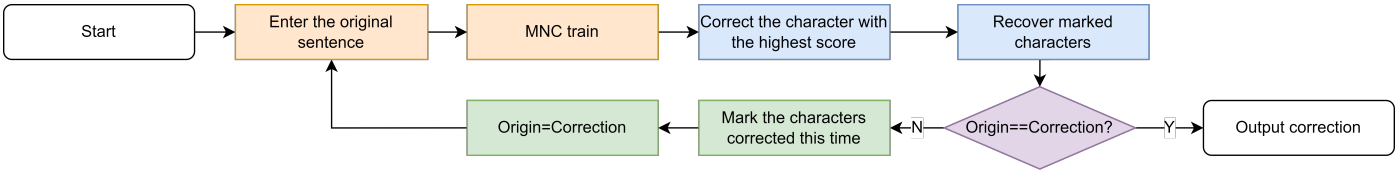


Fig. 3: Flowchart of Iterative Inference. A character will be corrected only once, and only the character with the highest score will be corrected each time.

TABLE IV: The performance of our approach and baseline models on SIGHAN15. The results at the bottom four rows are our implementation. ECOPO indicates the improvement of the model for REALISE. The symbol '-' means no results are reported in the corresponding papers.

Method	Character-level (%)						Sentence-level (%)					
	Detection			Correction			Detection			Correction		
	P	R	F	P	R	F	P	R	F	P	R	F
BERT [4]	90.9	84.9	87.8	95.6	81.2	87.8	68.4	77.6	72.7	66.0	74.9	70.2
PLOME [6]	94.5	87.4	90.8	97.2	84.3	90.3	77.4	81.5	79.4	75.3	79.3	77.2
REALISE [18]	-	-	-	-	-	-	77.3	81.3	79.3	75.9	79.9	77.8
ECOPO [14]	-	-	-	-	-	-	77.5	82.6	80.0	76.1	81.2	78.5
LEAD [15]	-	-	-	-	-	-	79.2	82.8	80.9	77.6	81.2	79.3
CoSPA [19]	**95.9**	**88.6**	**92.1**	**98.5**	85.3	91.4	79.0	82.4	80.7	76.7	80.0	78.3
PGBERT [20]	95.2	87.5	91.2	97.2	85.1	90.8	78.1	82.8	80.4	76.0	80.6	78.2
PLOME(cfs)	87.3	87.0	87.1	97.4	84.7	90.7	79.3	82.2	80.7	77.4	80.2	78.7
NCO-Spell	89.7	86.8	88.2	98.2	85.3	91.3	81.5	82.7	82.1	79.9	81.1	80.5
PLOME(iter)	88.6	88.1	88.3	98.3	**86.5**	**92.0**	82.1	83.3	82.7	80.5	81.6	81.0
NCO-Spell(iter)	88.7	88.1	88.4	98.3	**86.5**	**92.0**	**82.3**	**83.5**	**82.9**	**80.6**	**81.8**	**81.2**

Parameter Settings. The composition of the transformer encoder is the same as BERT, with a learning rate of 5e-5. These parameters are set empirically because pre-training is expensive. This article does not pre-train from scratch, but trains based on the pre-trained model released by the PLOME model. The threshold of the loss is set to 4.0, and only when the loss is lower than the threshold, the wrong character will be added to the confusion dictionary, and the confusion set is updated every 5 training epochs.

B. Fine-tuning Settings

Training Data. Following Liu et al. [6], the training data is composed of 10K manually annotated samples from SIGHAN [2], [3], [17] and 271K automatically generated samples from Wang et al. [9].

Evaluation Data. We use the latest SIGHAN test dataset [3] as in PLOME to evaluate the proposed approach.

Parameter Settings. Following Liu et al. [6], we set the maximum sentence length to 180, batch size to 32, and the learning rate to 5e-5. Take the most final training results with the best model in 10 rounds of evaluation effects.

Evaluation Metrics. Following previous work [5], [6], we use the precision (P), recall (R), and F1 scores as the evaluation metrics. Besides, for character-level evaluation, we also report sentence-level metrics on the detection and correction subtasks.

C. Baseline Models

We use the following methods for experimental evaluation. **1**) BERT [4]: The word embedding is used as the input of BERT, and the correct character is predicted based on the mask model. **2**) PLOME [6]: This method adopts a pre-training masking strategy with misspelling knowledge, and utilizes GRU networks to model such knowledge based on characters' phonics and strokes. **3**) REALISE [18]: This method captures the semantic, phonetic, and graphic information of the text, and mixes the information in these modalities to predict the output. **4**) ECOPO [14]: This method refines the knowledge representation of the pre-trained language model, it can be combined with existing CSC methods **5**) LEAD [15]: This method constructs positive and negative samples, and uses a contrastive learning-based training scheme to improve the representation. **6**) CoSPA [19]: This method proposes a variable replication mechanism and uses ResNet to mine the glyph information of characters to enhance the visual representation. **7**) PGBERT [20]: This method proposes a phonology and grapheme augmentation pre-training method.

In addition, we implement PLOME(cfs), which added dynamic confusion set based on PLOME. PLOME(iter) represents the result of inference based on the iteration of PLOME. NCO-Spell(iter) represents the result of inference based on the iteration of NCO-Spell.

V. Results and Analysis

A. Main Results

The main results are shown in Table IV. We have the following observations. **1**) At the sentence level, PLOME(cfs) has significantly improved compared to PLOME, and the accuracy rate has been greatly improved, indicating that the dynamic confusion set allows the model to have more semantic knowledge and the correct rate of corrected sentences is improved. **2**) Compared with the PLOME(cfs), the performance of NCO-Spell has improved in most indicators, and the detection and correction indicators at the sentence level have increased by 1.4% and 1.8% respectively. This shows that the pre-training data constructed by the multi-character masking strategy better fit the real data, which is not constructed by the previous masking strategy. **3**) ECOPO has added an error-driven probability comparison optimization module, RELEASE has integrated multi-channel information of characters, but their performance is still not as good as the performance of NCO-Spell, which shows the effectiveness of the multi-character masking strategy. **4**) Compared with NCO-Spell, the performance of most indicators of NCO-Spell(iter) has been improved, and other indicators have also achieved comparable performance. Compared with PLOME, the performance of most indicators of PLOME(iter) has been improved. Shows the effectiveness of iterative inference. **5**) The performance of PLOME(iter) and NCO-Spell(iter) is similar, because after

TABLE V: The performance of our approach and baseline model PLOME on the multi-typo subset of SIGHAN15. All the experimental results are using our implementation.

Method	Character-level (%)						Sentence-level (%)					
	Detection-level			Correction-level			Detection-level			Correction-level		
	P	R	F	P	R	F	P	R	F	P	R	F
PLOME	97.1	79.2	87.2	97.0	76.8	85.7	63.7	58.5	61.0	61.1	56.1	58.5
PLOME(iter)	97.1	**81.0**	88.3	**97.1**	**78.6**	**86.8**	73.6	65.9	69.5	69.1	61.8	65.2
NCO-Spell	**98.3**	77.5	86.6	96.9	75.1	84.6	67.3	60.2	63.5	62.7	56.1	59.2
NCO-Spell(iter)	97.5	**81.0**	**88.5**	**97.1**	**78.6**	**86.8**	**74.5**	**66.7**	**70.4**	**70.0**	**62.6**	**66.1**

TABLE VI: Ablation experiment results (sentence-level metrics) for different masking distances in multiple character masks.

Masking Distance	Whole Set						Multi-typo Set					
	Detection			Correction			Detection			Correction		
	P	R	F	P	R	F	P	R	F	P	R	F
0	79.3	82.2	80.7	77.4	80.2	78.7	63.7	58.5	61.0	61.1	**56.1**	58.5
1	**81.5**	**82.7**	**82.1**	**79.9**	**81.1**	**80.5**	**67.3**	**60.2**	**63.5**	**62.7**	**56.1**	**59.2**
2	80.0	80.9	80.5	77.9	78.7	78.3	63.4	57.7	60.4	61.6	**56.1**	58.7
3	80.8	80.9	80.8	78.6	78.7	78.7	64.9	58.5	61.5	60.4	54.5	57.3

iterating the algorithm in the inference stage, the result is less affected by noise, and the enhancement effect of the model on the noisy environment is covered to a certain extent.

B. Results for Multi-typo Subset

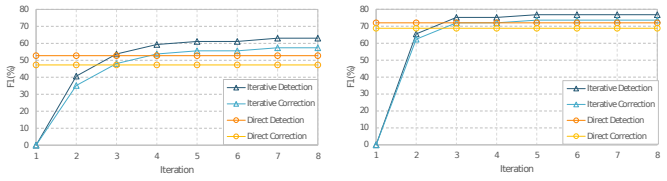
To better illustrate the role of NCO-Spell in multi-typo texts, we test on a test set containing multi-typos, which is a subset of SIGHAN15 and contains 123 test data. Table V presents the results. We observe that after adding the multi-character masking strategy, the sentence-level performance is significantly improved compared to the baseline model PLOME, which demonstrates the effectiveness of the strategy. Compared with NCO-Spell direct inference, after adding iterative inference, the sentence-level detection and correction indicators increased by 6.9% and 6.9% respectively; Compared with PLOME direct inference, after adding iterative inference, the sentence-level detection and correction indicators increased by 8.5% and 6.7% respectively. This verifies the effectiveness and model-agnostic of iterative inference.

C. Effects of Different Masking Distance

We suppose that when the noise distribution is denser, it has a greater impact on character correction. Therefore, when we perform sampling replacement for misspelled positions, the masking distance is set to 1/2/3 respectively. The results are shown in Table VI. We observe that the model achieves the best performance when the masking distance is 1. Therefore, it is used as the final selected hyperparameter. This result also validates our hypothesis in Section V-B that previous CSC models perform poorly on multi-typo texts due to the noise around typos.

D. Results of Iterative Inference for Continuous typos

Since the effect of sentences with multi-typos is mainly affected by noise, the data set is divided into continuous and discontinuous errors according to the noise distribution for experiments. The number of sentences is 59 and 64, respectively. The experimental results are shown in Fig. 4.



(a) Continuous (b) Discontinuous

Fig. 4: Iterative inference experiments with sentences containing continuous or discontinuous errors. Our method consistently achieves better performance.

In sentences containing continuous typos, the performance of the iterative inference method is significantly better than that of the direct inference method. At the third iteration, the performance achieved by the iterative inference method is close to that achieved by the direct inference method and continuously improves performance in constant iterations. Analyzing the reasons for this phenomenon, we argue that in sentences containing continuous typos, continuous typos are greatly affected by neighboring noise, and it is difficult for direct inference methods to fully correct them.

For sentences containing discontinuous typos, although the performance of iterative inference is improving as the number of iterations increases, the improvement is not obvious compared with the performance of the direct inference method. For a sentence only containing discontinuous typos, the adjacent characters of the wrong character are all correct characters. And the adjacent context noise is less, which is less affected. So the direct inference can correct more characters at one time.

At the same time, we can find that sentences containing discontinuous typos have an overall higher score than sentences containing continuous typos, because they are less affected by noise, which also explains that in our multi-character masking strategy, why setting the masking distance to 1.

VI. Conclusions

We propose NCO-Spell for CSC task in noisy contexts. We construct noisy context through a multi-character masking strategy with dynamic confusion sets in the pre-training stage. Furthermore, in order to improve the performance of character correction in noisy environments, we introduce an iterative inference method. Experimental results on a widely used benchmark show that our NCO-Spell outperforms all the compared baseline models. In the future, we will explore the large language models (LLMs) to enhance the performance of our method.

Acknowledgment

The authors would like to thank the editors and anonymous reviewers for their valuable comments on improving the final version of this article. This work was supported by High Performance Computing Platform of BUPT.

REFERENCES

[1] H. Wang, B. Wang, J. Duan, and J. Zhang, "Chinese spelling error detection using a fusion lattice lstm," *Transactions on Asian and Low-Resource Language Information Processing*, pp. 1–11, 2021.

[2] J. Yu and Z. Li, "Chinese spelling error detection and correction based on language model, pronunciation, and shape," in *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 2014, pp. 220–223.

[3] Y.-H. Tseng, L.-H. Lee, L.-P. Chang, and H.-H. Chen, "Introduction to sighan 2015 bake-off for chinese spelling check," in *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, 2015, pp. 32–37.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[5] X. Cheng, W. Xu, K. Chen, *et al.*, "SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 871–881.

[6] S. Liu, T. Yang, T. Yue, F. Zhang, and D. Wang, "Plome: Pre-training with misspelled knowledge for chinese spelling correction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2991–3000.

[7] B. Wang, W. Che, D. Wu, S. Wang, G. Hu, and T. Liu, "Dynamic connected networks for chinese spelling check," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2437–2446.

[8] H. Wu, S. Zhang, Y. Zhang, and H. Zhao, "Rethinking masked language modeling for Chinese spelling correction," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 743–10 756.

[9] D. Wang, Y. Song, J. Li, J. Han, and H. Zhang, "A hybrid approach to automatic corpus generation for chinese spelling check," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2517–2527.

[10] D. Wang, Y. Tay, and L. Zhong, "Confusionset-guided pointer networks for chinese spelling check," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5780–5785.

[11] J.-F. Yeh, S.-F. Li, M.-R. Wu, W.-Y. Chen, and M.-C. Su, "Chinese word spelling correction based on n-gram ranked inverted index list," in *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, 2013, pp. 43–48.

[12] Y. Hong, X. Yu, N. He, N. Liu, and J. Liu, "Faspell: A fast, adaptable, simple, powerful chinese spell checker based on dae-decoder paradigm," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 2019, pp. 160–169.

[13] L. Huang, J. Li, W. Jiang, *et al.*, "Phmospell: Phonological and morphological knowledge guided chinese spelling check," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5958–5967.

[14] Y. Li, Q. Zhou, Y. Li, *et al.*, "The past mistake is the future wisdom: Error-driven contrastive probability optimization for chinese spell checking," in *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, 2022, pp. 3202–3213.

[15] Y. Li, S. Ma, Q. Zhou, *et al.*, "Learning from the dictionary: Heterogeneous knowledge guided fine-tuning for chinese spell checking," in *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 2022, pp. 238–249.

[16] C. L. Liu, M. H. Lai, Y. H. Chuang, and C. Y. Lee, "Visually and phonologically similar characters in incorrect simplified chinese words," in *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, 2010.

[17] S.-H. Wu, C.-L. Liu, and L.-H. Lee, "Chinese spelling check evaluation at sighan bake-off 2013," in *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, 2013, pp. 35–42.

[18] H. Xu, Z. Li, Q. Zhou, *et al.*, "Read, listen, and see: Leveraging multimodal information helps chinese spell checking," in *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, 2021, pp. 716–728.

[19] S. Yang and L. Yu, "Cospa: An improved masked language model with copy mechanism for chinese spelling correction," in *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, 2022, pp. 2225–2234.

[20] L. Bao, X. Chen, J. Ren, Y. Liu, and C. Qi, "Pgbert: Phonology and glyph enhanced pre-training for chinese spelling correction," in *Natural Language Processing and Chinese Computing*, 2022, pp. 16–28.

GVDIE: A Zero-Shot Generative Information Extraction Method for Visual Documents Based on Large Language Models

Siyang Qi[1,2], Fei Wang[1], Hongzhi Sun[1], Yang Ge[1], Bo Xiao[2*]
[1] State Grid Shandong Electric Power Company Dezhou Power Supply Company, Dezhou, China
[2] Beijing University of Posts and Telecommunications, Beijing, China
E-mail: qisiyang@bupt.edu.cn, Wf6102@163.com, 328401563@qq.com, 15005340985@163.com, xiaobo@bupt.edu.cn
[*]Corresponding author

Abstract——Document Information Extraction aims to extract entities and relationships from visually rich documents. Traditional methods require significant annotation and lack generality. In this paper, we propose GVDIE, a Generative Visual Document Information Extraction method that leverages large language models for zero-shot information extraction. This method aligns document-level multimodal features with text layout features, enhancing performance. We introduce a layout augment module for key text block features and transform the document information extraction from sequence labeling task to a generative one. This approach improves flexibility, efficiency, and adaptability in processing diverse, unstructured documents.

I. INTRODUCTION

The task of Document Information Extraction (DIE) [1] represents a crucial application within the realm of artificial intelligence. Unlike traditional information extraction methodologies, DIE focuses on extracting entities and relationships from visually rich texts. This technology is intricately linked to real-world needs, particularly in industrial applications. In practice, especially within industrial supply chains, visually rich documents (VRDs) are the predominant data format. A significant volume of data is stored in unstructured formats, such as images, PDFs, and paper documents, including supply chain procurement contracts, work orders, receipts, and invoices. These documents are replete with visual information and possess substantial practical application value.

Conventional rule-based and machine learning-based visual document information extraction models necessitate considerable human effort and often lack both capability and generality. Common approaches typically involve using part-of-speech and dependency annotations for information extraction tasks. However, due to the limitations in the accuracy of Optical Character Recognition (OCR) tools and the disordered nature of text, the performance of these models is frequently suboptimal. Moreover, the fine-grained annotation of documents requires significant effort, thereby affecting model generalization.

With the advent of deep learning technology, contemporary visual document information extraction primarily follows this technological route: initially, OCR technology is employed to scan visual documents to obtain text and layout information. Subsequently, text, layout, and visual image information are organized and analyzed using large-scale pre-trained deep neural networks. However, this approach is constrained by its reliance on annotated datasets and its lack of generalization capability, posing difficulties in recognizing documents with undefined entity categories or varying layouts. In practical applications, due to the diversity of information formats and types, there are a large number of entity types that are not labeled. We need to extract such entities and discover new relationships. Additionally, this supervised fine-tuning (SFT) model confronts challenges related to limited computing resources and training samples. Consequently, exploring zero-shot document information extraction techniques is essential.

The emergence of large language models (LLMs) offers novel solutions to these challenges. LLMs have demonstrated exceptional capabilities in text understanding and generation, with the potential to generalize across various fields and tasks, effectively extracting unannotated entity information from documents. However, LLMs are designed to manage one-dimensional text sequences and require modifications to process text with significant two-dimensional structures, such as document images. General multimodal models capable of handling image information have not performed well with diverse, text-rich, and complexly formatted document images.

Therefore, effectively applying LLMs to VRDs for information extraction, particularly in the absence of targeted training, and extracting fine-grained visual features remains a significant challenge. Our method addresses the limitations of traditional models by amalgamating the strengths of visually rich document understanding models and LLMs. We propose a zero-shot generative visual document structured information

extraction method based on LLMs. This method integrates document-level multimodal features with text layout features to enhance the model's comprehension and processing capabilities. Additionally, leveraging the robust capabilities of LLMs for generation facilitates in-depth analysis of known data types and the continuous incorporation of new information and knowledge. This approach expands the knowledge graph to meet dynamic business requirements, enabling the model to swiftly adapt to new entity types and application scenarios, thus demonstrating exceptional adaptability and broad application prospects.

The main contributions of this paper can be summarized as follows:

1. We introduce the GVDIE, a zero-shot document information extraction method based LLMs. GVDIE integrates document-level multimodal features with text layout features. By employing feature fusion techniques, it significantly enhances zero-shot information extraction performance for visual documents.
2. We have designed a layout augment module grounded in LLMs. This module focuses on analyzing and enhancing key layout features of text blocks in documents, including size, global position, and relative position. Such in-depth layout analysis is crucial for understanding the structure and content of documents, thereby providing support for more accurate information extraction.
3. We reconceptualize the document information extraction task from traditional sequence labeling to a generative task, leveraging the powerful capabilities of LLMs. This transformation significantly improves the flexibility and efficiency of information processing, allowing the model to perform information extraction without predefining entity categories. Consequently, this enhances the model's performance in zero-shot tasks.

II. Related Work

A. Document information extractor

Document information extraction refers to the technology of extracting semantic entities and their relationships from a large amount of unstructured and visually rich document content. Traditional methods [2] heavily rely on predefined rules, templates, and manual features. With the advancement of deep learning technology, the performance of document information extraction methods has greatly improved. Sage et al. [3] proposed a text content-based method that utilizes a recurrent neural network (RNN) to extract entities from documents; however, the layout information is discarded during the text serialization process. Zhang et al. [4] modeled the document layout and utilized the token path in a directed graph to annotate the document content. Due to the knowledge transfer capabilities demonstrated by pre-trained models, many works now leverage these models. Xu et al. proposed LayoutLM [5], which combines semantic and layout information for model training. LayoutLMv2 [6] further integrates image information during the pre-training stage. However, the aforementioned methods treat the Document Information Extraction (DIE) task as a sequence labeling task, which is constrained by the reading order. In reality, the diverse layouts of documents result in varying reading sequences, thereby increasing the complexity of this task. Although some methods have optimized document reading order recognition, such as XYLayoutLM [7], which generates various appropriate reading orders using an enhanced XY cutting algorithm, and ERNIE Layout [8], which uses reading order prediction as a pre-training objective, DIE remains a challenging task that requires a large amount of labeled data and computing resources. Furthermore, these methods can only classify text into predefined labels and require fine-tuning and adaptation for downstream tasks, which limits their versatility.

B. Large Language Model

Large language models, such as GPT-3 [9], have demonstrated excellent performance across various natural language processing tasks. LLMs have been shown to effectively improve the zero-shot performance of unseen tasks. Research on LLMs for information extraction has primarily focused on the text domain [10], encompassing both general large models and those tailored for specific fields. One of the most straightforward approaches to using LLMs for document information extraction involves directly inputting the OCR results of example samples and the named entity task annotation results into the LLMs, allowing the model to predict the named entities in the test samples. However, this method underperforms as it ignores the positional relationship between text content and is highly sensitive to the selection of example samples. Although general multimodal large models that combine images and text exhibit strong capabilities in multimodal understanding, their performance in real-world document scenarios remains limited due to the lack of targeted training.

III. Method

In this section, we introduce the structure of the GVDIE model. As shown in Fig. 1, the model comprises five components: the Basic Information Extraction Module, the Layout-Augment LLM, the Multimodal Encoder, the Feature Adapter, and the LLM Backbone. Upon the input of a document image into the model, the Basic Information Extraction Module initially acquires the pertinent text and layout data. Subsequently, the Layout-Augment LLM and the Multimodal Encoder independently extract multi-dimensional document features. These features are then mapped into the feature space of the LLMs through the Feature Adapter and,

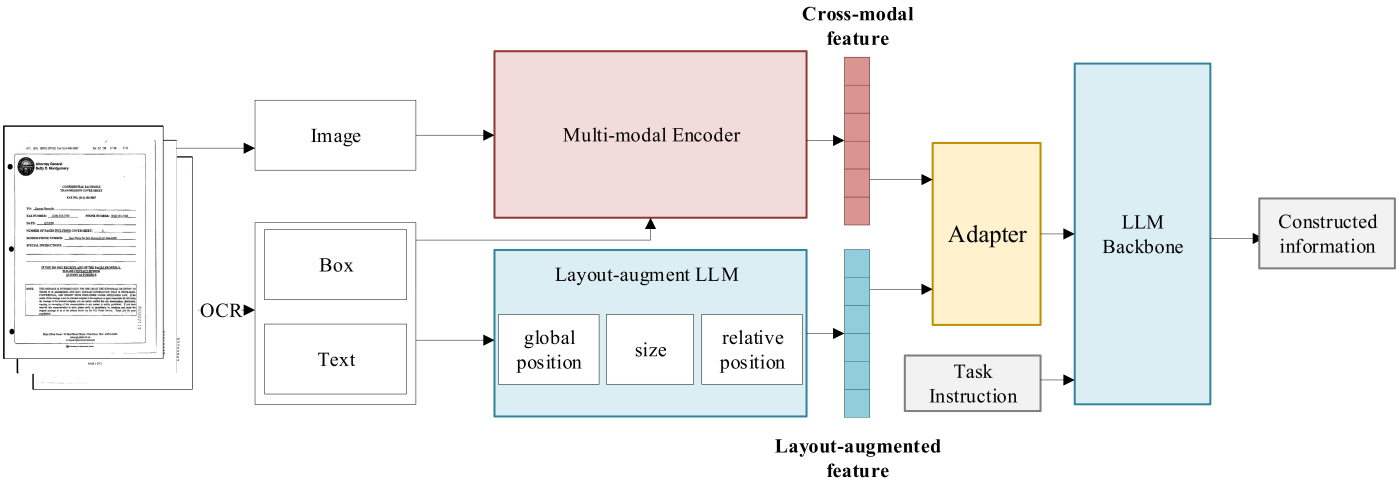


Fig. 1 The model structure of GVDIE

along with instruction embeddings, input into the LLMs to generate the final results.

A. Basic Information Extraction

To obtain the textual content and its corresponding layout information (referred to as "Box") contained within a visual document image, OCR technology is employed to scan the document image. The layout information describes the regions where the text is located within the image. The text regions identified by OCR are represented as rectangular boxes, defined by the coordinates of their corners. Specifically, each box is characterized by the coordinates of its top-left corner (x_{TL}, y_{TL}) and its bottom-right corner (x_{BR}, y_{BR}).

$$Box = (x_{TL}, y_{TL}, x_{BR}, y_{BR}) \quad (1)$$

B. Multimodal Encoder

Directly using text and layout information as input to a large language model (LLM) does not ensure its effective understanding. VRDs inherently align their images, text, and layouts. Extracting image features or text features in isolation fails to fully exploit the relationships between these modalities. Therefore, we employ a multimodal pre-trained model, LayoutLMv3[11], as the encoder module, referred to as the Multimodal Encoder (Mm-Encoder), to combine and align text features, layout features, and visual features. The image is scaled and segmented, and together with the basic information, input into the Mm-Encoder to obtain the multimodal feature embeddings of the document:

$$Fc = Mm - Encoder(Text, Box, resize(Image)) \quad (2)$$

C. Layout-augment LLM

Research has shown that layout features play a critical role in visual documents [12]. However, enabling LLMs to directly and effectively understand layout features extracted from basic OCR information is challenging. This limitation hinders the comprehension abilities of LLMs in zero-shot visual document information extraction tasks. LLMs are trained on vast corpora and exhibit strong in-context learning ability, allowing them to adapt to tasks using instruction information and contextual examples without fine-tuning. Therefore, we designed a layout feature extraction task, as shown in Table I, which enables the large language model to understand the positional relationship of the input text through in-context learning and enhance the layout of the newly input text. Specifically, the text and box information are input into the Layout-Augment LLM (LAM) with frozen parameters. The LLM further interprets and enhances text and layout information by analyzing the size of text blocks, their global positions within the document, and their relative positions to other text blocks. Finally, the encoded layout-enhanced features are output.

$$F_L = LAM(Text, \ Box) \quad (3)$$

TABLE I
Layout-augmented Description

Task:	**Layout feature extraction**
Instruction:You are a layout feature extractor and your task is to conduct layout feature analysis. Now I have a document that I parse with OCR. I will provide you with the OCR results as follows. Please analyze the size, global position, and relationship of these texts. Here are some examples **---Examples---** input:{{'words': 'AantieAnne's', 'boxes': {'x': [0.2, 1.0], 'y': [0.23, 0.34]}}…… output: "'AantieAnne's' has a box of {{'x': [0.2, 1.0], 'y': [0.23, 0.34]}} is located on the top middle of this document.And its surrounding is empty. ""SUB TOTAL' has a box of {{'x': [0.04, 0.26], 'y': [0.72, 0.75]}} is located on the middle left of this document.And it is at the same horizontal position as '17,000'.…… Now, please analyze the global position of the these words. And analyze the position relationship between the words and its surrounding.the inuput is： **--input--** Input:{{}} **--output--** The output is:	

Research [13] has indicated that the outputs generated by LLMs are more closely aligned with the data distribution observed during reasoning than with the training data itself. Consequently, this step enhances the LLM's ability to comprehend and reason about layout information, thereby improving the final information extraction result.

D. Feature Adapter

In our study, the Adapter module serves as a critical intermediary, bridging and transforming features from various modalities to meet the requirements of the LLM backbone. The primary objective of this module is to facilitate the seamless integration and processing of features extracted by the multimodal encoder and the layout-augment LLM. The adapter consists of three fully connected layers, each followed by a ReLU activation function. The initial layer transforms the input features into a hidden representation. This hidden representation is further refined by a second layer of the same size, enhancing the feature transformation process. Finally, the third layer projects the refined features to match the output feature space dimension required by the LLM backbone. The final output is cast to half-precision to optimize computational efficiency. Ultimately, these appropriately transformed and integrated features are input into the pre-trained large language model for subsequent task-specific training and information extraction.

$$H_F = \mathrm{Concat}(\mathrm{Adapter}(F_C, D_{LLM}), F_L) \quad (4)$$

Subsequently, we froze the parameters of all pre-trained models and updated the parameters of the adapter module through two distinct training tasks. These tasks aim to equip the GVDIE with the capability to comprehend the layout information of documents. The training tasks are as follows:

Text Region Box Prediction: The objective of this task is to predict the bounding box of a specific text region given the textual content. After inputting multimodal information and undergoing feature transformation, the GVDIE is trained to accurately predict the boundaries of the text region.

Coordinate Mapping Text Restoration: In this task, the goal is to restore the textual content within a given text region's bounding box. The GVDIE leverages multimodal features to reconstruct the text contained within the specified bounding box.

These training tasks are designed to refine the adapter module's ability to align and process diverse features, thereby enhancing the the proficiency of GVDIE in understanding and interpreting document layouts.

E. LLM Backbone

LLMs, pre-trained on extensive textual datasets, have acquired a profound understanding of linguistic patterns and knowledge. These models have demonstrated exceptional performance across various natural language processing tasks. Consequently, the LLM module receives processed features from the Adapter module, which integrates textual, visual, and potentially layout information. Leveraging these combined features, the LLM performs deep textual analysis, comprehending the document's deeper meaning and contextual relationships. Ultimately, based on the model's interpretative capabilities, it generates structured text, facilitating the extraction of document information.

IV. Experiments

A. Experiment Setup

We conducted experiments on a machine equipped with an INTEL XEON 8358P, 2.6 GHz, 64-core processor, and an NV HGX A800 GPU. During the Basic Information Extraction phase, we employed the PaddleOCR model [14] to extract text and bounding box information. For the foundational LLM in both the Layout-augment and Document Information Extraction (DIE) phases, we utilized the LLaMA2-7B-Chat model [15]. Developed by Meta AI, LLaMA2 is a cutting-edge artificial intelligence language model that leverages an efficient Transformer architecture. It offers excellent processing speed and memory optimization and has demonstrated superior performance in numerous natural language processing tasks. For the multimodal encoder, we selected the LayoutLMv3-base model. The adapter module was trained using the DocLocal4K [16] dataset, employing the AdamW optimizer. The training process used a learning rate of 1e-5, a batch size of 32, and a warmup ratio of 0.048. Additionally, we fine-tuned the LLM backbone and the Adapter module on the DocReason25K dataset.

B. Dataset

We mainly conduct experiments on two datasets: SROIE [17] and CORD [18]. The data statistics of the dataset are shown in Table II.

SROIE. This dataset consists of scanned receipt images and annotations for the competition on Scanned Receipts OCR and Key Information Extraction (SROIE). The test set comprises 347 receipt images, annotated with four entity types: company, date, address, and total.

CORD. This dataset contains thousands of Indonesian receipts, including images and box/text annotations for OCR, as well as multi-level semantic labels for parsing. Its test set consists of 100 receipt images, annotated with 30 entity types, including tax amount, total price, etc.

Table II
Statistic of SROIE/CORD Datasets

Dataset	Types	#Images
SROIE	Receipts	Train 526, Test 347
CORD	Receipts	Train 800, Val 100, Test 100

C. Main Result

We conducted evaluations on widely used document understanding benchmarks in zero-shot scenarios, focusing on large language models (LLMs) and multimodal large language models (MLLMs). The models evaluated include

Qwen-7B-Chat [19], InternLM-7B-Chat [20], and LLaVA-1.5-7B [21]. To prompt the LLMs and MLLMs for the DIE task, we converted the data format of the test datasets. Specifically, the entity annotations in the datasets were transformed into question-answer pairs. The models were assessed by querying which entities are linked to the annotated entities, using accuracy as the evaluation metric. As shown in Table III, the zero-shot document information extraction performance of GVDIE and the baseline is evaluated. Compared to the baseline, GVDIE demonstrates significant improvements across all datasets, confirming the effectiveness of our method. Furthermore, the performance of the multimodal large model LLaVA on the dataset is comparatively inferior to that of the plain text large model. This discrepancy could be attributed to the challenge of comprehending layout characteristics when processing pure image inputs, making them less advantageous than text obtained through OCR.

TABLE III
MAIN RESULT

Model	Modality	CORD	SROIE
Qwen-7B-Chat[19]	Text	34.15	36.53
InternLM-7B-Chat[20]	Text	60.05	40.78
LLaVA-1.5-7B[21]	Image	18.06	3.83
GVDIE(ours)	Layout+Image	67.07	62.41

D. Ablation Study

As demonstrated in Table IV, an ablation study was conducted to better verify the effectiveness of the layout-augmented module for GVDIE. We removed the layout enhancements from GVDIE and validated the model on the CORD and SROIE datasets. The results indicated a decrease in accuracy following the removal of layout enhancements, underscoring the importance of layout features in document understanding tasks, particularly in document information extraction.

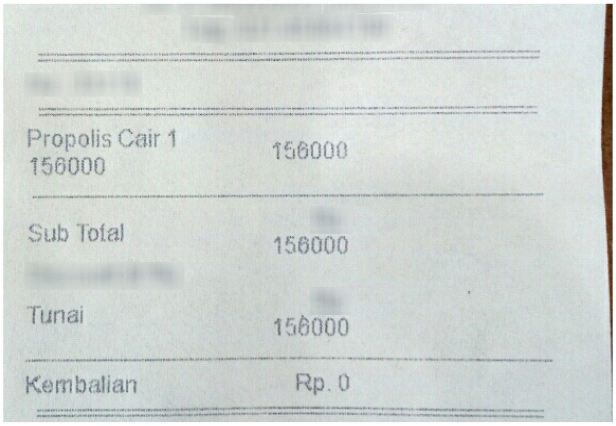
TABLE IV RESULT OF ABLATION STUDY

Method	CORD	SROIE
GVDIE	67.07	62.41
-Layout-augmented LLM	53.99	49.03

E. Case Study

In this section, we present a case study of the layout-augment module. Figure 2 displays an image of a receipt and its OCR results. The detailed process is presented in Table V. Without layout augmentation, the model incorrectly answers because it only interprets the coordinates of "Sub Total" without understanding the image layout. By employing the layout-augment module, the relevant information that "'156000' is at the same horizontal position as 'Sub Total'" is extracted, aiding in the analysis of the Sub Total value. Therefore, the answer after layout augmentation is correct. Without layout enhancement, the output would be incorrect as the model only considers the coordinates of 'Sub Total' without accounting for the image layout. With the implementation of the layout-augment module, relevant information such as "'156000' being at the same horizontal position as 'Sub Total'" is extracted, facilitating the analysis of the Sub Total value. Consequently, after applying layout enhancement, correct answers are obtained.

Image



OCR Results:

{'words': 'Propolis Cair 1', 'boxes': [26, 372, 189, 397]},
{'words': '156000', 'boxes': [292, 385, 379, 411]},
{'words': '156000', 'boxes': [27, 403, 114, 429]},
{'words': 'Sub Toial', 'boxes': [26, 477, 134, 503]},
{'words': '156000', 'boxes': [294, 491, 379, 517]},
{'words': 'Tunai', 'boxes': [25, 568, 91, 595]},
{'words': '156000', 'boxes': [293, 579, 381, 609]},
{'words': 'Kembalian', 'boxes': [23, 647, 147, 672]},
{'words': 'Rp.0', 'boxes': [315, 643, 384, 674]}

Fig.2 A case of receipt information extraction

TABLE V
CASE STUDY

Question：What is the 'Sub Tofal' in the given document?	
Original answer	Based on the OCR results provided, the word "Sub Total" appears in the document at position [26, 477, 134, 503]. Therefore, the answer to your question is: "Sub Tofal".
Layout augment	' Propolis Cair 1' has a box of {'x': [0.07, 0.49], 'y': [0.55, 0.59]} is located on the top left of this document. '156000' has a box of {'x': [0.76, 0.99], 'y': [0.57, 0.61]} is located on the middle right of this document. And it is at **the same horizontal position as**` Sub Total `. 'Sub Total' has a box of {'x': [0.07, 0.35], 'y': [0.71, 0.75]} is located on the bottom left of this document. And it is surrounded by '156000' on the top and 'Tunai' on the bottom. ……
New answer	Based on the OCR results provided, here are the answers to your questions: What is the value of 'Sub Total' in the given document? The value of 'Sub Toial' is \"15600\"

V. CONCLUSIONS

We propose an effective method termed GVDIE for the zero-shot document information extraction task. This approach transforms the visual document information extraction process from a traditional sequence labeling task into a sequence-to-sequence generative task, implemented using LLMs. Our method leverages the image and text layout of visual documents by employing a multimodal encoder and layout augmentation techniques to achieve zero-shot document information extraction. This transformation allows us to fully exploit the text understanding and generation capabilities of large models, facilitating information extraction without the need for predefined entity categories, thus enhancing the flexibility and efficiency of information processing. Experimental results demonstrate that our method consistently outperforms the baseline across all datasets, highlighting the effectiveness of the approach and emphasizing the critical role of layout information in visual document information extraction.

ACKNOWLEDGE

This work was financially supported by State Grid Technology Project (5700-202416236A-1-1-ZN) “Research on active semantic discovery technology based on SG-CIM and its application in power grid equipment supply chain optimization”.

REFERENCES

[1] Lei Cui, undefined., et al, "Document AI: Benchmarks, Models and Applications," 2021.

[2] Schuster, D., et al, "Intellix – End-User Trained Information Extraction for Document Archiving," in 2013 12th International Conference on Document Analysis and Recognition, 2013, pp. 101-105.

[3] Sage, C., et al, "Recurrent neural network approach for table field extraction in business documents," in 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1308–1313.

[4] Zhang, C., et al, "Reading Order Matters: Information Extraction from Visually-rich Documents by Token Path Prediction," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 13716–13730.

[5] Xu, Y., et al, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020,pp. 1192-1200.

[6] Xu, Y., et al, "LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 2579–2591.

[7] Gu, Z., et al, "XYLayoutLM: Towards Layout-Aware Multimodal Networks For Visually-Rich Document Understanding," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4573-4582.

[8] Peng, Q., et al, "ERNIE-Layout: Layout Knowledge Enhanced Pre-training for Visually-rich Document Understanding," in Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 3744–3756.

[9] Brown, T., et al, "Language models are few-shot learners," in Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, pp. 1877–1901.

[10] I. Keraghel, S. Morbieu, M. Nadif. "A survey on recent advances in named entity recognition," in arXiv preprint arXiv:2401.10825, 2024.

[11] Huang, Y., et al, "Layoutlmv3: Pre-training for document ai with unified text and image masking," in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4083–4091.

[12] Li, P., et al, "Selfdoc: Self-supervised document representation learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5652 – 5660.

[13] Cheng, X., et al. "Lift yourself up: Retrieval-augmented text generation with self-memory," in Advances in Neural Information Processing Systems, vol. 36, 2024.

[14] Du, Y., et al. "Pp-ocr: A practical ultra lightweight ocr system," in arXiv preprint arXiv:2009.09941, 2020.

[15] Touvron, H., et al. "Llama 2: Open foundation and fine-tuned chat models," in arXiv preprint arXiv:2307.09288, 2023.

[16] Hu, A., et al. "mplug-docowl 1.5: Unified structure learning for ocr-free document understanding," in arXiv preprint arXiv:2403.12895, 2024.

[17] Huang, Z., et al, "Icdar2019 competition on scanned receipt ocr and information extraction," in 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1516–1520.

[18] Park, S., et al, "CORD: a consolidated receipt dataset for post-OCR parsing," in Workshop on Document Intelligence at NeurIPS 2019, 2019.

[19] Bai, J., et al. "Qwen technical report," in arXiv preprint arXiv:2309.16609, 2023.

[20] I. Team, "Internlm: A multilingual language model with progressively enhanced capabilities," 2023.

[21] Liu, H., et al. "Visual instruction tuning," in Advances in neural information processing systems, vol. 36, 2024.

META: Text Detoxification by leveraging METAmorphic Relations and Deep Learning Methods

Alika Choo*, Arghya Pal†, Sailaja Rajanala‡ and Arkendu Sen§
Monash University Malaysia, Malaysia
E-mail: {alika.choo*, arghya.pal†, sailaja.rajanala‡, arkendu.sen§}@monash.edu

***Abstract*—In the world of online interactions, social communities face a significant challenge: the spread of offensive content and hate speech through toxic languages. Such issues led to growing research on text detoxification systems that can automatically rewrite toxic content. A systematic evaluation is required to ensure these systems produce high-quality detoxified text that modifies the original text to be non-toxic while preserving its content. However, this often relies on large amounts of labelled data and human judgement, which may not always be feasible. This limitation is typically known as the *oracle problem*. Metamorphic testing (MT) has conventionally been used to solve the oracle problem by deriving metamorphic relations (MRs) to test a program's functionality. A new MT approach focused on data validation showed that MRs incorporated with tools can be used to identify defects in machine translation services. This paper draws inspiration from this new MT perspective by presenting four metamorphic relations incorporated with tools to evaluate style transfer accuracy, content preservation, fluency, and a joint of these three. Our proposed approach effectively identifies defective behaviour in state-of-the-art text detoxification systems.**

I. Introduction

Warning: *This paper contains offensive language that commonly appears in hate speeches. Hate speech text present in this paper does not represent the views of the authors!*

♠ **Divisive hate speech and ideologies on a global scale** Social media has been serving as an outlet for people to display hate speech and harmful and offensive behaviour online, facilitated by user anonymity and lack of regulations on social media platforms. Such toxic content often contributes to cyberbullying – using technology to send hostile messages intended to inflict harm to a target individual or group – which has been postulated to be the cause of recent rising suicide rates in adolescents. Furthermore, the situation is so severe that the United Nations resolution on 18 June proclaimed the date to be remembered as the International Day for Countering Hate Speech. This paper, along with other sister works [2], [3], recognises the need to develop an AI-driven text detoxification method.

♠ **AI-driven Deep Learning based Hatetext Detoxification** This prevalence of online toxicity and the potential harm that it can bring to communities has prompted significant research in automatic detection systems for toxic speech and text detoxification systems [2][3]. Text detoxification is the autonomous rewriting of toxic content. It is understood as a style transfer task where the original text is rewritten to change its style (i.e. non-toxic) while preserving its content. Text detoxification systems are an encouraging approach for reducing harmful online behaviour and neutralising emotional comments [4].

♠ **Challenges to Find Correct Pipeline for AI-driven Detoxification** The evaluation of the degree of detoxification, i.e. how many words or sentences have been detoxified, mostly depends on metrics like the Human Turing Test [4] or word embedding-based methods [4], thus making it time-consuming, costly, and subjective [5]. To automate this testing process for text detoxification systems, parallel datasets such as ParaDetox [6] and APPDIA [7] have recently served as benchmark detoxification datasets. Nevertheless, referring to a single ground truth may not be ideal given the ever-growing complexity of online comments [8]. Thus, relating those studies to famous ***oracle problem*** [9]; An oracle (i.e. the evaluation) is a procedure determining whether a program (i.e. the AI-driven Deep Learning model) has produced the correct output (i.e. the detoxified text) [9]. The context of toxic language detoxification arises from the subjective nature of what constitutes an appropriate "detoxification" for a given text.

♠ **Metamorphic Testing: An Introduction** Metamorphic testing (MT) has conventionally addressed the oracle problem for software through metamorphic relations (MRs). MRs are relations we expect to hold over multiple inputs and their expected outputs [9] [10]. These relationships are more general compared to having single-ground truths, yet they set an expectation for what a valid output should be. A violation of an MR highlights an erroneous behaviour in the program. Conventional MT helps software test coverage. It relies on source test cases (a selected set of program inputs) and transformations [9] to build metamorphic test cases that share similar specifications with the source test cases. Finally, MT verifies the source and metamorphic test cases with their outputs against an appropriate MR [10]. However, conventional MT is not without its practical challenges. It often requires expert knowledge and the time-consuming generation and validation of metamorphic test cases [11]. Furthermore, with advancements in large language models (LLMs) like BERT,

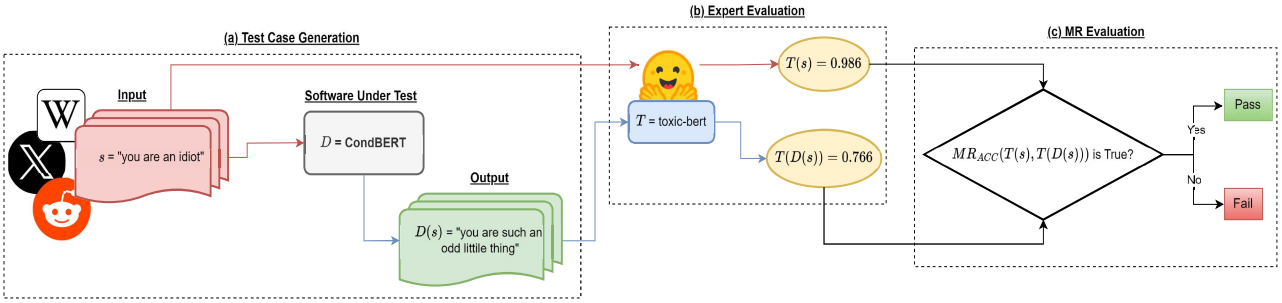


Fig. 1. MR for Data Validation approach [1] adapted for the text detoxification domain to evaluate style transfer accuracy. Our methodology, as discussed in Section IV has three distinct parts: (a) source inputs are passed to software under test (SUT) to generate detoxified outputs (b) experts compute a score for the source input and outputs individually (c) MR uses expert scores to determine whether the detoxified output is valid detoxification of the source input

the oracle problem is becoming less pronounced as these tools approach human-level judgment [3][4][6]. Hence, this leads us to the research question we aim to tackle for this paper *How can we incorporate such tools to act as "experts" for our testing processes to identify defects?*

♠ **Rethinking Metamorphic Testing for AI-driven Text Detoxification** Yan et al. [1] proposes a promising perspective focused on data validation, suggesting that MR violations might stem from input data quality rather than program faults. They demonstrated that this approach was helpful in automatically identifying poorly translated text messages using a sentiment analysis tool.

♠ **Our Contributions** In our paper, we emphasise less on data validation and focus back on identifying defects in AI-driven models under test (SUT), a concept not involved in the data validation MT approach [1]. We adapt their approach for the text detoxification domain and show the effectiveness of four MRs while incorporating state-of-the-art tools. The objectives of our study are to:

- Propose four MRs based on the Joint metric score proposed by Krishna et al. [12]:
  1) ACC - which classifies the level of non-toxicity
  2) SIM - which calculates the similarity between the original and translated text
  3) FL - which assesses fluency
  4) JOINT - combines ACC, SIM, and FL
- Suggest tools for evaluating detoxification models against toxicity removal, content preservation, and fluency

By adopting this approach, we can systematically test text detoxification systems for defective behaviour. Furthermore, an important step is to analyse MR violations which highlight cases of such model defective behaviour. This can include poor handling of toxicity removal, content preservation, or fluency during detoxification. This step may narrow the search for erroneous cases significantly, thus enabling practitioners to focus more on model defective behaviour rather than where defects occur.

II. Text Detoxification Models

Text detoxification models can be categorised as supervised or unsupervised. To show that our methodology applies to different text detoxification systems, we exemplify our approach on the following three state-of-the-art text detoxification models that exhibit unique characteristics:

- CondBERT [4]: An unsupervised method where toxic words in a sentence are masked, and the candidates are reranked based on their non-toxicity scores.
- ParaGedi [4]: An unsupervised method where the detoxification task is viewed as a paraphrasing task, but toxicity scores of the next token prediction are also taken into account during the generation step.
- ParaDetox-BART [6] : A BART (base) model trained on the parallel detoxification dataset ParaDetox.

These models have benchmarked past research works [3][13]. If we show that our metamorphic relations are effective on these state-of-the-art models, it may apply to similar architectures.

III. Dataset

We selected various datasets to test our three models and validate the effectiveness of our proposed metamorphic relations. We chose these datasets according to the criteria:

- Language - Limited to English datasets
- Release Date - Not older than ten years
- Distribution - Most comments are toxic
- Test Dataset Size - Allows sampling of 10,000 sentences
- Toxicity Rating - Scores can be deduced to 0-1 toxicity rating

From these criteria, we selected the following datasets:

- Jigsaw 2018[1] - An annotated dataset of comments from Wikipedia's talk page edits that includes scores measuring toxicity and several subtypes of toxicity.
- WikiDetox 2017[2] - A human-annotated dataset sourced from English Wikipedia of over 100,000 discussion comments.

To prepare the test dataset from Jigsaw, we took the same test dataset used by Dale et al. [4]. For WikiDetox, we preprocessed the dataset (removing irrelevant tokens, removing

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
[2] https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Toxicity/4563973/2

duplicate comments) then created a test dataset by first standardising the sentences to 200 characters or fewer (including whitespaces), then taking 10,000 of the most toxic sentences based on our chosen pre-trained classifier that achieved an AUC score of 0.9864[3]. Finally, we will have two test datasets that we can use to evaluate each model.

Most of each test dataset consisted of content with considerable levels of toxicity, validated by the toxicity distributions we analysed for Jigsaw and WikiDetox test datasets, which were right-skewed. Most of its sentences were highly toxic (> 0.8).

IV. Methodology

This paper adapts the MT approach presented by Yan et al. [1] for testing text detoxification models. Our approach focuses less on validating given datasets and more on identifying defects in text detoxification models. Figure 1 illustrates our approach in evaluating a detoxified output for style transfer accuracy. The process looks similar when evaluating content preservation and fluency. In the following sections, we discuss the distinct phases of our MT methodology.

A. Test Case Generation

This step generates outputs from a given text detoxification model D that, when paired together with a source input, represents a *test case* as shown in Figure 1(a). D can be referred to as SUT in this context, although Yan et al. mention SUT is not present in their data validation methodology [1]. Nonetheless, since D shares the closest meaning to SUT in this context, we labelled D under SUT to align with conventional MT terminology. Each test case is then passed onto the next stage of evaluation.

B. Expert Evaluation

This phase differentiates our methodology from conventional MT, as depicted in Figure 1(b). The Expert Evaluation step simulates the manual process an expert undertakes to validate input quality, except in this approach, this step is autonomous using an appropriate tool. In this paper, we introduce the concept of 'Experts', whose role is to evaluate the quality of its input against a particular metric. An Expert in this context refers to a toxicity classifier, a similarity scorer, or a fluency scorer. Each of these Experts returns scores that are passed to the next step for evaluation against an appropriate MR.

C. MR Evaluation

Finally, our test cases are evaluated against one MR at a time. Test cases are considered to have passed if an MR holds and fails if it does not. For example, in Figure 1, given that MR_{ACC} requires $T(D(s))$ to be lower than $T(s)$, the test case passed. In later sections, we describe other MRs to evaluate content preservation, fluency, and joint scores.

[3]https://www.https://huggingface.co/unitary/toxic-bert

V. Metamorphic Relations For Data Validation

In this section, we describe our proposed MRs and mention the tools we have chosen to use for our experiments. The main goal is to select a tool that can perform closest to human judgement concerning one of the three style transfer metrics. Hence, the violations highlighted by tools must precisely contain erroneous outputs. For the following MRs, let L_1 and L_2 be languages prone to toxic content and devoid of toxic content, respectively. Let s_1 and s_2 be texts written in L_1 and L_2, respectively.

A. Style Transfer Accuracy

The goal of MR_{ACC} is to evaluate the disparity in toxicity ratings between two comments:

$$MR_{ACC} : T(s_2) - T(s_1) < 0 \tag{1}$$

where T is an Expert in toxic comment classification that returns a score between 0-1 for both languages. Generally, for s_2 to be a valid detoxification of s_1, the toxicity rating of s_2, $T(s_2)$, should be less than $T(s_1)$. The detoxified text should not be more toxic than its original text. **Choosing Classifier T**: We use the BERT classifier from Section III and a roBERTa classifier, trained on Jigsaw's Wikipedia comments (2018) and Civil comments (2019) datasets. The roBERTa classifier achieves an AUC of 0.9374.

Our analysis showed no significant differences between the classifiers: both highlighted over 70% of the same CondBert violations and 100% of the ParaDetox-BART violations. However, roBERTa missed more ParaGedi violations. Given the negligible differences and BERT's slightly better performance, we choose the BERT model for our experiments.

B. Content Preservation

The goal of MR_{SIM} is to evaluate the similarity between two comments:

$$MR_{SIM} : S(s_1, s_2) \geq \alpha \tag{2}$$

where S is an Expert in similarity ratings that returns a score between 0-1 for both languages. For s_2 to a valid detoxification of s_1, the similarity score given for s_1 and s_2 given by S should be reasonable (i.e. $\alpha > 0.5$). Although preserving too much content risks the persistence of toxic content, simply removing toxic content inappropriately hinders content preservation. **Choosing Similarity Metric:** We compared three state-of-the-art metrics for MR_{SIM} at different α values in 0.1 increments: BLEU (bilingual evaluation understudy), METEOR (metric for evaluation of translation with explicit ordering), and LaBSE (language agnostic BERT sentence encoder). Our analysis shows that BLEU consistently underestimates the similarity of detoxified texts to the original. METEOR has a similar issue but to a lesser degree. These issues align with common pitfalls of n-gram-based metrics, which penalise translations for surface differences [14]. LaBSE performed closest to human judgment in assessing semantic similarity across all models.

C. Fluency

The goal of MR_{FL} is to evaluate the fluency difference between two comments:

$$MR_{FL} : F(s_2) - F(s_1) \geq -\beta \tag{3}$$

where F is an Expert in fluency ratings that returns a score between 0-1 for both languages. For s_2 to a valid detoxification of s_1, the fluency of s_2 given by F should not be worse than that of s_1 given by F. Substituting away toxic content at the expense of worse grammar may be suitable at times. However, output text with abysmal grammar may seem fake and unpleasant to read.

Choosing Fluency Expert: We compared BERT and roBERTa classifiers, both fine-tuned on the CoLA corpus, for judging grammar fluency. Analysis across different β thresholds revealed that roBERTa generally identified more violations and detected more extreme grammar issues in the detoxified text at $\beta = -0.8$. Therefore, we selected the roBERTa classifier for our experiments.

D. Joint

This MR combines the conditions from MR_{ACC}, MR_{SIM}, and MR_{FL}. Each comment is evaluated against

$$MR_{JOINT} : A \wedge S \wedge F, \tag{4}$$

where the statements A, S, and F are statements where the detoxified output complies with MR_{ACC}, MR_{SIM}, and MR_{FL} respectively.

VI. Experiment Design

This section aims to evaluate each MR from Section V in identifying defects in various models. Moreover, we consider the following metamorphic relations to be effective *if the violations that are highlighted precisely contain an erroneous detoxified output*.

A. Ablation Study

We determined reasonable thresholds for MR_{SIM} and MR_{FL} by experimenting with a range of α and β values to balance the detection of violations. We considered that if α is set too low, non-erroneous detoxifications might be flagged as violations, while a too-high α may risk missing some erroneous detoxifications. The same is true for β. We experimented with a scoring system where we selected the α/β values for a given model and dataset that produced a number of violations closest to the average number of violations across all α/β values tested. This approach helped us choose an α/β value roughly representing the average violations among different values.

B. Metamorphic Relation Design

The sections below define the conditions and Experts used for each metamorphic relation from Section V. To verify the effectiveness of each MR in highlighting defects, we generated sets of MR violations across different models and datasets, then analysed them in Section VII.

1) MR_{ACC}: Using the BERT-based classifier from Section V, each comment is evaluated against the condition defined in Section V-A. Moreover, if $T(s_2) - T(s_1)$ is zero or greater, s_2 is more toxic than s_1, and a violation has occurred.

2) MR_{SIM}: From Section VI-A, we found that the following condition highlighted a reasonable number of violations for poor content preservation across all test datasets and models. We use the LaBSE metric from Section V with an α value of 0.6. Thus, if $S(s_1, s_2)$ is less than 0.6, s_2 has not sufficiently preserved its original content from s_1, and a violation has occurred.

3) MR_{FL}: From Section VI-A, the following condition highlighted a reasonable number of violations for noticeable degradation of grammar across our test datasets and models. We use the roBERTa classifier from Section V and a β of 0.4. Thus, if $F(s_2) - F(s_1)$ is less than -0.4, the grammar of s_2 has degraded substantially after the detoxification of s_1, and a violation has occurred.

4) MR_{JOINT}: MR_{JOINT} follows the condition defined in (4). Thus, any ACC, SIM, or FL violation would also be a violation in MR_{JOINT}.

VII. Experimental Analysis

TABLE I
Metamorphic Relation Violations

Model	MR	Jigsaw	WikiDetox	All
CondBert	MR_{ACC}	106	2142	2248
	MR_{SIM}	571	368	939
	MR_{FL}	869	581	1450
	MR_{JOINT}	1467	2990	4457
ParaDetox-BART	MR_{ACC}	1479	1367	2846
	MR_{SIM}	969	2404	3373
	MR_{FL}	710	735	1445
	MR_{JOINT}	3046	4178	7224
ParaGedi	MR_{ACC}	82	280	362
	MR_{SIM}	1504	1839	3343
	MR_{FL}	1304	862	2166
	MR_{JOINT}	2690	2518	5208

This section demonstrates how the MRs from Section VI-B can be used on several models to highlight possible defects. Table I shows the frequency of MR violations across the three models on the two test datasets. For each violation (except of type MR_{JOINT}), each author were guided by three questions. (1) How many words were replaced from the original text? (2) If any, what were words replaced for? (3) Does this justify the type of violation? Similar errors were then grouped and an appropriate label was chosen. Table II shows the types of errors with examples[4]. For brevity, we label CondBert, ParaDetox-BART, and ParaGedi as CB, PD, and PG respectively.

A. MR_{ACC} Analysis

MR_{ACC} was used to evaluate the change in toxicity level in the detoxified comment. In general, most MR_{ACC} violations stemmed from the lack of replacement for toxic words, as exhibited by the minimal difference in toxicity between the

[4]For more examples, visit https://github.com/alikachoo/mt_research_2024

TABLE II
MODEL ERRORS AND TEST CASE EXAMPLES

ID	Error	CB	PD	PG	Example
1	replacement of toxic words for words with similar or greater level of toxicity	✓	✓	✓	"you are an embarrassment to the rest of us canadians" → "you are an enemy even to the rest of us canadians"
2	replacement of toxic words with different meaning	✓		✓	"dirty mind" → "well-read mind", "an idiot" → "an excellent writer", "F*cking nerd" → "Fick 'Nerd"
3	replacement of sentence meaning	✓		✓	":F*ck off you idiot" → "if you're like me, do it."
4	replacement of toxic words degrading overall grammar	✓	✓	✓	"you must really suck at life" → "you must really bad at life" "your country is sh*t poor and ugly" → "Your country's poor and unhappily"
5	avoiding the toxic or difficult part by generalising		✓	✓	"You're an idiot. Nuff said." → "that's what I meant.", "and apparently, youre a pile of horsesh*te" → "You are not good"
6	lack of replacements	✓	✓	✓	"GOOD RIDDANCE. GET LOST." → "good rid . get lost ."

original and the output text. Upon inspection of the violations that MR_{ACC} highlights, most of the detoxified outputs were erroneous. Additionally, we found that MR_{ACC} helps set expectations for each model's ability to remove toxicity content.

CondBert: MR violations were primarily due to the failure to replace toxic words (error 6). In the Jigsaw and WikiDetox datasets, 71 out of 106 and 1718 out of 2142 violations had zero toxicity difference, respectively. The model primarily made stylistic changes, such as modifying capitalisations and adding/removing spacing and punctuation, which did not sufficiently detoxify the content. The remaining errors were simply due to replacements of words (error 1) that were just as toxic or more toxic.

ParaDetox-BART: Error 6 was also a common mistake for ParaDetox-BART which only made stylistic changes to its text, thus resulting in minimal difference to the toxicity in the detoxified text. In other cases, ParaDetox-BART may also remove fragments of toxic content without replacing its entirety (e.g., "you die you die and you go to hell" → "you die you go to hell").

ParaGedi: Similar to the previous two models, ParaGedi failed to remove abusive language and hateful speech (racism, sexism, misogyny) in the output. Most text outputs showed the same types of stylistic changes as mentioned earlier and minor paraphrasing changes such as use of contractions (e.g., "are not" → "aren't").

B. MR_{SIM} Analysis

MR_{SIM} aims to quantify content preservation from the original text to the detoxified output. Most violations arose from replacement errors where the modified text diverged significantly from the original meaning. We inspected the violations that MR_{SIM} highlighted and found that most precisely failed to preserve content sufficiently in the output. Nonetheless, these models prioritise removing toxicity at the expense of worse content preservation.

CondBert: With regards to error 2, CondBert erroneously replaced toxic content for antonyms (e.g., "idiot" → "an even better friend"). As for error 3, CondBert may rewrite fragments of a sentence, changing the overall sentence meaning. Another case relates to how CondBert replaces toxic text with different punctuations when encountering sentences that are largely infested with abusive language. However, these behaviours may sometimes be appropriate, given that they often remove offensive words from the original text.

ParaDetox-BART: Most replacement errors were caused by avoiding toxic parts by generalisation (error 5). This behaviour often resulted in multiple distinct toxic texts being transformed into the same detoxified output, such as "go to hell with that idiot" and "so f*ck the f*ck off" both being reduced to "Go away.". While such generalisations can sometimes suffice in removing toxic content, they tend to strip away the specific target and content, undermining the comprehension of the original text's meaning.

ParaGedi: MR_{SIM} violations highlighted errors that overlap with a quantitative analysis Dale et al. [4] conducted (errors 2, 3, and 5). This can be attributed to ParaGedi's tendency to paraphrase the whole sentence, add punctuation or spacing, modify capitalisations, and replace fragments of toxic content.

C. MR_{FL} Analysis

MR_{FL} assesses whether a degradation of fluency has occurred during the detoxification process. MR_{FL} violations were predominantly due to errors 4. Upon analysing the violations that MR_{FL} highlights, we found that most outputs from the test cases precisely showed noticeable fluency degradation. Furthermore, these models may prioritise removing toxic content despite worsening fluency.

CondBert: MR_{FL} violations in this case were primarily due to replacement errors (error 4) which commonly involved inappropriately replacing proper nouns with abstract nouns (e.g., lesbians → lesbian, unfairly → unfair). Similar to some CondBert MR_{SIM} violations, this model may replace toxic text with punctuation and degrade fluency in the output.

ParaDetox-BART: Most MR_{FL} violations here stemmed from error 4, which typically removed toxic content by degrading grammar. This commonly involved ParaDetox-BART neglecting to put articles (e.g.,'a', 'an', 'the') in the output text to make it more fluent, even though the source input had high fluency (e.g., "This guy is jew" and "You are robot").

ParaGedi: ParaGedi exhibited the highest number of MR_{FL} violations across the two test sets due to poor paraphrasing that degraded grammatical fluency (error 4). This model often misused articles (e.g., "successful in the life") and inappropriately substituted noun types (e.g. "loser" → "lost"). Another common erroneous behaviour that degrades sentence

fluency is the inappropriate use of apostrophe s's (i.e. 's) and other contractions.

D. MR_{JOINT} Analysis

MR_{JOINT} violations consist of the violations from the other three MRs with no duplications. Although MR_{JOINT} does not provide new findings, it is still helpful in getting a consensus of how well a given text detoxification performed in transferring style on a test set.

VIII. Discussion

The approach taken for Section VII has two limitations. First, a comprehensive evaluation of model performance (e.g. proportion of error types, model comparison) was not included in this study. Secondly, confirming an MR violation contains an erroneous output is subjective. Therefore, this judgement may vary across different communities. Expanding analyses across a greater diversity of communities and analyses on model performance would be a priority in future works. Furthermore, our findings suggest there is utility in evaluating source inputs that led to our MR violations (erroneous outputs). For example, to design perturbations following a conventional MT under the same domain.

IX. Conclusion

The proposed metamorphic relations effectively identified defective behaviour across three models on different datasets. Moreover, we found that the MRs helped us gain a reasonable understanding of each model's behaviour and limitations toward text detoxification without detailed knowledge of its underlying architecture. This capability can enable developers to build and improve on the robustness of text detoxification systems, contributing to a safer world for everyone.

Acknowledgment

This work was supported by JC School of Medicine & Health Sciences, Monash University Malaysia Seed Grant (T & L) (Grant No: I-M010-STG-000192 and I-M010-SED-000229).

References

[1] Yan et al., "Metamorphic relations for data validation: A case study of translated text messages," 2019. DOI: 10.1109/met.2019.00018.

[2] P. Fortuna, J. Soler, and L. Wanner, "Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets," English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Calzolari et al., Ed., Marseille, France: European Language Resources Association, May 2020, pp. 6786–6794, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.838.

[3] Floto et al., "Diffudetox: A mixed diffusion model for text detoxification," in *Findings of the Association for Computational Linguistics: ACL-23*, Association for Computational Linguistics, 2023, pp. 7566–7574.

[4] Dale et al., "Text detoxification using large pre-trained neural models," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021. DOI: 10.18653/v1/2021.emnlp-main.629.

[5] L. Sun and Z. Q. Zhou, "Metamorphic testing for machine translations: Mt4mt," in *2018 25th Australasian Software Engineering Conference (ASWEC)*, 2018, pp. 96–100. DOI: 10.1109/ASWEC.2018.00021.

[6] Logacheva et al., "ParaDetox: Detoxification with parallel data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6804–6818. [Online]. Available: https://aclanthology.org/2022.acl-long.469.

[7] K. Atwell, S. Hassan, and M. Alikhani, "Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations," *ArXiv*, vol. abs/2209.08207, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252368208.

[8] I. Smirnov, "The digital flynn effect: Complexity of posts on social media increases over time," in *Social Informatics*. Springer International Publishing, 2017, pp. 24–30, ISBN: 9783319672564. DOI: 10.1007/978-3-319-67256-4_3.

[9] Barr et al., "The oracle problem in software testing: A survey," *IEEE Transactions on Software Engineering*, vol. 41, no. 5, pp. 507–525, 2015. DOI: 10.1109/TSE.2014.2372785.

[10] Chen et al., "Metamorphic testing: A review of challenges and opportunities," *ACM Comput. Surv.*, vol. 51, no. 1, Jan. 2018, ISSN: 0360-0300. DOI: 10.1145/3143561.

[11] Altamimi et al., "Metamorphic relation automation: Rationale, challenges, and solution directions," *Journal of Software: Evolution and Process*, vol. 35, no. 1, e2509, 2023. DOI: https://doi.org/10.1002/smr.2509.

[12] Krishna et al., "Reformulating unsupervised style transfer as paraphrase generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 737–762. DOI: 10.18653/v1/2020.emnlp-main.55.

[13] Pesaranghader et al., "Gpt-detox: An in-context learning-based paraphraser for text detoxification," in *IEE 2023 International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2023. DOI: 10.1109/icmla58977.2023.00230.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, *Bertscore: Evaluating text generation with bert*, 2020. arXiv: 1904.09675 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1904.09675.

Visual semantic alignment network based on pre-trained ViT for few-shot image classification

Jiaming Zhang* ,Jijie Wu* and Xiaoxu Li*
* University of Lanzhou University of Technology,Lanzhou,China
E-mail: lixiaoxu@lut.edu.cn

***Abstract*—Images often contain a wealth of information, but only a portion is relevant to image classification. In contrast, other irrelevant information may interfere, affecting the classification capability. We propose a category-aware feature alignment adaptation module (CAFAAM) to address this issue. The core idea of CAFAAM is to divide the image into multiple patches and process these patches with embedding. Unlike traditional embedding methods, CAFAAM introduces a type-aware embedding module during the embedding process, aiming to capture features related to the classification task more effectively. These embedding results are then fed into a feature fusion module, which uses a support set to construct a query set, thereby integrating feature information from different patches and enhancing classification accuracy. This method was tested on four datasets: CIFAR-FS, miniImageNet, tieredImageNet, and FC-100. The experimental results show that this method outperforms existing methods.**

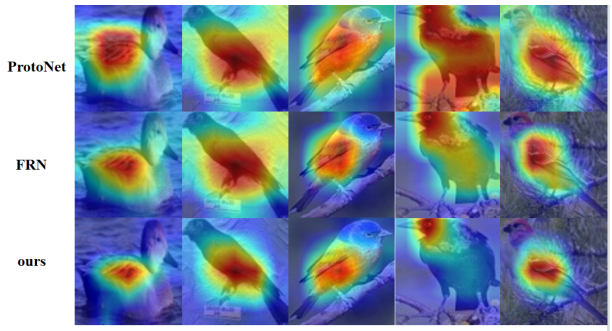


Fig. 1. We compare the ProtoNet [9] and FRN [10] models with our method on local regions of birds in the miniImageNet dataset. Visualization shows that regions with higher energy in the image represent more recognizable parts.

I. Introduction

In most datasets, each image is usually assigned a separate annotation. Although these annotations cannot describe all the information in the image, when there are enough annotated images in each category, the corresponding entities will always appear frequently in the image. In this way, even if the annotation only focuses on a part of the image, the model can learn a large amount of data to recognize and classify the image. However, in the case of few-shot classification, the model must recognize new categories that are different from the training phase, and each category has only a very small number of annotated images, which makes it difficult for the model to confirm which entities are the key information for judging the image, reducing the generalization ability of the model.

A promising approach is to use ViT, which can capture the relationship between different regions in an image through a self-attention mechanism. Unlike traditional convolutional neural networks (CNNs) [1], [2], ViT treats images as serialized inputs and uses a self-attention mechanism to process image pixel relationships. ViT performs well in tasks such as image detection and target classification [3], but ViT lacks the convolutional inductive bias in CNNs. CNNs have some inherent assumptions when processing images, which allow CNNs to work well even with fewer samples. In contrast, ViT lacks such assumptions, so ViT usually requires a lot of data for training. Recently, feature [4] provides an effective solution that enables architectures based entirely on ViT to be successfully generalized to small-scale image datasets.

Another common practice is to align semantically related regions [4]–[6]. For example, SAML [7] uses an activation-based attention mechanism to highlight category-related regions and suppress other areas, thereby enhancing the model's ability to generalize new categories. CNA [8] The core idea is to improve the model's ability to identify category-related regions in an image through an attention mechanism. DeepEMD [5] uses differentiable EMD as a metric to calculate the structural distance between image regions to judge the image's relevance.CTX [6] uses Transformer [6] to further enhance the expressive power of features through a self-attention mechanism, enabling the model to capture the complex relationships between different regions in the image. These feature methods dynamically select the most informative areas of the image and re-adjust the weights of these regions in the feature representation to optimize category recognition. While these techniques do a good job of reducing noise and avoiding over-reliance on training data, they also face challenges. First, the aligned regions may need to be more useful. Second, the small number of samples in the new category means that the model may mistakenly focus on areas that need to be more relevant to the target category, affecting the classification accuracy.

In this paper, we propose a new method, namely category-aware feature alignment adaptation module (CAFAAM), which makes patch embeddings relevant to categories by combining patch embeddings with category-aware embeddings [11] and then introduces a feature fusion module to process further and refine the features. As can be observed from the local region visualization shown in Figure 1, our method enhances the model's ability to identify category features. At the same

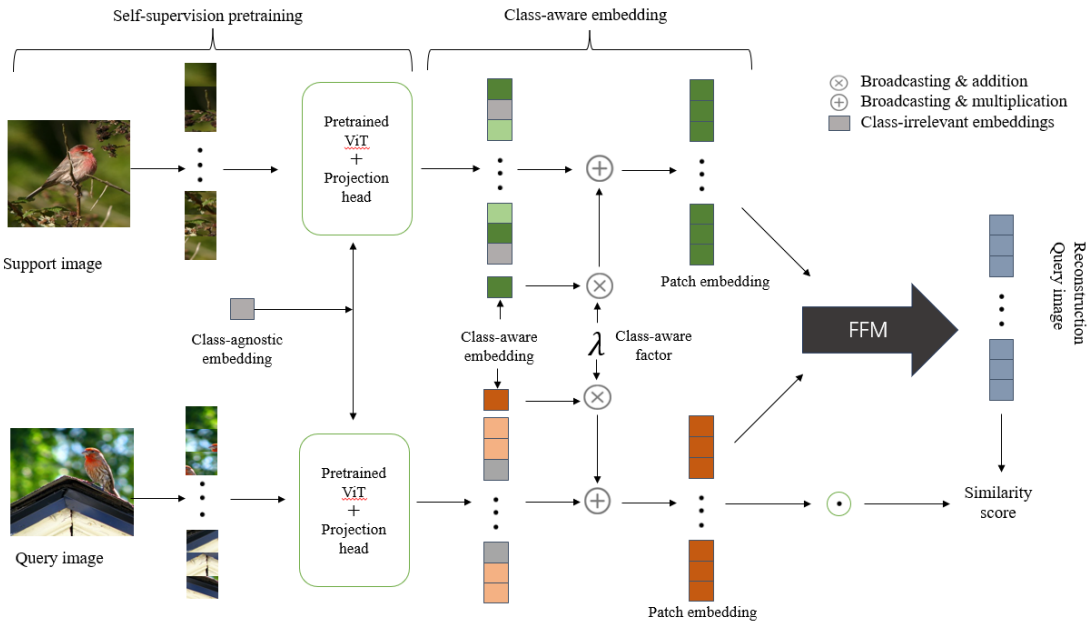


Fig. 2. Illustrates the framework of the class-aware Feature fusion method. The Vision Transformer (ViT) is initially pre-trained using Masked Image Modeling (MIM). Subsequently, a class-aware embedding module is introduced, which interacts and learns continuously with the image patch embeddings within the ViT. At the model's output end, these patch embeddings are adjusted with the class-aware embeddings to ensure they are associated with specific categories. Finally, we introduce a Feature fusion method that utilizes the reconstructed features for similarity measurement.

time, to avoid supervision collapse, we use self-supervised pre-training instead of traditional pre-training. Specifically, this method generates semantically meaningful patch embeddings through Masked Image Modeling (MIM) [12] as an excuse task. We introduce a category-independent embedding that is input into the Transformer model and the patch. This embedding retains the semantic information of the original features and enhances their category relevance.

Our main contributions are summarized as follows:

1) We take a new approach to deal with the problems caused by single-label annotation in a few-shot learning environment.
2) CAFAAM combines the patch embedding of the image with the category information so that the embedding vector contains not only the local features of the image but also the semantic information related to the category. This enables the model to capture the differences between categories more accurately during classification and improves classification accuracy.
3) The feature information of the support set is used to reconstruct the query set, thereby improving the classification performance of the query set.

II. Method

In this section, we first define the few-shot problem and then describe in detail the process of class-aware patch embedding adaptation and feature-feature fusion.

A. Problem Definition

Few-shot image classification aims to apply the patterns learned from the limited training effectively set C_{train} to the test set C_{test}, $C_{train} \cap C_{test} = \emptyset$. In this process, we follow an N-way K-shot classification task. N represents the total number of categories, K represents the number of labeled images available in each category, and Q represents the number of images used for testing each category. Precisely, it consists of two parts: the support set contains N categories, each with K images $X_s = (x_i, y_i)_{i=1}^{NK}$, which are used to train the model; the query set also contains N categories, but each category has only Q images, which are used to evaluate the generalization ability of the model $X_q = (x_i, y_i)_{i=1}^{NQ}$ so that a model that can learn from a small number of samples and generalize to new categories can be trained so that accurate predictions can still be made when facing new and unseen categories.

B. Overview

As shown in Figure 2, we first divide the input image into multiple small patches and map these patches into a high-dimensional feature space through linear projection. To preserve the spatial structure information of the picture, we add position encoding to the feature representation of these patches. In addition, we introduce a category-independent embedding vector to enhance the model's understanding of the overall characteristics of the image. Based on this innovative process, we also introduce a Feature fusion mechanism. Feature fusion generates richer feature representations through the interaction and relearning of internal features.

C. Class-aware Patch Embedding Adaptation

Self-supervised pre-training.In the few-shot image classification task, traditional supervised pre-training methods may cause overfitting [4], resulting in the inability of training data to generalize effectively to new categories. To address this problem, we adopt an innovative self-supervised pre-training method. Specifically, we divide the image into small patches, randomly mask some areas, and use a Vision Transformer (ViT) to encode these image patches to fuse the masked regions. This method, called Masked Image Modeling (MIM) [12], [13], not only generates semantically meaningful patch embedding [14]s but also forces the model to understand the overall structure and content of the image through the feature fusion task. Unlike self-supervised methods that rely on specific training courses, MIM aims to improve the model's deep understanding of images, enhancing its generalization ability when facing new categories. The advantage of this method is that it reduces the dependence on limited labeled data while providing a more robust feature.

Category patch embedding. During the pre-training process of self-supervised learning, the model usually generates many patch embeddings for the input image. These embeddings capture the characteristics of local areas of the image, but the patch embeddings themselves do not contain information about the image category. Some methods [15], [16] directly use the obtained patches as images to alleviate the problem of the number of images. Although there is some improvement, the effect could be clearer. We start from a new perspective and combine patch embedding with class-aware embedding: patch embeddings are not directly associated with any specific category in the initial state. To enhance these embeddings' category relevance, we propose adjusting these patch embeddings by interacting with category tokens. In this method, the category tokens are initially category-independent and do not contain category information at the input stage of the model. However, with continuous interaction with patch embeddings, tokens gradually gain category awareness. We combine local patch embedding with category-aware tokens, which not only consider the regional features of the image but also guide the model through the category token to learn how to integrate these local features into an understanding of the entire image category. This strategy helps the model better understand and classify images and can improve the performance of the model even when the number of images is limited:

$$\bar{z}_i = z_i + \lambda z_{class} \tag{1}$$

Where $\bar{z}_i$ represents the adaptive patch, z_i represents the i-th original patch extracted directly from the input data, and z_{class} represents the information related to the category. $\lambda > 0$ represents the class-aware factor of the correlation size, which adjusts the correlation between adaptive and class-aware embeddings. The larger the value of λ, the higher the correlation between the adapted patch embedding and the class of interest. This helps the model to utilize the information of known categories better when dealing with unknown categories.

D. FFM

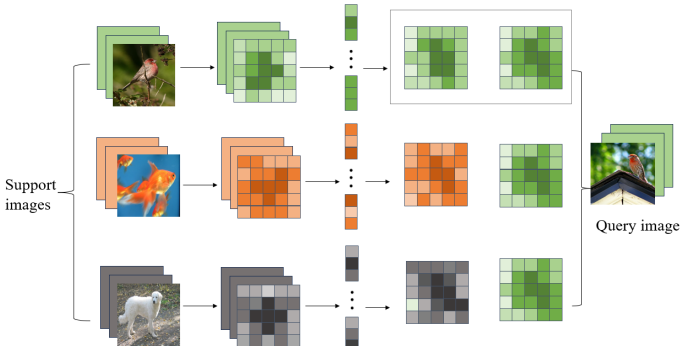


Fig. 3. Feature Fusion Module

Figure 3 shows our feature fusion module. As shown in the figure, for a C-class, K-times classification task, after the initial image block input, we can get the support features of the C-th class, denoted as $F_s = [\bar{z}_i^s] \in \mathbb{R}^{r \times d}$ $s \in \{1, \ldots, K\}$, where d represents the number of channels and r is the feature space resolution, that is, the height of the image multiplied by its width. For each class c, its query feature is $F_q = [\bar{z}_j^q] \in \mathbb{R}^{r \times d}$ $q \in \{1, \ldots, Q\}$, and the attention score S_{ij} is calculated as follows:

$$S_{ij} = \frac{(\bar{z}_j^q)^T \cdot \bar{z}_i^s}{\sqrt{D}} \cdot w \tag{2}$$

where w is the learnable weight parameter.

$$P_{ij} = \frac{\exp(S_{ij})}{\sum_{i=1}^{r} \exp(S_{ij})} \tag{3}$$

The attention probability P_{ij} is obtained by applying the softmax function to the attention scores S_{ij}, where the softmax function converts the raw scores into a probability distribution.

$$\hat{z}_j^q = \sum_{i=1}^{r} P_{ij} \bar{z}_j^q \tag{4}$$

The reconstructed query feature vector $\hat{Q}_i$ can be expressed as:

$$\hat{Q}_i = \begin{bmatrix} \hat{z}_1^q \\ \vdots \\ \hat{z}_r^q \end{bmatrix} \tag{5}$$

For the ith feature of the query set, the similarity matrix M between the reconstructed query set feature matrix $\hat{Q}_i$ and the original query set Q_i can be represented as:

$$M = \hat{Q}_i^T Q_i \tag{6}$$

We define a mean squared error (MSE) loss to measure the difference between $\hat{Q}_i$ and Q_i:

$$L_{\text{MSE}} = \frac{1}{Q} \sum_{j=1}^{r} \|\hat{z}_j^q - \bar{z}_j^q\|_2^2 \tag{7}$$

TABLE I
CLASSIFICATION ACCURACY OF FEW-SHOTS SET BY THIS METHOD ON MINIIMAGENET, TIEREDIMAGENET, CIFAR-FS, FC100 FOR 5-WAY 1-SHOT AND 5-WAY 5-SHOT.

Model	Backbone	$\approx \#Params$	miniImageNet		tieredImageNet		CIFAR-FS		FC100	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
SetFeat [17]	SetFeat-12	12.3 M	68.32±0.62	82.71±0.46	73.63±0.88	87.59±0.57	-	-	-	-
ProtoNet [9]	ResNet-12	12.4 M	62.29±0.33	79.46±0.48	68.25±0.23	84.01±0.56	-	-	41.54±0.76	57.08±0.76
FEAT [18]	ResNet-12	12.4 M	66.78±0.20	82.05±0.14	70.80±0.23	84.79±0.16	-	-	-	-
DeepEMDT [5]	ResNet-12	12.4 M	65.91±0.82	82.41±0.56	71.16±0.87	86.03±0.58	-	-	-	-
IEPT [19]	ResNet-12	12.4 M	67.05±0.44	82.90±0.30	72.24±0.50	86.73±0.34	-	-	-	-
MELR [20]	ResNet-12	12.4 M	67.40±0.43	83.40±0.28	72.14±0.51	87.01±0.35	-	-	-	-
FRN [10]	ResNet-12	12.4 M	66.45±0.19	82.83±0.13	72.06±0.22	86.89±0.14	-	-	-	-
CG [21]	ResNet-12	12.4 M	67.02±0.20	82.32±0.14	71.66±0.23	85.50±0.15	73.00±0.70	85.80±0.50	-	-
DMF [22]	ResNet-12	12.4 M	67.76±0.46	82.71±0.31	71.89±0.52	85.96±0.35	-	-	-	-
InfoPatch [23]	ResNet-12	12.4 M	67.67±0.45	82.44±0.31	-	-	-	-	-	-
BML [24]	ResNet-12	12.4 M	67.04±0.63	83.63±0.29	68.99±0.50	85.49±0.34	73.45±0.47	88.04±0.33	-	-
CNL [25]	ResNet-12	12.4 M	67.96±0.98	83.36±0.51	73.42±0.95	87.72±0.75	-	-	-	-
Meta-NVG [26]	ResNet-12	12.4 M	67.14±0.80	83.82±0.51	74.58±0.88	86.73±0.61	74.63±0.91	74.63±0.91	46.40±0.81	61.33±0.71
PAL [27]	ResNet-12	12.4 M	69.37±0.64	84.40±0.44	72.25±0.72	86.95±0.47	-	-	-	-
COSOC [28]	ResNet-12	12.4 M	69.28±0.49	85.16±0.42	73.57±0.43	87.57±0.10	-	-	-	-
Meta DeepBDC [21]	ResNet-12	12.4 M	67.34±0.43	84.46±0.28	72.34±0.49	87.31±0.32	-	-	-	-
LEO [29]	WRN-28-10	36.5 M	61.76±0.08	77.59±0.12	66.33±0.05	81.44±0.09	-	-	-	-
CC+rot [30]	WRN-28-10	36.5 M	62.93±0.45	79.87±0.33	70.53±0.51	84.98±0.36	73.62±0.31	86.05±0.22	-	-
FEAT [18]	WRN-28-10	36.5 M	65.10±0.20	81.11±0.14	70.41±0.23	84.38±0.16	-	-	-	-
PSST [31]	WRN-28-10	36.5 M	64.16±0.44	80.64±0.32	-	-	77.02±0.38	88.45±0.35	-	-
MetaQDA [32]	WRN-28-10	36.5 M	67.83±0.64	84.28±0.69	74.33±0.65	89.56±0.79	75.83±0.88	88.79±0.75	-	-
OM [33]	WRN-28-10	36.5 M	66.78±0.30	85.29±0.41	71.54±0.29	87.79±0.46	-	-	-	-
FewTURE [4]	ViT-S/16	22 M	68.02±0.88	84.51±0.53	72.96±0.92	86.43±0.67	76.10±0.88	86.14±0.64	46.20±0.79	63.14±0.73
CPEA [11]	ViT-S/16	22 M	71.97±0.65	87.06±0.38	76.93±0.70	90.12±0.45	77.82±0.66	88.98±0.45	**47.24±0.58**	65.02±0.60
Ours(CAFAAM)	ViT-S/16	22 M	**73.19 + 0.63**	**88.23±0.36**	**77.11±0.72**	**90.39±0.41**	**78.46±0.68**	**89.95±0.44**	47.19±0.58	**66.36±0.57**

TABLE II
IMPACT OF PROJECTION HEAD ON FEW-SHOT CLASSIFICATION

Projection head			miniImageNet		tieredImageNet		CIFAR-FS		FC100	
a	b	c	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
		✓	72.88±0.62	87.91±0.35	76.95±0.69	90.17±0.39	77.69±0.65	89.89±0.42	46.65±0.55	65.77±0.53
✓		✓	73.06±0.63	87.99±0.37	77.02±0.71	**90.68±0.41**	77.43±0.64	89.95±0.42	46.77±0.58	65.83±0.56
	✓	✓	72.95±0.62	87.96±0.36	77.04±0.71	90.24±0.41	77.80±0.67	89.93±0.43	46.65±0.57	65.84±0.56
✓	✓	✓	**73.19±0.63**	**88.23±0.36**	**77.11±0.72**	90.39±0.41	**78.46±0.68**	**89.95±0.44**	**47.19±0.58**	**66.36±0.57**

TABLE III
DIFFERENT INPUTS

	miniImageNet		tieredImageNet		CIFAR-FS		FC100	
$\mathbf{x}$	73.01±0.63	87.99±0.37	77.03±0.72	90.11±0.38	78.33±0.67	89.86±0.42	46.98±0.56	66.21±0.57
$\lvert\mathbf{x}\rvert$	73.17±0.68	**88.33±0.36**	77.11±0.68	**90.45±0.42**	78.37±0.67	89.93±0.44	**47.39±0.61**	66.24±0.58
$\mathbf{x^2}$	**73.19±0.63**	88.23±0.36	**77.13±0.72**	90.39±0.41	**78.47±0.68**	**89.95±0.44**	47.13±0.58	**66.36±0.57**

III. EXPERIMENTS

In this section, we first explain the purpose of the experiment and the results obtained, then conduct a comparative analysis with other researchers in the field, and finally perform an ablation experiment on the critical components of the experiment.

A. Experimental Settings

Datasets. We tested our method on four widely recognized few-shot learning benchmark datasets: miniImageNet [34], tieredImageNet [29], CIFAR-FS [35], and FC100 [36]. First, we split the datasets into training, validation, and test sets according to standard experimental practices. The class labels of these datasets are non-overlapping. The categories observed in the training phase will not appear in the validation and test phases. This approach ensures that the knowledge learned by the model in the training phase can be effectively generalized to unseen categories.

Backbone. This paper adopts the ViT-S/16 model as the core architecture of the study. This model is chosen because its number of parameters is comparable to other networks commonly used in few-shot image classification. In this configuration, we use a multi-layer perceptron (MLP) as the projection head, which contains two hidden layers. In the first hidden layer, we apply the GELU activation function, and after the second hidden layer, we use LayerNorm for normalization. Then, we introduce an innovative feature fusion strategy to calculate the attention score using the query vector and the key vector. We normalize it by the softmax function and the scaling factor to obtain the attention probability. With the calculated attention probability, we perform a weighted summation of the value vector to reconstruct the feature. Through the fusion

mechanism, the generalization ability of the model to samples is improved.

Experimental details. We adopted the strategy proposed in the literature [12], strictly set parameters, pre-trained the ViT-S/16 model and used 4 A100 40G GPUs to pre-train ViT-S/16. The total number of training times was 1600, and the batch size was 512. During the training process, we only used the trained part of the dataset to avoid overfitting the test data in the pre-training stage. After that, we trained an inverse ResNet as a decoder. The purpose of the decoder is to convert the high-dimensional feature values learned by the model back to the pixel space of the original image, with an output size of 3 × 84 × 84. At this stage, we used the Adam optimizer with an initial learning rate of 0.01. We set the batch size to 200 and the total number of training times to 1000.

Evaluation protocol. In both 5-way 1-shot and 5-way 5-shot image classification tasks, we evaluate the final performance of the model by randomly sampling 1000 samples from each test category and calculating their average accuracy, where there are 15 query images for each category.

B. Comparison of Results

Table I shows the performance comparison of different methods on four different benchmark datasets, including miniImageNet, tieredImageNet, CIFAR-FS, and FC100, using two settings: 5-way 1-shot and 5-way 5-shot. The experimental results show that the CAFAAM method helps improve the performance of the few-shot learning process in almost all cases. Compared with the state-of-the-art CPEA method, except for one result, the accuracy of our proposed method is improved by about 1% on average in all cases.

C. Ablation Study

Ablation experiments were performed on the critical components of the method proposed in this article, and experiments were conducted on CIFAR-FS, miniImageNet, tieredImageNet, and FC-100.

Projection head. The projection head increases the model's ability to express features by mapping the embedding vector to a higher-dimensional space, thereby improving the classification accuracy in the few-shot learning scenario. As shown in Table 2, we use different projection heads to obtain different experimental results.

Different inputs. As shown in Table 3, for the function fc(2), different input methods affect the performance of the model. The results show that taking the square and absolute value of the input can improve the performance. The square increases the similarity within the class and reduces the similarity between classes. The absolute value operation removes the directionality of the data and only retains the size information. This allows the model to achieve better performance and reduce the noise in the data.

IV. Conclusion

We have introduced a new method called CAFAAM, which enhances the model's efficiency in identifying critical features for image classification tasks through self-supervised learning and feature fusion. Experiments have shown that CAFAAM performs exceptionally well across multiple datasets.

References

[1] J. Lai, S. Yang, W. Liu, *et al.*, "Tsf: Transformer-based semantic filter for few-shot learning," in *European Conference on Computer Vision*, Springer, 2022, pp. 1–19.

[2] M. Ye, X. Lin, G. Burachas, A. Divakaran, and Y. Yao, "Hybrid consistency training with prototype adaptation for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2726–2735.

[3] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[4] M. Hiller, R. Ma, M. Harandi, and T. Drummond, "Rethinking generalization in few-shot classification," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3582–3595, 2022.

[5] C. Zhang, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 203–12 213.

[6] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: Spatially-aware few-shot transfer," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 981–21 993, 2020.

[7] F. Hao, F. He, J. Cheng, L. Wang, J. Cao, and D. Tao, "Collect and select: Semantic alignment metric learning for few-shot learning," in *Proceedings of the IEEE/CVF international Conference on Computer Vision*, 2019, pp. 8460–8469.

[8] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," *Advances in neural information processing systems*, vol. 32, 2019.

[9] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[10] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8012–8021.

[11] F. Hao, F. He, L. Liu, F. Wu, D. Tao, and J. Cheng, "Class-aware patch embedding adaptation for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 905–18 915.

[12] J. Zhou, C. Wei, H. Wang, *et al.*, "Ibot: Image bert pre-training with online tokenizer," *arXiv preprint arXiv:2111.07832*, 2021.

[13] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[14] M. Caron, H. Touvron, I. Misra, *et al.*, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[15] Y. Lifchitz, Y. Avrithis, and S. Picard, "Local propagation for few-shot learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 10 457–10 464.

[16] W. Li, J. Xu, J. Huo, L. Wang, Y. Gao, and J. Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 8642–8649.

[17] A. Afrasiyabi, H. Larochelle, J.-F. Lalonde, and C. Gagné, "Matching feature sets for few-shot image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9014–9024.

[18] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8808–8817.

[19] M. Zhang, J. Zhang, Z. Lu, T. Xiang, M. Ding, and S. Huang, "Iept: Instance-level and episode-level pretext tasks for few-shot learning," in *International Conference on Learning Representations*, 2021.

[20] N. Fei, Z. Lu, T. Xiang, and S. Huang, "Melr: Meta-learning via modeling episode-level relationships for few-shot learning," in *International Conference on Learning Representations*, 2021.

[21] J. Xie, F. Long, J. Lv, Q. Wang, and P. Li, "Joint distribution matters: Deep brownian distance covariance for few-shot classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7972–7981.

[22] C. Xu, Y. Fu, C. Liu, *et al.*, "Learning dynamic alignment via meta-filter for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5182–5191.

[23] C. Liu, Y. Fu, C. Xu, *et al.*, "Learning a few-shot embedding model with contrastive learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 8635–8643.

[24] Z. Zhou, X. Qiu, J. Xie, J. Wu, and C. Zhang, "Binocular mutual learning for improving few-shot classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8402–8411.

[25] J. Zhao, Y. Yang, X. Lin, J. Yang, and L. He, "Looking wider for better adaptive representation in few-shot learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 10 981–10 989.

[26] C. Zhang, H. Ding, G. Lin, R. Li, C. Wang, and C. Shen, "Meta navigator: Search for a good adaptation policy for few-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9435–9444.

[27] J. Ma, H. Xie, G. Han, S.-F. Chang, A. Galstyan, and W. Abd-Almageed, "Partner-assisted learning for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 573–10 582.

[28] X. Luo, L. Wei, L. Wen, *et al.*, "Rectifying the shortcut learning of background for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 073–13 085, 2021.

[29] A. A. Rusu, D. Rao, J. Sygnowski, *et al.*, "Meta-learning with latent embedding optimization," *arXiv preprint arXiv:1807.05960*, 2018.

[30] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord, "Boosting few-shot visual learning with self-supervision," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8059–8068.

[31] Z. Chen, J. Ge, H. Zhan, S. Huang, and D. Wang, "Pareto self-supervised training for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 663–13 672.

[32] X. Zhang, D. Meng, H. Gouk, and T. M. Hospedales, "Shallow bayesian meta learning for real-world few-shot recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 651–660.

[33] G. Qi, H. Yu, Z. Lu, and S. Li, "Transductive few-shot classification on the oblique manifold," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8412–8422.

[34] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[35] L. Bertinetto, J. F. Henriques, P. H. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," *arXiv preprint arXiv:1805.08136*, 2018.

[36] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," *Advances in neural information processing systems*, vol. 31, 2018.

Agent Attention Feature Reconstruction Network for Fine-Grained Few-Shot Image Classification

Dongfei Chang*, Jijie Wu* and Xiaoxu Li*
* Lanzhou University of Technology, Lanzhou, China
E-mail:lixiaoxu@lut.edu.cn

***Abstract*—To develop fine-grained few-shot image classification, it is crucial to study how to increase inter-class variation and reduce intra-class variation. A recently proposed method called the Bidirectional Feature Reconstruction Network (BiFRN) increases inter-class variation by using support sets to reconstruct query sets and reduces intra-class variation by query sets to reconstruct support sets. In our study, we found a problem with BiFRN: The calculation of self-attention weights is affected by noise or inaccurate information, which may cause the model to over-focus on specific locations or ignore features at other locations. To solve this problem, we proposed an Agent Attention Bidirectional Feature Reconstruction Network (AAFRN), which extends BiFRN by fusing an Agent Attention Model to generate a new feature map for each feature map. Compared with BiFRN, our method can enhance feature expressiveness, facilitate global contextual information capture within feature maps, and enable weighted feature fusion across various positions. The experimental results on three widely used fine-grained image classification datasets demonstrate that the proposed method achieves competitive performance compared to other methods.**

I. Introduction

Fine-grained few-shot image classification [1] has recently become an important approach to addressing the data scarcity problem that is widely faced by fine-grained analysis [2]. According to the traditional few-shot setting, effective model transfer is achieved when limited support samples are provided. Unlike conventional methods, Fine-grained few-shot image classification deals with problems with fine-grained features, which means that the model needs to focus on learning subtle but discriminative features. This allows classification not only in terms of categories but, more importantly, can distinguish fine-grained visual differences within a class to achieve fine-grained classification between instances within the class.

In fine-grained few-shot image classification, the self-attention mechanism plays an important role [3]. This mechanism improves the classification performance and discrimination ability of the model by emphasizing the key areas in the image. Vaswani et al. [3] introduced the self-attention mechanism, leading to the development of a novel network architecture known as the Transformer. This architecture not only plays a role in natural language processing [3]–[5], but also plays a role in image classification [5]–[7]. In light of the attention mechanism, some few-shot learning methods began to leverage the self-attention mechanism. Someone proposed a Transformer-based Few-shot Embedding Adaptation (FEAT) algorithm for Few-shot learning [8]. In paper [8], the author tried to establish the mapping relationship between the sets and found that the Transformer was the best choice. The Transformer model can effectively capture the interaction between different images in the set so that the coordinated adaptation between each image can be achieved, unlike the Transformer structure in FEAT, which is exclusively applied to support samples. CTX [9] employs a self-attention mechanism to compute spatial attention weights between the query sample and support sample, facilitating the learning of a category prototype aligning with the query. Subsequently, the classification of the query involves evaluating the distance from the query to the matched category prototype. The introduced few-shot classification method additionally integrates a self-attention mechanism to optimize reconstruction weights within the self-construction model and mutual reconstruction model. PVT [10] uses sparse attention patterns to reduce the computational load by decreasing the number of keys and values. While effective, these approaches sacrifice the ability to model long-range dependencies, remaining less powerful than the global self-attention mechanism. NAT [11] simulates the computation of convolution and calculates the attention of each featured neighborhood. DAT [12] constructs a deformable attention module for data-specific attention patterns. BiFormer [13] employs a two-tier routing attention mechanism to identify the region of interest for each query dynamically. Xiong et al. [14] demonstrated that the attention mechanism enables the model to capture subtle feature differences and achieve better performance in small samples. The integration of attention mechanisms in fine-grained few-shot image learning addresses data scarcity challenges, enhancing model accuracy and generalization [15]. Our research found that although BiFRN achieves the goal of increasing inter-class variation and reducing intra-class variation through the bidirectional reconstruction method, there is still room for improvement. Agent attention adopts agent tokens for aggregating and propagating global information, offering high expressiveness with low computational overhead. This method enhances feature expressiveness, facilitates global contextual information capture within feature maps, and enables weighted feature fusion across various positions.

Initially, efforts concentrated on crafting intricate network structures [1], [16], [17]. However, the outcomes fell short of expectations when juxtaposed with the original approaches. Recent trends have witnessed the surge in popularity of reconstruction-based techniques [9], [18], exemplified by approaches like few-shot Classification With Feature Map Re-

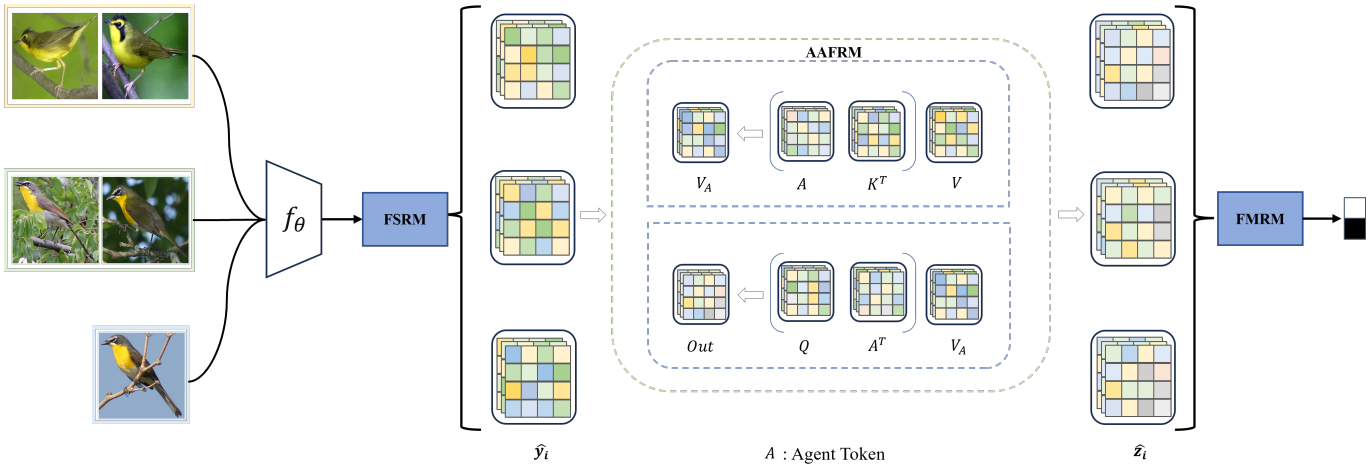


Fig. 1. The proposed agent attention bidirectional feature reconstruction network. f_θ is the feature extraction model. FSRM is a feature self-reconstruction model. AAFRM is the Agent Attention feature reconstruction model. FMRM is a feature mutual reconstruction model. After the image input, we first extract features $\hat{x}_i$ from the image through f_θ. Subsequently, we reconstruct these features to $\hat{y}_i$ through FSRM and input $\hat{y}_i$ into AAFRM. Afterward, we reconstruct the features to $\hat{z}_i$ through AAFRM. Finally, we input $\hat{z}_i$ to FMRM for mutual reconstruction and calculate the results through the Euclidean distance. ($\hat{x}_i \in \mathbb{R}^{d\times r}$, $\hat{y}_i$ and $\hat{z}_i \in \mathbb{R}^{r\times d}$, $r = h \times w$)

construction Networks (FRN) [19], showcasing superior performance. Through aligning support and query features, these models naturally foster fine-grained transfer. However, while these models can emphasize nuanced and distinct regions, the semantic content represented by these regions may differ across samples of the same category. This highlights the substantial intra-class discrepancies that persist within reconstruction-based methods, hindering comprehensive fine-grained learning. In essence, these approaches primarily augment inter-class disparities and fall short in mitigating intra-class variations.

Our agent attention feature reconstruction mainly consists of a feature extractor and three Models: 1) Feature Self-Reconstruction Model (FSRM). 2) Agent Attention Feature Reconstruction Model (AAFRM). 3) Feature Mutual Reconstruction Model (FMRM). With the cooperation of these modules, we can not only effectively improve the expressiveness of features and help the model to be better trained but also increase inter-class differences and reduce intra-class differences.

To summarise, our work makes triple contributions:

- It reveals that the critical issue for fine-grained few-shot image classification is to increase inter-class variation and reduce intra-class variation.
- By integrating agent attention, the expressiveness of feature maps is improved.
- The experimental structure proves the effectiveness of the method.

II. Methodology

In this section, we will introduce the proposed method in this paper, starting with the formula of the method, followed by an overview and an in-depth description of each model.

As shown in Fig. 1, our network mainly consists of a feature extractor and three Models. The feature extraction f_θ is used to extract deep convolutional image features. This part can be a residual network or a traditional convolutional network. The first model is the feature self-reconstruction Model (FSRM), which self-reconstructs the convolutional features of each image based on the self-attention mechanism. This model can make similar local features more similar and dissimilar local features more dissimilar, and it is also conducive to the mutual reconstruction of subsequent features. The second model is the Agent Attention Feature Reconstruction Model (AAFRM), which can improve the expressiveness of the model and help the model capture global context information in the feature map. The third model is the Feature Mutual Reconstruction Model (FMRM), which not only reconstructs the query sample using the support sample but also reconstructs the support sample using the query sample, thereby reducing the intra-class differences. Subsequently, the Euclidean distance [20] is calculated between the query sample and its reconstructed version, as well as between the support sample and its reconstructed version. The weighted sum of these two distances is used to classify the query sample.

A. feature self-construction Model (FSRM)

For a C-way K-shot task, we input $C \times (K + M)$ samples x_i into the feature extraction Model f_θ to extract features $\hat{x}_i = f_\theta(x_i) \in \mathbb{R}^{d\times r}$, where $r = h \times w$ and d is the channel number, h and w are the height and width of the feature. First, we shape the feature $\hat{x}_i$ as r local features in the spatial positions $[\hat{x}_i^1, \hat{x}_i^2, \cdots, \hat{x}_i^r]$, $r = h \times w$. Then the sum of the local feature $\hat{x}_i^j$ and the corresponding spatial position embedding $E_{pos} \in \mathbb{R}^{r\times d}$ is calculated as the input of the transformer[3], $y_i = [\hat{x}_i^1, \hat{x}_i^2, \cdots, \hat{x}_i^r] + E_{pos}$, where E_{pos} adopts sinusoidal position encoding[3]. The output of this part is calculated according to

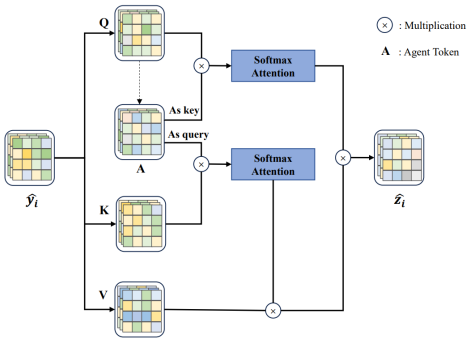


Fig. 2. Agent Attention Feature Reconstruction Model. The output $\hat{y_i}$ from FSRM was reconstructed to $\hat{z_i}$ through gent aggregation and agent broadcast as input to FMRM. ($\hat{y_i}$ and $\hat{z_i} \in \mathbb{R}^{r\times d}$)

the self-attention in the Transformer Encoder. Therefore, the output $\hat{y_i}$ of this Model was calculated using (2),

To simplify, we abbreviate softmax attention as (1),

$$\sigma(QK^T)V \triangleq Attn(Q,K,V) \tag{1}$$

$$\hat{y_i} = \sigma\left(y_i W_\phi^Q (y_i W_\phi^K)^T\right) y_i W_\phi^V, \hat{y_i} \in \mathbb{R}^{r\times d} \tag{2}$$

where $\sigma(\cdot)$ represents the Softmax function, $Attn(\cdot)$ is the self-attention mechanism, $\triangleq$ indicates equivalence, W_ϕ^Q, W_ϕ^K and W_ϕ^V are a set of learnable weight parameters with $d \times d$ size.

B. Agent Attention Feature Reconstruction Model (AAFRM)

We design an agent token A to act as the "agent" for the query Q and the key K, as shown in Fig. 2. Each feature $\hat{y_i}$ is the input of AAFRM, and the output of this Model is $\hat{z_i} \in \mathbb{R}^{r\times d}$ was calculated using (6). The Q, K and V was calculated using (3),

$$Q = yW_Q, K = yW_K, V = yW_V \tag{3}$$

where W_Q, W_K, W_V and W_A are a set of learnable weight parameters with $d \times d$ size. a represents the agent token in the agent attention, which is the result of Q adaptive average pooling was calculated using (4),

$$A = aW_A, a = Polling(Q) \tag{4}$$

where $polling(\cdot)$ is adaptive average pooling, $A \in \mathbb{R}^{r\times d}$, d is the dimension.

The agent attention consists of two softmaxes: agent aggregation and agent broadcast. Specifically, we treat A as query Q and perform an attention calculation between A, K, and V to aggregate agent features V_A from all values. We call this attention calculation aggregation. Afterward, we use A as keys, V_A as value and query matrix Q to perform a second attention calculation called agent broadcast. The final output $\hat{z}$ was calculated using (5),

$$O = \underbrace{Attn^S(Q, A, \underbrace{Attn^S(A,K,V)}_{\text{Agent Aggregation}})}_{\text{Agent Broadcast}} \tag{5}$$

It is equivalent to:

$$\hat{z} = \sigma(QA^T)\sigma(AK^T)V \tag{6}$$

where $\sigma(\cdot)$ represents the Softmax function.

C. Feature Mutual Reconstruction Model (FMRM)

The third Model is the Feature Mutual Reconstruction Model (FMRM), which contains two operations: reconstructing support features in one class given a query feature and reconstructing the query feature given support features in one class. After AAFRM, we can obtain the support features of the reconstructed c class, $S_c = [\hat{z}_k^c] \in \mathbb{R}^{kr\times d}$, where $k \in [1,\cdots,K]$, $c \in [1,\cdots,C]$ and reconstructed query feature $Q_i = \hat{z_i} \in \mathbb{R}^{r\times d}$, where $i \in [1,\cdots,C\times M]$. S_c multiplies by weights W_γ^Q, W_γ^K and W_γ^V, repectively, obtaining S_C^Q, S_C^K, S_C^V, where $W_\gamma^Q, W_\gamma^K, W_\gamma^V \in \mathbb{R}^{d\times d}$. Similarly, Q_i multiplies by weights W_γ^Q, W_γ^K and W_γ^V, repectively, obtaining Q_i^Q, Q_i^K, Q_i^V.

The i query feature $\hat{Q}$ reconstructed from the c class support feature S_c^V and the c class support feature $\hat{S}$ reconstructed from the i query were calculated by (7) and (8),

$$\hat{Q} = \sigma\left(Q_i^Q (S_c^K)^T\right) S_c^V, \hat{Q} \in \mathbb{R}^{r\times d} \tag{7}$$

$$\hat{S} = \sigma\left(S_c^Q \left(Q_i^K\right)^T\right) Q_i^V, \hat{S} \in \mathbb{R}^{kr\times d} \tag{8}$$

The last part is the Euclidean distance metric (EDM)[27]. The distance between the query sample Q_i and the support sample in class c and the distance between the support sample in class c and the query sample Q_i are called d_{QS} and d_{SQ}. d_{QS} and d_{SQ} were calculated using (9) and (10),

$$d_{QS} = ||Q_i^V - \hat{Q}||^2 \tag{9}$$

$$d_{SQ} = ||S_c^V - \hat{S}||^2 \tag{10}$$

where $||\cdot||$ is the Euclidean distance, λ_1 and λ_2 are learnable weight parameters and both of them are initialized as 0.5. τ is a learnable temperature factor. The total distance d_i^c was calculated by (11). Then normalize to get $\hat{d}_i^c$, which was calculated using (12),

$$d_i^c = \tau(\lambda_1 d_{QS} + \lambda_2 d_{SQ}) \tag{11}$$

$$\hat{d}_i^c = \frac{\mathrm{e}^{-d_i^c}}{\sum_{c=1}^{C} \mathrm{e}^{-d_c^i}} \tag{12}$$

Based on $\hat{d}_i^c$, the total loss in one C-way K-shot task was calculated using (13),

$$L = -\frac{1}{M\times C} \sum_{i=1}^{M\times C} \sum_{c=1}^{C} \mathbf{1}(y_i = c) \log \hat{d}_i^c \tag{13}$$

TABLE I
RESULTS ON 1-SHOT AND 5-SHOT TASKS IN THE CONV4 BACKBONE. THE DATASETS WE USE ARE CUB-200-2011 (CUB), STANFORD-DOGS (DOGS) AND STANFORD-CARS (CARS).

Model	CUB		Dogs		Cars	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet (NerulPS 17') [20]	64.82±0.23	85.74±0.14	46.66±0.21	70.77±0.16	50.88±0.23	74.89±0.18
Relation (CVPR 18') [21]	63.94±0.92	77.87±0.64	47.35±0.88	66.20±0.74	46.04±0.91	68.52±0.78
DN4 (CVPR 19') [22]	57.45±0.89	84.41±0.58	39.08±0.76	69.81±0.69	34.12±0.68	87.47±0.47
PARN (ICCV 19') [23]	74.43±0.95	83.11±0.67	55.86±0.97	68.06±0.72	66.01±0.94	73.74±0.70
SAML (ICCV 19') [24]	65.35±0.65	78.47±0.41	45.46±0.36	59.65±0.51	61.07±0.47	88.73±0.49
DeepEMD (CVPR 20') [25]	64.08±0.50	80.55±0.71	46.73±0.49	65.74±0.63	61.63±0.27	72.95±0.38
LRPABN (TMM 21') [1]	63.63±0.77	76.06±0.58	45.72±0.75	60.94±0.66	60.28±0.76	73.29±0.58
BSNet (TIP 21') [26]	62.84±0.95	85.39±0.56	43.42±0.86	71.90±0.68	40.89±0.77	86.88±0.50
CTX (NeurrIPS 20') [9]	72.61±0.21	86.23±0.14	57.86±0.21	73.59±0.16	66.35±0.21	82.25±0.14
FRN (CVPR 21') [19]	74.90±0.21	89.39±0.12	60.41±0.21	79.26±0.15	67.48±0.22	87.97±0.11
BiFRN (AAAI 23') [27]	79.08±0.20	92.22±0.10	64.74±0.22	81.29±0.14	75.74±0.20	91.58±0.09
Ours	**79.76±0.20**	**92.57±0.10**	**65.30±0.22**	**81.30±0.15**	**77.22±0.19**	**91.85±0.09**

TABLE II
RESULTS ON 1-SHOT AND 5-SHOT TASKS IN THE RESNET12 BACKBONE. THE DATASETS WE USE ARE CUB-200-2011 (CUB), STANFORD-DOGS (DOGS) AND STANFORD-CARS (CARS).

Model	CUB		Dogs		Cars	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet (NerulPS 17') [20]	81.02±0.20	91.93±0.11	73.81±0.21	87.39±0.12	85.46±0.19	95.08±0.08
CTX (NeurIPS 20') [9]	80.39±0.20	91.01±0.11	73.22±0.22	85.90±0.13	85.03±0.19	92.63±0.11
DeepEMD (CVPR 20') [25]	75.59±0.30	88.23±0.18	70.38±0.30	85.24±0.18	80.62±0.26	92.63±0.13
FRN (CVPR 21') [19]	84.30±0.18	93.34±0.10	76.76±0.21	**88.74±0.12**	88.01±0.17	95.75±0.07
BiFRN (AAAI 23') [27]	85.44±0.18	94.73±0.09	76.89±0.21	88.27±0.12	**90.44±0.15**	**97.49±0.05**
Ours	**86.16±0.17**	**94.82±0.07**	**76.89±0.21**	88.72±0.12	89.77±0.15	97.22±0.06

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Datasets

For performance evaluation, we assessed our method on three benchmark fine-grained datasets: CUB-200-2011 (CUB) [28], consisting of 11,788 images across 200 bird species; Stanford-Dogs (Dogs) [29], comprising 20,580 annotated images representing 120 dog breeds worldwide; and Stanford-Cars (Cars) [30], a dataset featuring 16,185 images covering 196 car types. All images were resized to 84×84.

B. Implementation Details

In our experiments, we tested two widely used backbone architectures: Conv-4 [19] and ResNet-12 [19]. The experiments were conducted using PyTorch on a single NVIDIA GeForce RTX 4090 GPU. We trained Conv-4 and ResNet-12 models for 1200 epochs with a weight shrinkage of 0.9, an initial learning rate of 0.1 and a 10x learning rate reduction every 400 epochs. For Conv-4 models, we employed 30-way 5-shot training, while for ResNet-12 models, we used 15-way 5-shot training. Regular data augmentation techniques such as center cropping, random horizontal flipping and color jittering were applied to enhance stability.

C. Experimental Results

We use Conv-4 and ResNet-12 as the backbone of all compared models and test 5-way 1-shot and 5-way 5-shot classification performance. The outcomes utilizing Conv-4 and ResNet-12 as backbone networks are detailed in Table I and Table II, correspondingly.

In Table I, we can see that when Conv-4 is adopted, our method achieves the highest accuracy on all three datasets. In Table II, Apart from the result on the cars [30] dataset where performance falls slightly behind the BiFRN method when the ResNet-12 is adopted, our method achieves the highest accuracy.

In summary, compared with other newly proposed methods, our method has achieved stable and excellent performance on three fine-grained image datasets for 5-way 1-shot and 5-way 5-shot classification tasks. The improvement in experimental results is mainly due to the rationality of network design. FSRM makes similar local features more similar and dissimilar local features more dissimilar, which is beneficial for the subsequent model in feature reconstruction. FMRM utilizes support samples to reconstruct query samples and employs query samples to reconstruct support samples, thereby helping to reduce the intra-class differences of features. More importantly, AAFRM further enhances the feature expressiveness,

TABLE III
RESULTS ON 1-SHOT AND 5-SHOT TASKS WITH DIFFERENT BACKBONES ON CUB, DOGS AND CARS DATASETS.

Backbone	Method	CUB		Dogs		Cars	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Conv-4	Baseline (ProtoNet)	64.82±0.23	85.74±0.14	46.66±0.21	70.77±0.16	50.88±0.23	74.89±0.18
	(FSRM)	79.08±0.20	92.22±0.10	64.74±0.22	81.29±0.14	75.74±0.20	91.58±0.09
	(AAFRM)	78.21±0.20	91.82±0.11	64.61±0.22	81.25±0.14	75.40±0.20	91.04±0.10
	(AAFRM, FSRM)	78.74±0.20	91.83±0.11	63.98±0.22	81.17±0.14	74.56±0.20	90.70±0.10
	(Ours)	**79.76±0.20**	**92.57±0.10**	**65.30±0.22**	**81.30±0.15**	**77.22±0.19**	**91.85±0.09**
ResNet-12	Baseline (ProtoNet)	81.02±0.20	91.93±0.11	73.81±0.21	87.39±0.12	85.46±0.19	95.08±0.08
	(FSRM)	85.44±0.18	94.73±0.09	76.89±0.21	88.27±0.12	**90.44±0.15**	**97.49±0.05**
	(AAFRM)	85.14±0.18	94.78±0.08	76.89±0.21	88.76±0.12	90.14±0.15	97.34±0.05
	(AAFRM, FSRM)	85.18±0.18	94.48±0.09	76.80±0.21	**88.93±0.11**	89.75±0.15	97.45±0.05
	(Ours)	**86.16±0.17**	**94.82±0.07**	**76.89±0.21**	88.72±0.12	89.77±0.15	97.22±0.06

helps capture global contextual information within the feature map and achieves weighted feature fusion across various locations.

D. Ablation Study

To evaluate our method and model, we employed Conv-4 and ResNet-12 as backbones and conducted ablation studies on three datasets.

First, we verify the validity of our approach by removing some models. Then, we run the two models, FSRM and AAFRM, in parallel to verify the method's effectiveness. As shown in Table III, we remove FSRM and AAFRM, respectively, and run FSRM and AAFRM in parallel. The experimental results in Table III show that our fusion approach's model performance is further improved when using Conv-4 as the backbone network. However, when using ResNet-12 as the backbone network, the model performance is only improved on a portion of the dataset.

IV. CONCLUSIONS

In this paper, we integrated an agent attention Model to enhance feature expressiveness, facilitate global contextual information capture within feature maps, and enable weighted feature fusion across various positions. Extensive experiments show that the proposed method performs well in three fine-grained image datasets and is very competitive.

REFERENCES

[1] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Transactions on Multimedia*, vol. 23, pp. 1666–1680, 2020.

[2] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, *et al.*, "Fine-grained image analysis with deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8927–8948, 2021.

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[7] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[8] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8808–8817.

[9] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: Spatially-aware few-shot transfer," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 981–21 993, 2020.

[10] W. Wang, E. Xie, X. Li, *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[11] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6185–6194.

[12] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4794–4803.

[13] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, "Biformer: Vision transformer with bi-level routing attention," in *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition, 2023, pp. 10 323–10 333.

[14] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *International conference on machine learning*, PMLR, 2020, pp. 5156–5165.

[15] Y. Xiong, Z. Zeng, R. Chakraborty, *et al.*, "Nyströmformer: A nyström-based algorithm for approximating self-attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 14 138–14 148.

[16] X. Sun, H. Xv, J. Dong, H. Zhou, C. Chen, and Q. Li, "Few-shot learning for domain-specific fine-grained image classification," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 4, pp. 3588–3598, 2020.

[17] Y. Zhu, C. Liu, S. Jiang, *et al.*, "Multi-attention meta learning for few-shot fine-grained image recognition.," in *IJCAI*, Beijing, 2020, pp. 1090–1096.

[18] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: Spatially-aware few-shot transfer," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 981–21 993, 2020.

[19] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8012–8021.

[20] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[21] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[22] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7260–7268.

[23] Z. Wu, Y. Li, L. Guo, and K. Jia, "Parn: Position-aware relation networks for few-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6659–6667.

[24] F. Hao, F. He, J. Cheng, L. Wang, J. Cao, and D. Tao, "Collect and select: Semantic alignment metric learning for few-shot learning," in *Proceedings of the IEEE/CVF international Conference on Computer Vision*, 2019, pp. 8460–8469.

[25] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 203–12 213.

[26] X. Li, J. Wu, Z. Sun, Z. Ma, J. Cao, and J.-H. Xue, "Bsnet: Bi-similarity network for few-shot fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 30, pp. 1318–1331, 2020.

[27] J. Wu, D. Chang, A. Sain, *et al.*, "Bi-directional feature reconstruction network for fine-grained few-shot image classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 2821–2829.

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[29] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, vol. 2, 2011.

[30] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

A Quasilinear-Time CVP Algorithm for Triangular Lattice Based Fuzzy Extractors and Fuzzy Signatures

Kenta Takahashi[*] and Wataru Nakamura[†]
[*]Hitachi, Ltd., Yokohama, Japan
E-mail: kenta.takahashi.bw@hitachi.com
[†]Hitachi, Ltd., Yokohama, Japan
E-mail: wataru.nakamura.va@hitachi.com

***Abstract*— Fuzzy Extractor (FE) and Fuzzy Signature (FS) are useful schemes for generating cryptographic keys from fuzzy data such as biometric features. To implement FE and FS for fuzzy data in a Euclidean space, such as facial feature vectors, several techniques have been proposed that use lattice-based error correction. In these techniques, it is necessary to find the closest vector in the lattice to a given fuzzy data at the time of key reproduction or signing. However, the closest vector problem (CVP) in a general lattice is known to be NP-hard. Therefore, it is necessary to use a special lattice in which an efficient CVP algorithm can be constructed. In this paper, we propose a faster algorithm for CVP in triangular lattices with $O(n \log n)$-time compared to the conventional $O(n^2)$-time algorithm.**

I. INTRODUCTION

Fuzzy Extractor (FE) [1] is a general method to extract the same random numbers from fuzzy data within a certain error margin. FE enables the realization of cryptographic systems using biometric information or PUFs (Physically Unclonable Function) [2] as key management mediums. Similarly, Fuzzy Signature (FS) [3] is a digital signature scheme that uses fuzzy data itself as a signing key.

The specific algorithms of FE and FS depend on the metric space to which the fuzzy data belong and the error range to be tolerated. For example, if the fuzzy data is in n-bit Hamming space and you want to tolerate up to t-bit errors, there are known FE algorithms using n-bit error correcting codes that can correct t-bit errors [1]. With respect to the strength of the keys to be extracted, the error correcting code (ECC) should have as many codewords as possible for given (n, t). In this sense, it is ideal for the ECC to be a *perfect code*[1], which divides the n-bit space completely with Hamming spheres of radius t.

On the other hand, in biometric authentication methods such as face recognition, features are often extracted as n-dimensional vectors by machine learning or other methods, and errors are often evaluated as Euclidean distance[2]. Applying FE and FS here requires an error correction mechanism in Euclidean space, but since n-dimensional space cannot be divided by n-dimensional hyperspheres, an approximate division must be used instead. As a technique for this purpose, a method using an n-dimensional lattice structure has been proposed [6,7,8]. The desirable properties of the lattice to be applied to FE and FS are as follows.

(1) The lattice gives a dense sphere packing.
(2) The closest lattice point to a given arbitrary vector can be computed efficiently.

The first property is important to ensure the strength of the key to be extracted. Such lattice structures have been studied for many years in the context of the *sphere packing problem* [9], but for the densest structures it is an open problem except in $n = 1, 2, 3, 8$ and 24 dimensions. The second requirement is also important to optimize the balance between security and performance. However, the CVP in general lattices is known to be NP-hard [10].

As a lattice structure that satisfies these properties to some extent, Yoneyama et al. proposed the use of triangular lattices and constructed an efficient algorithm for CVP with $O(n^2)$-time complexity for n-dimensions by exploiting the high symmetry of the lattice [6]. However, when applied to biometric systems, this algorithm may take too much time to compute. For example, the typical number of feature dimensions in face recognition is around $n = 128 \sim 512$, and when applied to a 1:N authentication system with $N = 100{,}000$, the evaluated processing time is $3.8 \sim 62$ seconds, as shown in sections IV.B.

In this paper, we propose a fast CVP algorithm for triangular lattices with a computational time complexity of $O(n \log n)$. The main ideas are as follows. The first is to speed up the coordinate transformation subroutine, which conventionally took $O(n^2)$-time, to $O(n)$-time by using a kind of memorization technique that exploits the properties of the expression of the basis matrix of the triangular lattice. The second is to speed up

[1] Although there are a limited number of (n, t) pairs for which a perfect code exists.

[2] Feature extraction methods such as CosFace [4] and ArcFace [5], where a feature vector is $L2$−normalized and the 'closeness' between two vectors $\boldsymbol{x}, \boldsymbol{y}$ is defined by the cosine similarity $S_C(\boldsymbol{x}, \boldsymbol{y}) := \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} = \boldsymbol{x} \cdot \boldsymbol{y}$, are often used for face recognition. This is equivalent to using Euclidean distance, since $\|\boldsymbol{x} - \boldsymbol{y}\|^2 = 2(1 - S_C(\boldsymbol{x}, \boldsymbol{y}))$.

the subroutine that searches for the closest vector from n candidates, which conventionally took $O(n^2)$-time, to $O(n)$-time by showing that the sequence of distances from the target vector to the suitably sorted candidates is convex and introducing a binary search.

II. Preliminaries and Related Works

A. Triangular Lattice

In an n-dimensional real Euclidean space $\mathbb{R}^n$, take n independent vectors $\boldsymbol{b}^1, \cdots, \boldsymbol{b}^n \in \mathbb{R}^n$ (basis vectors) and let $B = (\boldsymbol{b}^1, \cdots, \boldsymbol{b}^n)$ be a matrix (basis matrix). Then the lattice $L(B)$ is defined as

$$L(B) := \{B\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{Z}^n\}.$$

That is, $L(B)$ is the set of all vectors $\boldsymbol{b}^1, \cdots, \boldsymbol{b}^n$ linearly combined with integer coefficients.

A lattice L is called a triangular lattice if it has a basis $B = (\boldsymbol{b}^1, \cdots, \boldsymbol{b}^n)$ satisfying the following conditions.

$$\forall_i \left\|\boldsymbol{b}^i\right\| = c \text{ (constant)}, \qquad \forall_{i \neq j}\, \boldsymbol{b}^i \cdot \boldsymbol{b}^j = \frac{c^2}{2}.$$

The above conditions imply that an angle between any two basis vectors is π/3 (Figure 1). Triangular lattices have been proven to give the densest sphere-packing in two and three dimensions. It has also been shown experimentally to have relatively good performance in higher dimensions when applied to FE [6].

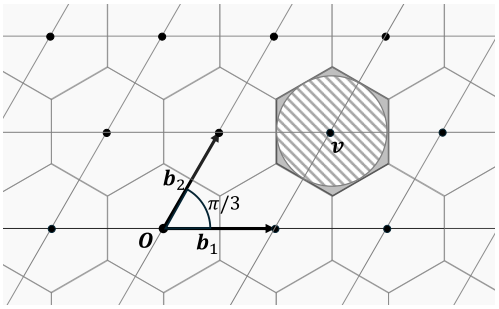


Figure 1: Black dots are the points of a triangular lattice in $\mathbb{R}^2$ spanned by $\boldsymbol{B} = (\boldsymbol{b}^1, \boldsymbol{b}^2)$. The gray hexagon represents a Voronoi cell centered at lattice point $\boldsymbol{v}$. The shaded disk represents a sphere with center $\boldsymbol{v}$ whose radius is half the minimum distance.

For simplicity, and without loss of generality, the following discussion deals with the triangular lattice with $c = 1$ and a basis matrix expressed as

$$B = \begin{pmatrix} \alpha_1 & \beta_1 & \beta_1 & \cdots & \beta_1 \\ 0 & \alpha_2 & \beta_2 & \cdots & \beta_2 \\ 0 & 0 & \alpha_3 & \cdots & \beta_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_n \end{pmatrix}. \tag{1}$$

where α_k and β_k are constants computed by the following simultaneous difference equations:

$$\alpha_k = \sqrt{1 - \sum_{i=1}^{k-1} \beta_i^2}, \qquad \beta_k = \frac{\frac{1}{2} - \sum_{i=1}^{k-1} \beta_i^2}{\alpha_k} \quad (k \in [n]), \tag{2}$$

where $[n]$ denotes the set $\{1, \cdots, n\}$. The first few values are as follows:

$$\alpha_1 = 1, \qquad \alpha_2 = \frac{\sqrt{3}}{2}, \qquad \alpha_3 = \frac{\sqrt{6}}{3}, \qquad \cdots,$$

$$\beta_1 = \frac{1}{2}, \qquad \beta_2 = \frac{\sqrt{3}}{6}, \qquad \beta_3 = \frac{\sqrt{6}}{12}, \qquad \cdots.$$

B. Conventional CVP Algorithm for Triangular Lattices

A CVP algorithm takes as input a target vector $\boldsymbol{y} \in \mathbb{R}^n$ and outputs the closest lattice point. In the following, we outline the conventional CVP algorithm for triangular lattices proposed by Yoneyama et al. See reference [6] for details.

Conventional CVP Algorithm $CV(\boldsymbol{y})$

Input: $\boldsymbol{y} \in \mathbb{R}^n$

1: $(x_1, \cdots, x_n)^t \leftarrow B^{-1}\boldsymbol{y}$

2: **for** $i \leftarrow 1, \cdots, n$ **do**

$w_i \leftarrow \lfloor x_i \rceil$

$r_i \leftarrow x_i - w_i$

end for

$\boldsymbol{w} \leftarrow (w_1, \cdots, w_n)^t$

3: $\left(r_{\sigma(1)}, \cdots, r_{\sigma(n)}\right) \leftarrow Sort(r_1, \cdots, r_n)$ // descending order

4: **for** $k \leftarrow 0, \cdots, n$ **do**

$\boldsymbol{z}^k \leftarrow (z_1^k, \cdots, z_n^k)^t$ where

$$z_j^k \leftarrow \begin{cases} 1, & \text{for } j \leftarrow \sigma(1), \cdots, \sigma(k) \\ 0, & \text{for } j \leftarrow \sigma(k+1), \cdots, \sigma(n) \end{cases}$$

end for

5: $\boldsymbol{z} \leftarrow argmin_{\boldsymbol{z}^k \in \{\boldsymbol{z}^0, \cdots, \boldsymbol{z}^n\}} \|B\boldsymbol{r} - B\boldsymbol{z}^k\|$

6: **return** $B(\boldsymbol{w} + \boldsymbol{z})$

The time complexity of each step is $O(n^2)$ for step 1, 4 and 6, $O(n)$ for step 2 and $O(n \log n)$ for step 3. Step 5 appears to take $O(n^3)$-time at first glance, but can be computed in $O(n^2)$-time by using the following difference equation:

$$B\boldsymbol{z}^n = 0,$$
$$B\boldsymbol{z}^{k-1} = B\boldsymbol{z}^k + \boldsymbol{b}_{\sigma(k)} \ (\text{for } k = n, n-1, \cdots, 1).$$

Therefore, the total complexity is $O(n^2)$.

III. Proposed Algorithm

Based on the conventional algorithm, let us consider speeding up steps 1, 4, 5, and 6, which take $O(n^2)$-time. Specifically, the coordinate transformation process (product of basis matrix and vector) in steps 1 and 6 and the closest vector search (enumeration of candidate vectors and distance calculation) in steps 4 and 5 are accelerated as subroutines, respectively.

A. Coordinate Transformation Subroutine

Let us consider algorithms for mutual transformation between $\boldsymbol{y} = (y_1, \cdots, y_n)^t$ (Cartesian coordinate system) and $\boldsymbol{x} = (x_1, \cdots, x_n)^t$ (triangular lattice basis coordinate system) that satisfies the following relation:

$$\boldsymbol{y} = B\boldsymbol{x} \ \Leftrightarrow \boldsymbol{x} = B^{-1}\boldsymbol{y}.$$

In general, the product of a matrix and a vector requires $O(n^2)$-time, but by utilizing the basis matrix representation in Eq (1), a linear-time algorithm can be constructed as follows. Defining s_k $(k \in [n])$ as

$$s_k := \sum_{i=k+1}^{n} x_k \ ,$$

we have

$$s_n = 0, \tag{3}$$
$$s_{k-1} = s_k + x_k \ (k = 2, \cdots, n). \tag{4}$$

Furthermore, from Eq (1) the following equation holds:

$$y_k = \alpha_k x_k + \beta_k s_k \ (k = 1, \cdots, n). \tag{5}$$

Therefore, given $(x_1, \cdots, x_n)$, $(y_1, \cdots, y_n)$ can be obtained by solving equation (3), (4) and (5) sequentially in the following order: $y_n \to s_{n-1} \to y_{n-1} \to s_{n-2} \to \cdots \to s_1 \to y_1$. Similarly, given $(y_1, \cdots, y_n)$, we can obtain $(x_1, \cdots, x_n)$ by solving equation (3), (4) and (5) sequentially in the following order: $x_n \to s_{n-1} \to x_{n-1} \to s_{n-2} \to \cdots \to s_1 \to x_1$. The time complexity of these algorithms is $O(n)$.

Triangular to Cartesian Coordinates $T2C(\boldsymbol{x})$

Input: $\boldsymbol{x} = (x_1, \cdots, x_n)^t \in \mathbb{R}^n$

1: $s_n \leftarrow 0$

2: **for** $k \leftarrow n, n-1, \cdots, 1$ **do**

$y_k \leftarrow \alpha_k x_k + \beta_k s_k$

$s_{k-1} \leftarrow s_k + x_k$

end for

3: **return** $\boldsymbol{y} = (y_1, \cdots, y_n)^t$

Cartesian to Triangular Coordinates $C2T(\boldsymbol{y})$

Input: $\boldsymbol{y} = (y_1, \cdots, y_n)^t \in \mathbb{R}^n$

1: $s_n \leftarrow 0$

2: **for** $k \leftarrow n, n-1, \cdots, 1$ **do**

$x_k \leftarrow \dfrac{y_k - \beta_k s_k}{\alpha_k}$

$s_{k-1} \leftarrow s_k + x_k$

end for

3: **return** $\boldsymbol{x} = (x_1, \cdots, x_n)^t$

B. Closest Vector Search Subroutine

Here we describe an idea to speed up steps 4 and 5 of the conventional CVP algorithm. The goal here is to efficiently determine the k that minimize $d_k := \|B\boldsymbol{r} - B\boldsymbol{z}^k\|^2$ without calculating $\boldsymbol{z}^k$ and d_k for each $k = 0, 1, \cdots, n$. d_k can be expanded as follow:

$$\begin{aligned} d_k &= \|B\boldsymbol{r}\|^2 + \|B\boldsymbol{z}^k\|^2 - 2B\boldsymbol{r} \cdot B\boldsymbol{z}^k \\ &= \|B\boldsymbol{r}\|^2 + \left\|\boldsymbol{b}^{\sigma(1)} + \cdots + \boldsymbol{b}^{\sigma(k)}\right\|^2 - 2(r_1\boldsymbol{b}^1 + \cdots r_n\boldsymbol{b}^n) \\ &\qquad \cdot \left(\boldsymbol{b}^{\sigma(1)} + \cdots + \boldsymbol{b}^{\sigma(k)}\right) \\ &= \|B\boldsymbol{r}\|^2 + \frac{1}{2}k(k+1) \\ &\qquad - k(r_1 + \cdots r_n) - \left(r_{\sigma(1)} + \cdots + r_{\sigma(k)}\right) \\ &= \|B\boldsymbol{r}\|^2 + \frac{1}{2}k^2 + k\left(\frac{1}{2} - r_{sum}\right) - \sum_{i=1}^{k} r_{\sigma(i)}, \end{aligned}$$

where $r_{sum} := \sum_{i=1}^{n} r_i$. Therefore the first and second order differences of d_k are:

$$\Delta_k := d_k - d_{k-1} = k - r_{sum} - r_{\sigma(k)},$$
$$\Delta_k^2 := \Delta_k - \Delta_{k-1} = 1 - \left(r_{\sigma(k)} - r_{\sigma(k-1)}\right).$$

Since $0 \le r_i < 1$ $(i \in [n])$ we have $\Delta_k^2 > 0$. Therefore, if we regard d_k as a (discrete) function of k, it is convex downwards, and the k that gives the minimum value is:

i. if $\Delta_1 \ge 0$ then $k = 0$,
ii. if $\Delta_n \le 0$ then $k = n$,
iii. otherwise the $k \in [n-1]$ satisfying $\Delta_k \le 0 \le \Delta_{k+1}$.

In case iii, k can be found in $O(\log n)$ time by a binary search for Δ_k.

C. *Fast CVP Algorithm*

The following is the proposed CVP algorithm that uses the subroutine described in sections III.A and III.B to speed up the conventional algorithm.

Proposed Fast CVP Algorithm $FastCV(\boldsymbol{y})$

Input: $\boldsymbol{y} \in \mathbb{R}^n$

1: $(x_1, \cdots, x_n)^t \leftarrow C2T(\boldsymbol{y})$
2: **for** $i \leftarrow 1, \cdots, n$ **do**
  $w_i \leftarrow \lfloor x_i \rceil$
  $r_i \leftarrow x_i - w_i$
  end for
  $\boldsymbol{w} \leftarrow (w_1, \cdots, w_n)^t$
3: $(r_{\sigma(1)}, \cdots, r_{\sigma(n)}) \leftarrow Sort(r_1, \cdots, r_n)$ //descending order.
4: $r_{sum} \leftarrow \sum_{i=1}^{n} r_i$
5: **If** $1 - r_{sum} - r_{\sigma(1)} \ge 0$ **then**
  $k \leftarrow 0$
  else if $n - r_{sum} - r_{\sigma(n)} \le 0$ **then**
  $k \leftarrow n$
  Otherwise
  determine $k \in [n-1]$ satisfying the following inequality by a binary search:
  $k - r_{\sigma(k)} \le r_{sum} \le k + 1 - r_{\sigma(k+1)}$
  end if
6: $\boldsymbol{z} \leftarrow (z_1, \cdots, z_n)^t$ where

$$z_j \leftarrow \begin{cases} 1, & \text{for } j \leftarrow \sigma(1), \cdots, \sigma(k) \\ 0, & \text{for } j \leftarrow \sigma(k+1), \cdots, \sigma(n) \end{cases}$$

7: **return** $T2C(\boldsymbol{w} + \boldsymbol{z})$

IV. Evaluation

A. *Asymptotic Complexity*

Let us evaluate the time complexity of the proposed CVP algorithm $FasctCV$. The complexity of coordinate transformation algorithms $C2T$ and $T2C$ in Steps 1 and 7 is $O(n)$-time as described in III.A. Step 3 can be computed in $O(n \log n)$-time. Step 5 is a subroutine described in III.B and can be computed in $O(\log n)$-time. Since other steps are $O(n)$-time, the total complexity of $FastCV$ is $O(n \log n)$-time.

B. *Experimental Evaluation*

We implemented the proposed CVP algorithm and the conventional algorithm to evaluate their computation times experimentally. We randomly selected 10,000 real vectors $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{10000}$ from the range $\boldsymbol{y}_i \in [-100, 100]^n$ as target vectors, executed $CV(\boldsymbol{y}_i)$ and $FastCV(\boldsymbol{y}_i)$ for all $\boldsymbol{y}_i$ $(i = 1, \cdots, 10000)$, measured the total computation time for each algorithm, and evaluated the average time per run for each algorithm. The results for $n = 128, 256$, and 512 are shown in Table 1 and Figure 2. The computing environment is as follows: OS: Windows 10 Pro, CPU: Intel(R) Core(TM) i7-8700K @ 3.70GHz, memory 32GB, storage: SSD.

The proposed algorithm is about 6.6 times faster than the conventional algorithm when $n = 128$, about 12 times faster when $n = 256$, and about 24 times faster when $n = 512$. The number of processes per second when $n = 512$ is 3.82×10^4 with the proposed algorithm while it is 1.60×10^3 with the conventional algorithm.

Table 1: Average Computation time of the algorithms (μs)

Algorithm	$n = 128$	$n = 256$	$n = 512$
Conventional	38.2	156	623
Proposed	5.8	12.5	26.2

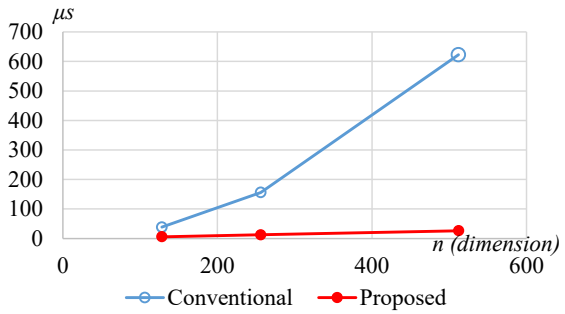


Fig. 2 Graphical comparison of the computation time of the algorithms.

V. Conclusions

In this paper, we proposed a fast algorithm for the closest vector problem (CVP) on high-dimensional triangular lattices. The time complexity of the proposed algorithm is $O(n \log n)$, whereas that of conventional algorithms is $O(n^2)$ for n dimensions. Experimental evaluation shows that the proposed algorithm achieves a speedup of about 24 times when $n = 512$. The CVP algorithm for high-dimensional triangular lattices is useful for constructing Fuzzy Extractors (FE) and Fuzzy Signatures (FS) using biometric data such as facial feature vectors. Therefore it is expected to realize real-time biometric identification for large-scale users with biometric template protection [11] based on FE and FS such as the Public Biometric

Infrastructure (PBI) [3]. A more detailed evaluation (e.g., using large-scale face datasets) is a subject for future work.

REFERENCES

[1] Yevgeniy Dodis, Rafail Ostrovsky, Leonid Reyzin and Adam Smith, "Fuzzy extractors: How to generate strong keys from biometrics and other noisy data," *SIAM Journal on Computing,* Vol.38, No.1, 97–139. 2008.

[2] Roel Maes, "Physically unclonable functions: Constructions, Properties and Applications," Springer. ISBN 978-3-642-41395-7, 2013.

[3] Kenta Takahashi, Takahiro Matsuda, Takao Murakami, Goichiro Hanaoka and Masakatsu Nishigaki, "Signature schemes with a fuzzy private key," in *International Journal of Information Security,* vol. 18 (2019), 581–617.

[4] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li and Wei Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in Proc. of CVPR, 2018.

[5] Jiankang Deng, Jia Guo, Niannan Xue and Stefanos Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in Proc. of CVPR, 2019.

[6] Yuta Yoneyama, Kenta Takahashi and Masakatsu Nishigaki, "Closest Vector Problem on Triangular Lattice and Its Application to Fuzzy Signature," in *IEICE Trans. Fundamentals* (Japanese Edition), Vol.J98-A, No.6, pp.427-435, 2015.

[7] Kaiyi Zhang, Hongrui Cui and Yu Yu, "Facial Template Protection via Lattice-based Fuzzy Extractors," IACR Cryptology ePrint Arch. 2021: 1559, 2021.

[8] Shuichi Katsumata, Takahiro Matsuda, Wataru Nakamura, Kazuma Ohara and Kenta Takahashi, "Revisiting Fuzzy Signatures: Towards a More Risk-Free Cryptographic Authentication System based on Biometrics," CCS 2021: 2046-2065, 2021.

[9] J. H. Conway and N. J. A. Sloane, "Sphere Packings, Lattices and Groups," second ed., Grundlehren der Mathematischen Wissenschaften Band 290, Springer-Verlag, New York, 1993.

[10] I. Dinur, G. Kindler, R. Raz and S. Safra, "Approximating CVP to Within Almost-Polynomial Factors is NP-Hard," in *Combinatorica,* vol. 23, pp. 205–243, 2003.

[11] ISO/IEC 24745:2022, "Information security, cybersecurity and privacy protection – Biometric information protection," Standard. International Organization for Standardization, Geneva, CH 2022.

ComplexFace: A Public Visible-Thermal Face Dataset with Real-Life Complexity

Jiajin He†, Chengxi Dong†, Yunqi Cai†, Dong Wang‡
† Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China
E-mail: Caiyq@kust.edu.cn
‡ Center for Speech, and Language Technologies (CSLT), BNRist at Tsinghua University, Beijing 100084, China
E-mail: wangdong99@mails.tsinghua.edu.cn

***Abstract*—Thermal face recognition (TFR) can address the limitations of traditional visible-light face recognition under complex lighting conditions. However, public thermal datasets are scarce, and the recording conditions of the few available datasets are relatively simple. In this paper, we present a new public thermal face dataset ComplexFace. The key feature of the new dataset is that it involves a wide range of real-life complexity, including variations in scenarios, sessions and devices. Comprehensive experiments were conducted with the new dataset to: (1) discover the most influential factors that impact TFR performance; and (2) discover the actual performance of existing TFR models when real-life complexity is involved. Our experiments demonstrated that variations in scenarios, sessions and devices all pose a significant impact on TFR, and the TFR performance of existing models is much worse than reported in the literature under real-life conditions.**

I. Introduction

With the rapid development and widespread application of biometric technology, thermal face recognition (TFR) has attracted increasing attention. Compared to traditional visible-light face recognition, TFR has unique advantages in complex lighting conditions, making it well-suited for a wide range of applications in security surveillance, healthcare, and military [1], [2].

Despite their great potential, thermal facial images lack texture details compared to visible-light images, resulting in significant challenges for recognition tasks. Various methods have been proposed to improve TFR performance. In the early feature-classifier paradigm, the research primarily focused on discovering robust features such as Local Binary Patterns (LBP) and Weber Local Descriptor (WLD) and others [3]–[5]. In recent years, deep learning methods have become mainstream, e.g., [6]–[8]. A key challenge of deep-learning approaches is that they require a large amount of training data, but thermal face images are far from abundant, partly due to the expense of thermal devices. To solve the problem, some research focuses on cross-modality data augmentation by synthesizing thermal facial images using methods such as visible-to-thermal conversion based on Generative Adversarial Networks (GANs) [9]–[11]. Another approach is domain transfer, which converts thermal images to visible-light images using models such as GANs and Diffusion Models, and then uses large-scale visible-light models to perform face recognition [12]–[15].

Although these methods have achieved some promising progress, limited thermal face data resources remain a bottleneck for both research and application development. Compared to abundant visible-light data, thermal facial data is usually scarce and mostly not publicly available due to factors such as collection costs. Currently, the most commonly used public thermal face datasets are SpeakingFaces, TUFTS, Terravic, IRIS, and Charlotte as shown in Table I. The common issues with these datasets are: (1) they are small in data volume; and (2) their recording conditions are often simple. The lack of large-scale and complex data resources poses a fundamental challenge and prevents training strong TFR models. Ironically, the results reported with these small and simple datasets are often incredibly good (more than 95% in accuracy), indicating a clear overestimation of TFR performance.

To change the situation and boost the research of TFR, we present a complex visible-thermal face dataset denoted as ComplexFace. The key feature of the new dataset is that it encompasses a wide range of real-life complexities, including variations in scenarios, sessions and devices. The purpose of the new dataset is: (1) to provide a rich data resource with sufficient real-life variations, enabling researchers to train more powerful TFR models; and (2) to offer a new TFR benchmark that involves sufficient real-life complexity, thus revealing the actual performance of the TFR system and encouraging researchers to focus on key challenges of the TFR task. ComplexFace is freely available to researchers[1].

II. The ComplexFace Dataset

A. Existing Thermal Datasets

Table I summarizes the main features of existing publicly available datasets. More details are presented below.

- **SpeakingFaces [16]:** A large-scale, publicly available visual, thermal and audio dataset contains data from

Jiajin He and Chengxi Dong have jointly contributed to the experimental work and made equal contributions to this paper. Yunqi Cai and Dong Wang have jointly supervised the experimental design and manuscript writing, both making significant contributions. Therefore, Jiajin He and Chengxi Dong are co-first authors, while Yunqi Cai and Dong Wang are co-corresponding authors.

[1] https://github.com/lovepreeminence/ComplexFace

TABLE I
COMPARISON OF AVAILABLE THERMAL FACE DATASETS

Dataset	Available Thermal Images	Subjects	Highest Resolution	Spectrum	Head Pose	Disguise	Sport	Multi-Session	Multi-Device
SpeakingFaces	7,668	142	464x348	VIS, LWIR	9	–	–	Two weeks	–
TUFTS	1,532	112	336x256	VIS, NIR, LWIR	9	Sunglasses	–	–	–
Terravic	23,333	20	320x240	Thermal	3	Glasses	–	–	–
IRIS	4,228	30	320x240	VIS, LWIR	11	–	–	–	–
Charlotte	10,376	10	640x480	LWIR	25	–	–	–	–
ComplexFace	55,890	70	640X512	VIS, LWIR	9	Mask,Glasses, Sunglasses	✓	Three months	✓

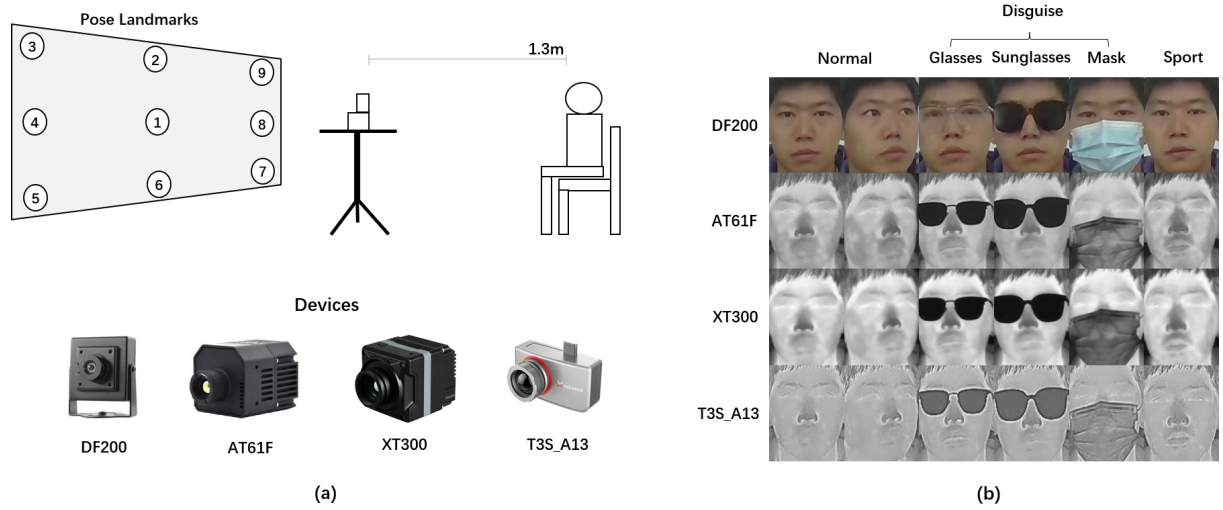


Fig. 1. (a) Illustration of the recording environment of the ComplexFace dataset and the four cameras used to collect the data. (b) Sample facial images recorded with variations in scenarios and devices. Each row represents a particular recording device. The first column shows a person's front face, the second column represents head pose variation, and columns 3-6 demonstrate scenario variation, including disguises (wearing glasses, sunglasses, and a mask) and sport.

142 individuals. Each individual has 54 thermal images, totaling 7,668 images[2]. The images were captured from cameras located in 9 positions. The dataset only includes head pose variations.

- **TUFTS [17]:** This is the most comprehensive large-scale multi-modality face dataset containing seven modalities. The thermal modality consists of 112 people and encompasses 9 head poses from left to right and 5 emotional expressions[3].
- **Terravic [18]:** This dataset includes 20 individuals, and the variations are poses (front, left and right), indoor/outdoor settings and with/without glasses or a hat.
- **IRIS [19]:** This dataset comprises 30 individuals, encompassing 3 different expressions, 5 different lighting angles and varying head poses.
- **Charlotte [20]:** This dataset only contains 10 individuals, involving variations in camera distances, head poses and ambient temperatures[4].

[2]These numbers are for the data that can be downloaded from https://github.com/IS2AI/SpeakingFaces, which are different from the numbers represented in the original paper [16].

[3]https://tdface.ece.tufts.edu/downloads/

[4]https://github.com/TeCSAR-UNCC/UNCC-ThermalFace

TABLE II
RECORDING DEVICES OF COMPLEXFACE

Device	Resolution	Focal length	Pixel Size	Thermal Sensitivity	Spectral Range
DF200	1920x1080	3.6mm	–	–	–
AT61F	640x512	19mm	12μm	<50mK	8∼14μm
XT300	640x480	18mm	12μm	<40mK	8∼14μm
T3S_A13	384x288	13mm	17μm	≤50mK	8∼14μm

B. Data Collection of ComplexFace

1) Recording Device: The ComplexFace dataset was collected using four devices: one visible-light device (DF200) and three thermal devices (Airui AT61F, Sat XT300 and Airui T3S_A13). The details of these devices are shown in Table II. According to the parameters, the quality of AT61F and XT300 is comparable, while T3S_A13 is of lower quality due to its lower resolution and shorter focal length.

2) Recording Process: The entire recording was conducted in two sessions, one in December 2023 and the other in March 2024. All subjects were required to attend the two sessions. The recording was conducted in an office with ceiling lights and natural lighting from outside windows. The four recording devices were set up 1.3 meters in front of the subject and

placed on a tripod. On the wall in front of the subject, there were 9 sticking labels used to guide the subject to change his/her head pose. Fig. 1 (a) shows the recording environment.

Before each recording started, a piece of cardboard was put in front of the cameras. After the subject was seated, we removed the cardboard to signify the starting time of the recording. During the post-processing step, the frame in which the cardboard was completely removed was regarded as the first frame in all four video streams, by which cross-device time synchronization is established.

Each subject was recorded for 5 rounds in each session to simulate different usage scenarios. For each round, the subject was recorded from 9 angles. As seen in Fig. 1 (a), there were 9 positions labeled on the wall, and the subject was asked to look at each position sequentially and stay for several seconds at each position. The 5 rounds of recording were as follows:

1) Round 1: Normal condition, calm and collaborative.
2) Round 2: Disguise by wearing glasses.
3) Round 3: Disguise by wearing sunglasses.
4) Round 4: Disguise by wearing a mask.
5) Round 5: Sport: jump rope 50 times and then return to recording immediately.

The 5 rounds of recordings were categorized into three scenarios: **Normal (Round 1), Disguise (Rounds 2-4), and Sport (Round 5)**. Fig. 1 (b) shows the sample images recorded in the three scenarios.

3) Face image selection: Ideally, 70 individuals participated in two collection sessions, providing three face images for each of the 9 head poses, and there are 5 rounds of recording to reflect scenario variation and 4 cameras representing device variation. This results in a total of $70 \times 9 \times 3 \times 5 \times 4 \times 2 = 75{,}600$ images. However, in the second session, two individuals could not participate for personal reasons, and the sunglasses scenario was not fully recorded for subject 31. Taking all the accidents into account, the final ComplexFace dataset contains a total of 74,412 images.

C. Features of ComplexFace

Compared to the existing databases, ComplexFace is more comprehensive. It contains 55,890 thermal face images from 70 individuals, making it the largest publicly available dataset for TFR so far in terms of the number of thermal images. The most significant feature of ComplexFace is that it involves complex variations that are often encountered in real-life applications. These variations include:

- **Device variation:** ComplexFace was recorded with four cameras simultaneously: one visible-light and three thermal. This multi-device feature is valuable for: (1) Comparison between visible-light and thermal FR; (2) Investigation for cross-device thermal FR; (3) Investigation for visible-thermal cross-modality FR [21]–[23]; and (4) Training and testing cross-modality generative models. In this study, we are mostly interested in the cross-device degradation for TFR. Note that ComplexFace is the only dataset available so far that contains simultaneous multi-device thermal face images.
- **Session variation:** ComplexFace involves face images of the same subject but recorded at different times, allowing for the investigation of session variation. This type of variation can be attributed to both subjective and objective factors. The former includes hairstyles, micro-expressions, sitting styles, etc., and the latter includes environmental temperatures, background lights, randomness in recording devices, and so on. The variation could be more evident with longer recording intervals. For existing datasets, only SpeakingFaces involves multiple sessions, and the recording interval is two weeks. For ComplexFace, the recording interval is three months, thus leading to more significant session variation. Note that it is difficult to alleviate session variation by the recorder's carefulness and the subjects' attention.
- **Scenario Variation:** Scenario variation refers to changes in a subject's appearance under different conditions or statuses. ComplexFace considers a Normal scenario and two uncommon scenarios: Disguise and Sport. Face disguise involves various types of facial occlusion, which often leads to significant accuracy reduction for TFR. In ComplexFace, three types of disguises commonly seen in real applications were recorded: wearing glasses, sunglasses, and a mask. Thermal faces tend to change when the temperature on the face increases due to varied physiological status of the subject, such as fever, shame, and sport. To investigate the impact of this factor, ComplexFace includes faces after sports. None of the existing datasets consider such complex scenario variation.

III. EXPERIMENTS

In this section, we conduct a series of TFR experiments using ComplexFace. The main purpose is two-fold: (1) to investigate the impact of different variations on TFR performance; and (2) to reveal the true performance of TFR models when real-life complexity is considered.

A. Setting

To clarify the data used in the experiments, we denote the data subset by the format 'Device/Session/Scenario', where 'Device' could be A: AT61F, X: XT300, U: T3S_A13, 'Session' could be S1 (Session 1) or S2 (Session 2), and 'Scenario' could be Normal, Disguise or Sport. For instance, data subset A/S1/Normal means thermal faces recorded by AT61F in the first session, and the scenario is Normal.

In the following experiments, S1 was divided into a training set of 49 people and a test set of 21 people, while S2 was divided into a training set of 48 people and a test set of 20 people. Since the cross-session experiment requires test subjects to appear in both S1 and S2, we removed the subject who is present in S1 but not in S2 from the *test set*.

During testing, we randomly selected 10 thermal images from the Normal scenario in the test set for enrollment and used the remaining for testing[5]. The mean vector of the em-

[5]We assume that the scenario for enrollment is mostly Normal, so only test this setting.

TABLE III
CROSS-SCENARIO RESULTS WITH THE S1 SUBSET FROM AT61F

Training set	Enroll set	Test set	Acc(%)	EER(%)
A/S1	A/S1/Normal	A/S1/Normal	**93.532(3.691)**	**4.425(0.703)**
		A/S1/Disguise	70.422(5.668)	12.535(0.999)
		A/S1/Sport	81.118(9.160)	8.515(2.309)
		A/S1/All	77.748(3.761)	10.510(0.795)

beddings from the enrollment images served as the enrollment vector, and the cosine similarity between the enrollment vector and the test embedding was used as the matching score. We evaluated the system performance using two metrics: Accuracy (Acc) and Equal Error Rate (EER). Accuracy is the proportion of correctly assigning a test face to the enrolled subjects according to the matching score. EER is calculated by setting a threshold on the matching score so that the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal, and the EER is the FAR/FRR value at this balance point. The trial list was created by cross-pairing enrolled individuals with all test thermal images.

The architecture of the recognition model used in the experiments was FaceNet [24]. The model was inherited from VGGFace2 [25] and fine-tuned using the training data of ComplexFace. The training process involved 15 epochs with a learning rate of 0.0001 and a batch size of 32. To mitigate the effect of model instability caused by training with limited data, the training process was repeated 10 times with different random initialization parameters. We report the mean values and variances of the Acc and EER obtained from the 10 replicates.

B. Cross-Scenario Experiment

The first experiment tests the impact of scenario variation, in particular the scenario shift between enrollment and test that is often observed in real-life applications. Here we used A/S1, i.e., images from S1 recorded by AT61F, to perform the test, thereby the session and device variations were eliminated. The model was trained on data from all three scenarios. During testing, the enrollment used 10 images from the Normal scenario, while the test was performed on all the remaining images.

Results in Table III show that the model performs much worse when tested on the Disguise and Sport scenarios compared to the Normal scenario. This is understandable because disguises such as glasses, sunglasses, and masks obscure key facial information. The recognition performance after sport is also lower than before sport, as the increased temperature after sport interferes with thermal imaging. These results demonstrate that scenario variation indeed poses a significant challenge for TFR.

C. Cross-Session Experiment

This section investigates the impact of session variation. To eliminate the impact of device variation, we used the data from AT61F only. Two models were trained, one with the A/S1 subset, and the other with both A/S1 and A/S2. The enrollment was based on A/S1/Normal as in the previous experiment, while the test could be within-session or cross-session.

TABLE IV
CROSS-SESSION RESULTS WITH DATA FROM AT61F

Training set	Enroll set	Test set	Acc(%)	EER(%)
A/S1	A/S1/Normal	A/S1/All	**78.759(3.426)**	**10.308(0.809)**
		A/S2/All	33.391(8.610)	25.415(2.715)
A/[S1+S2]	A/S1/Normal	A/S1/All	**76.324(8.241)**	**10.716(2.284)**
		A/S2/All	35.816(6.181)	24.387(2.826)

TABLE V
CROSS-DEVICE RESULTS WITH S1 SUBSET FROM THREE DEVICES: A: AIRUI AT61F, X: SAT XT300, U: AIRUI T3S_A13.

Training set	Enroll set	Test set	Acc(%)	EER(%)
A/S1	A/S1/Normal	A/S1/All	**77.748(3.761)**	**10.510(0.795)**
		X/S1/All	65.627(7.449)	14.209(3.093)
		U/S1/All	37.935(7.343)	24.148(2.328)
[A+X+U]/S1	A/S1/Normal	A/S1/All	**82.386(3.221)**	**8.270(0.644)**
		X/S1/All	78.578(4.764)	9.081(0.877)
		U/S1/All	64.336(7.640)	13.584(1.304)
	X/S1/Normal	A/S1/All	81.619(3.635)	8.792(0.768)
		X/S1/All	**81.679(2.376)**	**8.575(0.688)**
		U/S1/All	65.059(11.551)	13.718(0.936)
	U/S1/Normal	A/S1/All	**75.222(11.037)**	**11.946(1.178)**
		X/S1/All	72.917(12.917)	12.723(1.217)
		U/S1/All	74.716(8.106)	12.510(1.617)

The results are shown in Table IV. Firstly, focusing on the A/S1 model, it can be seen that when enrollment and testing were within the same session, i.e., both from A/S1, the model performed very well. However, when the enrollment images were from A/S1 and the test images were from A/S2, the model's performance dramatically declined. This indicates that session variation poses a very significant impact.

More information can be obtained by looking at the A/[S1+S2] model. Firstly we can observe the same trend as with the A/S1 model: cross-session leads to a severe performance drop. The most important observation is that the *cross-session performance was not clearly improved by using cross-session data in model training*. This indicates that the impact of session variation is due to the enrollment-test mismatch, and cannot be solved by multi-session and cross-session training. This in turn implies that the **underlying problem of session variation is individual behavior change, rather than global distribution shift**[6].

D. Cross-Device Experiment

This section explores the impact of device variation. To eliminate the impact of session variation, we used the data from S1 only. Two models were trained: model A/S1 was trained with the data from AT61F, and model [A+X+U]/S1 was trained with the data from all three thermal devices. As in previous experiments, enrollment was conducted with images from the Normal scenario, while the test involved images from all three scenarios.

[6]If the variation is caused by global distribution shift, the S1 + S2 model would have compensated for the shift by cross-session training.

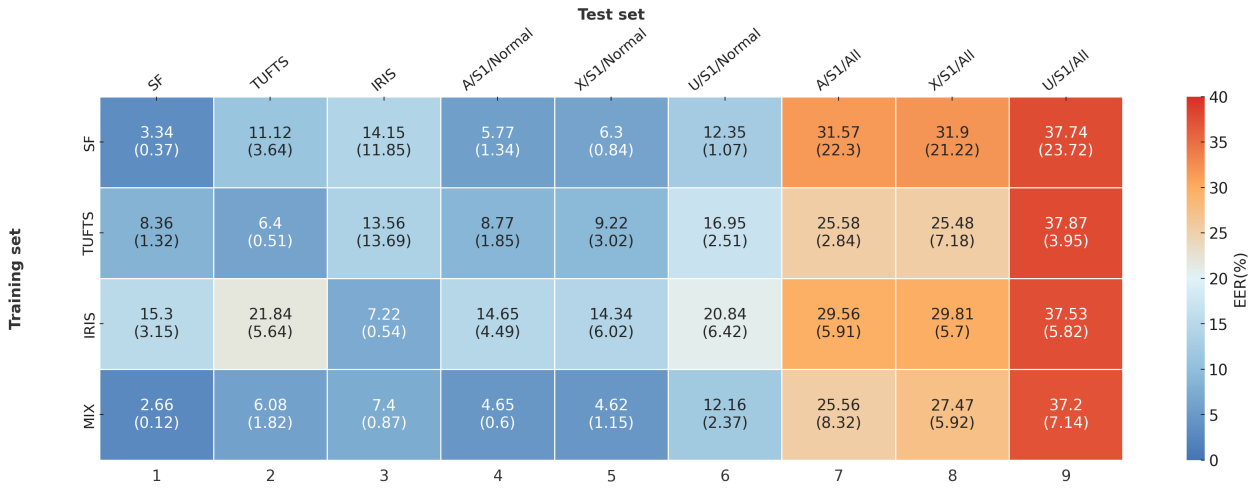


Fig. 2. Cross-dataset Results. Models were trained with three external datasets (rows 1-3) and their pooling (row 4), and were tested on their own test sets (columns 1-3) as well as the test sets of ComplexFace. The testing on ComplexFace was done in two conditions: without scenario variation (columns 4-6) and with scenario variation (columns 7-9).

The results are shown in Table V. Firstly, focusing on the AT61F model A/S1, it is clear that the model's performance dropped significantly in cross-device testing. This indicates that data characteristics vary greatly between the three devices, and the model's generalization ability across devices is weak. Differences in data characteristics between devices (e.g., resolution, focal length) are the main reasons for the decline in cross-device performance. Note that the performance drop is more significant when tested on T3S_A13 (U), indicating that the data characteristics of T3S_A13 (U) differ from AT61F (A) more significantly than XT300 (X).

Further examining the All-devices model [A+X+U]/S1. Firstly we found that it performed much better than the single-device model A/S1. If we pay attention to the results with A/S1/Normal as the enrollment data, one can see that for all three test sets, the performance was clearly improved compared to the results with the A/S1 model. This improvement is more evident in the other cross-device test. This suggests that cross-device discrepancy can be effectively mitigated by cross-device training[7]. This implies that **device variation is more or less a global distribution shift**, which contrasts with session variation.

E. Cross-dataset Experiment

In this section, we use ComplexFace to test models trained with three public datasets: SpeakingFaces (100 for training and 42 for testing), TUFTS (80 for training and 32 for testing) and IRIS (20 for training and 10 for testing). We also combined the training data of all three datasets to train a MIX model.

Firstly, we evaluated the performance on the corresponding SpeakingFaces, TUFTS and IRIS test sets (columns 1-3). Next, we tested the four models on the ComplexFace dataset, using the test sets of the three thermal devices from S1. The enrollment used the Normal Scenario (A/S1/Normal, X/S1/Normal and U/S1/Normal), and the test was on both the Normal scenario (columns 4-6) and the All scenario (columns 7-9).

The experimental results in terms of EER are shown in Figure 2. Models perform best when the training and test sets come from the same data source (e.g., results in cells 1-1, 2-2, 3-3), indicating better performance on homogeneous data. The MIX model (row 4) often results in lower EERs across different test sets, suggesting that collecting more diverse data improves the models' generalization ability.

Another key observation is that for tests on ComplexFace, performance under the 'ideal' condition (columns 4-6) that is free of variations in scenarios, sessions and devices, is quite good, comparable to results obtained with the ComplexFace model (Table III). This indicates that models can perform well with simple test data. However, performance significantly drops when tested under more complex conditions with scenario variation (columns 7-9), and is much worse than the results obtained with the ComplexFace model (Table V).

The first conclusion from these observations is that simple tests cannot differentiate between weak and strong models; stringent test conditions are required to evaluate true performance. Furthermore, to improve the quality of the TFR model, complex training data is necessary. Finally, the EER results from the four models are very poor when scenario variation is involved, approximately 37%. These high EER numbers indicate that the models are practically useless, despite their promising performance on their own test sets. This means that there is a significant gap between reported performance in the literature and real-world test results, demonstrating that existing TFR research is far from satisfactory and still has a long way to go before this technology becomes genuinely usable.

[7]The cross-device training means that during model training, the face images of the same subject but recorded by different devices are labeled as the same person.

IV. CONCLUSION

We present a publicly available dataset, ComplexFace, for thermal face recognition (TFR) research. This dataset contains over 70,000 face images from 70 individuals, encompassing complex real-life variations in scenarios, sessions and devices. Comprehensive experiments were conducted with ComplexFace and three public datasets: SpeakingFaces, TUFTS, and IRIS.

The first observation is that session variation is the most influential factor and is difficult to address even with cross-session training. Scenario variation is the second most influential factor, and cross-scenario training can partially alleviate its impact. Finally, device variation is also very influential but can be effectively addressed by cross-device training. The second observation is that models trained with existing public datasets perform well on ComplexFace under simple test conditions. However, when real-life complexity (e.g., scenario variation) is involved, all tested models become ineffective. This result demonstrates that the good performance reported in the literature is significantly overestimated, and true TFR performance should be evaluated using ComplexFace or other datasets with real-life complexity.

ACKNOWLEDGMENT

This work was supported by Yunnan Fundamental Research Projects (grant NO. 202301AT070274)

REFERENCES

[1] A. Wilson, K. A. Gupta, B. H. Koduru, A. Kumar, A. Jha, and L. R. Cenkeramaddi, "Recent advances in thermal imaging and its applications using machine learning: A review," *IEEE Sensors Journal*, vol. 23, no. 4, pp. 3395–3407, 2023.

[2] D. Mahouachi and M. A. Akhloufi, "Recent advances in infrared face analysis and recognition with deep learning," *AI*, vol. 4, no. 1, pp. 199–233, 2023.

[3] G. Hermosilla, J. Ruiz-del-Solar, R. Verschae, and M. Correa, "A comparative study of thermal face recognition methods in unconstrained environments," *Pattern Recognition*, vol. 45, no. 7, pp. 2445–2459, 2012.

[4] Y. Bi, M. Lv, Y. Wei, N. Guan, and W. Yi, "Multi-feature fusion for thermal face recognition," *Infrared Physics & Technology*, vol. 77, pp. 366–374, 2016.

[5] S. Kumar, S. K. Singh, and P. Peer, "Occluded thermal face recognition using bocnn and radial derivative gaussian feature descriptor," *Image and Vision Computing*, vol. 132, p. 104 646, 2023.

[6] P. Wang and X. Bai, "Regional parallel structure based cnn for thermal infrared face identification," *Integrated Computer-Aided Engineering*, vol. 25, no. 3, pp. 247–260, 2018.

[7] S. Kumar and S. K. Singh, "Occluded thermal face recognition using bag of cnn (Bo cnn)," *IEEE Signal Processing Letters*, vol. 27, pp. 975–979, 2020.

[8] N. Zaeri and R. Qasim, "Thermal image identification against pose and expression variations using deep learning," *Journal of Engineering Research*, 2023.

[9] G. Hermosilla, D.-I. H. Tapia, H. Allende-Cid, G. F. Castro, and E. Vera, "Thermal face generation using stylegan," *IEEE Access*, vol. 9, pp. 80 511–80 523, 2021.

[10] X. Cao, K. Lai, G.-S. J. Hsu, M. Smith, and S. N. Yanushkevich, "Cross-spectrum thermal face pattern generator," *IEEE Access*, vol. 10, pp. 9576–9586, 2022.

[11] C. Ordun, E. Raff, and S. Purushotham, "When visible-to-thermal facial gan beats conditional diffusion," in *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 181–185.

[12] Y. Luo, D. Pi, Y. Pan, L. Xie, W. Yu, and Y. Liu, "Clawgan: Claw connection-based generative adversarial networks for facial image translation in thermal to rgb visible light," *Expert Systems with Applications*, vol. 191, p. 116 269, 2022.

[13] N. K. Yadav, S. K. Singh, and S. R. Dubey, "Tva-gan: Attention guided generative adversarial network for thermal to visible image transformations," *Neural Computing and Applications*, vol. 35, no. 27, pp. 19 729–19 749, 2023.

[14] N. G. Nair and V. M. Patel, "T2v-ddpm: Thermal to visible face translation using denoising diffusion probabilistic models," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, IEEE, 2023, pp. 1–7.

[15] J. Lin, G. Zhao, J. Xu, *et al.*, "DiffTV: Identity-preserved thermal-to-visible face translation via feature alignment and dual-stage conditions," in *ACM Multimedia 2024*, 2024.

[16] M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, Y. Khassanov, M. Lewis, and H. A. Varol, "Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams," *Sensors*, vol. 21, no. 10, p. 3465, 2021.

[17] K. Panetta, Q. Wan, S. Agaian, *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 509–520, 2018.

[18] *Terravic facial infrared database*. [Online]. Available: http://vcipl-okstate.org/pbvs/bench/Data/04/download.html.

[19] *Iris thermal/visible face database*. [Online]. Available: http://vcipl-okstate.org/pbvs/bench/Data/02/download.html.

[20] R. Ashrafi, M. Azarbayjani, and H. Tabkhi, "Charlotte-thermalface: A fully annotated thermal infrared face dataset with various environmental conditions and distances," *Infrared Physics & Technology*, vol. 124, p. 104 209, 2022.

[21] C. N. Fondje, S. Hu, and B. S. Riggan, "Learning domain and pose invariance for thermal-to-visible face recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 1, pp. 15–28, 2022.

[22] K. Nikhal, N. Uzuegbunam, B. Kennedy, and B. S. Riggan, "Mitigating catastrophic interference using unsupervised multi-part attention for rgb-ir face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 335–344.

[23] A. George and S. Marcel, "Heterogeneous face recognition using domain invariant units," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 4780–4784.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[25] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, IEEE, 2018, pp. 67–74.

Multibiometrics Using a Single Face Image

Koichi Ito*, Taito Tonosaki*, Takafumi Aoki*, Tetsushi Ohki†, and Masakatsu Nishigaki†
* Graduate School of Information Sciences, Tohoku University, Japan
E-mail: ito@aoki.ecei.tohoku.ac.jp
† Faculty of Informatics, Shizuoka University, Japan

***Abstract*—Multibiometrics, which uses multiple biometric traits to improve recognition performance instead of using only one biometric trait to authenticate individuals, has been investigated. Previous studies have combined individually acquired biometric traits or have not fully considered the convenience of the system. Focusing on a single face image, we propose a novel multibiometric method that combines five biometric traits, i.e., face, iris, periocular, nose, eyebrow, that can be extracted from a single face image. The proposed method does not sacrifice the convenience of biometrics since only a single face image is used as input. Through a variety of experiments using the CASIA Iris Distance database, we demonstrate the effectiveness of the proposed multibiometrics method.**

I. Introduction

Biometric recognition is a technology that identifies individuals using physical characteristics such as face, iris, and fingerprints, or behavioral characteristics such as gait and handwriting [1]. Biometric recognition is reliable and convenient compared to passwords and personal identification numbers, however, environmental changes such as brightness and pose may significantly degrade the recognition performance. Therefore, multibiometrics, which uses multiple biometric traits to improve recognition performance instead of using only one biometric trait, has been investigated [2].

Multibiometrics improves recognition performance with more biometric traits to be combined. For example, the combination of face, fingerprint, and iris recognition has been investigated [3]. Face recognition is highly convenient since the user only needs to turn his/her face to the camera to be authenticated, while the performance of face recognition is degraded by aging, cosmetics, and changes in the face pose. Fingerprint and iris recognition require dedicated sensors, while their performance is more stable than that of face recognition. Combining face, fingerprint, and iris recognition can improve the performance and stability of user authentication, however, the convenience is significantly reduced because of acquiring face, fingerprint, and iris at the same time. In addition, the cost of the system increases since it ususally requires a camera for capturing the face, a sensor for capturing the fingerprint, and a dedicated camera for capturing the iris. Other multibiometrics that combine iris and periocular recognition have been investigated [4]. Periocular recognition uses images of the eye and its surrounding area, and therefore the same camera can be used to capture both periocular and iris images. Iris recognition has high performance, however, its performance depends on the accuracy of iris segmentation. Periocular recognition has stable performance because of the use of periocular images, however, its performance is degraded by environmental changes. Combining iris and periocular recognition improves the performance and stability due to their complementarity. On the other hand, the distance between the user and the system may be limited due to the need to capture the periocular images. Previous studies described above have combined individually acquired biometric traits or have not fully considered the convenience of the system.

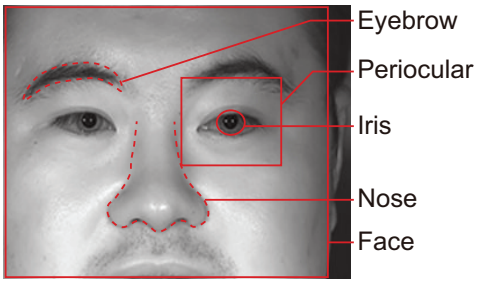


Fig. 1. Biometric traits extracted from a single face image, whose combination is considered in this paper.

Focusing on a single face image, we investigate a novel multibiometric method that combines five biometric traits, i.e., face, iris, periocular, nose, and eyebrow, that can be extracted from a single face image, as shown in Fig. 1. The proposed method combines multiple biometric traits without sacrificing the convenience of biometrics since only a single face image is used as input. To the best of our knowledge, there are no studies of multibiometrics, in which five biometric features are extracted from a single image. In the proposed method, we first detect keypoints of a face using Mediapipe FaceMesh[1] and obtain an image of each biometric trait based on these keypoints. Next, we extract features using Convolutional Neural Network (CNN) and calculate the matching score based on features for each trait. Then, we obtain the final matching score by fusing scores with a weighted sum. Among the biometric databases available so far, only the CASIA Iris Distance database[2] has a sufficiently large image size to extract images from face to iris. Through a variety of experiments using the CASIA Iris Distance database, we demonstrate the effectiveness of the proposed multibiometrics method.

[1]https://google.github.io/mediapipe
[2]http://biometrics.idealtest.org

II. Related Work

The study of multibiometrics requires databases that consist of multiple biometric traits. In the early days, XM2VTS [5] with face and voice and MCYT [6] with fingerprint and signature have been usually used for the study. NIST BSSR1 (Biometric Scores Set - Release 1)[3], which consists of matching scores obtained from fingerprint and face recognition, has been also available. In some cases, databases of a single biometric trait have been combined to create a pseudo multibiometric database for evaluation. These databases are based on the assumption that biometric traits are acquired by multiple sensors, which sacrifices the convenience of biometrics even if it improves the recognition accuracy. At that time, most studies have focused on investigating score fusion methods that improve recognition accuracy [2].

Recently, methods for multibiometrics using CNN have been proposed. Stokkenes et al. proposed to combine face and periocular recognition at the score level [7]. Telgad et al. proposed to combine three biometric features, i.e., face, fingerprint, and iris, at the score level [3]. Zhang et al. proposed to combine iris and periocular recognition using images captured by mobile devices [8]. Wang et al. proposed to combine UniNet for iris recognition and AttNet for periocular recognition [4]. Ng et al. proposed CMB-Net, which combines face and periocular features [9]. In addition, there are other methods that combine the face, left iris and right iris [10], and the periocular and iris [11], [12]. Since CNN can extract features with high discriminative performance and has improved the recognition accuracy of a single biometric trait, multibiometrics using CNN has also improved its recognition accuracy. However, they only combine two or three biometric traits, or require multiple sensors.

III. Multibiometrics Using a Single Face Image

We propose a multibiometric method that combines biometric traits extracted from a single face image: face, periocular, iris, nose, and eyebrow. Fig. 2 shows an overview of the proposed method. The proposed method consists of extraction of each biometric trait from a face image, feature extraction using CNN and calculation of a matching score for each trait, and score fusion by a weighted sum. In the following, we describe the details of each process.

A. Biometric Trait Extraction

This section describes the process of extracting biometric traits from face images. We first detect keypoints on the face image using Mediapipe FaceMesh[1]. Fig. 3 shows an example of the detected keypoints. The proposed method uses the keypoints at the center of both eyes, the center of the nose, and the center of the left eyebrow among these keypoints. Next, the rotation of the face image is normalized so that the line segment connecting the keypoints of both eyes is horizontal, and the normalized image is used as the face image for face recognition. The center of the left eye is used to crop a rectangular region to obtain the periocular image used for periocular recognition. Similarly, the center of the nose and the center of the left eyebrow are used to obtain the nose and eyebrow images for nose and eyebrow recognition, respectively. The iris images are extracted from the periocular image using the method proposed by kawakami et al [13] as shown in Fig. 4. The details of the process are omitted due to space limitation. For more details, refer to [13]. From the above process we obtain a face image, a periocular image, four normalized iris and mask images, a nose image, and an eyebrow image.

[3]https://www.nist.gov/itl/iad/image-group/nist-biometric-scores-set-bssr1

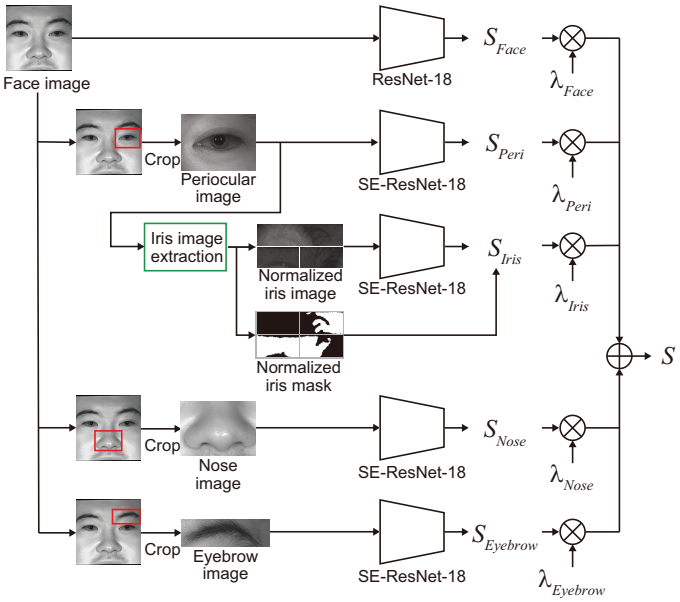


Fig. 2. Overview of the proposed multibiometrics approach using a single face image.

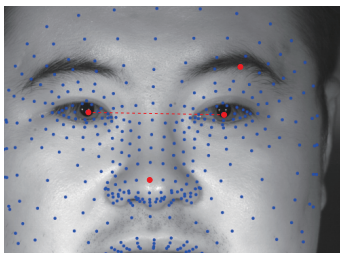


Fig. 3. Keypoints extracted using Mediapipe FaceMesh. The proposed method uses the center of both eyes, the center of the nose, and the center of the left eyebrow among these key points, which are indicated by red-colored points.

B. Feature Extraction Using CNN

We describe CNN for feature extraction from each biometric trait. As for face recognition, we use ResNet-18 [14] with ArcFace [15] as shown in Fig. 5, which is a popular approach in face recognition and achieves high recognition accuracy. As for other biometric traits, we use SE-ResNet-18 [14], [16] as shown in Fig. 6. SE-ResNet-18 is a CNN model that introduces a Squeeze-and-Excitation block (SE block), which is an attention mechanism, into ResNet. In SE block, feature

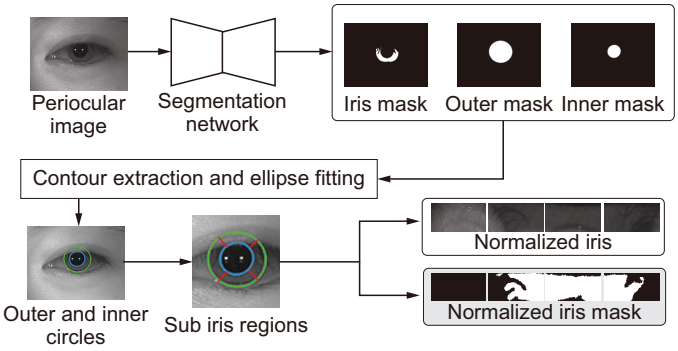


Fig. 4. Flow of iris extraction from a periocular image [13].

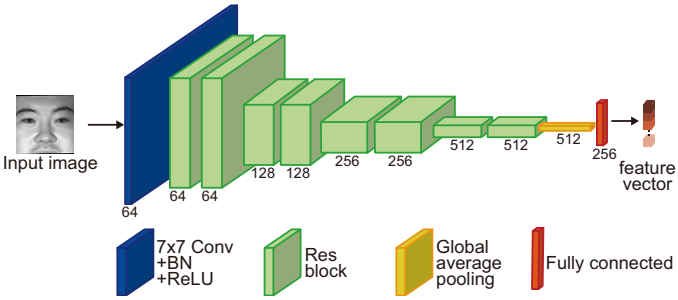


Fig. 5. Network architecture of ResNet-18 for face recognition in the proposed method.

maps output from the convolution layer are averaged in the channel direction and aggregated into a vector, and then passed through two fully-connected layers and activation functions (ReLU and Sigmoid) to convert it into a vector that represents the weights for each channel. The weight vectors are then multiplied by the feature map to obtain a refined feature map. The Cross Entropy (CE) loss is used as the loss function for SE-ResNet-18 in the proposed method. Note that for the iris image consisting of four subimages, we employ the method proposed by Kawakami et al. [13] for feature extraction.

C. Score Fusion

We describe a fusion approach of the matching scores obtained from each biometric trait. In the proposed method, the matching score is calculated by the cosine similarity between the features extracted from each biometric trait using CNN. As for iris recognition, we use a weighted average of the matching score of each subimage, where each weight is calculated based on their mask ratio, resulting in a score that gives more weight to subimages with less occlusion.

The final matching score is calculated by a weighed sum of all the matching score as follows

$$\begin{aligned} S &= \lambda_{face} S_{face} + \lambda_{peri} S_{peri} + \lambda_{iris} S_{iris} \\ &\quad + \lambda_{nose} S_{nose} + \lambda_{eyebrow} S_{eyebrow}, \end{aligned} \tag{1}$$

where S_* is the matching score for each biometric trait, and λ_* is the weight of each, which satisfies the following equation

$$\lambda_{face} + \lambda_{peri} + \lambda_{iris} + \lambda_{nose} + \lambda_{eyebrow} = 1. \tag{2}$$

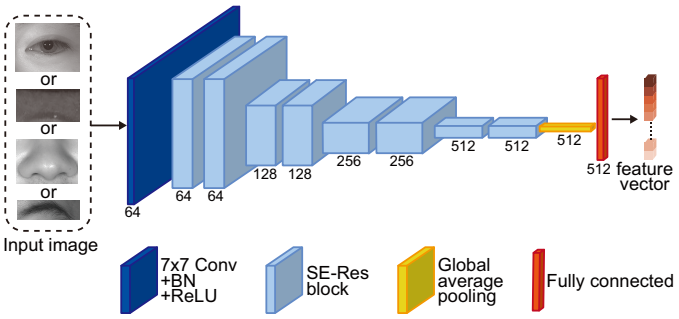


Fig. 6. Network architecture of SE-ResNet-18 for periocular, iris, nose, and eyebrow recognition in the proposed method.

IV. Experiments and Discussion

In this section, we describe the evaluation experiments of the proposed method using the CASIA Iris Distance database[2]. In this experiment, we evaluate the recognition accuracy of each biometric trait alone and combinations of traits. For the combination, we compare the proposed method with feature-level fusion in addition to the score-level fusion described above to demonstrate the effectiveness of the proposed method.

A. Dataset

In this experiment, we use the CASIA Iris Distance database[2]. This database consists of images taken of the upper part of faces of the subjects with a near-infrared camera from a distance of 2.4m∼3.0m, and contains a total of 2,567 images taking from 142 faces. Most of the images are blurred and do not show the iris clearly, making iris recognition particularly difficult. In the experiments, we divide this database into three subsets: training, validation, and test, and use them to evaluate the recognition accuracy of single biometrics and multibiometrics. The training data consists of 84 subjects (1,531 images) whose IDs are from S4000 to S4083, the validation data consists of 28 subjects (497 images) whose IDs are from S4084 to S4111, and the test data consists of 30 subjects (513 images) whose IDs are from S4112 to S4141.

B. Evaluation Metrics

We describe the evaluation metrics used in the experiments. False Rejection Rate (FRR) is calculated from the matching score of genuine pairs and False Acceptance Rate (FAR) is calculated from the matching score of impostor pairs. We use Equal Error Rate (EER), which is equal to FRR and FAR. In practical authentication systems that require high security, such as immigration control systems, it is necessary to operate the system so that the FAR is low. Therefore, we use FRR@FAR0.1% and FRR@FAR0.01%, which are FRR when the threshold is set so that FAR is 0.1% and 0.01%, respectively. The smaller the value of EER and FRRs, the higher the recognition accuracy.

C. Training Setting

We describe the training settings of the network used in the proposed method. ResNet-18 and SE-ResNet-18 are pre-

TABLE I
RESULTS FOR A SINGLE BIOMETRIC TRAIT WITH DIFFERENT COMBINATION OF CNN AND ITS LOSS FUNCTION, WHERE R AND S FOR CNN INDICATE RESNET-18 AND SE-RESNET-18, RESPECTIVELY, AND A AND C FOR LOSS FUNCTIONS INDICATE ARCFACE AND CE, RESPECTIVELY. THE UNDERLINES ARE THE PROPOSED COMBINATIONS, AND THE BOLD NUMBERS INDICATE THE BEST FOR EACH BIOMETRIC TRAIT. THE UNITS FOR EER AND FRR ARE PERCENTAGES.

Trait	CNN + Loss	EER	FRR	
			@FAR0.1	@FAR0.01
Face	R+A	5.636	**20.479**	**33.033**
	S+C	5.386	26.792	40.484
	S+A	**5.290**	44.542	53.180
Periocular	R+A	**4.747**	26.507	35.074
	S+C	4.836	**19.839**	**33.721**
	S+A	6.415	31.433	41.291
Iris	R+A	11.618	53.654	82.223
	S+C	**9.095**	**36.047**	**49.241**
	S+A	18.916	83.626	96.156
Nose	R+A	2.804	**9.611**	**16.255**
	S+C	**2.719**	12.838	24.608
	S+A	4.124	13.242	18.201
Eyebrow	R+A	3.316	**15.116**	**26.602**
	S+C	**2.642**	16.137	27.171
	S+A	3.654	20.812	34.362

trained using ImageNet and made fine-tuning using the training data of CASIA Iris Distance. The initial value of the learning rate is set to 0.0001, and the learning rate is multiplied by 0.1 if the loss to the validation data does not improve for 10 consecutive epochs. If the loss to the validation data does not improve for 30 consecutive epochs, training is terminated and the test data is evaluated using the weights from the epoch with the lowest loss. The batch size is set to 4 and Adam [17] is used as optimizer. When training CNN for periocular recognition, random crop, color jitter, random erasing, and random perspective are used for data augmentation and when training CNN for iris recognition, random erasing and random horizontal shift are used for data augmentation. The random horizontal shift is a circular shift within a range of 10 pixels on the left and right sides, respectively. All the data augmentation methods are applied with a probability of 50%. The segmentation network used for iris region extraction in Fig. 4 is pre-trained using a subset of CASIA Iris Distance created by [11]. This subset is annotated with the ground truth labels of the iris, outer, and inner masks for iris segmentation. The iris region is extracted by using the rectangle normalization module of OSIRIS [18].

D. Single Biometrics

First, we evaluate the recognition accuracy of the proposed method on a single biometric trait. To confirm the effectiveness of the combination of CNN and loss function employed in the proposed method, we use ResNet-18 and SE-ResNet-18 as CNNs, and ArcFace and CE as loss functions. Since ResNet-18 has been demonstrated to be effective in face recognition [15], we use only ArcFace as the loss function. Table I shows the experimental results for a single biometric trait. When EER is selected as the first priority, SE-ResNet-18+CE has the highest recognition accuracy for iris, nose, and eyebrow recognition. For face recognition, SE-ResNet-18+ArcFace has the lowest EER, however, we select ResNet-18+ArcFace, which has almost the same EER and lower FRRs. For periocular recognition, we use SE-ResNet-18+CE instead of ResNet-18+ArcFace, as is the case with face. As mentioned above, EERs for iris recognition are higher than those of the other methods because of the blurred images. Face recognition has high EERs since the whole face is not captured in the image. On the other hand, nose and eyebrow recognition are more accurate. This is because they are extracted more stably than other biometric traits, and they have sufficient individuality.

E. Multibiometrics

Next, we evaluate the recognition accuracy of the proposed method on two to five biometric traits. To confirm the effectiveness of the score-level fusion, we compare its performance with feature-level fusion. In feature-level fusion, we concatenate the features extracted from each biometric trait and produce a 1,024 dimensional feature through two fully-connected layers. Table II shows the experimental results for the proposed method with score-level fusion using two to five biometric traits. The best combination for two biometric traits is periocular and nose, which improves EER up to 0.793%. In particular, the lowest FRR@FAR0.1% for a single biometric trait is 9.611%, while the combination of two improves it to 3.109%. The best combination of three is periocular, nose, and eyebrow, which improves EER up to 0.463%, FRR@FAR0.1% up to 1.590% and FRR@FAR0.01% up to 4.105%. The best combination of the four is to exclude face, which improves EER up to 0.337%, FRR@FAR0.1% up to 1.495%, and FRR@FAR0.01% up to 3.251%. The combination of all five gives the same result as the combination of the four, since the weight for face is 0. The interesting point is that the use of facial parts with individual characteristics, such as the periocular, nose, and eyebrow, is sufficient to recognize a person without the use of the entire face. Table III shows the experimental results for the proposed method with feature-level fusion using two to five biometric traits. By increasing the number of biometric traits combined, the recognition accuracy does not necessarily improve, and even in the case of the best improvement, EER is 1.397%, which is higher than that of score-level fusion. There is a possibility that the accuracy can be improved to the same level as that of score-level fusion by modifying the approach of feature-level fusion used in this experiment, however, we consider that it is not cost-effective compared to score-level fusion.

F. Comparison with Conventional Methods

There is no method that combines the five biometric traits from a face image as the proposed method. Therefore, we compare the proposed method with the methods used to evaluate multibiometrics in the CASIA Iris Distance database. Note that the experimental conditions are different from those of the conventional methods. In particular, for iris and periocular recognition, the right eye is flipped to the left eye, i.e., the number of data is twice. Table IV shows the comparison

TABLE II
RESULTS OF MULTIBIOMETRICS USING SCORE-LEVEL FUSION, WHERE THE BOLD NUMBERS INDICATE THE BEST FOR EACH COMBINATION.

Trait					EER [%]	FRR [%]	
Face	Periocular	Iris	Nose	Eyebrow		@FAR0.1%	@FAR0.01%
0.5	0.5	—	—	—	1.760	11.319	17.062
0.5	—	0.5	—	—	3.027	14.713	25.344
0.4	—	—	0.6	—	1.295	6.004	11.153
0.4	—	—	—	0.6	2.231	7.143	15.567
—	0.6	0.4	—	—	3.528	13.645	18.296
—	0.4	—	0.6	—	**0.793**	**3.109**	7.546
—	0.5	—	—	0.5	1.757	5.363	12.174
—	—	0.4	0.6	—	1.563	6.407	10.631
—	—	0.5	—	0.5	1.699	7.000	16.327
—	—	—	0.6	0.4	1.019	3.227	**5.482**
0.4	0.3	0.3	—	—	1.431	9.729	15.496
0.1	0.3	—	0.6	—	0.696	2.871	6.692
0.3	0.5	—	—	0.2	1.384	6.478	12.340
0.3	—	0.3	0.4	—	0.979	3.251	6.407
0.2	—	0.5	—	0.3	1.522	7.262	12.648
0.1	—	—	0.6	0.3	0.971	3.322	5.339
—	0.3	0.2	0.5	—	0.719	2.254	5.411
—	0.2	0.4	—	0.4	1.309	3.916	9.397
—	0.3	—	0.4	0.3	**0.463**	**1.590**	**4.105**
—	—	0.4	0.3	0.3	0.485	2.515	5.695
0.2	0.2	0.1	0.5	—	0.693	1.685	6.122
0.2	0.3	0.2	—	0.3	1.105	3.844	7.546
0.0	0.3	—	0.4	0.3	0.463	1.590	4.105
0.0	—	0.4	0.3	0.3	0.485	2.515	5.695
—	0.1	0.3	0.3	0.3	**0.337**	**1.495**	**3.251**
0.0	0.1	0.3	0.3	0.3	**0.337**	**1.495**	**3.251**

TABLE III
RESULTS OF MULTIBIOMETRICS USING FEATURE-LEVEL FUSION, WHERE THE BOLD NUMBERS INDICATE THE BEST FOR EACH COMBINATION.

Trait					EER [%]	FRR [%]	
Face	Periocular	Iris	Nose	Eyebrow		@FAR0.1%	@FAR0.01%
√	√	—	—	—	7.489	30.778	42.572
√	—	√	—	—	15.916	71.239	83.056
√	—	—	√	—	2.633	12.814	19.696
√	—	—	—	√	3.590	13.075	21.974
—	√	√	—	—	12.586	54.438	70.069
—	√	—	√	—	1.992	9.967	18.889
—	√	—	—	√	2.654	13.787	26.341
—	—	√	√	—	4.911	31.728	45.444
—	—	√	—	√	6.414	35.928	52.421
—	—	—	√	√	**1.851**	**4.865**	**7.333**
√	√	√	—	—	12.142	58.851	72.188
√	√	—	√	—	**1.781**	9.516	14.642
√	√	—	—	√	2.376	12.506	22.164
√	—	√	√	—	4.744	26.863	39.203
√	—	√	—	√	4.690	30.897	55.007
√	—	—	√	√	2.042	5.529	**7.665**
—	√	√	√	—	5.791	31.870	43.901
—	√	√	—	√	5.227	33.318	59.326
—	√	—	√	√	1.637	**8.685**	16.255
—	—	√	√	√	3.735	27.290	44.186
√	√	√	√	—	6.328	34.006	49.051
√	√	√	—	√	5.868	32.677	47.698
√	√	—	√	√	**1.397**	**9.112**	**17.205**
√	—	√	√	√	2.118	16.279	38.443
—	√	√	√	√	2.789	20.242	35.999
√	√	√	√	√	**2.797**	**19.767**	**30.636**

results between the conventional methods and the proposed method. Verma [19] has low recognition accuracy due to its combination of hand-crafted features. The other methods use CNN and achieve high recognition accuracy. Considering that the number of iris and periocular data is twice for the conventional methods, the proposed method is demonstrated to have high recognition accuracy.

TABLE IV
COMPARISON WITH CONVENTIONAL METHODS AND THE PROPOSED METHOD. THE UNITS FOR EER AND FRR ARE PERCENTAGES. (F: FACE, P: PERIOCULAR, I: IRIS, LI: LEFT IRIS, RI: RIGHT IRIS, N: NOSE, E: EYEBROW)

Method	Trait	EER	FRR	
			@FAR0.1	@FAR0.01
Verma [19]	P+I	—	38.5	—
Ammour [10]	F+LI+RI	0.24	—	—
Wang [11]	P+I	2.29	—	13.7
Tonosaki [12]	P+I	1.13	—	11.2
Proposed	P+I+N+E	0.34	1.50	3.25

V. CONCLUSION

In this paper, we proposed a novel multibiometrics method that combines face, iris, periocular, nose, and eyebrow, that can be extracted from a single face image. This method does not sacrifice the convenience of biometrics since only a single face image is used as input. Through a set of experiments using the CASIA Iris Distance database, we demonstrated that the method with score-level fusion of the four biometric traits except face has the highest recognition accuracy and also demonstrated the effectiveness of the proposed method by comparing it with conventional methods.

REFERENCES

[1] A. Jain, P. Flynn, and A. Ross, *Handbook of Biometrics*. Springer, 2008.

[2] A. Ross, K. Nandakumar, and A. Jain, *Handbool of Multibiometrics*. Springer, 2006.

[3] R. Telgad, A. Siddiqui, S. Lothe, P. Deshmukh, and G. Jadhao, "Development of an efficient secure biometric system by using iris, fingerprint, face," *Proc. Int'l Conf. Intelligent Systems and Information Management*, pp. 100–104, 2017.

[4] K. Wang and A. Kumar, "Periocular-assisted multi-feature collaboration for dynamic iris recognition," *IEEE Trans. Information Forensics and Security*, vol. 16, pp. 1060–1064, Sep. 2021.

[5] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maître, "XM2VTSDB: The extended M2VTS database," *Proc. Int'l Conf. Audio- and Video-Based Biometric Person Authentication*, pp. 72–77, 1999.

[6] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, *et al.*, "MCYT baseline corpus: A bimodal biometric database," *IEE Proc. - Vision, Image and Signal Processing*, vol. 150, no. 6, pp. 395–401, Dec. 2003.

[7] M. Stokkenes, R. Ramachandra, M. Sigaard, K. Raja, M. Gomez-Barrero, and C. Busch, "Multi-biometric template protection - A security analysis of binarized statistical features for bloom filters on smartphones," *Proc. Int'l Conf. Image Processing Theory, Tools and Applications*, pp. 1–6, Dec. 2016.

[8] Q. Zhang, H. Li, Z. Sun, and T. Tan, "Deep feature fusion for iris and periocular biometrics on mobile devices," *IEEE Trans. Information Forensics and Security*, pp. 2897–2912, 2018.

[9] T. Ng, C. Low, J. Long Chai, and A. Beng Jin Teoh, "Conditional multimodal biometrics embedding learning for periocular and face in the wild," *Proc. Int'l Conf. Pattern Recognition*, pp. 812–818, 2022.

[10] B. Ammour, T. Bouden, and L. Boubchir, "Face-iris multimodal biometric system based on hybrid level fusion," *Proc. Int'l Conf. Telecommunications and Signal Processing*, pp. 298–302, Jul. 2018.

[11] C. Wang, J. Muhammad, Y. Wang, Z. He, and Z. Sun, "Towards complete and accurate iris segmentation using deep multi-task attention for non-cooperative iris recognition," *IEEE Trans. Information Forensics and Security*, pp. 2944–2959, Mar. 2020.

[12] T. Tonosaki, S. Kawakami, H. Kawai, *et al.*, "Eye biometrics combined with periocular and iris recognition using cnn," *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conf.*, pp. 1804–1810, Nov. 2023.

[13] S. Kawakami, H. Kawai, K. Ito, *et al.*, "A simple and accurate cnn for iris recognition," *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conf.*, pp. 1563–1568, Nov. 2022.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 770–778, Jun. 2016.

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4685–4694, Jun. 2019.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 7132–7141, Jun. 2018.

[17] D. Kingma and L. Ba, "Adam: A method for stochastic optimization," *Proc. Int'l Conf. Learning Representations*, pp. 1–15, May 2015.

[18] N. Othman, B. Dorizzi, and S. Garcia-Salicetti, "OSIRIS: An open source iris recognition software," *Pattern Recognition Letters*, vol. 82, no. 2, pp. 124–131, Oct. 2016.

[19] S. Verma, P. Mittal, M. Vatsa, and R. Singh, "At-a-distance person recognition via combining ocular features," *Proc. Int'l Conf. Image Processing*, pp. 3131–3135, Sep. 2016.

Comparative Analysis of Voice Mimicry Attacks by High- and Low-Skilled Imitators on Speaker Verification Systems

Koji Iwano*, Wakana Komuro*, and Manami Gomi*
* Tokyo City University, Yokohama, Japan
E-mail: iwano@tcu.ac.jp Tel: +81-45-910-2598

Abstract— **Given that "voice mimicry" constitutes a fundamental attack against speaker verification systems, it is essential to comprehend the impact of these attacks on the system and their characteristics. This study evaluates and compares the effects of voice mimicry attacks by high- and low-skilled imitators on speaker verification systems under impartial conditions. Two types of speaker verification systems are used in the analysis: a classical system based on GMM-UBM method and a deep-learning-based system using x-vector. Results of the comparative analysis, using a dataset originally constructed, indicate that the attack of the high-skilled imitator has higher spoofing ability against both systems than the low-skilled imitators. Furthermore, the analysis of the verification scores obtained from the x-vector-based speaker verification system revealed differences in the characteristics of high- and low-skilled imitators in voice mimicry attacks.**

I. INTRODUCTION

Voice-based person authentication, namely speaker verification, is one of the most promising biometric authentication technologies due to its advantages, such as not requiring any special input devices and having a wider application area compared to other biometrics [1]. To promote the practical application of this technology, it is essential to have a comprehensive understanding of its vulnerability to various spoofing attacks. Possible attacks include impersonation (voice mimicry by humans), replay of recordings, speech synthesis (text-to-speech), and voice conversion [2]. For the spoofing attacks other than voice mimicry, various countermeasure methods have been proposed. These methods mainly detect and use acoustical properties of the recording voice (e.g. [3, 4]) or artefacts by signal processing in speech synthesis or voice conversion (e.g. [5, 6, 7]). On the other hand, although voice mimicry is a simple attack that does not require expertise technical knowledge, it cannot be effectively counteracted by the same approach because it is produced by live human. In order to consider countermeasures against voice mimicry attacks, their threats and characteristics should be thoroughly investigated.

In this context, several studies have investigated the impact of voice mimicry attacks on speaker verification systems [8, 9, 10]. In paper [8], Lau et al. have shown that the attacks by speakers whose natural voice is closed to that of the target speaker can cause performance degradation for classical GMM-UBM-based speaker verification systems. The work in Hautamäki et al. [9] investigated the effects of mimicry attacks by a professional celebrity impersonator on the performance of i-vector-based speaker verification system. As a result, it has been confirmed that the impersonation can increase the FAR (False Acceptance Rates) of the verification systems. In the work by Mariéthoz and Bengiothe [10], mimicry attacks for GMM-UBM-based speaker verification system was analyzed when the attacker was each of professional or non-professional impersonators. Their experimental results suggest that professional impersonators can spoof more effectively than non-professional imitators. However, this comparison of abilities is not considered to be fair because the target speakers were public figures who the professional impersonator felt was best able to imitate.

Therefore, we conduct experiments to investigate the voice mimicry attacks by both a high-skilled imitator (a professional celebrity impersonator with a large repertoire) and low-skilled imitators (regular university students with little to no experience in performing imitations) when the target speaker is "a person who has never been imitated by both attackers before", allowing us to fairly compare their spoofing abilities. In the experiments, two types of speaker verification systems are used: one is based on the classical GMM-UBM method [11] and the other is based on deep-learning-based method using x-vector [12]. We first investigate the performance degradation of the two verification systems due to the attacks. Then, we analyze the differences in characteristics between high- and low-skilled imitators using the verification scores obtained from the x-vector-based system.

The rest of the paper is structured as follows: Section II details the construction of the original speech dataset. In Section III, we explain two types of the verification systems utilized in the analysis. Section IV reports the results of a comparative analysis of spoofing abilities by both high- and low-skilled imitators. Section V reveals the differences in the characteristics of both imitators. Finally, Section VI concludes the paper.

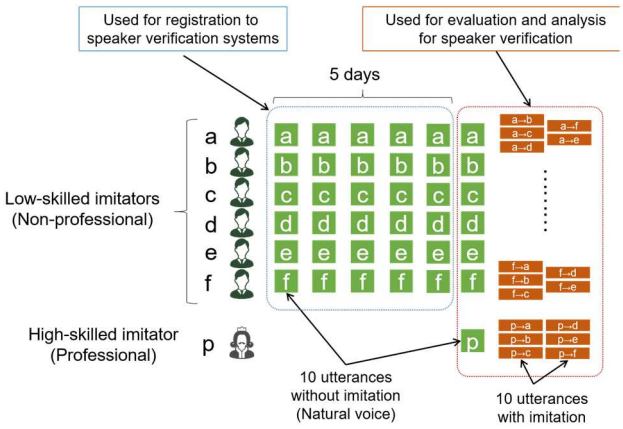


Fig. 1 Overview of the constructed dataset.

II. Dataset Construction

A dataset was created by recording Japanese four-connected-digit speech; each utterance was like "2598" (in English, "two-five-nine-eight") composed of random digits. Fig. 1 shows an overview of the constructed dataset.

The utterances used for registration in the verification systems were connected-digit strings which were naturally uttered by six male speakers (a~f) without imitation. The speakers were Japanese university students in 21-22 years old. Each speaker uttered 10 strings per day, with interval of several days. A total of five days' recording data (50 utterances per speaker) was used as the registration voice. On a different day from these five days, each speaker uttered 10 strings without imitation and 50 utterances with imitation (each speaker imitates each of the other five speakers ten times). These utterances were used as the input voice for attacks by low-skilled imitators. Note that each speaker frequently heard the voice of the other speakers, but had no experience imitating them.

In addition, we requested a famous Japanese male impersonator (p), who frequently appears on TV and has a repertoire of around 200 different characters, to imitate the six students mentioned above, and recorded his imitation. He was in his 40s and had a career spanning about 20 years. He uttered 10 strings in his own natural voice and 60 strings with imitation (ten times per student). These utterances are used as inputs for high-skilled imitator's attacks. He is not acquainted with the six students prior to the recording, and this was the first time for them to imitate the students' voices. Before recording, he was asked to listen to each student's voice several times and practice for a few minutes.

The recording was done in a relatively quiet room, using a microphone (Audio-Technica AT-UD6) and a digital converter (Roland UA-101). The data was recorded in 16bits single-channel PCM with 16kHz sampling frequency.

III. Speaker Verification Systems for Analysis

A. Overview of the Verification Process

Both of two speaker verification systems for the analysis utilize a log-likelihood ratio (LLR) score ℓ,

$$\ell = \log \frac{p(\mathbf{X}|H_T)}{p(\mathbf{X}|H_{NT})} \quad (1)$$

where $\mathbf{X} = X_1, \cdots, X_t$ is an input sequence of acoustic feature vector, and X_t indicates a feature vector of t-th frame. H_T indicates the target hypothesis "input is generated by a target (claimed) speaker", and H_{NT} indicates the alternative hypothesis "input is generated by a non-target (different) speaker". $p(\mathbf{X}|H_T)$ and $p(\mathbf{X}|H_{NT})$ denote the likelihood of each hypothesis. If the score l is larger than a pre-determined threshold θ, the speaker is accepted as the target speaker.

B. GMM-UBM-based Speaker Verification System

In the original GMM-UBM method [11], the acoustic features of the speech from a target speaker and a non-target speaker are modeled as a target speaker model and a speaker independent model (UBM: Universal Background Model), respectively, using Gaussian Mixture Model (GMM). In the implementation of this study, a three-state Hidden Markov Model (HMM) is used for each model. Since each state of the HMM is represented by a GMM, the framework corresponds to the GMM-UBM method. At the time of authentication, a sequence of features $\mathbf{X}$ is input to the target speaker model (C) and the UBM (U), and then the likelihoods $p(\mathbf{X}|C)$ and $p(\mathbf{X}|U)$ are calculated, respectively, using the Viterbi algorithm. By considering $p(\mathbf{X}|C)$ as $p(\mathbf{X}|H_T)$ and $p(\mathbf{X}|U)$ as $p(\mathbf{X}|H_{NT})$ in Eq. (1), the score ℓ can be obtained using the following equation:

$$\ell = \log \frac{p(\mathbf{X}|C)}{p(\mathbf{X}|U)} \quad (2)$$

The target speaker models for the six registrants are trained using 5 days of data (50 utterances) as described in Section II, and the UBM is trained using 300 utterances of all six registrants. The acoustic features are 25-dimensional vectors based on Mel-Frequency Cepstrum Coefficients (MFCCs) which consists of 12-dimensional MFCCs, their first derivative components, and the first derivative component of log power. The frame shift for feature extraction is 10ms and the analysis window length is 25ms. The number of mixtures in the GMMs was optimized to maximize verification performance for natural voice and set to 64.

C. X-vector-based Speaker Verification system

In x-vector-based speaker verification system [12], a sequence of acoustic features $\mathbf{X}$ is input to a deep neural network for speaker identification that has been trained from the large amount of speech data by many speakers. The output obtained from the intermediate layer is used as a vector (x-vector) that is regarded as an efficient representation of the

speaker. The input acoustic feature X_t is a 30 dimensional MFCC extracted with a window length of 25ms and a frame shift of 10ms. The output nodes of the network generate the probabilities (likelihoods) of each speaker. The features are first input into the TDNN (Time Delay Neural Network) layers frame by frame. After the entire frame have been input, the statistics (mean and standard deviation) are calculated as a vector by the statistics pooling layer. The vectors are input to the embedding layers, and then 512-dimensional x-vectors are extracted from the intermediate nodes. This deep neural network was trained on more than 1 million utterances by 7,323 speakers in the speaker recognition database VoxCeleb1, 2 [13]. In training process, Kaldi toolkit [14] and the recipes provided in [15] were used. According to the recipes, data augmentation using additive noises and reverberation was conducted for the training data. We use the MUSAN dataset [16] for additive noises, and room impulse responses (RIRs) database described in [17] for reverberation.

The score ℓ is defined as similarity between the x-vectors of the input voice (x_i) and the registered voice of target speaker (x_r) and calculated based on PLDA (Probabilistic Linear Discriminant Analysis) [18]. Specifically, the score is defined by Eq. (3), $p(x_i, x_r|H_T)$, $p(x_i|H_{NT})$, and $p(x_r|H_{NT})$ are obtained from generative models trained through the process of PLDA.

$$\ell = \log \frac{p(x_i, x_r|H_T)}{p(x_i|H_{NT}) \cdot p(x_r|H_{NT})} \quad (3)$$

Since there are actually 50 utterances for each target speaker as registered voice, the similarity scores of the input voice to all registered voices are calculated, and the average is used as the final verification score.

IV. Analysis of Speaker Verification Performance against Voice Mimicry Attacks

We investigate the performance of two speaker verification systems when the following four types of the voice is used for spoofing attacks:

Case 1: Utterances in the natural voice of non-professional impersonators (students) without imitation

Case 2: Utterances by non-professional impersonators (students) with imitation

Case 3: Utterances in the natural voice of a professional impersonator without imitation

Case 4: Utterances by a professional impersonator with imitation

The performance evaluation results are shown as changes in FAR (false acceptance rate; success probability of spoofing) relative to the threshold θ. For reference, we also show the change in FRR (false reject rate) when the registrant's own natural voice is input. Figures 2 and 3 show the evaluation results for the GMM-UBM-based system and the x-vector-based system, respectively. These results show that, for both systems, Case 2 (voice mimicry by the low-skilled imitators) slightly increases the FARs compared to Case 1. On the other hand, imitation attacks by Case 4 (voice mimicry by the high-skilled imitator) significantly increase the FARs on both systems compared to Case 3.

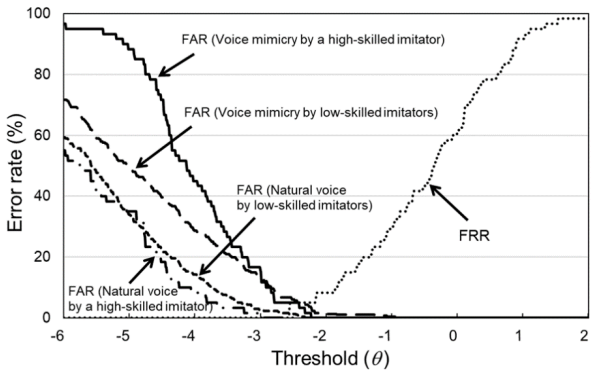


Fig. 2 Performance evaluation results of the GMM-UBM-based speaker verification system against four types of spoofing attacks.

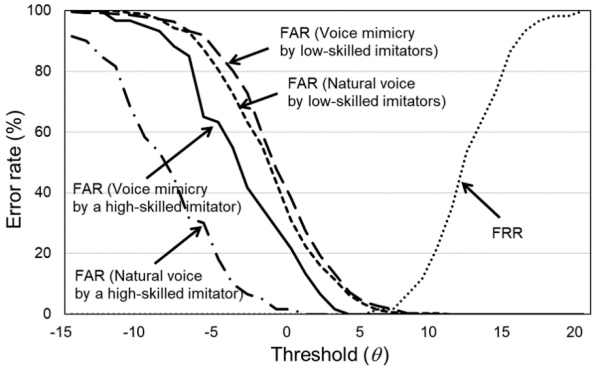


Fig. 3 Performance evaluation results of the x-vector-based speaker verification system against four types of spoofing attacks.

Fig. 2 shows that the mimicry attacks by the professional impersonator tend to yield higher verification scores than those by the students in the GMM-UBM-based system. However, Fig. 3 demonstrates that it is difficult for him to achieve higher scores compared to those of the low-skilled imitators in the x-vector-based system. This may be due to the fact that the use of x-vector greatly improves the discrimination performance between the natural voice of the professional impersonator and the registered voices of the six students, as indicated by the notably low FARs in Case 3. Nevertheless, the fact that the professional impersonator can substantially increase the FAR through imitation suggests that he has higher skill than the non-professional impersonators in changing his voice to approach the target speaker. Therefore, if the attacker's natural voice is close to that of the target, voice imitation by a high-skilled imitator is still expected to pose a threat, even to deep-learning-based verification systems.

Table 1 Average verification scores of each imitator's natural voice for each target speaker, obtained from the x-vector-based system.

Imitator		Target Speaker					
		a	b	c	d	e	f
Low-skilled	a	10.35	−3.57	−1.41	1.25	1.68	−4.11
	b	−2.51	10.60	−4.55	−5.09	−4.61	3.58
	c	0.54	−2.03	12.79	−1.08	−1.31	−1.85
	d	3.72	−2.62	−2.71	15.51	−1.84	−4.79
	e	2.22	−0.73	−4.40	−1.27	9.72	−1.35
	f	−2.81	2.27	−2.32	−4.52	−4.35	13.28
High-skilled	p	−10.94	−4.40	−12.36	−11.82	−8.41	−4.76

V. Analysis of Characteristics of Voice Mimicry Attacks Using Verification Scores

A. *Preliminary analysis on distance between natural voices*

The differences in the characteristics of voice mimicry attacks between high- and low-skilled imitators are analyzed using the verification scores. We chose to use the x-vector-based system for this analysis due to its superior discrimination ability over the GMM-UBM-based system.

Before this analysis, we preliminarily investigated the extent to which the natural voice of each imitator differs from that of each target speaker. Table 1 shows the average scores of each imitator's natural voice for each target speaker. This result shows that there are some pairs of speakers with slightly higher scores (e.g., a-d), indicating that the natural voices of these speakers are considered to be similar. On the other hand, we see that the high-skilled imitator (p) is quite far from any speaker. In the following sections, we will analyze the imitations of low-skilled speakers by dividing them into three groups (A, B, and C) based on the scores observed in the preliminary analysis:

A) Imitations between speakers with close natural voices (a-c, a-d, a-e, b-f, d-e)
B) Imitations between speakers with intermediately distant natural voices (b-e, c-d, c-e, c-f, e-f)
C) Imitations between speakers with distant natural voices (a-b, a-f, b-c, b-d, d-f)

B. *Comparison of verification scores between high- and low-skilled imitators*

Fig. 4 shows box-and-whisker plots of the verification scores for utterances imitated by high- and low-skilled imitators. For the low-skilled imitators, it can be seen that imitation between speakers with close natural voices (Group A) tends to yield higher scores. For the high-skilled imitator, the scores are comparable to those of low-skilled imitators in Group C, even though the high-skilled imitator's natural voice is quite different from that of the target speaker. Fig. 5 shows box plots of the differential scores for each imitated utterance, compared to the average score of the imitator's own natural voice for the target speaker. This shows that the scores increased for the low-skilled imitators only when they imitated a target speaker whose natural voice is close to theirs. For the high-skilled imitator, the scores increased to a greater extent compared to the low-skilled imitators.

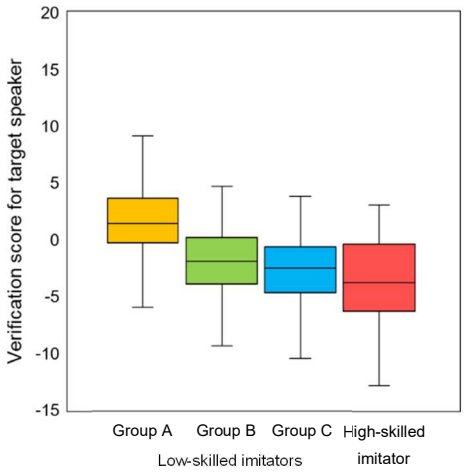


Fig. 4 Box plots of the verification scores for utterances imitated by high- and low-skilled impostors. Groups A, B, and C indicate that the distance between natural voices of the low-skilled imitator and target is close, intermediate, and distant, respectively.

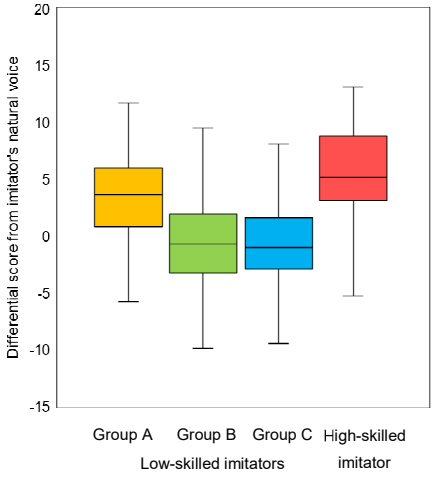


Fig. 5 Box plots of the differential scores for each imitated utterance compared to the average score of the imitator's own natural voice for the target speaker. Groups A, B, and C indicate that the distance between natural voices of the low-skilled imitator and target is close, intermediate, and distant, respectively.

C. *Analysis based on feature movements in registered speaker space*

We analyze how the verification score from x-vector-based system changes with imitation, considering scores not only for the target speaker but also for other registered speakers. First, each utterance used in the evaluation is transformed into a six-dimensional feature vector, where each dimension represents the average verification scores for one of the six registered speakers.

Let $\boldsymbol{u}_{i\to j}^{S}$ denote a feature vector of an utterance by imitator i intended to spoof target speaker j. We also denote a feature

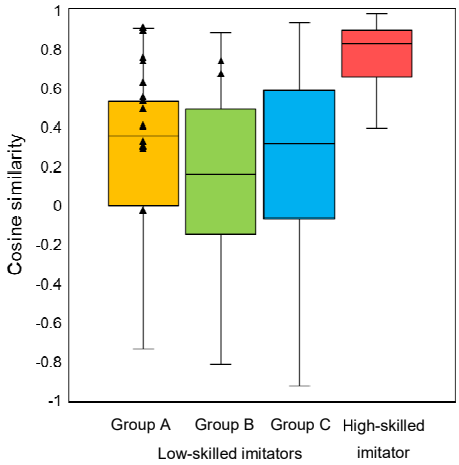


Fig. 6 Box plots of the cosine similarities for each imitated utterance defined by equation (4). Groups A, B, and C indicate that the distance between natural voices of the low-skilled imitator and target is close, intermediate, and distant, respectively. The black triangles in the figures represent the imitated utterances with verification scores in the top 5%.

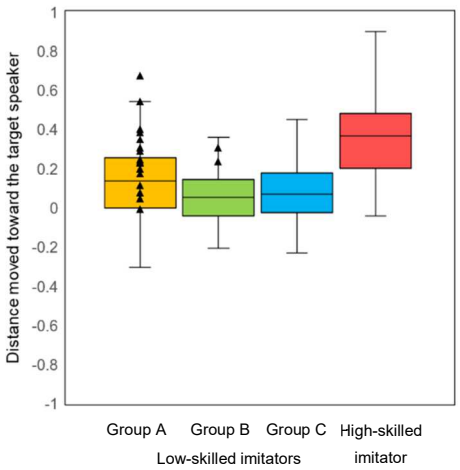


Fig. 7 Box plots of the distances moved toward the target speaker for each imitated utterance defined by equation (5). Groups A, B, and C indicate that the distance between natural voices of the low-skilled imitator and target is close, intermediate, and distant, respectively. The black triangles in the figures represent the imitated utterances with verification scores in the top 5%.

vector of an utterance in speaker i's natural voice by $\boldsymbol{u}_i^N$. The feature movement in the 6-dimensional speaker space by the imitation $\boldsymbol{u}_{i\to j}^S$ can be expressed as $\boldsymbol{M}(\boldsymbol{u}_{i\to j}^S) = \boldsymbol{u}_{i\to j}^S - \bar{\boldsymbol{u}}_i^N$ and ideal movement that the imitator i should aim for can be express as $\boldsymbol{M_B}(\boldsymbol{u}_{i\to j}^S) = \bar{\boldsymbol{u}}_j^N - \bar{\boldsymbol{u}}_i^N$. Therefore, the gap between the two movement vectors is evaluated using cosine similarity between them (denoted as CS) and the distance moved toward the target speaker (denoted as D). These are defined by the following equations:

$$CS\left(\boldsymbol{u}_{i\to j}^S\right) = \frac{\boldsymbol{M}(\boldsymbol{u}_{i\to j}^S)\cdot\boldsymbol{M_B}(\boldsymbol{u}_{i\to j}^S)}{\left\|\boldsymbol{M}(\boldsymbol{u}_{i\to j}^S)\right\|\left\|\boldsymbol{M_B}(\boldsymbol{u}_{i\to j}^S)\right\|} \tag{4}$$

$$D\left(\boldsymbol{u}_{i\to j}^S\right) = \frac{\boldsymbol{M}(\boldsymbol{u}_{i\to j}^S)\cdot\boldsymbol{M_B}(\boldsymbol{u}_{i\to j}^S)}{\left\|\boldsymbol{M_B}(\boldsymbol{u}_{i\to j}^S)\right\|^2} \tag{5}$$

Note that the distance D is normalized by the distance between natural voices of speakers i and j, i.e., $\left\|\boldsymbol{M_B}(\boldsymbol{u}_{i\to j}^S)\right\|$.

Figures 6 and 7 show the box-and-whisker plots of the cosine similarities and distances, respectively. The results show that the high-skilled imitator achieves significantly greater cosine similarity (CS) and distance (D) than the low-skilled imitators, indicating that his imitated utterances more effectively approach the target speaker's voice. Note that there are no significant differences in cosine similarity among the three groups of low-skilled imitators at the 1% significance level.

The black triangles in the figures represent the imitated utterances with verification scores in the top 5%, indicating they are more likely to be successfully spoofed. It is found that these utterances are concentrated in Group A, and some have very small similarity and distance. They indicate the following:

- Spoofing by low-skilled imitators is more likely to be successful when their natural voice is close to the target speaker.
- There are two types of cases in which spoofing by low-skilled imitators can be successful: (1) those that certainly approach the target speaker's natural voice, and (2) those that do not approach the target speaker's natural voice but still produce high verification scores.

VI. Conclusions

In this study, we conducted a comparison of voice mimicry attacks performed by a high-skilled imitator (a professional celebrity impersonator) and low-skilled imitators (university students) under fair conditions, utilizing two types of speaker verification systems: a GMM-UBM-based system and an x-vector-based system. The results confirmed that the high-skilled imitator exhibited greater spoofing ability (i.e., a higher increase in FAR due to voice mimicry attack) in both systems.

Furthermore, a detailed analysis of high- and low-skilled imitators' mimicry attacks using scores obtained from the x-vector-based speaker verification system was conducted. The results show that spoofing by the high-skilled imitator more effectively changes his voice to approach the target speaker's voice compared to the low-skilled imitators. They also show that there exist imitated utterances by the low-skilled imitators that are not significantly close to the target speaker, some of which may still produce high verification scores.

Since these conclusions are derived from only one professional impersonator, its generality may be limited. However, our study provides evidence of the existence of such high-skilled imitators. Our future work includes developing countermeasures against voice spoofing attacks that takes into account the results of these analyses.

VII. ACKNOWLEDGMENT

We extend our sincere gratitude to the impersonators who collaborated with us in conducting this experiment. This research was supported by JSPS Grant-in-Aid for Scientific Research (C) 19K12051.

REFERENCES

[1] M. Gomez-Barrero, P. Drozdowski, C. Rathgeb, J. Patino, M. Todisco, M. Nautsch, N. Damer, J. Priesnitz, N. Evans, and C. Busch, "Biometrics in the era of COVID-19: Challenges and opportunities," IEEE Trans. Technology and Society, 3(4), 307-322, 2022.

[2] Z. Wu, N. Evans, T, Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," Speech Communication, 66, 130-153, 2015.

[3] W. Shang and M. Stevenson, "Score normalization in playback attack detection," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2010.

[4] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," In C. Vielhauer, J. Dittmann, A. Drygajlo, N. Juul, M. Fairhurst (Eds.), Biometrics and ID Management, Lecture Notes in Computer Science, Springer, 274-285, 2011.

[5] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," Proc. European Conference on Speech Communication and Technology (Eurospeech), 2001.

[6] L.W. Chen, W. Guo, and L.R. Dai, "Speaker verification against synthetic speech," Proc. 7th Int. Symposium on Chinese Spoken Language Processing (ISCSLP), 2010.

[7] Z. Wu, E.S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," Proc. Interspeech, 2012.

[8] Y.W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," Proc. Int. Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.

[9] R.G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, A.M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry," Proc. Interspeech, 2013.

[10] J. Mariéthoz and S. Bengio, "Can a professional imitator fool a GMM-based speaker verification system?" IDIAP Research Report, Idiap-RR 05-61, 2006.

[11] D. Reynolds, T. Quatieri, and R.B. Dunnet, "Speaker verification using adapted Gaussian Mixture Models," Digital Signal Processing, 10, 19-41, 2000.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 5329-5333, 2018.

[13] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/

[14] https://kaldi-asr.org/

[15] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

[16] D. Snyder, G Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," arXiv:1510.08484v1, 2015.

[17] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," Proc. IEEE Int. Conf., Acoustics, Speech and Signal Processing (ICASSP), pp. 5220–5224, 2017.

[18] P. Li, Y. Fu, U. Mohammed, J.H. Elder, and S.J. Prince, "Probabilistic models for inference about identity," IEEE Trans. Pattern Analysis Machine Intelligence, 34(1), 144–157, 2012.

Multi-Observed Authentication: A secure and usable authentication based on multi-point observation of a single physical credential

Wataru Hatakeyama[1], Shinnosuke Nozaki[2], Ayumi Serizawa[1], Mizuho Yoshihira[1],
Masahiro Fujita[2], Ayako Yoshimura[2], Tetsushi Ohki[1], Masakatsu Nishigaki[1]
1 Shizuoka University, 3-5-1 Johoku, Chuo, Hamamatsu, Shizuoka, 432-8011, Japan.
2 Mitsubishi Electric Corporation, 5-1-1 Ofuna, Kamakura, Kanagawa, 247-8501, Japan.
E-mail: nisigaki@inf.shizuoka.ac.jp

Abstract— **Nowadays, it is no longer uncommon for a PC to be infected with malware. User authentication based on only one legitimate credential (such as a password) may be insufficient for judging whether a user is legitimate. A typical solution to this problem is two-factor authentication, which is a method of authentication based on the presentation of two factors by the user. Generally, the first factor is input into a PC and the second into a smartphone. However, this reduces usability by forcing the user to present multiple factors at each authentication. Therefore, we propose multi-observed authentication as a new method based on the concept of "confirmation of the user's intention at the time of authentication, in addition to the validity of the credentials (passwords or authentication tokens)." The proposed method captures the user's physical authentication actions as physical events and simultaneously observes each event at multiple points, ensuring that the credential is not sent by malware residing in the PC but is physically input by a real human.**

I. INTRODUCTION

In recent years, business systems have been changing to better meet the digital transformation strategy. Correspondingly, the use of malware for stealing user-specific information (hereinafter referred to as credentials) has continued to grow against users (individual and businesses) who are not fully prepared for such changes [1]. In the digital transformation environment, employees log into cloud business systems from their PCs to work. When malware resides on a PC, it is difficult for an authentication server to judge whether the credentials received from the PC are from a legitimate user or the malware. Therefore, inputting only one piece of legitimate user-specific information, such as a password, is often insufficient for judging whether the user is legitimate.

A typical solution to this problem is two-factor authentication. Currently, two-factor authentication is used to prove that a user is legitimate by presenting credentials for two of the following three elements [2].

1. Knowledge factor: based on something the user knows, such as passwords.
2. Possession factor: based on something the user has, such as ID cards, smartphones, etc.
3. Biometric factor: based on something the user is, such as through face recognition, fingerprint recognition, etc.

In general, the user is authenticated by first inputting the credential into the PC and then into a smartphone. In other words, two-factor authentication is a method of strengthening security by multiplexing authentication [3], [4]. However, it reduces usability by forcing the user to present multiple credentials at each authentication. If only improvement in usability is considered, the automatic confirmation of the smartphone's proximity to the PC could be an option for the second factor (possession). However, with this method, if malware fraudulently presents the first credential while a legitimate user is working on a PC, the second factor will also be passed automatically. In terms of security, this is synonymous with single-factor authentication.

Rephrasing it in another way, if the purpose of requiring two-factor authentication is to increase the security of the PC against a malware infection, another credential is not always necessary. Instead, it should be sufficient to verify that a human (not malware) has inputted the legitimate credential. With this method, the user only needs to input a single credential into the PC as before, while maintaining usability and improving security.

Therefore, in this paper[1], a new user authentication method called multi-observed authentication, is proposed based on the concept of confirming that a human (not malware) has inputted the legitimate credential. The proposed method achieves this by simultaneously observing "the input of a single legitimate credential by the user" on both the PC and smartphone. Judging from the fact that both the PC and smartphone have received the legitimate credential from a single input device at the same

[1] This paper is a reorganization of the concept reported in NBiS 2022 [5] with the addition of experimental results presented in a poster at NDSS 2024 [6].

time, it is deduced that the credential is not sent by malware residing in the PC but physically inputted by a real human. As a result, both PC and smartphone can play a role in authentication factors to verify the user's legitimacy. Thus, the proposed method achieves authentication equivalent to two-factor authentication with both usability and security.

Depending on the legitimate user model and/or threat model, the information that should be confirmed by the smartphone may be changed to "a human (not malware) has inputted what appears to be a credential" or "a legitimate user has inputted the legitimate credential." This study, in Section V, also proposes a variation of the method for these cases. For the sake of simplicity, the focus is on two-factor authentication, but the proposed method can be extended to multi-factor authentication with a single credential observed at three or more points.

II. Two-Factor Authentication

With the spread of COVID-19, working from home has become widespread. Employees work on their PC from home and access the information on the assets of the company's ICT resources, as needed. Each time the employees access an asset, they are required to authenticate. In line with this, malware that infects the employee's PC and steals user authentication credentials is rapidly increasing [1]. This is why user authentication methods must be hardened and two-factor authentication being more widely used across many companies at present. Hereafter, the employee is referred to as the user.

A. Single-Factor and Two-Factor Authentication

In single-factor authentication, the server judges the legitimacy of the user by the presentation of a factor (knowledge/possession/biometric) that can only belong to the legitimate user. The use of possession or biometric factors is recommended from the viewpoint of impersonation. However, single-factor authentication using a knowledge factor has been more common until now due to ease of implementation (input via standard input devices of ICT equipment) and privacy concerns (biometric information being leaked).

When users working with a PC access a company resource using a knowledge factor, e.g., a password (PW), the flow of the single-factor authentication is as follows;

1. The user inputs the PW into the PC.
2. The PC sends the PW to the authentication server.
3. The server authenticates whether the user is legitimate or not based on the PW received from the PC.

In Fig. 1, the PWs in Steps 1 and 2 are written as PW and PW_{PC}, respectively, to distinguish them, but actually PW and PW_{PC} are identical.

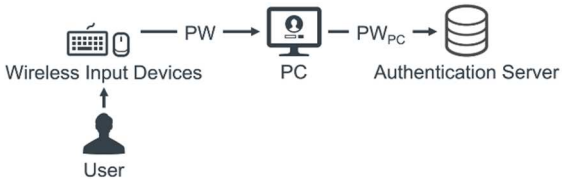


Fig. 1. Single-factor Authentication

If a legitimate user's PC is infected by malware, the premise that the credential is possessed by only a legitimate user is broken. Therefore, the server cannot judge if the user is legitimate simply from the input of the PW. A typical solution to this problem is two-factor authentication, which uses two credentials.

A user's PC can be infected with malware and manipulated either autonomously or by the remote control of a rogue actor. Therefore, it is not a good approach to register each user's PC as the second credential to confirm the possession of a legitimate PC. Moreover, the malware can steal a credential inputted into the infected PC. Therefore, it is also not a good idea to send the second credential to the authentication server via the PC[2]. For these reasons, two-factor authentication, in which the smartphone is used as a means to submit the second authentication factor, has become common. By securing a different authentication route from the PC, even if the PC is infected with malware and the first credential (PW) is stolen, impersonation can be prevented as long as an attacker does not break though the authentication of the smartphone.

In the following sections, a two-factor authentication system consisting of a PC and smartphone (Fig. 2) is described, as a concrete example. As shown in Fig. 2, the first credential is a PW and the second is a PIN, but credentials other than knowledge factors may be used. In the actual authentication procedure, the authentication server instructs the smartphone to request the user to input the PIN only when the PW from the PC arrives at the authentication server; however, this is omitted in Fig. 2. In Fig. 2, to distinguish between the credentials inputted by the user and those forwarded to the authentication server, PW, PW_{PC}, PIN, and PIN_{SP}, are specified, respectively, but actually $PW = PW_{PC}$ and $PIN = PIN_{SP}$.

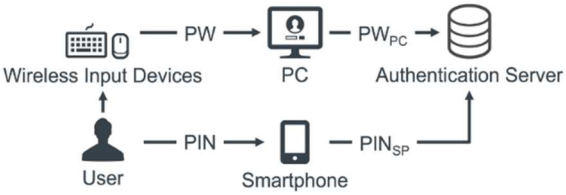


Fig. 2. Two-factor Authentication

B. Issue of Two-Factor Authentication

Figure 2 simply shows only two credentials (PW and PIN). However, in practice, the actual operation of the user is to input

[2] The exception is when a One-time password (OTP) is used as the second credential. For example, in [4], the user inputs the OTP sent to the smartphone from the authentication server to the authentication server via the PC, which constitutes two-factor authentication.

the PW after logging in to the PC, and to input the PIN after activating the smartphone. To note this precisely, Fig. 2 is reconfigured as shown in Fig. 3. Notations are provided below.

- AC_{PC}: The activation credential that the user inputs to the PC when using the PC.
- PW: The first credential that the user inputs into the PC when accessing the information asset. The PC sends the first credential to the authentication server.
- AC_{SP}: The activation credential that the user inputs into the smartphone when using the smartphone.
- PIN: The second credential that the user inputs into the smartphone when accessing the information asset. The smartphone sends the second credential to the authentication server.

In Fig. 3, just like in Fig. 2, $PW = PW_{PC}$ and $PIN = PIN_{SP}$.

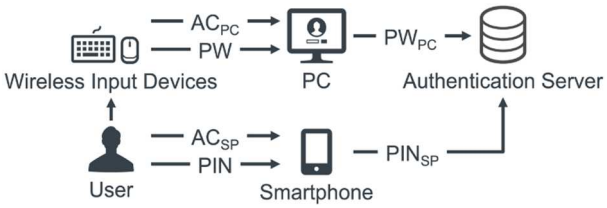


Fig. 3. Details of Two-factor Authentication

While the users (employee at their job) are working on their computer, the user is in front of the PC and the user's PC is always activated. In other words, there is no need to input the AC_{PC} to the PC again when the user accesses company's internal data. Thus, the PW (PW_{PC}) is an important security factor on the PC side. In general, it is not unusual at all, even for the legitimate user, to be prompted for the password before accessing confidential data. Therefore, inputting the PW into the PC is not a significant burden to the user. In contrast, as the user is unlikely to utilize the smartphone while working on the PC, the smartphone is not activated. Thus, the AC_{SP} must be input into the smartphone each time the user accesses company's resources. In practice, it is possible to keep the PIN in the smartphone and have the smartphone automatically send the PIN (PIN_{SP}) to the authentication server only when the AC_{SP} input from the user has been confirmed. Thus, the AC_{SP} is an important security factor on the smartphone side.

As a result, the tedious part of the procedure in Fig. 3 is inputting the AC_{SP}. Each time the user must be authenticated, the user must be aware of the smartphone, which has nothing to do with the user's work. With the widespread practice of working from home, company's internal data are being accessed from outside more frequently. Current two-factor authentication requires activation of the smartphone for each authentication. This significantly increases user effort and reduces work efficiency. Therefore, in terms of usability, there is a need for two-factor authentication that can avoid making the legitimate user aware of the smartphone (referred to as "Requirement 1").

C. *Existing Research on Improving the Usability of Two-Factor Authentication*

Bardram et al. proposed the idea of context-aware authentication in which the authentication method dynamically changes according to the user's situation [7]. This method confirms whether a client to be authenticated is in close proximity to a smart card held by a user. If the client and the smart card are in close proximity, user authentication is performed automatically. If proximity cannot be confirmed, the user is asked to input a password. By using context-aware authentication, the second factor can be employed automatically. When the PW (PW_{PC}) arrives at the authentication server, the server sends the second factor request to the smartphone. If the smartphone confirms its proximity to the PC, the PIN (PIN_{SP}) is sent automatically from the smartphone to the authentication server.

Fathy et al. investigated the usefulness of fully automatic face recognition system using videos recorded by the front camera of a smartphone [8]. Fully automatic face recognition can be used to automate receipt of the second factor for authentication. When the PW (PW_{PC}) arrives at the authentication server, the server sends the second factor request to the smartphone. If the smartphone is able to confirm the user's presence by means of a fully automatic facial recognition system, the PIN (PIN_{SP}) is sent from the smartphone to the authentication server.

As in these existing studies, if the smartphone automatically authenticates the user, meaning it remains active, it is possible to mitigate the usability degradation in two-factor authentication. However, in situations where the user's PC is infected by malware, such automated measures are insufficient. If malware sends the PW (PW_{PC}) to the authentication server in the background while the user (twiddling with their smartphone) is working on the PC, the smartphone will confirm the legitimate user and automatically send the PIN (PIN_{SP}) to the server. In other words, malware can easily break through the two-factor authentication while the user is working on their PC. Even when the second factor authentication is automated, it is still necessary to confirm the intention of the user to authenticate. Therefore, in terms of security, there is a need for two-factor authentication to confirm the intention of a legitimate user to authenticate (referred to as "Requirement 2").

III. MULTI-OBSERVED AUTHENTICATION

A. *Concept*

The requirements explained in the previous section for two-factor authentication are summarized below.

(Requirement 1) The legitimate user is not made aware of the smartphone.

(Requirement 2) The intention of the legitimate user to authenticate can be confirmed.

The key point of two-factor authentication is to establish a different authentication route from that of the PC. Therefore, we propose a new user authentication method called multi-observed authentication in which a single legitimate credential physically entered by the user into the PC is observed simultaneously on the smartphone. While the user inputs a credential into the PC, the legitimacy is authenticated via two routes (the PC and smartphone). Thus, Requirement 1 is satisfied [3]. Judging from the fact that both the PC and smartphone have received the legitimate credential from a single input device at the same time, it is expected that the credential is not sent by malware residing in the PC but physically inputted by a real human. Thus, Requirement 2 is also satisfied. As a result, the proposed method achieves authentication equivalent to two-factor authentication with both usability and security[4].

B. Authentication Procedure

This section describes the specific authentication procedure of the proposed method (Fig. 4). A variation of the authentication procedure are also described in Section V. The authentication procedure discussed in this section is referred to as Method A, and the variation in Section V.A as Methods B, respectively.

The basic flow of the proposed Method A is outlined below (Fig. 4). The credential used in the example is a password. To realize Method A, it is assumed that keyboard and mouse inputs to the PC are simultaneously input to the smartphone. Specifically, it is assumed that the wireless keyboard and wireless mouse (referred to as wireless input devices in this paper) are connected to the PC, and that the keyboard and mouse operations are modified so that they can also be received by the smartphone.

1. The user is working with the PC (i.e., the PC is already in activated state).
2. The user requests the authentication server to access information assets via the PC.
3. The authentication server instructs the smartphone to start listening to input from wireless input devices to the PC.
4. The smartphone starts receiving signals from the wireless input devices.
5. The authentication server instructs the PC to display a login page.
6. The PC displays the login page.
7. The user inputs the PW into the PC by using the wireless input devices.
8. The PC sends the PW to the authentication server.
9. The PW inputted by the user in Step 7 is simultaneously received by the smartphone. The smartphone also sends the PW to the authentication server.
10. The authentication server confirms the legitimacy of the PW received from the PC.
11. The authentication server confirms the legitimacy of the PW received from the smartphone.
12. If both Steps 10 and 11 are verified, the user is determined to be legitimate.
13. The authentication server instructs the smartphone to stop listening to input from wireless input devices to the PC.
14. The smartphone finishes receiving signals from the wireless input devices.

In Fig. 4, the PWs in Steps 7, 8, and 9 are shown as PW, PW_{PC}, and PW_{SP}, respectively, to distinguish them. However, in fact, PW, PW_{PC}, and PW_{SP} are identical. AC_{PC} is a credential for PC activation. By the time the user starts to work on the PC (prior to Step 1), the user has inputted AC_{PC} and logged into the PC. As explained in Section III.C, AC_{SP} (for smartphone activation) is not required in Method A.

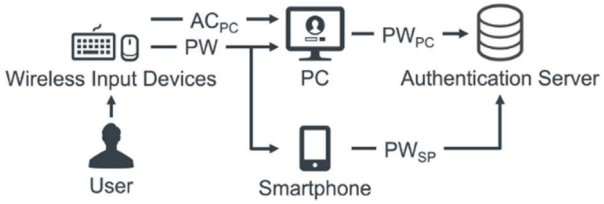


Fig. 4. Two-observed Two-factor Authentication

C. Evaluation

Here, Method A is evaluated in terms of usability (Requirement 1) and security (Requirement 2).

In Method A, all that is required of the user is to input the PW into the PC (in step 7). Thus, Method A satisfies Requirement 1. The user can use Method A in the same way as one-factor authentication (password authentication). As password authentication is familiar to the user, there are no mental barriers to the introduction of Method A. In contrast, the fact that the input to the PC is also sent to the smartphone may be a concern in terms of privacy. However, Method A is expected to limit the user's psychological burden, as the smartphone listens to the user's key input to the PC only during the period of credential input (from Step 4 to Step 14).

[3] One of the key challenges with password authentication is the risk of fraud and abuse through phishing, particularly because the stolen credentials can be reused. To protect against Adversary-in-the-Middle (AiTM) attacks, credentials are sent from the PC to the authentication server using client-authenticated TLS, allowing the server to verify the legitimacy of the PC.

[4] By allowing only the user's own smartphone to be used, the proposed method becomes a two-factor authentication system, combining what-you-know (a password) with what-you-have (a smartphone).

Notably, malware lurking in the PC can illegally send PW in Step 8. However, malware cannot physically manipulate the wireless input devices, so Step 7 cannot be performed. Therefore, the PW received by the smartphone in Step 9 can be determined as the PW inputted by the user in Step 7. In other words, the receipt of the PW from the smartphone in Step 9 is the basis for the assumption that the user intended to input the credential. Therefore, Method A also satisfies Requirement 2.

In existing two-factor authentication (Fig. 3), AC_{SP}, the credential for activating the smartphone, is required to confirm the user's intention for authentication using the smartphone. In contrast, Method A does not need to request AC_{SP} from the user, because the intention to authenticate is confirmed in Step 9. In other words, Method A can omit the confirmation of AC_{SP} for activating the smartphone (for this reason, the AC_{SP} is not shown in Fig. 4). This fulfills Requirement 2 (the legitimate user is not made aware of the smartphone) while satisfying Requirement 1 (the intention of the legitimate user to authenticate can be confirmed).

D. Authorization Procedure

To enhance convenience in multi-factor authentication, practices that separate authentication and authorization are widely adopted [9]. Authentication tokens are issued to users who have successfully completed multi-factor authentication upon initial access to information assets. These tokens have an expiry date, and within this period, presenting the authentication token (authorization) is sufficient for repeated access by the same user. However, this approach effectively reduces the authorization phase to single-factor authentication (only confirming possession of the authentication token). Consequently, malware on the PC can easily misuse the authentication token by simply waiting until the legitimate user obtains it.

This means that it is crucial to confirm the user's intention at the time of not only authentication but also authorization. This suggests that the concept and procedures of the proposed multi-observed authentication can be directly applied into authorization to enhance both usability and security of authorization. The flow of multi-observed authorization is similar to that of Method A, but due to space constraints, the details are omitted.

IV. User Experiment

To verify the convenience, privacy, and safety of the multi-observed authentication/authorization method proposed in the previous section, we implemented a file management system equipped with four types of security mechanisms: the conventional two-factor authentication/authorization using commonly used PCs and smartphones, and the proposed two-observed authentication/authorization method. We had 20 experiment participants (university students majoring in engineering/informatics) perform comparative experiments. After each participant had experienced the system, we conducted subjective assessments. We employed a questionnaire method for the assessment, asked them to rate the method on a 7-point Likert scale ("-3": Support for the conventional method, "0": neither, "3": Support for the proposed method) and to provide reasons for their assessments.

A. Authentication

In response to the question "Which method do you want to use from the perspectives of convenience and privacy?", the results were "-2" for 5%, "-1" for 30%, "0" for 5%, "1" for 30%, "2" for 20%, and "3" for 10%. The reason the experiment participants evaluated the proposed method favorably was that no operation on a smartphone was needed. The reason the experiment participants appraised the conventional method was their feelings of mistrust towards the transmission of credentials to a terminal different from their own operating PC in the proposed method. From this, it was confirmed that ensuring transparency is essential in multi-observed authentication. Furthermore, when asked how their respective assessments would change if the frequency of authentication increased, the result was a shift in the assessment toward the proposed method. From this, it was confirmed that in scenarios where the demand for convenience increases, the assessment of the proposed method (two-observed authentication) would rise.

B. Authorization

In response to the question "Which method do you want to use from the perspectives of convenience and safety?", the results were "-2" for 5%, "0" for 5%, "1" for 30%, "2" for 25%, and "3" for 35%. The reason the experiment participants evaluated the proposed method favorably was that safety improved with the trivial addition of a single mouse click. The reason the experiment participants appraised the conventional method was that even though it needed only a single click to confirm the user's intention, the conventional method was still more convenient than the proposed one. Furthermore, when asked how their respective assessments would change if the frequency of file access increased, the results were evenly split. This indicated that even in scenarios where the demand for convenience increased, half the users still supported the benefits of the proposed method (two-observed authorization). Making users aware of the proposed method's merit that "safety is ensured just by a single click" will contribute to enhancing the social acceptability of the proposed method.

V. Discussion

A. Modified Method with Keystroke Recognition (Method B)

Consider the case where a malicious user in collusion with the malware can physically access the legitimate user's PC. In that case, the malicious user can directly input the PW stolen by the malware into the PC while the legitimate user is away from the PC. In such an environment, it is not sufficient to

simply confirm that a legitimate credential (PW) was inputted into the PC. It becomes necessary to confirm that the PW, which is a legitimate credential, was inputted by a legitimate user (not by malware nor a malicious user). In such a case, in addition to confirming the legitimacy of the PW, another credential is needed to verify that the user who inputted the PW is legitimate.

To address the problem of requiring another credential, we propose a modification to Method A (Section III.B) by adding keystroke dynamics. Specifically, keystroke recognition is added to Step 11 of Method A using the characteristics of the keystroke operations when inputting the PW. This allows the authentication server to simultaneously achieve confirmation of the legitimacy of the PW (whether the input PW is legitimate) and legitimacy of the user (whether the user who input the PW is legitimate) in Step 11 while the legitimate user is still requested to simply input a single credential into the PC in Step 7. This enhances the password inspection in Step 11 of Method A. In this study, this was referred to as Method B.

B. Authentication

The proposed methods (Methods A and B) are authentication methods that assume the use of wireless input devices connected to both the PC and smartphone. As Bluetooth pairing is based on a one-to-one connection, a remodeling of the wireless communication protocol is required to implement the proposed method. Therefore, from the viewpoint of feasibility, the barriers to introducing the proposed method cannot be ignored.

In addition, vulnerabilities in wireless communication protocols pose a risk to the proposed method. Bad USB is a typical example of such a vulnerability, whereby the device drivers to be exploited are tampered with by malware without permission [10]. Under normal conditions, wireless input device communication is unidirectional (wireless input devices to PC). However, if malware replaces these with bidirectional device drivers, it is possible for malware residing in the PC to manipulate the wireless input devices. The solution will be discussed in a future study.

VI. CONCLUSIONS

In this paper, we proposed multi-observed authentication was proposed as a solution to the problem of reduced usability caused by the introduction of two-factor authentication. This was based on the concept of verifying a legitimate credential inputted by a human. After presenting the specific procedures of the proposed method (Method A), we evaluated Method A in terms of usability and security. Also, we implemented the proposed method (two-observed authentication/authorization) and carried out user experiments with 20 participants. With regards to future research directions, we plan to investigate ways to ensure transparency in the multi-observed authentication method and ways to increase users' awareness of the advantages of the multi-observed authorization method.

VII. ACKNOWLEDGMENT

This study was supported in part by JST Moonshot-type R&D project JPMJMS2215 and JSPS KAKENHI Grant Number JP23K28084.

REFERENCES

[1] Kaspersky Emotet modules and recent attacks, SECURELIST, Kaspersky (online), https://securelist.com/emotet-modules-and-recent-attacks/106290/ (accessed 2022-4-21).

[2] Ometov, A., Bezzateev, S., Mäkitalo, N., Andreev, S., Mikkonen, T., Koucheryavy, Y. "Multi-Factor Authentication: A Survey," Cryptography, vol. 2, no. 1, 2018. https://doi.org/10.3390/cryptography2010001.

[3] Microsoft: How it works: Azure AD Multi-Factor Authentication, Microsoft (online), https://docs.microsoft.com/en-us/azure/active-directory/authentication/concept-mfa-howitworks (accessed 2021-12-23).

[4] Amazon: Adding MFA to a user pool, Amazon (online), https://docs.aws.amazon.com/cognito/latest/developerguide/user-pool-settings-mfa.html (accessed 2021-12-23).

[5] Nozaki, S., Serizawa, A., Yoshihira, M., Fujita, M., Shibata, Y., Yamanaka, T., Matsuda, N., Ohki, T., Nishigaki, M.: Multi-observed Multi-factor Authentication: A Multi-factor Authentication Using Single Credential, International Conference on Network-Based Information Systems (NBiS), pp201-211, Sep. 2022. https://doi.org/10.1007/978-3-031-14314-4_20.

[6] Nozaki, S., Takaiwa, T., Fujita, M., Yoshimura, A., Ohki, T., Nishigaki, M.: Multi-Observed Authentication: A secure and usable authentication/authorization based on multi-point observation of physical events, Poster presentation in Network and Distributed System Security Symposium (NDSS), Mar. 2024.

[7] Bardram, J. E., Kjær, R.E., Pedersen, M. Ø.: Context-aware user authentication- supporting proximity-based login in pervasive computing. International Conference on Ubiquitous Computing. Springer, Berlin, Heidelberg, 2003. https://doi.org/10.1007/978-3-540-39653-6_8

[8] Fathy, M. E., Patel, V. M., Chellappa, R.: Face-based Active Authentication on mobile devices. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1687–1691, Apr. 2015. https://doi.org/10.1109/ICASSP.2015.7178258.

[9] "Kerberos Authentication Overview". https://learn.microsoft.com/enus/windowsserver/security/kerberos/kerberos-authentication-overview, (accessed 2023-11-28)

[10] Choi, B., Suh, T.: A Security Program to Protect against Keyboard-Emulating BadUSB. Journal of the Korea Institute of Information Security & Cryptology, vol. 26, no. 6, pp. 1483–1492, Dec. 2016.

Enhancing Remote Adversarial Patch Attacks on Face Detectors with Tiling and Scaling

Masora Okano*, Koichi Ito†, Masakatsu Nishigaki* and Tetsushi Ohki*
* Shizuoka University, Shizuoka, Japan
E-mail: okano@sec.inf.shizoka.ac.jp, {nisigaki, ohki}@inf.shizuoka.ac.jp
† Tohoku University, Miyagi, Japan
E-mail: ito@aoki.ecei.tohoku.ac.jp
‡ RIKEN AIP, Tokyo, Japan

***Abstract*—This paper discusses the attack feasibility of Remote Adversarial Patch (RAP) targeting face detectors. The RAP that targets face detectors is similar to the RAP that targets general object detectors, but the former has multiple issues in the attack process the latter does not. (1) It is possible to detect objects of various scales. In particular, the area of small objects that are convolved during feature extraction by CNN is small,so the area that affects the inference results is also small. (2) It is a two-class classification, so there is a large gap in characteristics between the classes. This makes it difficult to attack the inference results by directing them to a different class. In this paper, we propose a new patch placement method and loss function for each problem. The patches targeting the proposed face detector showed superior detection obstruct effects compared to the patches targeting the general object detector.**

I. Introduction

Deep neural network (DNN) models are susceptible to malicious manipulation of input data. Previous researches have employed a variety of approaches, including the use of specially designed spectacles that obscure key feature points of an object by reflecting light [1], as well as the development of sophisticated Adversarial Examples (AE) that subtly alter pixel values to cause the model to misclassify inputs intentionally. The threat of DNN attacks has emerged as a significant concern in light of the increasing incorporation of DNN-based models in a multitude of systems. Additionally, researches have explored the Remote Adversarial Patch (RAP), an attack that does not depend on the specific object being targeted. Unlike methods that directly alter features, RAP can remotely influence the model's output.

RAP was defined by Mirsky et al[2]. They identified two key characterisitics in their definition of RAP. One characteristic is the semantic alteration of model inference caused by patches, and the other is the ability of patches to be remotely attacked. Among these characteristics, the remote attackability of RAP is particularly significant. Here, the word " remote " means that the AP exists outside the target object's area. Compared to an AP that is placed on the target object, a RAP that does not directly modify the target object's characteristics is more difficult from an attack perspective. For the sake of simplicity, this paper defines RAP solely based on the characteristic of remote attackability.

This study focuses on the application of RAP to face detection. Although RAP has been extensively discussed for general object detection (multi-class object detection)[3], [4], the feasibility of RAP for face detection remains underexplored. Focusing on face detection is particularly important because obstructing it significantly enhances privacy protection. For example, applying a RAP to images before publication or capture could prevent unauthorized third parties from replicating the face region.

Two key challenges in applying RAP to face detection are as follows. **(1) the scale of detectable objects is diverse.** In the field of face detection, unlike the general object detection field, there is a requirement to detect obscure and distant objects, such as surveillance camera images. As a result, there is a tendency to focus on small features so that even extremely small faces can be detected. Small features have limited characteristics in the area around the face in the image that are included in the convolution. Consequently, the area that affects face detection is also small. Thus, the restrictions on where the patches can be placed are strict and are considered difficult to attack by RAP. **(2) fewer classes to classify.** Face detection involves fewer classes compared to multi-class object detection. In addition, because RAP operates from a distance and does not alter the object's intrinsic features directly. Therefore, it is difficult to redirect a class given to a particular object to another class with similar features, which limits the approaches that can be taken in an attack.

To address this issue, this study proposes a novel RAP method which involves two main processes: scaling and tiling. The **scaling process** adjusts the size of patches to correspond with different face scales during training, enhancing optimization across varying scales. The **tiling process** arranges the patches in a grid pattern, ensuring that any cropped region of the image will contain part of a patch. This approach addresses the issue by applying a tiling process that ensures any cropped region contains patches to some extent, and a scaling process that adjusts their relative size according to the scale variation of the face.

In addition to optimizing patches using the proposed patch applying method, we introduce a novel loss function for face detection obfuscation, called the **Borderline False Positives Loss**. The Borderline False Positives Loss is designed to

increase the number of false positives near the true face region and to disturb the coordinates of the true face. This enables false inferences on the face detection model without forcing the inference class towards the background class, which is significantly different from the face class. In other words, we overcome the attack limitations of having (2) fewer classes to classify, thereby realizing an effective attack method.

We conducted comprehensive experiments on the proposed method using multiple datasets and various RAP methods. Through these experiments, we demonstrated the effectiveness of the proposed method in terms of its obstruction performance against patch-based face detectors and its performance across diverse scales.

The contributions of this paper are listed below.

- We proposed a novel RAP method based on patch tiling and borderline false positive loss.
- Our comprehensive experiments showed consistent obstruction performance.
- Our proposed method also demonstrated consistent obstruction performance across datasets with varying face scales, showing robustness to face scale variation.

II. Related work

A. Adversarial Attack

Data that exploits vulnerabilities in a Deep Neural Network (DNN) model when attached to an image is called an Adversarial Patch (AP). Although AP generation methods were initially discussed for image classification models, Song et al. proposed the first AP generation method for object detection models [5]. Compared to image classification tasks, object detection tasks detect and label multiple objects in a scene, making it more difficult for AP to obstruct detection. APs have also been studied for person detection tasks[6]. Person detection is considered to be difficult to obstruct due to the large diversity within the person class.

However, all the studies discussed so far had the limitation that they must be located on a defined region of the detected object.To overcome the region limitation, Liu et al. proposed an adversarial patch, called DPatch[3], so that the effect of the attack is independent of the location.Their method is similar to that of Liu et al. but differs in that Liu et al. optimize the patch so that its region of presence is the only region proposed, whereas Lee et al. maximize the losses used by the model during training[4].

On the other hand, Mirsky et al. proposed a patch generation method that can apply AP to segmentation models and simultaneously defined RAP[2]. According to the definition given by Mirsky et al., the methods of Lee et al. and Liu et al. do not strictly fit the definition of RAP. However, in this paper, we define RAP as a method that can be attacked remotely, and therefore include these two methods as RAP.

B. Face Detection Obstruction

Obstruction of face detection has been discussed for the purpose of preventing unauthorized face image leakage due to unintentional capture of face images. Yamada et al. [1] proposed a detection jamming method that does not interfere with facial expression communication, in which facial features are modified by glasses irradiating near-infrared signals. AE [7] and AP [8] in face detection tasks have also been studied, but all of these methods involve processing of the face areas to be disturbed. Therefore, their applications are limited for privacy protection purposes. On the other hand, the patches generated by the proposed method do not require processing of face regions.

III. Method

The proposed method consists of a patch application process and a learning process. Each process is explained in the sections Section III-A and Section III-B, respectively. The overall process is shown in Fig. 1. Let I be a dataset consisting of N images, where each image can be represented as $\mathcal{I} = \{I_i \mid i = 0, \cdots, N-1\}$. The face detector F takes an image I_i as input and returns an inference result $d_i = \{d_{ik} \mid k = 0, \cdots, M-1\}$. Here, the inference result d_i consists of $M \geq 0$ face regions d_{ik}. Each face region d_{ik} includes the rectangular coordinates (x, y) of the rectangle's center, the width and height (w, h), as well as the confidence value p for the detection. It can be expressed as follows:

$$d_{ik} = (p_{ik}, x_{ik}, y_{ik}, w_{ik}, h_{ik}). \tag{1}$$

In addition, the Ground Truth (GT) of the correct face detection inference result for each image is g_i.

Let P be a adversarial patch, and let (w_P, h_P) be its width and height. Using the embedding function A to embed the patch in the image, the patched image I_i can be expressed as follows.

$$\tilde{I}_i = A(I_i, P). \tag{2}$$

Furtermore, let the face detection results for $\tilde{I}_i$ be denoted as $\tilde{d}_i$, with each element expressed similarly to d_i as follows.

$$\tilde{d}_i = \{\tilde{d}_{ij} \mid j = 0, \cdots, M-1\} \tag{3}$$

$$\tilde{d}_{ij} = (\tilde{p}_{ij}, \tilde{x}_{ij}, \tilde{y}_{ij}, \tilde{w}_{ij}, \tilde{h}_{ij}). \tag{4}$$

A. Learning Methods

1) Definition of Obstruction: For the detection $\tilde{d}_{ij}$, True Positive (TP) and False Positive (FP) are defined based on the IoU threshold value θ_D, and the number of cases where the detection of the ground truth (GT) fails is defined as False Negative (FN), as follows.

$$\tilde{d}_{ij} \text{ is TP} \iff \exists g_{ik} \text{ such that } IoU(\tilde{d}_{ij}, g_{ik}) \geq \theta_D \tag{5}$$

$$\tilde{d}_{ij} \text{ is FP} \iff \forall g_{ik} \text{ such that } IoU(\tilde{d}_{ij}, g_{ik}) < \theta_D \tag{6}$$

$$g_{ik} \text{ is FN} \iff \forall \tilde{d}_{ij} \text{ such that } IoU(\tilde{d}_{ij}, g_{ik}) < \theta_D \tag{7}$$

True Negative (TN) indicates whether the system can correctly identify areas where faces do not exist. However, in this paper, we will be focusing on the True Positive (TP), False Positive (FP) and False Negative (FN) values related to the face areas detected by the detector. A decrease in TP indicates

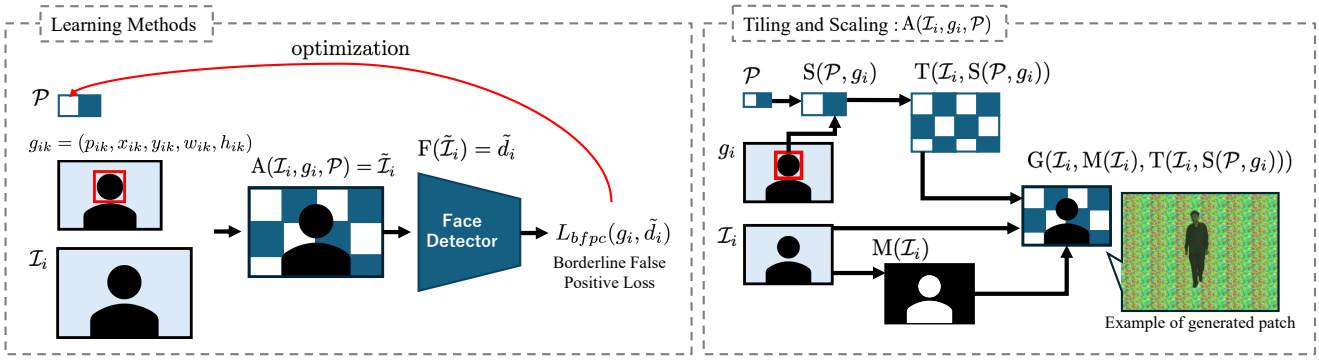


Fig. 1: Schematic diagram of the proposed patch generation method. We propose a patch application method and its learning method to explore the feasibility of generating RAPs to attack face detectors

that the face detector is failing to correctly identify face areas, while an increase in FP suggests that it is becoming more difficult to extract the correct face areas from the detection results. Thus, the efficiency of obstructing face detection can be evaluated based on the decrease in TP and the increase in FP.

2) Borderline False Positive Loss: In this study, we propose **Borderline False Positive Loss** to reduce True Positives (TP) and increase False Positives (FP). This loss function increases FPs around the boundary of the rectangle in g_i by obstructing the true face area, thereby reducing TP. Let $\mathcal{L}_{bfpc}$ denote the Borderline False Positive Loss, Equation (8), the objective function, Equation (9) the definition of loss function.

$$\min_{P}\{\mathcal{L}_{bfp}(g_i, \tilde{d}_i)\}. \tag{8}$$

$$\mathcal{L}_{bfpc}(g_i, \tilde{d}_i) = -\sum_{j=0}^{M} b_{ij} * log(1 - \tilde{p}_{ij}). \tag{9}$$

Here, b_{ij} is shown in the formula for the borderline variable in the Equation (10). b_{ij} is a borderline judgment variable that is 1 when the inference result for the image with the patch added, $\tilde{d}_{ij}$, is $\theta_T > a_{ij} \geq \theta_F$, which is the boundary between TP and FP, and 0 otherwise.

$$b_{ij} = \begin{cases} 0 & (a_{ij} \geq \theta_T) \ or \ (a_{ij} < \theta_F) \\ 1 & \theta_T > a_{ij} \geq \theta_F \end{cases}. \tag{10}$$

The values of θ_T and θ_F are IoU thresholds, set separately from the threshold θ_D used to classify $\tilde{d}ij$ as TP or FP. The following inequality must hold for each threshold: $\theta_T > \theta_D > \theta_F$. As shown in Equation (11), a_{ij} represents the maximum IoU value calculated between $g_i = \{g_{ik} \mid k = 0, \cdots, M-1\}$ and $\tilde{d}_{ij}$.

$$a_{ij} = \max_{g_{ik} \in g_i} IoU(g_{ik}, \tilde{d}_{ij}). \tag{11}$$

The loss function, $\mathcal{L}_{bfpc}$, increases the confidence value of the inference results that may be false positives based on the borderline decision variable, b_{ij}. The boundary judgment variable b_{ij} is a variable that classifies the inference results to be the judgment boundaries between TP and FP based on the maximum IoU value a_{ij} obtained between g_i and $\tilde{d}_{ij}$. The loss function, $\mathcal{L}_{bfpc}$, is designed to mislead the coordinates of the inference results that would otherwise be true positives into false positives by increasing the number of false positives around the true face region using b_{ij}.

B. Patch Applying Process

The expression for the patch application process A is shown as Equation (12), and the process of the patch application process is shown in the figure Fig. 1 on the right.

$$A(I_i, g_i, P) = G(I_i, M(I_i), T(I_i, S(P, g_i))). \tag{12}$$

The function M takes the image I_i as an argument and converts it into a reference image for masking only the foreground. The function G takes as arguments three images(foreground, mask and background) of equal height and width, and performs foreground-background composition. The scaling function S and the tiling function T are both functions that perform transformations on the patches that are optimized. In applying patches, the scaling and tiling processes play a particularly important role.

1) Scaling: The scaling process is the $S(P, g_i)$ on the rigth side of Fig. 1 . By making the scaling function S align the area ratio of patches and faces, it encourages patches to be able to obstruct faces of various scales uniformly during learning. In the scaling process, patch P is resized to have a width of w_{PS} and a height of h_{PS}. Using patch P and the corresponding GT g_i, the width w_{PS} and height h_{PS} of the scaled patch are expressed by the following equations.

$$\begin{aligned} (w_{PS}, h_{PS}) &= S(P, g_i) \\ &= (round(w_P * s), round(h_P * s)). \end{aligned} \tag{13}$$

$$s = \sqrt{\frac{\alpha * w_{ik} * h_{ik}}{h_P * w_P}}. \tag{14}$$

$$k = \arg\max_{k} w_{ik} * h_{ik}. \tag{15}$$

2) Tiling: The $T(I_i, S(P, g_i))$ on the right side of Fig. 1 is a tiling process. The tiling function T is a process that converts a patch P^S into a patch tile P^T of the same height and width as the image by tiling the patch P^S horizontally and vertically.

Reducing the dependence of face detection on face position by ensuring that the pixels of the patch are included when any face in any area is extracted as a feature. Each coordinate $P^T[x, y]$ of P^T can be determined using the tiling function T as follows.

$$P^T = T(I, P^S). \quad (16)$$

$$P^T[x, y] = P^S[x \bmod w_{PS}, y \bmod h_{PS}]. \quad (17)$$

The edges of the patches created by tiling will be cut off to match the image size.

IV. Experiment

A. Experimental Setup

1) Dataset: Two datasets are utilized in the experiment: CASIA Gait B (CGB) [9] and FaceForensics++ (FFP) [10]. CGB is designed for gait recognition and includes images captured in a controlled indoor environment. The dataset features 124 subjects, each captured from 11 different angles. FFP, on the other hand, is a dataset for detecting DeepFake videos and comprises 1000 videos featuring frontal faces without occlusion, collected from Youtube press conference videos. In the experiment, only frontal faces were extracted from the CGB to match the experimental conditions with the FFP containing frontal faces. Images without facial regions were removed from these videos using S3FD to avoid duplication, and the training and validation datasets of 3000 images each were extracted. In addition, for the purpose of detection obstruction, GT is the inference result d_i of the image I_i before patch application. Therefore, in the experiment, $d_i = g_i$.

2) Preprocessing: For these moving images, a mask image and $\hat{D}_i$, the GT, are created in advance. In this experiment, rembg[11] was used to create the mask image and S3FD was used to create $\hat{D}_i$.

3) Learning Patch: For the various hyper-parameters, the IoU threshold $\theta_D = 0.5$, and the scaling parameter $\alpha = 5.58$ and the loss function parameters $\theta_T = 0.6, \theta_F = 0.3$ are determined from empirical results. For optimization, the Nesterov Iterative Fast Gradient Sign Method (NI-FGSM)[12] is used in the proposed method. In addition, the methods of DPatch[3] and Lee et al.[4] are used for comparison.

4) Evaluation Methods: The F value and the Average Precision (AP) value are used to compare how much the detection performance is degraded compared to the situation without patch obstructions. The number of TPs, FPs, and GTs are also used to compare obstruction performance. The reason for not simply using F and AP values is that F values are insensitive to changes in the number of TPs when the number of FPs becomes extremely large relative to the number of TPs, and AP values are insensitive to increases in the number of FPs, making it difficult to correctly compare obstruction performance.

B. Detection Obstruction Performance Evaluation

We present a comparison of the proposed RAP generation method with other methods. DPatch[3] and Lee et al.'s method[4] are used for the comparison. The coordinates on the image where the two patches to be compared are placed are determined randomly, as in the case of both patches generation.S3FD is used for face detection. The experimental results are shown in I. Table I shows that the proposed method has fewer TP than other methods on both datasets. This indicates that Borderline False Positive Loss can successfully disturb the coordinate of the true face. However, as shown in Table I, the increase in FP with the proposed method is not as pronounced as in the method by Lee et al. This suggests that further increasing FP may enhance the effect of the Borderline False Positive Loss.

TABLE I: Comparison of detection obstruction performance

dataset	method	F	AP	GT	TP	FP
CGB	Dpatch	9.96e-1	1.0	3000	2977	0
	Lee	1.34e-1	4.14e-2	3000	2967	38235
	Proposed	7.91e-1	9.99e-1	3000	**1967**	7
FFP	Dpatch	9.13e-1	9.87e-1	3315	3305	617
	Lee	1.96e-1	9.29e-2	3315	3287	26866
	Proposed	9.78e-1	9.99e-1	3315	**3227**	56

When CGB is used as the training data, the patches generated by the proposed method reduce TP when verified using CGB. However, when FFP is used to learn and verify patches, the degree of suppression of TP is less than when CGB is used. We think that the reason for this difference in the dataset is that CGB contains many small faces, while FFP contains almost large faces. For smaller faces included in the CGB, the IoU fluctuates steeply in overlapping regions, making it easier to meet the threshold requirements.

C. Positional Robustness Evaluation

When detection is obstructed for different images, the relative positional fluctuations between the patch and the obstructed object due to the difference in the coordinates of the obstructed object for each image are considered to affect the detection obstruction performance. Therefore, we verify whether detection obstruction is robust to such position variations by comparing its position independence with the DPatch[3] and Lee et al.'s method[4]. The inference results for all images in the test set with the patch applied are classified into TP and FP based on the ground truth, and the frequency for each upper right coordinate is tabulated and plotted as a heat map. The two patches to be compared are fixed at the coordinate $(0, 0)$ to illustrate the effect of the patch on detection. Ideally, the verification should be performed on a dataset where the coordinates of the obstacle targets are evenly distributed throughout the image. In reality, however, it is difficult to prepare such a dataset. Therefore, we created a coordinate uniform data set that met the ideal conditions in a pseudo-way, and conducted a verification.

Coordinate Uniform Data Set In order to show that it is possible to obstruct detection no matter where the face is located in the image, we created a dataset in which the positions of the face regions are uniformly distributed throughout the image, based on 10 images randomly extracted from the

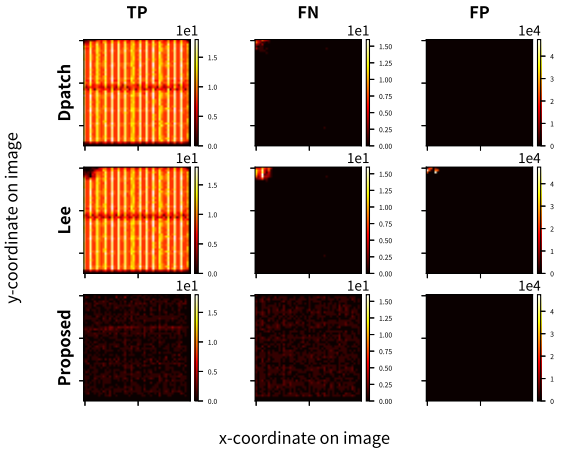


Fig. 2: Detection results for the coordinate uniform dataset. Top row: DPatch, middle row: patches by Lee et al. and bottom row: patches by the proposed method. From left to right: TP, FN, FP

CASIA Gait B test set. The coordinate uniform dataset was created by shifting the image based on the top left coordinate of the face region and repeating the operation from the top left to the bottom right of the image. In this case, the width of each movement was set to 25 pixels. The coordinate-uniform dataset was created by inferring the face region before applying the tiling adversarial background image, and then excluding images for which there were no inference results. The total number of images in the dataset is 24,654. The results of the experiment are shown in Fig. 2.

The verification results show that the Dpatch and Lee et al. patches each have a concentration of FN in the vicinity of the coordinates where the patches are located $(x, y) = (0, 0)$. In contrast, the proposed method distributes the FNs throughout the image. This result demonstrates the effectiveness of the proposed method's tiling process, and shows one advantage over previous research. This result is also one basis for supporting the hypothesis proposed in Section I that the placement coordinates of the patches are highly constrained when disturbing face detection.

D. Dataset Transferability

For the purpose of protecting privacy, it is desirable that the generated patches can obstruct face detection in various situations. Therefore, we will swap the data sets between the training and validation phases, and verify whether the obstructing performance of the patches learned with the known data set can obstruct the unknown evaluation data set as well as the known training data set. If it is possible to obstruct face detection in various situations, it should be possible to obstruct other datasets with different shooting conditions in the same way as the training dataset. In addition, we will consider the possibility that the transferability to unknown datasets may differ depending on the face detection model used during training, and we will conduct verification using multiple models.

The patch is tested using a different dataset from the one used for learning. We conduct experiments for each of the three models MTCNN[13], S3FD[14], and RetinaFace[15] and compare them. The experimental results are shown in TableII.

TABLE II: Table for dataset transferability

train/test dataset	model	F	AP	GT	TP	FP
CGB / CGB	MTCNN	2.92e-1	1.0	3000	513	0
	S3FD	7.91e-1	9.99e-1	3000	1967	7
	RetinaFace	7.87e-1	1.0	3000	1947	0
CGB / FFP	MTCNN	9.14e-1	9.99e-1	3315	3135	407
	S3FD	9.78e-1	9.99e-1	3315	3226	58
	RetinaFace	9.72e-1	9.99e-1	3315	3147	10
FFP / FFP	MTCNN	9.18e-1	9.99e-1	3315	3132	371
	S3FD	9.78e-1	9.99e-1	3315	3227	56
	RetinaFace	9.72e-1	9.99e-1	3315	3147	10
FFP / CGB	MTCNN	2.92e-1	1.0	3000	512	0
	S3FD	7.97e-1	9.99e-1	3000	1993	10
	RetinaFace	7.92e-1	9.99e-1	3000	1968	1

The patches generated using CGB have kept the TP below 2,000 in the verification using CGB. In addition, patches generated using FFP kept the TP to about 3,200 in the verification using FFP. All the models tended to have the same tendency in terms of transferability to unknown data sets. When referring to the results of verifying the patches learned with FP using CGB, TP is less than 2000 for all models. In contrast, when the same verification process is conducted using CGB with FFP, the results do not exhibit a similar trend.

The reason for this difference is thought to be that FFP contains many high-resolution images of faces of various sizes compared to CGB, and is more difficult as a detection obstruction task. The patches trained on FFP have high-resolution and obvious facial features, which may have contributed to the enhancement of universal optimization independent of the dataset, although the optimization did not work well for the faces included in FFP. As a result, it may have shown higher interference only for CGB containing low-resolution, ambiguous facial features. Since the purpose of scaling is to find patterns that can obstruct the detection of faces of any scale, we believe that the obstructing performance of FFP against CGB is a result that shows the effectiveness of the scaling of the proposed method.

V. Discussion

Multiple faces in the same scene : In this experiment, we did not consider the case where there are multiple faces in the scene to be disturbed, but only the case where there is a single face. This is because scenes where there is a single face are assumed to be online meetings and press conferences. By performing face detection obstruction on these videos, it is possible to protect the privacy of the photographed person, and it is effective. We plan to consider the case where there are multiple faces in the future.

Improvement of the loss function : One possible reason for the smaller number of FP cases in Section IV-B than in

other methods is that the optimization process did not progress because the loss value often took the value 0 in the early stages of optimization. In order for the value calculated by the Borderline False Positive Loss to be non-zero, the IoU calculated between the correct face region and the inferred detection needs to be less than $\theta_T = 0.6$. On the other hand, for images without patches applied, the IoU will be close to 1. Therefore, it is likely that the loss value did not change from 0 and optimization did not progress. To solve this, it seems that it is necessary to have an objective function that reduces the iou between the correct face region and the estimated detection results from 1 to around 0.6.

VI. Conclusion

We discussed the difficulty of implementing RAP that targets face detectors and presented the diversity of face region scales and the small number of classification classes as reasons for this. In contrast, this study proposed a unique patch placement method and loss function as a solution to each of these problems.

As a result of patching using the proposed method, it showed fewer TP than other methods. In addition, the patches learned on a dataset that tends to include large faces showed a certain level of obstruction performance on a dataset that tends to include small faces, indicating that they can cope with the diversity of face sizes. Comparison with patches generated by other methods shows that the proposed disturbance method contributes to TP reduction even for face detectors with a small number of classification classes.

The proposed method still has room for improvement in both TP and FP, and further examination is needed regarding the refinement of the loss function and parameter optimization.

Acknowledgement

This work was supported in part by JSPS KAKENHI JP 23H00463, JP 23K28084 and JST Moonshot R&D Grant Number JPMJMS2215.

References

[1] T. Yamada, S. Gohshi, and I. Echizen, "Privacy visor: Method for preventing face image detection by using differences in human and device sensitivity," in *Communications and Multimedia Security: 14th IFIP TC 6/TC 11 International Conference, CMS 2013, Magdeburg, Germany, September 25-26, 2013. Proceedings*, vol. 8099, books.google.com, 2013, p. 152.

[2] Y. Mirsky, "Ipatch: A remote adversarial patch," *Cybersecurity*, vol. 6, no. 1, p. 18, 2023.

[3] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, and Y. Chen, *DPatch: An adversarial patch attack on object detectors*, arXiv/1806.02299, Jun. 2018.

[4] M. Lee and Z. Kolter, *On physical adversarial patches for object detection*, arXiv/1906.11897, Jun. 2019.

[5] D. Song, K. Eykholt, I. Evtimov, *et al.*, "Physical adversarial examples for object detectors," in *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018.

[6] S. Thys, W. Van Ranst, and T. Goedeme, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2019.

[7] C. Zhang, Y. Qi, and H. Kameda, "Multi-scale perturbation fusion adversarial attack on MTCNN face detection system," in *2022 4th International Conference on Communications, Information System and Computer Engineering (CISCE)*, IEEE, May 2022, pp. 142–146.

[8] X. Yang, F. Wei, H. Zhang, and J. Zhu, "Design and interpretation of universal adversarial patches in face detection," in *Computer Vision – ECCV 2020*, ser. Lecture notes in computer science, Cham: Springer International Publishing, 2020, pp. 174–191.

[9] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *2011 18th IEEE International Conference on Image Processing*, Sep. 2011, pp. 2073–2076.

[10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 1–11.

[11] D. Gatis, *Rembg: Rembg is a tool to remove images background*, en.

[12] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, *Nesterov accelerated gradient and scale invariance for adversarial attacks*, arXiv/1908.06281, Aug. 2019.

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[14] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: Single shot scale-invariant face detector," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 192–201.

[15] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, openaccess.thecvf.com, 2020, pp. 5203–5212.

Detecting Abnormal Machine Sounds Using An Ensemble Approach with Data Augmentation Techniques

Po Cheng Chan[1,2], Chung Li Lu[1], and Jia Ching Wang[2]

[1]Advanced Technology Laboratory, Chunghwa Telecom Laboratories, Taoyuan, Taiwan
[2]Department of Computer Science and Information Engineering, National Central University
[1]Email: *cbc@cht.com.tw, chungli@cht.com.tw*
[2]Email: *jcw@csie.ncu.edu.tw*

***Abstract*—In this paper, we present our methodology in the DCASE 2021 Challenge, specifically Task 2: Unsupervised Detection of Anomalous Sounds in Machine Condition Monitoring under Domain Shift Conditions. The primary goal was to ascertain whether sounds emitted by a machine are normal or anomalous, utilizing a test dataset devoid of any abnormal sounds in its training counterpart. To tackle this challenge, we utilized deep learning models, namely MobileNetV2 and ResNetV2-50, which were enhanced with data augmentation techniques like mix-up to improve anomaly detection in machine sounds. The results of our evaluation are presents the harmonic mean AUC and pAUC across various machines including ToyCar, ToyTrain, Fan, Gearbox, Pump, and Slider. Notable results are an AUC of 54.77 and pAUC of 58.83 for ToyCar, with other machines showing varied performance metrics, such as a 44.10 AUC for Gearbox and a 59.02 AUC for Slider.**

I. Introduction

The DCASE 2021 Challenge Task 2 addresses the issue of misclassifying normal machine sounds as anomalous due to shifts in normal operating conditions, also known as domain shifts [1]. In real-world scenarios, various factors such as changes in production demands or environmental conditions can alter how a machine operates. The core challenge is to develop a model that can effectively differentiate between normal and abnormal machine sounds. Changes in production demands may include variations in operation speed, machine load, viscosity, heating temperatures, and environmental noise, which can fluctuate across different seasons. For example, a pump with a minor leak, a slide rail lacking grease, or a fan experiencing voltage fluctuations might still appear functional on visual inspection, but can be acoustically identified as faulty due to unique sound patterns. Such conditions can reduce the signal-to-noise ratio, complicating the detection of anomalous operations amid background industrial noise. Deep learning approaches are advantageous as they can extract hierarchical, discriminated features from data, a method that diverges from traditional machine learning techniques and is highly adaptable across various scenarios [2]. These models do not require extensive engineering knowledge of the system and can handle the unknown dynamics of machine faults [3].

The task presents three primary challenges: First, identify an effective deep neural network architecture for detecting abnormal sounds. Second, acquire a robust dataset that provides both a significant quantity and a variety of sound samples for training. Third, train the chosen architecture from start to finish, specifically optimized for this task.

II. DCASE 2020 Challenge Task 2 Database and Setup

The data set for DCASE 2021 Task 2 comprises ToyADMOS2 [4] and MIMII [5] DUE datasets, featuring 10-second audio recordings from various machines and their associated equipment, along with environmental sounds. The dataset categorizes into seven machine types, including ToyCar and ToyTrain from ToyADMOS2, and Valve, Pump, Fan, and Slider from MIMII DUE. As indicated in Table I, each category consists of two domains—source and target—each providing a complete set of training and testing data. The training data set for each section includes approximately 1,000 clips of normal sounds from the source domain and three clips of normal sounds from the target domain. For testing, each domain contains 100 clips of normal sounds and 100 clips of real anomalous sounds. The spectrograms of a fan in normal and abnormal conditions are shown in Figure 1 and 2. We must distinguish the subtle differences in the spectrograms to classify the abnormal sounds.

TABLE I: OVERVIEW OF DOMAIN DATASET IN TRAINING AND TESTING.

Dataset / Domain	Source		Target	
	Normal	**Abnormal**	**Normal**	**Abnormal**
Training	1000	0	3	0
Testing	100	100	100	100

The anomaly score calculator A with parameter θ. The input of A is the audio clip x and additional information, and one anomaly score $A_\theta(x)$ is output. Then, the machine is determined to be anomalous when the anomaly score $A_\theta(x)$ exceeds a pre-defined threshold value as $\emptyset$:

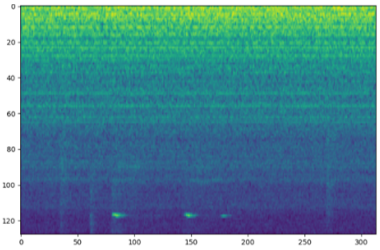


Fig. 1: Fan in normal condition

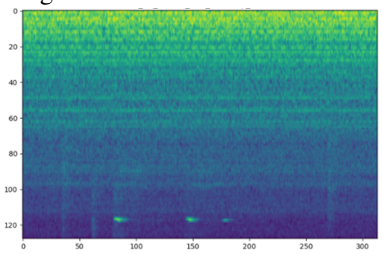


Fig. 2: Fan in abnormal condition

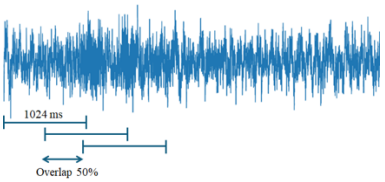


Fig. 3: Waveform segmentation

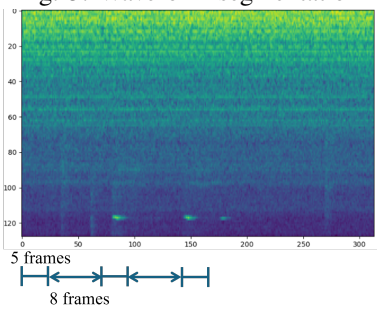


Fig. 4: Mel spectrograms segmentation

$$\text{Decision} = \begin{cases} \text{Anomaly} & \text{if } A_g(x) > \phi \\ \text{Normal} & \text{otherwise} \end{cases} \tag{1}$$

This task is evaluated with the AUC and the pAUC. The pAUC is an AUC calculated from a portion of the ROC curve over the pre-specified range of interest. In our metric, the pAUC is calculated as the AUC over a low false-positive-rate (FPR) range $[0, p]$. The reason for the additional use of the pAUC is based on practical requirements. If an ASD system frequently gives false alarms frequently, we cannot trust it. Therefore, it is important to increase the true-positive rate under low FPR conditions. In this task, we will use $p = 0.1$.

III. Methodology

In this section, we begin by detailing each sub-task of our ensemble individually, followed by an explanation of how these elements are integrated to produce the final scoring.

A. Feature Extraction

For the spectrogram generation, we utilized the librosa toolkit to extract Mel spectrograms from the input audio. Each audio sample is represented by a Mel spectrogram with 128 frequency bins, using a window size of 1024 ms and an overlap of 50% as shown in Figure 3. The spectrogram data is log-scaled. To derive the acoustic feature, we concatenate the log Mel filterbank outputs of several preceding and succeeding frames (5 frames in total) . After calculating the log Mel spectrogram for each audio clip, the acoustic feature is formed by this concatenation process involving 5 frames of log Mel filterbank outputs.

B. Data Augmentation

For data augmentation, we tried as follow: The random noise from data:

$$y[n] = x[n] + 0.005 \cdot z[n] \tag{2}$$

where z is an random uniform $[0, 1]$ noise from audio, y is the audio signal from machine ID's signal and z.

The mixed output is combined by audio and Gaussian noise that RMS of the noise generated:

$$y[n] = x[n] + \text{RMS}_{\text{noise}} \tag{3}$$

$$\text{RMS}_{\text{noise}} = \sqrt{\frac{P_{\text{signal}}}{10^{\text{SNR}/10}}} \tag{4}$$

The time stretching is stretch a spectrogram in time for a given rate r:

$$y_{\text{speed}}[n] = x\left[\frac{n}{r}\right] \tag{5}$$

For frame shift, we used the toolkit numpy to shift 16000 points to right:

$$y[n] = x[n + 16000] \tag{6}$$

Mix up:

$$y[n] = \alpha \cdot x_1[n] + (1 - \alpha) \cdot x_2[n] \tag{7}$$

where y is a mixture features of x_1 and x_2 are log-mel-filterbank outputs of machine section ID. We also multiply the one-hot encoded label with α.

C. Autoencoder Model Architecture

The autoencoder (AE) model, identical to that used in DCASE 2020 Task 2 and released by [6], follows a specific structure in its layers. The architecture of the hidden units in each layer is organized as follows: [128, 128, 128, 128, 32, 128, 128, 128, 128]. This configuration compresses 5 frames of 128 mel-energies into an 8-dimensional space as shown in Figure 4. The model incorporates Batch Normalization and ReLU non-linearities for activation, and it is trained using the Adam optimizer.

D. MobileNetV2 and ResNetV2-50 Models Architecture

The DCASE 2021 Task 2 utilizes the MobileNetV2 model, as detailed in [7], and the ResNetV2-50 model, referenced in [8], both of which are extensively used in computer vision tasks like object detection, classification, and semantic segmentation. These models are employed to compress 5 frames of 128 mel-energies into an 8-dimensional space as as shown in Figure 4. Training is conducted using the ADAM optimizer with a learning rate of 0.00001. The training process is halted after 20 epochs, and a batch size of 32 is maintained. Each machine type is addressed individually, with models being trained using normal audio clips from all sections pertinent to that specific machine type.

IV. EXPERIMENTAL RESULTS

We also experimented with varying the spectrogram settings for each machine type, but these alterations did not surpass the performance of the standard audio representation as shown in Table II and III. Notably, time stretching demonstrated effective results in the ToyCar category when combined with development dataset training and other data augmentation methods such as noise addition, Gaussian noise, and frame shifting, although these latter techniques did not enhance accuracy.

TABLE II: Default Parameters Setting

Parameter	AutoEncoder	MobileNetV2	ResNetV2-50
batch size	512	32	32
no. Mels	128	128	128
no. FFT	1024	1024	1024
hop	512	512	512
hop frame	1	8	8

We evaluated four distinct systems that utilized a combination of data augmentation, additional training datasets, and ensemble models. The arithmetic mean AUC scores for each machine type across these four systems are presented in Tables IV and V. Through our data augmentation trials, we discovered that the mix-up technique yielded superior performance for certain machine types. However, models trained with both the development and additional datasets underperformed compared to the baseline.

Ultimately, Our evaluation results are as shown in Table VI, each selected based on the best performance observed for individual machine types using both the development and additional datasets.

REFERENCES

[1] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, and D. N. et al., "Descritpion and discussion on dcase 2021 challenge task2: Unsupersied anomalous sound detection for machine condition monitoring under domain shifted conditions," in *arXiv:2106.04492*, 2021.

[2] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," in *arXiv:1901.03407*, 2019.

[3] G. S. Galloway, V. M. Catterson, T. Fay, A. Robb, and C. Love, "Diagnosis of tidal turbine vibration data through deep neural networks," in *Proceedings of the 3rd European Conference of the Prognostics and Health Management Societry*, IEEE, 2016.

[4] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Y. et al., "Toyadmos2: Another dataset of miniaturemachine perating sounds for anomalous sound detection under domain shift conditions," in *arXiv:2106.02369*, 2021.

[5] R. Tanabe, H. Purohit, K. Dohi, and Y. i. e. a. T. Endo, "Mimii due: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due o changes in operational and environmental conditions," in *arXiv:2105.02702*, 2021.

[6] H. Purohit, R. Tanabe, K. Ichige, T. Endo, and Y. N. et al., "Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *arXiv:1909.09347*, 2019.

[7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlen ecks," in *In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, 2018, pp. 4510–4520.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *In Proceedings of European Conference on computer vision*, Springer, 2016, pp. 630–645.

TABLE III: Varied Machine Parameters Setting

	Machine						
Parameter	**ToyCar**	**ToyTrain**	**Fan**	**Gearbox**	**Pump**	**Slider**	**Valve**
batch size	64	64	64	64	64	64	128
no. Mels	128	128	128	128	128	256	128
no. FFT	1024	1024	1024	1024	1024	1024	1024
hop	80	512	512	512	80	80	512
fmin	10	0	1	0	100	10	0
fmax	4000	4000	4000	8000	7700	7700	8000

TABLE IV: Arithmetic mean AUC for baseline and system. The results of AE-Baseline1, MobileNetV2-Baseline, System1 (MobileNetV2 with mix up), System2 (ResNet50V2 with mix up) and System3 (MobileNetV2 and ResNet50V2 with mix up) are in the same dev set. The results of System4 (MobileNetV2 and ResNet50V2 with mix up) that training set is composed of dev set and additional set.

System	**Ae-Baseline**		**MobileNetv2 Baseline**		**System1**		**System2**		**System3**		**System4**	
Machine	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**
ToyCar	62.81	52.59	56.17	57.32	57.94	58.02	58.96	55.84	60.56	56.88	56.27	61.17
ToyTrain	62.93	55.01	60.55	51.74	54.02	50.23	49.83	50.86	51.67	51.09	47.84	51.93
Fan	64.30	53.56	64.32	65.57	68.66	67.97	66.96	67.10	68.04	67.75	37.04	53.40
Gearbox	66.63	52.26	61.73	54.26	63.39	53.95	65.41	58.24	69.81	62.02	40.63	49.25
Pump	64.21	54.65	65.95	59.96	65.97	60.14	73.48	64.38	71.86	63.41	29.81	49.27
Slider	69.59	56.31	64.65	57.39	63.33	57.40	58.99	60.70	65.62	57.67	35.66	51.02
Valve	53.68	50.78	57.88	55.88	59.15	55.79	63.48	55.57	61.61	55.69	41.56	49.12

TABLE V: Harmonic mean AUC for baseline and system. The results of AE-Baseline1, MobileNetV2-Baseline, System1 (MobileNetV2 with mix up), System2 (ResNet50V2 with mix up), and System3 (MobileNetV2 and ResNet50V2 with mix up) are in the same dev set. The results of System4 (MobileNetV2 and ResNet50V2 with mix up) indicate that the training set is composed of both dev set and additional set.

System	**Ae-Baseline**		**MobileNetv2 Baseline**		**System1**		**System2**		**System3**		**System4**	
Machine	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**
ToyCar	62.17	52.53	53.91	56.26	54.65	56.63	56.34	54.90	58.54	55.81	51.58	60.00
ToyTrain	61.79	53.87	56.66	51.59	52.36	50.08	47.46	50.71	49.77	50.90	45.61	51.53
Fan	63.50	53.38	61.09	63.62	64.27	65.75	60.51	64.67	62.93	65.55	30.26	52.78
Gearbox	65.73	52.23	59.75	53.62	59.41	53.04	61.53	57.30	66.53	60.93	40.28	49.22
Pump	62.85	54.38	64.68	59.09	64.94	59.16	72.20	63.26	70.75	62.32	20.28	49.22
Slider	67.35	55.79	61.40	56.28	59.65	56.51	53.90	59.80	63.47	56.96	27.45	50.73
Valve	53.35	50.69	57.36	55.58	58.51	55.26	62.83	55.21	60.82	55.21	36.27	49.06

TABLE VI: Harmonic AUC and pAUC in evaluation

Harmonic mean												
ToyCar		**ToyTrain**		**Fan**		**Gearbox**		**Pump**		**Slider**	
AUC	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**	**AUC**	**pAUC**
54.77	58.83	47.67	50.37	47.82	50.56	44.10	51.72	51.86	51.07	59.02	53.30

Leveraging Attention Mechanisms for Breast Cancer Diagnosis

Akumalla Brahma Reddy*, Bach-Tung Pham*, Tung-Yu Zhuang *, Bima Paristao ‡, Pao-Chi Chang†, Jia-Ching Wang*
*Department of Computer Science and Information Engineering, National Central University, Taiwan
†Department of Communication Engineering, National Central University, Taiwan
†Department of Informatics, Institut Teknologi Kalimantan, Indonesia
Email: jcw@csie.ncu.edu.tw

***Abstract*—Recent advancements in deep learning have significantly impacted breast cancer diagnosis, enhancing the accuracy and efficiency of detection and classification. In this study, we applied various attention mechanisms in different configurations to evaluate their effects on our model's performance, specifically in terms of F1 score and accuracy. We utilized ConvNeXt for initial feature extraction, followed by the implementation of attention mechanisms such as SRM (Spatial Relation Module), GCT (Global Contextual Transformer), LCT (Local Contextual Transformer), and Triplet Attention. These attention mechanisms were arranged both in parallel and series configurations to explore their impact on feature extraction. Our experiments demonstrated that the choice and arrangement of attention mechanisms significantly influenced the model's performance, with the Triplet_LCT in the parallel configuration achieving the highest accuracy and F1 score.**

I. Introduction

Breast cancer is a disease where cells in the breast grow uncontrollably. It is one of the most common cancers affecting women worldwide, though it can also occur in men. Breast cancer can manifest in various forms, primarily categorized into benign (non-cancerous) and malignant (cancerous) masses and calcifications.Benign masses are non-cancerous growths in the breast. They typically have smooth, well-defined edges and are often oval or circular in shape. Common types of benign breast masses include fibroadenomas and cysts. While benign masses are not life-threatening, they may require monitoring or removal if they cause discomfort or other issues.Malignant masses, on the other hand, are cancerous and can invade surrounding tissues. These masses often have irregular shapes and poorly defined edges. Malignant tumors can spread (metastasize) to other parts of the body, making early detection and treatment crucial. The most common types of malignant breast tumors include invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC)[1].

Breast calcifications are small deposits of calcium that appear as white spots on a mammogram. They are common, especially in women over 50, and can be classified into two types: microcalcifications and macrocalcifications.Microcalcifications are tiny calcium deposits that may indicate the presence of breast cancer, particularly ductal carcinoma in situ (DCIS), which is an early form of breast cancer. These calcifications often appear in clusters and may require further investigation through a biopsy to determine if they are benign or malignant .Macrocalcifications are larger, coarser calcium deposits that are usually benign. They are often associated with aging, past injuries, or benign breast conditions such as fibroadenomas or cysts. Macrocalcifications typically do not require further testing unless they appear in a suspicious pattern.

Machine learning (ML) has become a pivotal tool in the fight against breast cancer, offering significant advancements in early detection, diagnosis, and treatment planning. Here are some key areas where machine learning is making an impact:

Machine learning algorithms, particularly deep learning models like convolutional neural networks (CNNs), are being used to analyze mammograms, ultrasounds, and MRI scans to detect breast cancer at an early stage. These models can identify patterns and anomalies that may be indicative of cancer, often with higher accuracy than traditional methods. For instance, models like Inception-V3 and ResNet-50 have shown superior performance in identifying various breast tissue types and detecting abnormalities[2].

ML techniques are also employed to classify breast lesions as benign or malignant. This involves analyzing features from medical images to determine the nature of the lesion. Advanced models, such as those incorporating transfer learning and ensemble methods, have demonstrated high accuracy in distinguishing between benign and malignant masses. For example, a study using a deep ensemble-based model achieved 99% accuracy on the MIAS datase[3]. Predictive models using machine learning can assess the risk of breast cancer recurrence. By analyzing patient data, including tumor characteristics and treatment history, these models can provide valuable insights into the likelihood of cancer returning. Algorithms like AdaBoost have been used to develop robust prediction models, helping clinicians make informed decisions about follow-up care and treatment[4].

Attention mechanisms have become a crucial component in enhancing the performance of deep learning models for breast cancer detection and classification. These mechanisms

help models focus on the most relevant parts of the input data, improving accuracy and interpretability. Here are some key applications of attention mechanisms in breast cancer research: Attention mechanisms can refine the feature extraction process by emphasizing important regions in medical images. For instance, the Efficient Channel Spatial Attention Network (ECSAnet) integrates a convolutional block attention module (CBAM) to improve the classification of histopathological images. This approach has shown superior performance compared to traditional models like AlexNet and ResNet50[5].Incorporating attention mechanisms into deep learning models can significantly boost classification accuracy. The Residual Attention Neural Network for Breast Cancer Classification (RANN-BCC) utilizes residual neural networks with attention layers to classify invasive ductal carcinoma (IDC) and non-IDC. This model achieved high accuracy, recall, precision, and F1 score, outperforming other models such as CNN and AlexNet[6].

II. PREPROCESSING

A. Data

The DDSM (Digital Database for Screening Mammography) and the CDD (Curated Breast Imaging Subset of DDSM) datasets are essential resources in breast cancer research and diagnosis.

The DDSM dataset comprises 2,620 scanned film mammography studies, including normal, benign, and malignant cases with verified pathology information. Its extensive scale and ground truth validation make it invaluable for developing and testing decision support systems. However, historical limitations such as non-standard compression files and imprecise lesion annotations have hindered direct comparisons and replication of results.

In contrast, the CDD dataset is an updated and standardized version of DDSM, curated by trained mammographers. It includes decompressed images converted to DICOM format, improved ROI segmentation, and precise pathologic diagnoses for training data. By addressing the limitations of DDSM, CDD provides a consistent evaluation platform for future CADx (computer-aided diagnosis) and CADe (computer-aided detection) research in mammography.

By integrating the diverse breast tissue variations from DDSM with the structured clinical information provided by CDD, researchers can develop sophisticated machine learning models. These models can accurately differentiate between benign and malignant lesions, leading to earlier detection and improved treatment strategies. Ultimately, this combination enhances breast cancer diagnosis and patient care.

B. Preprocessing

The original datasets contained small regions extracted from entire breast images, including some artifacts and irrelevant portions. To improve data quality, we meticulously removed these non-useful regions, ensuring that the dataset only contained relevant information for training and evaluation.

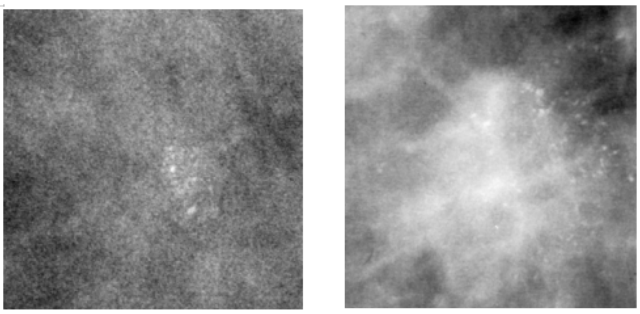


Fig. 1. Calcification images

We carefully divided the data into training, validation, and test subsets, selecting samples from both mass (benign/malignant) and calcification (benign/malignant) cases. This balanced approach allowed the model to learn from a diverse set of scenarios, ensuring robustness. Additionally, we performed augmentations such as flipping, rotation, and brightness adjustments to enhance the dataset. Images in Figure 1 depict calcification, while images in Figure 2 show masses. To enhance model generalization, various augmentations were applied to the training data. Images were horizontally or vertically flipped to create additional variations. Random rotations were introduced to simulate different orientations, and brightness adjustments were made to account for varying lighting conditions. These augmentations help the model learn from a more diverse set of scenarios, improving its robustness and performance.

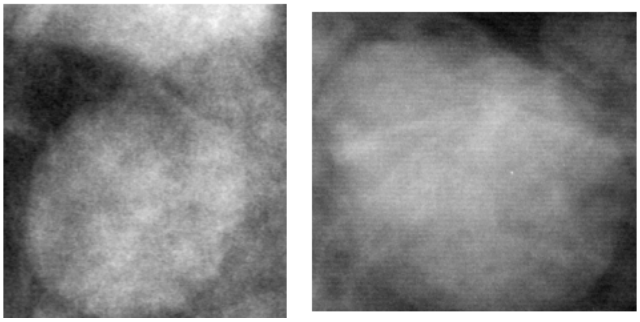


Fig. 2. Mass images

C. Augmentations

To enhance model generalization, various augmentations were applied to the training data. Images were horizontally or vertically flipped to create additional variations. Random rotations were introduced to simulate different orientations, and brightness adjustments were made to account for varying lighting conditions. These augmentations help the model learn from a more diverse set of scenarios, improving its robustness and performance.

III. METHODOLOGY

For our methodology, we began by utilizing ConvNeXt for feature extraction. ConvNeXt, known for its robust perfor-

TABLE I
DATA DISTRIBUTION

Partition	Calcification (Benign)	Mass (Benign)	Calcification (Malignant)	Mass (Malignant)	Pathology (Malignant/Benign)	Total
Train	1007	768	541	655	1775/1196	2971
Test	110	99	75	87	209/162	371
Validation	110	99	74	85	209/162	371

mance, allowed us to effectively capture and represent the essential features of our data. This initial step was crucial in ensuring that our subsequent processes were built on a strong foundation. Following the feature extraction, we employed

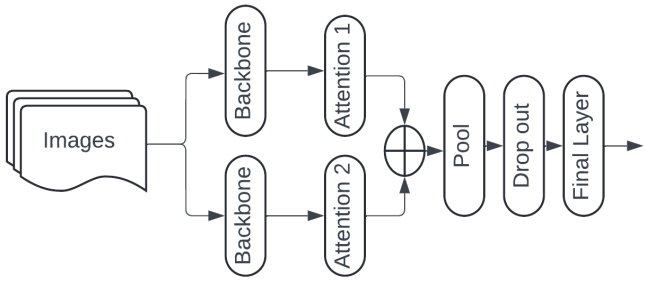


Fig. 3. Parrel Configaration

attention mechanisms to further refine and extract features in various forms. Attention mechanisms have the ability to focus on different parts of the input data, making them highly effective for our purposes. We explored different configurations by arranging two attention mechanisms both in parallel and in series. This approach enabled us to investigate how different arrangements could impact the feature extraction process.

In our experiments, we tested several types of attention mechanisms, each offering unique advantages. The first attention mechanism we tested was the Spatial Relation Module (SRM). SRM focuses on capturing spatial relationships within the data, which can be particularly useful for tasks that require an understanding of spatial context. Next, we experimented

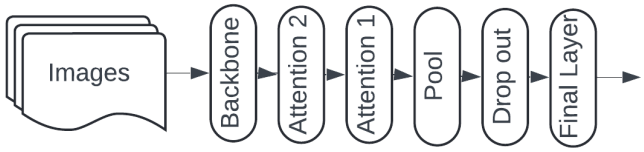


Fig. 4. Series Configaration

with the Global Contextual Transformer (GCT). GCT is designed to capture global contextual information, allowing the model to understand the broader context of the data. This can be beneficial for tasks that require a comprehensive understanding of the entire input.

We also tested the Local Contextual Transformer (LCT), which focuses on capturing local contextual information. LCT is particularly useful for tasks that require a detailed understanding of specific parts of the input data. By combining LCT with other attention mechanisms, we aimed to achieve a balance between local and global context.

Lastly, we experimented with Triplet Attention, which is designed to capture relationships between triplets of data points. This unique approach allowed us to explore how interactions between multiple data points could enhance the feature extraction process. By systematically testing these different attention mechanisms, we were able to identify the most effective configurations for our specific use case, ultimately enhancing the performance of our model.

IV. EXPERIMENT SETTING

Our current experiment was conducted on a Linux operating system, utilizing a GTX 2090 GPU to ensure efficient processing and model training. We employed Python and PyTorch as our primary programming and deep learning frameworks, respectively. These tools provided the flexibility and robustness needed to implement and test various attention mechanisms in our model.

For the optimization process, we used the Adam optimizer combined with a cosine annealing learning rate schedule. This approach helped in achieving a smooth convergence during training. The model was trained for 100 epochs with a batch size of 8, allowing us to effectively balance computational efficiency and model performance. This setup enabled us to thoroughly evaluate the impact of different attention mechanisms on our model's accuracy and F1 score.

V. METRIC

Accuracy is a metric that measures how often a machine learning model correctly predicts the outcome. It is calculated as the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances. The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

In binary classification, it can also be expressed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN : False Negatives

The **F1 score** is a metric that combines precision and recall into a single value, providing a balance between the two. It

Model	type	Accuracy	F1 Score	Loss	F1 Score (Test)
Base model	-	0.78	0.7655	0.557	0.74
Parrel	SRM_GCT	0.787	0.7736	0.5017	0.74
	tripple_gct	0.787	0.7736	**0.4812**	0.72
	tripple_lct	**0.7924**	**0.7818**	0.5583	0.75
Series	tripple_gct	0.787	0.7669	0.6313	0.75
	tripple_lct	0.7843	0.7575	0.6793	0.752
	SRM_GCT	0.7789	0.7696	0.4913	**0.758**

TABLE II
MODEL PERFORMANCE COMPARISON

is particularly useful when you need to account for both false positives and false negatives. The F1 score is the harmonic mean of precision and recall, calculated as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

- **Precision** is the ratio of true positive predictions to the total number of positive predictions (true positives + false positives).
- **Recall** (or sensitivity) is the ratio of true positive predictions to the total number of actual positive instances (true positives + false negatives).

VI. RESULTS

The Base model achieved an accuracy of 0.78, an F1 Score of 0.7655, a loss of 0.557, and an F1 Score (Test) of 0.74. This serves as a benchmark for evaluating the performance of the other models.

In the Parallel configuration, the Triplet_LCT model stood out with the highest accuracy of 0.7924 and an F1 Score of 0.7818. However, it also had a relatively high loss of 0.5583. The SRM_GCT model in this configuration had a slightly lower accuracy of 0.787 and an F1 Score of 0.7736, but it achieved the lowest loss of 0.5017 among the parallel models. The Triplet_ACT model also performed well with an accuracy of 0.787, an F1 Score of 0.7669, and a loss of 0.6313.

In the Series configuration, the SRM_GCT model achieved an accuracy of 0.7789, an F1 Score of 0.7696, and the lowest loss of 0.4913 among all models. The Triplet_LCT model in this configuration had a higher accuracy of 0.7843 and an F1 Score of 0.7575, but it also had the highest loss of 0.6793. Overall, the parallel configuration models generally performed better in terms of accuracy and F1 Score compared to the series configuration models. The choice of attention mechanism and its arrangement significantly impacted the performance metrics, highlighting the importance of selecting the right configuration for the specific use case.

VII. CONCLUSION

In our experiments, the addition of attention mechanisms to the models demonstrated noticeable improvements in performance metrics such as accuracy and F1 score. By incorporating attention mechanisms like SRM, GCT, and LCT, we were able to enhance the model's ability to focus on relevant features, leading to better overall performance. These attention mechanisms allowed the model to capture both local and global contextual information, which is crucial for accurate feature extraction in complex datasets.

Among the various configurations tested, the combinations involving Triplet Attention yielded the most significant performance gains. The Triplet Attention mechanism, when combined with other attention mechanisms like GCT and LCT, provided a more comprehensive understanding of the relationships between data points. This resulted in higher accuracy and F1 scores compared to other configurations. The parallel configuration of Triplet_LCT, in particular, achieved the highest accuracy and F1 score, highlighting the effectiveness of this combination in improving model performance.Overall, the integration of attention mechanisms, especially in combination with Triplet Attention, proved to be a valuable approach in enhancing the performance of our models.

REFERENCES

[1] J. Hao, R. Jin, Y. Yi, *et al.*, "Development of a humanized anti-fabp4 monoclonal antibody for potential treatment of breast cancer," *Breast Cancer Research*, vol. 26, no. 1, pp. 1–15, 2024.

[2] D. Zuo, L. Yang, Y. Jin, H. Qi, Y. Liu, and L. Ren, "Machine learning-based models for the prediction of breast cancer recurrence risk," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 276, 2023.

[3] A. Khalid, A. Mehmood, A. Alabrah, *et al.*, "Breast cancer detection and prevention using machine learning," *Diagnostics*, vol. 13, no. 19, p. 3113, 2023.

[4] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine learning with applications in breast cancer diagnosis and prognosis," *Designs*, vol. 2, no. 2, p. 13, 2018.

[5] L. A. Aldakhil, H. F. Alhasson, and S. S. Alharbi, "Attention-based deep learning approach for breast cancer histopathological image multi-classification," *Diagnostics*, vol. 14, no. 13, p. 1402, 2024.

[6] C. K. Toa, M. Elsayed, and K. S. Sim, "Deep residual learning with attention mechanism for breast cancer classification," *Soft Computing*, pp. 1–11, 2023.

Enhanced Detection of Illegally Parked Vehicles Using YOLO and Good Feature to Track Methods

Fauzan Maftuh Alwafi*, Boby Mugi Pratama*, Phuong Thi Le†, Bima Prihasto*, and Jia-Ching Wang‡
* Department of Informatics, Institut Teknologi Kalimantan, Indonesia
† Department of Computer Science and Information Engineering, Fu Jen Catholic University, Taiwan
‡ Department of Computer Science and Information Engineering, National Central University, Taiwan

***Abstract*—The study addresses illegal parking, a problem that disrupts traffic and poses safety risks under Indonesia's legal system. It proposes using the YOLO technique for detection and Optical Flow for vehicle speed assessment to identify illegally parked vehicles. The objective is to improve detection precision and efficiency in areas prone to illegal parking by leveraging deep learning, particularly Convolutional Neural Networks (CNN). The study highlights the potential of noninvasive computer vision technology to enhance public awareness and law adherence. The findings aim to support the development of a monitoring system to help law enforcement in managing illegal parking in public spaces.**

Keywords: Illegal Parking, Object Detection, YOLO, Optical Flow, Computer Vision

I. Introduction

Illegal parking occurs in areas not designated for parking and disrupts traffic and safety. Traffic regulations prohibit such activities to maintain order and safety. Illegal parking not only causes inconvenience, but also poses security risks by obstructing emergency routes and facilitating criminal activities [1]. In 2022, 30,801 vehicles were penalized for illegal parking in DKI Jakarta [2], and 350 illegal parking spots were identified in Balikpapan [3]. Factors contributing to illegal parking include limited parking spaces, high costs of official facilities, and inadequate public transportation [4]. Weak law enforcement in Indonesia exacerbates the problem, with patrols issuing only warnings instead of sanctions [5]. Raising public awareness and adherence to parking regulations is essential.

In recent years, significant advancements in artificial intelligence (AI) have been made, particularly in deep learning methods such as convolutional neural networks (CNNs) and recurrent models, which are utilized for solving various computer vision and digital image processing problems [6]. Deep learning has become prominent in machine learning due to its effectiveness in modeling complex data like images and sounds. A notable advancement in image recognition has come from CNNs, which mimic the human visual cortex's image recognition system, thus excelling in image processing. This is exemplified by the You Only Look Once (YOLO) algorithm, which specializes in real-time object detection. Research has indicated that the YOLO algorithm can achieve approximately 90% accuracy in vehicle type detection, influenced by factors such as video quality, dataset quality, and camera angle variations [7]. Higher accuracy is linked to a richer and more diverse dataset. An application of YOLO in tracking parking time violations demonstrated the potential for reducing labor and costs in various parking settings [8]. YOLO's capability extends across several versions, employing localization and optical flow techniques to enable real-time detection and motion tracking [9].

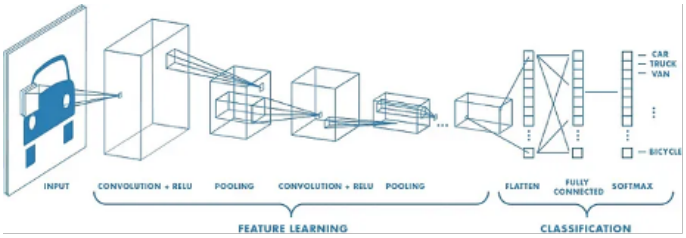


Fig. 1: Architecture of YOLO[11]

The study tackles illegal parking issues in Indonesia, emphasizing the disruption of traffic and safety hazards. It proposes using YOLO for vehicle detection and Optical Flow for speed analysis to identify illegally parked vehicles. By leveraging deep learning through Convolutional Neural Networks (CNN), the study aims to enhance detection precision and efficiency in high-risk areas. It underscores the potential of non-invasive computer vision technology to improve public awareness and legal compliance. The findings are intended to assist in developing a monitoring system to aid law enforcement in managing illegal parking.

II. Method

A. Convolutional Neural Network.

Convolutional Neural Network (CNN) is a type of artificial neural network designed for image processing tasks. Inspired by human visual perception, it has been proven effective in identifying and classifying images [10]. A CNN consists of several layers: convolution, activation, pooling, and fully connected layers.

B. Optical Flow

Optical flow is a technique in image processing and computer vision used to capture the motion of objects based on a sequence of images. In processing a sequence of images, the optical flow reflects the movement of pixels in the image or the actual visualization seen[12]. Optical flow provides a depiction of how each pixel in an image experiences movement or

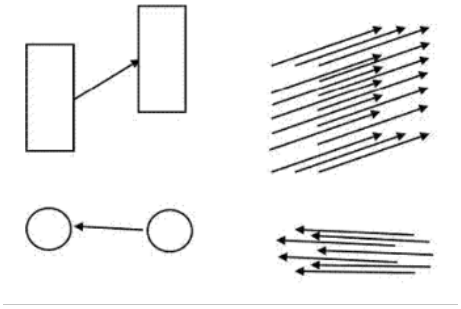


Fig. 2: The movement of objects in two frames and the optical flow vectors for each object[12].

changes over time. If an object in the image moves, variations in the brightness of certain pixels will create an optical flow pattern that reflects the direction and speed of the object's movement. A pixel in the first frame of an image sequence has coordinates (x, y) and light intensity $I(x, y, t)$. The assumption is that the pixel being tracked will maintain the same light intensity even after it has moved and time has passed, and neighboring pixels have similar movements.

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{1}$$

$$I(x, y, t) = I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt \tag{2}$$

$$\frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt = 0 \tag{3}$$

$$I_x dx + I_y dy + I_t dt = 0 \tag{4}$$

$$f_x u + f_y v + f_t = 0 \tag{5}$$

Equation [1-5] describes the basic principle of optical flow, where the brightness of a feature over time is balanced by movement in space or optical flow, so the total change in the brightness of a feature along the object's path is zero. The variables f_x and f_y represent the image intensity gradients or changes in pixel intensity from one location to another, while f_t is the gradient over time. The solution to this equation cannot be determined without finding the variable values of (u, v). Therefore, methods such as the Lucas-Kanade method can be used to solve it, which is effective in matching points by focusing on specific points or points with special characteristics.

C. Region of Interest

Region Of Interest (ROI) is an area specified within an image or video that serves as a target area for further analysis. The main goal of ROI is to focus computational resources and analysis techniques on the most relevant and significant parts. The ROI defined by the user is used as a specific feature for further processing. This approach can improve efficiency and reduce the amount of processing required for unnecessary parts [13].

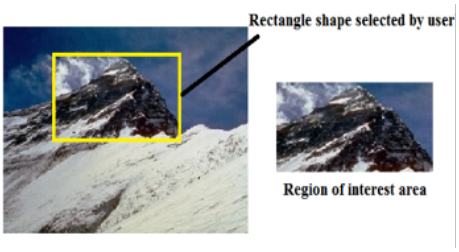


Fig. 3: Visualization of ROI Implementation[13]

III. Object Detection using YoloV8 and Optical Flow

This research employs the YOLO (You Only Look Once) algorithm and Optical Flow for object detection and motion tracking. YOLO, introduced by [11], uses a single neural network to process the entire image, dividing it into a grid for object detection. Optical Flow is utilized to detect object movement across consecutive frames in images or videos, aiding in understanding dynamic motion sequences. The research involved multiple stages to achieve its objectives. Initially, data was collected from various sources and meticulously labeled to ensure accuracy. YOLOv8[14] was integrated to enhance object detection capabilities. OpenCV was used to create bounding boxes around illegal parking areas to focus on vehicles in these zones. An indication system was implemented to track if a vehicle remained in the designated area for a specific duration using optical flow. Vehicles staying stationary in the illegal parking zones for the designated time were classified as illegally parked, while those that did not were not classified as such.

A. YOLOv8

The research utilizes YOLO version 8, which comprises three architectural parts: the backbone, the neck, and the head. The backbone extracts features from the input image, as illustrated in Figure 5.

The neck in YOLOv8 combines outputs from various layers of the backbone to produce features with different resolutions. These features are then used by the head to detect objects in the image. The architecture of the neck section is illustrated in Figure 6.

The head section of the model is crucial for object detection in images. It serves two primary functions: predicting bounding boxes, object classes, and probabilities for each object, and managing the removal of overlapping or colliding bounding boxes. The architecture for the YOLOv8's head section is depicted in Figure 7.

B. Lucas-Kanade

The Lucas-Kanade method is a popular technique in optical flow calculation. It determines the optical flow vector by assuming that the vector will be similar within a small

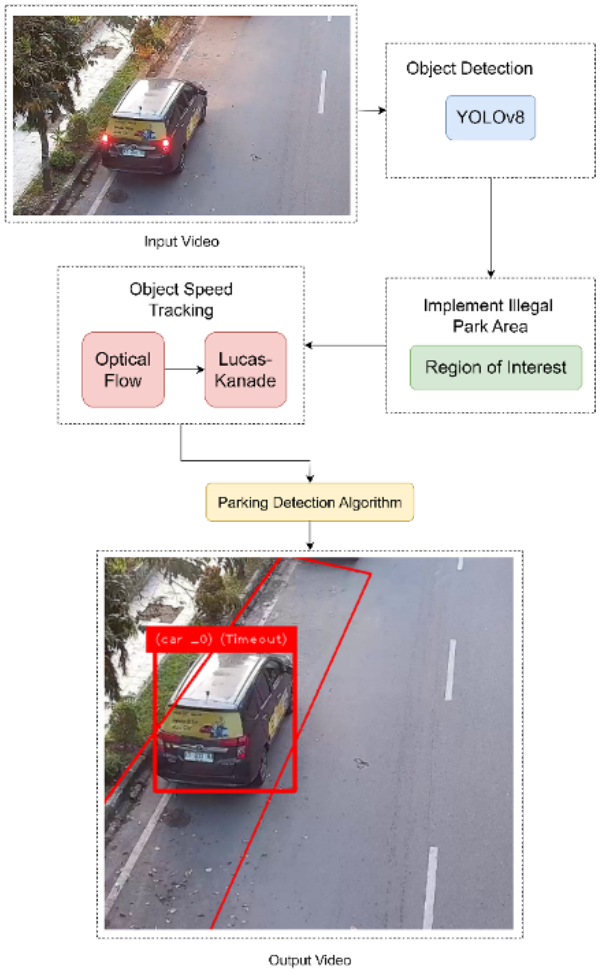


Fig. 4: Research Flow

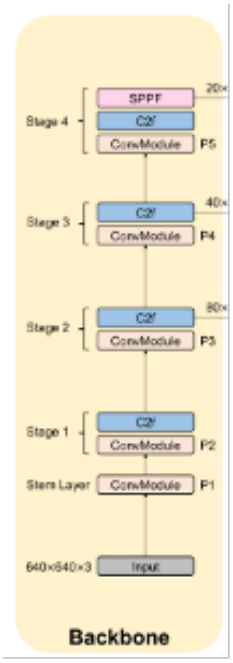


Fig. 5: Architecture of Neck[14]

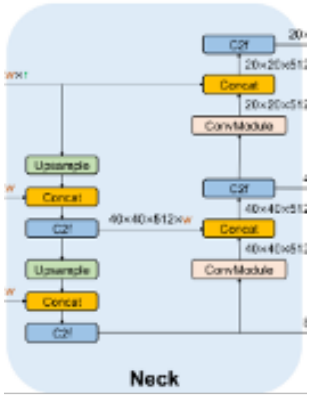


Fig. 6: Architecture of Backbone[14]

neighborhood around the pixel. Specifically, it uses a 3x3 patch around a point, where all nine points are assumed to have similar motion. The method searches for (f_x, f_y, f_t) at these nine points.

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i f_x^2 & \sum_i f_x f_y \\ \sum_i f_x f_y & \sum_i f_y^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i f_x f_t \\ -\sum_i f_y f_t \end{bmatrix} \quad (6)$$

The Lucas-Kanade method tracks small movements by obtaining the optical flow vector of determined points. However, it loses tracked features with large movements. To address this, the pyramid concept is used, allowing large movements to be transformed into small movements as you go up in the pyramid.

IV. Experiment

A. Data Preparation

During the data collection phase, images of various types of vehicles, such as cars, were collected from the Roboflow website platform [15]. The vehicles represented are found throughout the country. The data is split into 80% for training, 10% for testing, and 10% for validation before the model can make predictions on image or video input.

B. Model Performance Evaluation

The evaluation is performed using evaluation metrics, namely precision, recall, and mAP50. Table 1 shows the evaluation results of the YOLOv8 model in detecting cars on the road.

Class	Precision	Recall	mAP50
Car	91,6%	82,7%	87,5%

TABLE I: Evaluation YOLO model result

Regarding the model performance, the visualization based on evaluation metrics for the validation data is as follows.

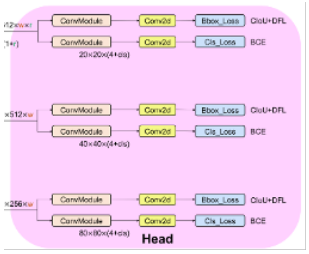
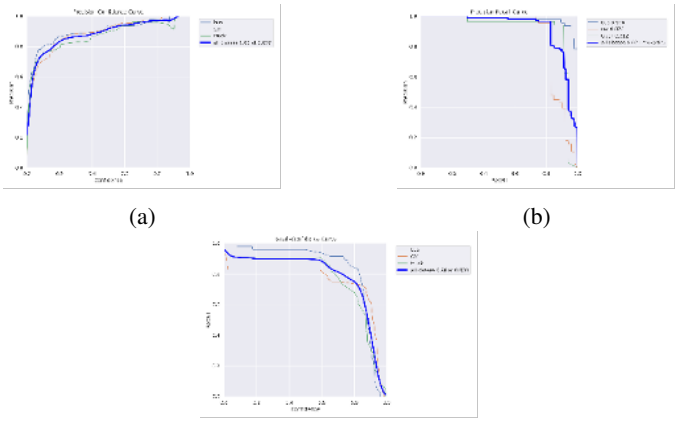


Fig. 7: Architecture of Head[14]



(a) (b)

(c)

Fig. 8: (a) Precision Curve, (b) Precision-Recall Curve, and (c) Recall Curve

C. Implementation of Illegal Parking Area Detection

The implementation for detecting illegal parking areas utilizes the OpenCV library to create focused regions designated as illegal parking zones. The process involves applying the Region of Interest (ROI) concept, which concentrates on specific illegal parking areas defined by four coordinate points. The Yolo bounding boxes used for detection have a format consisting of the coordinates x_1, y_1 for the top-left point, x_2, y_2 for the bottom-right point, a confidence score (conf), and a class label (cls).



Fig. 9: Illustration of applying the indicated illegal parking area

To determine the region, four coordinate points are required: $(x1, y1)$, $(x2, y2)$, $(x3, y3)$, and $(x4, y4)$, forming a quadrilateral. OpenCV's polylines function is used to draw the sides connecting these points. Applying the region of interest allows for focused detection of illegal parking. Test samples from CCTV footage on public roads are used for this detection.

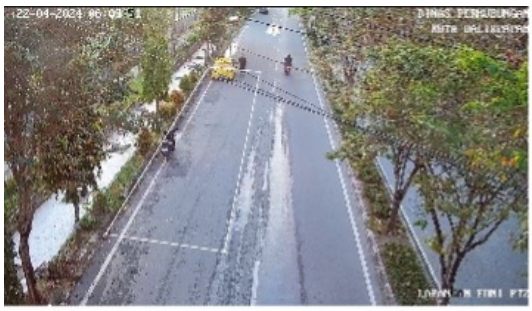


Fig. 10: Sample Video Test Illegal Park

D. Implementation of Parking Detection Algorithm Using Optical Flow

The implementation of the illegal parking detection algorithm involves several steps using YOLO detection and the optical flow method with the OpenCV library. After defining the area suspected of illegal parking and selecting the points, the algorithm enters a loop, processing incrementally every 3 frames to enhance efficiency. The video is resized to 1200 x 720 pixels for consistency. YOLO is used to detect vehicles entering the illegal parking zone, and the Lucas-Kanade method then checks for vehicle movement by calculating the displacement in pixels per frame from the midpoint reference detected by YOLO. This process helps determine if a vehicle is stationary, indicating illegal parking.



Fig. 11: Implementation of Lucas-Canade Optical Flow Method

The implementation of optical flow uses a black dot, representing the midpoint of an object detected by YOLO, to calculate displacement per pixel. This enables the determination of vehicle movement rates based on pixel displacement speed per frame. If a vehicle is detected in a predefined illegal parking

area with zero displacement per frame, a countdown from a predetermined time limit begins. If this limit is exceeded, the vehicle is flagged as illegally parked using a red bounding box; otherwise, a blue bounding box indicates the vehicle is either moving or has not yet exceeded the time limit. Figure 12 illustrates a vehicle moving in the illegal parking area.

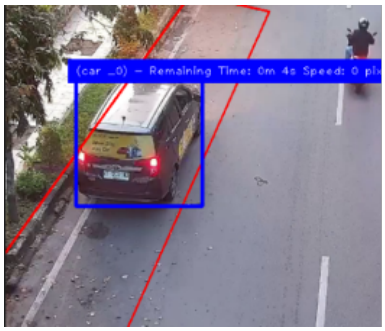


Fig. 12: The vehicle is still moving in the area.

Then, when the vehicle is in the area and is no longer moving, or has a speed of 0 pixels per frame, the time for how long it has been stationary will be counted down.



Fig. 13: The vehicle in the area is not moving but has not yet exceeded the time limit and the timer is counting

Next, when the vehicle is in the illegal parking area, not moving or with a speed of 0 pixels per frame, and has also exceeded the specified time limit.

E. Testing Of Illegal Parking Detection

In testing the illegal parking detection algorithm, we will test 1 location at 3 different times using a scenario parking duration of 1 min. The footage is recorded for a duration of 5 minutes. The comparison will be made by displaying the original data alongside the data detected using the illegal parking algorithm using video footage from CCTV provided by the Department of Transportation of Balikpapan City.

V. Conclusions

The study shows that combining the YOLO and optical flow methods effectively detects illegally parked vehicles. YOLO is used for object detection, and optical flow tracks vehicle movement, accurately identifying vehicles that remain stationary in



Fig. 14: The vehicle inside the area is not moving and has exceeded the specified time limit.

TABLE II: Testing CCTV Video 06.05 – 06.10

Data	Detection	Ilegall Park
Prediction		2
Actual		2

TABLE III: Testing CCTV Video 12.55 – 13.00

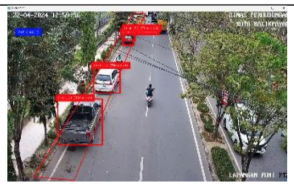
Data	Detection	Ilegall Park
Prediction		5
Actual		4

TABLE IV: Testing CCTV Video 17.27 – 17.32

Data	Detection	Ilegall Park
Prediction		5
Actual		3

restricted areas beyond the allowed time. Challenges include difficulty in vehicle ID recognition since YOLO processes each frame independently, potentially restarting the nonmovement time calculation upon timeout. Despite this, YOLO is effective in object identification with trained data, and the Lucas-Kanade optical flow method accurately computes object movement by measuring pixel displacement speed.

REFERENCES

[1] F. Fatmasari and N. Wahyuning Retno Mumpuni, "Effectiveness of the application of administrative criminal sanctions against illegal parking in the malioboro area," *Legal Literacy*, Oct. 2023. DOI: 10.31002/lh.v7i2.7836.

[2] D. Alfa, A. Bachtiar, and C. Candi, "Analysis of health-based transportation system for health transformation in dki jakarta," *Journal of Social Research*, Oct. 2023. DOI: 10.55324/josr.v2i11.1568.

[3] I. Setiawan, M. Ilham, and M. Nawawi, "Smart governance implementation in balikpapan city, east kalimantan," *Journal of Borneo-Kalimantan*, vol. 6, no. 1, Jul. 2020. DOI: 10.33736/JBK.2412.2020.

[4] R. W. Triputro, A. Giawa, S. Suharyanto, and J. H. Wijaya, "Government policy in illegal parking charges at public spaces," *Journal of Governance and Public Policy*, vol. 10, no. 2, Jun. 2023. DOI: 10.18196/jgpp.v10i2.17880.

[5] T. R. Isnawan, "Legal basis of imposing sanctions on illegal parking retribution collectors from the maqāṣid perspective," *Indonesian Journal of Maqasid and Fiqh Muqaran*, vol. 2, no. 1, Jun. 2023. DOI: 10.22373/ijomafim.v2i1.3056.

[6] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, Jul. 2019. DOI: 10.1186/s40537-019-0197-0. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0.

[7] D. I. Mulyana and M. A. Rofik, "Implementation of real time detection of vehicle type classification in indonesia using the yolov5 method," *Jurnal Pendidikan Tambusai*, vol. 6, no. 3, pp. 13 971–13 982, Jul. 2022. DOI: 10.31004/jptam.v6i3.4825. [Online]. Available: https://doi.org/10.31004/jptam.v6i3.4825.

[8] N. Sharma, S. Baral, M. P. Paing, and R. Chawuthai, "Parking time violation tracking using yolov8 and tracking algorithms," *Sensors*, vol. 23, no. 13, p. 5843, Jun. 2023. DOI: 10.3390/s23135843. [Online]. Available: https://doi.org/10.3390/s23135843.

[9] L. Jiao, F. Zhang, F. Liu, *et al.*, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128 868, Jan. 2019. DOI: 10.1109/access.2019.2939201. [Online]. Available: https://doi.org/10.1109/access.2019.2939201.

[10] J. Gu, Z. Wang, J. Kuen, *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, May 2018. DOI: 10.1016/j.patcog.2017.10.013. [Online]. Available: https://doi.org/10.1016/j.patcog.2017.10.013.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Jun. 2016. DOI: 10.1109/cvpr.2016.91. [Online]. Available: https://doi.org/10.1109/cvpr.2016.91.

[12] P. J. O'donovan, "Optical flow : Techniques and applications," 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:16986085.

[13] S. S. Elkapelli and L. B. Damahe, "A review: Region of interest based image retrieval," Nov. 2016. DOI: 10.1109/get.2016.7916731. [Online]. Available: https://doi.org/10.1109/get.2016.7916731.

[14] R.-Y. Ju and W. Cai, "Fracture detection in pediatric wrist trauma x-ray images using yolov8 algorithm," *Scientific Reports*, vol. 13, no. 1, Nov. 2023. DOI: 10.1038/s41598-023-47460-7. [Online]. Available: https://doi.org/10.1038/s41598-023-47460-7.

[15] H. T. Dwyer B Nelson J, *Roboflow: Computer vision tools for developers and enterprises — roboflow.com*, https://roboflow.com/, [Accessed 05-08-2024].

Integrating VGGSK and BEATs for Enhanced Sound Event Detection: A Semi-Supervised GRU-Based System with Weak Labels and Synthetic Soundscapes

Po Cheng Chan[1,2], Wei-Yu Chen[1], Chung Li Lu[1], Hsiang-Feng Chuang[1], Yu-Han Cheng[1], and Jia Ching Wang[2]

[1]Advanced Technology Laboratory, Chunghwa Telecom Laboratories, Taoyuan, Taiwan
[2]Department of Computer Science and Information Engineering, National Central University
[1]Email: *cbc, weiweichen, chungli, gotop, henacheng@cht.com.tw*
[2]Email: *jcw@csie.ncu.edu.tw*

***Abstract*—This paper presents the architecture we developed for the Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge, specifically Task 4 on Sound Event Detection using Weak Labels and Synthetic Soundscapes. We integrated embeddings from VGGSK and BEATs, and employed a GRU-based model to classify sound events in each time frame. The system employs thresholding and smoothing techniques in its post-processing phase. For semi-supervised learning, we used the mean teacher approach with an Exponential Moving Average (EMA) strategy to update the teacher model's parameters. Pseudo-labels, generated by the student model, help leverage unlabeled data. Additionally, we applied data augmentation methods including mix-up, Gaussian noise, and embedding masking. With additional training data, our system achieved a Polyphonic Sound Detection Score (PSDS) of 0.529 for PSDS1 and 0.78 for PSDS2 on the validation dataset.**

I. Introduction

The objective of sound event detection is to identify the occurrence of specific sound events within an audio clip, including their onset and termination times. A significant challenge in employing supervised learning for this task lies in the high cost and potential bias associated with manually annotating sound data after its collection, as annotations can vary greatly among different reviewers. To address this, the DESED dataset [1] comprises audio files collected via two approaches: authentic ambient recordings and synthetic sounds. The dataset categorizes labels into three types: strong, weak, and unlabeled. Participants in the challenge are prompted to utilize the DESED dataset and are permitted to incorporate external datasets or pre-trained embeddings, although one submitted system must rely solely on DESED without using these additional resources, and at least one system should avoid using ensemble methods.

For enhancing the system that can utilize external datasets or pre-trained embeddings, the Bidirectional Encoder Representation from Audio Transformers (BEATs) [2] based on the Transformer architecture is employed to derive embedding features. Conversely, the system exclusively trains on the VGGSK model [3][4], a CNN-based architecture. The Exponential Moving Average (EMA) strategy is employed to update the VGGSK and Gated Recurrent Unit (GRU) [5] components during the teacher-student model update process. Data augmentation techniques such as mix-up [6], Gaussian noise, and ICT [7] are implemented on VGGSK inputs, while masking is applied to BEATs embeddings to enhance feature extraction. The primary aim is to boost predictive accuracy without extending inference time.

II. Methodology

This section will discuss the methods we used. Before employing the mean-teacher approach, we will first conduct supervised training on the CRNN and BEATs-VGGSK models. Subsequently, we will train the teacher model using a semi-supervised method to enhance the model's performance.

A. Feature Extraction

Due to variations in sampling rates, channel numbers, and audio file lengths in the DESED dataset, we employ the librosa library to normalize all audio files to a consistent 16000 Hz sampling rate and convert them to a mono channel format. Each audio file is padded to reach a standard length of 10 seconds through zero-padding. Subsequently, these audio files are converted into Mel-spectrograms by transforming the waveform signals and applying logarithmic scaling. We conduct a Short-Time Fourier Transform (STFT) on these signals using a window size of 2048 and a hop length of 256. The Mel-spectrograms produced are of dimensions 768 (time) and 128 (frequency), created using a bank of 128 Mel filters.

B. Baseline

The baseline model, a convolutional recurrent neural network (CRNN) [8], integrates elements of both convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This CRNN structure features a CNN component consisting of seven blocks, each equipped with filters in a

sequence of 16, 32, 64, 128, 128, 128, and 128. Every block uses a 3 × 3 kernel size and employs average-pooling operations with configurations of [2, 2], [2, 2], [2, 1], [2, 1], [2, 1], [2, 1], [2, 1] across the layers. The RNN segment of the architecture includes two layers of 128 bidirectional gated recurrent units (Bi-GRUs) [5]. An attention pooling layer follows the RNN portion, featuring a linear layer with softmax functions followed by multiplication with another linear layer that utilizes sigmoid activations.

C. DataAugmentation

The mixed output is combined by audio and Gaussian noise that RMS of the noise generated:

$$y[n] = x[n] + \text{RMS}_{\text{noise}} \tag{1}$$

$$\text{RMS}_{\text{noise}} = \sqrt{\frac{P_{\text{signal}}}{10^{\text{SNR}/10}}} \tag{2}$$

Mixup [6] is a technique that performs linear combinations of pairs of data points and their labels. The proportion of each data point in the combination is sampled from a Beta distribution, using a parameter alpha to control the interpolation strength between the two samples:

$$y[n] = \alpha \cdot x_1[n] + (1 - \alpha) \cdot x_2[n] \tag{3}$$

where y is a mixture features of x_1 and x_2 are log-mel-filterbank outputs of class ID. We also multiply the one-hot encoded label with α.

Embedding masking in the context of neural networks is a technique used to prevent over-fitting and improve generalization by randomly setting elements of the embedding vector to zero during training. This can be represented mathematically as follows:

$$\mathbf{E}' = \mathbf{E} \odot \mathbf{M} \tag{4}$$

where $\odot$ denotes the element-wise product. Each element of $\mathbf{M}$ is generated independently based on a certain probability p, which is a hyperparameter determining the likelihood that an element of the embedding vector is masked. Typically, each m_{ij} in $\mathbf{M}$ is sampled from a β distribution:

$$m_{ij} \sim \beta(1 - p) \tag{5}$$

where $\mathbf{E}$ is the original embedding matrix. $\mathbf{M}$ is the mask matrix with the same shape as $\mathbf{E}$. $\mathbf{E}'$ is the masked embedding matrix, used during the forward pass in training. p is the probability of an element being set to zero (masked).

D. BEATS

The BEATs model [2], trained on extensive datasets such as Audioset [9], represents a significant advancement in the field of sound event classification. Central to the BEATs architecture is the innovative use of acoustic tokenizers, a component designed to transform raw audio signals into a series of discrete tokens. This transformation is achieved by extracting meaningful audio features from the raw data, which are then quantized into tokens that capture the fundamental acoustic properties of the input signals. This tokenization process simplifies the audio data, converting it into manageable and semantically rich units.

E. VGGSK

The network architecture primarily utilizes VGGSK [3] [4]and BEATs. In this configuration, data augmentation is conducted during the preprocessing phase of VGGSK. Embeddings are derived from the BEATs model, which masking is applied to enhance the embedding representation. The VGGSK component of the architecture includes a VGG block and four residual blocks that employ selective kernels (SK). For the supervised learning, Binary Cross-Entropy (BCE) serves as the loss function to assess model performance.

F. Interpolation Consistency Training

The interpolation consistency training (ICT) [7] is to perform interpolation calculation on the prediction results of the model. ICT involves mixing any two data samples from the dataset to create a new input for the student model, which then makes predictions based on this mixed data. Simultaneously, the original two data samples are used as inputs for the teacher model, which also makes predictions. These predictions are then blended together. Finally, the predictions made by the student model on the mixed data are regularized for consistency with the blended predictions of the teacher model, and an additional ICT loss value is calculated to measure this consistency.

G. Semi-supervised learning

The Mean Teacher model, employed within a semi-supervised learning framework, utilizes a sophisticated dual-model architecture to enhance model performance and stability using both labeled and unlabeled data. This framework consists of two main components: a student model and a teacher model, which share identical network structures but differ significantly in their parameter updating methods. The student model is conventionally trained through gradient descent, processing both labeled and unlabeled data. For unlabeled data, it generates predictions that are used as pseudo labels for training the teacher model. This phase not only optimizes the student model's parameters but also prepares it to update the teacher model using the Exponential Moving Average (EMA) strategy [10]. This EMA strategy smooths the parameters of the teacher model, reducing the impact of volatile updates from the student model and ensuring a more stable parameter evolution. Moreover, the teacher model is not trained directly. Instead, its parameters are continuously refined based on the EMA of the student model's parameters, enhancing the stability and reducing discrepancies between the models, which are quantified using Mean Square Error (MSE). To further improve recognition accuracy, Inter-Class Training (ICT) is integrated into the loss function, leveraging the structured yet flexible nature of this framework to effectively harness the potential of both labeled and unlabeled datasets in semi-supervised learning scenarios.

The Exponential Moving Average (EMA) is used to stabilize and update model parameters during the training process as follow:

$$\theta_{\text{ema}} = \beta \cdot \theta_{\text{ema}} + (1 - \beta) \cdot \theta \tag{6}$$

where $\theta_e ma$ is the exponential moving average of the model parameters at the current step. β is the decay factor, controlling the extent to which the previous values of ema_θ influence the new value. θ is the current parameter value.

H. Loss Function

The supervised loss $L_{\text{supervised}}$ is formulated by summing the binary cross-entropy losses from two distinct segments of the dataset, each processed by separate parameterized models. Specifically, we define the loss as:

$$L_{\text{supervised}} = \text{BCE}(\theta_s(X_s), Y_s) + \text{BCE}(\theta_w(X_w), Y_w) \tag{7}$$

where θ_s and θ_w denote the model parameters tailored to specific subsets of data, denoted as X_s and X_w respectively. The Y_s and Y_w represent the true binary labels corresponding to these subsets. The Binary Cross-Entropy (BCE) loss is utilized to measure the discrepancy between the predicted probabilities and the actual labels across both model configurations

$$L_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8}$$

The consistency loss $L_{\text{consistency}}$ in our framework is defined using a combination of weighted mean squared errors (MSE) that compare predictions made by the baseline model parameters θ_s and θ_w, with those made by their perturbed counterparts θ_s and θ_w. This formulation is expressed as:

$$L_{\text{consistency}} = W \cdot \{\text{MSE}(\theta_s(X), \theta'_s(X')) + \text{MSE}(\theta_w(X), \theta'_w(X'))\} \tag{9}$$

where X and X denote the original and perturbed versions of the input data. θ_s and θ_w represent the original model parameters, whereas θ_s and θ_w are their respective perturbed versions. W serves as a scalar that adjusts the impact of the MSE terms on the total loss value.

III. EXPERIMENTS

A. DataSet

The dataset consists of a training set along with validation and evaluation components, as listed in Table I. The occurrences of events within these components are listed in Table II. The training set includes: Weakly Labeled Training Set: This subset comprises 1578 clips with 2244 class occurrences, where each audio clip is labeled with the class of audio events but without specific timestamps. Unlabeled In-Domain Training Set: This subset includes 14412 clips, significantly larger in size compared to the weakly labeled data, and it does not contain labels. Synthetic Strongly Labeled Set: Composed of 10000 clips created using the Scaper soundscape synthesis and augmentation library, this subset is strongly labeled with timestamps indicating the sound events. Real Strongly Labeled Training Set: This subset contains 3470 audio clips sourced from Audioset, each strongly labeled with timestamps of sound events. This set is regarded as an external dataset. Subset of Audioset: Selected from Audioset, this segment includes 32975 clips. Labels were removed, and the clips are used as an external dataset. The validation dataset contains 1168 clips, each strongly labeled with timestamps. The evaluation dataset comprises 699 clips. As shown in Figure 1, it illustrates the key distinction between strong and weak labels in audio event annotation. Strong labels provide precise timestamps marking the start and end points of specific sound events within an audio track, as shown in the waveform and spectrogram. This labeling is detailed, indicating exactly when events like speech or the sound of a blender occur. On the other hand, weak labels only confirm the presence of these sound events within the entire audio track without specifying their temporal boundaries. The visual representation highlights that strong labels are time-bound and specific, whereas weak labels are more general and only indicate occurrence.

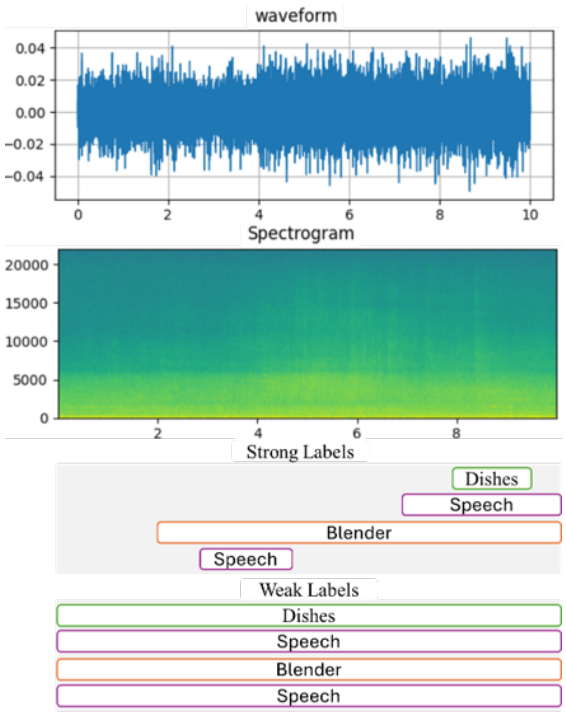


Fig. 1. The differences between strong and weak labels in audio event annotation.

B. Evaluation Metric

The evaluation of our systems was based on the recently introducedthreshold-independent [11] implementation of the polyphonic sound event detection scores (PSDS).

C. Single System

Table 1 shows the performance of each stage of our single model in the submission system. The baseline model initially utilized both the CRNN [8] and BEATs models [2]. However, it

TABLE I
DESED DATASET

Dataset	Label Type	Audio Clips	Sampling Rate(kHz)	Type
Training Set	**Strong Label**	13,470	44.1/16	**Record / Synthetic**
	Weak Label	1,578	44.1	Record
	Unlabeled	10,000	44.1	Record
Validation	**Strong Label**	1,168	44.1	Record
Evaluation		699	44.1	Record

TABLE II
OCCURRENCES OF DESED EVENT CLASSES

Class	Occurrences
Alarm bell ringing	2143
Blender	313
Cat	781
Dishes	2576
Dog	1949
Electric shaver toothbrush	279
Frying	620
Running water	833
Speech	9998
Vacuum Cleaner	178

was modified by replacing the CRNN model with the VGGSK model. This change resulted in an improvement in PSDS1, increasing it from 0.500 to 0.517. After incorporating the strong real dataset into the training process, an improvement was observed in PSDS2, with the value increasing from 0.764 to 0.775. By incorporating the ICT method, significant improvements were achieved in the results. PSDS1 improved to 0.529, indicating a substantial enhancement, while PSDS2 saw a remarkable improvement to 0.780.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT MODEL CONFIGURATIONS ON PSDS1 AND PSDS2 METRICS.

Model Configuration	PSDS1	PSDS2
CRNN+BEATs (Baseline)	0.500	0.762
VGGSK+BEATs	0.517	0.764
+Strong Real Dataset	0.516	0.775
+ICT	0.529	0.780

D. Ensemble System

As listed in Table ,the performance of System 2, 3 and 4 based on BEATs-VGGSK model in validation set with external data. Systems 3 and 4 were developed using the same model structure but were differentiated by employing distinct data augmentation techniques. System 3 achieved the highest PSDS1 score of 0.552. Similarly, System 4 recorded the highest PSDS2 score of 0.799.

IV. CONCLUSION

Our evaluation of the single and ensemble systems has significant improvement with down stream task of pre-trained model and data augmentation techniques. Initially, our single model system, the BEATs-VGGSK model leads an increase in PSDS1 from 0.500 to 0.517. Further incorporation of the strong real dataset significantly enhanced PSDS2, elevating it from 0.764 to 0.775. The application of the ICT method subsequently marked substantial improvements, with PSDS1 reaching 0.529 and PSDS2 peaking at 0.780. The ensemble systems, specifically Systems 3 and 4, built on the BEATs-VGGSK model and differentiated by data augmentation strategies, achieved robust performance. In conclusion, the updates to model architecture and the integration of advanced data augmentation and training techniques like ICT significantly contributed to the improvements in model performance.

TABLE IV
PERFORMANCE OF VARIOUS SYSTEMS ON PSDS1 AND PSDS2 METRICS WITH OPTIONAL EXTRA DATA

System	Extra data	PSDS1	PSDS2
CRNN (Baseline)		0.359	0.562
CRNN+BEATs (Baseline)	✓	0.500	0.762
System 1		0.424	0.633
System 2	✓	0.529	0.780
System 3	✓	0.552	0.794
System 4	✓	0.542	0.799

REFERENCES

[1] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *In Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[2] S. Chen, Y. Wu, C. Wang, S. Liu, Z. C. D. ompkins, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *arXiv:2212.09058*, 2018.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *arXiv:1409.1556.*, 2014.

[4] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[5] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, and F. B. et.al, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *In Proceeding of Empirical Methods in Natural Language Processing*, 2019, pp. 1724–1734.

[6] H. Zhang, M. Cisse, Y. N. Dauphin, and D. L. Paz, "Mixup: Beyond empirical risk minimization," in *arXiv:1710.09412*, 2017.

[7] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, and Y. B. et.al, "Interpolation consistency training for semi-supervised learning," in *arXiv:1903.03825*, 2019, pp. 271–350.

[8] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for dcase 2019 task 4 technical report," in *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2019.

[9] A. Tarvainen and H. Valpola, "Audio set: An ontology and human-labeled dataset for audio events," in *In Proceeding of IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 1195–1204.

[10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *In Proceeding of Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 1195–1204.

[11] J. Ebbers, R. H. Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores," in *In Proceeding of IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 1021–1025.

Seismic-ionospheric Precursor Prediction Using Deep Learning

Bach-Tung Pham*, Pao-Chi Chang†, and Jia-Ching Wang*
*Department of Computer Science and Information Engineering, National Central University, Taiwan
†Department of Communication Engineering, National Central University, Taiwan
E-mail: jcw@csie.ncu.edu.tw

***Abstract*- Numerous studies have indicated the potential involvement of seismic-ionospheric precursors (SIPs) in catastrophic earthquakes. These SIPs have been extensively analyzed in terms of characteristics such as the degree of decrease or increase, timing of occurrence, and duration, with a specific focus on electron density or total electron content (TEC). While deep neural networks have demonstrated remarkable accuracy in various applications, their capability to estimate SIPs has been severely limited. In this study, we developed a novel AI prediction model for forecasting SIPs based on a six-channel data sequence: TEC, median TEC, lower bound TEC, upper bound TEC, positive polarity, and negative polarity. To leverage the power of deep learning image classification methods, we treated this data as a six-dimensional image. This approach yielded a maximum accuracy of 67.1 percent.**

I. Introduction

Seismo-Ionospheric precursors (SIPs), occurring before large earthquakes, have been well-documented in previous studies [1],[2],[3]. These notable anomalies typically appear a few days prior to major earthquakes [2],[4],[5]. In this study, we employ a deep learning technique to detect SIPs by analyzing temporal variations in the ionosphere. The experiment is focused on the Japan region due to its high frequency of earthquakes (over 1000 events with magnitudes larger than 5.0) between 1999 and 2021.

II. Data analysis

The Center for Orbit Determination in Europe (CODE) utilizes spherical harmonics expansion and VTEC (vertical total electron content) observations from several hundred ground-based GNSS receiving stations to generate the global ionospheric map (GIM) [6]. The GIM offers a comprehensive global distribution of ionospheric VTEC, covering latitudes from 87.5 degrees to -87.5 degrees and longitudes from -180 degrees to 180 degrees. It provides a latitude resolution of 2.5 degrees and a longitude resolution of 5 degrees. With hourly updates, the GIM enables monitoring of global ionospheric variations and analysis of the ionospheric horizontal structure. For our analysis, we have selected the location at 36.5N and 142E for GIM TEC assessment.

For deep learning training, we utilize six categories: TEC, median TEC, lower bound TEC, upper bound TEC, positive polarity, and negative polarity. Each category is structured as a matrix with dimensions of 24 hours * 30 days. The 24-hour resolution data represents the TEC values for each hour of the day, while the 30-day duration signifies the data collected in the 30-day period leading up to the earthquake. The TEC category represents the GIM TEC at the selected location, providing the ionospheric conditions prior to the earthquake. The median TEC represents the median TEC value calculated from the previous 15 days, serving as a reference state for the ionosphere. The upper bound and lower bound TEC correspond to the boundaries of ionospheric anomaly TEC values, as defined by the following equations,

$$UB=M+k(UQ-M).$$

$$LB=M-k(M-LQ).$$

, which UB is TEC upper boundary, LB is TEC lower boundary, M is the median TEC of the previous 15 days, UQ is the third quartile TEC of the previous 15 days, LQ is the first quartile TEC of the previous 15 days, and k is threshold constant. The positive and negative polarities represent ionospheric states that exceed or fall below the upper or lower boundary, respectively, indicating positive and negative ionospheric anomalies. The six categories utilized in this study consist of the ionospheric state (TEC), ionospheric reference (median TEC, lower bound TEC, upper bound TEC), and ionospheric anomaly polarity (positive polarity, negative polarity). This comprehensive set of categories is

well-suited for training a deep learning model to accurately classify SIAs features.

III. Experiment and Methods:

1. Proposed Method:

This study explores the application of EfficientNetV2 [7], a highly advanced image classification method, for a binary classification task. The original EfficientNetV2 model was slightly modified to suit the specific requirements of this task. The modifications included adjusting the "strides" parameter in the first stem convolution and the first block to one. This modification allows the model to capture more detailed information in the initial stages before down sampling, enhancing its ability to extract relevant features. To mitigate overfitting, the model was equipped with high dropout regularization layers. For binary classification, the sigmoid function was employed as the activation function in the final output layer.

To compare the performance of the modified EfficientNetV2 with other methods, a range of models was selected, varying from lightweight to highly effective. The lightweight model chosen was MobileNetV3-Large [8], while popular methods such as VGG16 [9], DenseNet121 [10], and ResNet50 [11] were also included. Additionally, ConvNext [12] and ResNet-RS [13], which are state-of-the-art convolutional neural networks for classification tasks, were incorporated. By including these diverse models, the study aimed to provide a comprehensive assessment of the performance of the modified EfficientNetV2 in comparison to established and state-of-the-art approaches for classification tasks.

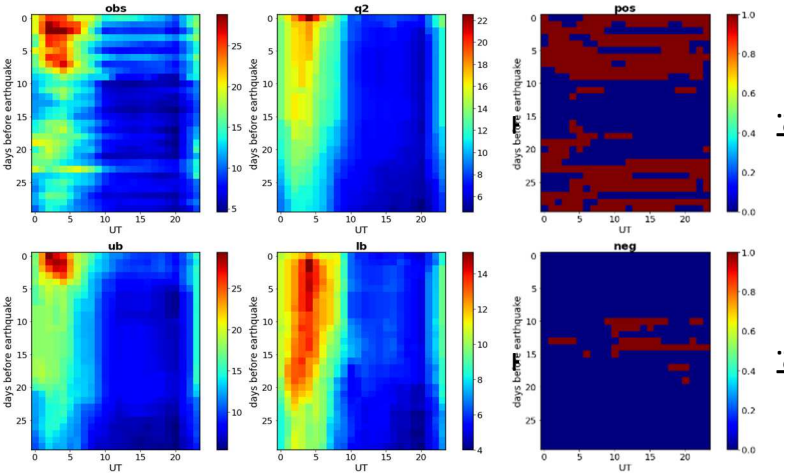
2. Experiments:

2.1. Data Preprocessing:

For this study, we conducted a random selection of 2356 samples. These samples were divided equally into two groups: normal events and SIPs events. Each sample is represented by a 30x24x6 array, indicating a time series with six channels. To approach the problem as an image classification task, we treated the time series data as a 6-channel image array. This means that each channel of the time series corresponds to a particular aspect or feature, much like the different color channels in an image.

To input the data into the model, we resized the 30x24 array to 32x32. When resizing, we added zero padding to maintain the aspect ratio between the width and height of the data. By resizing the data to 32x32, we transformed the time series to a format compatible with image-based models so that they could take advantage of the image classification capabilities of convolutional neural networks (CNNs).

Figure 1. A sample of the dataset (30x24x6)



2.2. Data augmentation:

CutMix [14] and Mixup [15] are widely employed image augmentation techniques in deep learning for enhancing model performance and generalization. CutMix combines cutout [16] and Mixup methods by replacing a portion of an image with a patch from another, introducing diversity and perturbations. On the other hand, Mixup blends pairs of images and their corresponding labels. These techniques effectively improve model regularization and generalization, and they have found extensive usage in various computer vision tasks.

2.3. Data division:

We divided the data into three groups: Test, Validation, and Training. Initially, we randomly assigned 20 percent of the data to the test set. The next 10 percent were allocated to the validation set, while the remaining data were designated for the training set.

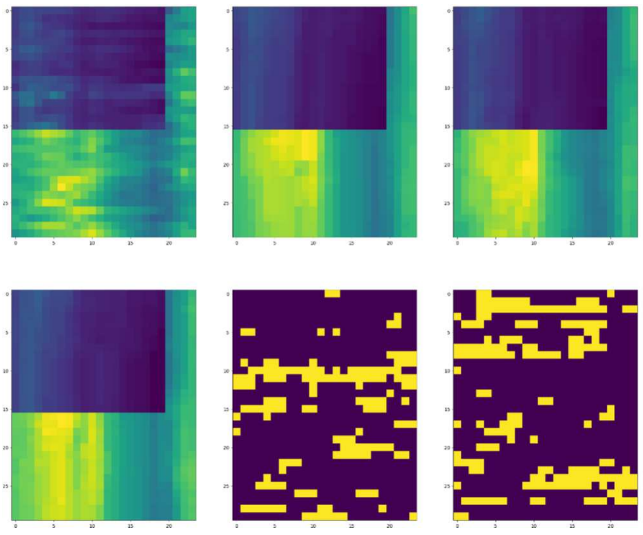


Figure 2. Cutmix augmentation

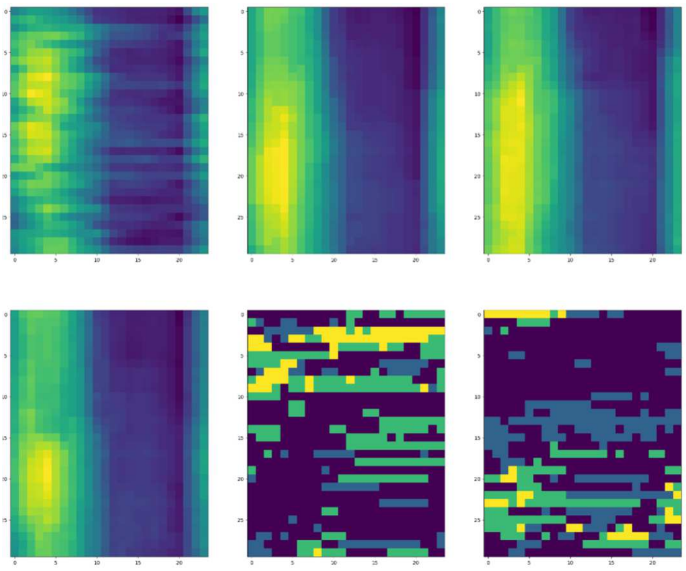


Figure 3. Mixup augmentation

2.4. Experiment setting:

In this study, we implemented the following approach: We initialized all models with pre-trained ImageNet weights and set the input size to 32x32x6. Each model underwent ten training iterations, with each iteration comprising 80 epochs. Throughout the training process, we utilized the Adam optimizer with a learning rate of 1e-4 and employed binary cross-entropy loss, which is commonly used for binary classification tasks. Following the ten training iterations, we selected the best-performing model and conducted a comparative analysis using the test data.

In the second experiment, we utilized the same settings and compared the utilization of CutMix and Mixup augmentations in combination with our method.

3. Results:

Table 1. Comparison with state-of-the art methods on Test Set

Models	ResnetRS50	ResNet50	VGG16	DenseNet121	Mobinet_v3	Efficient Net-v2	ConvNext
Accuracy	0,6208	0,6186	0,5826	0,6123	0,5530	**0,6716**	0,5000

Among the models evaluated in our study, the modified EfficientNet-v2 demonstrated the highest performance with an accuracy of 0.6716. It outperformed the other models, including Resnet-RS50 (accuracy of 0.6208) and its predecessor ResNet50 (accuracy of 0.6186). DenseNet121 achieved an accuracy of 0.6123, VGG16 had an accuracy of 0.5826, and MobinetV3-Large had an accuracy of 0.5530. However, it is worth noting that the ConvNext model did not effectively converge during the experiment.

Overall, our study demonstrates the superior performance of the modified EfficientNet-v2, followed by Resnet-RS50 and ResNet50, while noting the convergence issues encountered by the ConvNext model.

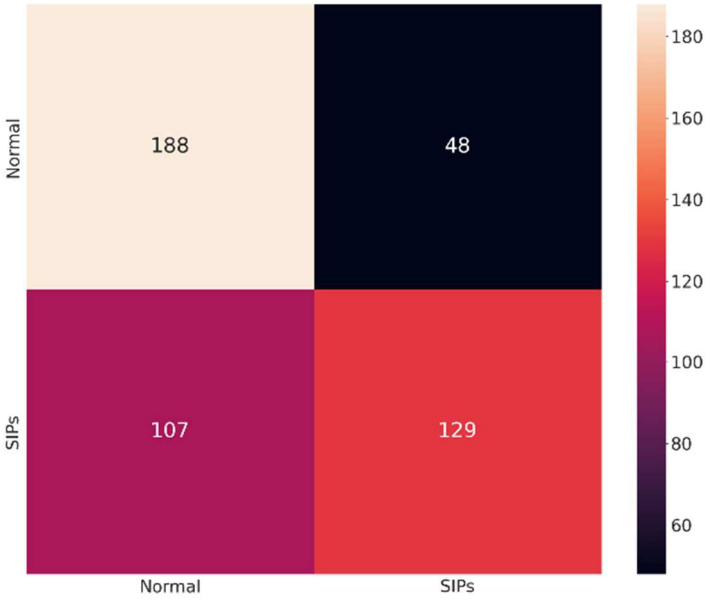


Figure 4. The confusion matrix

The Figure 4 present the confusion matrix of the efficientNet-v2.

Table 2. Comparison with the use of Cutmix and Mixup Augmentation

Num of Experiment	EfficientNet-V2 Cutmix+Mixup	EfficientNet-V2
1	**0.6504**	0,6398
2	0.6292	0,6335
3	0.6186	0,6356
4	**0.6504**	0,6102
5	0.6377	0,6398
6	0.6186	0,6229
7	0.6462	0,6483
8	0.6462	0,6144
9	0.6335	**0,6716**
10	0.6271	0,6314
Avg ± std	**0.6358 ± 0.012**	0.6348 ± 0.017

Table 2 demonstrates the significance of the CutMix and Mixup augmentation methods. While they did not directly improve the performance to achieve the new best result, they contributed to enhancing the stability of the performance. This improvement was observed not only in the average results but also in reducing the standard deviation, indicating a more consistent performance.

IV. Conclusions:

In this study, we propose an image representation

approach for SIPs time sequence data, achieving a promising performance of 0.67 despite the limited data availability. To the further enhance our methodology; we intend to explore enhancement techniques, self-supervised and unsupervised methods, as well as meta-learning approaches. These strategies aim to improve the model's capability to work with limited data and enable its application in real-world domains.

Reference

[1] Liu J Y, Chen Y I, Pulinets S A, et al. (2000) Seismo-ionospheric signatures prior to M>6.0 Taiwan earthquakes. Geophysical Research Letters, 27:3113-3116

[2] Liu J Y, Chen Y I, Chuo Y J, Tsai H F. (2001) Variations of ionospheric total electron content during the Chi-Chi earthquake. Geophys. Res. Lett. , 28:1383-1386

[3] Hayakawa M, Molchanov O A. (2002) Seismo Electromagnetics: Lithosphere Atmosphere Ionosphere Coupling. Tokyo: TERRAPUB, 477

[4] Chen Y I, Liu J Y, Tsai Y B, Chen C S (2004) Statistical tests for pre-earthquake ionospheric anomaly. Terr. Atmos. Ocean. Sci., 15:385-396

[5] Liu J Y, Chen Y I, Chuo Y J, Chen C S (2006) A statistical investigation of pre-earthquake ionospheric anomaly. J. Geophys. Res., 2006, 111, A05304, doi:10.1029/2005JA011333

[6] Schaer, S. (1999) Mapping and predicting the Earth's ionosphere using theGlobal Positioniong System,Geod. Geophys. Arb. Schweiz,59, 205 pp.

[7] Tan, Mingxing and Quoc V. Le. "EfficientNetV2: Smaller Models and Faster Training." ArXiv abs/2104.00298 (2021): n. pag.

[8] Howard, Andrew G. et al. "Searching for MobileNetV3." 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019): 1314-1324.

[9] Simonyan, Karen and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." CoRR abs/1409.1556 (2014).

[10] Huang, Gao et al. "Densely Connected Convolutional Networks." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 2261-2269.

[11] He, Kaiming et al. "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 770-778.

[12] Liu, Zhuang et al. "A ConvNet for the 2020s." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022): 11966-11976.

[13] Bello, Irwan et al. "Revisiting ResNets: Improved Training and Scaling Strategies." ArXiv abs/2103.07579 (2021)

[14] Yun, Sangdoo et al. "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features." 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019): 6022-6031.

[15] Zhang, Hongyi et al. "mixup: Beyond Empirical Risk Minimization." ArXiv abs/1710.09412 (2017)

[16] Devries, Terrance and Graham W. Taylor. "Improved Regularization of Convolutional Neural Networks with Cutout." ArXiv abs/1708.04552 (2017)



APSIPA
Signal and Information Processing